Volume 12 Number 23 1984 Nucleic Acids Research

Optimizing selection of restriction enzymes in the search for DNA variants

Ellen M.Wijsman

Department of Genetics, Stanford University, Stanford, CA 94305, USA

Received 27 February 1984; Revised and Accepted 1 November 1984

ABSTRACT

A model is developed for predicting the relative efficiencies of different enzymes for detecting DNA variants when such variants are the result of single base-pair changes. 71 enzymes are analyzed for this ability in human DNA. Their relative ranked efficiencies are influenced by the sizes of the probes used, and the size of the smallest detectable fragment produced.

INTRODUCTION

Restriction fragment length polymorphisms (RFLP'S) are becoming increasingly useful as genetic tools, but the search for such polymorphisms can be tedious. There are currently dozens of different restriction enzymes, all of which can potentially be used to define polymorphisms. It would be helpful to develop some guidelines for making rational choices from this long list, i.e., for maximizing the probability of finding a polymorphism with a minimum amount of work. Empirical estimates of the relative efficiencies of different enzymes in detecting polymorphisms would be the best since they do not depend on approximations or assumptions and they include all sources of experimental error. However, empirical estimates are only good if they are based on large amounts of data in particular organisms. They may not transfer well from one species to another, and do not predict the utility of a new enzyme in detecting RFLP'S. It is useful, therefore, to develop a model which can be used to make predictions of the relative efficiencies of enzymes in detecting RFLP'S.

To date, the only model which exists is one which uses the expected size-distribution of fragments resulting from digestion with a particular enzyme (1, 2) to aid in the choice of enzymes for detecting insertion or deletion RFLP'S (2). However, a large class of RFLP'S consists of point mutations. For example, only 4 of the 159 RFLP'S currently known in humans are insertions or deletions (3). It is desirable to define some properties which make enzymes useful in a search for polymorphisms which result from

single base-pair changes and to apply these properties to some commonly available enzymes. This is the topic we will discuss in this paper. In doing so, we will determine the relative abilities of enzymes in the available pool to detect RFLP's, and examine some properties of DNA composition which we need to make such predictions.

THE MODEL AND ASSUMPTIONS

The objective of developing the following model is to derive a quantity which can be used to predict the relative abilities of different enzymes to detect RFLP's. In the model we will assume that the probability of detecting a polymorphism is proportional to the probability of detecting a mutation, i.e., there is no distinction between mutations and substitutions. The problem of defining the efficiencies of enzymes for detecting RFLP's can therefore be restated as: what is the probability that a mutation will cause a detectable change in the restriction pattern? This can occur in two ways. A mutation in an existing restriction site may cause loss of the site, or a mutation in a potential site may cause a gain of a restriction site. A potential site is one which differs from a restriction site by exactly one base (4). Both of these types of events are potential sources of polymorphisms, and must be considered together to quantify the efficiency of a particular enzyme for detecting RFLP's.

The following six assumptions underly the model.

i) It is possible to predict the number of restriction and potential sites in a given length of DNA by using a simple measure of DNA composition.

ii) Polymorphisms occur with equal probability at each of the four bases and are randomly distributed along the DNA. This implies, for example, that if the frequency of an arbitrary base, B, is p_B, out of all polymorphisms a fraction p_B are expected to be the result of changes at bases B; p_B need not be .25.

iii) Mutations or polymorphisms occur independently of each other; gene conversion does not produce multiple base-pair changes within short regions, and the presence of one mutation does not cause a second mutation.

iv) There is no bias in the direction of mutation or substitution. For example, a change of A to T is as likely as a change of A to G. This assumption may not be true as is indicated by biases in interspecies comparisons of mtDNA substitutions (5), but as is discussed later, violations of this assumption do not alter the conclusions obtained from the model.

v) All fragments are resolvable: no 'hidden' bands exist and all

fragments of different length above a minimum size are distinguishable. It is not necessary, however, to assume that all size-classes of fragments are detectable.

vi) Modifications of DNA such as methylation do not affect the restriction pattern. This assumption is not always accurate for specific enzymes, but is adequate when an isoschizomer which is insensitive to such modifications is available, or when the DNA is unmodified. If neither condition is met in a particular situation, it may be necessary to eliminate from consideration enzymes which are sensitive to DNA modifications.

Of these assumptions, it will be possible to test (i) and the importance of (iv). A large compilation of DNA sequences in the GenBank(TM) files provides the data for comparison of the observed vs. expected number of sites, and for determination of the effect of violations of assumption (iv) on the probability of detecting mutations. Assumptions (ii) and (iii) are in principle testable in the same manner, i.e., by comparing the observed to the expected number of occurrences of an event. Sufficient data on mutation and substitution rates and their directions at the level of the nucleotide are not yet available to do this. Assumptions (v) and (vi) are not readily testable, but are necessary to bring the model to a workable level.

Now consider the following model. Suppose an arbitrary enzyme, E, has a recognition sequence of length n_E. Let a site for enzyme E be any sequence of n_E bases along the DNA. If the n_E bases match the recognition sequence precisely, the site is a recognition site; if there is one mismatch it is a potential site; if there is more than one mismatch it is an invisible site.

Let r_E be the probability that a site is a restriction site for enzyme E, and let $p_E(i)$ be the probability that it is a potential site for enzyme E which differs from the recognition sequence at the i^{th} base (the i^{th} potential site), counting from the 5' end. The number of sites is expected to follow a Poisson distribution with Lr_E restriction sites in L bases of DNA and $Lp_E(i)$ potential sites at base i in L bases of DNA.

At each restriction site, there are a number of bases which can change to produce a modified restriction pattern, while at each potential site there is only one such base. Let $\rho_E(i)$ be the probability that a variant at base i in a restriction site produces a loss in the restriction site, and let $\pi_E(i)$ be the probability that a variant at the i^{th} base in the i^{th} potential site produces a gain in a restriction site. If assumption (iv) is valid,

$\rho_E(i)=1$, $\pi_E(i)=1/3$ if the i^{th} base in the recognition sequence must be one of A, T, G, or C;

$\rho_E(i)=\pi_E(i)=2/3$ if the i^{th} base must be either of a particular pair of bases; (1)

$\rho_E(i)=1/3$, $\pi_E(i)=1$ if the i^{th} base must be one of a specific trio of bases;

$\rho_E(i)=0$ if the i^{th} base may be any base.

The total expected number of detectable variants resulting from a loss of restriction sites in L bases is then

$$T_{rE} = \mu L r_E \sum \rho_E(i), \quad (2)$$

where μ is the proportion of all bases which vary, and the total expected number of detectable variants which produce gains in restriction sites is

$$T_{pE} = \mu L \sum p_E(i)\pi_E(i). \quad (3)$$

If we ignore overlaps between restriction sites and potential sites, the total number of detectable variants in L bases is

$$\begin{aligned} T_E &= T_{rE} + T_{pE} \\ &= \mu L\{r_E \sum \rho_E(i) + \sum p_E(i)\pi_E(i)\}. \end{aligned} \quad (4)$$

Overlaps between sites are readily treated. The probability of finding two or more overlapping sites is generally exceedingly low, in particular for overlaps between restriction and potential sites of a single enzyme. Thus the estimates which result from ignoring overlaps are very close to those obtained by considering the overlaps. There are, however, a few readily identified pairs of enzymes for which overlap is significant. These will be noted.

The procedure for counting overlaps is as follows. An overlap can include two restriction sites, two potential sites, or one restriction and one potential site. To determine the expected number of mutations which are counted twice in pooling the results for more than one kind of site, one needs to enumerate all possible overlaps, to find the probability of each resulting combined sequence, and to find the number of mutations in the combined sequence which are counted twice. For example, the two restriction sites Taq I (5'TCGA) and Hga I (5'GACGC) overlap in the combined sequence 5'TCGACGC. Because of its length, this combined sequence has a much lower probability than either of the individual recognition sites. All mutations at bases 3 and 4 in the combined sequence are counted as losses in a restriction site for both enzymes. If we take the sum of the number of identifiable mutations obtained for each enzyme, we therefore have overestimated the number of observable mutations by counting some twice. To correct such a sum for mutations which are counted twice, one merely needs to subtract the total number of such mutations from the sum. The efficiencies presented in Table 1

for the different enzymes have been corrected for overlaps between potential and restriction sites.

Thus far, we have a model which predicts the (relative) number of identifiable variants within a fixed length of DNA if we assume that we can detect all fragments produced in a digest. However, the amount of DNA examined with a given probe varies for different enzymes, and since small fragments tend to get lost, we cannot detect all the fragments. Both of these factors will affect the total number of variants we may expect to see with a particular enzyme and make the above model only a first approximation to our desired result.

The loss of small fragments will affect the proportion of detectable variants in a length of DNA as follows. Let m be the maximum length of a fragment which cannot be detected for technical reasons. A variant will go undetected if either 1) the variant changes a site from a potential to a restriction site and the potential site occurs in a fragment of size m or less, or 2) the variant changes a site from a restriction to a potential site and occurs at a restriction site between two fragments whose sum is m or less. If f_p is the fraction of the total DNA which falls in fragments of length m or less, the expected number of lost variants, V_p, which change potential to restriction sites is by (3), $V_p = \mu f_p L \sum p_E(i)\pi_E(i)$. If f_r is the fraction of total DNA which falls in two contiguous fragments which total m bases of less, the expected number of lost variants, V_r, which change restriction to potential sites is by (2), $V_r = \mu f_r L r_E \sum \rho_E(i)$.

We find f_p and f_r as follows. Enzyme E generates fragments of random length y whose distribution is very close to the negative exponential (2), $Pr(y>Y)=\exp(-Yr_E)$, where Y is a particular fragment length and $\exp(x)=e^x$.

$$\begin{aligned} Pr(y) &= Pr(y=Y) \\ &= Pr(y>Y-1) - Pr(y>Y) \\ &= \exp(-Yr_E)\{\exp(r_E)-1\}. \end{aligned}$$

The fraction of all bases which occur in fragments of no more than m bases is

$$\begin{aligned} f_p &= \int_0^m y\ Pr(y)\ dy \ /\ \int_0^\infty y\ Pr(y)\ dy \\ &= 1 - \exp(-mr_E)\{1 + mr_E\}, \end{aligned}$$

and the fraction of all bases which occur in contiguous fragments which together are less than m bases is

$$\begin{aligned} f_r &= \int_0^m \int_0^{m-y} (y+z)\ Pr(y)\ Pr(z)\ dzdy \ /\ \int_0^\infty \int_0^\infty (y+z)\ Pr(y)\ Pr(z)\ dzdy \\ &= 1 - \exp(-mr_E)\{1 + mr_E + mr_E^2/2\}. \end{aligned}$$

So in a length of DNA, D_E, we can expect to see T_E' variants,

$$T_E' = D_E\{r_E\sum\rho_E(i) + \sum p_E(i)\pi_E(i)\} - \mu D_E f_p \sum p_E(i)\pi_E(i) - \mu D_E f_E r_E \sum\rho_E(i)$$
$$= \mu D_E\{r_E(1-f_r)\sum\rho_E(i) + (1-f_p)\sum p_E(i)\pi_E(i)\}.$$

D_E, however, is not constant for different enzymes. Bishop et al (2) show that the expected total length, D_E, of DNA potentially detectable by enzyme E with a probe of length L is

$$D_E = L+2/r_E.$$

Thus the expected number of variants we can expect to see with enzyme E is

$$T_E' = \mu(L+2/r_E)\{r_E(1-f_r)\sum\rho_E(i) + (1-f_p)\sum p_E(i)\pi_E(i)\} \quad (5)$$
$$= \mu RE_E \quad (6)$$

where RE_E is the relative efficiency of enzyme E and is proportional to the number of variants we can detect with the use of enzyme E.

TESTS OF ASSUMPTIONS

One of the more important assumptions in this analysis is that the number of restriction and potential sites in the DNA is predictable. The data used to test this are 95666 bases in 81 partial or complete sequences of human DNA from the GenBank(TM) sequences. The sequences are those of independent, unique sequence, coding regions. The observed number of restriction and potential sites in these sequences was determined for each of the enzymes in Table 1. The expected number of restriction and potential sites for each of the enzymes was computed with either the base frequencies or dinucleotide frequencies in these sequences. With the base frequencies, the probability, Q, of a site of length n is the product of the individual base frequencies, or

$$Q = \prod_{i=1}^{n} q_i$$

where q_i is the probability that the i^{th} base is correct in the site. With the dinucleotide frequencies, Q becomes

$$Q = q_1 \prod_{i=2}^{n} C_{i,i-1}$$

where $C_{i,i-1}$ is the conditional probability that the i^{th} base is correct given the $(i-1)^{th}$ base is correct. It was necessary to examine the effect of using the dinucleotide frequencies in addition to the base frequencies because of the observation that the dinucleotide frequencies are not those expected on the basis of the base frequencies (6, 7, 8, 9).

Figure 1 shows the fit of the observed numbers of restriction and potential sites for each enzyme in Table 1 to those predicted on the basis of the base frequencies (1a) and dinucleotide frequencies (1b). It is obvious that

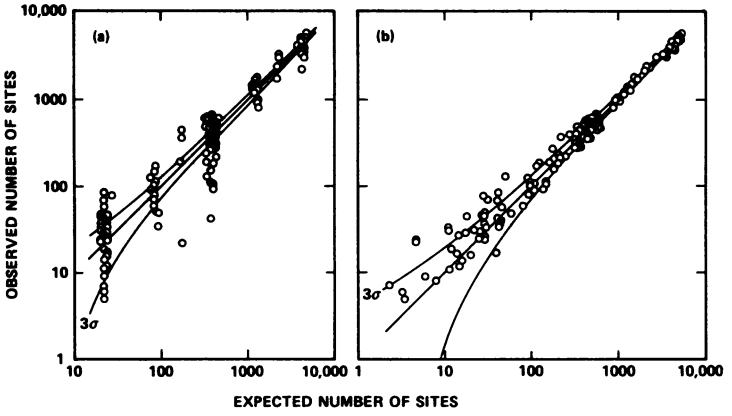

Figure 1. Observed vs. expected number of restriction and potential sites in 95666 bases of human DNA. For display purposes, all potential sites defined by each enzyme have been pooled into one number. Restriction and potential sites have not been pooled. a: Expectations computed with base frequencies. b: Expectations computed with dinucleotide frequencies.

the dinucleotide frequencies give a much better fit to the observed data than the base frequencies. For the base frequencies, the square root of the mean square error is 135.9, while for the dinucleotide frequencies it is 55.4. This is not the result of a few disparate observations: for only 107 out of 458 total restriction and potential sites is the observed number of sites closer to the number expected on the basis of the base frequencies than on the basis of the dinucleotide frequencies ($P<.0001$). It is worth noting, however, that even with the dinucleotide frequencies the observed numbers of sites are significantly non-random. 140 out of the 458 observed total sites are more than 3σ from the expected values computed with the dinucleotide frequencies (257/458 are more than 3σ from the expectations computed with the base frequencies). This is in agreement with earlier conclusions that much but not all of the apparent non-random distribution of sites disappears when predictions are made with the dinucleotide frequencies (4). The observed distribution of the trinucleotide frequencies is consistent with that predicted with the dinucleotide frequencies, making it unlikely that the use of the trinucleotide frequencies would produce substantial improvement in the fit of the observed to predicted restriction sites. For purposes of the current analysis, the residual non-random distribution is not sufficiently severe to have a major impact on the estimated relative efficiencies of the enzymes in detecting RFLP's.

Table 1. Relative efficiencies of enzymes for detecting RFLP's in human DNA under various experimental conditions.

Enzyme	Recognition[a] Sequence	MDF[b]	2 Kb probe: 200	800	12 Kb probe: 200	500	800
Bgl I	GCCN$_5$GGC		22.0	21.3	46.9	46.6	45.5
Aha II	AAATTT		28.1	26.8	69.9	68.8	66.9
Bst XI	CCAN$_6$TGG		29.0	27.5	75.8	74.3	71.9
Sph I	GCATGC		29.1	28.5	60.1	59.6	58.9
Apa I	GGGCCC		29.4	27.9	75.2	73.8	71.5
Bam HI	GGATCC		29.4	28.6	64.3	63.7	62.7
Sst I	GAGCTC		29.4	28.7	64.3	63.7	62.7
Xba I	TCTAGA		29.5	28.9	60.5	60.0	59.3
Aha III	TTTAAA		29.6	28.9	63.5	63.0	62.1
Eco RI	GAATTC		29.7	29.1	61.7	61.2	60.5
Bcl I	TGATCA		29.8	29.1	65.8	65.2	64.1
Bst EII	GGTNACC		29.9	29.5	57.9	57.6	57.0
TthIII II	GACN$_3$GTC		30.1	29.9	50.0	49.9	49.7
Xmn I	GAAN$_4$TTC		30.1	29.4	65.0	64.5	63.5
Hind III	AAGCTT		30.1	28.9	74.0	72.8	71.0
Bal I	TGGCCA		30.3	28.8	78.1	76.6	74.2
Bgl II	AGATCT		30.3	29.5	67.2	66.6	65.5
Nco I	CCATGG		30.4	29.1	76.3	75.0	72.9
Nde I	CATATG		30.4	30.1	57.4	57.2	56.7
Hpa I	GTTAAC		30.5	30.3	50.5	50.3	50.1
Pvu II	CAGCTG		30.5	29.1	76.3	75.0	73.0
Pst I	CTGCAG		30.5	29.1	76.3	75.0	73.0
Sau I	CCTNAGG		30.6	28.9	80.3	78.7	76.0
Stu I	AGGCCT		30.6	29.1	79.4	77.9	75.4
Eco RV	GATATC		31.1	30.9	52.1	52.0	51.7
Kpn I	GGTACC		31.3	31.0	54.3	54.1	53.8
Sma I	CCCGGG		31.5	31.2	56.5	56.3	55.9
Nae I	GCCGGC		32.1	32.0	49.1	49.1	48.9
Bde I	GGCGCC		34.0	33.9	52.0	51.9	51.8
Xho I	CTCGAG		34.9	34.7	54.8	54.8	54.6
Mst I	TGCGCA		35.4	35.3	54.4	54.3	54.2
Cla I	ATCGAT		36.1	36.0	50.1	50.1	50.0
Acc I	GTSWAC		37.0	35.2	94.8	93.0	90.1
Hph I	GGTGA		37.8	31.3	131.2	121.8	108.7
Sal I	GTCGAC		37.8	37.8	48.8	48.8	48.7
Hind II	GTYRAC		38.3	35.1	111.8	108.1	102.5
Sst II	CCGCGG		38.4	38.4	47.4	47.4	47.4
Xma III	CGGCCG		38.4	38.4	47.4	47.4	47.4
Sfa NI	GATGC		38.5	31.9	133.9	124.3	110.9
HinGU II	GGATG		39.0	30.7	142.7	129.7	112.3
Hga I	GACGC		39.7	38.8	86.7	85.9	84.6
Nci I	CCZGG		39.8	21.5	165.2	129.7	89.3
Mbo II	GAAGA		40.1	30.5	150.2	134.6	114.2
Nru I	TCGCGA		40.1	40.1	47.1	47.1	47.1
Hgi CI	GGYRCC		41.1	34.8	139.7	130.7	118.0
Hgi AI	GQGCQC		41.5	33.9	146.5	135.2	119.6
Ava I	CYCGRG		41.7	38.1	122.1	118.0	111.7
Acy I	GRCGYC		41.9	40.7	95.8	94.7	92.9
Hae II	RGCGCY		42.3	40.1	111.0	108.8	105.1
Xor II	CGATCG		42.8	42.8	49.8	49.8	49.8

Table 1 (cont)

Enzyme	Recognition Sequence	MDF:	2 Kb probe 200	800	12 Kb probe 200	500	800
Hgi JII	GRGCYC		45.7	31.2	183.3	156.8	125.0
Tha I	CGCG		49.7	46.2	138.2	134.4	128.5
Mlu I	ACGCGT		50.8	50.8	56.8	56.8	56.8
Bbv I	GCQGC		52.5	24.0	236.6	170.5	108.4
Ava II	GGQCC		53.4	27.0	237.7	178.1	120.0
Hha I	GCGC		55.0	39.9	212.0	186.2	153.6
Sau 96I	GGNCC		56.9	0.4	296.1	83.8	1.9
Fnu4H I	GCNGC		57.3	14.4	277.2	155.9	69.4
Rsa I	GTAC		57.4	32.9	246.7	195.6	141.6
Msp I	CCGG		58.3	30.4	256.5	195.2	133.5
Taq I	TCGA		59.5	40.0	241.1	204.8	162.0
Scr FI	CCNGG		61.2	3.9	318.4	106.1	20.4
Eco RII	CCQGG		61.2	14.8	302.7	161.5	72.9
Hae III	GGCC		61.7	5.1	319.7	113.6	26.4
Mbo I	GATC		63.0	17.1	308.4	173.8	83.5
Hinf I	GANTC		63.7	14.2	317.1	163.4	70.5
Eco RI*	AATT		63.9	7.9	326.7	133.9	40.3
Mnl I	CCTC		64.0	2.2	339.2	89.4	11.9
Dde I	CTNAG		64.9	6.1	339.2	117.2	31.9
Alu I	AGCT		65.1	6.9	338.6	122.8	35.6
Eco R1'	RRATYY		75.5	6.8	391.0	141.0	35.2

[a] [b] Code: Y=T/C; R=A/G; Q=A/T; S=A/C; Z=G/C; W=G/T; N=any base.
MDF: Minimum detectable fragment size in base pairs.

The GenBank(TM) sequences also allow examination of the impact of violations of assumption (iv) on predictions of efficiencies of enzymes in detecting RFLP's. A search of 80374 bases for each enzyme assigned each base into one of three categories. The first category is bases for which no mutation will change the restriction pattern and the second is bases for which any mutation will change the restriction pattern. The third category is those bases for which some mutations will change the restriction pattern; in this category, a count was also kept of each original base and all identifiable changes. The fraction of all mutations which could change the restriction pattern for each enzyme was then computed assuming either 1) equal mutation rates from any base to any other base (.33μ) or 2) a transition bias of 0.95μ where μ is the probability that a particular base mutates. This bias was chosen because of the observation that mitochondrial DNA shows a substitution bias of approximately 95% (5, 10, 11).

If the presence of a transition bias presents a serious problem, then the number or fraction of mutations actually detectable with a particular enzyme is expected to differ from that predicted under the assumption that no such bias exists. The above search allows us to assess the importance of a

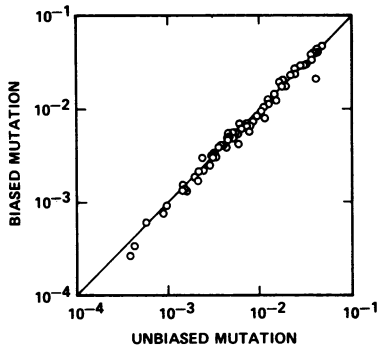

Figure 2. Similarity of fraction of mutations detectable under two assumptions about relative mutation rates. Fraction of detectable mutations assuming no bias in rates and direction of mutation is plotted along ordinate; fraction assuming a 95% transition bias is plotted along axis. 71 enzymes from Table 1 are plotted. Fraction of detectable mutations is calculated on the basis of 80374 bases of human DNA.

possible transition bias in predicting the number of detectable RFLP's. For the enzymes in Table 1, Figure 2 shows the fraction of detectable mutations in the 80374 bases under the assumption that there is no transition bias plotted against the fraction detectable under the assumption that the transition bias is 95%. For any given enzyme, the fraction of detectable mutations is almost identical in the two cases (ρ=.98). This similarity becomes even stronger as the results from several enzymes are pooled in the search for RFLP's. If one were to use all the enzymes in Table 1 to search the 95666 bases of human sequence DNA mentioned earlier, the empirical probability of detecting a mutation (if all fragments can be detected) is 45.2% with or without the transition bias. In other words, even a very strong transition bias has little effect on the probability of detecting a variant, especially if several enzymes are employed in the search. Since most experiments will use more than one or two enzymes, it seems likely that the validity of assumption (iv) is not terribly important for the current problem.

RELATIVE EFFICIENCY OF THE ENZYMES

The following investigation of the predicted efficiencies of different enzymes in detecting RFLP's is based on the dinucleotide frequencies. Let relative efficiency be a number which is proportional to the expected number of observable variants which will produce a detectable change in the restriction pattern produced with a particular enzyme. The expected number of observable variants produced with a specific enzyme in a comparison of DNA from two individuals is easily found by multiplying the relative efficiency by the fraction of variable bases, μ, as equation (6) shows. This requires an estimate of μ which may be crude at best. However, an estimate of μ is

not required for choosing one enzyme over another in a search for RFLP's.

The relative efficiency is influenced not only be the length and composition of the recognition sequence, but also by two experimental factors: the size of the probe and the size of the smallest detectable fragment. These conditions vary among experiments, making it impossible to rank the relative efficiencies in a fashion which will be valid under all experimental conditions. We will, however, examine some more likely situations for their effects on the efficiencies.

Probes vary widely in length. Inspection of Table 1 shows the effect on the relative efficiency of increasing the probe size from 2 Kb to 12 Kb. This increase causes an approximate doubling of the efficiency of enzymes with infrequent recognition sites, and about a six-fold increase in the efficiency of enzymes with frequent recognition sites. It has very little effect on the ranked order of the efficiencies.

The minimum detectable fragment size is of considerably more importance than the probe size in determining the ranked efficiencies of the enzymes. Enzymes with frequent restriction sites will produce a large number of small fragments, many of which may fall below the detectable threshold. This will effectively reduce the total amount of DNA which is under examination; the amount of reduction will be a function of the minimum detectable fragment size and the frequency with which an enzyme cuts the DNA. Figure 3 and Table 1 show the effect on the relative efficiencies of enzymes of increasing the minimum detectable fragment size. This has considerable effect on the efficiencies of the enzymes with frequent recognition sites: many of these become relatively poor at detecting variation if small fragments are not detectable.

The final factor which may affect the efficiency is overlap of sites. In many experiments it is necessary to try several enzymes before a polymorphism appears. An approximate estimate of the efficiency of a group of enzymes is the sum of the individual efficiencies. One might assume that certain pairs of enzymes might overlap considerably in the polymorphisms they recognize becasue of some significant overlap in their recognition sequences. This would cause the overall estimate of efficiency as calculated above to be an overestimate. In fact, pooling the estimates of efficiency of two enzymes with two or even three bases of overlap in the recognition sequences does not cause much of an overestimate in the joint efficiency of the enzymes. Table 2 lists those pairs of enzymes for which the joint probability of detecting a mutation in a defined length of DNA is less than 90% of the probability

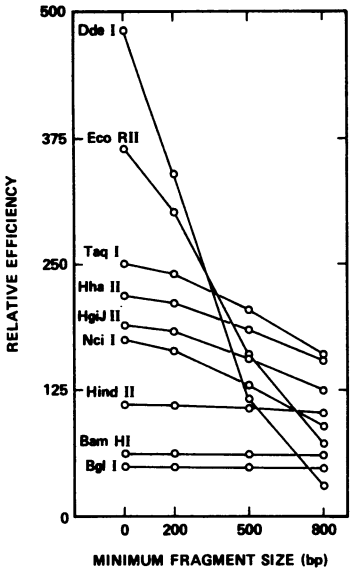

Figure 3. Relative efficiency of selected enzymes for detecting RFLP's with a 12 Kb probe and a range of minimum detectable fragment sizes.

obtained by taking the sum of the individual probabilites. The total number of such pairs is only about 1% of the total number of all of the pairs of the enzymes in Table 1. The pairs in Table 2 generally have almost identical recognition sequences. Even when one recognition sequence of a pair is completely contained in a longer recognition sequence, the true efficiency is on the order of 95% that predicted by the sum of the individual efficiencies because so much of the total information is supplied by the enzyme with the shorter recognition sequence. Therefore, the sum of the individual efficien-

Table 2. Pairs of enzymes from Table 1 for whic the percent of mutations detectable with both enzymes (percent overlap) is at least 10% of the total number of mutations detectable with either enzymes.

Enzyme Pair	Percent Overlap	Enzyme Pair	Percent Overlap	Enzyme Pair	Percent Overlap
Acc I/Sal I	14.5	Bde I/Hae II	19.5	HgiJ I/Sau 96I	10.9
Acy I/Bde I	23.6	Bde I/HgiC I	12.7	HgiJ I/Sst I	17.2
Acy I/Hae II	13.0	Eco RI*/Eco RI'	12.3	Hha I/Tha I	13.1
Acy I/Hga I	12.9	Eco RI'/Mbo I	10.3	Hind II/Hpa I	19.0
Ava I/Nci I	15.4	Eco RII/Nci I	10.3	Hind II/Sal I	10.6
Ava I/Sma I	22.4	Eco RII/Scr FI	39.9	HinGU II/Sfa NI	10.8
Ava I/Xho I	17.7	Hae II/Hha I	25.0	Msp I/Nci I	31.6
Ava II/Sau 96I	30.4	Hae III/Sau 96I	10.9	Msp I/Scr FI	14.8
Apa I/HgiJ II	22.5	HgiA I/HgiJ I	12.7	Nci I/Sma I	22.0
Apa I/Sau 96I	15.1	HgiA I/Sst I	22.4	Nci I/Scr FI	28.0
Bbv I/Fnu4H I	40.6	HgiC I/Kpn I	16.3		

cies is, to a close approximation, an adequate estimate of the efficiency of a group of enzymes.

A TEST OF THE MODEL

A test of the model can be made by comparing for many enzymes the observed number of polymorphic sites to a function of the total number of probes examined. Note that this requires the knowledge of failures as well as successes in finding polymorphisms with the enzymes. A few recent studies have reported both positive and negative results (12-36); these will be used here to test the model.

A simple test of the ability of the model to adequately predict the observed number of variable sites and a number which is the total rela-

Table 3. Observed vs. expected number of variable sites for 46 enzymes, where the observed number of variants are taken from references 12-36, and the expected numbers are those predicted either by regression of the observed number of variants onto the total relative efficiencies (LS), or as the average of all variable sites out of the sum of all total relative efficiencies (AV). Total relative efficiency (Total RE) for each enzyme is the sum over all the probes examined with that enzyme of the relative efficiencies calculated for the individual probes, assuming a minimum detectable fragment size of .5 Kb.

Enzyme	Obs. Var.	Total RE	Exp. Var. (LS)	Exp. Var. (AV)	Total Probes	Enzyme	Obs. Var.	Total RE	Exp. Var. (LS)	Exp. Var. (AV)	Total Probes
AhaII	0	403.67	0.88	0.80	5	KpnI	0	1316.18	2.88	2.61	34
AluI	1	237.72	0.52	0.42	9	MboI	9	2853.77	6.25	5.67	35
ApaI	0	968.43	2.12	1.92	11	MboII	0	833.96	1.83	1.66	7
AvaII	3	2398.99	5.25	4.77	14	MnlI	0	176.99	0.39	0.35	1
BamHI	9	2955.60	6.47	5.87	65	MspI	21	6891.18	15.09	13.69	95
BclI	0	2268.58	4.97	4.51	30	NcoI	0	958.51	1.47	1.34	23
BglI	4	1520.35	3.33	3.02	17	NdeI	0	672.87	1.47	1.34	22
BglII	11	3326.58	7.29	6.61	82	PstI	6	2825.64	6.19	5.61	57
BstEII	9	970.34	2.13	1.93	11	PvuII	6	3758.56	8.23	7.47	92
BstXI	0	913.17	2.00	1.81	23	RsaI	5	1847.60	4.05	3.67	10
DdeI	0	270.87	0.59	0.54	2	SalI	0	69.90	0.15	0.14	2
EcoRI	4	4217.87	9.24	8.38	84	SauI	0	292.16	0.64	0.58	1
EcoRII	1	944.80	2.07	1.88	24	Sau96I	2	16.34	0.04	0.03	1
EcoRV	1	742.48	1.63	1.48	24	ScrFI	1	14.50	0.03	0.03	1
Fnu4HI	0	312.73	0.68	0.62	1	SmaI	0	29.20	0.06	0.06	1
HaeIII	5	412.34	0.90	0.82	18	SphI	1	954.56	2.09	1.90	23
HgiAI	0	333.52	0.73	0.66	1	SstI	6	2074.05	4.54	4.12	42
HgiJII	0	2113.33	4.63	4.20	11	SstII	0	487.30	1.07	0.97	2
HindII	6	2227.46	4.88	4.43	19	StuI	3	1774.13	3.89	3.52	32
HindIII	6	3592.81	7.87	7.14	86	TaqI	17	7202.95	15.77	14.31	87
HinfI	0	569.21	1.25	1.13	19	XbaI	1	1616.90	3.54	3.21	46
HpaI	1	90.28	0.20	0.18	3	XhoI	0	65.00	0.14	1.13	2
HphI	2	2129.23	4.66	4.23	30	XmnI	2	1322.71	2.90	2.63	31

tive efficiency from equation (6), summed over all probes examined with each enzyme. This latter number is, of course, proportional to the expected number of variants with a constant of proportionality, μ. This total will vary widely from enzyme to enzyme, depending on the lengths and number of probes examined with each of the different enzymes. To test the model, it is necessary to have the sizes of the probes used for screening. These were available for most of the studies, but had to be estimated from a description of a partial set of inserts for a few studies (13, 16, 17, 31). For the latter, the inserts were taken to be 2 Kb (13, 16, 17) and 1 Kb (31). Table 3 gives the observed number of variable sites with the total relative efficiency for each enzyme. The correlation between these is $\rho=.84$, which is highly significant ($Z=10.7$, $P<<.0001$). The constant μ, as estimated by least squares from the data, is 2.19×10^{-3} when the line of best fit is forced through the origin. This gives a standard deviation of 2.46 for the error distribution of the observed RFLP's. A second estimate of μ is 1.99×10^{-3}, obtained by taking the total number of RFLP's and dividing by the sum of the total RE_E's. This gives a corresponding standard deviation for the error of the observed RFLP's of 2.52.

A second, more robust way of looking at the same data is to compute the ranked correlation coefficient, Kendall's Tau. This eliminates the possibility that the above results might in part be an artifact of the widely disparate number of probes examined. For the data in Table 3, Kendall's Tau is 0.54, again highly significant ($Z=4.77$, $P<<.0001$).

DISCUSSION

The model developed in this paper predicts the relative efficiencies of enzymes in detecting human RFLP's under a range of possible experimental conditions. Results not shown here indicate similar efficiencies in mouse DNA. The efficiencies of these enzymes may differ for detecting RFLP's in other organisms with different DNA composition. The results are dependent on certain assumptions. The two which were testable seem valid: that the number of restriction and potential sites is predictable with sufficient accuracy with the dinucleotide frequencies, and that bias in the direction of mutation is qualitatively unimportant. The unimportance of such a bias is probably the result of the composition of the recognition sequences: the equal number of pyrimidines and purines in most recognition sequences means the observed mutation rate should be roughly the average mutation rate, regardless of the amount of transition bias. There are, however, other

currently untestable assumptions which might change the results, although this is not likely in view of the good fit of experimental data to the predictions based on the model. If some of these are not correct, it should be easy to modify the estimates of efficiency to account for deviations from the assumptions. Unless the deviations are large, their effect on the relative efficiencies should not be major.

One untested assumption is that the number of variants in a given region of DNA is independent of the location of the region of DNA. It is quite possible that because of functional constraints, more variants exist outside than inside coding regions. Since probes generally identify coding regions, the analyses above might underestimate the efficiency of enzymes for recognizing polymorphisms determined with such probes and with enzymes which have infrequent recognition sites because these are more likely to cover a region of DNA which is outside the coding region. This bias is easy to correct. Equation (5) can be used twice: once for the number of bases inside and one for the number outside the coding region, with the latter multiplied by a constant to take account of this bias. The constant would simply be the ratio of the number of variable bases ouside to inside the coding region. A second, related assumption is that the DNA composition is constant inside and outside the region covered by the probe. Russell et al (37) found that different fractions of DNA (unique sequence, repetitive, etc.) in the genome had very similar, although not identical, compositions. It therefore seems unlikely that this would have a significant effect on the results.

Another assumption which might be an approximation is that variants occur at all bases (A, T, G, and C) with equal probability. It is possible that some variation exists in the mutation rates. Bird (38) suggested that there is an elevated mutation rate at 5-methylcytosine. This may mean that enzymes which have a CG in their recognition site or potential sites are more efficient at detecting variants than Table 1 suggests (see also 13, 39). A few such enzymes (e.g., Taq I and Msp I) are included in Table 3. Taq I does not show any obvious elevation in its ability to detect RFLP's; Msp I <u>may</u>, although as is argued below, the observed excess in RFLP's detected by Msp I is probably not significant. Unless such a bias is severe, it is unlikely to have a major effect on the results. The inclusion of the potential sites in the analysis means that even if a recognition sequence only contains two out of four bases, the set of bases which may cause detectable changes consists of all four bases, which would tend to smooth out the effects of irregularities in mutation rates among the bases. Again, such a bias, if it exists,

could easily be incorporated into the mutation rates used in the model.

None of the enzymes examined in Table 3, including Msp I, show a significant excess or deficiency in actual vs. predicted abilities to detect RFLP's. This conclusion can be obtained as follows. Although it has been suggested that Msp I exposes an unusually high number of RFLP's (40), we cannot directly test this hypothesis here because the current analysis includes some of the data used to make this suggestion. The best we can do is to ascertain whether or not any of the 46 enzymes examined detect significantly more or fewer RFLP's than predicted on the basis of the model. With 46 enzymes, we expect 2-3 by chance to show deviations at the 5% level even if the model is adequate for all enzymes. To reject the null hypothesis with 95% confidence, we must demand that the excess or deficiency be at least $\pm 3.29\sigma$ from expectation. This is the number of standard deviations corresponding to the confidence level of 0.05/46 which is required to reject any one of the 46 enzymes as fitting the model (41). None of the enzymes shows this much deviation. This test has very low power. It is, of course, in principle possible to test the null hypothesis with greater power for a preselected group of enzymes by collecting an independent data set; this will have to await further experiments.

Of the remaining assumptions, the problem of methylation of CG pairs cannot be treated analytically, although it is often possible to circumvent the problem by using an isoschizomer which is not sensitive to methylation. The problem of resolution of fragments and of hidden bands is not readily treated with the techniques presented here. It would however, be interesting in future studies to examine the problem of loss of information because of poor resolution or hidden bands. It may also be useful to modify the current approach to take into consideration information such as a preliminary restriction digest.

The analysis presented here demonstrates the usefulness of the model as a tool for predicting which enzymes will detect the most DNA variants under a defined set of conditions. A few general guidelines follow immediately from these results. First, because an increase in size of the cloned insert has such a marked effect on the number of variants expected to be detected with the probe, larger probes should be used if possible. This, for example, indicates that in the isolation of random unique-sequence probes (12, 13, 16, 17), methods of isolation which favor larger inserts are preferred. Second, the results indicate that much of the information on variability is potentially contained in fragments of low molecular weight. This suggests that

for enzymes which produce many small fragments, it would be beneficial to alter experimental conditions to increase detection of small fragments: change gel concentrations, take care to use freshly labelled nucleotides, etc. Finally, it is possible to use the model to predict the relative efficiencies in detecting RFLP's of a group of enzymes under a particular set of experimental conditions. This makes it possible to choose enzymes which are optimal for detecting DNA variants under the experimental conditions at hand. This should take some of the arbitrariness out of choosing enzymes which will be used to search for DNA variants.

ACKNOWLEDGEMENTS

I wish to thank J. Feder and M.J. Johnson for comments on the manuscript and discussion during the analysis of this problem, J. Richards for comments on the manuscript, and L.L. Cavalli-Sforza for the original stimulation of the problem. The GenBank(TM) sequences are publicly available: GenBank c/o Computer Systems Div., Bolt Beranek & Newman, Inc., 10 Moulton St., Cambridge, Massachusets 02238. This work was supported in part by NIH grant GM 28428.

REFERENCES

1. Bastie-Sigeas, F. and G. Lucotte (1983) Hum. Genet. 63:162-165.
2. Bishop, T.D., J.A. Williamson and M.H. Skolnick (1983) Am. J. Hum. Genet. 35:795-815.
3. Skolnick, J.H., H.F. Willard and L.A. Menlove (1984) Cytyogenetics and Cell Genetics 37:210-273.
4. Adams, J. and E.D. Rothman (1982) Proc. Natl. Acad. Sci. 79:3560-3564.
5. Brown, W.M., E.M. Prager, A. Wang and A.C. Wilson (1982) J. Mol. Evol. 18:225-239.
6. Josse, J., A.D. Kaiser and A. Kornberg (1961) J. Biol. Chem. 236:864-875.
7. Swartz, M.N., T.A. Trautner and A. Kornberg (1962) J. Biol. Chem. 237:1961-1967.
8. Subak-Sharpe, J.H., R.R. Burk, L.V. Crawford, J.M. Morrison, J. Hay and H.M. Keir (1966) Cold Spring Harbor Symp. Quant. Biol. 31:737-747.
9. Nussinov, R. (1980) Nucleic Acids Res. 8:4545-4562.
10. Aquadro, C.F. and B.D. Greenberg (1983) Genetics 103:287-312.
11. Greenberg, B.D., J.E. Newbold and A. Sugino (1983) Gene 21:33-49.
12. Feder, J., L. Yen, L. Wang, L. Wilkins, J. Schroder, E. Wijsman, N. Spurr, H. Cann, M. Blumenberg and L. Cavalli-Sforza (1984) manuscript submitted.
13. Barker, D., M. Schafer and R. White (1984) Cell 36:131-138
14. Michelson, A.M. and S.H. Orkin (1980) Cell 22:371-377.
15. Driscoll, M.C., M. Birch and A. Bank (1981) J. Clin. Invest. 68:915-919.
16. Aldridge, J., L. Kunkel, G. Bruns, U. Tantravahi, M. Lalande, T. Brewster, E. Moreau, M. Wilson, W. Bromley, T. Roderick and S.A. Latt (1984) Am. J. Hum. Genet. 36:546-564.

17. Dryja, T.P., J.M. Rapaport, R. Weichselbaum and G.A.P. Bruns (1984) Hum. Genet. 65:320-324.
18. Murray, J.C., R.M. Demopulos and A.G. Motulsky (1983) Proc. Natl. Acad. Sci. 80:5951-5955.
19. Antonarkis, S.E., D.D. Boehn, P.J.V. Guardina and H.H. Kazazian (1982) Proc. Natl. Acad. Sci. 79:137-141.
20. Jeffreys, A.J. (1979) Cell 18:1-10.
21. Johnson, J.J. (1984) unpublished Ph.D. dissertation.
22. Owerbach, D., G.I. Bell, W.J. Rutter, J.A. Brown and T.B. Shows (1981) Diabetes 30:267-270.
23. Goldfarb, M., K. Smimizu, M. Perucho and M. Wigler (1982) Nature 296:404-409.
24. Tuan, D., P.A. Biro, J.K. deRiel, H. Lazarus and B.G. Forget (1979) Nucleic Acis Res. 6:2519-2544.
25. Higgs, D.R., S.E.Y. Goodbourn, J.S. Wainscoat, J.B. Clegg and D.J. Weatherall (1981) Nucleic Acids Res. 9:4213-4224.
26. Beutler, E., W. Kuhl and C. Johnson (1981) Proc. Natl. Acad. Sci. 78:7056-7058.
27. Nussbaum, R.L., W.E. Crowder, W. Nyhan and C.T. Caskey (1983) Proc. Natl. Acad. Sci. 80:4035-4039.
28. Humphries, P., D. Barton, A.M. McKay, M.M. Humphries and B. Carritt (1983) Molec. Gen. Genet. 190:143-149.
29. Wilson, G.N., L.L. Szura, C. Rushford, D. Jackson and J. Erickson (1982) Am. J. Hum. Genet. 34:32-49.
30. Murray, J.M., K.E. Davies, P.S. Harper, L. Meredith, C.R. Muette and R. Williamson (1982) Nature 300:69-71.
31. Holden, J., J. Beckett, L. Mulligan, A. Phillips, N. Simpson, M. Partington, J. Hamerton, H.-S. Wang, L. Donald and B. White (1983) Am. J. Hum. Genet. 35:174A.
32. Kan, Y.W. and A.M. Dozy (1978) Proc. Natl. Acad. Sci. 75:5631-5635.
33. Naylor, S.L., A.Y. Sakaguchi, T.B. Shows, M.L. Law, D.V. Goeddel and P.W. Gray (1983) J. Exp. Med. 157:1020-1027.
34. Old, J.M. and J.S. Wainscoat (1983) Br. J. Haematol. 53:337-341.
35. Darby, J.K., J. Feder, M. Selby, V Riccardi, R. Ferrell, D. Siao, K. Goslin, W. Rutter, E.M. Shooter and L.L. Cavalli-Sforza (1984) manuscript submitted.
36. Daiger, S.P., N.S. Hoffman, R.S. Wildin and T.-S. Su (1984) Am. J. Hum. Genet. 36:736-749.
37. Russell, G.J., P.M.B. Walder, R.A. Elton and J.H Subak-Sharpe (1976) J. Mol. Biol. 108:1-23.
38. Fird, A.P. (1980) Nucleic Acids Res. 8:1499-1504.
39. White, R., M. Schafer, D. Baker, A. Wyman and M. Skolnick (1982) in Human Genetics, Part A: The Unfolding Genome (B. Bonne-Tamir, ed.) pp. 67-77.
40. Skolnick, M. and R. White (1982) Cytogenetics and Cell Genetics 32:58-67.
41. Snedecor, G.W. and W.G. Chochran (1980) Statistical Methods, seventh edition. The Iowa State University Press, Ames, pp. 166-167.