Genetic Epidemiology 16:337–343 (1999)

Method for Constructing Confidently Ordered Linkage Maps

Karl W. Broman* and James L. Weber

Marshfield Medical Research Foundation, Marshfield, Wisconsin

We describe a method for identifying, from a comprehensive genetic map, the most dense framework of confidently ordered markers. The approach uses the number of observed recombination events between each pair of markers, and finds the largest subset of markers for which adjacent loci are separated by at least one recombination. We illustrate the approach using a short region of chromosome 7p. Genet. Epidemiol. 16:337–343, 1999. © 1999 Wiley-Liss, Inc.

Key words: genetic map; dynamic programming; framework map

INTRODUCTION

Multipoint analyses to identify disease susceptibility genes require that genetic markers be correctly ordered; an incorrect ordering of markers compromises the linkage evidence for a disease gene. While the order of genetic markers will ultimately be obtained from sequence data, currently one must rely on genetic linkage or radiation hybrid data. Generally, genetic maps are presented with a framework map of markers that are confidently ordered, additional markers being assigned to likely intervals. The choice of markers placed on the framework map can be ad hoc.

We describe an algorithm that, beginning with a comprehensive genetic map, identifies the largest possible framework of confidently ordered markers. The algorithm uses the pairwise recombination information described by Fain et al. [1995, 1996], and determines the largest subset of markers having adjacent anchor markers separated by at least one (or two) recombination events. A two-recombinant rule is somewhat more conservative than the LOD 3.0 criterion in common use [Fain et al., 1995].

*Correspondence to: Karl W. Broman, Center for Medical Genetics, Marshfield Medical Research Foundation, 1000 N. Oak Ave., Marshfield, WI 54449. E-mail: BromanK@cmg.mfldclin.edu

Received 1 May 1998; Revised 17 July 1998; Accepted 18 July 1998

Contract grant sponsor: NHLBI; Contract grant number: N01-HV-48141.

The algorithm is an example of dynamic programming [Bellman, 1957; Bellman and Dreyfus, 1962], which uses recursion and the storage of intermediate results to optimize a function (in this case, to maximize the number of markers in the framework map) without performing a complete enumeration of all possibilities. Dynamic programming has also been used to align sets of DNA or amino acid sequences [Needleman and Wunsch, 1970]. The algorithm requires a comprehensive map with the markers in approximately the correct order and a table containing the number of observed recombinations between each pair of markers. Such a table may be obtained using the output from the *chrompic* option of the CRI-MAP program [Green et al., 1990].

In this communication, we describe our algorithm and illustrate its use on a small region of chromosome 7p.

METHODS

Consider a set of linked markers numbered 1,2,..., M, in approximately the correct order. Assign a weight w_i to marker i. (For example, $w_i = 1$ for all i or $w_i = -\log(1 - \text{het}_i)$ where het_i is the heterozygosity of marker i.) Let $R(i,j)$ be the number of observed recombination events between markers i and j.

We wish to identify the subset of markers $\{i_1, i_2,...,i_k\}$, with $i_1 < i_2 < ... < i_k$, where $R(i_j,i_{j+1}) \geq R_{\min}$ for all $j = 1,..., k - 1$ and where $\Sigma_{j=1}^{k} w_{i_j}$ is maximized.

When $w_i = 1$ for all i, this is equivalent to choosing the largest subset for which adjacent markers have at least $R_{\min}$ observed recombinations. When the markers do not have identical weights, some markers are preferred over others. For example, if the markers are in linkage equilibrium in a population, and if three nearby markers have heterozygosities het_1, het_2, and het_3 satisfying $w(\text{het}_1) + w(\text{het}_2) = w(\text{het}_3)$ where $w(\text{het}) = -\log(1 - \text{het})$, then the chance that an individual is heterozygous for at least one of markers 1 and 2 is equal to the chance that the individual is heterozygous for marker 3. In such a situation, one may consider marker 3 to be equivalent to the pair of markers 1 and 2.

Our solution is based on the following. Let p_j denote the optimal subset of $\{1,..., j\}$ containing the marker j. (That is, p_j is the subset of $\{1,..., j\}$ containing j, and with adjacent markers showing at least $R_{\min}$ recombinations, for which the sum of the weights is maximized.) If the overall optimal subset of $\{1,..., M\}$ contains the marker j, then the portion of this overall optimal subset which is $\leq j$ must be exactly p_j. Thus we may build up the optimal subset from left to right along the map, at marker j storing the optimal subset of $\{1,..., j\}$ containing j.

Now let p_j be defined as above, and let b_j be the total weight of the subset p_j. The p_j are formed in a stepwise fashion. First, let $p_1 = \{1\}$ with weight $b_1 = w_1$. Next suppose we have formed $p_1,..., p_{j-1}$ with weights $b_1,..., b_{j-1}$. If there is no $i < j$ with $R(i,j) \geq R_{\min}$, then $p_j = \{j\}$ with weight $b_j = w_j$. Otherwise, let

$$i^* = \arg \max_{\substack{i < j \\ R(i,j) \geq R_{\min}}} b_i$$

TABLE I. Numbers of Observed Recombinations Between Each Pair of Markers, for a Set of 25 Genetic Markers From Chromosome 7p*

	R_{min}				Marker																								
No.	1	2	Marker	het	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	•	•	AFMb035xb9	0.67		1	2	1	1	3	3	4	3	2	5	2	6	7	6	8	8	9	3	5	13	19	14	4	18
2	•		UT5195	0.64			1	1	2	1	1	4	3	0	4	3	5	5	5	3	5	1	3	5	6	9	8	9	8
3	•		AFMa090xg1	0.77				1	1	0	0	3	3	0	4	2	6	5	6	10	8	5	1	1	7	13	14	5	13
4	•		AFMb017yh1	0.75					0	0	1	2	2	0	3	1	3	2	4	10	6	6	0	2	9	13	15	5	14
5			D7S21	0.66						0	1	2	2	0	3	2	3	3	3	2	3	1	2	3	4	6	6	6	6
6		•	AFM185yh2	0.73							2	3	3	1	4	1	4	4	5	8	7	8	2	4	10	15	14	5	15
7	•	•	AFM254yc9	0.77								0	0	0	2	1	3	2	2	7	5	4	1	3	10	13	14	3	17
8			AFMa136zh1	0.60									0	0	2	2	4	3	3	7	5	4	1	3	10	13	11	5	13
9			AFMc027xb5	0.66										0	2	1	3	2	1	6	3	5	1	1	8	10	11	1	15
10			AFMb286yc6	0.50											2	1	3	2	2	6	4	3	0	2	9	11	10	3	11
11			AFM210xc7	0.80												0	0	0	0	2	0	1	1	3	5	7	7	3	9
12			Mfd172	0.61													0	0	0	0	0	0	1	3	2	3	4	3	5
13			GATA24F03	0.73														0	1	5	3	3	1	3	7	10	12	2	14
14	•	•	AFM225xa1	0.83															1	3	2	1	1	3	3	7	7	3	7
15	•		AFMa224wh9	0.79																5	3	3	1	3	6	9	12	3	14
16	•	•	AFM049xe3	0.84																	0	0	1	3	3	4	7	3	10
17			GATA61G06	0.62																		0	1	1	0	1	4	0	6
18			AFMb040zb5	0.63																			1	1	1	2	4	1	7
19			UT626	0.26																				0	0	0	1	0	1
20			UT7600	0.63																					0	0	1	0	1
21			AFMa062yf9	0.60																						0	2	0	4
22	•	•	AFMc011yc9	0.85																							3	1	7
23	•	•	AFM224yb6	0.72																								0	3
24			UT5023	0.47																									0
25	•	•	GATA119B03	0.69																									

*The second and third columns, labeled ''R_{min},'' indicate which markers were chosen in the framework maps when requiring at least one or two recombinations between adjacent markers. The column labeled ''het'' gives the heterozygosities of the markers.

Then $p_j = p_{i^*} \cup \{j\}$with total weight, $b_j = b_{i^*} + w_j$.

Finally, having formed $p_1,..., p_M$ with total weights $b_1,..., b_M$, let

$$j^* = \arg \max_{1 \le j \le M} b_j.$$

Then p_{j^*} is the optimal subset of $\{1,..., M\}$. The optimal subset is not necessarily unique, though the above algorithm is guaranteed to produce one of the optimal subsets.

RESULTS AND DISCUSSION

We illustrate our method using 25 markers from chromosome 7p, taken from a recent comprehensive genetic map of the human genome [Broman et al., 1998]. These markers span approximately 17 cM. Table I contains the markers in their approximate order, their estimated heterozygosities, and, for each pair of markers, the number of observed recombinations in eight of the CEPH families (1331, 1332, 1347, 1362, 1413, 1416, 884, and 102), which comprise approximately 180 meioses. Figure 1 contains a representation of the sex-averaged map of these markers, taken from Broman et al. [1998]. The map was formed using the CRI-MAP program [Green et al., 1990]. Several groups of markers map to the same locus. LOD scores, indicating the local support for the marker order, were obtained using the *flips* option of CRI-MAP: adjacent pairs of markers were exchanged and the change in likelihood noted. In this process, markers mapping to exactly the same locus were kept together.

The second and third columns in Table I, labelled "$R_{\min}$," indicate the subsets of markers chosen by the algorithm when $R_{\min} = 1$ and 2. When $R_{\min} = 1$, a greater number of markers are used in forming the framework map, but with a greater risk for errors in the marker order. The genetic map of the subset of markers obtained using $R_{\min} = 2$ is displayed in Figure 1, along with LOD scores indicating the local support for the marker order.

Table II contains a list of the subsets p_j and their total weights b_j for the data in Table I, when using $R_{\min} = 2$. In forming, for example, p_{12}, one notes that marker 12 shows two or more recombinations with markers 1, 2, 3, 5 and 8. Among p_1, p_2, p_3, p_5 and p_8, the subset p_8 has the greatest weight, and so $p_{12} = p_8 \cup \{12\} = \{1, 3, 8, 12\}$, with weight $b_{12} = b_8 + w_{12} = 3.5 - \log(1 - 0.61) = 4.4$.

It is important to note that the table of observed recombinations depends on the initial ordering of the markers, since the order of the markers is sometimes used to infer phase, and phase is assumed to be correctly known when counting the numbers of recombinations. In addition, the method we describe here does not revise the order of the markers, but rather extracts the largest subset of markers that are believed to be confidently ordered. Thus, the algorithm requires an initial marker order of reasonable quality.

What is important, in the initial comprehensive map, is that the recombination events be correctly placed. Thus, the required quality of the initial order depends on the resolution provided by the available meiotic data. Groups of markers that did not recombine may be ordered arbitrarily, but if two markers are both informative in a

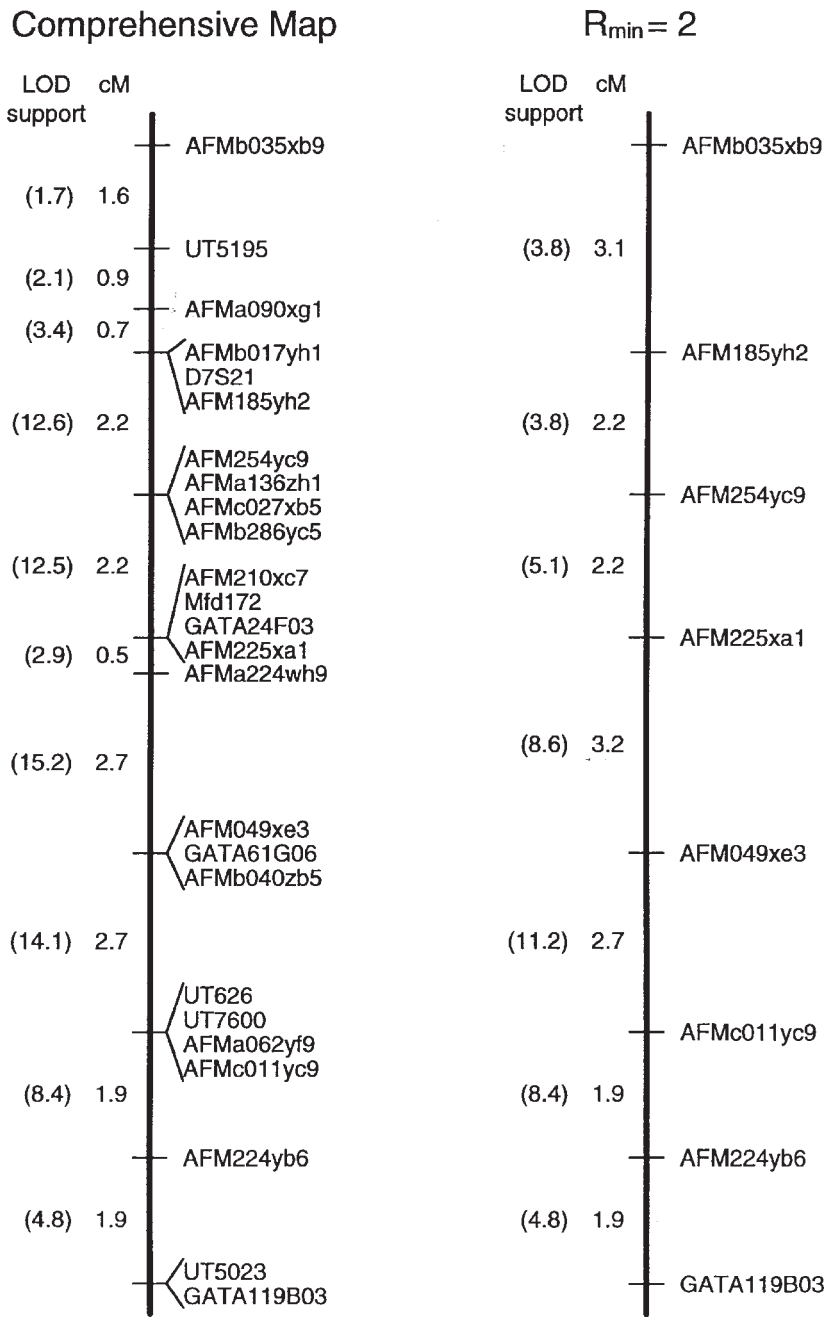

Fig. 1. Comprehensive genetic map of a set of 25 markers from chromosome 7p and the framework map obtained using $R_{min} = 2$. LOD support values, indicating the local support for order, were obtained using the *flips* option of CRI-MAP.

TABLE II. Subsets p_j and Their Total Weights b_j Formed When Constructing the Overall Optimal Subset of Markers for the Data in Table I When Using $R_{min} = 2$

j	p_j	b_j
1	{1}	1.1
2	{2}	1.0
3	{1, 3}	2.6
4	{4}	1.4
5	{2, 5}	2.1
6	{1, 6}	2.4
7	{1, 6, 7}	3.9
8	{1, 3, 8}	3.5
9	{1, 3, 9}	3.7
10	{1, 10}	1.8
11	{1, 6, 7, 11}	5.5
12	{1, 3, 8, 12}	4.4
13	{1, 6, 7, 13}	5.2
14	{1, 6, 7, 14}	5.7
15	{1, 6, 7, 15}	5.4
16	{1, 6, 7, 14, 16}	7.5
17	{1, 6, 7, 14, 17}	6.7
18	{1, 6, 7, 15, 18}	6.4
19	{1, 6, 19}	2.7
20	{1, 6, 7, 14, 16, 20}	8.5
21	{1, 6, 7, 14, 16, 21}	8.4
22	{1, 6, 7, 14, 16, 22}	9.4
23	{1, 6, 7, 14, 16, 22, 23}	10.6
24	{1, 6, 7, 14, 16, 24}	8.1
25	{1, 6, 7, 14, 16, 22, 23, 25}	11.8

meiosis in which a recombination occurred between them, their initial order needs to be correct. If the initial placement of a marker is far from its true location, our algorithm is likely to incorporate that marker into the framework map, and it will be placed in the incorrect position.

While the optimal subset of markers (by our criterion) is not necessarily unique, a simple modification of our algorithm will allow the identification of all of the optimal subsets: one retains, at each step in the algorithm, information on all of the optimal subsets p_j rather than a single optimal subset.

Our method should prove useful in obtaining longer and more robust framework maps than can be generated by hand. The algorithm is most applicable when markers are at a density of less than around 2 cM, since otherwise most adjacent markers will be separated by more than two recombination events. While we describe its use for genetic maps, the approach could also be used for radiation hybrid maps and other maps that rely on breakpoint- or recombination-like events.

We have implemented our algorithm in an Internet query program available at the Marshfield Web site (http://www.marshmed.org/genetics). The user submits a list of markers from a single chromosome and indicates whether the markers should be weighted equally or according to –log(1 – het) and whether to use $R_{min} = 1$ or 2, and receives the locations of the markers on the Marshfield comprehensive maps, the table of observed recombinations, and a framework map for those markers, as determined by this algorithm.

ACKNOWLEDGMENTS

Mark Neff generously provided comments for the improvement of this manuscript. This work was supported in part by NHLBI contract N01-HV-48141 for the Mammalian Genotyping Service.

REFERENCES

Bellman RE. 1957. "Dynamic Programming." Princeton, NJ: Princeton University Press.

Bellman RE, Dreyfus SE. 1962. "Applied Dynamic Programming." Princeton, NJ: Princeton University Press.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. Am J Hum Genet 63:861–869.

Fain PR, Kort EN, Chance PF, Nguyen K, Redd DF, Econs MJ, Barker DF. 1995. A 2D crossover-based map of the human X chromosome as a model for map integration. Nat Genet 9:261–266.

Fain PR, Kort EN, Yousry C, James MR, Litt M. 1996. A high resolution CEPH crossover mapping panel and integrated map of chromosome 11. Hum Mol Genet 5:1631–1636.

Green P, Fall K, Crooks S. 1990. Documentation for CRI-MAP, version 2.4.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 48:443–453.