

Sequence Alignment

BCB 567 Fall 2007

The Problem

- Given two biological sequences, we wish to describe the similarity and differences between the two sequences.
 - A score
 - A number describing the sequences' similarity
 - An alignment
 - A way in which to show which regions of one sequences correspond to which regions on the other sequence

Why Alignments?

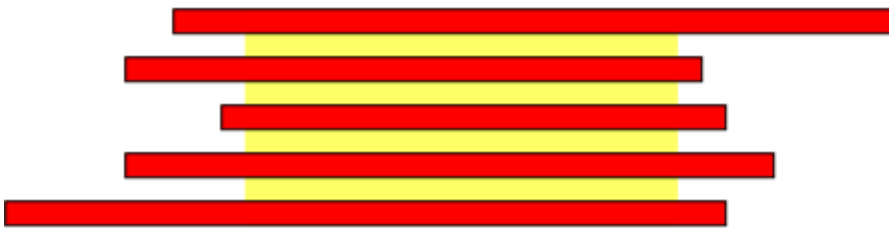
- Decide how sequences are evolutionarily related
- Discover how short sequences from a sequencing project fit together, to form larger sequences
- Discover for which genes an RNA codes
- Find commonalities among sequences

Types of Alignment

Pairwise



Multiple



Global



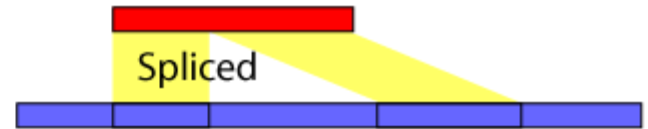
Suffix-Prefix



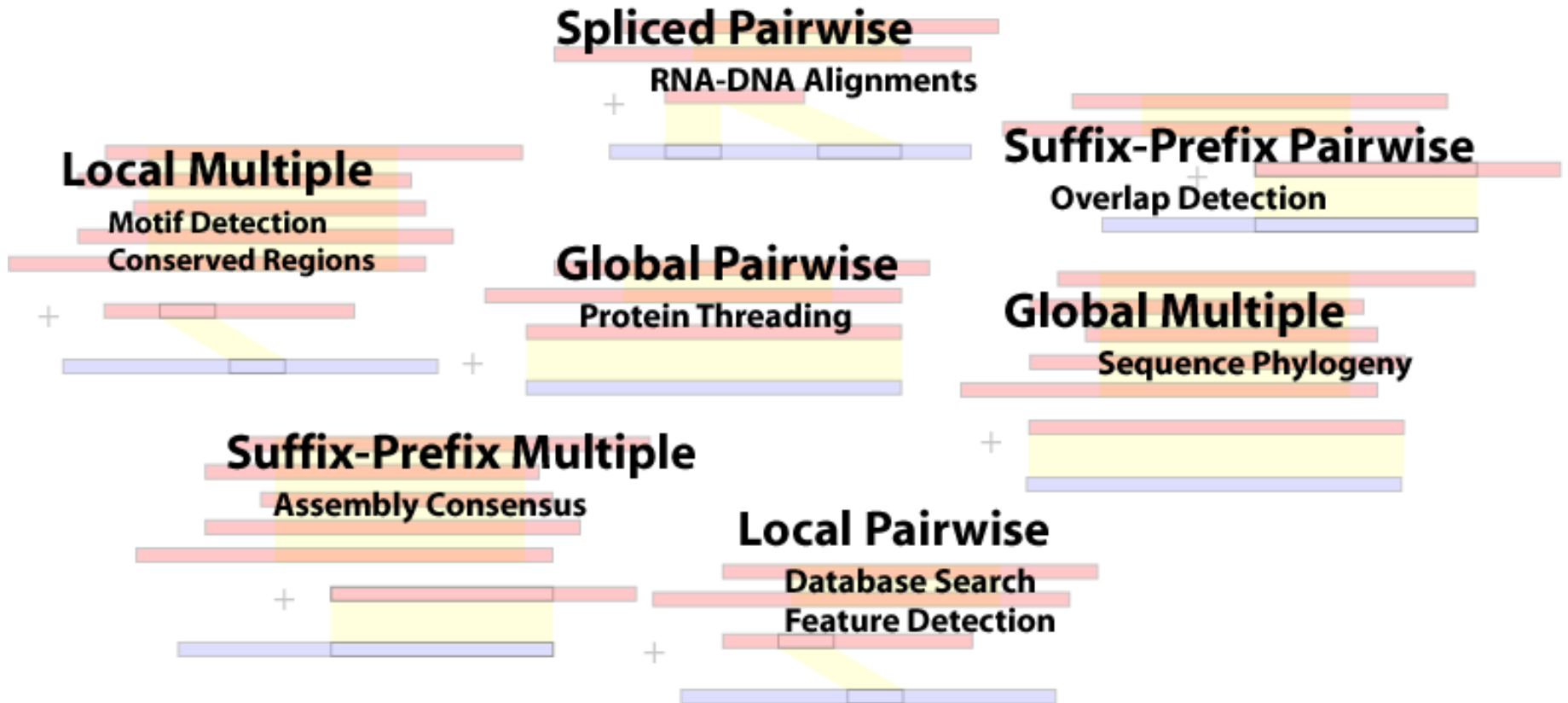
Local



Spliced



Examples of Alignment Uses



Strings

Alphabet {A, C, T, G}

String S = **A C T C A G A G T C**
 0 1 2 3 4 5 6 7 8 9

Index S[5] = G

Substring S[3,5] = C A G A

Prefix S[0,3] = A C T

Suffix S[6,9] = G T C

S[0,9] = A C T C A G A G T C



1-length of S

Global Alignment

- Input: Two strings, labeled A and B
 - Length of A we call n
 - Length of B we call m
- Output: Two strings A_A and B_A such that:
 - A_A and B_A are of equal length L ,
 - The characters of A_A consist only characters in A in the same order, with a special gap character '-' possibly inserted in some positions.
 - The characters of B_A consist only characters in B in the same order, with a special gap character '-' possibly inserted in some positions.
 - If $A_A[i] = '-'$, then $B_A[i] \neq '-'$
 - If $B_A[i] = '-'$, then $A_A[i] \neq '-'$

Input Sequences

ACTGACATAG

AGAGGCG

Valid Alignments

ACTGA--CATAG

A--GAGGC---G

ACTG-A-CATAG

A--GAGGC--AG

ACTGACATAG

A--GA-GGCG

Invalid Alignments

A-CTGATAG

AGA-GGCG-

A-CTGACATAG

A-GAG---GCG

ACTGACATAG

A-GAGGCG

Optimal Alignments

- Lets say that we are given some scoring function that evaluates the “goodness” of an alignment:

$$E(A_A, B_A)$$

- Now, we wish to find an optimal alignment, or an A_A and B_A that maximize this score.
- We will denote the alignment that maximizes this function as $opt(A, B)$

What if??

- $E(A_A, B_A) = E(A_A[0,i], B_A[0,i]) + E(A_A[i+1,L], B_A[i+1,L])$

$$E\left(\begin{array}{|c|c|} \hline \text{red} & \text{red} \\ \hline 0 & L \\ \hline \text{yellow} & \text{yellow} \\ \hline \text{blue} & \text{blue} \\ \hline \end{array}\right) = E\left(\begin{array}{|c|c|} \hline \text{red} & \text{red} \\ \hline 0 & i \\ \hline \text{yellow} & \text{yellow} \\ \hline \text{blue} & \text{blue} \\ \hline \end{array}\right) + E\left(\begin{array}{|c|c|} \hline \text{red} & \text{red} \\ \hline i+1 & L \\ \hline \text{yellow} & \text{yellow} \\ \hline \text{blue} & \text{blue} \\ \hline \end{array}\right)$$

$$E\left(\begin{array}{c} \text{ACTAC-CTG} \\ \text{AC-AGACTA} \end{array}\right) = E\left(\begin{array}{c} \text{ACTAC} \\ \text{AC-AG} \end{array}\right) + E\left(\begin{array}{c} \text{-CTG} \\ \text{ACTA} \end{array}\right)$$

\uparrow $\text{opt}(A, B)$ \uparrow $\text{opt}(\text{ACTAC}, \text{ACAG})$ \uparrow $\text{opt}(\text{CTG}, \text{ACTA})$

$$E\left(\begin{array}{|c|} \hline \text{red} \\ \hline \text{0} \quad \quad \quad \text{L} \\ \hline \text{blue} \\ \hline \end{array}\right) = E\left(\begin{array}{|c|} \hline \text{red} \\ \hline \text{0} \quad \quad \quad \text{L-1} \\ \hline \text{blue} \\ \hline \end{array}\right) + E\left(\begin{array}{|c|} \hline \text{red} \\ \hline \text{L} \\ \hline \text{blue} \\ \hline \end{array}\right)$$

$$E\left(\begin{array}{|c|} \hline \text{red} \\ \hline \text{0} \quad \text{Opt(A,B)} \quad \text{L} \\ \hline \text{blue} \\ \hline \end{array}\right) =$$

XXXXXXXXXXXXXXXXXXXX C

XXXXXXXXXXXXXXXXXXXX C

XXXXXXXXXXXXXXXXXXXX -

XXXXXXXXXXXXXXXXXXXX C

XXXXXXXXXXXXXXXXXXXX C

XXXXXXXXXXXXXXXXXXXX -

$$E \left(\begin{array}{c} \text{---} \\ 0 \quad \text{Opt}(A,B) \quad L \\ \text{---} \end{array} \right) =$$

maximum of

XXXXXXXX –
XXXXXXXX C

$$E \left(\begin{array}{c} \text{Opt}(A[0,n-1],B[0,m-2]) \\ \text{---} \\ 0 \quad \quad \quad L-1 \\ \text{---} \end{array} \right) + E \left(\begin{array}{c} - \\ \text{---} \\ B[m-1] \\ \text{---} \end{array} \right)$$

XXXXXXXX C
XXXXXXXX –

$$E \left(\begin{array}{c} \text{Opt}(A[0,n-2],B[0,m-1]) \\ \text{---} \\ 0 \quad \quad \quad L-1 \\ \text{---} \end{array} \right) + E \left(\begin{array}{c} A[n-1] \\ \text{---} \\ - \\ \text{---} \end{array} \right)$$

XXXXXXXX C
XXXXXXXX C

$$E \left(\begin{array}{c} \text{Opt}(A[0,n-2],B[0,m-2]) \\ \text{---} \\ 0 \quad \quad \quad L-1 \\ \text{---} \end{array} \right) + E \left(\begin{array}{c} A[n-1] \\ \text{---} \\ B[m-1] \\ \text{---} \end{array} \right)$$

Notation

- $S(i,j) = E(\text{Opt}(A[0,i],B[0,j]))$
 - “ $S(i,j)$ is the evaluated score of the optimal alignment between the prefix of A ending at position i and the prefix of B ending at position j .”

$$S(n-1, m-1) = \max \begin{cases} S(n-2, m-2) + E(A[n-1], B[m-1]) \\ S(n-1, m-2) + E('-', B[m-1]) \\ S(n-2, m-1) + E(A[n-1], '-') \end{cases}$$

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + E(A[i], B[j]) \\ S(i, j-1) + E('-', B[j]) \\ S(i-1, j) + E(A[i], '-') \end{cases}$$

Simple Scoring Method

- Motivations
 - Sequencing Error
 - Evolutionary Model
- Gap Penalty: γ
- Mismatch Penalty: β
- Match Bonus: α

ACTAC-CTG

AC-AGACTA

$$S(A_A, B_A) = 5\alpha + 2\beta + 2\gamma$$

$$\text{for: } \alpha=2, \beta=-1, \gamma=-1$$

$$S(A_A, B_A) = 10 - 2 - 2 = 6$$

$$\delta(i, j) = \begin{cases} \alpha & A[i] = B[j] \\ \beta & A[i] \neq B[j] \end{cases}$$

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + E(A[i], B[j]) \\ S(i, j-1) + E('-', B[j]) \\ S(i-1, j) + E(A[i], '-') \end{cases}$$

Becomes

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + \delta(i, j) \\ S(i, j-1) + \gamma \\ S(i-1, j) + \gamma \end{cases}$$

$$\delta(i, j) = \begin{cases} \alpha & A[i] = B[j] \\ \beta & A[i] \neq B[j] \end{cases}$$

The Dynamic Programming Table

		Sequence B											
		A	G	C	T	A	A	G	C	T	A	A	
Sequence A	$s[0,0]$ →	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
	A	-1	2	1	0	-1	-2	-3	-4	-5	-6	-7	-8
	G	-2	1	4	3	2	1	0	-1	-2	-3	-4	-5
	C	-3	0	3	6	5	4	3	2	1	0	-1	-2
	C	-4	-1	2	5	5	4	3	2	4	3	2	1
	T	-5	-2	→	→	7	6	5	4	3	6	5	4
	A	-6	-3	0	3	6	9	8	7	6	5	8	7
	G	-7	-4	-1	2	5	8	7	10	9	8	7	6
	C	-8	-5	-2	1	4	7	6	9	12	11	10	9
	C	-9	-6	-3	0	3	6	5	8	11	11	10	9
	A	-10	-7	-4	-1	2	5	8	7	10	10	13	12
A	-11	-8	-5	-2	1	4	7	6	9	9	12	15	

column



$$S[i,j] = S[i-1,j-1]$$



row

$$S[i,0] = i\gamma$$

$$S[0,j] = j\gamma$$

$$S[i,j] = \max \begin{cases} S[i-1,j-1] + \delta(i-1,j-1) \\ S[i,j-1] + \gamma \\ S[i-1,j] + \gamma \end{cases}$$