18.417 Introduction to Computational Molecular Biology

Lecture 8: October 4, 2001 Scribe: Caroline Cutting
Lecturer: Lecturer Bonnie Berger Editor: Ian Chan

Physical Mapping

Why Physical Mapping?

Previously we have seen how genetic mapping techniques based on phenotypic observations may be used to obtain information about the arrangement og genes in the genome. While these techniques are often accurate to within around 3cM (3 million base pairs), genetic mapping tools alone are insufficient for a number of reasons. First, these techniques often depend upon the observation of phenotypes associated with fairly rare recombination events. Thus genes which do not correspond to distinct, easily recognized, phenotypes are difficult to track. In addition, recombination events do not occur with constant frequency throughout the genome. Rather, the genome is peppered with "hot" and "cold" spotsber of physical mapping techniques.

To compensate for many of these shortcomings, geneticists have cultivated a number of physical mapping techniques. These techniques are designed with the aim of identifying accurate and uniformly distributed markers within the genome. In general, the results of these techniques improve upon genetic mapping results and estimate gene location to within just 1cM.

A physical map is often defined as an ordered list of distinct genetic markers and the distances between them (given in megabases, Mb). Alternately, a physical map may be represented by a library of DNA pieces, ranging from 50Kb to 1Mb in length, which span a region of the genome. These DNA fragments are known as clones. A number of techniques exist to generate both of these types of physical maps.

General Method for Constructing a Physical Map

The procedure for creating a physical map begins by obtaining a copy of the DNA segment of interest. The segment is then replicated multiple times (using PCR techniques or a similar technology) and these multiple copies are broken up into clones.

Thus creating a set of DNA fragments with overlapping sequences. The clones are then isolated and, depending on the specific type of physical mapping technique used, they are categorized by a number of uniquely identifying characteristics to create a clone "fingerprint". Analysis of the similarities and identities among the clone fingerprints is used to reconstruct the order in which the overlapping clones were originally aligned on the genome.

When complete, this reconstruction may not include the entire genetic sequence of the original DNA segment, but it will provide information about the relative order and distance between markers on the clones within the sequence. Three popular techniques for performing physical mapping are Clone-by-Clone Reconstruction, Restriction Analysis, and Hybridization.

Clone-by-Clone Reconstruction

In Clone-by-Clone Reconstruction, replicated DNA segments are parsed into a set of large clones, around 130 Kb in length, called Bacterial Artificial Chromosomes (BACs). This parsing may be accomplished in a random way by forcing the DNA through a syringe repeatedly. Once the set of BACs is obtained, sequencing techniques are used to determine the 500 base pair sequence at each end of each BAC. Overlaps in these end-sequences are then used identify clones which came from adjacent segments of the original DNA. Once the relative order of all clones has been determined, a minimum tiling subset of the large clone population can be identified. This subset is chosen so that the entire genetic sequence of the original DNA segment is represented with minimum redundancy created by overlapping clones. Sequencing techniques such as shotgunning can then be used to determine the DNA sequence of any areas of interest in a very efficient manner.

Restriction Mapping

In Restriction Mapping, BAC clones representing the original DNA segment are broken into smaller segments via restriction enzyme digestion. Restriction enzymes are proteins which recognize specific, short nucleotide sequences and cut DNA only at those sites. The sizes of the DNA segments remaining after a clone is digested with a given restriction enzyme may be visualized via gel electrophoresis. The lengths of these segments represent the distances between restriction enzyme sites in the original clone and can be used to uniquely identify, or fingerprint, it. When two overlapping clones undergo digestion by a restriction enzyme, their fingerprints will share a com-

mon subset of segment lengths. The original order of the clones on the DNA segment may be reconstructed by analyzing the common subsets of clone fingerprints. Maynard and Olson have developed a greedy algorithm for reconstructing DNA segments from clone fingerprints. This algorithm dictates that whenever restriction fragment sizes are shared between two clones, those clones should be considered consecutive.

While Restriction Mapping techniques do not always allow for the unique determination of clone order, they do provide fairly robust results even when clones are analyzed in a different order, as is illustrated in the following example. In addition, Restriction Mapping techniques can be enhanced by incorporating digestion with more than one enzyme. This creates an even more unique fingerprint of fragment lengths for each clone and increases the likelihood of determining an accurate ordering.

A shortcoming of the Restriction Mapping technique is that when a clone is digested with a restriction enzyme, fragments are created from each of its ends. Because these fragments are not terminated by restriction sites, but rather by the end of the clone, their lengths will not match the lengths of those generated within an overlapping clone. Theses fragments of non-matching lengths might lead to confusion when considering possible orderings of two or more overlapping clones.

An Example

To illustrate the usefulness of Maynard and Olson algorithm consider the following example. Three clones, C1, C2, and C3, are digested with a single restriction enzyme, yielding fragments of the following lengths (as determined by gel electrophoresis):

Clone	Restriction Fragments (Sorted by Length)
C1	2,2,3,3,4,5,6,7
C2	1,2,3,3,4,8,9
C3	1,2,2,3,4,6,8

Solution A:

Using the Maynard Olson algorithm we decide to first align clones C1 and C3 as follows.

C1:	3,5,7	2,2,3,4,6	
C3:		2,2,3,4,6	1,8

We next add C2 to the alignment.

C1:	3,5,7	2,2,3,4,6			
C3:		2,2,3,4,6	1,8		
C2:		2,3,4,?	1,8	3,9	

We find C2 is incompatible with the first alignment, so we re-evaluate, and obtain:

C1:	3,5,7	2,6	2,3,4		
C3:		2,6	2,3,4	1,8	
C2:			2,3,4	1,8	3,9

We conclude that that the order of the markers, by group, is: (3,5,7)(2,6)(2,3,4)(1,8)(3,9)

Solution B:

Observe that we could also obtain a physical map by first aligning clones C1 and C2.

C1:	2,5,6,7	2,2,3,4		
C2:		2,3,3,4	1,8,9	

We next add C3 to the alignment.

C1:	2,5,6,7	2,2,3,4		
C2:		2,3,3,4	1,8,9	
C3:	2,6	2,?,3,4	1,8	

This alignment doesn't work so we re-order. We obtain the same final answer as in Solution A.

C1:	3,5,7	2,6	2,3,4		
C2:			2,3,4	1,8	3,9
C3:		2,6	2,3,4	1,8	

Under ideal data conditions Restriction Mapping and the Maynard and Olson algorithm provide a useful method for determining the order of clones from a DNA segment. However it is important to note that in practice the Maynard Olson algorithm may not always provide a unique solution for clone order.

Hybridization Mapping

In Hybridization Mapping, Sequence Tag Site markers are used to determine the unique fingerprint of each DNA clone.

A Sequence Tag Site (STS) is a sequence, roughly 200 base pairs in length, which

occurs in only several locations throughout the genome. Once an STS sequence has been identified, markers can be made to find that sequence in other locations. This is done by first amplifying the sequence using PCR and then synthesizing a complimentary strand.

In Hybridization Mapping, these complimentary strands are labeled with flourescent markers and then mixed with BAC clones from the DNA segment of interest in a hybridization reaction. The locations where these strands bind represent occurrences of the STS in clone DNA. The clone fingerprint then consists of the list of all STS markers which it binds. To determine the ordering of the clones note that clones which share identical STS presence and ordering probably repre,,sent overlapping segments of DNA. By using this information to determine the order of the STSs on the original DNA, we can infer the order of the clones. The Fulkerson-Gross Algorithm and the Booth-Lucker Algorithm both offer methods for determining the order of STS markers. These algorithms hinge upon the manipulation of a matrix such as the one given below, which represents the results of a hybridization experiment using three clones (C1, C2, and C3) and four STS markers (m1, m2, m3, m4). In the matrix, ones indicate the presence of a given marker in a clone, and zeros indicate its absence.

		Clones		
Markers		C1	C2	C3
	M1	1	0	1
	M2	0	1	0
	M3	1	1	1
	M4	0	0	1

A solution for the ordering of STS markers is obtained by permuting the rows of the matrix until all the ones in a given column occur consecutively. This is called the "consecutive ones property". For the matrix given above we can find the solution:

		Clones		
Markers		C1	C2	C3
	M2	0	1	0
	M3	1	1	1
	M1	1	0	1
	M4	0	0	1

which corresponds to STS marker ordering m2, m3, m4, m1 and clone ordering C2, C1, C3. Note that the reverse ordering (here C3, C2, C1) will also always be a solution.

The Fulkerson-Gross Algorithm is fairly involved and allows for the determination of

a solution matrix in order m+n time (where m and n are matrix dimensions). The speed of the Booth-Lucker Algorithm depends upon the number of ones in the matrix (r) and requires n+m+r time to find a solution.

Regardless of the algorithm used, the results of some hybridization experiments will not provide enough information to uniquely determination of clone order. When conceiving a hybridization experiment it is important to select STSs which occur relatively infrequently in the genome, and to use an adequate number of unique markers.

Complications

When undertaking a hybridization experiment there are a number of ways in which error might be introduced into the analysis.

During the hybridization reaction it is possible that two unrelated clones could fuse together, creating a "chimera" clone. This adds extra ones to the matrix, making it appear that certain markers are erroneously adjacent, or it may cause the matrix to be unsolvable.

The matrix might also become unsolvable or misleading as a result of false zeroes. When constructing the matrix, a zero might be recorded for a given clone and marker even if the marker is truly present in the clone, if the marker was not detected by the reaction. This might happen if some base pairs were deleted from the clone DNA so the STS marker could no loner hybridize. It might also arise from errors in the PCR reaction when the markers were generated, or errors in the conditions of the hybridization reaction which did allow the markers to properly adhere

Similarly false ones might occur in the matrix if a clone is erroneously determined to contain a given marker. This is especially likely to happen if the STSes used are not chosen carefully enough. If an non-unique STS is used then it might appear on multiple non-overlapping clones. Alternatively, an insertion into the clone DNA during the reaction might cause it to have a site which closely resemble the STS and binds with the marker.

Finally, errors in the mechanization of the entire hybridization procedure could lead to errors in the matrix. The physical results of the hybridization process are analog but in order to be analyzed they must be stored digitally. Thus it is inevitable that results will sometimes misinterpreted or misrepresented.

To accommodate for these potential errors am algorithm for "consecutive ones with

errors" has been developed. Recognizing that when the matrix has errors it will not always be solvable for consecutive ones, this algorithm aims to minimize the number of gaps when the order of markers is permuted. This problem has been show to be NP-complete (Karp).