

Gene and genome sequencing

Dideoxy chain termination sequencing

was invented by **Fred Sanger** in the 70s

If the reaction is carried out using **dideoxynucleotide triphosphates** rather than the **normal deoxy** form, further DNA synthesis is prevented

This is because there is no **hydroxyl group** for the next nucleotide to form a **phosphodiester bond** to

If one performs a DNA synthesis from a primer, the DNA is extended one base at a time

The **single stranded** DNA we want to sequence

A specific primer

To sequence, prepare a mixture containing

A small amount of radioactive tracer (e.g.: radiolabelled nucleotides or primers)

A small amount of **ONE** of the four dideoxynucleotides

Synthesis will then stop

In the population of molecules, a terminated synthesis at every one of that particular base is represented

Occasionally, a dideoxy nucleotide will be incorporated during DNA synthesis

Since they are all initiated from the same primer, these all differ by a single base

We do this four times, with all four bases

The four reactions are then run side by side on a gel, and the sequence is read off

Four different colours, one for each of the reactions

After DNA synthesis, they can then all be pooled

And then run of a single lane and detected by a laser

More reliable and higher throughput

Fluorescently labelled primers are used

Fluorescently labelled dideoxy-NTPs have been developed

All four reactions can now be performed in a single tube

Polymer-filled capillaries that allow massive throughput are now used

Today, the system is completely automated, with a computer automatically interpreting the output to give the sequence and its quality

Sequencing is *not* an error-free process

Modernisations

DNA sequencing

Comparative genomics

The comparison of genome sequences, or the set of proteins they encode, between two or more organisms

This can be very informative - for example, comparing *myobacterium tuberculosis* to *myobacterium bovis* can identify the sets of genes that are potentially involved in the pathogenicity of the TB causing bug

Where the sequences of many closely related organisms are available, the analysis of **conserved sequences** can be a powerful way of finding regulatory sequences in the genome

In mammals, comparative maps of genome organisation at the level of each chromosome reveals how the genome has been rearranged during evolution

Identify such **syntenic blocks** (segments of chromosome carrying the same genes) can help in genome annotation and inferring gene function

Comparative genomics

Annotating genomes

This involves the identification of genes (coding regions) in genome sequences

We just look for **open reading frames**

These are sequences separated by start and stop codons

Find the probability of having an ORF for all 6 reading frames (sense, antisense, 3 each)

The graph will have lots of short "blips" (too small to be a gene)

It'll also have large humps - this is where the genes are

We do this as follows

Domains

Start + stop codons

Length

Probability worked out using

Then compare with database!

About 40% of genes in E-Coli have unknown function

Perhaps it's because they're essential - this means that if they go, the bacteria dies and we therefore can't investigate them easily

Genes structures are complex

Our understanding of the sequences that regulate, for example, differential splicing is very poor

Introns may be very large, which makes joining up genes hard

Genes can be nested in other genes (for example hiding in their introns)

Finding genes in **metazoan** (animal) genomes is much harder

Use the sequence of mRNAs (obtained from cDNA clones)

Use comparative analysis with less complex genomes

Some of the most successful ways are as follows

Annotating genomes

Genome sequencing

The throughput of sequencing means that many genomes can now be sequenced

In 1976, the first genome (of the small RNA Phage MS2) was sequenced

It took 20 years of long, slow throughput sequencing until the genome of the first free-living organism (the bacteria *H. influenzae*) was sequenced

In 2001/2003, the human genome project was available

Most sequencing efforts are prokaryotic, but there are an increasing number of eukaryotic, particularly vertebrate, efforts

There are basically two strategies to sequencing a genome

The clone-based approach

A set of large insert clones (e.g. BACs) are first ordered with respect to the genome

Lyse each piece of DNA from the clones, and get a series of restriction fragments

Find another fragment which hybridises with the *end* of the bit we're talking about, but not the bit further in

TTT is the next one in the order

This is a **gene walk**

This forms a physical map of overlapping clones that span each genome

Individual clones are then broken down and sequenced

This is advantageous in that everything can be stitched together rather easily

Repeats aren't so much of a problem, because the fragments are much larger

Making the large-scale BAC map, however, is very slow

The shotgun approach

The entire genome is randomly fragmented

Must be random (e.g.: via ultrasound)

Restriction enzymes would cut at a particular point

First put into vector

PCR used, using primers on either side of the cut in the vector

Lots of small bits of DNA are sequenced

Can be problematic if there's not a 100% overlap in between the two primers - not all the bit gets sequenced

Sophisticated computer algorithms are then used to compare all of these sequences at attempt to assemble them into a contiguous block of sequence

Faster, but *very* difficult computationally - has only been possible recently

Another problem with this approach involves sequencing complex eukaryotic genomes

These have a lot of repetitive DNA (there are over 1 million copies of the Alu repeat in every human genome)

These may flank unique DNA, present at a single site in the genome

Thus, aligning short sequence clones that may contain repeats can be very difficult

Genome coverage

Sequencing reactions often yield poor results near the start and end of sequence runs, due to technical issues

This corresponds to the bits of DNA beginning and ending clones, that are the most useful in assembly

For this reason, when sequencing, much more DNA sequence than is needed for a single coverage is generated (lots of overlapping bits)

High quality genome sequence is generally 8-10 X coverage

Prokaryotes

With relatively small (under 10 mb) and fairly repeat free genomes, many bacteria yield well to the shotgun approach

With related bacteria, it is frequently easy to assemble a genome sequence with low coverage (~ 3 X) if a high quality (10 X) sequence of a relative is available

This is also true, to a certain extent, of eukaryotic genomes

These are more problematic and usually approached using the clone-based method

Eukaryotes

For example, the genome of the nematode *C. Elegans* was the first to be sequenced

It took as tremendous effort

Built a map of overlapping clones, and then sequenced them

Mostly at the Sanger centre in Cambridge

Next was *Drosophila melanogaster* - also began using a slow clone based approach

Went rather faster when Celera Genomics announced they were going to do it faster using a shotgun approach

However, they were much helped by an extensive clone map already constructed

The sequence of 11 other drosophila species is almost complete, helping us greatly in our study of evolution

Human genome

Started off in the same way as *C. Elegans*, with a BAC library

Celera accelerated things!

Our current sequence is 99.999% accurate

To date, about 300 gaps remain

These are slowly being filled, but it's a tedious process, since DNA in these gaps is difficult to clone

Long-range PCR and comparative analysis with chimpanzee sequence are methods being employed today

No sequence is really available from heterochromatin

The human genome sequence contains many less genes than was expected

Genome sequencing

Genome sequencing - examples