

2nd and 4th February, 2009
IRTISHA SINGH

DNA Sequencing:

It is the process to determine the order of bases (A, T, C and G) in the DNA molecule. The order of the bases helps to determine the genotypic variation which is utilized for personalized medicine etc.

It has been a challenging task to determine the sequence because the machines available are not efficient enough to sequence DNA pieces of more than 900 bases.

DNA sequencing methods:

Cloning:

The DNA to be sequenced is cut in fragments by enzymes and then inserted into the vectors. These vectors are then inserted into bacteria where they replicate. After the replication, DNA fragments from the vectors are isolated and then they are sequenced.

1. Chain Termination method (Maxam Gilbert sequencing, used for 300-500 bases, 1975).

Requirements: A single-stranded DNA template (to be sequenced), a DNA primer, a DNA polymerase, radioactively or fluorescently labeled nucleotides, and modified nucleotides (ddNTPs) that terminate DNA strand elongation.

Procedure: The DNA template is divided into four separate sequencing reactions, containing four deoxynucleotides (which are dATP, dGTP, dCTP and dTTP) and the DNA polymerase. One of the four Dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is added to each reaction. These dideoxynucleotides are used as the chain-terminating nucleotides which lacks a 3'-OH group required for elongation. Incorporation of a dideoxynucleotide into the DNA strand therefore terminates DNA strand extension, resulting in various other DNA fragments of varying length.

Gel Electrophoresis: Each of the four DNA synthesis reactions is run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be read.

2. Shotgun Sequencing:

- The target sequence is amplified by inserting into a vector.
- The amplified DNA is randomly cut many times producing a redundant set of fragments (spanning the target several times).
- Each sequence obtained in this way is called a read. Since, there is redundancy in target sequence and the shearing is performed in a random way, the reads will overlap to a different extent.
- Algorithms are used to assemble the reads to get the contiguous sequence.

Challenges with Shotgun sequencing: None of these algorithms can produce correct assemblies if the sequence contains repeats. This problem of repeats can be resolved by clustering or linking them. It is good when the repeats are located in different pieces of DNA fragments.

Coverage: Coverage of a base position x is defined as the event whereby one or more sequence reads span x . The expected fractional coverage (FC) is

$$FC = 1 - \exp(-C), \text{ where } C = NL/G$$

N – Number of reads.

L – Length of reads.

G – Length of genomic segment.

Whole Genome Sequencing:

1. Hierarchical Sequencing: The genome is broken into pieces. Each of these pieces is mapped on to the genome by Physical Mapping. A minimum tiling path is selected and each of the clones in the path is sequenced with shotgun. The clones are assembled.

Methods of Physical Mapping:

Physical mapping: To make a map of the locations of each clone relative to one another and then use the map to select a minimal set of clones to sequence the genome.

Physical mapping can be done in two ways:

i) Restriction Mapping: Restriction enzymes are used to cut the DNA at specific sites. Mapping of restriction sites of a cutting enzyme on the target DNA based on the lengths of fragments is called Restriction Mapping.

Restriction Mapping Problem:

- Given a subset E of ΔX , construct X from E
- $\Delta X = \{ | X_i - X_j | : X_i, X_j \in X \}$
- Find $|\Delta X|$

Partial Digest: Only one enzyme is used cutting each clone. Fragments of all possible lengths are obtained.

Partial Digest Problem: X has to be constructed from E, where $E = \Delta X$. Example: if $E = \{3, 11, 17, 19, 8, 14, 16, 6, 8, 2\}$ then X comes to be

3	8	6	2
---	---	---	---

To solve this problem Polynomial time algorithm is not known till now. This problem is not known to be NP complete.

Algorithm to determine the restriction sites:

1. Find the longest distance in ΔX , and delete that distance from ΔX .
2. Repeatedly position the longest remaining distance of ΔX
 - Determine between two possible positions (one of the most outermost points) for the point.
 - For each of these two points, check whether all the distances from the position to the points already positioned are in ΔX .
 - If they are, delete all those distances from ΔX and proceed.
 - Backtrack if they are not for both of the two positions.

Double Digest Problem: In Double Digest, two enzymes A, B and their mixture A+B are used for cutting the clone. There are many equivalent solutions to this problem using the above algorithm. Double Digestion gives more information as it helps in catching more restriction sites in a DNA fragment.

2. Hybridization: In this process, the clones are mapped based on the hybridization data.

- Probes complementary to the clones are constructed.
- Each of the clones is treated with these probes.
- The probes that attach with particular clones are recorded.
- If the same probe attaches to two clones, then there is an overlap between those two clones.

The hybridization data is stored in a $n \times m$ matrix M, where $M(i,j) = 1$, if probe p_j attaches to the clone C_i ; otherwise $M(i,j) = 0$.

Mapping of clones here is equivalent to shortest string problem.

Shortest String problem: This problem deals with finding the shortest string in the alphabet of probes that covers all clones using the hybridization data M. A string S covers a clone C, if a substring of S contains exactly the same set of probes as C. For solving this problem the matrix should satisfy Consecutive 1's property. Error free hybridization matrix

should satisfy C1P property because a clone is a chunk and so there should not be any 0 in between the 1's. Thus, the matrix is permuted in such a manner that it satisfies the C1P.

C1P problem: This problem deals with permuting the matrix M ($n * m$) in such a manner that it satisfies the C1P.

Assumptions:

- All rows are different i.e. no two clones are same.
- No row has all 0's i.e. at least one probe binds to a particular clone.

Algorithm for the C1P problem:

- S_i is the set of column numbers (k) for which $M_{i,k}=1$ for each row i in M .
- For two rows i and j
 - 1) $S_i \cap S_j = \emptyset$
 - 2) S_i is a subset of S_j or vice versa
 - 3) $S_i \cap S_j \neq \emptyset$ and none of them is a subset of the other.

1. The rows are separated into components, where component is a connected set of clones (rows), G_c . A $G_c = (V_c, E_c)$ is created from M such that each vertex V_c represents a clone (row) and $(i,j) \in E_c$, if $S_i \cap S_j \neq \emptyset$ and none of them is a subset of the other.

2. The columns of the components are permuted.

3. The components are then joined together where a component α is connected to β , $\alpha \rightarrow \beta$, if S_i for all $i \in \beta$ are contained in at least one set of S_j of α .