J. theor. Biol. (1986) **119**, 397–407

Allele and Locus Classification in Electrophoretic Population Studies

DAVID SANKOFF, VINCENT FERRETTI AND DOMINIQUE ROBY

Centre de recherches mathématiques, Université de Montreal, C.P. 6128,
succ. "*A*", *Montréal, Québec, H3C* 3*J7, Canada and*
Département d'Océanographie, Université du Québec à Rimouski

(*Received* 12 *February* 1985, *and in revised form* 11 *July* 1985)

The electrophoretic separation of protein variants having slightly different mobilities is a basic tool of biochemical population genetics. In certain situations it is difficult to determine how to classify the variants as alleles of a number of genetic loci, that is, as variant subsets within each of which the Mendelian laws hold. In this article, we develop and analyze a series of algorithms for solving various versions and generalizations of this problem of optimal classification.

Electrophoretic separation of the m variants of a protein having slightly different mobilities on a starch, agarose or acrylamide gel is a basic tool of biochemical population genetics (Hartl, 1980, pp. 72–84). Crude tissue extracts from the n individuals in a sample are placed at intervals along a *sample line* or a series of slots across a rectangular gel. Under the influence of an electric field oriented perpendicular to the sample line, the components of the extract including the proteins migrate away from this line towards the anode or cathode, depending on their molecular charge. At a time depending on the mobility of the protein of interest, the field is removed and the current positions of the variant forms in each extract are revealed by protein-specific staining reactions. As illustrated by the (artificial) data in Fig. 1, each of the n individuals in the sample will then be characterized

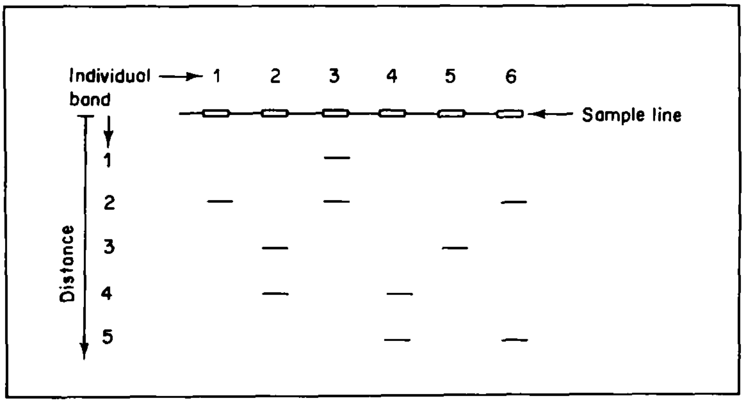

FIG. 1. Electrophoretic patterns of six individuals for a protein with five variants. Individuals 1 and 5 have only one band each; the others all have two.

0022-5193/86/080397+11 $03.00/0

by a pattern on the gel contained in a narrow region perpendicular to the sample line, consisting of one or more *bands*, at various distances from this line, each band indicating the presence of some form of the protein under study. The presence of each variant of the protein will be reflected by the appearance of a band at the same relative position in the pattern of all the individuals containing this variant.

Thus the data produced by an electrophoretic assay are basically as in Table 1; an $m \times n$ matrix of 0's and 1's, indicating absence or presence of the m variants in the n individuals, respectively, together with an increasing sequence of m numbers measuring the distance (positive or negative) traveled from the sample line by each of the m variants represented by the m rows of the matrix.

TABLE 1

Data abstracted from electrophoretic assay of Fig. 1

	1	2	3	4	5	6	Distances
1	0	0	1	0	0	0	6
2	1	0	1	0	0	1	15
3	0	1	0	0	1	0	22
4	0	1	0	1	0	0	34
5	0	0	0	1	0	1	45

In the simplest case the m variants correspond to the m alleles coded by a single genetic locus for an enzyme or other protein which occurs in simple *monomeric* form (i.e. is not composed of two or more subunits). Then the Mendelian distribution of alleles ensures, as in the above example, that each column of the matrix contain either one or two 1's and $m-1$ or $m-2$ 0's. A column with only one 1 indicates an individual homozygotic for the locus, while two 1's indicate a heterozygotic individual.

More generally, several genetic loci may code the same type of protein, so that the m variants are not all alleles of the same locus but must be partitioned among L loci. For example, in a compendium of 76 enzymes used in human genetics studies, Harris & Hopkinson (1976, pp. 1-2) found for about 25% that $L > 1$. Then there may be between L and $2L$ 1's in a column. When L is unknown and each column contains many 1's, it is not always obvious how to partition the set of m alleles into a number of subsets within each of which the Mendelian restriction holds: that is, each individual manifests exactly one or two alleles from each locus. In practice the problem is further complicated by *null* or *silent alleles*: for various reasons it is possible for an individual to manifest *no* allele for a given locus (Harris & Hopkinson, 1976, pp. 1-12). Various biochemical techniques may provide information about which alleles belong in which loci, and analysis of family history, in human genetics, or genetic experiments involving cross-breeding, in botany or entomology, can generally resolve the issue. These are tedious and time-consuming, however, and it would be helpful to have an algorithm capable of suggesting a principled solution using the basic electrophoretic data only.

In its most general form, this problem must be solved through an exhaustive search of all possible partitions to see which, if any, best satisfy the Mendelian laws. This rapidly becomes impractical as m increases. However, alleles belonging to the same locus tend to have similar mobilities so that the search may be appropriately constrained, permitting a much more efficient algorithm. The goal of this paper is to develop such algorithms for various versions and generalizations of the partition problem, allowing null alleles, but imposing constraints on intralocus mobility differences. The approach is that of dynamic programming for the comparison of sequences (Sankoff & Kruskal, 1983), although at the outset only one sequence is explicitly involved.

The Basic Problem and Algorithm

Let **M** be an $m \times n$ matrix of 0's and 1's and $\mu(1), \ldots, \mu(m)$ an increasing sequence of positive real numbers. We are required to find $L \geq 1$ and a partition of the integers $1, \ldots, m$ into L subsets $\sigma_1, \ldots, \sigma_L$ satisfying:

(i) Mendelian condition: if i_1, i_2 and $i_3 \in \sigma_l$ then for $1 \leq j \leq n$, at least one of $M_{i_1 j}$, $M_{i_2 j}$, $M_{i_3 j}$ is zero.

(ii) disjoint locus mobilities: each σ_l consists of consecutive integers.

(iii) Optimality: for some given $\alpha \geq 0$, $\beta > 0$

$$\alpha \sum_{l=1}^{L} \max_{\substack{i_1 \in \sigma_l \\ i_2 \in \sigma_l}} |\mu(i_1) - \mu(i_2)| + \beta \sum_{l=1}^{L} N(\sigma_l)$$

is minimized, where $N(\sigma_l)$ is the number of individuals manifesting a null allele for locus l.

In this formulation, the disjoint mobilities condition and $\alpha > 0$ both reflect the idea that alleles of the same locus should have closely related mobilities. The parameter β must be positive and sufficiently large with respect to α, otherwise allowing null alleles would permit the trivial solution $L = m$.

To illustrate, consider the (artificial) data in Table 2. Conditions (i) and (ii) permit bands 3, 4 and 5, bands 4, 5 and 6 or bands 5, 6 and 7 to be grouped into loci, as well, of course, as any pair of consecutive bands, or any single band. As β increases from zero, the number of loci in the solution satisfying (iii) decreases from 7 to 4 to 3.

TABLE 2

Data set with alternate solutions depending on α and β

<table>
<tr><th></th><th colspan="6">j</th><th></th><th colspan="3">Solutions</th></tr>
<tr><th>i</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>μ</th><th>α = 1
β = 10</th><th>α = 1
β = 1</th><th>α = 1
β = 0·1</th></tr>
<tr><td>1</td><td>1</td><td>1</td><td>0</td><td>1</td><td>0</td><td>1</td><td>15</td><td rowspan="2">}</td><td rowspan="2">}</td><td>}</td></tr>
<tr><td>2</td><td>0</td><td>0</td><td>1</td><td>1</td><td>1</td><td>1</td><td>16</td><td>}</td></tr>
<tr><td>3</td><td>1</td><td>0</td><td>0</td><td>1</td><td>1</td><td>0</td><td>17</td><td rowspan="2">}</td><td rowspan="2">}</td><td>}</td></tr>
<tr><td>4</td><td>1</td><td>1</td><td>1</td><td>0</td><td>1</td><td>0</td><td>18</td><td>}</td></tr>
<tr><td>5</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>1</td><td>25</td><td rowspan="3">}</td><td>}</td><td>}</td></tr>
<tr><td>6</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>1</td><td>31</td><td rowspan="2">}</td><td>}</td></tr>
<tr><td>7</td><td>0</td><td>1</td><td>1</td><td>0</td><td>1</td><td>0</td><td>32</td><td>}</td></tr>
</table>

A rapid solution to the problem is embodied in the recursion

$$C(i)=\min_{0\leq j<i}[C(j)+\alpha(\mu(i)-\mu(j+1))+\beta\bar{N}(\{j+1,\ldots,i\})] \qquad (1)$$

with initial condition $C(0)=0$, where $\bar{N}(\sigma)=N(\sigma)$ if σ is Mendelian and $\bar{N}(\sigma)=\infty$ otherwise. The validity of equation (1) is assured by the additive nature of the criterion in (iii) and the principle of optimality of dynamic programming.

Values of $\mathbf{C}(i)$, for $i=0,\ldots,7$ for the example in Table 2 are 0, 20, 1, 31, 12, 9, 28 and 19 when $\alpha=1$ and $\beta=10$.

In applying this recursion to $i=1,2,\ldots,m$ at each step values of $C(j)$ for $0\leq j<i$ have already been calculated. The computational effort at each step depends on the time it takes to calculate $\bar{N}$. An efficient way to do this is through the precalculation of the column partial sums matrix $\mathbf{S}$ where $\mathbf{S}_{ij}=\sum_{k=1}^{i}\mathbf{M}_{kj}$. Then $N(\{i_1,\ldots,i_2\})$ is just the number of individuals j for whom $\mathbf{S}_{i_1-1,j}=\mathbf{S}_{i_2j}$ and $\bar{N}(\{i_1,\ldots,i_2\})$ is ∞ or $N(\{i_1,\ldots,i_2\})$ depending on whether or not for any individual j, $\mathbf{S}_{i_2j}>\mathbf{S}_{i_1-1,j}+2$. These checks are carried out in time proportional to n, independent of i_2-i_1. Thus the algorithm consisting of applying the recursion to $i=1,\ldots,m$ requires time proportional to m^2n. If there is any motivation for limiting *a priori* the number of alleles per locus, the minimization in the recursion is taken over a fixed number of j only, so that the overall computing time becomes proportional to mn only.

The calculation of C does not complete the solution of the partitioning problem, just the most difficult part of it. In addition, at each application of the recursion, to calculate $C(i)$ we must keep track of the value(s) of j which minimize(s) it. We do this by storing these values in a pointer vector $\mathbf{P}$

$$\mathbf{P}(i)=\{j|0\leq j<i,\ C(j)+\alpha(\mu(i)-\mu(j+1))+\beta\bar{N}(\{j+1,\ldots,i\})\ \text{is minimized}\}$$

For the example in Table 2 ($\alpha=1$, $\beta=10$), the $\mathbf{P}(i)$ for $i=1,\ldots,7$ are {0}, {0}, {2}, {2}, {4}, {4}, {4}. Once $C(m)$ and $\mathbf{P}(m)$ have been calculated, then an optimal partition is reconstructed by

$$\sigma_L=\{p_L+1,\ldots,m\},\ \text{for any } p_L\in\mathbf{P}(m)$$

$$\sigma_{L-1}=\{p_{L-1}+1,\ldots,p_L\},\ \text{for any } p_{L-1}\in\mathbf{P}(p_L)$$

$$\cdots$$

Alternative Treatment of Null Alleles

The linearity of the criterion in (iii) with respect to $N(\sigma_l)$ is convenient, but somewhat arbitrary. In avoiding solutions with null alleles, if it is necessary to admit one individual with a null allele at a given locus, it may not seem much worse to admit two or three. If this treatment is preferred, N should be replaced by I where $I(\sigma_l)=0$ if σ_l has no null alleles and $I(\sigma_l)=1$ if it has one or more. It is no more difficult to compute I than N, and all the algorithms in this paper may make use of either one.

Overlapping Loci

In some cases, the disjoint locus mobility condition (ii) above may be too restrictive an interpretation of the fact that the alleles of a locus have similar mobilities. We may want to allow two adjacent loci σ_l and σ_{l-1} to overlap; some of the $i \in \sigma_l$ may be less than some of the $j \in \sigma_{l-1}$, i.e. some of the $\mu(i)$ may be less than some of the $\mu(j)$. For example, in Table 2 ($\alpha = 1$, $\beta = 10$), two loci containing variants 1, 2 and 4, and 3, 5, 6 and 7, respectively, satisfy (i) and the criterion in (iii) is just 18, compared to 19 under the disjoint locus mobilities condition (ii). Let

$$V = \max\{j - i + 1 \mid i \in \sigma_l, j \in \sigma_{l-1}\}.$$

Then $V \equiv 0$ is just the disjoint locus mobility condition. The overlapping solution just presented for the data in Table 2 has $V = 2$ (N.B. $V = 1$ is impossible). To appropriately weaken condition (ii) we replace it by an upper bound on V,

(ii′) limited overlap: $V \leq V^*$.

In allowing overlap, we risk finding bizarre solutions such as a number of loci all contained within the range of another locus, or even several loci each one *nested* in the previous one. To avoid this, we impose

(iv) nesting prohibition: $\inf \sigma_l < \inf \sigma_{l+1}$, $\sup \sigma_l < \sup \sigma_{l+1}$

To solve the partitioning problem under conditions (i), (ii′), (iii) and (iv), we calculate the following recursion for $i = 3, \ldots, m-1$ and over all subsets $\Sigma_i \subseteq \{i - W, \ldots, i-1\}$, where $W = \min(V^* - 1, i-2)$:

$$C^{(i)}_{\Sigma_i} = \min_{0 \leq j < \inf \Sigma_i - 1} \min_{\Sigma_j} [C_{\Sigma_j}(j) + \alpha(\mu(i) - \mu(\inf \Sigma_j)) + \beta \bar{N}(\Sigma_j \cup \{j+1, \ldots, i\}/\Sigma_i)] \qquad (2)$$

with the interpretation that if $\Sigma_i = \varnothing$, then $\inf \Sigma_i = i+1$. For $i = 0$, 1, 2 and m, we set $C_{\Sigma_i}(i) = \infty$ unless $\Sigma_i = \varnothing$, with the initial condition $C_{\varnothing}(0) = 0$.

In this recursion j and i are considered candidates for $\sup \sigma_{l-1}$ and $\sup \sigma_l$ in some optimizing solution of the partitioning problem. The elements in Σ_j are less than j but are considered part of σ_l. Similarly the elements in Σ_i are excluded from σ_l in preparation for their inclusion in σ_{l+1}. Condition (ii′) is verified since

$$\begin{aligned} \max\{y - x + 1 \mid x \in \sigma_l, y \in \sigma_{l-1}\} &\leq j - \inf \Sigma_j + 1 \\ &= j - (j - V^* + 1) + 1 \\ &= V^*. \end{aligned}$$

Condition (iv) is verified since $j < \inf \Sigma_i - 1$, so that at least one element in σ_l, namely $j+1$, is less than all elements of Σ_i, and hence of σ_{l+1}.

Along with recursion (2), pointers must be calculated and stored for each combination of i and Σ_i, containing all the optimizing j and Σ_j.

The computation for recursion (2) requires time proportional to $4^{V^*-1}{}_m{}^2{}_n$. In the case $V^*=2$, recursion (2) becomes

$$\begin{aligned} C_1(i) &= \min_{0\le j<i-2} \min \begin{cases} C_1(j)+\alpha(\mu(i)-\mu(j-1))+\beta\bar{N}(\{j-1,j+1,\ldots,i-2,i\}) \\ C_\varnothing(j)+\alpha(\mu(i)-\mu(j+1))+\beta\bar{N}(\{j+1,\ldots,i-2,i\}) \end{cases} \\ C_\varnothing(i) &= \min_{0\le j<i} \min \begin{cases} C_1(j)+\alpha(\mu(i)-\mu(j-1))+\beta\bar{N}(\{j-1,j+1,\ldots,i\}) \\ C_\varnothing(j)+\alpha(\mu(i)-\mu(j+1))+\beta\bar{N}(\{j+1,\ldots,i\}) \end{cases} \end{aligned} \tag{3}$$

where $C_1(i)$ is the value of the criterion when $i=\sup\sigma_l$ and σ_l overlaps with σ_{l+1}, while $C_\varnothing$ represents no overlap between these loci. Initial conditions are $C_\varnothing(0)=0$; $C_1(0)=C_1(1)=C_1(2)=\infty$. For the example in Table 2 when $\alpha=1$ and $\beta=10$, the results of applying recursion (3) are given in Table 3. The optimizing solution is the partition consisting of {1, 2, 4} and {3, 5, 6, 7}.

TABLE 3

Results of applying recursion (3) *to Table* 2 *data*

i	0	1	2	3	4	5	6	7
$C_\varnothing(i)$	0	20	1	31	12	9	28	18
opt. j	—	0	0	1 (or 2)	2	2	4	4
opt. Σ_j	—	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	{3}
$C_1(i)$	∞	∞	∞	12	3	19	54	∞
opt. j	—	—	—	0	0	2	3	—
opt. Σ_j	—	—	—	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	—

Gel Alignment

It is sometimes necessary to simultaneously analyze two (or more) electrophoretic assays on different samples of a population or on different populations. In this case we have two matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ of sizes $m_1\times n_1$ and $m_2\times n_2$, respectively, together with corresponding distance vectors $\mu^{(1)}$ and $\mu^{(2)}$. We must find not only $L^{(1)}$ and $L^{(2)}$ and the corresponding partitions, but also an *alignment* of the m_1 rows of $\mathbf{M}^{(1)}$ with the m_2 rows of $\mathbf{M}^{(2)}$, i.e. a suitable set A of pairs such that if $(i_1, i_2)\in A$ and $(j_1, j_2)\in A$, then

$$\begin{gathered} 1\le i_1<j_1\le m_1 \text{ and } 1\le i_2<j_2\le m_2 \\ \text{or} \\ 1\le j_1<i_1\le m_1 \text{ and } 1\le j_2<i_2\le m_2 \end{gathered} \tag{4}$$

It is not necessary that all rows be aligned; e.g. for $1\le i_1\le m_1$ possibly no $(i_1, i_2)\in A$, but rows from the same locus in $\mathbf{M}^{(1)}$ may only be aligned with rows from a single locus of $\mathbf{M}^{(2)}$, and vice-versa. Then along with intra-locus distance differences and null alleles, we try to minimize the number of non-aligned rows (i.e. in no pair in A), the number of loci for which no row is aligned, and the $|\mu^{(1)}(i_1)-\mu^{(2)}(i_2)|$ for $(i_1, i_2)\in A$.

The solution of the partitioning problem is thus to be carried out at the same time as a most parsimonious inference of the genetic differences between the populations (or samples) and of the differences in experimental conditions under which the two gels were obtained. The genetic differences are inferred to be simply the non-aligned alleles and/or loci, while the experimental differences are reflected by the $|\mu^{(1)}(i_1)-\mu^{(2)}(i_2)|$.

For simplicity of notation we present a solution for the non-overlapping case only, though there are no real problems in extending it for $V^*>0$.

We replace (iii) by

(iii′) partition and alignment optimality:

Let A be an alignment of $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ where X loci in one gel or the other are non-aligned, and a further Y alleles are non-aligned in loci which themselves are aligned. The optimality criterion is then

$$C^{(1)}+C^{(2)}+\Delta X+\delta Y+\gamma\sum_{(i_1,i_2)\in A}|\mu^{(1)}(i_1)-\mu^{(2)}(i_2)|$$

for some given constants $\Delta>0$, $\delta>0$ and $\gamma>0$, where $C^{(1)}$ and $C^{(2)}$ are the partitioning criteria for $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ respectively, given by (iii).

Recursion (5) represents a solution to the partition and alignment problem under conditions (i), (ii) and (iii′)

$$C(i_1,i_2)=\min\begin{cases}\min\limits_{0\le j_1<i_1}[C(j_1,i_2)+\alpha(\mu^{(1)}(i_1)-\mu^{(1)}(j_1+1))\\ \qquad+\beta\bar{N}^{(1)}(\{j_1+1,\ldots,i\})+\Delta]\\ \min\limits_{0\le j_2<i_2}[C(i_1,j_2)+\alpha(\mu^{(2)}(i_2)-\mu^{(2)}(j_2+1))\\ \qquad+\beta\bar{N}^{(2)}(\{j+1,\ldots,i\})+\Delta]\\ \min\limits_{\substack{0\le j_1<i_1\\ 0\le j_2<i_2}}\begin{cases}C(j_1,j_2)+\alpha(\mu^{(1)}(i_1)-\mu^{(1)}(j_1+1)\\ \quad+\mu^{(2)}(i_2)-\mu^{(2)}(j_2+1))\\ \quad+\beta(\bar{N}^{(1)}(\{j_1+1,\ldots,i_1\})\\ \quad+\bar{N}^{(2)}(\{j_2+1,\ldots,i_2\}))\\ \quad+d(j_1+1,\ldots,i_1;j_2+1,\ldots,i_2)\end{cases}\end{cases}\tag{5}$$

with initial condition $C(0,0)=0$, and where

$$d(j_1,\ldots,i_1;j_2,\ldots,i_2)=\min\begin{cases}d(j_1,\ldots,i_1-1;j_2,\ldots,i_2)+\delta\\ d(j_1,\ldots,i_1-1;j_2,\ldots,i_2-1)\\ \quad+\gamma|\mu^{(1)}(i_1)-\mu^{(2)}(i_2)|\\ d(j_1,\ldots,i_1;j_2,\ldots,i_2-1)+\delta\end{cases}\tag{6}$$

for $j_1<i_1$ and $j_2<i_2$,

$$d(i_1;j_2,\ldots,i_2)=\min\begin{cases}(i_2-j_2)\delta+\gamma|\mu^{(1)}(i_1)-\mu^{(2)}(i_2)|\\ d(i_1;j_2,\ldots,i_2-1)+\delta\end{cases}$$

for $j_2<i_2$,

$$d(j_1, \ldots, i_1;\, i_2) = \min\begin{cases} d(j_1, \ldots, i_1 - 1;\, i_2) + \delta \\ (i_1 - j_1)\delta + \gamma|\mu^{(1)}(i_1) - \mu^{(2)}(i_2)| \end{cases}$$

for $j_1 < i_1$, and

$$d(i_1;\, i_2) = \min\begin{cases} 2\delta \\ \gamma|\mu^{(1)}(i_1) - \mu^{(2)}(i_2)|. \end{cases}$$

We interpret $d(j_1, \ldots, i_1; j_2, \ldots, i_2)$ as the minimum cost of any alignment satisfying (4) between rows $j_1, \ldots, i_1$ of $\mathbf{M}^{(1)}$ and $j_2, \ldots, i_2$ of $\mathbf{M}^{(2)}$, a non-aligned row costing δ and two aligned rows k_1 and k_2 costing $\gamma|m^{(1)}(k_1) = \mu^{(2)}(k_2)|$. Recursion (6) is the standard dynamic programming solution to this type of alignment problem (Sankoff & Kruskal, 1983).

In recursion (5), the partition criterion is evaluated separately for $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ whichever of the three options is chosen. The first two options involve a non-aligned locus, with cost Δ, and the third involves two aligned loci. In this alignment, each non-aligned row contributes δ to the cost and each alignment pair (k_1, k_2) contributes $\gamma|\mu^{(1)}(k_1) - \mu^{(2)}(k_2)|$, so that the criterion in (iii′) is indeed optimized by recursion (5).

Pointer arrays must be stored for each (i_1, i_2) indicating which of the three options in recursion (5) is (are) optimal, as well as the optimizing j_1 and/or j_2. In addition pointers must be stored for all (j_1, i_1, j_2, i_2) while calculating d, so that the optimizing alignment may be reconstructed once $C(m_1, m_2)$ is obtained.

Computing time for recursion (5) is proportional to $m_1^2 n_1 m_2^2 n_2$. Recursion (6) also requires time proportional to $m_1^2 m_2^2$. If the number of alleles per locus is bounded above, the computing times are proportional to $m_1 n_1 m_2 n_2$ and $m_1 m_2$, respectively.

There is a technique in assay methodology specifically intended to attenuate difficulties in aligning gels. A marker dye of rapid mobility is added to each of the samples before the assay. After the assay is completed, the distance between the sample line and the marker band is used to normalize the two or more gels to a common distance scale. These normalized distances are easily incorporated into recursions (5) and (6).

Hardy–Weinberg Proportions

We have seen that the solution to the partitioning problem can be highly dependent on the values of the parameters α and β. These may be calibrated empirically by testing the algorithm on electrophoretic patterns where the classification of alleles into loci is known beforehand. Another approach is based on the Hardy-Weinberg proportions for alleles in a genetically stable population. If $p_1, p_2, \ldots, p_m$ are the proportions of the m *alleles* of a locus in the population (counting two different alleles per heterozygotic individual, and the same allele twice for a homozygote), then the proportion of homozygotic *individuals* in the population tends to p_1^2, $p_2^2, \ldots, p_m^2$, respectively, while the proportion in the population for each heterozygotic combination of allele i with allele j is $2p_ip_j$.

If there is reason to believe that the population is relatively stable genetically with respect to the protein under study, then the alternative solution of the partitioning problem under various values of α and β may be evaluated for their overall 'fit' to Hardy-Weinberg proportions using a chi-square goodness-of-fit statistic or log-likelihood ratios.

Alternatively, as suggested to us by J. Felsenstein, the goodness-of-fit statistic itself may form part of the optimality criterion, so that instead of the criterion in (iii) we have (iii″)

$$\alpha \sum_{l=1}^{L} \max_{\substack{i_1 \in \sigma_l \\ i_2 \in \sigma_l}} |\mu(i_1) - \mu(i_2)| + \beta \sum_{l=1}^{L} \bar{N}(\sigma_l) + \tau \sum_{l=1}^{L} \chi_l^2$$

where

$$\chi_l^2 = \sum_{i \in \sigma_l} \frac{(np_i^2 - O_i)^2}{np_i^2} + \sum_{i_1 \in \sigma_l} \sum_{i_2 \in \sigma_l} \frac{(2np_{i_1}p_{i_2} - O_{i_1 i_2})^2}{2np_{i_1}p_{i_2}}$$

$$p_i = \frac{1}{2n}\left[2O_i + \sum_{i' \in \sigma_l} O_{ii'}\right]$$

and the O_i and the $O_{i_1 i_2}$ represent the number of homozygotes and heterozygotes of each type on the sample.

Though (iii″) may seem to exacerbate the profusion of parameters in these criteria, in fact the imposition of (ii) largely obviates the need for a non-zero α in many cases, and the parameter τ lessens the need for precise knowledge of β, since solutions with many null alleles will generally not fit the Hardy-Weinberg laws, and will be disfavoured by large values of τ.

Recursion (1) is easily adapted to include the Hardy-Weinberg criterion

$$C(i) = \min_{0<j<i} [C(j) + \alpha(\mu(i) - \mu(j+1)) + \beta \bar{N}(\{j+1, \ldots, i\}) + \tau\chi^2\{j+1, \ldots, i\}]. \qquad (7)$$

The necessity of recalulating the O_i, the $O_{ii'}$, and the p_i for all $i \in \sigma_l$ for all possible σ_l can be computationally costly, however, proportional to $m^4 n$.

Multimeric Proteins

Many types of protein are not monomeric, but are rather composed of $k > 1$ subunits in association, dimers ($k = 2$) and tetromers ($k = 4$) being frequent. In assembling the subunits into an active protein, a single heterozygotic individual will manifest not only two kinds of *homomer* containing k identical subunits of one allelic type or the other, but also $k - 1$ kinds of *heteromer*, containing h subunits of one allelic type and $k - h$ of the other, where $1 \leq h < k$. Thus while homozygotic individuals will still manifest single-band electrophoretic patterns for such loci, heterozygotes will generally display $k + 1$ equally-spaced bands. If μ_1 and μ_2 are the positions of the homomers, then the heteromers will appear at positions $\mu_1 + (1/k)(\mu_2 - \mu_1)$, $\mu_1 + (2/k)(\mu_2 - \mu_1), \ldots, \mu_1 + (k - 1/k)(\mu_2 - \mu_1)$.

If the protein being analyzed is known to be a k-mer, then the Mendelian condition (i) can be altered to assure exactly $k+1$ bands per locus for each heterozygotic individual.

(i′) Mendelian heteromers: For each locus σ_l, and each j, where $i \leq j \leq n$, $\mathbf{M}_{ij}=0$ for all $i \in \sigma_l$ (null allele), or exactly one $\mathbf{M}_{ij}=1$ for some $i \in \sigma_l$ (homozygote), or else $\mathbf{M}_{i_1 j}, \ldots, \mathbf{M}_{i_{k+1} j}$ are all equal to 1, for some $i_1, \ldots, i_{k+1} \in \sigma_l$ (heterozygote).

Then $\bar{N}(\{i_1, \ldots, i_2\})$ is readily redefined so that recursion (1) satisfies conditions (i′), (ii) and (iii), being infinite unless for each individual j, the column partial sums matrix satisfies

$$S_{i_2 j} = S_{i_1-1,j} \quad \text{or} \quad S_{i_2 j} = S_{i_1-1,j}+1 \quad \text{or} \quad S_{i_2 j} = S_{i_1-1,j}+k+1.$$

In theory, the analysis of multimeric proteins could be made more rigorous by evaluating in the recursion how equally spaced the heteromeric bands are in a heterozygotic individual. In practice, however, this requirement would tend to exceed the resolution of the assay method. Indeed, condition (i′) as it stands is too strong to be realistic—it may be more meaningful to allow 0, 1, 2, . . . , $k+1$ bands per locus per individual rather than 0, 1 or $k+1$ only.

One methodological reason for the difficulty in evaluating heteromeric patterns is that some of the bands are necessarily much less intense than others. If π_1 and π_2 are the proportions of the two allelic subunit types produced by a heterozygote, then the proportion of the two types of homomeric protein will be π_1^k and π_2^k, while a heteromer containing h type 1 subunits and $k-h$ type 2 subunits will appear in proportion $\binom{k}{h} \pi_1^h \pi_2^{(k-h)}$. Thus for a heterozygotic tetramer where production determined by the two alleles is balanced, $\pi_1 = \pi_2$, and the proportion of each homomer is less than 0·07 while the heteromer with two subunits of each type is six times more abundant. Thus even if the resolution of the assay permits the distinction between the five closely spaced bands of a heterozygotic individual, the staining technique may not detect the outer two (homomeric) bands.

One approach to resolving these difficulties would be to incorporate a goodness-of-fit criterion in the recursion which would measure how well heterozygotic band patterns for multimers resemble predicted intensity profiles. Again, there is no essential difficulty in incorporating this into the recursion, though the utility of such an algorithm might depend on methodological refinements in quantifying band intensities and calibrating the amount of protein in the sample.

Discussion

We have proposed a dynamic programming approach for resolving ambiguities in the assignment of alleles to loci in electrophoretic population studies. We have shown how this approach may be extended to a variety of related problems. It is true that these problems are most often encountered in practice in a form in which the correct solution can be guessed at without use of formal criteria. It is equally true, however, that protein systems in which there is a multiplicity of variants and ambiguity in their classification are often avoided by experimentalists precisely because of these problems. Moreover, as electrophoretic methodology increases in precision we may expect better resolution to lead to better separation of more variants, adding to the potential relevance of the techniques we describe here. In any case our introduction of optimization criteria and conditions on solutions provides a precise and coherent language for discussing desirable properties in electrophoretic pattern recognition.

We have written and tested experimental computer programs which carry out recursions (1), (2) with $V^* \leq 3$, and (5), and plan to undertake a systematic evaluation of the applicability of the algorithms to various types of electrophoretic data.

We have not exhausted the topics which could be treated within this framework. For example, the algorithm discussed in the previous section could be used in deciding whether a given protein occurs in monomeric, dimeric or other form, when this is unknown (cf. Harris & Hopkinson, 1976, pp. 1-7). The appropriate goodness-of-fit criterion would presumably peak at the correct value of k when the recursion is repeated for $k = 1, 2, \ldots$ in condition (i'). Another problem involves the analysis of heteromers whose subunits can come from different loci. Finally, the strict prohibition against nesting (iv) could be relaxed somewhat to allow one level of nesting at most, i.e. σ_l could be nested in σ_{l+1}, but then no locus could be nested in σ_l nor could σ_{l+1} be nested in any other.

REFERENCES

HARRIS, H & HOPKINSON, D. A. (1976). *Handbook of Enzyme Electrophoresis in Human Genetics.* Amsterdam: North-Holland.

HARTL, D. L. (1980). *Principles of Population Genetics.* Sunderland, Mass: Sinauer Associates Inc.

SANKOFF, D. & KRUSKAL, J. B. (1983). *Time Warps, String Edits, and Macromolecules.* Reading, Mass: Addision-Wesley.