Am. J. Hum. Genet. 41:180–188, 1987

Multipoint Gene Mapping Using Seriation. I. General Methods

KENNETH H. BUETOW AND ARAVINDA CHAKRAVARTI

Department of Biostatistics, Human Genetics Program, University of Pittsburgh, Pittsburgh

SUMMARY

Initial and accurate inference of locus order and estimates of interlocus distances and interference can be obtained using seriation techniques. The analysis requires a matrix of recombination values that can be estimated by standard pairwise linkage analysis. This allows combination of results from individual investigators without reanalysis of basic pedigree material. Seriation can be performed without the use of a computer.

INTRODUCTION

The construction of a comprehensive genetic map of each human chromosome has long been a goal of human genetics. However, plagued by the paucity of polymorphic markers, lack of informative crosses, and small sibship sizes, mapping by the family-study method has in the past been difficult and limited. The continuing discovery of DNA markers has provided a virtually limitless source of polymorphic markers for use in linkage studies. These markers, together with reference pedigrees, such as those available through the Centre d'Étude du Polymorphisme Humain (CEPH), have made it possible to construct a multilocus linkage map of each human chromosome. Although many different strategies have been suggested for multipoint mapping (Meyers et al. 1976; Lalouel 1977; Rao et al. 1979; Lathrop et al. 1984, 1985), there is at present no widely accepted multipoint methodology.

The major goals of multipoint mapping are to obtain locus order and to estimate map distances between loci. These goals are fundamentally different

Received August 26, 1986; revision received February 23, 1987.

Address for correspondence and reviews: Dr. K. H. Buetow, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA 19111.

 0002-9297/87/4102-0009$02.00

from the analysis of two-locus data, a type of analysis whose primary goal is to establish linkage. This change in priority is due in part to the increasing availability of chromosome-specific DNA probes. Multipoint methods must not only detect linkage but must also allow inference about the process of recombination. It is therefore important that the methodologies used for constructing multipoint maps make a minimum of simplifying assumptions and are applicable to a large number of loci. These requirements are necessary since very little is known about recombination mechanisms in higher organisms. Thus, specific models of the recombination process may discard important mapping information. More important, a multilocus mapping method must also be constrained by several practical considerations. Specifically, the method must allow linkage data to be combined and communicated by independent investigators and should not be limited by the availability of extensive computer resources.

In the present paper a new multipoint mapping methodology called seriation is presented. This method uses the results of pairwise linkage analysis to determine locus order and to estimate map distances. It is applicable to an arbitrarily large number of loci, makes few assumptions about the underlying structure of the data, and, at the same time, is conceptually simple and computationally easy.

OBTAINING LOCUS ORDER BY SERIATION

The Pairwise Distance Matrix

The seriation algorithm uses as input a matrix of distances between all possible pairs of loci to be mapped. The distance metric used must display two basic properties, monotonicity and symmetry. In gene mapping the metric of choice is the recombination value, which is easily estimated by the lod-score method of Morton (1955) using computer programs such as LIPED (Ott 1974) and LINKAGE (Lathrop et al. 1984, 1985). A major advantage of the lod-score method is the ease with which linkage data may be reported and summarized from material collected by independent investigators.

The Seriation Algorithm

In 1971 Gelfand (1971) presented an algorithm by which a collection of n objects could be linearly arranged from knowledge of the similarity between pairs of objects. The goal of this algorithm was to order the set of points so that the matrix of similarities between all possible pairs of objects was monotonically arranged or in "Robinson" form (Gelfand 1971). In Gelfand's original paper (Gelfand 1971), seriation was used to recover the temporal order of grave sites in which archeological artifacts had been located, using as a measure of similarity the proportion of artifacts common to two graves. The application of this procedure to multilocus gene mapping is straightforward. In gene mapping the natural metric is a distance, the recombination value between pairs of loci, rather than a similarity measure as suggested by Gelfand (1971).

Consider a distance matrix of pairwise recombination values for n loci where θ_{ij} is the estimated recombination value between the ith and jth locus in the matrix.

For each locus L_i, $i = 1, 2, \ldots, n$ (referred to as the reference locus),
1. Write locus L_i.
2. Consider the distance between L_i and the other $(n - 1)$ loci. Select the locus (L_j) with the smallest distance from L_i and place it to the right of L_i, i.e., L_iL_j.

For the remaining $(n - 2)$ loci in the row referenced by L_i, the following procedure is repeated:
1. Choose the locus L_k from the remaining unplaced loci in that row with the smallest distance to L_i.
2. Compare the distance of L_k with the two loci currently external in the cluster of placed loci, L_l (the locus on the left side) and L_r (the locus on the right side), i.e., $L_l, \ldots, L_r$.

If $\theta_{kr} > \theta_{kl}$, place L_k to the left of the cluster of currently placed loci, i.e., $L_kL_l, \ldots, L_r$, or, if $\theta_{kr} < \theta_{kl}$, place L_k to the right of the cluster of currently placed loci, i.e., $L_l, \ldots, L_rL_k$.

In matrices generated from small samples or from very closely linked loci, it is possible that some of the estimated recombination values will be identical. These identical distances may result in ties at various points in the ordering algorithm. Therefore, we present a set of general rules for resolving ties.

If at any point in considering the row referenced by L_i a tie is encountered when trying to select the next locus—i.e., $\theta_{ij} = \theta_{ik}$, if L_j and L_k are placed to different sides of the ordered locus cluster (e.g., $L_jL_l, \ldots, L_rL_k$)—the tie requires no further consideration; on the other hand, if they are both placed to the same side—i.e., $L_l, \ldots, L_r$ (L_j,L_k)—the two loci should be ordered with respect to the locus most external on that side in the ordered locus cluster (L_r or L_l). If this fails to resolve the tie, the next two internal loci in the ordered cluster should be considered, i.e., (L_rL_r* or L_l*L_l). If, after considering all loci already placed, the tie cannot be resolved, the group may be ordered with respect to a locus external to the tied loci and placed locus cluster.

To use an external reference locus, L_e, one selects the next closest locus in the row indicated by the reference locus (L_i) that is not included in the tied group. This locus is then placed with respect to the ordered locus cluster—$L_r, \ldots, L_l$ (L_j,L_k) L_e—and the tied loci are ordered with respect to their distance from this locus, θ_{je} and θ_{ke}. If this locus fails to resolve the tie, a new external locus is chosen. At the point that all loci in the reference row have been attempted without breaking the tie, the tie is considered unresolvable.

A second situation in which a tie may be observed is when, in trying to place locus L_k, the distance from the rightmost and leftmost loci in the ordered locus cluster is found to be the same, i.e., $\theta_{kr} = \theta_{kl}$. This again can be resolved by first considering the more internal loci in the ordered locus cluster and then, if unsuccessful, using an external reference locus.

Determination of Locus Order

After deriving an order with respect to each locus, it is necessary to reduce these to a single order. If the estimated matrix of recombination values is

monotonic (or Robinson), obtaining the final order is straightforward, since all locus-specific orders will differ only in orientation and therefore be mirror images of each other—i.e., A-B-C-D versus D-C-B-A (Gelfand 1971). Unfortunately, many observed data matrices will not be optimal (monotonic), and a variety of locus-specific orders may be observed. There are two approaches that can be taken to obtain the final order.

First, one may obtain an average order by using rank scores (Gelfand 1971)—namely, after each order has been oriented in the same direction, the average position of each locus in the observed orders is determined. This is calculated as the sum of ranks of this locus in each of the derived orders. The final order is obtained by arranging the loci in the order corresponding to the magnitude of the rank sum for each locus. If ties in the rank sums are observed, no single final order is derived and a set of orders obtained by placing the tied loci in the alternative positions are suggested. The final order is then judged by other criteria.

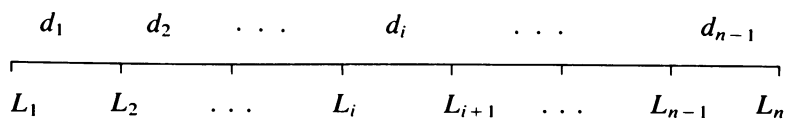
Second, one may choose that order that comes closest to monotonically arranging the elements of the recombination-value matrix. Such an order can be obtained by measuring the goodness-of-fit to monotonicity. Gelfand (1971) defined one possible monotonicity measure, the continuity index (CI), in which the observed metric distance between two loci, θ_{ij}, is compared with the ordinal distance between their locations in the order, namely,

$$CI = \sum_{i<j} \left[\frac{\theta_{ij}}{(j-i)^2} \right] . \tag{1}$$

The CI value is computed for each order using equation (1), the best order being the one with the lowest CI. One shortcoming of the CI is that the ratio of metric to ordinal values does not assume a singular value for different optimal distance matrices. Therefore, although the measure is useful in deciding which of several orders provides the best fit, it is not possible to infer whether another unspecified order may provide a better fit. Additionally, large θ_{ij}'s with small interlocus ordinal differences $(j - i)$ will contribute more to the overall magnitude of the CI than will loci with small θ_{ij}'s and large $(j - i)$ differences.

Least-Squares Estimation of Map Distances

After obtaining the locus order, the interlocus map distances can be obtained from the pairwise distances by means of least squares. In this procedure, the ordered distance matrix of recombination values is transformed to map distances by means of an appropriate mapping function. From this transformed matrix, estimates of interlocus distances between adjacent loci (d_i) can be obtained. Consider the seriated order of a linkage group of n loci with distances d_i between adjacent loci as follows:



Let $D(j, k)$ be the observed map distance between loci j and k, where $j \neq k$; $j = 1, 2, \ldots, n-1$ and $k = j+1, \ldots, n$. The least-squares estimates of d_i are then (see Appendix):

$$\hat{d}_i = \frac{2\Phi_i - (\Phi_{i-1} + \Phi_{i+1})}{n} \tag{2}$$

for $i = 1, 2, \ldots, n-1$, where,

$$\Phi_i = \sum_{j=1}^{i} \sum_{k=i+1}^{n} D(j, k) \tag{3}$$

and $\Phi_0 = \Phi_n = 0$.

Estimating the Mapping Function

The results of the least-squares estimation depend critically on the choice of an appropriate mapping function. Historically this choice has been either based on interference relationships observed in experimental organisms (Rao et al. 1977; Ott 1985) or fixed for computational convenience (Lathrop et al. 1984). An alternative strategy is to estimate this function from the present observations, as described below.

Map distances are additive, so that one can define an index that measures how adequately a given mapping function succeeds in transforming the observed pairwise recombination values to meet this additivity criterion. We define one such index, stress (S).

Let, $D(i, j)$ and d_i be defined as above. The expected distance between two loci, $\mathrm{E}(i, j)$ is then the sum of the adjacent intervals between these two points, or,

$$\mathrm{E}(i, j) = d_i + d_{i+1} + \ldots + d_{j-1} \ . \tag{4}$$

Stress can then be defined as (Lalouel 1977; Kruskal and Wish 1978)

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{[D(i, j) - \mathrm{E}(i, j)]^2}{\mathrm{E}(i, j)^2} \ . \tag{5}$$

Although equation (5) appears to be similar in form to the χ^2 goodness-of-fit statistic, it is not a χ^2. The terms are squared to make the index independent of scale of measurement.

The best mapping function is then determined by a method similar to that first suggested by Lalouel (1977). In brief, successive values of the mapping parameter (p) in Rao's generalized mapping function (Rao et al. 1977) are used to transform the observed recombination values to map distances. The value of p that minimizes the stress in equation (5) determines the best mapping function for this data set.

DISCUSSION

Initial and accurate solutions to multilocus ordering and distance estimation can be obtained from pairwise linkage data by using the seriation procedure (Buetow et al. 1986). The mapping results obtained by using seriation on simulated and empirical data sets are comparable to those determined by other multipoint techniques and are described in the companion article (Buetow and Chakravarti 1987) and in Buetow et al. (1986).

The use of pairwise linkage data and seriation methods, although less efficient than true multipoint analysis (Lathrop et al. 1984), offers several advantages in constructing preliminary maps. First, the use of lod scores provides a standard means of data communication and summary between investigators. This aspect cannot be abandoned and will become more critical both as linkage data accumulates and as it becomes impractical for any one investigator to test all markers on a given chromosome. Seriation is a simple method for deriving multilocus orders from compiled groups of data, such as those maintained by Keats (Keats et al. 1979; Keats 1981). The addition of new marker or disease loci to an established genetic map may be accomplished without reanalysis of the basic pedigree material. At most, in an existing map of n loci, only n pairwise tests will need to be performed. Also, unlike the location-score method (Lathrop et al. 1984), it is not necessary to keep the previously mapped locations fixed. Another advantage of seriation is that it requires few biological assumptions. Since the locus order is derived from recombinations values, it is not necessary to assume a mapping function. In fact, it is possible to estimate interference levels and to determine the most appropriate mapping function for a particular data set.

Seriation offers several practical advantages as well. First, the method is applicable to an arbitrarily large number of loci. Second, the algorithm does not require the use of a computer and, if desired, can be performed by hand, even for large linkage groups. Simplicity in computation is one way in which seriation differs from multidimensional scaling techniques (Kruskal and Wish 1978). Seriation is also different from multidimensional scaling in that it uses the recombination values as observed rather than reestimating them for each trial configuration. This characteristic makes seriation less sensitive to 50%-recombination values because it does not have to reconcile loci of various underlying map distances with the same observed value.

The seriation algorithm is limited in two respects. First, it requires that all possible pairs of distances between loci be available. When extracting data from the literature, it may be difficult to recover a complete set of pairwise comparisons. This normally can be overcome by seriating smaller sets of overlapping loci and could be eliminated by more comprehensive data reporting. Second, seriation gives equal weights to all observations in the pairwise distance matrix. This should not present a significant problem when reasonable sample sizes are used to estimate recombination values.

The least-squares procedure implicitly assumes independence of all pairwise distances. This assumption is justified when recombination values are estimated from independent studies and is approximately valid when they are

estimated from the same set of families. This is true since there is a low probability that multiple loci will be jointly informative in single families.

The seriation procedure is proposed as a means of providing initial gene orders and interlocus distance estimates. These in turn may be used as input for more time-consuming, maximum-likelihood multipoint techniques and will assist in both reducing the number of possible orders that must be considered and providing initial estimates for interlocus distances.

ACKNOWLEDGMENT

We wish to thank Dr. C. R. Rao for discussions of the seriation methods. A portion of this research was supported by National Institutes of Health grant GM33771 to A.C.

APPENDIX

LEAST-SQUARES ESTIMATION OF INTERLOCUS MAP DISTANCES

We assume that $L_1, L_2, \ldots, L_n$ is the order of n loci within a linkage group with map distance d_i ($i = 1, 2, \ldots, n-1$) between loci L_i and L_{i+1} ($i = 1, 2, \ldots, n-1$). Furthermore, let D_{jk} be the observed map distance between any two locus pairs j and k, where $j \neq k = 1, 2, \ldots, n$. Our objective is to estimate the set of $(n-1)$ parameters (d_i) from the set of $n(n-1)/2$ pairwise observations (D_{jk}) by means of the least-squares method. We will assume that the observations D_{jk} are independent of each other. Then, the expectation of D_{jk} is $E(D_{jk}) = d_j + d_{j+1} + \ldots + d_{k-1}$, where $j = 1, 2, \ldots, n-1$ and $k = j+1, \ldots, n$. In matrix notation, $E(\underset{\sim}{D}) = X\underset{\sim}{d}$, where $\underset{\sim}{D}$ is the m $\times$ 1 column vector of the elements D_{jk} arranged in dictionary order (i.e., $D_{12}, D_{13}, \ldots, D_{1n}, D_{23}, D_{24}, \ldots, \ldots, D_{n-1,n}$), $\underset{\sim}{d}$ is the $(n-1) \times 1$ column vector $(d_1, d_2, \ldots, d_{n-1})'$, and X is an $m \times (n-1)$ matrix of zeros and ones such that the jth row corresponding to the element D_{jk}, for example, contains a 1 in every column between j and k; also $m = n(n-1)/2$. The least-squares estimate of the vector $(\underset{\sim}{d}\ \hat{\underset{\sim}{d}})$ is obtained by minimizing $Q = (\underset{\sim}{D} - X\underset{\sim}{d})'(\underset{\sim}{D} - X\underset{\sim}{d})$ with solution [16],

$$\hat{\underset{\sim}{d}} = (X'X)^{-1}(X'\underset{\sim}{D}), \tag{A1}$$

or, alternatively, solving the equation

$$(X'X)\hat{\underset{\sim}{d}} = X'\underset{\sim}{D} \tag{A2}$$

Equation (A2) is easier to solve than equations (A1) since from (A2) we have

$$\begin{bmatrix} n-1 & n-2 & n-3 & \ldots & n-i & \ldots & 1 \\ n-2 & 2(n-2) & 2(n-3) & \ldots & 2(n-i) & \ldots & 2 \\ n-3 & 2(n-3) & 3(n-3) & \ldots & 3(n-i) & \ldots & 3 \\ \ldots & \ldots & \ldots & & \ldots & & \ldots \\ \ldots & \ldots & \ldots & & \ldots & & \ldots \\ \ldots & \ldots & \ldots & & \ldots & & \ldots \\ n-i & 2(n-i) & 3(n-i) & \ldots & i(n-i) & \ldots & i \\ \ldots & \ldots & \ldots & & \ldots & & \ldots \\ \ldots & \ldots & \ldots & & \ldots & & \ldots \\ \ldots & \ldots & \ldots & & \ldots & & \ldots \\ 1 & 2 & 3 & \ldots & i & & n-1 \end{bmatrix} \begin{bmatrix} \hat{d}_1 \\ \hat{d}_2 \\ \hat{d}_3 \\ \ldots \\ \ldots \\ \ldots \\ \hat{d}_i \\ \ldots \\ \ldots \\ \ldots \\ \hat{d}_{n-1} \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \ldots \\ \ldots \\ \ldots \\ \phi_i \\ \ldots \\ \ldots \\ \ldots \\ \phi_{n-1} \end{bmatrix}, \tag{A3}$$

where

$$\phi_i = \sum_{j=1}^{i} \sum_{k=i+1}^{n} D_{jk}$$

and represents the sum of all observed pairwise distances containing the interval d_i. Also, we define $\phi_0 = \phi_n = 0$. Note that from formula (A3)

$$\phi_i = (n - i) \sum_{\alpha=1}^{i} \alpha\, d_\alpha + i \sum_{\beta=i+1}^{n-1} (n - \beta) d_\beta \ ,$$

so that

$$\phi_i - \phi_{i-1} = \sum_{\beta=i}^{n-1} (n - \beta) d_\beta - \sum_{\alpha=1}^{i-1} \alpha\, d_\alpha \ , \tag{A4}$$

and

$$\phi_i - \phi_{i+1} = \sum_{\alpha=1}^{i} \alpha\, d_\alpha - \sum_{\beta=i+1}^{n-1} (n - \beta) d_\beta \ . \tag{A5}$$

On summing equations (A4) and (A5), we obtain $2\phi_i - \phi_{i-1} - \phi_{i+1} = n\hat{d}_i$, so that

$$\hat{d}_i = [2\phi_i - (\phi_{i-1} + \phi_{i+1})]/n \ . \tag{A6}$$

REFERENCES

Buetow, K. H., and A. Chakravarti. 1987. Multipoint mapping using seriation. II. Analysis of simulated and empirical linkage data. Am. J. Hum. Genet. **41**:189–201.

Buetow, K. H., A. Chakravarti, and S. Cole. 1986. A genetic map of human chromosome 11p. Genet. Epidemiol. [Suppl.] **1**:135–140.

Gelfand, A. E. 1971. Seriation. Pp. 186–201 *in* F. R. Hodson, D. G. Kendall, and P. Tauta, eds. Mathematics in the archaeological and historical sciences. Edinburgh University Press, Edinburgh.

Keats, B. J. B. 1981. Linkage and chromosome mapping in man. The University Press of Hawaii, Honolulu.

Keats, B. J. B., N. E. Morton, D. C. Rao, and W. R. Williams. 1979. A source book for linkage in man. The Johns Hopkins University Press, Baltimore.

Kruskal, J. B., and M. Wish. 1978. Multidimensional scaling. Sage, Beverly Hills, CA.

Lalouel, J. M. 1977. Linkage mapping from pairwise recombination data. Heredity **38**:61–77.

Lathrop, G. M., J. M. Lalouel, C. Julier, and J. Ott. 1984. Strategies for multilocus linkage analysis in humans. Proc. Natl. Acad. Sci. USA **81**:3443–3446.

———. 1985. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am. J. Hum. Genet. **37**:482–498.

Meyers, D. A., P. M. Conneally, and E. W. Louvrien. 1976. Linkage group I: the simultaneous estimation of recombination and interference. Pp. 335–339 *in* Baltimore Conference (1975): Third International Workshop on Human Gene Mapping. Birth Defects: Original Article Series.

Morton, N. E. 1955. Sequential tests for the detection of linkage. Am. J. Hum. Genet. **7**:277–318.

Ott, J. 1974. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am. J. Hum. Genet. **26**:588–597.

———. 1985. Analysis of human genetic linkage. The Johns Hopkins University Press, Baltimore.

Rao, C. R. 1973. Linear statistical inference and its applications. Wiley, New York.

Rao, D. C., B. J. Keats, J. M. Lalouel, N. E. Morton, and S. Yee. 1979. A maximum likelihood map of chromosome 1. Am. J. Hum. Genet. **31**:680–696.

Rao, D. C., N. E. Morton, J. Lindsten, M. Hultén, and S. Yee. 1977. A mapping function for man. Hum. Hered. **27**:38–51.