

BCB 567/CprE 548
Fall 2007
Homework 5
Solutions

1. The maximum possible SP score occurs when all sequences are the same. In this case the total score is $n\binom{k}{2}\alpha$. This is because each column is perfectly matched:

```
AAAA
AAAA
....
AAAA
```

The worst possible SP score occurs when each string consists of a single unique character from Σ . In this case, the SP score becomes one of two equations. If all sequences are aligning (for a MSA length of n), the score is:

$$n\binom{k}{2}\beta = \frac{nk(k-1)\beta}{2}$$

```
AAAA
BBBB
....
KKKK
```

If all sequences are aligned with gaps (total MSA length of nk), the score is:

$$nk(k-1)\gamma = \frac{nk(k-1)2\gamma}{2} = n\binom{k}{2}2\gamma$$

```
AAAA----...----
----BBBB...----
...
-----...KKKK
```

Either of these cases could have lower total score depending on whether or not $\beta < 2\gamma$.

2. Example: XYX , YXY , and YXX . $\alpha = 1, \gamma = -2, \beta = -1$ In this case, the multiple alignment:

XYX
 YXY
 YXX

is optimal with score of -3. However, obviously the pairwise alignment

XYX
 YXY

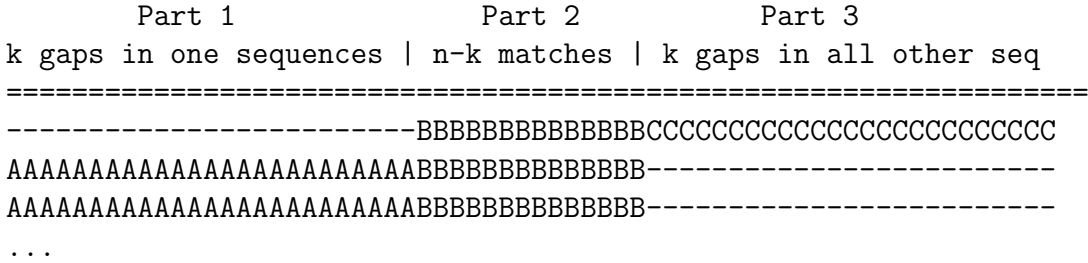
with a score of -3 is not optimal.

$XYX-$
 $-YXY$

with a score of -2 is optimal.

3. The first thing we should do is generalize the k-band to d dimensions. Recall that we defined the k-band to be all cells such that $|i-j| < k$. For d demensions a cell's coordinates can be written $[i_1, i_2, \dots, i_d]$. We will say that the k-band consists of all cells such that $|i_j - i_h| < k$ for all pairs of coordinates.

To derive a stopping condition for the k-band algorithm, we must find an equation for the highest scoring theoretical alignment that travels outside of the k-band, or *best-outside-k*. We see that a highest scoring such alignment will be of the form:



The score for Part 1 is:

$$k(d-1)\gamma + k \binom{d-1}{2} \alpha$$

The score for Part 2 is:

$$\binom{d}{2} (n-k) \alpha$$

The score for Part 3 is:

$$k(d-1)\gamma$$

Therefore, the total score of the best alignment outside of the k-band is:

$$2k(d-1)\gamma + \left(k \binom{d-1}{2} + (n-k) \binom{d}{2} \right) \alpha$$

(Notice that if we set $d = 2$, we end up with the equation $2k\gamma + (n-k)\alpha$, which is the score of the *best-outside-k* for two sequences, so our generalization holds up.)

The rest of the k-band algorithm proceeds as normal. Notice that because the second term grows much more quickly than the first, the k-band is much less useful for many sequences than it was for 2.

4. We will show that each character from S_i has exactly one possible position in the multiple alignment, independent of the order in which we insert S_i .

Consider the character $S_i[j]$. If $S_i[j]$ is aligned with $S_c[k]$, then $S_i[j]$ must appear in the same column as $S_c[k]$ in the multiple alignment, otherwise the induced alignment between S_i and S_c would not be optimal. If $S_i[j]$ is aligned with a gap in the optimal alignment between S_i and S_c , then $S_i[j]$ will appear in the leftmost possible column of the multiple sequence alignment, under the condition that $S_i[j]$ be aligned with a gap in S_c in the multiple alignment.

This is because when we insert the characters from S_i into the multiple sequence alignment, we do so in a left to right fashion, selecting the first valid column. Thus, if gaps in S_c are long enough, we will insert characters in the left most positions of the gap. If the gaps are not long enough we will extend them to the right.

Therefore there is exactly one configuration for each sequence in the multiple sequence alignment. See below.

Pairwise alignment:

S_C: CCCC--CCC-CCCC-----CCCCCCC---CCCC-C

S_i: CC--CAACCCACCCCAAAAAACCCCCCAAACC-CAC <-- the only possible configuration

In the multiple sequence alignment:

S_C: CCCC----CCC---CCCC-----CCCC---CCC-----CCCC-C

S_i: CC--CAA--CCCA--CCCCAAAAACCCC---CCCAAA---CC-CAC

Not Possible:

S_i: CCCC-AA-CCC

^ This gap will never occur, no matter the order in which the sequence is inserted

5. Let σ be the size of the alphabet (number of rows in the profile).

We will use dynamic programming to find the alignment between the sequence and the profile. Create an $n \times m$ table \mathcal{S} . We will choose the simplest interpretation of this problem, where gaps are scored exactly γ . We then fill the table \mathcal{S} with the following equations:

$$\begin{aligned}\mathcal{S}[0, 0] &= 0 \\ \mathcal{S}[i, 0] &= i\gamma \\ \mathcal{S}[0, j] &= j\gamma \\ \mathcal{S}[i, j] &= \max \begin{cases} \mathcal{S}[i-1, j] + \gamma \\ \mathcal{S}[i, j-1] + \gamma \\ \mathcal{S}[i-1, j-1] + \delta(T_i, j) \end{cases} \\ \delta(a, j) &= \sum_b w_{ab}P(b, j)\end{aligned}$$

The optimal alignment score is found in the bottom right cell, and using traceback (as for pairwise sequence alignment) we can construct the alignment.

The runtime of this algorithm is $O(\sigma)$ per cell, for a total runtime of $O(nm\sigma)$. The space requirement is $O(m(\sigma + n))$.