New topic: DNA sequencing

How we obtain the sequence of nucleotides of a species

<table>
<thead>
<tr>
<th>AGCTGACTGAGGACCGTG</th>
<th>CAGCTACAGCTAGCAGCTAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGACTGAGACTGACTGGGT</td>
<td>TATATATATACGTCGTCGT</td>
</tr>
<tr>
<td>CTAGCTAGACTACGTTTTA</td>
<td>ACTGATGACTAGATTACAG</td>
</tr>
<tr>
<td>ACTGATTTAGATACCTGAC</td>
<td>TGATTTTAAAAATTATT...</td>
</tr>
</tbody>
</table>

Which representative of the species?

Which human?

Answer one:

Answer two: it doesn’t matter

Polymorphism rate: number of letter changes between two different members of a species

Humans: ~1/1,000 – 1/10,000

Other organisms have much higher polymorphism rates

Why humans are so similar

A small population that interbred reduced the genetic variation

Out of Africa ~ 100,000 years ago

Migration of human variation

Early *Homo sapiens sapiens* in Africa

150,000 to 100,000 BP

*Homo sapiens sapiens* colonizing south west Asia

~100,000 BP
DNA Sequencing

Goal:
Find the complete sequence of A, C, G, T's in DNA

Challenge:
There is no machine that takes long DNA as an input, and gives
the complete sequence as output
Can only sequence ~500 letters at a time

DNA sequencing – vectors

DNA sequencing – gel electrophoresis

Different types of vectors

<table>
<thead>
<tr>
<th>VECTOR</th>
<th>Size of insert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasmid</td>
<td>2,000-10,000</td>
</tr>
<tr>
<td>Cosmid</td>
<td>40,000</td>
</tr>
<tr>
<td>BAC (Bacterial Artificial Chromosome)</td>
<td>70,000-300,000</td>
</tr>
<tr>
<td>YAC (Yeast Artificial Chromosome)</td>
<td>&gt;300,000 Not used much recently</td>
</tr>
</tbody>
</table>

Electrophoresis diagrams
Reading an electropherogram

1. Filtering
2. Smoothening
3. Correction for length compressions
4. A method for calling the letters – PHRED

PHRED = PHill's read editor (by Phil Green)
Based on dynamic programming
Several better methods exist, but labs are reluctant to change

Output of PHRAP: a read

500-700 nucleotides
A C G A A T C A G .... A
16 18 21 23 25 15 28 30 32 21
Quality scores: -10log₁₀ Prob(Error)
 Reads can be obtained from leftmost, rightmost ends of the insert

Double-barreled sequencing:
Both leftmost & rightmost ends are sequenced

Method to sequence segments longer than 500

Genomic segment
cut many times at random (Shotgun)

Get one or two reads from each segment

~500 bp
~500 bp
Reconstructing the Sequence (Fragment Assembly)

Cover region with ~7-fold redundancy (7X)

Overlap reads and extend to reconstruct the original genomic region

Definition of Coverage

Length of genomic segment: \( L \)
Number of reads: \( n \)
Length of each read: \( l \)
Coverage \( C = \frac{nL}{L} \)

How much coverage is enough?
(Lander-Waterman model): Assuming uniform distribution of reads, \( C = 10 \) results in 1 gapped region in 1,000,000 nucleotides

Challenges with Fragment Assembly

- Sequencing errors
  - ~1-2% of bases are wrong
- Repeats

- Computation: \( \sim O(N^2) \) where \( N \) = # reads

Repeats

- Bacterial genomes: 5%
- Mammals: 90%
- Repeat types:
  - Low-Complexity DNA (e.g. ATATATACATA...)
  - Microsatellite repeats: \( (a_1...a_k)^N \) where \( k \sim 3-6 \)
    - (e.g. CAGCAGTAGCAGCACCAG)
  - Common Repeat Families:
    - SINE (Short Interspersed Nuclear Elements)
    - LINE (Long Interspersed Nuclear Elements)
    - MIR
    - LTR/Retroviral
    - Other:
      - Genes that are duplicated & then diverge (paralogs)
      - Recent duplications, ~100,000-long, very similar copies

Strategies for sequencing a whole genome

1. Hierarchical – Clone-by-clone
   i. Break genome into many long pieces
   ii. Map each long piece onto the genome
   Example: Yeast, Worm, Human, Rat

2. Online version of [1] – Walking
   i. Break genome into many long pieces
   ii. Start sequencing each piece with shotgun
   iii. Construct map as you go
   Example: Rice genome

3. Whole genome shotgun
   One large shotgun pass on the whole genome
   Example: Drosophila, Human (Celera), Neurospora, Mouse, Rat, Fugu
Hierarchical Sequencing Strategy

1. Obtain a large collection of BAC clones
2. Map them onto the genome (Physical Mapping)
3. Select a minimum tiling path
4. Sequence each clone in the path with shotgun
5. Assemble
6. Put everything together

Methods of physical mapping

Goal:
Make a map of the locations of each clone relative to one another
Use the map to select a minimal set of clones to sequence

Methods:
- Hybridization
- Digestion

1. Hybridization

Short words, the probes, attach to complementary words

1. Construct many probes
2. Treat each BAC with all probes
3. Record which ones attach to it
4. Same words attaching to BACS X, Y ⇒ overlap

Hybridization – Computational Challenge

Matrix:
$m$ probes × $n$ clones

$p_{ij}$: 1, if $p_i$ hybridizes to $C_j$
0, otherwise

Definition: Consecutive ones matrix
A matrix is are consecutive

Computational problem:
Reorder the probes so that matrix is in consecutive-ones form
Can be solved in $O(m^3)$ time ($m >> n$)

If we put the matrix in consecutive-ones form,
then we can deduce the order of the clones & which pairs of clones overlap

Hybridization – Computational Challenge

Additional challenge:
A probe (short word) can hybridize in many places in the genome

Computational Problem:
Find the order of probes that implies the minimal probe repetition
Equivalent: find the shortest string of probes such that each clone appears as a substring

APX-hard

Solutions:
Greedy, Probabilistic, etc.
Lots of manual curatation
2. Digestion

Restriction enzymes cut DNA where specific words appear
1. Cut each clone separately with an enzyme
2. Run fragments on a gel and measure length
3. Clones $C_a, C_b$ have fragments of length $l_i, l_j, l_k$ ⇒ overlap

Double digestion:
Cut with enzyme A, enzyme B, then enzymes A + B

The Walking Method

1. Build a very redundant library of BACs with sequenced clone-ends (cheap to build)
2. Sequence some "seed" clones
3. "Walk" from seeds using clone-ends to pick library clones that extend left & right

Walking: An Example

Advantages & Disadvantages of Hierarchical Sequencing

Hierarchical Sequencing
- ADV. Easy assembly
- DIS. Build library & physical map; redundant sequencing

Whole Genome Shotgun (WGS)
- ADV. No mapping, no redundant sequencing
- DIS. Difficult to assemble and resolve repeats

The Walking method – motivation
Sequence the genome clone-by-clone without a physical map
The only costs involved are:
- Library of end-sequenced clones (CHEAP)
- Sequencing

Walking off a Single Seed

- Low redundant sequencing
- Too many sequential steps
Walking off Several Seeds in Parallel

- Few sequential steps
- Additional redundant sequencing

In general, can sequence a genome in ~5 walking steps, with <20% redundant sequencing

Next Lecture

- Whole-genome shotgun sequencing
  - Currently, the most popular method for sequencing a genome
- Computational assembly of a genome
  - Putting a large puzzle together