

Total Number of Possible Global Alignments between Two Sequences

Zhiping Weng
Boston University



Definitions

Two sequences $\vec{a} = a_1a_2\dots a_m$ and $\vec{b} = b_1b_2\dots b_n$ ($m \geq n$)

$f(m, n)$: The total number of possible alignments between \vec{a} and \vec{b}

$g(m, n)$: The total number of possible **non-redundant** alignments between \vec{a} and \vec{b}

L : The length of an alignment (with padded gaps)



$$n=m=1$$

$$L=1 \quad L=2$$

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}$$

$$\begin{bmatrix} -a_1 \\ b_1 - \end{bmatrix}$$

$$\begin{bmatrix} a_1 - \\ -b_1 \end{bmatrix}$$

$$f(1,1)=3$$

$$g(1,1)=2$$

BE561 9/19/00
Zhiping Weng



$$m=2; n=1$$

$$L=2 \quad L=3$$

$$\begin{bmatrix} a_1 a_2 \\ b_1 - \end{bmatrix}$$

$$\begin{bmatrix} -a_1 a_2 \\ b_1 - - \end{bmatrix}$$

$$\begin{bmatrix} a_1 a_2 \\ -b_1 \end{bmatrix}$$

$$\begin{bmatrix} a_1 a_2 - \\ - - b_1 \end{bmatrix}$$

$$\begin{bmatrix} a_1 - a_2 \\ -b_1 - \end{bmatrix}$$

$$f(1,2)=5$$

$$g(1,2)=3$$

BE561 9/19/00
Zhiping Weng



$$m=2; n=2$$

$L=2$

$$\begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$$

$L=3$

$$\begin{bmatrix} -a_1 & a_2 \\ b_1 & -b_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & -a_2 \\ -b_1 & b_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & -a_2 \\ b_1 & b_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & a_2 \\ b_1 & -b_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & a_2 \\ -b_1 & b_2 \end{bmatrix}$$

$$\begin{bmatrix} -a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$$

$L=4$

$$\begin{bmatrix} a_1 & -a_2 \\ -b_1 & b_2 \end{bmatrix}$$

$$\begin{bmatrix} -a_1 & -a_2 \\ b_1 & -b_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & -a_2 \\ -b_1 & -b_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & a_2 \\ -b_1 & b_2 \end{bmatrix}$$

$$\begin{bmatrix} -a_1 & a_2 \\ b_1 & -b_2 \end{bmatrix}$$

$$\begin{bmatrix} -a_1 & -a_2 \\ b_1 & b_2 \end{bmatrix}$$

$$f(2,2)=13$$

BE561 9/19/00
Zhiping Weng

$$g(2,2)=6$$



$$f(n,m)$$

	n=0	1	2	3	4	5
m=0	1	1	1	1	1	1
1	1	3	5	7	9	11
2	1	5	13	25	41	61
3	1	7	25	63	129	231
4	1	9	41	129	321	681
5	1	11	61	231	681	1683

$$\begin{array}{ccc} f(m-1, n-1) & & f(m, n-1) \\ & \searrow & \downarrow \\ f(m-1, n) & \rightarrow & f(m, n) \end{array}$$

BE561 9/19/00
Zhiping Weng

$$f(m, n) = f(n-1, m) + f(n, m-1) + f(n-1, m-1)$$



f(n,m)

3 ways an alignment can be made:

$$\begin{array}{ccc} \dots a_m & \dots a_m & \dots - \\ \dots b_n & \dots - & \dots b_n \end{array}$$

$$\begin{aligned} f(m,n) &= f(n-1,m) + f(n,m-1) + f(n-1,m-1) \\ f(0,0) &= f(m,0) = f(0,n) = 1 \end{aligned}$$

$$f(m,n) = \sum_{k=0}^n \binom{k+m}{n} \cdot \binom{n}{k}$$

$$\lim_{n \rightarrow \infty} f(n,n) = \sqrt{n}(1+\sqrt{2})^{2n+1}$$

$$f(1000,1000) = 10^{767.4}$$

BE561 9/19/00
Zhiping Weng



g(n,m)

	n=0	1	2	3	4	5
m=0	1	1	1	1	1	1
1	1	2	3	4	5	6
2	1	3	6	10	15	21
3	1	4	10	20	35	56
4	1	5	15	35	70	126
5	1	6	21	56	126	252

$$\begin{array}{ccc} & & g(m,n-1) \\ & & \downarrow \\ g(m-1,n) & \rightarrow & g(m,n) \end{array}$$

$$g(m,n) = g(n-1,m) + g(n,m-1)$$

BE561 9/19/00
Zhiping Weng



$g(n,m)$

$\dots a_m - \dots - a_m$ are the same alignment
 $\dots - b_n \dots b_n -$

$$\begin{aligned} g(m,n) &= g(n-1,m) + g(n,m-1) \\ f(0,0) &= f(m,0) = f(0,n) = 1 \end{aligned}$$

$$g(m,n) = \binom{m+n}{n}$$

$$\lim_{n \rightarrow \infty} g(n,n) = \frac{(2n)!}{(n!)^2} = \frac{\sqrt{2\pi 2n} (2n)^{2n} e^{-2n}}{(\sqrt{2\pi n} (n)^n e^{-n})^2} = \frac{2^{2n}}{\sqrt{n\pi}}$$

$$g(1000,1000) = 10^{600}$$

BE561 9/19/00
Zhiping Weng

Pairwise Sequence Alignment

Local Alignment

Semi-Global Alignment

Zhiping Weng
Boston University



Dynamic Programming End Gaps

End Gaps: gaps appear before the first or after the last nucleotide/amino acid in a sequence

```
CAGCA-CTTGGATTCTCGG
  || | |||
---CAGCGTGG-----
```

BE561 9/19/00
Zhiping Weng



End Gaps vs. Internal Gaps

```
CAGCA-CTTGGATTCTCGG
  || | |||
---CAGCGTGG-----
```

6 matches, 1 mismatch, 1 internal gap, 11 end gaps

```
CAGCACTTGGATTCTCGG
||||      ||  ||
CAGC-----G-T----GG
```

8 matches, 0 mismatch, 10 internal gaps, 0 end gap

BE561 9/19/00
Zhiping Weng



Local Alignment

A local Alignment between sequence s and sequence t is an alignment with maximum similarity between a substring of s and a substring of t .

T. F. Smith & M. S. Waterman, "Identification of Common Molecular Subsequences", J. Mol. Biol., 147:195-197

BE561 9/19/00
Zhiping Weng



Local Alignment

For any entry (i,j) in the dynamic programming matrix, there is always the alignment between the empty suffixes of $s[1..i]$ and $t[1..j]$, which has score zero. Therefore the dynamic programming matrix will have all entries greater than or equal to zero.

BE561 9/19/00
Zhiping Weng



Dynamic Programming Local Alignment

$$S_{ij} = \max \begin{cases} S_{i-1, j-1} + s_{ij} \\ S_{i-1, j} + w \\ S_{i, j-1} + w \\ 0 \end{cases}$$

S_{ij} : score for the alignment ending at i in sequence 1 and j in sequence 2

s_{ij} : score for aligning base/amino acid i in sequence 1 with base/amino acid j in sequence 2

w : the score for making a gap

No gap penalties for the boundaries

The end of the optimal path may be found anywhere in the dynamic

BE561 9/19/00 programming matrix
Zhiping Weng



Dynamic Programming: local Alignment Match = 5, Mismatch = -4, Gap $w = -7$

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

BE561 9/19/00
Zhiping Weng



Dynamic Programming: local Alignment

Match = 5, Mismatch = -4, Gap w = -7

	0	G	C	T	G	G	A	A	G	G	C	A	T
0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	$s_{G,G}$											
C	0												
A	0												
G	0												
A	0												
G	0												
C	0												
A	0												
C	0												
T	0												

$$s_{G,G} = \max \begin{cases} s_{0,0} + s_{G,G} = 0 + 5 = 5 \\ s_{G,0} + w = 0 - 7 = -7 \\ s_{0,G} + w = 0 - 7 = -7 \\ 0 \end{cases} = 5$$

BE561 9/19/00
Zhiping Weng



Dynamic Programming: local Alignment

Match = 5, Mismatch = -4, Gap w = -7

	0	G	C	T	G	G	A	A	G	G	C	A	T
0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	$s_{G,C}$										
C	0												
A	0												
G	0												
A	0												
G	0												
C	0												
A	0												
C	0												
T	0												

$$s_{G,C} = \max \begin{cases} s_{0,G} + s_{G,C} = 0 - 4 = -4 \\ s_{G,G} + w = 5 - 7 = -2 \\ s_{0,C} + w = 0 - 7 = -7 \\ 0 \end{cases} = 0$$

BE561 9/19/00
Zhiping Weng



Dynamic Programming: local Alignment

Match = 5, Mismatch = -4, Gap w= -7

	0	G	C	T	G	G	A	A	G	G	C	A	T
0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	0										
C	0	0	$S_{C,C}$										
A	0												
G	0												
A	0												
G	0												
C	0												
A	0												
C	0												
T	0												

$$S_{C,C} = \max \left\{ \begin{array}{l} S_{G,G} + S_{C,C} = 5 + 5 = 10 \\ S_{C,G} + w = 0 - 7 = -7 \\ S_{G,C} + w = 0 - 7 = -7 \\ 0 \end{array} \right\} = 10$$

BE561 9/19/00
Zhiping Weng



Dynamic Programming: local Alignment

Match = 5, Mismatch = -4, Gap w= -7

	0	G	C	T	G	G	A	A	G	G	C	A	T
0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	0	0	5	5	0	0	5	5	0	0	0
C	0	0	10	3	0	1	1	0	0	1	10	3	0
A	0	0	3	6	0	0	6	6	0	0	3	15	8
G	0	5	0	0	11	5	0	2	11	5	0	8	11
A	0	0	1	0	4	7	10	5	4	7	1	5	4
G	0	5	0	0	5	9	3	6	10	9	3	0	1
C	0	0	10	3	0	2	5	0	3	6	14	7	0
A	0	0	3	6	0	0	7	10	3	0	7	19	12
C	0	0	5	0	2	0	0	3	6	0	5	12	15
T	0	0	0	10	3	0	0	0	0	2	0	5	17

BE561 9/19/00
Zhiping Weng



Dynamic Programming: local Alignment

Match = 5, Mismatch = -4, Gap w= -7

	0	G	C	T	G	G	A	A	G	G	C	A	T
0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	0	0	5	5	0	0	5	5	0	0	0
C	0	0	10	3	0	1	1	0	0	1	10	3	0
A	0	0	3	6	0	0	6	0	0	3	15	8	0
G	0	5	0	0	11	5	0	2	11	5	0	8	11
A	0	0	1	0	4	7	10	5	4	7	1	5	4
G	0	5	0	0	5	9	3	6	10	9	3	0	1
C	0	0	10	3	0	2	5	0	3	6	14	7	0
A	0	0	3	6	0	0	7	10	3	0	7	19	12
C	0	0	5	0	2	0	0	3	6	0	5	12	15
T	0	0	0	10	3	0	0	0	0	2	0	5	17

BE561 9/19/00
Zhiping Weng



Dynamic Programming: local Alignment

Match = 5, Mismatch = -4, Gap w= -7

G	A	A	G	-	G	C	A
G	C	A	G	A	G	C	A

6 matches: $6 * 5 = 30$

1 mismatch: -4

1 gaps: $1 * -7 = -7$

Total: 19

BE561 9/19/00
Zhiping Weng



Dynamic Programming: local Alignment

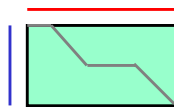
Match = 5, Mismatch = -4, Gap w= -7

	O	G	C	T	G	G	A	A	G	G	C	A	T
O	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	0	0	5	0*	0	0	5	5	0	0	0
C	0	0	10	3	0	1	0*	0	0	1	10	3	0
A	0	0	3	6	0	0	6	0*	0	0	3	15	8
G	0	5	0	0	11	5	0	2	0*	5	0	8	11
A	0	0	1	0	4	7	10	5	0*	0*	1	5	4
G	0	5	0	0	5	9	3	6	10	0*	0*	0	1
C	0	0	10	3	0	2	5	0	3	6	0*	0*	0
A	0	0	3	6	0	0	7	10	3	0	2*	0*	0*
C	0	0	5	0	2	0	0	3	6	0	5	0*	0*
T	0	0	0	10	3	0	0	0	0	2	0	1*	5*

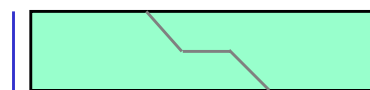
BE561 9/19/00
Zhiping Weng



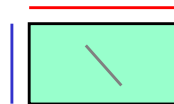
Four Basic Forms of the Optimal Pairwise Alignment by Dynamic Programming



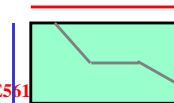
Global, Needleman-Wunch



Best in long
Don't penalize unmatched long ends.



Best Local optimal, Smith-Waterman



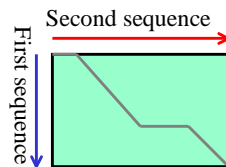
Optimal overlap,
No cost non-overlap

BE561
Zhiping Weng



Dynamic Programming End Gap Variations

Where gaps are not charged for	Action
Beginning of first sequence	Initialize first row with zeros
End of first sequence	Look for maximum in last row
Beginning of second sequence	Initialize first column with zeros
End of second sequence	Look for maximum in last column



BE561 9/19/00
Zhiping Weng