

**Problem Set 1****Due Thursday, September 17th**

Collaboration is allowed on this homework. You must hand in homeworks individually and list the names of the people you worked with.

You may *not* use a program to do this homework. Turn in your handwritten answers on the attached sheets. Extra alignment templates will be available on the website.

## 1. Pairwise alignment:

- (a) Compute the local alignment of “SUBWAYTRAIN” with “TRANSUBSTANTIATION”, using the following scoring system: matches = 1, mismatches = -1, indels = -1. Show your alignment matrix with scores and traceback on the attached alignment template.
- (b) What is the score of the optimal local alignment? How many different optimal alignments are there? Show them.

*There are three optimal local alignments with a score of 3:*

TRA  
TRA

TRAIN  
TRA-N

*and*

SUB  
SUB

Local pairwise alignment

(a)

$s(i,i)$  1  
 $s(i,j)$  -1  
 $d$  -1

	O	T	R	A	N	S	U	B	S	T	A	N	T	I	A	T	I	O	N
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
U	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	1	3	2	1	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	2	2	1	0	0	0	0	0	0	0	0	0
A	0	0	0	1	0	0	0	1	1	1	2	1	0	0	1	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
T	0	1	0	0	0	0	0	0	0	1	0	0	2	1	0	1	0	0	0
R	0	0	2	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
A	0	0	1	3	2	1	0	0	0	0	1	0	0	0	2	1	0	0	0
I	0	0	0	2	2	1	0	0	0	0	0	0	0	1	1	1	2	1	0
N	0	0	0	1	3	2	1	0	0	0	0	1	0	0	0	0	1	1	2

b.

T	R	A	-	N		S	U	B	
1	1	1	-1	1	3	1	1	1	3
T	R	A	I	N		S	U	B	

  

T	R	A	
1	1	1	3
T	R	A	

## 2. Semiglobal alignment:

- (a) Using the same scoring function, compute the semiglobal alignment of “SUBWAY-TRAIN” with “TRANSUBSTANTIATION”, with no penalty for aligning indels with the leading characters of “TRANSUBSTANTIATION”. (That is, indels can be inserted at the beginning of “SUBWAYTRAIN” with no cost.)

Show your alignment matrix with scores and traceback on the template provided. To reduce the amount of work you have to do, low scoring areas of the matrix have been blacked out on the template. You do not need to fill in those cells.

- (b) What is the score of the optimal semiglobal alignment? How many different optimal alignments are there? Show them.

*There are 2 alignments with optimal score of 2.*

```
TRANSUBSTANTIATION
-----SUBW-AYTRA-I-N
```

*and*

```
TRANSUBSTANTIATION
-----SUB-WAYTRA-I-N
```

## Semiglobal alignments

$s(i,i)$  1  
 $s(i,j)$  -1  
 $d$  -1

	<b>O</b>	<b>T</b>	<b>R</b>	<b>A</b>	<b>N</b>	<b>S</b>	<b>U</b>	<b>B</b>	<b>S</b>	<b>T</b>	<b>A</b>	<b>N</b>	<b>T</b>	<b>I</b>	<b>A</b>	<b>T</b>	<b>I</b>	<b>O</b>	<b>N</b>	
<b>O</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<b>S</b>	-1	-1	-1	-1	-1	1	0	-1	1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
<b>U</b>	-2	-2	-2	-2	-2	0	2	1	0	0	-1	-2	-2	-2	-2	-2	-2	-2	-2	-2
<b>B</b>	-3	-3	-3	-3	-3	-1	1	3	2	1	0	-1	-2	-3	-3	-3	-3	-3	-3	-3
<b>W</b>	-4	-4	-4	-4	-4	-2	0	2	2	1	0	-1	-2	-3	-4	-4	-4	-4	-4	-4
<b>A</b>	-5	-5	-5	-3	-4	-3	-1	1	1	1	2	1	0	-1	-2	-3	-4	-5	-5	-5
<b>Y</b>	-6	-6	-6	-4	-4	-4	-2	0	0	0	1	1	0	-1	-2	-3	-4	-5	-6	-6
<b>T</b>	-7	-5	-6	-5	-5	-5	-3	-1	-1	1	0	0	2	1	0	-1	-2	-3	-4	-4
<b>R</b>	-8	-6	-4	-5	-6	-6	-4	-2	-2	0	0	-1	1	1	0	-1	-2	-3	-4	-4
<b>A</b>	-9	-7	-5	-3	-4	-5	-5	-3	-3	-1	1	0	0	0	2	1	0	-1	-2	-2
<b>I</b>	-10	-8	-6	-4	-4	-5	-6	-4	-4	-2	0	0	-1	1	1	1	2	1	0	0
<b>N</b>	-11	-9	-7	-5	-3	-4	-5	-5	-5	-3	-1	1	0	0	0	0	1	1	2	2

<b>T</b>	<b>R</b>	<b>A</b>	<b>N</b>	<b>S</b>	<b>U</b>	<b>B</b>	<b>S</b>	<b>T</b>	<b>A</b>	<b>N</b>	<b>T</b>	<b>I</b>	<b>A</b>	<b>T</b>	<b>I</b>	<b>O</b>	<b>N</b>			
				1	1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1			2
				<b>S</b>	<b>U</b>	<b>B</b>	<b>_</b>	<b>W</b>	<b>A</b>	<b>Y</b>	<b>T</b>	<b>R</b>	<b>A</b>	<b>_</b>	<b>I</b>	<b>_</b>	<b>N</b>			

<b>T</b>	<b>R</b>	<b>A</b>	<b>N</b>	<b>S</b>	<b>U</b>	<b>B</b>	<b>S</b>	<b>T</b>	<b>A</b>	<b>N</b>	<b>T</b>	<b>I</b>	<b>A</b>	<b>T</b>	<b>I</b>	<b>O</b>	<b>N</b>			
				1	1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1			2
				<b>S</b>	<b>U</b>	<b>B</b>	<b>W</b>	<b>_</b>	<b>A</b>	<b>Y</b>	<b>T</b>	<b>R</b>	<b>A</b>	<b>_</b>	<b>I</b>	<b>_</b>	<b>N</b>			

### 3. Comparing semiglobal and local alignments

- (a) There is a set of scoring functions for which optimal local pairwise alignment of “SUBWAYTRAIN” and “TRANSUBSTANTIATION” yields the same alignment as the one you obtained using semiglobal alignment in Problem 2 (disregarding the leading overhang.) Give one such scoring function. Your alignment won’t necessarily have the same score as in Problem 2.

*Any scoring function that ensures that the score remains greater than 0 between “SUB” and “TRAIN” will give the same alignment as Problem 2. There are many. Here is one: matches = 3, mismatches = -1, and indels = -1.*

- (b) The alignment in Problem 2 contains one of the optimal local alignments found in Problem 1 but not the others. Is there any scoring function that could result in a semiglobal alignment that incorporates two non-overlapping local alignments found in Problem 1? If so, give it. If not, explain why not.

*No. One optimal local alignment aligns the beginning of “SUBWAYTRAIN” with the middle of “TRANSUBSTANTIATION”. The second local alignment aligns the end of “SUBWAYTRAIN” with the beginning of “TRANSUBSTANTIATION”. Because they are not in the same order, it is not possible to include both local alignments in a single global alignment.*



## 4. Multiple sequence alignment:

For this problem, use the sum-of-pairs metric and follow the “once a gap, always a gap” rule. Consider the following three sequences.

- (1) ACGTC
- (2) TCAT
- (3) ACGTCAT

- (a) Compute all three optimal pairwise alignments assuming a similarity scoring function where  $p(x, x) = 0$  and  $p(x, y) = -3, x \neq y$ . The score for an indel is  $g = -2$ . Give the score and the alignment for each pair of sequences.

```
ACGTC
TCAT-
-8
```

```
ACGTCAT
---TCAT
-6
```

```
ACGTC--
ACGTCAT
-4
```

- (b) Each pairwise global alignment in (a) yields a sequence over the 24 letter alphabet  $\Sigma^* \times \Sigma^* - \{(-)\}$  (i.e.,  $\{AA, AC, AG, AT, CA, \dots\}$ ). Durbin calls this a profile. We can obtain a multiple alignment by aligning a profile  $s[1, \dots, m]$  with a sequence  $t[1, \dots, n]$ , where elements in  $s[i]$  are of the form  $xy$ ,  $x-$  or  $-y$ , and elements in  $t[i]$  are of the form  $z$ , where  $x, y, z \in \Sigma$ . The score for aligning  $s[i]$  with  $t[j]$  is the sum of the scores for aligning the first character in  $s[i]$  with  $t[j]$  and the second character in  $s[i]$  with  $t[j]$ .

Write down the recurrence relation for calculating the alignment matrix  $a[i, j]$ , for the case where  $s[i]$  is of the form  $xy$ , in terms of  $p(\cdot, \cdot)$ ,  $g$ ,  $x$ ,  $y$  and  $t[\cdot]$ .

$$a[i, j] = \max \left\{ \begin{array}{l} a[i-1, j-1] + p(x, t[j]) + p(y, t[j]), \\ a[i, j-1] + 2g, \\ a[i-1, j] + 2g \end{array} \right\}$$

Write down the recurrence relation for calculating  $a[i, j]$ , for the case where  $s[i]$  is of the form  $x-$  or  $-y$ .

$$a[i, j] = \max \left\{ \begin{array}{l} a[i-1, j-1] + p(x, t[j]) + g, \\ a[i, j-1] + 2g, \\ a[i-1, j] + g \end{array} \right\}$$

- (c) Using the recurrence relations from (b), align the profile with the best score from (a) against the remaining sequence. Show your alignment matrix with scores and traceback on the attached template.

Your traceback will give you a pairwise alignment between a profile and a sequence which is equivalent to a multiple alignment of three sequences. Write down the multiple alignment and compute the sum-of-pairs MSA score.

```
---TCAT
ACGTC--
ACGTCAT
```

*The sum of pairs score is -20.*

Does this give you the same score as the entry in the lower righthand corner of your alignment matrix? If not, why not?

*The profile-sequence alignment score is -16.*

*No, Sum of Pairs does not account for matches or mismatches between the sequences in the profile.*

Repeat problem (c), but this time align the profile with the *worst* score from (a) against the remaining sequence. Write down the multiple alignment and compute the sum-of-pairs MSA score.

beginquotation

```
TCAT---  
ACGTC--  
ACGTCAT
```

*The sum of pairs score is -24.*

Compare the MSA and MSA score you obtained from (c) and (d). Do they agree? If not, why do they differ? What does this tell you about progressive multiple alignment methods?

*The progressive alignment heuristic can lock the gaps into the wrong place early. This is why we typically get better alignments with this heuristic if we align the most closely related sequences first.*

	0	A	C	G	T	C
0	<b>0</b>	-2	-4	-6	-8	-10
T	-2	<b>-3</b>	-5	-7	-6	-8
C	-4	-5	<b>-3</b>	-5	-7	-9
A	-6	-4	-5	<b>-6</b>	-8	-10
T	-8	-6	-7	-8	<b>-6</b>	<b>-8</b>

	0	A	C	G	T	C
0	0	-2	-4	-6	-8	-10
A	-2	<b>0</b>	-2	-4	-6	-8
C	-4	-2	<b>0</b>	-2	-4	-6
G	-6	-4	-2	<b>0</b>	-2	-4
T	-8	-6	-4	-2	<b>0</b>	-2
C	-10	-8	-6	-4	-2	<b>0</b>
A	-12	-10	-8	-6	-4	<b>-2</b>
T	-14	-12	-10	-8	-6	<b>-4</b>

	0	T	C	A	T
0	0	-2	-4	-6	-8
A	<b>-2</b>	-3	-5	-4	-6
C	<b>-4</b>	-5	-3	-5	-6
G	<b>-6</b>	-7	-5	-6	-8
T	-8	<b>-6</b>	-7	-8	-6
C	-10	-8	<b>-6</b>	-8	-8
A	-12	-10	-8	<b>-6</b>	-8
T	-14	-12	-10	-8	<b>-6</b>

5. Give the recurrence relation for a dynamic programming algorithm for aligning three sequences,  $r[1, \dots, l]$ ,  $s[1, \dots, m]$  and  $t[1, \dots, n]$ , using sum-of-pairs and edit distance as a scoring function.

$$a[i, j] = \min \left\{ \begin{array}{l} a[i-1, j-1, k-1] + \delta(r[i], s[j]) + \delta(s[j], t[k]) + \delta(r[i], t[k]) \\ a[i-1, j, k-1] + \delta(r[i], t[k]) + 2, \\ a[i, j-1, k-1] + \delta(s[j], t[k]) + 2, \\ a[i-1, j-1, k] + \delta(r[i], s[j]) + 2, \\ a[i, j, k-1] + 2, \\ a[i, j-1, k] + 2, \\ a[i-1, j, k] + 2, \end{array} \right\}$$

where  $\delta(x, y) = 1$  if  $x \neq y$  and zero otherwise.