Sequence Alignment in Bioinformatics

Overview

- Sequence Distance
  - Hamming Distance
  - Edit Distance
- Dynamic Programming
- Extensions
  - Local Alignment
  - Multi-sequence Alignment
  - Protein Alignment
- BLAST

Hamming Distance

- Game of Doublets from Lewis Carroll
  - CAT → DOG
  - HEAD→ TAIL
- Number of nucleotide changes
  - Distance between TGCATAT → ATCCGAT
  - Distance between ATATATA → TATATAT

Edit Distance

- Base our distance on evolution of sequences
  - Insertion
  - Deletion
  - Substitution
- Edit Distance between TGCATAT → ATCCGAT
- How to find minimum edit distance?
  - Minimum edit distance = best alignment

Alignment: Dynamic Programming

- Best alignment of strings length *n*
  - Best way to align strings length *n – 1* plus
  - Best way to align last characters
- Recursive Problem (defined in terms of self)
  - Base Case, when one or both sequences are zero characters long.
- Solve from the bottom up
  - Best alignment of first characters, then build

Alignment Grid

		A	T	A	G
	0	X ←	X ←	X ←	X ←
T	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
A	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
C	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
G	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

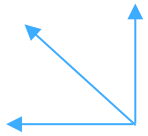
Solution

Alignment: Scores

- Use *i* and *j* to denote squares in grid
- $s_{0,0} = 0$
- Alignment option:
  - Borders = 1
  - Middle = 3

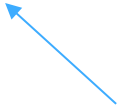
$$s_{i,j} = \max \begin{cases} s_{i-1,j} - \text{gap penalty} \\ s_{i,j-1} - \text{gap penalty} \\ s_{i-1,j-1} + \text{match if } v_i = w_j \\ \qquad\qquad - \text{mismatch if } v_i \mathrel{!}= w_j \end{cases}$$

Alignment: Backtracking

Arrows  show where the score originated

 gap in the top

 gap in the left

 match or mismatch

Alignment Grid

		A	T	A	G
	0	X ←	X ←	X ←	X ←
T	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
A	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
C	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
G	↑ X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X ← -1	X ← -2	X ← -3	X ← -4
T	↑ -1 X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
A	↑ -2 X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
C	↑ -3 X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
G	↑ -4 X	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X -1	X -2	X -3	X -4
T	↑ -1 / X / -1	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←
A	↑ -2 / X / -2	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←
C	↑ -3 / X / -3	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←
G	↑ -4 / X / -4	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←	↑ / ↖ / ←

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X -1	-1 X -2	-2 X -3	-3 X -4
T	↑ -1 X	-1	-1		
A	↑ -2 X	-2			
C	↑ -3 X	-3			
G	↑ -4 X	-4			

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X / ← -1 / -1	X / ← -2 / -2	X / ← -3 / -3	X / ← -4 / -4
T	↑ -1 / X / -1	↑ -2 / ↖ -1 / ← -2 / -1	↑ -3 / ↖ 0 / ← -2 / 0	↑ / ↖ / ← /	↑ / ↖ / ← /
A	↑ -2 / X / -2	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /
C	↑ -3 / X / -3	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /
G	↑ -4 / X / -4	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X / ← -1 / -1	X / ← -2 / -2	X / ← -3 / -3	X / ← -4 / -4
T	↑ -1 / X / -1	↑ -2 / ↖ -1 / ← -2 / -1	↑ -3 / ↖ 0 / ← -2 / 0	↑ -4 / ↖ -3 / ← -1 / -1	↑ -5 / ↖ -4 / ← -2 / -2
A	↑ -2 / X / -2	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /
C	↑ -3 / X / -3	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /
G	↑ -4 / X / -4	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X ← -1 -1	X ← -2 -2	X ← -3 -3	X ← -4 -4
T	↑ -1 X -1	↑ -2 ↖ -1 ← -2 -1	↑ -3 ↖ 0 ← -2 0	↑ -4 ↖ -3 ← -1 -1	↑ -5 ↖ -4 ← -2 -2
A	↑ -2 X -2	↑ -2 ↖ 0 ← -3 0	↑ -1 ↖ -2 ← -1 -1	↑ -2 ↖ 1 ← -2 1	↑ -3 ↖ -2 ← 0 0
C	↑ -3 X -3	↑ ↖ ← 	↑ ↖ ← 	↑ ↖ ← 	↑ ↖ ←
G	↑ -4 X -4	↑ ↖ ← 	↑ ↖ ← 	↑ ↖ ← 	↑ ↖ ←

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X ← -1 / -1	X ← -2 / -2	X ← -3 / -3	X ← -4 / -4
T	↑ -1 / X / -1	↑ -2 / ↖ -1 / ← -2 / -1	↑ -3 / ↖ 0 / ← -2 / 0	↑ -4 / ↖ -3 / ← -1 / -1	↑ -5 / ↖ -4 / ← -2 / -2
A	↑ -2 / X / -2	↑ -2 / ↖ 0 / ← -3 / 0	↑ -1 / ↖ -2 / ← -1 / -1	↑ -2 / ↖ 1 / ← -2 / 1	↑ -3 / ↖ -2 / ← 0 / 0
C	↑ -3 / X / -3	↑ -1 / ↖ -3 / ← -4 / -1	↑ -2 / ↖ -1 / ← -2 / -1	↑ 0 / ↖ -2 / ← -2 / 0	↑ -1 / ↖ 0 / ← -1 / 0
G	↑ -4 / X / -4	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /	↑ / ↖ / ← /

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Alignment Grid

		A	T	A	G
	0	X -1 / ← -1	-1	X -2 / ← -2	-2	X -3 / ← -3	-3	X -4 / ← -4	-4
T	↑ -1 / X	-1	↑ -2 / ↖ -1 / ← -2	-1	↑ -3 / ↖ 0 / ← -2	0	↑ -4 / ↖ -3 / ← -1	-1	↑ -5 / ↖ -4 / ← -2	-2
A	↑ -2 / X	-2	↑ -2 / ↖ 0 / ← -3	0	↑ -1 / ↖ -2 / ← -1	-1	↑ -2 / ↖ 1 / ← -2	1	↑ -3 / ↖ -2 / ← 0	0
C	↑ -3 / X	-3	↑ -1 / ↖ -3 / ← -4	-1	↑ -2 / ↖ -1 / ← -2	-1	↑ 0 / ↖ -2 / ← -2	0	↑ -1 / ↖ 0 / ← -1	0
G	↑ -4 / X	-4	↑ -2 / ↖ -4 / ← -5	-2	↑ -2 / ↖ -2 / ← -3	-2	↑ -1 / ↖ -2 / ← -3	-1	↑ -1 / ↖ 1 / ← -2	1

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution

Final Alignment

		A		T		A		G	
	0	X ← -1	-1	X ← -2	-2	X ← -3	-3	X ← -4	-4
T	↑ -1 X	↑ -2 ↖ -1 ← -2	-1	↑ -3 ↖ 0 ← -2	0	↑ -4 ↖ -3 ← -1	-1	↑ -5 ↖ -4 ← -2	-2
A	↑ -2 X	↑ -2 ↖ 0 ← -3	0	↑ -1 ↖ -2 ← -1	-1	↑ -2 ↖ 1 ← -2	1	↑ -3 ↖ -2 ← 0	0
C	↑ -3 X	↑ -1 ↖ -3 ← -4	-1	↑ -2 ↖ -1 ← -2	-1	↑ 0 ↖ -2 ← -2	0	↑ -1 ↖ 0 ← -1	0
G	↑ -4 X	↑ -2 ↖ -4 ← -5	-2	↑ -2 ↖ -2 ← -3	-2	↑ -1 ↖ -2 ← -3	-1	↑ -1 ↖ 1 ← -2	1

Substitution Matrix	
-1	Substitution Mismatch
-1	Gap Introduction
+1	Match
Bias	
Left, Up, Diagonal	

Solution
ATA_G
_TACG

Try This Yourself

		T	C	A	G
	0	X ←	X ←	X ←	X ←
T	↑ X ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
A	↑ X ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
G	↑ X ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←
G	↑ X ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←	↑ ↖ ←

Substitution Matrix	
-3	Substitution Mismatch
-1	Gap Introduction
+4	Match
Bias	
Left, Up, Diagonal	

Solution

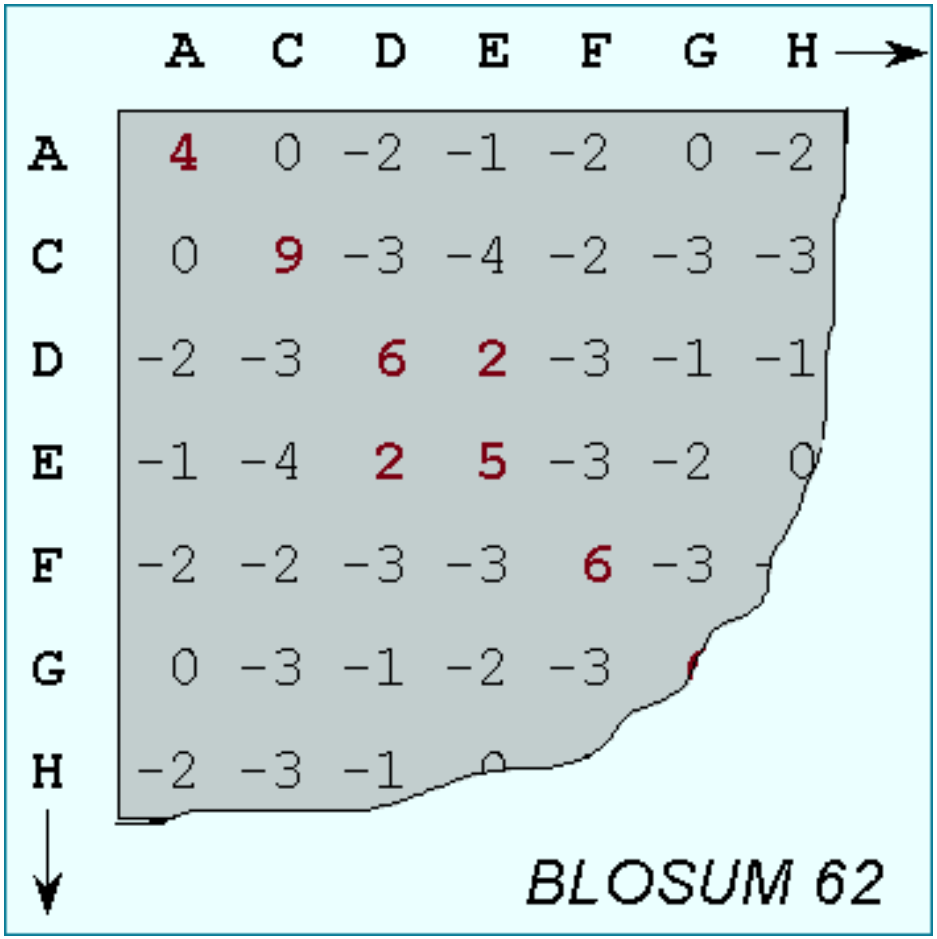
Local Alignment

- Not always good to align from start to end on both sequences
  - TAG
  - ATCCACTAGAGG
- Algorithm
  - Perform Dynamic Programming as before
  - Find highest square
  - Trace back until score becomes negative

Protein Alignment

- Substitution Matrices for Proteins
  - PAM (Point Accepted Mutation)
    - Based on small changes over short timescale
    - Extrapolated to larger divergence
  - BLOSUM (BLock SUbstitution Matrix)
    - Based on alignment of divergent proteins
    - BLOSUM62 default in BLAST
- Vary from less to more divergent evolution

Portion of BLOSUM62



http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html

BLAST

- Basic Local Alignment Search Tool
  - http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Nucleotides
- Proteins
- FASTA format for genomic data