CS1810

9/16/2025

Ch1. SEQUENCE ALIGNMENT ALGORITHMS

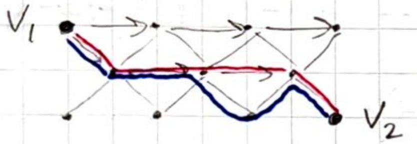
1.1. The Global Alignment Alg   Needleman-Wunsch (1970)
1.2. Heuristic Interpretation of alignment score as "likelihood"
1.3. Scoring scheme ≡ mathematical models of evolution

Principle of optimality

In a directed graph with weights/costs/lengths, the maximum cost path (optimum) is made out of optimal/max cost subpaths.

Proof: By contradiction

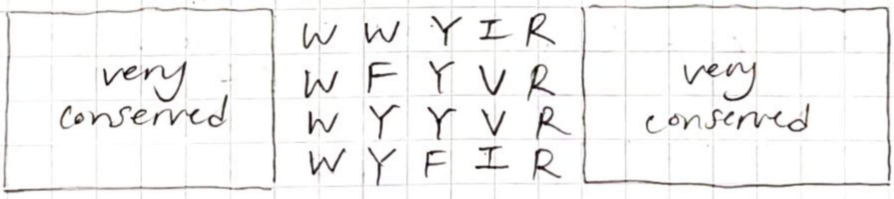
We have an optimal path between V_1 and V_2.



max cost path between V_1 and V_2
path cost > path cost
CONTRADICTION.

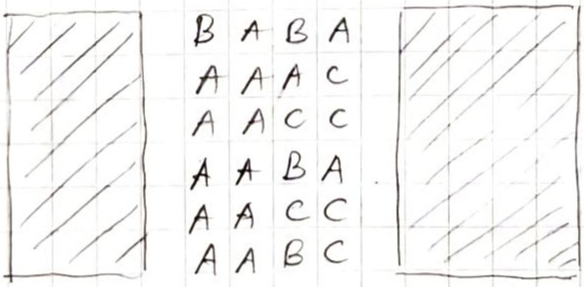
Scoring schemes
BLOSUM – Henikoff et al.
PAM – Margaret Dayhoff

Probabilistic Models of Biomolecular Sequences (DNA, RNA, Proteins)



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

3 amino acids A, B, C



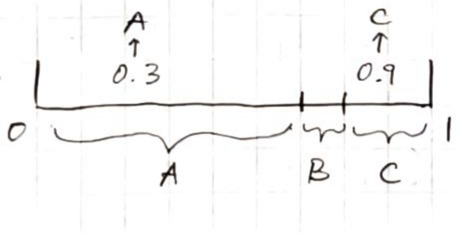
B A B A
A A A C
A A C C
A A B A
A A C C
A A B C

ungapped multiple alignment

Extract the rules of evolution

Random models
- for protein sequences ✓
- for pairwise alignment of protein sequences

# of A's :	14	$prob(A) = \frac{14}{24}$
B's :	4	$prob(B) = \frac{4}{24}$
C's :	6	$prob(C) = \frac{6}{24}$
sum	24	



0.3, 0.5, 0.1, 0.8, ...

4 columns
each column 6 letters

$4\cdot\binom{6}{2} = 4\cdot\frac{6\cdot5}{2} = 60$ aligned pairs of letters

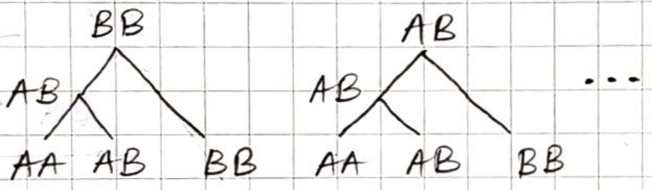
Aligned pairs	Frequency observed	Expected frequency	$2\log_2 \frac{\text{obs. freq}}{\text{exp. freq}}$
A to A	$\frac{26}{60}$	$\frac{14}{24}\times\frac{14}{24} = \frac{196}{576}$	0.70
A to B	$\frac{8}{60}$	$2\times\frac{14}{24}\times\frac{4}{24} = \frac{112}{576}$	−1.09
A to C	$\frac{10}{60}$	$\vdots = \frac{168}{576}$	−1.61
B to B	$\frac{3}{60}$	$= \frac{16}{576}$	1.70
B to C	$\frac{6}{60}$	$= \frac{48}{576}$	0.53
C to C	$\frac{7}{60}$	$= \frac{36}{576}$	1.86

These are used to calculate the "estimated likelihood ratio"
"$2\log_2\left(\frac{\text{observed}}{\text{expected}}\right)$"

$$\begin{array}{c|ccc} & A & B & C \\ \hline A & 1 & -1 & -2 \\ B & -1 & 2 & 1 \\ C & -2 & 1 & 2 \end{array}$$

matrix

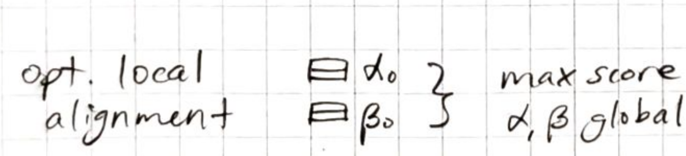
AA
AB
BB



15 trees

1.4. Local Alignment Algorithm

Smith-Waterman Alg



GLOBAL

opt: $O(N)$ length of alignment

LOCAL opt: $O(\log N)$

α ⊟
β ⊟ global alignment

opt. local alignment ⊟ α_0 ⊟ β_0 $\}$ max score among all α, β global alignments

Prefixes
Suffixes

X = ACCG

Pref	Suff	substrings
A	G	CC
AC	CG	C
ACC	CCG	
ACCG	ACCG	
= Pref (X)	= Suff (X)	= substrings (X)