Some Combinatorial Problems of DNA Restriction Fragment Length Polymorphisms

KENNETH LANGE[1] AND MICHAEL BOEHNKE

SUMMARY

Recombinant DNA techniques provide a means of defining new polymorphisms at the DNA sequence level. Polymorphisms arise when individuals differ in the location and number of sites where restriction endonucleases can cleave their DNA. Each such site exhibits two possible states: one for the presence of a specific endonuclease recognition sequence, the other for its absence. The states of a system of adjacent sites can be revealed experimentally by cleaving a person's DNA into a set of fragments. For experimentally well-understood systems of sites, we consider problems of counting numbers of possible fragments, haplotypes, genotypes, and phenotypes, and the means of resolving phenotype-genotype ambiguities. The degree of polymorphism generated by such systems and the importance to gene mapping are discussed.

INTRODUCTION

The application of recombinant DNA methods promises to have a profound impact on human gene mapping [1–7]. In the past, only DNA polymorphisms leading to distinguishable differences in gene products were detectable. Such gene product polymorphisms include differences at the clinical or anatomical level, differences in the presence or absence of antigens, and differences in the electrophoretic mobility of enzymes. The new recombinant DNA methods detect differences at the DNA sequence level. Many of these differences occur in regions of DNA that do not code for proteins. Nevertheless, the basic combinatorial problems of counting genotypes and phenotypes do carry over into the new context. The purpose of this paper is to consider some of these combinatorial problems for

Received March 23, 1982; revised July 7, 1982.

This research was supported in part by the University of California, Los Angeles; National Institutes of Health Research Career Development Award K04-HD00307; and United States Public Health Service-National Research Service Award 07104.

[1] Both authors: Department of Biomathematics, University of California, Los Angeles, CA 90024.

experimentally well-understood systems. Our arguments will not be of direct relevance to those interested in strategies for defining new polymorphisms.

The current method for detecting polymorphisms at the DNA sequence level depends on cutting human DNA into small fragments, sorting the fragments on the basis of length, and then selectively imaging some of the fragments by autoradiography [1, 2, 4, 6]. Type II restriction endonucleases perform the first step in this program. These enzymes have the ability to recognize a specific sequence in a double-stranded DNA molecule and to cleave both strands of the molecule everywhere the sequence occurs. Typically, these sequences are 4–6 base pairs long. For example, the restriction endonuclease Eco RI recognizes the base-pair sequence GAATTC/CTTAAG.

The second step of the program is accomplished by electrophoresis of the resulting fragment. Shorter fragments naturally migrate farther than longer fragments. Once the fragments have separated adequately, they are denatured and bound to a solid support by the Southern transfer procedure [8]. Denaturation splits the double-stranded fragments into single-stranded fragments. In the final step of the program, the single-stranded fragments are incubated with radioactive complementary DNA probes. These human probes hybridize only with those fragments sharing extended DNA sequence homologies. If probes are employed that contain little repetitive DNA, hybridization is confined to those fragments that come from regions of the human genome overlapping the probe. As a consequence, only an extremely small fraction of the fragments will show up in autoradioagraphy. Polymorphisms arise when individuals differ in the location and number of sites where restriction endonucleases can cleave their DNA. While many such polymorphisms arise from insertions and deletions in the DNA, we shall consider only those polymorphisms due to base-pair changes within the sequences recognized by the restriction enzymes. In figure 1 we have depicted a DNA probe arranged alongside the homologous portion of a chromosome. The numbers represent sites where restriction endonuclease recognition sequences can occur. (Some authors use "site" to denote the presence of a recognition sequence, that is, in the sense of allele. We prefer to use site in the sense of locus, and "state" in the sense of allele.) If a site is polymorphic, then in some individuals the necessary recognition sequence is present while in others it is lacking. Thus, one chromosome may have the sequence GAATTC/CTTAAG at site number 1 and will be cleaved by Eco RI at this positive site. Another chromosome may have TAATTC/ATTAAG at site 1, a single base-pair substitution relative to the first chromosome, and will not be cleaved at this negative site. Thus, each polymorphic site considered separately

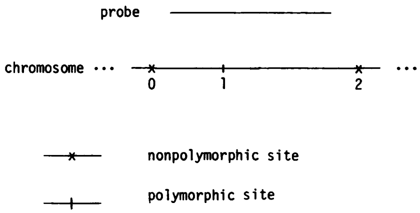


FIG. 1.—Single polymorphic site

has two possible states, one for the presence of a specific recognition sequence and one for its absence.

In figure 1, suppose sites 0 and 2 are nonpolymorphic Eco RI sites that fall beyond the ends of the probe. If site 1 is Eco RI polymorphic, fragment patterns for individuals fall into three phenotypes. Homozygotes for the presence of the recognition sequence at site 1 show two fragments corresponding to the intervals from 0 to 1 and 1 to 2. Homozygotes for the absence of this recognition sequence show the single fragment from 0 to 2. Heterozygotes show all three possible fragments. Thus, there is a one-to-one correspondence between genotype and phenotype in this simple case.

The interesting combinatorial problems arise when there are multiple, closely spaced polymorphic sites determined by one or more restriction endonucleases. For some purposes, it may be necessary or advantageous to perform digests and phenotyping using a single endonuclease at a time. This might be the case, for example, if optimal conditions of salt concentration or pH vary too much from endonuclease to endonuclease. On the other hand, digests using several different endonucleases simultaneously have certain economic advantages. Such multiple digests can also be particularly helpful in determining haplotype. Adopting the terminology used for the HLA complex, a combination of states at the various polymorphic sites along a region of a chromosome may be called a haplotype. In the HLA complex, genotypes at the various loci are typed independently, and haplotypes can be inferred only from the phenotypes of relatives. In contrast to HLA typing, restriction fragments can convey partial phase information in addition to the determination of the two chromosomal states at each site. This partial phase information comes about because each fragment reveals something about the common chromosomal origin of its midsites and endsites. In many cases, haplotypes can be determined fully without reference to the phenotypes of relatives. For instance, often a clever choice of the right combination of endonucleases for a digest can resolve all ambiguities.

The next section sets the stage for counting numbers of possible fragments, haplotypes, genotypes, and phenotypes under the assumption that all relevant endonucleases are used simultaneously. This assumption will be relaxed when we consider methods for haplotype determination. Counting problems for a series of digests with different subsets of endonucleases will be taken up in a later section. Departures from the idealized model will be considered in the DISCUSSION. Such departures include fragment identifiability problems and the application of multiple probes.

MODEL FORMULATION AND SOME EXAMPLES

The most general model, allowing for multiple polymorphic and nonpolymorphic sites, is illustrated in figure 2. As mentioned earlier, the different sites may correspond to different restriction endonucleases. The region of the chromosome that hybridizes with the probe is assumed to carry m runs of polymorphic sites ($m \geq 1$) with the ith run containing n_i such sites ($n_i \geq 0$). These runs of polymorphic sites are separated by $m - 1$ nonpolymorphic sites. The case $n_i = 0$ corresponds to adjacent nonpolymorphic sites. Sites in the region of the chromosome hybridizing

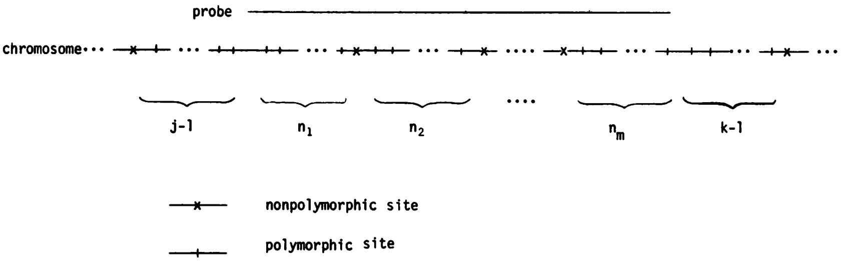


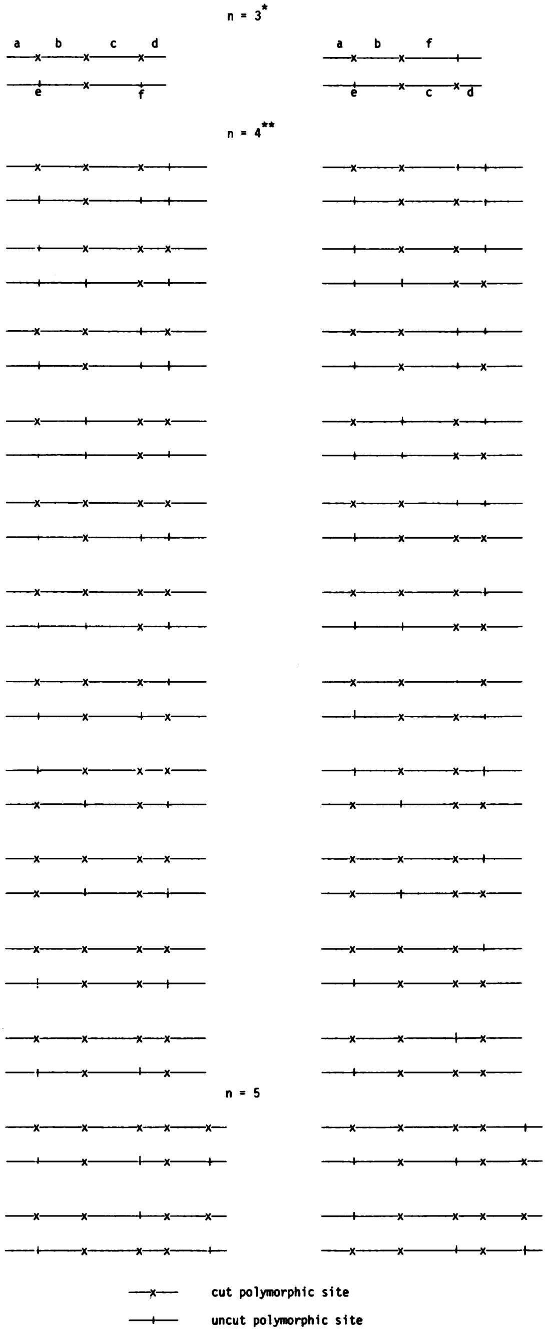
FIG. 2.—General model: multiple polymorphic and nonpolymorphic sites

with the probe we shall call "internal." Enzyme cuts at internal sites result in detectable fragments. The chromosomal regions flanking the probe extend to nonpolymorphic sites or chromosome ends. These flanking regions are assumed also to contain runs of $j - 1$ and $k - 1$ polymorphic sites ($j, k \geq 1$). Enzyme cuts at any of these external sites will generate detectable fragments only if no other external site closer to the hybridizing region is cut. It is for this reason that any site beyond the first external nonpolymorphic site will not be detected and need not be considered. When there are no external polymorphic sites ($j = k = 1$), we say we are dealing with fixed endpoints; otherwise, with variable endpoints. The simplest interesting case has a single internal polymorphic site and fixed endpoints ($j = k = m = n_1 = 1$), as illustrated in figure 1.

Let us consider in figure 3 some examples in which different genotypes (i.e., pairs of haplotypes) correspond to the same phenotype. In all of these examples, we assume fixed endpoints separated by n internal polymorphic sites. For simplicity, we omit the external nonpolymorphic sites. When $n = 3$, there is one pair of genotypes that corresponds to the same phenotype; when $n = 4$, there are 11 such pairs. For $n = 5$, all four displayed genotypes correspond to the same phenotype. There are an additional 75 pairs of ambiguous genotypes not shown.

Despite the possible ambiguities, it is clear that any given phenotypic pattern of fragments strongly limits the number of compatible genotypes. In practice, the number of polymorphic sites will usually be small, and most phenotypes will correspond to a single genotype. Even with an ambiguous phenotypic pattern of fragments, it may still be possible to determine the individual's genotype. Suppose one is given a phenotype corresponding to more than one genotype. For simplicity, let the nearest external site for each set of endonucleases be nonpolymorphic. The first example, $n = 3$, of figure 3 will do. From the resulting pattern of fragments, it will be clear which sites are heterozygous. In the present example, the first and third are. To determine the genotype of a person with the given phenotype, it suffices to digest his DNA with the set of restriction endonucleases appropriate to these heterozygous sites. This may not always be possible if a restriction endonuclease acts at more than one site. Suppose it is possible, however, to cleave his DNA at only those sites that are heterozygous.

Consider the homologous chromosomes of two individuals with the same phenotype but different genotypes. Digest their DNA with the enzymes appropriate to the heterozygous sites. Figure 4 depicts an example of this process. Proceeding



* For the case n = 3, if we identify the fragments with the letters a through f, it becomes clear that the two distinct genotypes produce the same phenotype.

** The two genotypes of each row produce the same phenotype.

FIG. 3.—Examples of genotypes that correspond to the same phenotypes

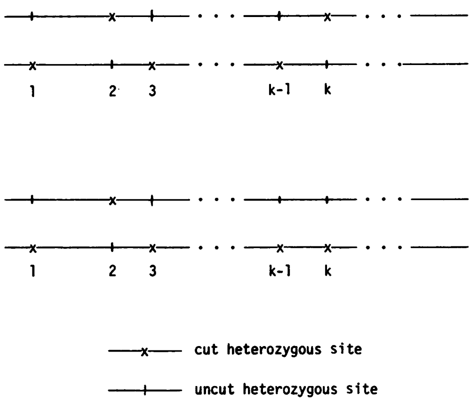


FIG. 4.—Haplotype determination by cleavage at heterozygous sites

from left to right, the recognition sequences at the heterozygous sites occur concordantly for awhile, some on the upper chromosome, some on the lower. Eventually, at heterozygous site k in the figure, the recognition sequences must be discordant. If this were not the case, the two individuals would have the same genotype. Examining the phenotypic fragment patterns of the two individuals, one will possess the fragment with endpoints $k - 1$ and k, while the other will not. Hence, the fragment patterns of the two individuals are distinguishable.

A variant of this haplotyping procedure would be to perform a sequence of digests with endonucleases appropriate to pairs of adjacent heterozygous sites. In figure 4, one could first use the pair of endonucleases appropriate to sites 1 and 2. For both pairs of chromosomes, this would establish that the recognition sequences were trans. The next digest would involve the endonucleases appropriate to sites 2 and 3. Continuing in this fashion, the $(k - 1)$th digest would involve the endonucleases appropriate to sites $k - 1$ and k. Examining the fragment patterns, it would then be clear that the recognition sequences at sites $k - 1$ and k were trans for the first pair of chromosomes and cis for the second.

The variant procedure can work when the original procedure fails. For example, in figure 4, suppose sites 1, 2, and 3 are Msp I, Msp I, and Taq I sites, respectively. (These endonucleases will be abbreviated Msp and Taq.) If Taq cuts both chromosomes at another site between 1 and 2, then digesting with Msp and Taq will not reveal the phase relation between sites 1 and 2. Digestion with Msp alone, however, will then reveal the phase relation. An Msp-Taq double digest will then reveal the phase relation between 2 and 3 and so forth. The variant procedure entails economic disadvantages, however, since it requires several digests per individual.

FRAGMENTS, HAPLOTYPES, AND GENOTYPES

We now revert to the assumption that all endonucleases are used simultaneously in a single digest. To count the total number of possible fragments, consider the case $m = 1$, where there are j external sites adjacent to one end of the probe, k external sites adjacent to the other, and a single run of n internal polymorphic sites. If $\binom{a}{b}$ is the binomial coefficient specifying the number of possible subsets of size b given a set of size a, then there are $\binom{n}{2}$ possible fragments with both endpoints internal, $n(j + k)$ possible fragments with one endpoint internal and one external, and jk possible fragments with both endpoints external. The total of $\binom{n}{2} + n(j + k) + jk$ fragments reduces to $\binom{n+1}{2} + j\binom{n+1}{2}$ when $k = 1$, and to $\binom{n+2}{2}$ when $j = k = 1$. It is easy to show that a single individual will exhibit no more than ½(3n + 2) of these $\binom{n+2}{2}$ possible fragments when n is even, or ½(3n + 3) when n is odd. When the number of runs $m > 1$, the total number of possible fragments is clearly

$$\binom{n_1+1}{2} + j(n_1 + 1) + \sum_{i=2}^{m-1} \binom{n_i+2}{2} + \binom{n_m+1}{2} + k\binom{n_m+1}{2} ,$$

since no internal nonpolymorphic site can serve as a midsite of a fragment. In other words, we can group fragments according to which run they originate from.

Counting the number of possible haplotypes is equally straightforward. At each of the

$$\sum_{i=1}^{m} n_i$$

internal polymorphic sites, there are two possible states. In addition, there are j and k possible choices for the two external cuts. Altogether there should be

$$l = j \cdot 2^{\sum_{i=1}^{m} n_i} \cdot k$$

possible haplotypes. These haplotypes generate $\binom{l+1}{2}$ genotypes. Table 1 gives the numbers of possible fragments, haplotypes, and genotypes in the simplest case $m = j = k = 1$.

PHENOTYPES

We next turn to the problem of counting phenotypes. Consider first the simplest case of fixed endpoints and no internal nonpolymorphic sites. Let P_n be the number of possible phenotypes with n internal polymorphic sites. We view these internal sites in sequence from left to right and number them 1 to n. It is convenient to label all phenotypes as either single-cut or double-cut. These terms refer to the number of chromosomes cut at the right-most site where at least one of the two homologous chromosomes is actually cut. Obviously, this right-most site is not the nth if neither chromosome is cleaved at site n. Let S_n and D_n be the number of single-cut and double-cut phenotypes, respectively. Clearly,

$$\begin{bmatrix} S_0 \\ D_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} ,$$

since the left external site is nonpolymorphic.

Now suppose we know

$$\begin{bmatrix} S_{n-1} \\ D_{n-1} \end{bmatrix}$$

and wish to calculate

$$\begin{bmatrix} S_n \\ D_n \end{bmatrix} .$$

Imagine adding the nth internal polymorphic site. To arrive at a single-cut phenotype for n polymorphic sites, we can begin with a single-cut phenotype for $n - 1$ sites, and at site n: (1) cleave neither chromosome, (2) cleave the most recently cut chromosome, or (3) cleave the other chromosome; or we can begin with a double-cut phenotype for $n - 1$ sites and at site n cleave one chromosome. Thus, $S_n = 3S_{n-1} + D_{n-1}$. Note that actions (2) and (3) above add different fragments to the overall phenotypic pattern. To arrive at a double-cut phenotype for n polymorphic sites, we can begin with a single-cut phenotype for $n - 1$ polymorphic sites and at site n cleave both chromosomes; or we can begin with a double-cut phenotype for $n - 1$ sites, and at site n: (1) cleave neither chromosome, or (2) cleave both chromosomes. Thus, $D_n = S_{n-1} + 2D_{n-1}$. In matrix notation,

$$\begin{bmatrix} S_n \\ D_n \end{bmatrix} = M \begin{bmatrix} S_{n-1} \\ D_{n-1} \end{bmatrix} \text{, where } M = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} .$$

Iterating this formula gives

$$\begin{bmatrix} S_n \\ D_n \end{bmatrix} = M^n \begin{bmatrix} S_0 \\ D_0 \end{bmatrix} = M^n \begin{bmatrix} 0 \\ 1 \end{bmatrix} ,$$

and, consequently,

$$P_n = S_n + D_n = [1 \ \ 1]\, M^n \begin{bmatrix} 0 \\ 1 \end{bmatrix} .$$

Table 1 lists the numbers of possible phenotypes for various numbers of internal polymorphic sites.

Counting phenotypes for the variable endpoint model is a straightforward extension of the previous argument. The only difference is in the number of ways to choose the endpoints. Define P_n^{jk} as the number of possible phenotypes assuming variable endpoints and no internal nonpolymorphic sites. By convention, $P_n^{11} = P_n$. In analogy to the fixed endpoint case, P_n^{jk} may be calculated by pre- and postmultiplication of M^n by vectors, these vectors giving the numbers of ways to end and start the phenotypes, respectively.

Given j external sites adjacent to one end of the probe, there are j detectably different ways to begin with a double-cut phenotype. As before, only the cuts

TABLE 1

SIMPLEST MODEL: FIXED ENDPOINTS, NO INTERNAL NONPOLYMORPHIC SITES

No. internal polymorphic sites	No. fragments	No. haplotypes	No. genotypes	No. phenotypes
0	1	1	1	1
1	3	2	3	3
2	6	4	10	10
3	10	8	36	35
4	15	16	136	125
5	21	32	528	450
6	28	64	2080	1625
7	36	128	8256	5875
8	45	256	32896	21250
9	55	512	131328	76875
10	66	1024	524800	278125
n	$\binom{n+2}{2}$	2^n	$2^{n-1}(2^n+1)$	$(1\ \ 1)\begin{pmatrix}3 & 1\\ 1 & 2\end{pmatrix}^n \begin{pmatrix}0\\ 1\end{pmatrix}$

closest to the probe on each chromosome will be detected. There are also $\binom{j}{2}$ distinct ways to begin with a single-cut phenotype because the chromosomes are at this stage indistinguishable. Thus,

$$\begin{bmatrix}S_0\\ D_0\end{bmatrix}$$

in the formula for P_n should be replaced by

$$\begin{bmatrix}\binom{j}{2}\\ j\end{bmatrix}.$$

To determine the number of ways we can end a phenotype, we must consider the n site single-cut and double-cut phenotypes separately. We assume there are k sites beyond the end of the probe. A phenotype that is double-cut after the first $j + n$ sites can be completed as either a double-cut or single-cut phenotype. There are k ways to double-cut and $\binom{k}{2}$ distinct ways to single-cut, for a total of $k + \binom{k}{2} = \binom{k+1}{2}$ total ways. A phenotype that is single-cut after $j + n$ sits can be completed in the same two ways. Again there are k ways to double-cut. However, because of the asymmetry imposed by the single-cut phenotype, there are $k(k - 1)$ distinct ways to cut at two different sites, for a total of $k + k(k + 1) = k^2$ ways. Thus $(1\ \ 1)$ is replaced in the previous formula by $[k^2\ \binom{k+1}{2}]$, and we have shown

$$P_n^{jk} = \left[k^2\ \binom{k+1}{2}\right] \begin{bmatrix}3 & 1\\ 1 & 2\end{bmatrix}^n \begin{bmatrix}\binom{j}{2}\\ j\end{bmatrix}.$$

At first glance, the formula for P_n^{jk} does not appear to be symmetric in j and k. This would be disturbing since the number of phenotypes must be the same whether we begin counting with j sites external to the probe and end with k or vice versa. The symmetry in j and k will be demonstrated shortly.

Determining the number of phenotypes for the complete model with both variable endpoints and internal nonpolymorphic sites requires only one additional fact: nonpolymorphic sites internal to the probe divide the overall counting problem into independent subproblems. With each internal nonpolymorphic site, the counting problem is restarted. Hence, the total number of phenotypes for the complete model is simply the product of the number of phenotypes for each run of polymorphic sites. With the notation of figure 2, the number of possible phenotypes is

$$P_{n_1}^{\,j1} \cdot \prod_{i=2}^{m-1} P_{n_i} \cdot P_{n_m}^{\,1k} \qquad (m \geq 2)\ .$$

To derive a simple recurrence relation for the numbers of phenotypes, it is convenient to introduce generating functions (see, for example, [9]). For fixed j and k define

$$\begin{aligned} Q^{jk}(s) &= \sum_{n=0}^{\infty} P_n^{jk} s^n \\ &= \left[k^2 \binom{k+1}{2}\right]\left\{\sum_{n=0}^{\infty}\begin{bmatrix}3 & 1\\ 1 & 2\end{bmatrix}^n s^n\right\}\begin{bmatrix}\binom{j}{2}\\ j\end{bmatrix}. \end{aligned}$$

For $|s|$ small, Neumann's lemma [10] implies

$$\sum_{n=0}^{\infty}\left\{s\begin{bmatrix}3 & 1\\ 1 & 2\end{bmatrix}\right\}^n = \left\{I - s\begin{bmatrix}3 & 1\\ 1 & 2\end{bmatrix}\right\}^{-1}.$$

Inverting this matrix explicitly, substituting and simplifying yields

$$\begin{aligned} Q^{jk}(s) &= \frac{1}{1 - 5s + 5s^2}\left[k^2 \binom{k+1}{2}\right]\begin{bmatrix}1-2s & s\\ s & 1-3s\end{bmatrix}\begin{bmatrix}\binom{j}{2}\\ j\end{bmatrix} \\ &= \frac{a^{jk} - b^{jk}s}{1 - 5s + 5s^2}\ , \end{aligned}$$

where $a^{jk} = \frac{1}{2}jk(jk + 1)$ and $b^{jk} = \frac{1}{4}jk(7 - j - k + 3jk)$. Multiplying the last equation for $Q^{jk}(s)$ by $1 - 5s + 5s^2$ and equating coefficients for the powers of s gives

$$\begin{aligned} P_0^{jk} &= \tfrac{1}{2}jk(jk + 1) \\ P_1^{jk} &= \tfrac{1}{4}jk(3 + j + k + 7jk) \\ P_n^{jk} &= 5(P_{n-1}^{jk} - P_{n-2}^{jk}) \qquad (n \geq 2)\ . \end{aligned}$$

Note that these formulas are symmetric in j and k. In the case of fixed endpoints ($j = k = 1$), we have

$$\begin{aligned} P_0 &= 1 \\ P_1 &= 3 \\ P_n &= 5(P_{n-1} - P_{n-2}) \qquad (n \geq 2) \ . \end{aligned}$$

It would be interesting to have a direct combinatorial interpretation of this second-order linear difference equation. None is obvious to us.

It is also of interest to consider the asymptotic behavior of the numbers of genotypes and the numbers of phenotypes. For notational simplicity, let $j = k = m = 1$. We showed that the number of genotypes for this fixed endpoint model is $G_n = 2^{n-1}(2^n + 1)$ for n internal polymorphic sites. Thus,

$$\lim_{n \to \infty} \frac{G_n}{G_{n-1}} = 4 \ .$$

Calculation of the corresponding

$$\lim_{n \to \infty} \frac{P_n}{P_{n-1}}$$

is achieved by appeal to the Perron-Frobenius theorem [11]. Since each of the entries of

$$M = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

is positive, $\lim_{n \to \infty} \lambda^{-n} M^n = M_\infty$, where λ is the dominant eigenvalue of M and M_∞ is some constant matrix with positive entries. It is then clear that

$$\lim_{n \to \infty} \frac{P_n}{P_{n-1}} = \lambda \ .$$

λ can be computed as the larger root of $\det(sI - M) = s^2 - 5s + 5 = 0$, where det represents the matrix determinant. Hence, $\lambda = \frac{1}{2}(5 + \sqrt{5}) \doteq 3.618$, and the number of phenotypes also increases geometrically. The difference in the asymptotic rates of increase is small, but the cumulative effect is significant, as illustrated in table 1.

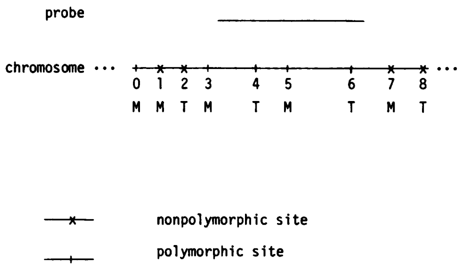
DIGESTS WITH ALL POSSIBLE SUBSETS

Instead of digesting with all pertinent endonucleases simultaneously, it may be advantageous to digest with different subsets of the endonucleases. For each subset, there arise counting problems that can be solved exactly as outlined in the previous sections. One subtlety is that the collection of relevant sites changes from one digest to the next. In figure 5, for example, it is possible to digest with Msp alone, with Taq alone, or with Msp and Taq together. To count fragments or phenotypes for an Msp digest, we must ignore the Taq sites, and vice versa. Now note an interesting complication. An Msp digest can give some of the same fragments and phenotypes as an Msp-Taq double digest. If we want to accurately count all fragments and phenotypes for all possible digests, we must correct for this repetition.

Before proceeding further, it is useful to make one more observation. Digests using only a subset of the endonucleases may expose external sites that lie beyond the closest nonpolymorphic sites for the full set of endonucleases. This is the case with the external Taq site 8 in figure 5. It is exposed when Taq alone is employed. For the Msp-Taq double digest, the Msp site 7 intervenes. We will call a site accessible if it is internal or if no nonpolymorphic site for the same endonuclease intervenes between it and the probe region. In figure 5, all sites depicted are accessible except site 0.

To count the number of possible fragments, the simplest procedure is to consider each relevant pair of accessible sites. Both sites of a pair can be external provided that they lie on opposite sides of the probe region. Thus, the pair of sites 2, 3 is disallowed. Each pair can generate a fragment only if there are no intervening nonpolymorphic sites between them for the relevant endonuclease or pair of endonucleases. For instance, the fragment 5–7 is possible while 5–8 is not. In figure 5, there are 19 possible fragments.

The problem of counting possible haplotypes is partially a definitional one. We take the view that any two haplotypes that are experimentally indistinguishable should be considered the same. For instance, if there were another external Msp site, say 3*, between sites 3 and 4 in figure 5, then no digest or combination of digests could distinguish a haplotype with both 3 and 3* Msp positive from a haplotype with 3* Msp positive and 3 Msp negative. In general, the most that can be detected for each endonuclease is the location of the two positive sites closest to the probe region. From this perspective, the number of possible haplotypes is simply the product of the number of haplotypes for each endonuclease alone. In figure 5, there are $2 \cdot 2^1 \cdot 1$ Msp haplotypes and $1 \cdot 2^2 \cdot 1$ Taq haplotypes. This gives $4 \cdot 4 = 16$ possible haplotypes and $\binom{16+1}{2} = 136$ possible genotypes.



M stands for Msp site and T for Taq site. ·

FIG. 5.—Digests using different subsets of endonucleases

The last and hardest problem is to count phenotypes. It is possible to correct for errors in counting phenotypes multiple times by applying an inclusion-exclusion (or sieve) formula from elementary combinatorics [9, 12]. To display this formula, it is necessary to introduce more notation. Let S be the full set of endonucleases. The family of all subsets of S will be denoted $\mathscr{P}(S)$. Thus, $B\epsilon\mathscr{P}(S)$ if and only if B is a subset of S. Let ϕ_B be the set of phenotypic patterns possible for a digest using the subset $B\subset S$ of endonucleases. If the function $m(\cdot)$ counts numbers of phenotypes, then the quantity we wish to calculate is

$$\overline{m}[\mathscr{P}(S)] = m(\bigcap_{B\epsilon\mathscr{P}(S)} \phi_B) .$$

According to the classical inclusion-exclusion principle,

$$\begin{aligned}\overline{m}[\mathscr{P}(S)] &= \sum_{\mathscr{C}}(-1)^{|\mathscr{C}|+1}\, m(\bigcap_{C\epsilon\mathscr{C}} \phi_C) \\ &= \sum_{\mathscr{C}}(-1)^{|\mathscr{C}|+1}\, \underline{m}(\mathscr{C}) ,\end{aligned}$$

where

$$\underline{m}(\mathscr{C}) = m(\bigcap_{C\epsilon\mathscr{C}} \phi_C) .$$

In this formula, $\mathscr{C}$ ranges over all collections of subsets, $|\mathscr{C}|$ denotes the number of subsets in the collection $\mathscr{C}$, and C is a typical subset of S in $\mathscr{C}$. For example, a possible $\mathscr{C}$ is {{Msp}, {Msp,Taq}}, with $|\mathscr{C}| = 2$ and one of the two subsets C being {Msp}. Note that if S has q elements, then the 2^q subsets in $\mathscr{P}(S)$ generate 2^{2^q} collections $\mathscr{C}$.

To apply the inclusion-exclusion formula, it is crucial to be able to compute $\underline{m}(\mathscr{C})$. First, define

$$D = \bigcap_{C\epsilon\mathscr{C}} C ,$$

$$E = \bigcap_{C\epsilon\mathscr{C}} C\backslash D .$$

D is the subset of endonucleases common to all the different subsets $C\epsilon\mathscr{C}$. E is the residue. For the specific collection $\mathscr{C}$ above, $D = \{\text{Msp}\}$, $E = \{\text{Taq}\}$. It is convenient to label sites as D sites or E sites depending on whether the corresponding endonuclease belongs to D or E.

A phenotype cannot be in

$$\bigcap_{C\epsilon\mathscr{C}} \phi_C$$

if it contains a fragment with an endsite that is an E site. Thus,

$$\bigcap_{C\in\mathscr{C}} \phi_C$$

is automatically empty if there is a nonpolymorphic internal E site. Similarly, it is empty if to the left (right) of the probe region, a nonpolymorphic E site occurs before any D sites. Finally, it is empty whenever D is empty.

If

$$\bigcap_{C\in\mathscr{C}} \phi_C$$

is not automatically empty, $\underline{m}(\mathscr{C})$ can be computed by the following steps: (1) Eliminate from consideration all external D sites that lie beyond the two closest nonpolymorphic external E sites. (2) Force the two extreme external D sites that remain to be nonpolymorphic. (3) Eliminate all E sites. (4) Finally, count the possible phenotypes for a D subset digest by the methods of the previous section (PHENOTYPES).

Let us illustrate these steps for figure 5 and the choice of $\mathscr{C}$ above. Step (1) forces the elimination of site 1 and the inaccessible site 0. Step (2) turns site 3 into a nonpolymorphic site. Step (3) eliminates sites 2, 4, 6, and 8. Step (4) produces a count of three phenotypes for the remaining sites 3, 5, and 7.

Table 2 lists the nonzero contributions to the sum for $\overline{m}[\mathscr{P}(S)]$. Most collections of subsets make no contribution because D is empty. The collection {{Taq}, {Msp,Taq}} makes no contribution since an Msp-Taq double digest always produces one fragment with the Msp site 7 as an endsite.

The total of 142 possible phenotypes can all be explicitly displayed. The required combinatorial constructions are implicit in our counting procedures. We will not enter into details here. With more than two or three endonucleases, the sheer number of collections $\mathscr{C}$ becomes prohibitively large.

DISCUSSION

Perhaps the most striking feature of these multiple restriction site systems is their potentially high degree of polymorphism. Given the relatively short distances involved, crossovers between the sites of a system will be exceedingly rare and, for

TABLE 2

COMPUTATION OF ALL PHENOTYPES FOR FIGURE 5

$\mathscr{C}$	$(-1)^{\lvert\mathscr{C}\rvert+1}\underline{m}(\mathscr{C})$
{{Msp}}	10
{{Taq}}.............................	10
{{Msp, Taq}}	125
{{Msp}, {Msp, Taq}}..............	−3
Total.............................	142

NOTE: $\underline{m}(\mathscr{C})$ is the number of phenotypes consistent with all possible digests using the different subsets of endonucleases in $\mathscr{C}$.

purposes of genetic analysis, haplotypes may be regarded as alleles. These systems could rival or even surpass the degree of polymorphism exhibited by the individual loci of the HLA complex. This extensive polymorphism would be particularly useful in human gene mapping. In practice, the rarity of crossovers may restrict the observed number of haplotypes and the actual degree of polymorphism. Such a case of linkage disequilibrium was recently reported by Antonarakis et al. [13]. They described a five restriction site system distributed over 32 kilobases of the β-globin gene family. In this system, three of the possible $2^5 = 32$ haplotypes account for 92% of the 89 chromosomes examined. The observed frequencies of these haplotypes were .64, .15, and .13, in contrast to the expected frequencies of .20, .006, and .005, respectively, assuming linkage equilibrium among the sites.

Another practical limitation is in the number of restriction enzymes that may be used simultaneously to cleave the DNA. Too many enzymes may result in fragments too small for efficient hybridization and transfer, or too homogeneous in length for efficient differentiation using electrophoresis [2, 6–8]. Additionally, restriction enzymes with recognition sequences differing by only a single base pair should not be used together. This technique can actually mask an existing polymorphism. Consider, for example, the enzymes Sst I and Hha I with recognition sequences GAGCTC/CTCGAG and GCGC/CGCG. Used separately, these enzymes distinguish the sequences GAGCTC/CTCGAG and GCGCTC/CGCGAG, with Sst I cutting only the first sequence, and Hha I only the second. Used together, the enzymes cut both sequences and fail to detect the polymorphism.

Because of these technical problems, it may not be practical to determine an individual's genotype using all enzymes simultaneously. For complete haplotype determination, it will usually suffice to carry out a sequence of parallel procedures using pairs of restriction enzymes, one pair of enzymes for each pair of neighboring heterozygous sites. When this is not possible—for example, when one enzyme acts both at a heterozygous site and an adjacent homozygous site—information on relatives may prove definitive, as in HLA typing. Using as few enzymes as possible, one can therefore avoid some of the technical problems described above and still identify haplotypes completely in many cases. This approach entails economic disadvantages, however, since it requires multiple DNA digests per individual.

In conclusion, we would like to stress the limitations of the model discussed. There are still serious issues to be faced by human geneticists in dealing with multiple restriction site systems. For instance, the question of haplotype definition is considerably obscured in the presence of insertions and deletions. Furthermore, the set of relevant external sites depends on the battery of endonucleases employed. Digests with only a subset of enzymes may expose external sites that lie beyond the closest nonpolymorphic sites for the full set of enzymes. Until haplotypes are clearly defined, we cannot hope to adequately classify and catalog these systems. These ambiguities will ultimately impinge on our ability to record and use family data. Finally, this is an area of very rapidly changing technology. There have already been experiments employing adjoining probes to cover a series of closely spaced polymorphic systems [1, 14]. There also exist techniques to evict uninformative fragments and to use subprobes of larger probes. It will take time before typing procedures are standardized.

ACKNOWLEDGMENTS

We thank Drs. Richard Gatti, Kenneth Kidd, Steven Matthysse, and Mark Skolnick for reading the manuscript and suggesting several improvements.

REFERENCES

1. Botstein D, White RL, Skolnick M, Davis RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331, 1980
2. Davies KE: The application of DNA recombinant technology to the analysis of the human genome and genetic disease. *Hum Genet* 58:351–357, 1981
3. Gusella JF, Keys C, Varsanyi-Breiner A, et al.: Isolation and localization of DNA segments from specific human chromosomes. *Proc Natl Acad Sci USA* 77:2829–2833, 1980
4. Housman D, Gusella J: Use of recombinant DNA techniques for linkage studies in genetically based neurological disorders, in *Genetic Research Strategies for Psychobiology and Psychiatry*, edited by Gershon E, Matthysse S, Breakfield XO, Ciaranello RD, Pacific Grove, Calif., Boxwood, 1981, pp 17–24
5. Ruddle FH: A new era in mammalian gene mapping: somatic cell genetics and recombinant DNA methodologies. *Nature* 294:115–120, 1981
6. White R: In search of DNA polymorphism in humans, in *Banbury Report 4: Cancer Incidence in Defined Populations*, edited by Cairns J, Lyon J, Skolnick M, Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory, 1980, pp 409–420
7. Wyman A, White R: A highly polymorphic locus in human DNA. *Proc Natl Acad Sci USA* 77:6754–6758, 1980
8. Southern EM: Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503–517, 1975
9. Riordan J: *An Introduction to Combinatorial Analysis.* New York, John Wiley, 1958, pp 19–49
10. Ortega JM: *Numerical Analysis: A Second Course.* New York, Academic Press, 1972, p 26
11. Seneta E: *Non-Negative Matrices: An Introduction to Theory and Applications.* New York, John Wiley, 1973, pp 1–8
12. Berge C: *Principles of Combinatorics.* New York, Academic Press, 1971, pp 88–95
13. Antonarakis SE, Boehm CD, Giardina PJV, Kazazian HH: Non-random association of polymorphic restriction sites in the β-globin gene cluster. *Am J Hum Genet* 33:35A, 1981
14. Phillips JA, Panny SR, Kazazian HH, Boehm CD, Scott AF, Smith KD: Prenatal diagnosis of sickle cell anemia by restriction endonuclease analysis: Hind III polymorphisms in γ-globin genes extend test applicability. *Proc Natl Acad Sci USA* 77:2853–2856, 1980