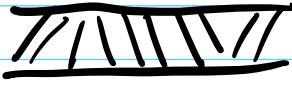
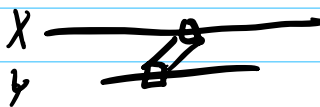


Lecture 8

10/6/26

Ch 1: Pairwise Sequence Alignment Algorithms

Global Alignment Algorithms 

Local Alignment Algorithms  X and Y can be any length

Scoring Matrices = Mathematical models of evaluation

Random simulations
Random model of protein sequences
• Ungapped multiple alignment

W	W	Y	I	R
W	F	Y	V	R
W	Y	Y	V	R
W	Y	F	I	R
W	Y	Y	T	R

Observed prob:

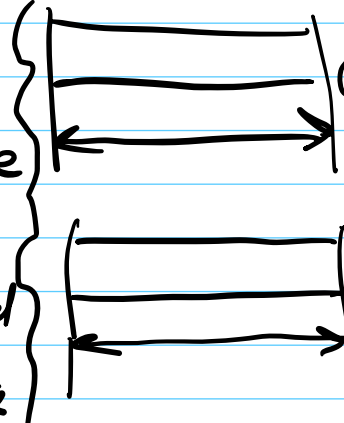
$$P(W) = 6/25$$

$$P(R) = 5/25$$

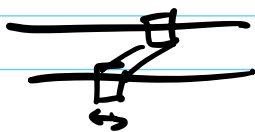
Random Protein Sequences

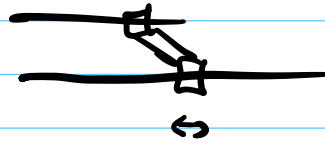
* Global Alignments of Pairs of Random Proteins with 5 Scoring Scheme

$$M=N$$

$D(N) = \text{Average length of } 1000 \text{ optimal alignments}$


* Local Alignments





- 1000 average of the lengths of the optimal local alignment
- $\max \{ \log(N), \log(M) \}$

★ PHASE-TRANSITION

$$\underset{\text{Global}}{O(N)} \rightarrow \underset{\text{Local}}{O(\log N)}$$

★ Karlin & Aetschul Statistic

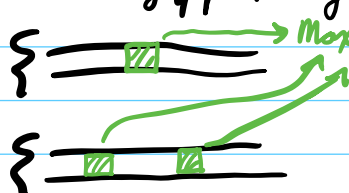
5 axioms:

- 1) A positive score in the scoring matrices
- 2) The expected score of the scoring matrix is NEGATIVE

$$S = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{bmatrix} \sum_{i,j=1}^k a_{ij} p_i p_j < 0$$

- 3) Letters in sequences are independently & identically distributed
- 4) Sequences are infinitely long
- 5) No gaps

= Ungapped alignments



Max score pairs: $2ij$
↓ Put into a database

$\begin{matrix} \xi = \\ \xi = \\ \xi = \\ \vdots \\ \xi = \end{matrix}$

Compute the distribution of amino acid pairs aligned in these max scoring segments

$$Z_{ij} = p_i p_j e^{-\lambda Z_{ij}}$$

↳ score of aligning i & j in the max scoring segments distributed

The Lady Tasting Tea Puzzle: introduces null hypothesis

- Lady claimed she can recognize ^{whether} tea or milk was placed in cup first based on just tasting it
- Fisher mathematician decided to test her skill
- Fisher gave her 8 cups.
 - 4 cups: milk
 - She passes only if she can identify all 8

Pertains to local alignment
as it is like the random model

Sir Fisher introduced the concept of the null hypothesis:
"never proved, or established, but is possibly disproved in the course of the experimentation."

Lady tasting tea: "Design of experiments":

- We provide lady 8 randomly ordered cups of tea/milk
 - 4 of them prepared by adding milk first
 - 4 of them prepared by adding tea first
- Her task: select 4 cups prepared by one method
- She was fully informed of the experiment
- The NULL HYPOTHESIS: The lady had no ability to

distinguish the teas.

- The test statistic: A simple count of the # of successes in selecting the 4 cups

- Success = the # of cups of the given type successfully selected

- Need to do random model to test H_0

- The distribution of possible number of successes observing the null hypothesis, was completed using the number of permutations.

- Using the combinatorial formula for $m=8$ total cups and $k=4$ cups chosen, there are:

$$\binom{8}{4} = \frac{8!}{4!(8-4)!} = 70 \text{ possible combinations}$$

$$\text{"8 choose 4"} = \frac{8!}{4!4!} = \frac{5 \cdot 6 \cdot 7 \cdot 8}{1 \cdot 2 \cdot 3 \cdot 4} = 70$$

- The frequencies of the possible number of successes are as follows.

For 0 successes: $\begin{matrix} 0 = \text{failure} \\ X = \text{success} \end{matrix}$ There is clearly 1 set of 4 choices (namely all incorrect) 0000

For 1 success & 3 failures: 000X, 00X0, 0X00, X000
• There are 4 correct cups of which 1 is selected: The number of all is:

$$\binom{4}{1} = 4 \text{ different ways}$$

- And independently of that, there are 4

incorrect cups of which we pick 3
 $\binom{4}{3} = 4$

• TOTAL $4 \cdot 4 = 16$

For 2 successes & 2 failures: $00XX, 0XX0, 0X0X, XX00, X0X0, X00X$
 $\therefore \binom{4}{2} = 6, \binom{4}{2} = 6$
 TOTAL = $6 \cdot 6 = 36$

For 3 successes & 1 failure: $\binom{4}{3} \cdot \binom{4}{1} = 4 \cdot 4 = 16$

For 4 successes & 0 failures: $XXXX$
 $\therefore \text{TOTAL} = 1$

Success count	# of permutations
0	1
1	16
2	36
3	16
4	1
	TOTAL = 70

• The # of successes is distributed according to the hypergeometric distribution

• The critical region for rejecting the null hypothesis of no ability to distinguish (tea vs. milk put in first) was the 4 successes

- The conventional probability criterion is $< 5\%$
 - For the lady, the 4 successes $\frac{1}{70} \approx 1.4\% < 5\%$
 - At least 3 successes $\frac{16+1}{70} \approx 24.3\% > 5\%$
 - That's why can't set ≥ 3 successes as the benchmark