Algorithmic Approaches for Biological Data, Lecture #20

Katherine St. John

City University of New York
American Museum of Natural History

20 April 2016

Outline

- Aligning with Gaps and Substitution Matrices

Outline

- Aligning with Gaps and Substitution Matrices
- Global versus Local Alignment

Outline

- Aligning with Gaps and Substitution Matrices
- Global versus Local Alignment
- Searching Graphs: Breadth First & Depth First

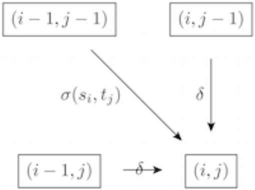
Pairwise Sequence Alignment

		A	G	A	G
	0	-1	-2	-3	-4
A	-1	1			
G	-2				
G	-3				

Pairwise Sequence Alignment

		A	G	A	G
	0	-1	-2	-3	-4
A	-1	1			
G	-2				
G	-3				

- Pictorially:



Pairwise Sequence Alignment

		A	G	A	G
	0	-1	-2	-3	-4
A	-1	1			
G	-2				
G	-3				

- Pictorially:

$$\begin{tikzcd}
\boxed{(i-1,j-1)} \arrow[dr, "\sigma(s_i,t_j)"'] & \boxed{(i,j-1)} \arrow[d, "\delta"'] \\
\boxed{(i-1,j)} \arrow[r, "\delta"] & \boxed{(i,j)}
\end{tikzcd}$$

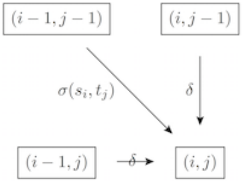
- As equations:

$$S(s_{0..i}, t_{0..j}) = \max \begin{cases} \sigma(s_i, t_j) + S(s_{0..i-1}, t_{0..j-1}) \\ -\delta + S(s_{0..i-1}, t_{0..j}) \\ -\delta + S(s_{0..i}, t_{0..j-1}) \end{cases}$$

where:

$$\delta = 1 \text{ and } \sigma(s,t) = \begin{cases} 1 \text{ if } s = t \\ -1 \text{ otherwise} \end{cases} .$$

Aligning with Gaps and Substitution Matrices



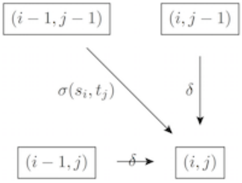
$$S(s_{0..i},t_{0..j}) = \max \begin{cases} \sigma(s_i,t_j) + S(s_{0..i-1},t_{0..j-1}) \\ -\delta + S(s_{0..i-1},t_{0..j}) \\ -\delta + S(s_{0..i},t_{0..j-1}) \end{cases}$$

where:

$$\delta = 1 \text{ and } \sigma(s,t) = \begin{cases} 1 \text{ if } s=t \\ -1 \text{ otherwise} \end{cases}.$$

- The basic dynamic programming format can be adjusted for different gaps and substitutions models.

Aligning with Gaps and Substitution Matrices



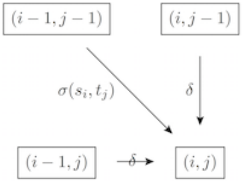
$$S(s_{0..i},t_{0..j}) = \max \begin{cases} \sigma(s_i,t_j) + S(s_{0..i-1},t_{0..j-1}) \\ -\delta + S(s_{0..i-1},t_{0..j}) \\ -\delta + S(s_{0..i},t_{0..j-1}) \end{cases}$$

where:

$$\delta = 1 \text{ and } \sigma(s,t) = \begin{cases} 1 \text{ if } s = t \\ -1 \text{ otherwise} \end{cases}.$$

- The basic dynamic programming format can be adjusted for different gaps and substitutions models.
- δ: the gap penalty

Aligning with Gaps and Substitution Matrices



$$S(s_{0..i},t_{0..j}) = \max \begin{cases} \sigma(s_i,t_j) + S(s_{0..i-1},t_{0..j-1}) \\ -\delta + S(s_{0..i-1},t_{0..j}) \\ -\delta + S(s_{0..i},t_{0..j-1}) \end{cases}$$

where:

$$\delta = 1 \text{ and } \sigma(s,t) = \begin{cases} 1 \text{ if } s = t \\ -1 \text{ otherwise} \end{cases} .$$

- The basic dynamic programming format can be adjusted for different gaps and substitutions models.
- δ: the gap penalty
- σ: scores matches/mismatches.

Gaps Are Treated Equally

		A	G	A	G
	0	-1	-2	-3	-4
A	-1	1			
G	-2				
G	-3				

- Commonly use affine gap penalty function:

Gaps Are Treated Equally

		A	G	A	G
	0	-1	-2	-3	-4
A	-1	1			
G	-2				
G	-3				

- Commonly use affine gap penalty function:
    - *h*: penalty associated with opening a gap
    - *g*: (smaller) penalty associated with extending the gap.

Gaps Are Treated Equally

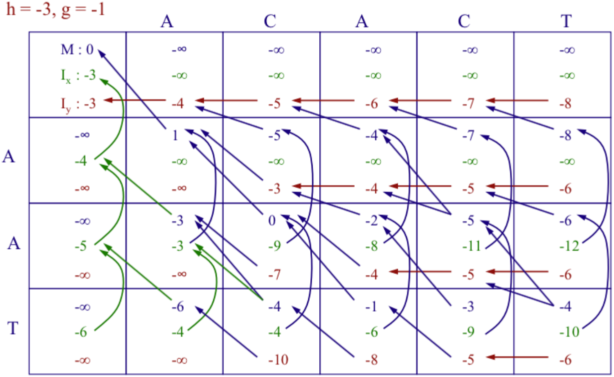
		A	G	A	G
	0	-1	-2	-3	-4
A	-1	1			
G	-2				
G	-3				

- Commonly use affine gap penalty function:
  - h: penalty associated with opening a gap
  - g: (smaller) penalty associated with extending the gap.
- To implement this efficiently, use 2 additional matrices that keeps track of the gaps (one for each sequence).

Affine Gap

Global Alignment Example

(Affine Gap Penalty)



h = -3, g = -1

		A	C	A	C	T
	M : 0 I_x : -3 I_y : -3	-∞ -∞ -4	-∞ -∞ -5	-∞ -∞ -6	-∞ -∞ -7	-∞ -∞ -8
A	-∞ -4 -∞	1 -∞ -∞	-5 -∞ -3	-4 -∞ -4	-7 -∞ -5	-8 -∞ -6
A	-∞ -5 -∞	-3 -3 -∞	0 -9 -7	-2 -8 -4	-5 -11 -5	-6 -12 -6
T	-∞ -6 -∞	-6 -4 -∞	-4 -4 -10	-1 -6 -8	-3 -9 -5	-4 -10 -6

Burr Settles, U Wisconsin, 2008

Using Substitution Matrices

- Can view $\sigma(i, j)$ as a substitution matrix.

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Using Substitution Matrices

- Can view $\sigma(i, j)$ as a substitution matrix.
- Substitution matrices commonly used for protein seqeunces.

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Using Substitution Matrices

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

- Can view $\sigma(i, j)$ as a substitution matrix.
- Substitution matrices commonly used for protein seqeunces.
- PAM $=$ Percent Accepted Mutation

Using Substitution Matrices

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

- Can view $\sigma(i, j)$ as a substitution matrix.
- Substitution matrices commonly used for protein seqeunces.
- PAM = Percent Accepted Mutation
  - Dayhoff *et al.*, 1978
  - Used for closely related protein sequences
  - Based on global alignment

Using Substitution Matrices

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

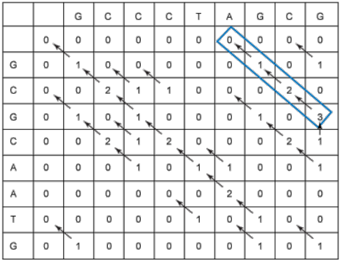
- Can view $\sigma(i, j)$ as a substitution matrix.
- Substitution matrices commonly used for protein seqeunces.
- PAM = Percent Accepted Mutation
  - Dayhoff *et al.*, 1978
  - Used for closely related protein sequences
  - Based on global alignment
- BLOSUM = Blocks Substitution Matrix

Using Substitution Matrices

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

- Can view $\sigma(i,j)$ as a substitution matrix.
- Substitution matrices commonly used for protein seqeunces.
- PAM = Percent Accepted Mutation
  - Dayhoff *et al.*, 1978
  - Used for closely related protein sequences
  - Based on global alignment
- BLOSUM = Blocks Substitution Matrix
  - Henikoff & Henikoff, 1992
  - Used for more divergent sequences
  - Based on local alignment

Global versus Local Alignment

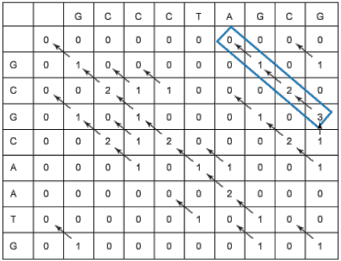


		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Paul Reiners, IBM, 2008

- Global: Needleman & Wunsch, 1970.
- Local: Smith & Waterman, 1981.
- Instead of looking for the global best score, look for the best score for subsequences of the initial sequences.

Global versus Local Alignment

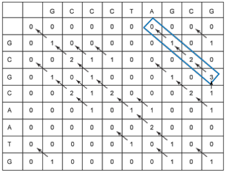


		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Paul Reiners, IBM, 2008

- Global: Needleman & Wunsch, 1970.
- Local: Smith & Waterman, 1981.
- Instead of looking for the global best score, look for the best score for subsequences of the initial sequences.
- Examples:
  - finding motifs (conserved patterns) across sequences,
  - comparing sequences against longer sequences (e.g. blast search).

Smith-Waterman Algorithm

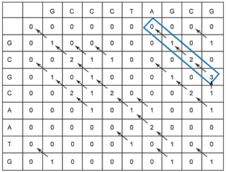


		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Paul Reiners, IBM, 2008

- The equation is slightly different:

Smith-Waterman Algorithm



		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

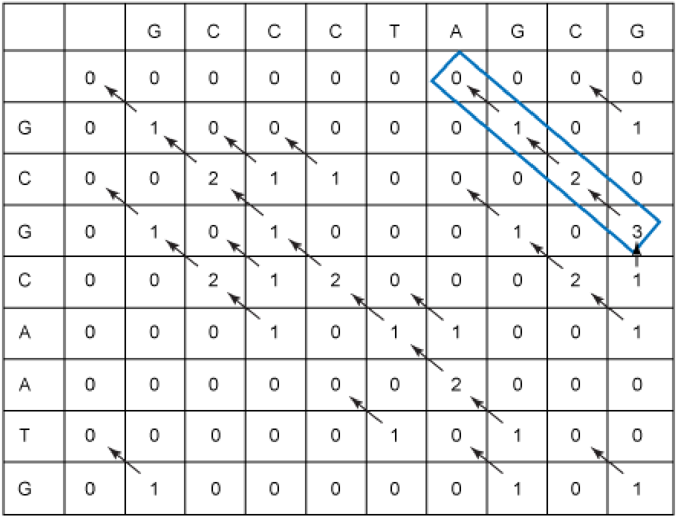
Paul Reiners, IBM, 2008

- The equation is slightly different:

$$s(i,j) = max \begin{cases} \sigma(i,j) + s(i-1,j-1) \\ -\delta + s(i,j-1) \\ -\delta + s(i-1,j) \\ 0 \end{cases}$$

- Initialize: first row and first column set to 0's
- Traceback: find maximum value of $s(i,j)$ anywhere in the the matrix, stop when we get to a cell with 0.

Smith-Waterman Algorithm



		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Paul Reiners, IBM, 2008

In Pairs: Local Alignment

		A	A	G	A
T					
T					
A					
A					
G					

- Use σ from Monday, but $\delta = 2$.

In Pairs: Local Alignment

		A	A	G	A
T					
T					
A					
A					
G					

- Use σ from Monday, but $\delta = 2$.
- What are the best local alignments?

In Pairs: Local Alignment

		A	A	G	A
	0	0	0	0	0
T	0				
T	0				
A	0				
A	0				
G	0				

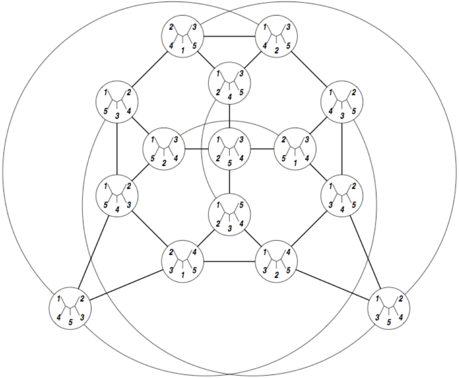
- Use σ from Monday, but $\delta = 2$.
- What are the best local alignments?

In Pairs: Local Alignment

		A	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
T	0	0	0	0	0
A	0	1	1	0	1
A	0	1	2	0	1
G	0	0	0	3	1

- Use σ from Monday, but $\delta = 2$.
- What are the best local alignments?

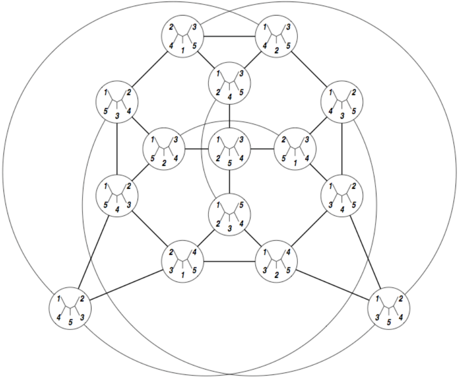
In Pairs: Searching Graphs



Bastert *et al.*, 2002

- Develop a strategy to visit every node of the graph
  (i.e. what data structures are needed?)

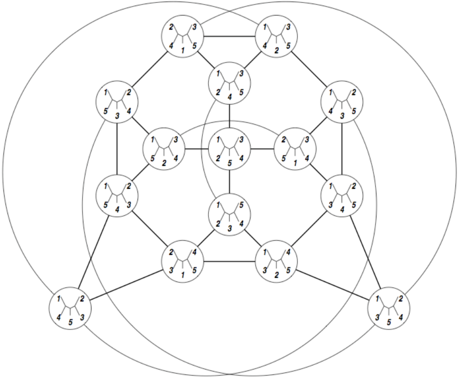
In Pairs: Searching Graphs



Bastert *et al.*, 2002

- Develop a strategy to visit every node of the graph
  (i.e. what data structures are needed?)
- The bookkeeping is important.

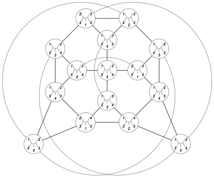
In Pairs: Searching Graphs



Bastert *et al.*, 2002

- Develop a strategy to visit every node of the graph
  (i.e. what data structures are needed?)
- The bookkeeping is important.

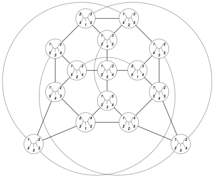
In Pairs: Searching Graphs



Bastert *et al.*, 2002

- Two common strategies:

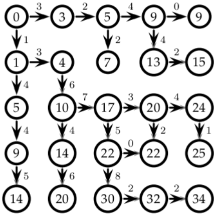
In Pairs: Searching Graphs



Bastert *et al.*, 2002

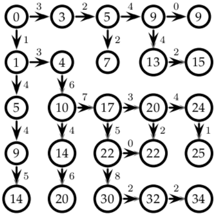
- Two common strategies:
  - Breadth First Search (BFS): visit all the neighbors, then visit all the neighbors' neighbors, etc.
  - Depth First Search (DFS): for each neighbor, visit its' neighbors, and continue as far down as possible.
- Bookkeeping is important:
  - Keep a "To Do" list (priority queue) of nodes still to visit.
  - Mark nodes as you visit them, so, you know not to visit again.

Recap



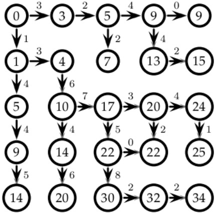
- Dynamic Programming: will do local & global alignments in lab today.

Recap



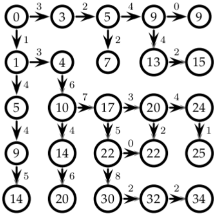
- Dynamic Programming: will do local & global alignments in lab today.
- More on searching graphs on Monday.

Recap



- Dynamic Programming: will do local & global alignments in lab today.
- More on searching graphs on Monday.
- Email lab reports to `kstjohn@amnh.org`.

Recap



- Dynamic Programming: will do local & global alignments in lab today.
- More on searching graphs on Monday.
- Email lab reports to `kstjohn@amnh.org`.
- Challenges available at `rosalind.info`.