Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 2363–2367, April 1987
Genetics

Construction of multilocus genetic linkage maps in humans

(restriction fragment length polymorphism/EM algorithm/genetic reconstruction algorithm/human genetics/plant genetics)

ERIC S. LANDER†‡§ AND PHILIP GREEN¶

†Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142; ‡Massachusetts Institute of Technology, Cambridge, MA 02139; §Harvard University, Cambridge, MA 02138; and ¶Human Genetics Department, Collaborative Research, 2 Oak Park, Bedford, MA 01730

Communicated by David Botstein, November 24, 1986

ABSTRACT **Human genetic linkage maps are most accurately constructed by using information from many loci simultaneously. Traditional methods for such multilocus linkage analysis are computationally prohibitive in general, even with supercomputers. The problem has acquired practical importance because of the current international collaboration aimed at constructing a complete human linkage map of DNA markers through the study of three-generation pedigrees. We describe here several alternative algorithms for constructing human linkage maps given a specified gene order. One method allows maximum-likelihood multilocus linkage maps for dozens of DNA markers in such three-generation pedigrees to be constructed in minutes.**

A fundamental problem with constructing genetic linkage maps in humans is that important data are often missing. Whereas a *Drosophila* geneticist may arrange crosses to avoid or resolve any potential ambiguities, human geneticists must take crosses as they find them in natural populations. Human geneticists thus cannot simply "count recombinants" in a cross, since they typically lack the information needed to identify unambiguously where recombination events occurred. The reasons are three: (*i*) Parents are typically homozygous, and thus uninformative, for some of the loci of interest. (*ii*) Even where parents are heterozygous, it is often unknown which alleles at various loci are in cis and which are in trans (i.e., the linkage phase is unknown). (*iii*) Genotype cannot always be uniquely inferred from phenotype.

To address this problem, Fisher (1), Haldane and Smith (2), and Morton (3) developed a theoretical approach based on the method of maximum likelihood: considering all possibilities for the missing data, map distances are chosen to maximize the probability that the observed data would have occurred. When no data are missing, the approach reduces to counting recombinants. Elston and Stewart (4) proposed a general algorithm for computing the required likelihoods. Using this algorithm, Ott (5) produced a computer program, LIPED, that allowed a geneticist efficiently to determine the recombination fraction θ between a pair of genetic loci. The dearth of adequately polymorphic human genetic markers made it unnecessary to consider any but two-point crosses.

Recent advances in molecular biology, however, have made it practical to score hundreds of genetic markers in humans: each a variation in DNA sequence conveniently observed as a restriction fragment length polymorphism (RFLP). Botstein *et al.* (6) suggested that RFLPs could be used for the systematic study of human heredity and proposed the construction of a true linkage map of the entire human genome. Lander and Botstein (7, 8) have shown that such an RFLP linkage map would allow more powerful analytical strategies for the study of human diseases and traits.

Genetic maps are most accurately constructed by using multipoint crosses. In humans, the case for multilocus analysis is even stronger: gathering enough information to map, for example, a disease-causing locus may require pooling data from many families, each informative for a different set of marker loci. Studying a dozen or more loci simultaneously may thus often be desirable.

Such multilocus analysis, however, faces severe computational obstacles: (*i*) With m loci under study, there are $\frac{1}{2}m!$ potential gene orders. (*ii*) For even a single gene order, the traditional approach to constructing human linkage maps requires computing time that scales exponentially with the number of loci studied. Many hours of computer time may be required to analyze four or five loci in a single order, despite excellent computer programs (9, 10). For a larger number of loci, "simultaneous analysis with current algorithms is prohibitively time-consuming, even on a supercomputer" (12).

This paper addresses the second problem: given a gene order, we explore ways to make multilocus linkage analysis and computation of likelihoods practical, even for dozens of loci. The main ideas are (*i*) a different search principle and (*ii*) an algorithm for each step of the search that scales linearly rather than exponentially with the number of loci studied. Provided that the pedigrees under study are not too large, the simultaneous study of any number of loci becomes feasible.

When gene order is not known, the methods can be used to compare the likelihood of alternative gene orders.

Statement of Problem

Let $M_1, \ldots, M_m$ be m genetic loci, listed in the correct (or assumed) chromosomal order. Given information about the phenotypes of members of several pedigrees, we wish to construct the best genetic map. Specifically, let θ_i denote the recombination fraction between adjacent loci M_i and M_{i+1}. We want to find the value of $\theta = (\theta_1, \ldots, \theta_{m-1})$ that maximizes the chance of the data having arisen. For simplicity, we shall ignore crossover interference; i.e., assume complete independence of recombination between all chromosomal intervals. Although a useful starting point, this assumption requires future scrutiny, since interference certainly exists in well-studied organisms such as *Drosophila melanogaster*. Also, we shall suppose here that phenotypes due to different loci are not epistatic.

Finding θ requires searching a multidimensional space. An iterative procedure must be specified to replace a previous guess θ^{old} by a revised guess θ^{new}, at which the likelihood is (one hopes) higher.

Traditional Approach. The traditional approach (9, 10, 13) is to approximate the derivative of the likelihood function $L(\theta)$ at θ^{old} by computing the likelihood at θ^{old}, and at $m - 1$ further points each displaced slightly in a different coordinate

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: RFLP, restriction fragment length polymorphism.

direction. A so-called "quasi-Newton" method (13–15) is then used to choose θ^{new}.

The method has three drawbacks: (*i*) On each iteration, likelihoods must be computed at m different points. Each likelihood calculation is very time-consuming when many loci are involved. Indeed, Newton's method itself is not used for this very reason: it involves second derivatives, which require computing likelihoods at m^2 points. (*ii*) θ^{new} may occasionally have lower likelihood. (*iii*) If the initial guess is far from the maximum, initial convergence may be very slow (14), especially in high-dimensional spaces. We therefore discuss an alternative search technique.

Search Via the EM Algorithm. Human genetic map-making can be viewed as a problem of missing data. In experimental organisms, geneticists arrange to observe complete data: the number of recombinant and nonrecombinant meioses that occurred in each of the intervals (M_i, M_{i+1}). Given these complete data, the maximum likelihood θ_i is determined simply by counting recombinants: θ_i is the ratio of the number of recombinants to total meioses.

Human geneticists must estimate the parameters θ_i by using only incomplete data—data that do not uniquely determine the number of recombinant and nonrecombinant meioses. The EM algorithm (16, 17) offers a powerful general approach to obtaining maximum likelihood estimates from incomplete data. Applied to linkage analysis, it prescribes the following:

(*i*) Make an initial guess, $\theta^{old} = (\theta_1, \ldots, \theta_{m-1})$.

(*ii*) Expectation step. Using θ^{old} as if it were the true recombination fraction, compute the expected value for the complete data—i.e., the expected number of recombinant and nonrecombinant meioses in each interval.

(*iii*) Maximization step. Using the expected value of the complete data as if it were the true value, compute the maximum likelihood estimate θ^{new} for the recombination fractions.

(*iv*) Iterate the E and M steps until the likelihood converges to a maximum.

The EM algorithm is not truly an algorithm, since it specifies no procedure for performing the E and M steps. For each application, appropriate algorithms must be fashioned. For genetic map-making, the M step is trivial: $(\theta^{new})_i$ is just the ratio of the expected number of meioses recombinant for the ith interval (recombinant meioses for short) to the total number of meioses.

The difficulty is the E step, which we call the "genetic reconstruction problem": given the recombination fractions θ, compute the expected number of recombinant meioses for each meiosis. Several approaches to the genetic reconstruction problem are discussed below. While the traditional search method requires m likelihood calculations per iteration, we shall show that genetic reconstruction can be accomplished with the equivalent of only two traditional likelihood calculations. (For two- and three-generation pedigrees, we shall also describe even faster methods.)

Advantages of EM Search. (*i*) *Likelihood increases monotonically.* Since the probability distribution for the complete data comes from an exponential-family form, the following result holds (16, 17).

THEOREM. *Successive estimates of θ generated by the EM algorithm have increasing likelihoods and converge to a point θ^* at which the derivative of the likelihood is zero.*

As for all general optimization procedures, there is no guarantee that the limit point is the global maximum; several initial guesses should be tried. In our experience, however, most human linkage problems appear to have a single local maximum, for a given gene order. The exceptions involve either very small or unlikely data sets.

(ii) Convergence properties of EM are roughly opposite those of Newton searches. Unlike Newton searches, EM searches tend to converge quickly to the vicinity of the maximum, even when started at a distant point. Thus, EM is often favored in multidimensional searches, an extreme example being the reconstruction of a positron emission tomography scan image involving maximizing >15,000 variables (18). Our tests, described below, show that EM is similarly effective at solving genetic linkage problems: 5–20 iterations typically suffice for problems involving dozens of loci.

In the vicinity of the maximum, each EM iteration covers only a constant fraction of the remaining distance on each iteration (i.e., linear convergence) (16, 17). Newton-type methods converge more quickly in the final stages and thus are preferable when many decimal places of accuracy are required. In human genetics, such accuracy is unnecessary and typically spurious. Nevertheless, one can easily accelerate the convergence of EM in the final stages by using the fact that the EM method yields the exact derivatives of the likelihood function for no extra work. If θ^{old} and θ^{new} are the initial and revised estimates according to EM, then

$$\frac{\partial \log L(\theta^{old})}{\partial \theta_i} = \frac{n(\theta_i^{new} - \theta_i^{old})}{\theta_i(1 - \theta_i)}, \quad [1]$$

where n is the total number of meioses in the pedigrees. (Eq. 1 amounts to a special case of formula 2.13 in ref. 16; it also follows directly from differentiating the expression for the likelihood.) Thus, one can switch to a Newton rule in the vicinity of the maximum, using EM to generate the required derivatives. More subtle acceleration methods for EM are also known and involve predicting the target of the linear convergence (16). Our experiments (described below) suggest that such methods can roughly halve the number of iterations required for satisfactory convergence in practical problems.

Second derivatives, and thus the information matrix, can also be computed exactly via the EM approach (16).

(iii) Being intuitive, EM is easy to generalize. Sex-specific estimates, θ^{male} and θ^{female}, of the recombination fractions can be found with only a minor modification of the above: simply "count expected recombinants" separately in male and female meiosis to obtain sex-specific revised guesses. The computation time per iteration is unchanged, even though twice as many variables are being estimated.

Summary. The theoretical advantages of the EM search are (*i*) less computation time per iteration; (*ii*) increased likelihood on each iteration; (*iii*) good initial convergence properties; (*iv*) exact expressions for derivatives of the likelihood; and (*v*) ease of generalization.

Genetic Reconstruction

The practicality of the EM approach rests entirely on efficient solutions to the genetic reconstruction problem: Given phenotype data for the loci $M_1, \ldots, M_m$ in a pedigree and given the recombination fractions $\theta = (\theta_1, \ldots, \theta_{m-1})$ between consecutive loci, determine the expected number of recombinations that occurred in each interval (M_i, M_{i+1}).

The answer depends on the nature of the data. We discuss three algorithms suited to different situations.

Reconstruction: A Special Case

Known Genotypes. Suppose that we can completely observe the genotype of each individual in a pedigree, including which alleles are on the paternally and maternally derived chromosomes. This is frequently possible for most meioses in multigeneration families. For each meiosis, we can then tell by inspection whether a recombination occurred between

any two loci for which the parent is informative (i.e., heterozygous). The data are incomplete only in that some loci are uninformative.

For example, suppose that M_1 and M_3 are informative, but locus M_2 is not. If M_1 and M_3 recombined in a meiosis, we cannot tell whether a recombination occurred in the interval (M_1, M_2) or in (M_2, M_3). Nevertheless, it is easy to determine the expected number of recombinations in each interval (here, just the probability of a recombination). For (M_1, M_2), it is $p_1 = \theta_1(1 - \theta_2)/[\theta_1(1 - \theta_2) + (1 - \theta_1)\theta_2]$. For (M_2, M_3) it is $p_2 = 1 - p_1$. On the other hand, if M_1 and M_3 did not recombine, then the chance that a recombination occurred in either of the basic intervals is $p_1 = p_2 = \theta_1\theta_2/[\theta_1\theta_2 + (1 - \theta_1)(1 - \theta_2)]$.

Similarly, a recombination or nonrecombination observed in a larger interval can be apportioned into expected recombinations and nonrecombinations in each of the subintervals. Genetic reconstruction consists of performing this process for each meiosis.

Computational Complexity. For the sake of efficiency, observations concerning the same interval from different meioses should first be aggregated. Recombinations and nonrecombinations in each interval starting at M_1 should next be apportioned between (M_1, M_2) and the remaining subinterval. Then intervals starting at M_2 should be apportioned and so on. Structured in this way, the computing time is proportional to m^2. (If observations were not aggregated, the running time would be proportional to mk, where k is the number of individuals under study. Typically, $k \geq m$.)

We now turn to the general case.

Reconstruction: Via Elston–Stewart Algorithm

When only a few loci are considered, genetic reconstruction can be efficiently performed via a slight modification of the Elston–Stewart algorithm (4, 19). In brief, the Elston–Stewart algorithm proceeds recursively up the family tree computing probabilities for each possible genotype of each child, conditional on the genotypes of his parents, the phenotype of the child, and the phenotypes for the child's descendants.

Genetic reconstruction may be performed as follows: (*i*) Having performed the Elston–Stewart algorithm, descend the pedigree computing the probability distribution over the possible genotypes for each triple consisting of a mother, father, and child, via Bayes' theorem. (*ii*) For each triple (x, y, z) of genotypes, count the expected number of triples consisting of a mother, father, and child having genotypes (x, y, z), respectively. (*iii*) Each triple (x, y, z) corresponds to one of the 2^{2m-2} possible patterns of recombination for the $(m - 1)$ intervals in male and female meiosis; add up the expected number of occurrences of each pattern. (*iv*) For each interval, add up expected occurrences of crossover patterns with a recombination in the interval.

Computational Complexity. For m loci having a alleles each in a pedigree with n nonoriginal individuals and no inbreeding, the Elston–Stewart algorithm requires on the order of $a^{6m}n$ multiplications and $a^{6m}n$ additions (see ref. 19). The four steps of genetic reconstruction then require on the order of (*i*) $a^{6m}n$ multiplications and $a^{6m}n$ additions; (*ii*) $a^{6m}n$ additions; (*iii*) a^{6m} additions; and (*iv*) $m2^{2m-1}$ additions, respectively. Genetic reconstruction thus essentially doubles the asymptotic computation time for the Elston–Stewart algorithm, as claimed above.

Since the computation time scales with a^{6m}, the Elston–Stewart algorithm becomes impractical for more than four or five loci. This provoked us to develop an alternative approach.

Reconstruction: Via Hidden Markov Chains

Whereas the Elston–Stewart method is appropriate for pedigrees of arbitrary size but only few loci, the following approach will handle arbitrarily many loci but only pedigrees of limited size.

The Inheritance Vector. Consider a pedigree containing k nonoriginals—that is, individuals with at least one parent in the pedigree. For a locus M_i, define the inheritance vector $\mathbf{v}_i$ to be a binary vector of length $2k$, with coordinates corresponding to the $2k$ gametes that gave rise to the nonoriginals. A coordinate is 0 if the gamete carried DNA from the parent's paternally derived chromosome; otherwise, it is 1. The *a priori* chance that any given coordinate differs between $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ is the recombination fraction θ_i. In other words, the inheritance vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ arise from an inhomogeneous Markov chain with known transition matrices: the transition $T(\theta_i)$ between M_i and M_{i+1} is the Kronecker product of the 2×2 transition matrices corresponding to transitions in each of the $2k$ coordinates.

Human geneticists observe only phenotype data at each locus M_i, from which the inheritance vector $\mathbf{v}_i$ cannot be uniquely inferred. (If the inheritance vector could be uniquely inferred, genetic reconstruction would be trivial: the number of recombinants in the ith interval would simply be the number of coordinates at which $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ differ.) However, it is easy to compute the probability that the phenotype data at M_i would have been observed, given each of the possible values for $\mathbf{v}_i$. Let q_i denote a row vector of these probabilities, with coordinates indexed by the possible values for $\mathbf{v}_i$. Applying Bayes' theorem (with all inheritance vectors equally probable *a priori*), one can then compute the probability distribution p_i over the possible values for $\mathbf{v}_i$, conditional on the phenotype data for M_i. As for q_i, view p_i as a row vector indexed by possible values for $\mathbf{v}_i$. Although in the worst case q_i and p_i could have 2^{2k} nonzero coordinates, typically the support is over a much smaller set—since the phenotype data automatically exclude many possibilities. (For efficiency, a locus that is completely uniformative in a family, and thus for which no possibilities may be excluded, should be skipped over. Expected recombinations in the resulting larger interval may then be apportioned using the approach in the first reconstruction algorithm.) Thus, it may be practical to enumerate q_i and p_i even if k is fairly large ($k \leq 20$).

To reconstruct the expected number of meioses recombinant for a given interval, we proceed as follows:

(*i*) Recursively compute the left-conditioned probability distribution p_i^L for $\mathbf{v}_i$ conditional on all data for loci $M_1, \ldots, M_i$. Given p_i^L, q_{i+1}, and $T(\theta_i)$, apply Bayes' theorem to obtain p_{i+1}^L:

$$p_{i+1}^L = \frac{[p_i^L T(\theta_i)] \circ [q_{i+1}]}{[p_i^L T(\theta_i)] \cdot [q_{i+1}]}, \qquad [2]$$

where $\circ$ denotes componentwise product of vectors, and $\cdot$ represents dot product.

(*ii*) Compute the right-conditioned probabilities analogously.

(*iii*) Define a matrix $T^*(\theta_i)$ as follows. Let t_{vw} be the entry of the transition matrix $T(\theta_i)$ corresponding to the transition from inheritance vector v to w and let $d(v, w)$ be the number of coordinates at which v and w differ. Define $t^*_{vw} = d(v, w)$ t_{vw} and $T^*(\theta_i) = (t^*_{vw})$. By Bayes' theorem, the expected number of recombinations between M_i and M_{i+1} is

$$\frac{[p_i^L T^*(\theta_i)] \circ [p_{i+1}^R]}{[p_i^L T(\theta_i)] \cdot [p_{i+1}^R]}.$$

This completes genetic reconstruction.

(*iv*) Note that the denominator in Eq. 2, $L_{i+1} = [p_i^L T(\theta_i)] \cdot [p_{i+1}]$, is simply the conditional probability for the data at M_{i+1}, conditioned on the data for $M_1, \ldots, M_i$. Thus, the overall likelihood is just $L(\theta_1, \ldots, \theta_{m-1}) = L_2 L_3 \ldots L_m$. Thus, the algorithm accomplishes both genetic reconstruction and likelihood calculation.

Computational Complexity. The initial probability distribution p_i may be computed using the basic approach of the Elston–Stewart algorithm for the single locus M_i. Since the p_i do not depend on the θ_i, they may be precomputed off-line when the data are first entered.

Steps *i–iii* of the algorithm require a total of $3(m-1)$ matrix multiplications of matrices of size 2^{2k}. If the p_i are sparse distributions, with support on a set of cardinality s_i, then $(s_1 s_2 + s_2 s_3 + \ldots + s_{m-1} s_m)$ multiplications are needed. On the other hand, suppose that the p_i are dense. Since the matrices $T(\theta_i)$ and $T^*(\theta_i)$ are built from Kronecker products of 2×2 matrices, each matrix multiplication can be performed with $2k2^{2k}$ multiplications using a simple "divide and conquer" approach (20), rather than 2^{4k} multiplications. The worst case computation time is then $O(6mk2^{2k})$, although considerably less time is needed the more that is known about the inheritance vectors.

For a given pedigree, the computation time scales linearly with the number of loci studied, rather than exponentially as in the case of Elston–Stewart: studying 10 intervals takes only 10 times as long as studying 1 interval. Of course, the scaling constant limits the size of pedigrees that may be studied: no more than 10–25 nonoriginals is probably practical, the exact limit depending on the informativeness of the phenotypes. Nevertheless, a great many pedigrees of interest fall into this class.

Practical Implementation

As part of an international collaboration coordinated by the Centre d'Etude du Polymorphisme Humaine (CEPH), human geneticists are currently scoring hundreds of RFLPs in 40 three-generation families consisting of four grandparents, two parents, and many children. To explore the practicality of the approaches described above, we wrote preliminary computer programs implementing them for CEPH-type pedigrees. The programs were used to study segregation data for some 60 RFLP loci on human chromosome 7 in ≈25 CEPH families, gathered by Donis-Keller and colleagues at Collaborative Research. For any given probe, about one-third of meioses were informative, of which about one-half were phase-known. The results of the genetic mapping will be reported elsewhere (David Barker, P.G., Robert Knowlton, James Schumm, Arnold Oliphant, E.L., Gina Akots, Valerie Brown, Thomas Gravius, Cynthia Helms, Christopher Nelson, Carol Parker, Kenneth Rediker, and Helen Donis-Keller, unpublished results).

(*i*) We first wrote a computer program, called MAPMAKER, to analyze an unbiased subset consisting of genotype- and phase-known data, using the first genetic reconstruction algorithm described above. Fig. 1 shows a representative example (from among >20,000 uses): studying 16 loci simultaneously, the program converged to the maximum likelihood map in 9 sec, after 12 iterations. (Convergence was declared when the $\log_{10}$ likelihood increased by <0.01, after having shown clear linear approach. We frequently performed a further 50 iterations to confirm that convergence was complete.)

The number of iterations required for convergence varied with the informativeness and, less importantly, with the number of markers. In general, 20–30 iterations were sufficient when about a dozen RFLPs were mapped simultaneously. Using a simple acceleration technique (p. 24 in ref. 16) to project the target of the linear convergence, the number of iterations was roughly halved to 10–15.

(*ii*) To study the meioses with ambiguous phase, we wrote an extension to MAPMAKER implementing an EM search using the hidden Markov-chain approach. The program typically required 3–5 min to converge (running on an HP9000 computer) when 16 loci were studied. Slightly fewer iterations were typically required than in the phase-known case, presumably because more data were being included (16). By contrast, the traditional approach would have required years of computer time.

Based on these results, an EM search using a hidden Markov-chain approach for genetic reconstruction appears to be the method of choice for simultaneous analysis of any number of RFLP markers in the CEPH pedigrees. We are now rewriting the MAPMAKER program for general distribution to interested investigators. [We should note that other nontraditional approaches are being pursued by other investigators (cf. ref. 11).]

We have not yet implemented this approach for arbitrary genetic systems or general pedigrees. Although the theory demonstrates the favorable asymptotic scaling properties, the practical limitations upon pedigree size will only be known when complete computer programs are written.

	Iteration	Recombination Fractions															log(Likelihood)
Initial	0	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	-351.45
	2	.01	.05	.18	.09	.13	.10	.08	.09	.08	.08	.14	.09	.09	.08	.21	-306.68
	4	.01	.04	.22	.08	.16	.11	.08	.10	.08	.08	.18	.08	.08	.06	.24	-304.25
	6	.01	.03	.24	.07	.17	.11	.08	.11	.08	.08	.20	.08	.08	.06	.25	-303.66
	8	.01	.03	.25	.06	.18	.11	.07	.11	.08	.08	.22	.08	.08	.05	.25	-303.43
	10	.01	.03	.25	.06	.18	.11	.07	.11	.08	.07	.22	.07	.08	.05	.25	-303.34
Final	12	.01	.03	.25	.05	.19	.11	.07	.11	.08	.07	.23	.07	.08	.05	.25	-303.28

FIG. 1. Example of multipoint linkage analysis using EM algorithm, showing convergence to maximum-likelihood genetic map for 16 RFLPs on human chromosome 7, studied in CEPH families (see text). The initial assumption of 5% recombination between consecutive RFLPs corresponded to a $\log_{10}$ likelihood of −351.45. After 12 iterations, the recombination fractions converged to a map that was $\approx 10^{48}$ times more likely to have produced the observed data. The analysis used the first genetic reconstruction algorithm discussed in the text, involving only genotype-known data, and it required ≈9 sec on an HP9000 minicomputer. Analysis of the full data set, using the hidden Markov-chain reconstruction algorithm, required ≈4 min and did not alter the recombination fractions significantly.

Determining Gene Order

Gene order is typically not known. Combinatorial optimization techniques, however, can be used together with the methods described above to find the gene orders yielding maximum-likelihood maps with the highest likelihoods. We have found the following satisfactory: (*i*) exhaustive search for up to eight loci; (*ii*) branch-and-bound search (21), with likelihood as the criterion for bounding and with the most informative loci tried first; and (*iii*) simulated annealing (22) with log likelihood used as energy function and with a random walk over gene orders generated by transpositions.

A number of excellent techniques using criteria other than likelihood have also been proposed, including crossover minimization and seriation (23).

Discussion

The construction of multilocus linkage maps in humans is formulated above as a "missing data" problem, amenable to solution by the EM algorithm. To apply the method, one requires an efficient solution to the genetic reconstruction problem. Three algorithms are described above, each highly efficient in certain situations: (*i*) a simple allocation scheme applicable to data in which genotypes and phases are known; (*ii*) a modification of the Elston–Stewart algorithm appropriate for studying a few loci in pedigrees of any size; (*iii*) a hidden Markov-chain algorithm appropriate for studying any number of loci in pedigrees with fewer than ≈20 nonoriginals.

We should note that Ott (24, 25) explored an EM-type algorithm for linkage analysis over a decade ago, but explicitly rejected it as impractical. Ott defined θ^{new} via Eq. 1, rather than using separate E and M steps. Since an expression for the derivative of the likelihood function is unavailable for most problems, Ott eventually decided (26) that the method was of very limited usefulness. [EM was suggested for phase-known data, however, by Thompson (27)]. By highlighting the availability of genetic reconstruction algorithms, we hope to revive interest in the potential uses of the EM method in human linkage mapping, most of which Ott foresaw in his important papers (24, 25).

For CEPH pedigrees, any number of RFLPS may be simultaneously mapped in minutes by using the hidden Markov-chain approach. This solves the computational bottleneck in constructing a complete RFLP linkage of the human genome. The power and limitations of such methods remain to be explored for more general pedigrees and genetic systems.

Finally, even in experimental organisms such as maize, RFLP maps are most efficiently made via F_2 intercrosses, despite the fact that some phases remain ambiguous. Multilocus analysis, using Markov reconstruction, provides an efficient way to extract the full information from the data.

We thank Persi Diaconis, David Botstein, and Helen Donis-Keller for many helpful discussions. We are grateful to two referees, whose comments led to the clarification of several points in the paper. Aaron Barlow, Lee Newburg, and Mark Daly provided invaluable assistance programming and offered many insightful conversations. This work was partially supported by grants from the System Development Foundation and National Science Foundation (E.S.L.).

1. Fisher, R. A. (1935) *Ann. Eugen.* **6,** 187–201.
2. Haldane, J. B. S. & Smith, C. A. B. (1947) *Ann. Eugen.* **14,** 10–31.
3. Morton, N. (1955) *Am. J. Hum. Genet.* **7,** 277–318.
4. Elston, R. C. & Stewart, J. (1971) *Hum. Hered.* **21,** 523–542.
5. Ott, J. (1976) *Am. J. Hum. Genet.* **28,** 528–529.
6. Botstein, D., White, R. L., Skolnick, M. H. & Davis, R. W. (1980) *Am. J. Hum. Genet.* **32,** 314–331.
7. Lander, E. S. & Botstein, D. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7353–7357.
8. Lander, E. S. & Botstein, D. (1986) *Cold Spring Harbor Symp. Quant. Biol.*, in press.
9. Lathrop, G. M., Lalouel, J. M., Julier, C. & Ott, J. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3443–3446.
10. Lathrop, G. M. & Lalouel, J. M. (1984) *Am. J. Hum. Genet.* **36,** 460–465.
11. Lathrop, G. M., Lalouel, J. M. & White, R. L. (1986) *Genet. Epidemiol.* **3,** 39–52.
12. Morton, N. E., MacLean, C. J., Lew, R. & Yee, S. (1986) *Am. J. Hum. Genet.* **38,** 868–883.
13. Lalouel, J. M. (1979) *Technical Report No. 14* (Univ. Utah, Salt Lake City, UT).
14. Ralston, A. (1965) *A First Course in Numerical Analysis* (McGraw-Hill, New York).
15. Foulds, L. R. (1981) *Optimization Techniques* (Springer, New York).
16. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *J. R. Statist. Soc. Ser.* **39,** 1–38.
17. Wu, C. F. J. (1983) *Ann. Stat.* **11,** 95–103.
18. Vardi, Y., Shepp, L. A. & Kaufman, L. (1985) *J. Am. Stat. Assoc.* **80,** 8–37.
19. Lange, K. & Elston, R. C. (1975) *Hum. Hered.* **25,** 95–105.
20. Aho, A. V., Hopcroft, J. E. & Ullman, J. D. (1974) *The Design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, MA).
21. Knuth, D. (1968) *The Art of Computer Programming: Fundamental Algorithms* (Addison-Wesley, Reading, MA).
22. Kirkpatrick, S., Gellatt, C. D. & Vecchi, M. P. (1983) *Science* **220,** 671–680.
23. Buetow, K., Chakravarti, A., Murray, J. & Ferrel, R. (1985) *Am. J. Hum. Genet.* **37,** Suppl. A190.
24. Ott, J. (1977) *Ann. Hum. Genet.* **40,** 443–454.
25. Ott, J. (1979) *Am. J. Hum. Genet.* **31,** 161–175.
26. Ott, J. (1985) *Analysis of Human Genetic Linkage* (Johns Hopkins, Baltimore).
27. Thompson, E. (1984) *IMAJ Math. Appl. Med. Biol.* **1,** 31–50.