The materials are from Mark Blaxter's lecture notes on Sequencing strategies and Primary Analysis

Genome Sequencing Overview

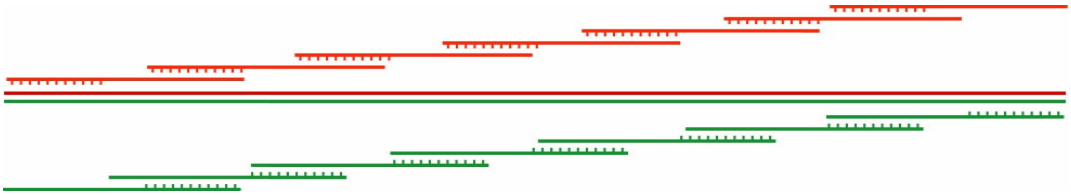
The goal: the correct sequence of a genetic element of interest

What does "correct" mean? We want the sequence to be reliable (no errors) and contiguous (no breaks).

The level of reliability is determined by the reliability of the methods used to derive the sequence, the number of times a particular base has been sequenced, and the possibility of polymorphism (real biological polymorphism) in the sequence. Contiguity is achieved by ensuring that all areas of the gene to be sequenced are accessible to our methods.

For smaller pieces of DNA (individual clones, small viruses, plasmids) it is possible to sequence them to completion. The small pieces of DNA that can be sequenced to completion are called sequence reads. It is usual to sequence a piece of DNA on both strands, and to generate six or more overlapping reads for each base.

To sequence a clone larger than the average read length, it is possible to use a shotgun approach. The idea is to pepper the DNA with sequence reads such that they overlap, and yield, when assembled, the complete sequence of the clone.



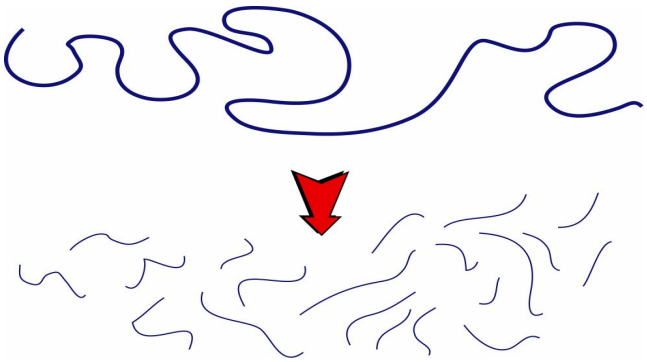
The shotgun part comes from the way the clone is prepared for sequencing: it is randomly sheared into small pieces (usually about 1 kb) and subcloned into a "universal" cloning vector. The library of subfragments is sampled at random, and a number of sequence reads generated (using a universal primer directing sequencing from within the cloning vector). These sequence reads are then assembled into contigs, and the complete sequence of the clone generated.

Physical mapping

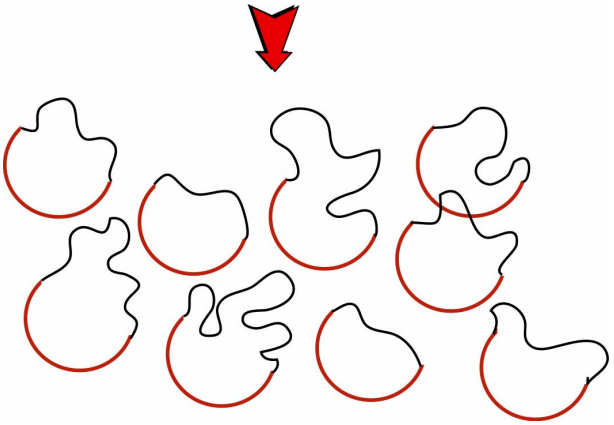
A physical map is a set of cloned DNA fragments whose position relative to each other in the genome is known. The complete DNA sequence of a gene or genome is the ultimate physical map. However, it is useful to construct intermediate level physical maps from cloned fragments: these cloned fragments can subsequently be used for sequencing or other manipulations.

Shotgun sequencing

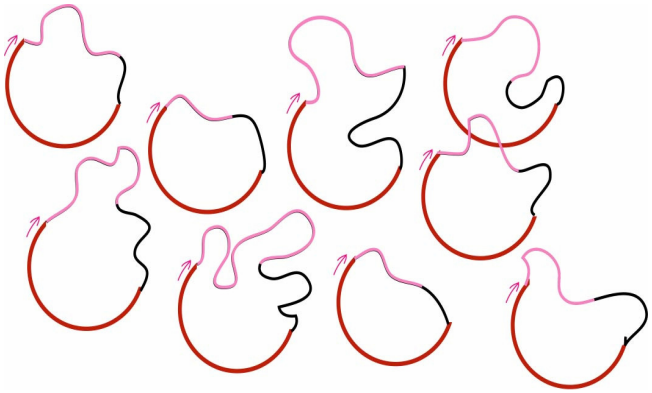
Genomic DNA is sheared or restricted to yield random fragments of the required size.



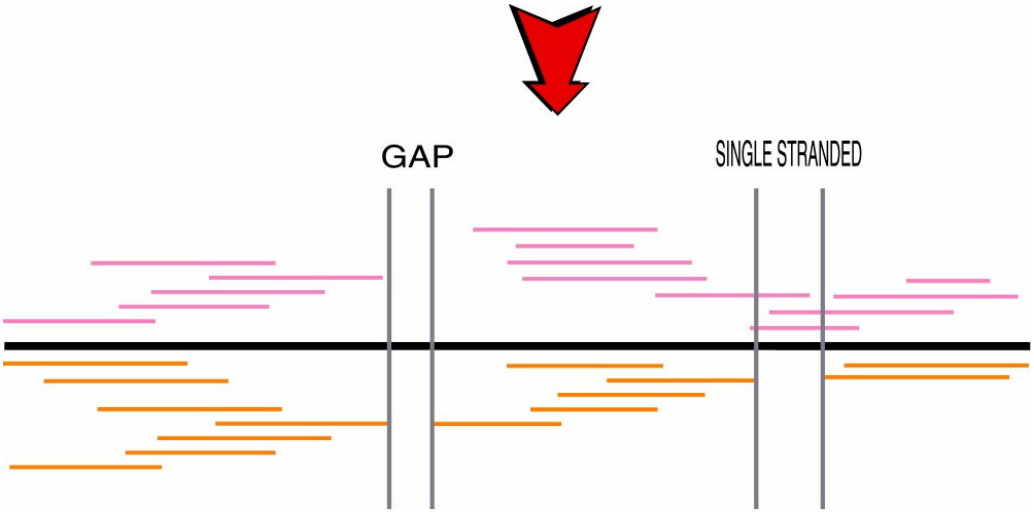
The fragments are cloned in a universal vector



Sequencing reactions are performed with a universal primer on a random selection of the clones in the shotgun library.



These sequencing reads are assembled in to contigs, identifying gaps (where there is no sequence avaialble) and single-stranded regions (where there is sequence for only one strand).



The gaps and single-stranded regions are then targeted for sequencing to produce the full sequenced molecule.

This shotgun methodology is applicable to many DNA sizes, from cosmids to YACs.

Assembly

In the assembly phase, all the sequence reads from the clone are first compared to each other. Identities between the sequences of different reads are noted, and these identities are used to align the sequences into sequence contigs. The sequences of two different reads of the same segment of DNA may not be identical because of the quality of the sequencing reaction analysis (technical issues). Thus for each base in the contig it is usual to require that it is independently confirmed from multiple overlapping reads from both directions. Contig building software has been designed that takes into account the "quality" of each base in a read (where quality is a measure of the confidence the software has that the base has been called correctly). Any gaps, discrepancies or ambiguities in the sequence can be flagged for resequencing, possibly using alternate chemistry.

Chromosomal Sequencing

Most bacterial genomes are in the range 1-5 Mb. This size of genome is sequencable using whole-chromosome shotgun strategies. For many eukaryotes, such as fission and budding yeast, and protozoan parasites, individual chromosomes are in this range (though the genomes may be bigger in total). The main limitation in this whole genome shotgun approach are the generation of fully representative small insert libraries and the computing power needed to generate the fully contiguous sequence from tens to hundreds of thousands of individual sequence reads.

Sequencing Larger Genomes

For larger genomes or chromosomes, and for genomes where the repetitive DNA content is significant, a map-then-sequence approach has been used.

A physical map is first built using large insert clones. The minimum set of clones that cover the whole genome is then selected (called the minimum tiling path). These are chosen such that there is overlap, but minimal overlap between the end of one clone and its neighbor. Each clone is sequenced and assembled in isolation using a shotgun strategy. The individual clones are then contiguated for production of the final sequence.

For sequencing genomes where the chromosomes are very large (such as most eukaryotes) it may be possible to separate the chromosomes by fluorescence activated sorting, and make chromosome specific libraries. Mapping can then be performed on using these less-redundant libraries.

This map-and-sequence strategy has been used for many bacterial genomes, for yeast and for *C. elegans.*

Sequencing Genomes: Problems and Prospects (Optional)

Problems

The sequence of some spans of DNA is difficult to generate. This can be because of a biased base content (this can result in failure to be cloned, poor stability in the chosen host-vector system, or inability of the polymerase to reliably copy the sequence). For example AT-rich DNA clones poorly in bacteria. GC rich DNA is difficult to sequence and often requires the use of inosine substitution for G in the reactions, and careful monitoring sequencing gel conditions.

The presence of poisonous sequence: sequence that interferes with the biology of the host organism. This could be an operator that reduced effective concentrations of an essential DNA binding regulator, or an open reading frame that expresses a toxic protein. These problems are commoner for bacterial genomes cloned in bacterial vector-host systems, but can affect eukaryotic cloning too.

The presence of repetitive DNA can result in deletion of sequences from the clone, particularly in bacterial systems, and thus the sequenced DNA will not correspond to the situation in the original source organism.

Malaria Genomics as an example:

The malaria parasite *Plasmodium falciparum* has a 35 Mb genome with approximately 80% AT. It is organized as a set of small chromosomes from 800 kb to >3 kb. In some regions the sequence is >95% AT for several hundred bases. This DNA has proved very difficult to clone in bacterial vectors, though it appears to be quite stable in yeast. The genome has been mapped using YAC clones, and sequencing is proceeding chromosome by chromosome, using a whole chromosome shotgun strategy. It is happening at three sites, the Sanger Centre,

TIGR and Stanford. To sequence the intergenic (very high AT) regions of this genome it was necessary to use very small insert libraries.

The Sanger Centre project started with Chromosome 3. The results have been published in Nature (see the Sanger website for details) It was purified from pulsed-field gels and subcloned in a variety of small insert libraries. These were shotgun sequenced, and the sequence assembled. The Sanger website has a detailed description of the strategy used. There were significant problems with the AT-richness of the chromosome, particularly in the centromeric region.

Whole genome shotgun sequencing above 3 Mb

Currently, the largest genome assembled from whole genome shotgun is bacterial (~3 Mb). TIGR (an academic genomics company) and Celera Genomics (a commercial genomics company) have proposed that, given appropriate sequencing throughput and computing power, it will be possible to sequence larger (much larger) genomes using the whole genome shotgun strategy.