

Sequence Alignment

Global Alignment
and
Multiple Sequence Alignment

Recall:

Local Sequence Alignment

- Optimal Local Alignment of two sequences
 - Use substitution matrix
 - log-odds values
 - best alignment has highest score
- Uses 0 in matrix if score is less than 0
- Look for largest value of any subsequence in the matrix

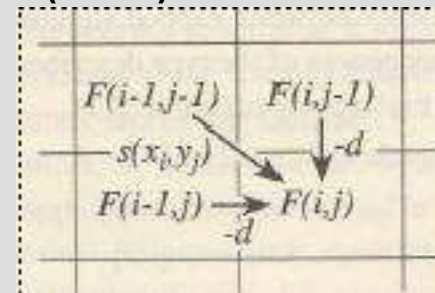
Global Sequence Alignment

- Find the highest scoring global alignment, allowing gaps
- Again, we design a substitution matrix
 - Rank matches
 - Account for substitutions and gaps
 - Can be built recursively or iteratively
- Dynamic programs are more accurate than heuristic models

Needleman-Wunsch Algorithm

- Dynamic programming solution
- To begin, build F recursively, from F(0,0):

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d. \end{cases}$$



- d is the 'gap-open' penalty
- s is the 'log-odds' ratio
- Maintain reference to cell that $F(i, j)$ is derived from

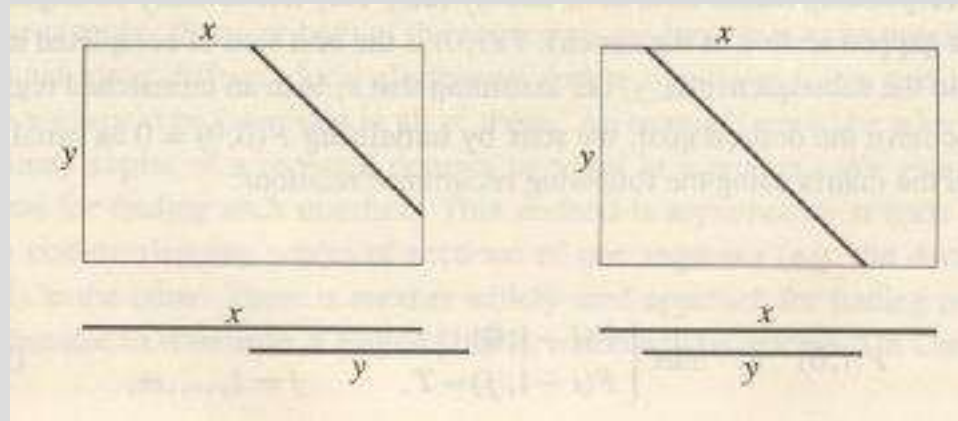
Needleman-Wunsch Cont.

	H	E	A	G	A	W	G	H	E	E	
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

- 'Traceback' from the final cell to the initial one, following the pointers.
 - Add symbols to the alignment

Overlap matches

- Occurs when one sequence contains the other, or they overlap



- Start on top or left side
- Finish on bottom or right side

Overlap matches algorithm

- New function for scoring top and left edge

$$F(i,0) = \max \begin{cases} F(i-1,0), \\ F(i-1,m) - T; \end{cases}$$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i,y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

	H	E	A	G	A	W	G	H	E	E
P	0	0	0	0	0	0	0	0	0	0
A	0	-2	-1	-1	-2	-1	-4	-2	-2	-1
W	0	-3	-5	-4	1	-4	18	10	2	6
H	0	10	2	6	-6	-1	10	16	20	12
E	0	2	16	8	0	7	2	8	16	26
A	0	-2	8	21	13	5	3	2	8	18
E	0	0	4	13	18	12	4	4	2	14

Multiple Sequence Alignment

- Global
 - Pair-wise alignment extended
 - Dynamic programming can extend to 3 sequences

Challenges of Multiple Sequence Alignment

- Computation needed (memory)
 - Minimize the computations (approximate)
- Properly reflecting ancestry of divergent sequences
 - Weighted algorithms

MSA

- Optimal alignment using dynamic programming
- Resource Intensive
- Small number of shorter sequences
- Epsilon and Delta – how closely related or distantly related the sequence pairs are

Progressive Alignment

- Uses dynamic programming to build a closely related multiple sequence alignment
- Progressively adds less-related sequences
- Programs using progressive methods
 - CLUSTALW
 - PILEUP
 - T-COFFEE

CLUSTAL(W)

- Create and use a phylogenetic tree
- Neighbor join sequences according to the tree with dynamic programming
- Weighted
- Tries to place gaps between conserved regions
- Fast

PILEUP

- Needleman-Wunsch algorithm for sequence alignment
- Standard scoring matrix used
- No weighting

T-COFFEE

- Tree-based Consistency based Objective Function For alignment Evaluation
- Uses CLUSTAL for an optimal global alignment of all possible sequence pairs
- Uses LALIGN for local alignment
- Good at reproducing known alignments
- Slow

Local Multiple Sequence Alignment

- Local Alignment
 - Look for blocks (ungapped regions)
 - Pattern searching