Analysis of Path-Based Approaches to Genomic Physical Mapping

by
Alan Rimm-Kaufman
James Orlin

WP #3742-94 November 1994
revised December 1994

Analysis of path-based approaches to genomic physical mapping

Alan Rimm-Kaufman and James B. Orlin[1]

[1] A. Rimm-Kaufman, Operations Research Center, MIT, Cambridge, MA. 02139, USA

J. B. Orlin, Sloan School of Management and the Whitehead Institute MIT Center for Genome Research, MIT, Cambridge, MA. 02139, USA

Abstract

An integrated genomic physical map combines multiple sources of data to position landmarks and clones along a genome. "Path-based" strategies for constructing integrated maps employ paths of overlapping clones to bridge intervals between genetically mapped markers. There is a difficulty with path-based strategies: a small rate of clone overlap error may create numerous spurious paths and result in many false linkages in the physical map. We propose an analytical method to evaluate path-based approaches in light of this difficulty, and demonstrate the method on an integrated map of the human genome.

Integrated genomic physical maps combine many varieties of structural and positional data to order clones and landmarks along the genome. Genetically-mapped sequence tagged sites (STSs) can provide a backbone for an integrated physical map, with overlapping yeast artificial chromosomes (YACs) used to provide coverage of the genetic intervals between adjacent markers. The overlaps between YACs may be detected through a variety of means including fingerprinting and hybridization data.

In order to discuss the results of our analysis to integrated physical maps, we first introduce some terminology. We consider a data set in which each STS is assigned to a genetic locus. A path of length k between genetic locus i and genetic locus j is defined as two STSs, s_i and s_j, and k YACs, y_1, y_2, ..., y_k such that (i) s_i genetically maps to locus i, (ii) s_j maps to locus j, (iii) s_i lies in y_1 and s_j lies in y_k by STS content mapping, and (iv) for each step (y_i,y_{i+1}) the data indicate that the YACs y_i and y_{i+1} overlap. Paths of length 1 include only one YAC and correspond to traditional STS content mapping, while longer paths depend on the detection of YAC overlaps. An integrated map will assign a YAC to an interval connecting genetically close loci if there is a short path joining the two loci that contains the YAC.

A serious concern with such path-based mapping strategies is that falsely detected YAC overlaps can create short false paths between genetically-mapped STSs, resulting in spurious assignment of YACs to genetic intervals, and to spurious coverage of the genome. It is known in random graph theory (1) that, in certain random

structures, paths of bounded length suffice to connect essentially all pairs of points. This phenomenon has recently gained popular attention through John Guare's award-winning play, "Six Degrees of Separation," in which it is asserted that any two people in the world can be connected through a path of at most six acquaintances. Within an integrated genomic physical map, this same phenomenon could occur even if the majority of the clone overlaps are valid. A small rate of falsely declared clone overlaps can lead to spurious coverage of large regions of the genome.

To address this concern, we developed a methodology to assess the probability that a shortest path joining two loci is correct. We then applied this methodology to the CEPH-Genethon 'first-generation' integrated physical map of the human genome (2). We describe our methodology in the context of this integrated physical map.

Briefly, the CEPH-Genethon physical mapping data involved screening the 33,000-clone CEPH mega-YAC library by two different methods, STS content mapping and Alu-PCR probe hybridization. In the first method, 2100 genetically-mapped STSs (3) were screened against the YAC library (with half the STSs screened completely and half screened partially to obtain 1-2 positives). In the second method, Alu-PCR products were prepared from 5300 individual YACs and were screened by hybridization against spotted Alu-PCR products from a subset of 25,000 YACs and monochromosomal hybrid cell lines. In addition, many YACs were subjected to hybridization-based 'fingerprinting' (4).

A *path* of length k between genetic locus i and genetic locus j is defined as before. Two YACs are determined to overlap if (a) at least one of the two YACs was used as an Alu-PCR probe and hybridized to the other YAC, or (b) the two YACs had similar restriction fragment fingerprints (4). A *chromosomally allowable path* is defined as a path where (i) loci i and j lie on the same chromosome, c, and (ii) each y_i that was used as an Alu-PCR probe either gave no signal when hybridized to the monochromosomal panel or hybridized to a set of chromosomes that included chromosome c. (For technical reasons, chromosomal assignments were not always unique: 50% could be assigned to a single chromosome, 19% hybridized to multiple chromosomes, and 31% could not be assigned to any chromosome.)

The CEPH-Genethon integrated physical map was defined (2) to be the set of all chromosomally allowable paths of a specified length connecting loci within 10 cM. Neither analytic nor experimental analysis was advanced to suggest that most paths of a given length were correct. It was noted that as longer paths are allowed, the coverage of the genome increases. With paths of length 1, 3, 5 and 7 YACs, the strategy covered 11%, 30%, 70%, and 87%, respectively, of the total genetic length of the genome. (The percentage coverage is defined in (2) as the proportion of total centiMorgans lying between connected loci. This is a conservative measure of coverage when restricted to valid paths, since the coverage does not account for ends of contigs that extend beyond the genetic loci. For example, YACs creating a path between two

genetically indistinguishable STSs do not provide incremental coverage on the genome by this definition.)

We set out to evaluate the CEPH-Genethon paths using data from the February, 1994 CEPH-Genethon data release (5) and using the publically released computer package QUICKMAP (6) which accompanies the CEPH-Genethon data. We first constructed the minimum length chromosomally allowable path connecting every pair of loci on the same chromosome -- regardless of the genetic distance between them (7). Figure 1 shows the proportion of loci that may be connected, as a function of the path length and the width of the genetic interval between the loci. We were interested in determining what fraction of these observed paths were valid and what fraction were spurious.

A simple way to estimate the proportion of false connections is to consider loci separated by ≥50 cM and connected by short paths. Such paths must surely be spurious inasmuch as the average YAC length is 1 Mb, corresponding to only about 1 cM in the human genome. The proportion of such widely-separated loci pairs connected by chromosomally allowable paths of length 1, 3, 5, and 7 is 0.05, 4, 33, and 63%, respectively. In particular, the curve rises dramatically for path lengths exceeding four, indicating that random connections dominate at these distances. The proportion of connected loci pairs at distances 5-10, 10-20, and 20-50 cM was not much higher than for loci pairs at distances ≥ 50 cM. This suggests that many of these paths are also spurious, and that such

spurious paths span close loci pairs at approximately the same rate as distant loci pairs.

We attempted to estimate the proportions of valid paths. We partitioned the pairs of loci into six groups on the basis of their inter-locus distance: the groups consisted of loci at 0-2, 2-5, 5-10, 10-20, 20-50, and > 50 cM, respectively. For a randomly chosen pair of loci at distance range d, let A_d denote the length of the shortest path between them, let V_d denote the length of the shortest valid path between them, and let I_d denote the length of the shortest invalid path between them. With these definitions, $A_d = \min(V_d, I_d)$. Figure 1, then, may be viewed as the cumulative probability distribution of A_d for each of the six groups.

We next made two assumptions: (i) the frequency of spurious paths between two loci is independent of the distance between them, and (ii) the length of the shortest spurious path joining two loci is independent of the length of the shortest valid path. Under these assumptions, one can estimate an empirical joint probability density for V_d and A_d.

Let $P_d(a)$ = the probability that two points in distance range d are connected by a valid path of length $\leq a$. Let $S_d(a)$ = the probability that two points in distance range d are connected by a spurious path of length $\leq a$. Let $\pi_d(a)$ = the probability that two points in distance range d are connected by any path of length $\leq a$. It follows from the independence of I_d and V_d that $\pi_d(a) = P_d(a) + S_d(a) - P_d(a)S_d(a)$. Thus, $P_d(a) = (\pi_d(a) - S_d(a))/(1-S_d(a))$. The quantity $\pi_d(a)$ is directly observable. We assume that $S_d(a)$ is

independent of d, and so can be calculated from points separated by at least 50 cM. Therefore, $P_d(a)$ can be calculated.

One can then estimate, for each distance group and each path length, the probability $P(V_d=a \mid A_d=a)$ that there is at least one valid shortest path among the observed shortest paths (Figure 2). This is the probability $P(V_d=a \ \& \ A_d=a)/P(A_d=a)$. The denominator is $\pi_d(a) - \pi_d(a-1)$. The numerator is the probability there is a valid path of length a and there is no path (valid or invalid) of length at most a-1. Under the independence of V_d and I_d, the numerator is equal to $(P_d(a) - P_d(a-1))(1-S_d(a-1))$. With this probability, we provide a corrected estimate of the observed proportion of loci pairs connected by at least one valid path (Figure 3) by subtracting out the expected fraction of loci pairs whose shortest paths are invalid.

The results indicate that for two loci within 5 cM of one another and spanned by shortest paths of ≤ 3 YACs, the shortest paths are likely to include at least one valid path. On the other hand, if the shortest paths are ≥ 4 YACs or if the two loci are more than 5 cM apart, then the shortest paths are likely to be invalid. Considering only shortest paths of at most 3 YACs between loci within 5 cM, and using the definition of coverage given in (2), the CEPH/Genethon data cover only about 27% of the human genome. This coverage of the genome is far less than indicated in (2).

The previous analysis was carried out on the complete data set including both Alu-PCR connections and fingerprint data. We conducted two additional sets of analyses -- one with Alu-PCR but

no fingerprints, and one with fingerprint data but no Alu-PCR -- to investigate whether the coverage by valid paths might increase. The amount of valid coverage may possibly increase with the omission of a data source if the omitted data has provided many spurious connections and few valid ones. The results in both of these analyses were qualitatively the same, and both additional cases resulted in a decreased coverage of the genome as compared to the first analysis. In the analysis involving Alu-PCR connections but no fingerprints, for two loci within 5 cM of one another and spanned by shortest paths of ≤ 3 YACs, the shortest paths are likely to include at least one valid path. On the other hand, if the shortest paths are ≥ 4 YACs or if the two loci are more than 5 cM apart, then the shortest paths are likely to be invalid. (For example, for two loci that are within 2 cM of each other, and spanned by a shortest path of 4 YACs, the probability is 55% that the shortest path is invalid.) Considering only shortest paths of at most 3 YACs between loci within 5 cM, the Alu-PCR data (without fingerprints) cover 24% of the human genome. In the analysis including fingerprints data but no Alu-PCR connections, the fingerprint data generate longer valid paths: loci within 5 cM of one another and spanned by shortest fingerprint-only paths of ≤ 5 YACs are likely to be covered by at least one valid fingerprint-only path. While the fingerprint data support longer valid paths, the relative paucity of these paths leads to lower genomic coverage overall. Fingerprint-only paths of ≤ 5 YACs between loci within 5 cM cover only 21% of the human genome.

These specific analyses address neither the quality of the CEPH/Genethon data nor alternative path-based strategies.(8) These calculations consider only the quality of shortest paths in the CEPH/Genethon data as generated from QUICKMAP. A larger fraction of reliable paths might be obtained using alternative analysis strategies, as well as additional data since added by CEPH and Genethon to their February, 1994 release. Identifying and removing noisy data, modifying the CEPH/Genethon definition of a path, considering longer length paths in addition to the minimum length paths, or employing a denser genetic map are possible alternatives that might yield a larger fraction of veracious paths covering more of the genome. Further, while our analysis can indicate the probability that there is a valid path, it does not provide a means to distinguish valid paths from invalid paths. Only additional laboratory experiments can provide final confirmation or invalidation of a path.

Aside from an application to a specific data set, we have outlined a general method for estimating the percentage of valid paths in the presence of false connections leading to invalid paths. In order to estimate the probability that a shortest path is valid, one needs to subtract out the probability that it is invalid. This, in turn, is estimated by evaluating the percentage of distance loci pairs connected by shortest paths of each length. This methodology extends to other analyses as well, including analyses using alternative criteria for determining YAC overlaps or alternative definitions of paths.

References and Notes

(1) B. Bollobas, *Random Graphs*. (Academic Press, London, 1985).

(2) D. Cohen, I. Chumakov, and J. Weissenbach, *Nature*, **366**, 698 (1993).

(3) J. Weissenbach et al., *Nature*, **359**, 794 (1992).

(4) C. Bellanne-Chantelot et al., *Cell*, **70**, 1059 (1992).

(5) These data included 6602 Alu-PCR probes and 3450 STSs. 2035 of the STSs were genetically mapped to a specific chromosomal location; the resulting genetic map contained 1266 loci. These data are the same data employed in (1), with an additional 1270 Alu-PCR probes added to the dataset by CEPH/Genethon beyond that described in (1).

(6) P. Rigault et al., available from rigault@ceph.cephb.fr.

(7) We employed CLONESPATH, one module of the QUICKMAP package, to generate chromosomally allowable paths between selected loci. CLONESPATH offers a variety of options and parameters; the results described in this paper correspond to the default CLONESPATH settings, which in turn correspond to the path creation rules described in (1). For the "link-type" parameter, we conducted three analyses: once using ALU and fingerprint links, once using only ALU links, and once using only fingerprint links. The figures correspond to using ALU and fingerprint links; the results were qualitatively the same using ALU links only.

(8) Varying the criteria used to declare YAC overlaps or modifying the criteria used to declare a genetic interval to be covered leads to alternative path-based construction strategies. A few examples involving overlap detection include (i) requiring two or more ALU probes to link two YACs, and (ii) requiring both an ALU probe and a fingerprint overlap to link two YACs. Examples involving the definition

of covering a genetic interval include (iii) requiring two or more YAC-disjoint paths to cover the genetic interval, and (iv) restricting the use of each Alu-PCR probe to a specific region on the genome.

(9) We thank D. Cohen, I. Chumakov, and J. Weissenbach for sharing this valuable data with us and the scientific community at large. We thank E. Lander, D. Page, L. Kruglyak, L. Stein, D. Cohen, I. Chumakov, and P. Rigault for helpful discussions and comments on the manuscript. Supported in part by NIH grant HG00098.

Figure Legends

Figure 1. Observed cumulative proportion of connected loci pairs, by inter-loci distance and path length. Minimal paths were constructed between all intra-chromosomal loci pairs.

Figure 2. Estimated probability that there is at least one valid shortest path spanning a inter-locus interval among all the observed shortest paths spanning that interval, by inter-locus distance and path length.

Figure 3. Estimated true proportion of connected loci pairs, by inter-locus distance and path length. The probability an interval is covered correctly is the probability that at least one of the paths spanning the interval is valid.

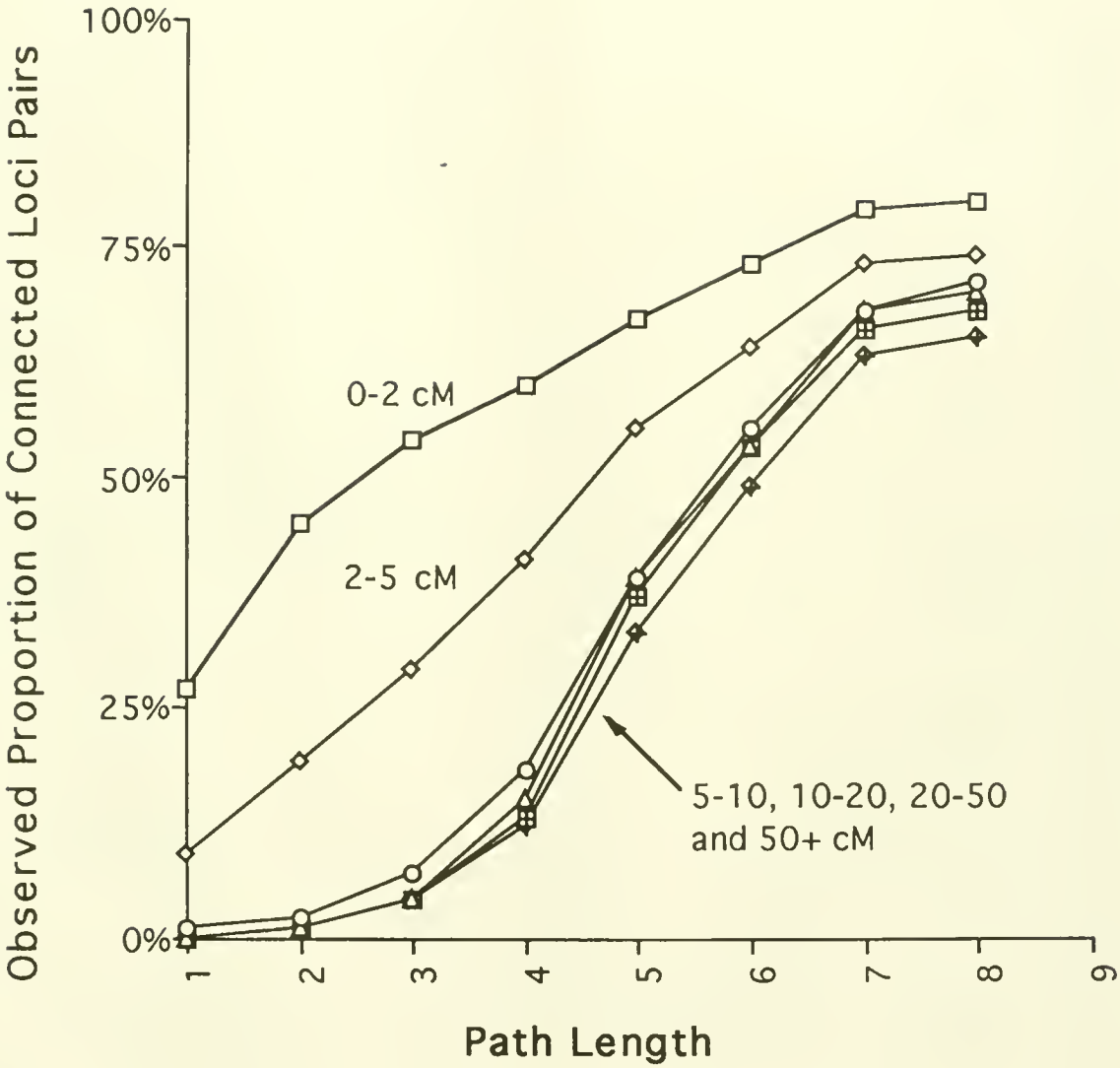
100%
75%
50%
25%
0%
Observed Proportion of Connected Loci Pairs
0-2 cM
2-5 cM
5-10, 10-20, 20-50
and 50+ cM
1
2
3
4
5
6
7
8
9
Path Length

Figure 1

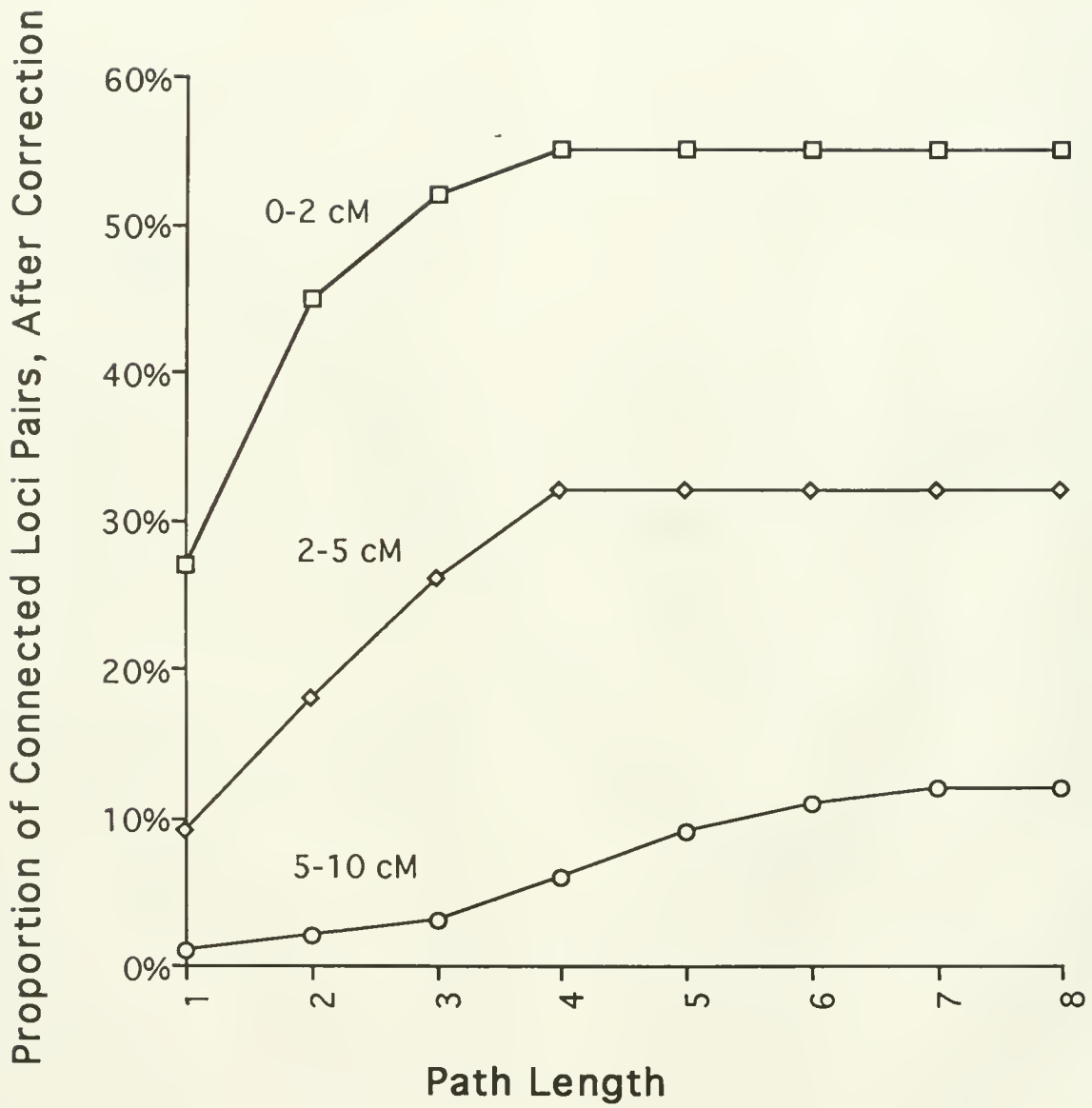
Proportion of Connected Loci Pairs, After Correction
60%
50%
40%
30%
20%
10%
0%
0-2 cM
2-5 cM
5-10 cM
1
2
3
4
5
6
7
8
Path Length

Figure 2

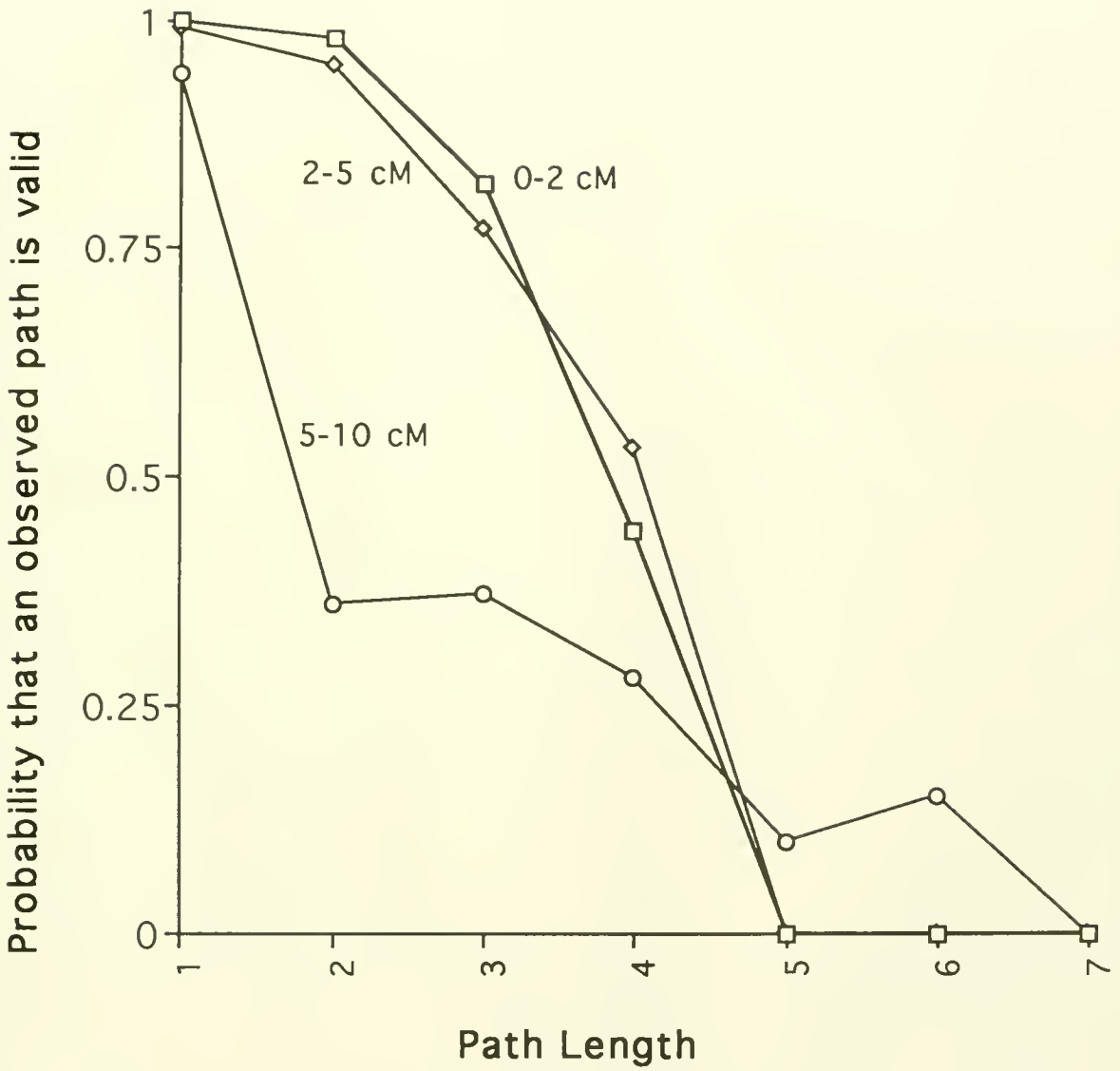

Figure 3.