Capacity of shotgun sequencing channel.

↓

C = rate at which message is transmitted with lowest probability possible

- $C = \max_{P_X(x)} I(X;Y)$

  ↓ input message — ↓ output code words

Shotgun sequencing :-

- Breaking the DNA into smaller fragments without any specification.

Running the sanger method / Sioxy method on the smaller fragments generated.

Sequence the smaller segments individually

These fragments are known as contigs.

After getting the sequence from all these different fragments,

the length and data of the sequence generated from the fragments is provided to the computer. The computer will find the overlapping regions present in these fragments.

After finding the overlapping regions, complete stretch of DNA seq can be obtained.

AGG|CT|
CT|CC|CAC|
CAC|TT|AGG|
AGG|C GAT

⇒ During reconstruction there is p(error) as some sequences obtained may get erased.

Reliable sequence reconstruction

Read length

length of the DNA fragments that are sequenced during the shotgun seq process

Read length ↑
Contiguous coverage of the genome ↑
more challenging & costly to achieve in practice.

Coverage length (total no. of reads)

extent to which the genome is covered by seq. fragments.

measure of how many times a specific position in a genome is covered by sequenced reads
coverage length ↑
↓
comprehensive rep of the genome

DNA mol – taking seq – codeword from a predefined codebook

Utilizing coding techniques in DNA-based storage, number of reads required for reliable reconstruction of a binary seq. can be reduced from $O(n)$ to $O(n/\log n)$ making storage system more efficient & enabling higher capacity with fewer reads of shorter length

i.i.d : Independent and Identically distributed

x^n – independent and identically distributed

Each random variable in the set is statistically independent of one another and follow same prob. distribution

$x^n \sim Ber(1/2)$

$x^n \to 0$ @ 1/2

$\to 1$ @ 1/2

$n \to \infty$ $\quad \ell = I \log n$ $\quad C = \frac{K\ell}{n}$

$\downarrow$

coverage depth

when $I < 2 \to$ reconstruction is impossible

$I > 2 \to$ " possible $\quad c_{LW} = \ln\left(\frac{n}{\epsilon}\right)$

$c_{LW} = \ln\left(\frac{n}{\epsilon}\right)$

all symbols in x^n are seq. with prob atleast $1-\epsilon$.

Feasibility region

K → factor that quantifies the fraction of the genome that is effectively covered by each read.

$O(n)$ / $O(n/\log n)$ → used to represent scaling behaviour or growth rate of certain quantities wrt size of input, which is denoted by n

Shuffling - Sampling channel

↓

The transmitted symbols are randomly rearranged or permuted during the comm. process.

no. of possible distinct length sequences

↓

$$2^{\bar{L}\log n} = n^{\bar{L}} = o(n/\log n) \Rightarrow o(K)$$

$$\frac{n^{\bar{L}}\log n}{n} \to 0 \text{ as } n \to \infty \qquad \frac{n^{\bar{L}}}{K} \to 0 \text{ as } n \to \infty$$

~~$\log n \to 0$ as $n \to \infty$~~

$$n^{\bar{L}} = o(n/\log n) \qquad \frac{n^{\bar{L}}}{K} \to 0 \text{ as } n \to \infty$$

when $\bar{L} < 1$

probability that a given symbol in X^n is not seq. by any of the K reads

$$= \left(1 - \frac{K}{n}\right)^K$$

$$= \left(1 - \frac{L}{n}\right)^{cn/L} = \left(1 - \frac{L}{n}\right)^{\frac{c}{L/n}}$$

$$= \lim_{n\to\infty} \left(1 - \frac{L}{n}\right)^{\frac{c}{L/n}}$$

$$= \lim_{n\to\infty} \log e^{\frac{-cK}{n}\left(\frac{n}{K}\right)}$$

$$= e^{-c}$$

$$c = \frac{k\bar{l}\log n}{n} = \frac{c}{\bar{l}}$$

$$c' = \frac{k(\bar{l}-1)\log n}{n}$$

$$c' = \frac{c}{\bar{l}}(\bar{l}-1)$$

$$\boxed{c' = c\left(1-\frac{1}{\bar{l}}\right)}$$

Even when $\log n$ bits are removed, still the c of the channel would remain the same, which signifies the fact $\log n$ is not providing any new information and is just helping in the ordering of the substrings.