

DNA sequencing

DNA sequencing is the process of determining the exact bases A, T, C and G in a strand of DNA. It is important in many aspects including understanding the blueprint of life and personalized medicines.

General Method

1. Dissolve the DNA into fragments through various means
2. Incorporate the DNA fragments into vectors by DNA recombination and the use of restriction enzymes
3. Clone the DNA fragment through replication
4. Isolate the DNA fragments from the vectors
5. Sequence the DNA fragments
6. Assemble the fragments

Determine the base pairs: Chain-Termination Method

1. Excise out the DNA fragment from the vector.
2. Extend the DNA fragment by polymerase.
3. Add in dideoxynucleotide (ddNTPs) which prevents polymerase from extending the DNA replication.
4. ddNTPs are added randomly thus one would obtain all DNA of all different lengths
5. Separate the products with length by gel electrophoresis.

Sequence Assembly

- Using overlap and extend reads to reconstruct the original genomic region

- Coverage: $C = N L / G$
 - Length of genomic segment G
 - Number of reads: N
 - Length of each read: L
- How much coverage is enough?
 - The Lander-Waterman model: $\text{Prob}(\text{not covered bp}) = e^{-C}$

Challenges: Repeats

- Types of repeats:
 - Low complexity DNA: ATATATATATCATA
 - Microsatellite repeats $(a_1 \dots a_k)^N$, $k \sim 3-6$: CAGCACCTAGCAGCACCAG
 - Common Repeat Families
 - ◆ SINE
 - ◆ LINE
 - ◆ MIR
 - ◆ LTR/Retroviral
 - Other
 - ◆ Paralogs
 - ◆ Recent duplications
- Solution
 - Cluster the reads
 - Link the reads

Hierarchical Sequencing

- Obtain a large collection of BAC clones
- Physical mapping – map them onto the genome
- Select a minimum tiling path
- Sequence each clone in the path with shotgun
- Assemble

Physical mapping

- Restriction Mapping

Given a subset $E \subseteq \Delta X$, construct X from E

$$\Delta X = \{|x_i - x_j| : x_i, x_j \in X\}$$

- Partial Digestion: one enzyme ($E = \Delta X$)

- ◆ Cut each clone separately by restriction enzyme
- ◆ Determine length by running on a gel
- ◆ Clone C_a, C_b have fragments of lengths $\{l_i, l_j, l_k\} \rightarrow$ overlap
- ◆ Skiena et al's backtracking algorithm
 - Find the longest distance in ΔX , and delete that distance from ΔX
 - Repeatedly position the longest remaining distance of ΔX
 - Determine between 2 possible positions for the point
 - For each of the 2 positions, check whether all distances from other position are in ΔX
 - If they are delete all those distances from ΔX
 - Backtrack if they are not for both of the two positions

- Double Digestion: two enzymes

- ◆ Cut with enzyme A, enzyme B, then both enzymes A+B

- Hybridization

- Construct many probes (short words)
- Treat each BAC with all probes
- Record which ones attach to it
- Same words attaching to BACs $X, Y \rightarrow$ overlap