

4/29- Lecture 8: DNA Sequencing

Quick aside on upcoming topics: sequencing, gene recognition- applying HMMs, large scale genomic alignment (human, mouse), multiple alignment, protein alignment across species- many short sequences, DNA microarrays, finding regulatory motifs, (time permitting) phylogeny & rearrangements, RNA structural folding. Exciting stuff! 😊

Organisms are characterized by their genomes- specific long sequence of 4 nucleotides- A, C, G, T. This sequence captures all information, in genes, needed for the animal to reproduce, develop, and perform all biological functions. In each cell, there is a full copy of the genome. Different genes are “expressed,” i.e. “turned on” in different types of cells.

Genes are transcribed and then translated into proteins.

Challenge: find the exact sequence of nucleotides in a given organism, e.g. human.

Challenging both for technology and computational methods.

Frequent question: which human was sequenced?

Two answers: 1. Craig Venter- former CEO of Celera. “Anyone who doesn’t want his genome sequenced shouldn’t be in this business.” (paraphrased)

2. It doesn’t matter- we’re all very similar.

Polymorphism rate: the # of letters that differ between 2 organisms of the same species. In human, rate is very low- $\sim 1/1,000$ - $1/10,000$

SNP- single nucleotide polymorphism. (pronounced “snip”) There are also areas with longer differences.

Pathological cases: extra copies of an entire chromosome, or fusion of two chromosomes- cause disease states.

We’ll focus on SNP’s, which is only every few thousand in humans- relatively low rate.

Small sea creature organism can vary by as much as 10%. →

SNP consortium- project to identify all human SNPs.

Why are humans so similar? Generally, if you look at a small population, genetic variation is reduced with each successive generation.

Mate AA and BB → AB, AB.

Mate AB and AB → 50% chance AB, 25% AA, 25% BB

With enough generations, you’re likely to lose either A or B.

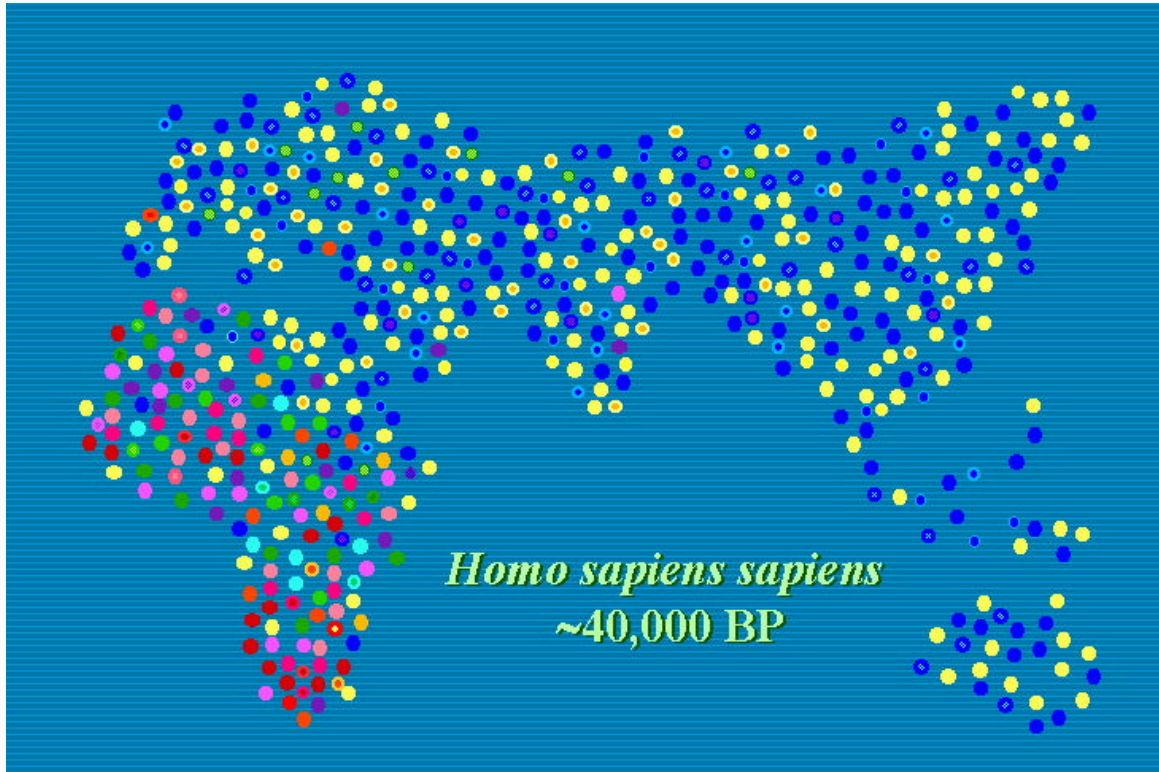
Humans: small population in Africa interbred for a while. Went on to populate the rest of the earth, having already lost much variation.

~130k years ago, humans left Africa.

20k years ago, humans entered North America.



Chart showing genetic variation in Africa ~150k years ago. Subset left, went to Mesopotamia- smaller genetic variation in this group, and they were the ones who populated the rest of the earth. Interestingly, this means there is MORE variation within Africa than you see in the rest of the world.



Different colors in graphic above indicate genetic variation.

How much of this variation has to do with skin color? Not much! Interbreeding for ~1000 years can change skin color in a population.

Biologists who do studies on human variation are considering doing their studies only on Africans- more for their money, as it were.

How do we sequence DNA? Can't just stick it in a machine, get 1M reads out. Can only sequence ~500 at a time.

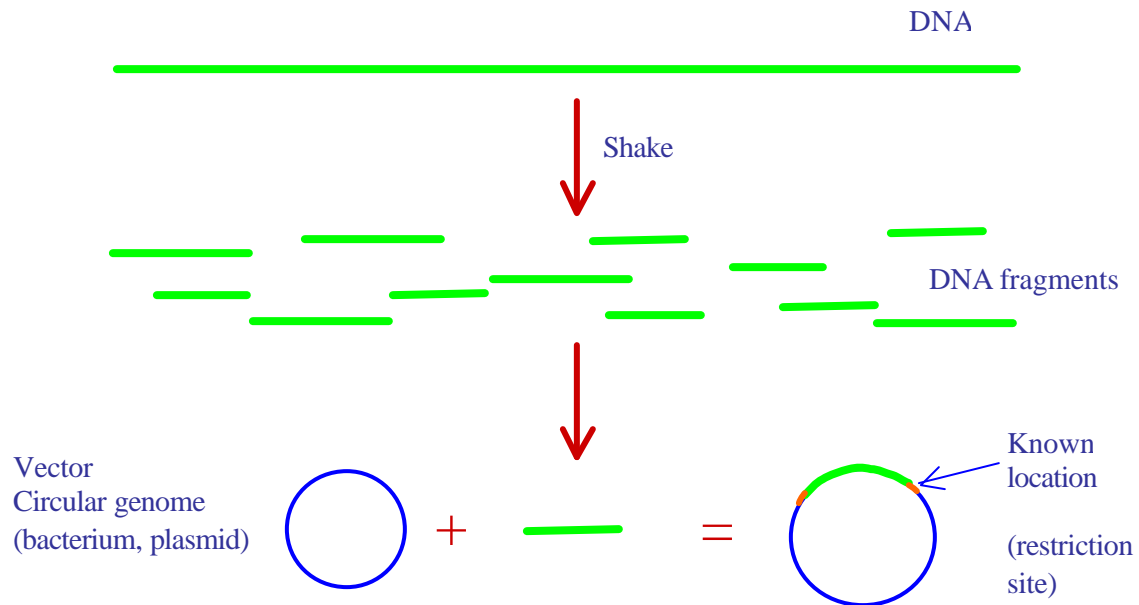
Start by breaking DNA into pieces by shaking it. (though this doesn't mean that when you jump up and down, your DNA breaks apart.) Do this with many copies of the genome, and you get overlapping pieces.

Then incorporate those fragments into biological hosts- generally circular genomes where we can insert DNA into a known location.

BAC- Bacterial artificial chromosomes.

Plasmids, YACs.

Each type incorporates a different sized fragment when mixed.



This helps so that we know the approximate size of a fragment. Not precise- sometimes chimerics where 2 pieces join together, etc. but we know approximate length.

Plasmids- 2,000-10,000

Cosmid- 40,000.

BAC- 70-300k

YAC- >300k

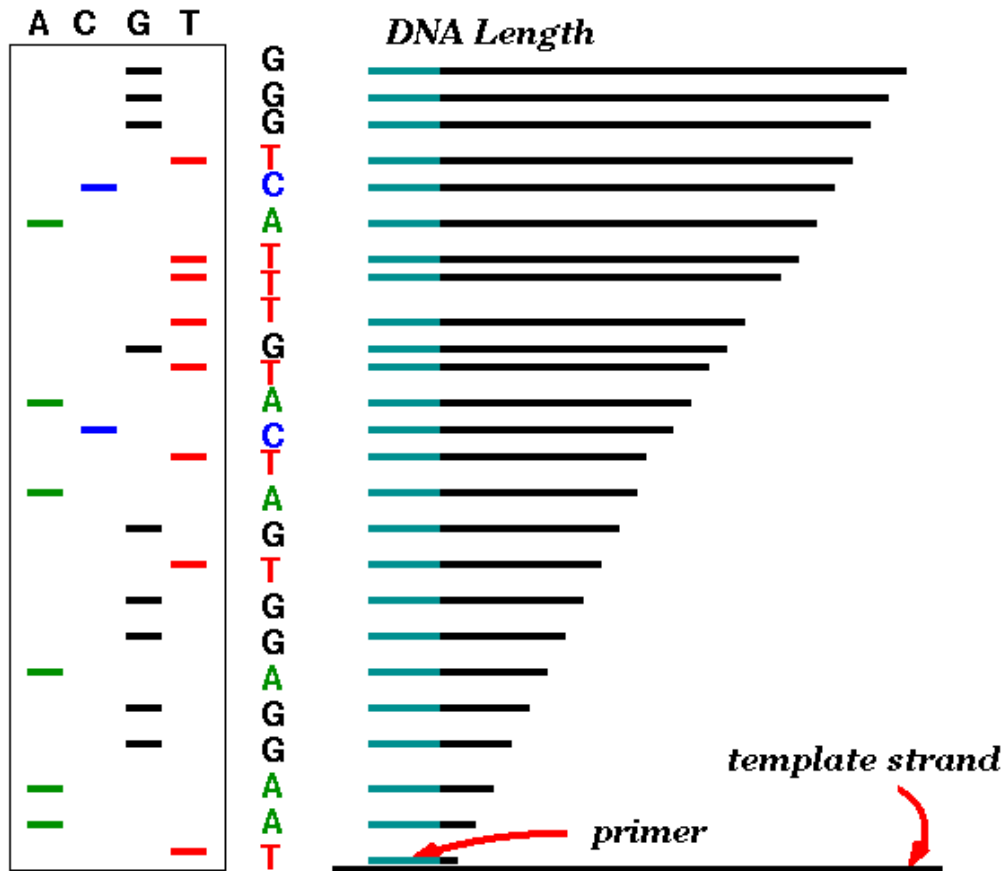
Synthesize DNA from a restriction site. A primer sticks to complementary site, starts transcription. This is done in a "DNA soup" which contains many, many individual nucleotides. It also incorporate one type of di-deoxynucleotide (A's, G's, C's, or T's) which, when incorporated, causes transcription to end at that point.

Di-deoxynucleotides are marked such that they can be seen in a gel. At the end, you have fragments of all sizes, and you know what nucleotide they end in.

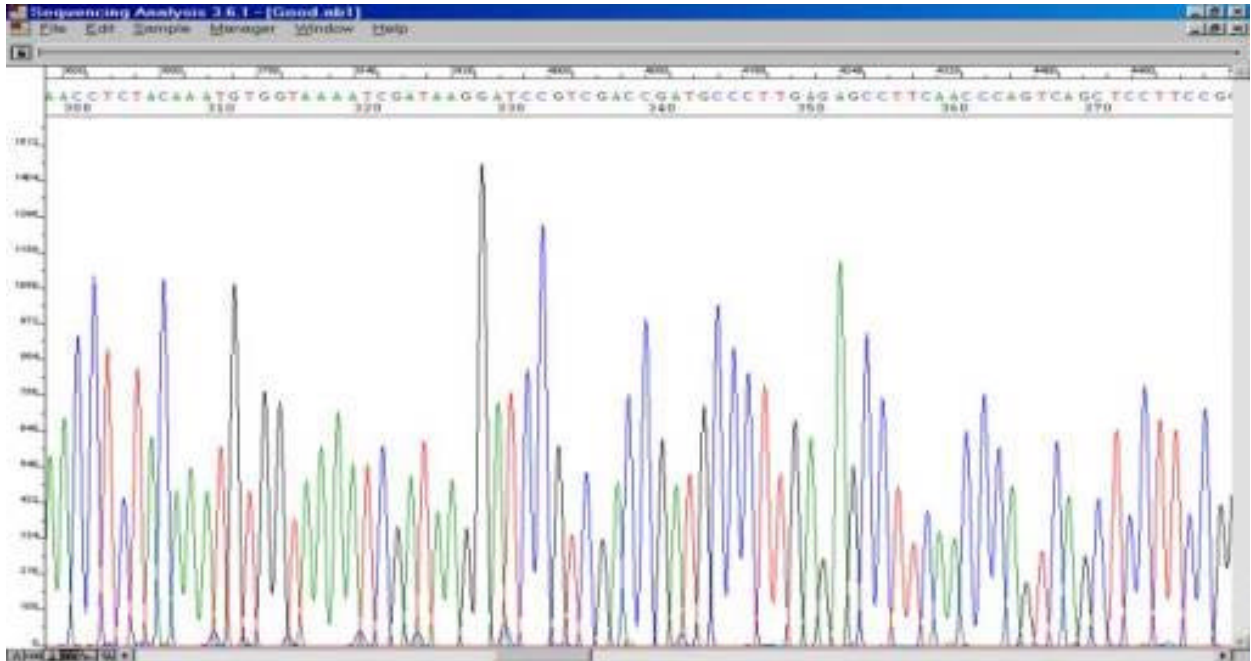
Run these fragments from one side of a gel to the other. In each column, put a mixture of fragments that end in a different nucleotide, introduce current. Bigger fragments travel slower- don't get as far.

At the end, we can see bands where the different sizes ended up.

This is why we can only sequence so many at a time- it's easy to tell the difference between molecules 1 base long and 2 base long. This gets impossible to measure for DNA that's 1000 vs. 1001 base pairs long.



Slide below shows (extremely clean) output for DNA sequencing. This data has been filtered, smoothed, corrected for concentration. As you'd expect, you'll see fewer long molecules- this is corrected for. Y axis represents strength of signal read by the machine.



Question: how long does this take? Dr. B isn't sure, but gives the example: a lab can sequence $\sim 7x$ coverage of a mammalian genome ($\sim 30B$ reads) in ~ 1 year.

Very interesting signal processing problem- do the best you can reading the signal.

Electropherograms- output of reading.

PHRED- method for calling the letters ("A", "G", etc.)- used by almost all labs. (By Phil Green at UW.) There are many better methods out there now, but inertia makes labs reluctant to switch, despite the potential gain in accuracy.

Output of PHRED is a read- $\sim 500-700$ long.

Also gives quality scores: $-10 \cdot \log_{10} \text{Prob}(\text{error})$

So, score of 30 means $\sim 1/1000$ reads are wrong.

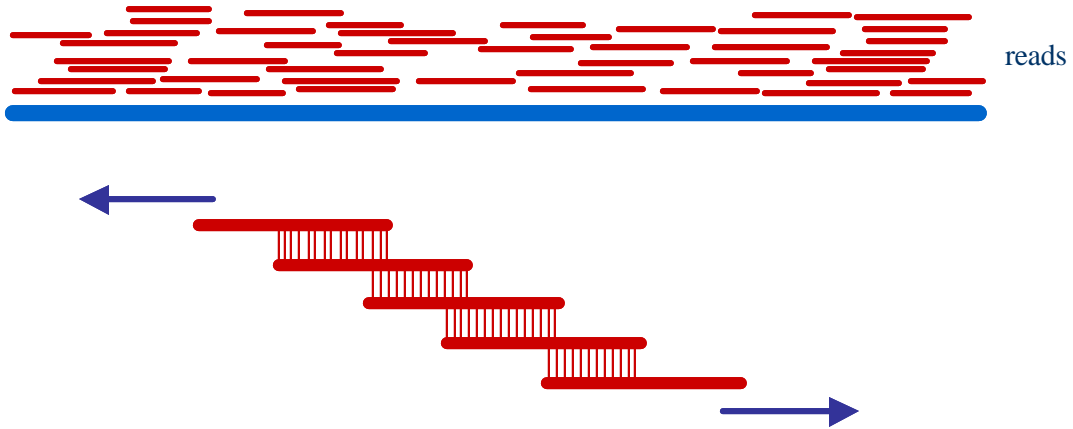
Sequencing from both ends is referred to as double barreled sequencing.

How to sequence segments longer than 500? Shotgun sequencing.

Cut into many pieces, $\sim 7x$ coverage. Then sequence one or both ends of each fragment.

Do this many times so that you have overlaps between reads.

Each time you find an overlap like this, you may be able to stitch these reads together.



Sometimes get surprises as to the length- e.g. a specific archaea (vs. eukaryotes and bacteria)- expected ~2M, got ~4.5-5M.

Coverage: need enough redundancy. Can calculate statistically what's needed- Lander-Waterman method. Coverage = nl/L , i.e. number of reads times average length of reads, divided by the length of the genome.

Redundancy of 10, read 500 long, expect 1 gap per million letters- pretty good- considered gold standard.

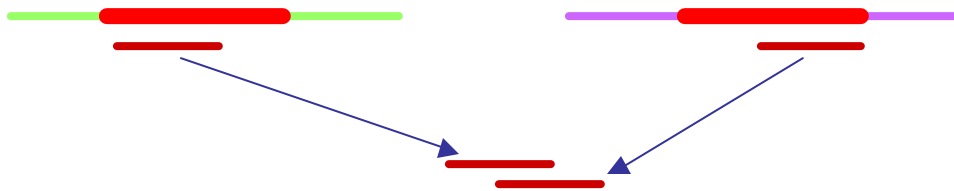
Frequency of gaps depends on coverage and length of segments.

So, method sounds simple enough. But it's not! Challenges:

Technical:

- errors by sequencing machines- PHRED, etc. 1-2% wrong- usually substitutions rather than insertions or deletions
- merging of 2 reads into 1- chimerism- when 2 reads fuse
- contamination- get a little human or bacterium mixed in with mouse
- others

Repeats- big problem! If you take repeated regions and mistakenly merge them, can end up with sequences that don't actually occur in the genome, as illustrated below.



Also, comparing all possible reads requires N^2 comparisons, where N can be ~ millions.

So, what's the extent of repeats? Some organisms- relatively low. Many genomes, e.g. bacteria, drosphila- ~5%.

Human is ~50% repeats! (and of the non-repeated, most is useless, at least that's the current understanding.)

Kinds of repeats:

- Low density- ATATAT, etc.
- Microsatellites- $(a_1 \dots a_k)^N$ where $k \sim 3-6$ (e.g. CAGCAGTAGCAGCACCAG)
- Common repeat families- usually transposons-
 - o SINE- (Short Interspersed Nuclear Elements) e.g. Alu- ~300 long repeat sequence that repeats ~1M times in the genome!
 - o LINE- (Long Interspersed Nuclear Elements)
 - o MIR
 - o LTR/Retroviral

How do these repeats occur in the genome? Have a short sequence that by chance is able to transcribe itself and "fool" the genetic machinery into "gluing" it back into another place in the genome. Once that happens, there's no reason why it won't keep propagating itself. The most fit of those will keep going...

There may develop an "arms race" between the organism and these repeated segments.

If you look at a genome and see a bunch of similar repeats, but not identical, you know this sequence has "died", i.e. is no longer replicating itself. If it was still replicating itself, there would have been recent activity and you'd see some exact repeats.

Paralogs –genes that duplicate and then diverge, both forms of which are functional. There may be selective pressure to do so.

So, because of all these challenges, various hierarchical strategies have been used to sequence the genome (sometimes in combination).

Hierarchical- clone by clone.

Break the genome into pieces with high redundancy, and incorporate these pieces into, a "vector," say, BAC clones. Each is ~200,000. Redundancy ~20x.

Getting these BAC clones is cheap. The expensive part is doing the shotgun sequencing part. So, you want to minimize how much of that you have to do. (doing the human genome once requires ~\$300-500 million.)

So, before sequencing, we want to map them onto the genome. That is, order them relative to each other and find out which ones overlap- "Minimal tiling path."

Then select BACs with minimum overlaps. (but no gaps!) "small overlap" means ~10k. Sequence that minimum tiling path, then put them all together.

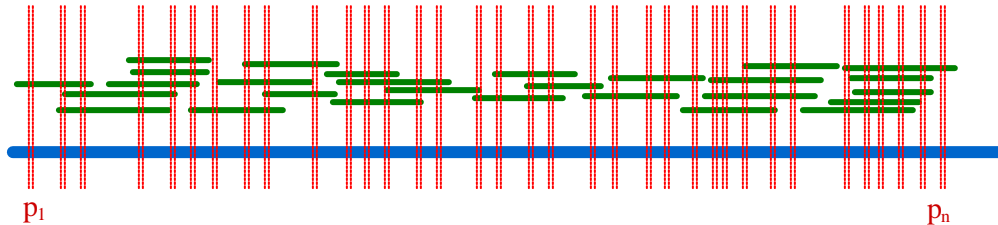
How do you map them, given that you don't know the sequence of the pieces, or of the genome? A few years ago we'd spend a few lectures on this area, but most of this has been figured out- less important now.

Several methods for mapping:

Hybridization

In image below, red lines are probes, essentially a short word that is likely to be complimentary somewhere in the genome. At that point it will hybridize to that part of the genome.

Treat each BAC with all probes. Then we can see which probes stick to the same BAC. If two different probes stick to two different BACs, there's a good chance those BACs overlap.



Given m probes, n clones. Create matrix of 0/1 of hybridizes/doesn't.

If we assume that each probe hybridizes to only one place in the genome, and we're able to order of the probes in the order they appear in the genome, then ordering the BACs becomes a trivial problem.

Then each clones is a row with a bunch of 0's, then a bunch of 1's, then 0's.

"Consecutive ones" property of a matrix.- holds if at every row, all the 1's are in consecutive order.

If the probes always hybridize correctly, and always only hybridize in one place, then this problem can be solved in cubic time.

However, there are many technical difficulties: Chimeric clones, probe false positives, probe false negatives, short probes will statistically hybridize to >1 place.

So, we're forced to turn to heuristics- if two clones have many common probes, then they probably overlap.

One way to formulate this problem computationally is maximum parsimony. Assume that when results can be explained by overlapping clones, call them overlapping. i.e. result implies largest possible overlap of clones.

Or, find shortest string of probes that "explains" all the clones. i.e. all probes appear uninterrupted.

NP complete problem. APX hard.

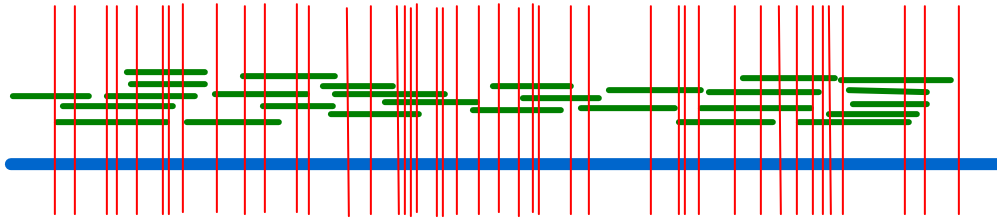
Solutions to this problem are greedy, probabilistic, etc. with much manual curation.

Digestion

Certain restriction enzymes exist that can be used to cut the clones in various places.

Use these enzymes to treat all the clones, end up with short intervals of given lengths.

Then each clone is a set of interval lengths. Run these over gels to determine length. If two clones have overlapping set of intervals, likely the clones overlap.



Double digestion: first digest all clones with enzyme A, then with B, then with both.

Each approach above has its own noise factors.

Next time: Walking method.