

Bioinformatics 1, 2011 -- Homework 5

Align two profiles using dynamic programming

due thursday, Nov 10

Follow these steps to align two profiles (see Homework 4), using local dynamic programming (see Homework 2). Use a novel, position specific gap penalty.

Summary:

- (1) Read two profiles and two MSAs.
- (2) Calculate the scoring matrix and position-specific gap penalty.
- (3) Sum the alignment matrix.
- (4) Trace back to get the optimal alignment.
- (5) Output the aligned MSAs as one MSA. (extra credit for block output)
- (6) Compare your alignment to UGENE.

===== Detailed instructions =====

(1) **Read data files**, two MSAs in FASTA format and two profiles. Save two sets of sequences, including gap characters. Each MSA must have a corresponding profile in the following format.

i c_{li} $P(A)$ $P(C)$ $P(D)$ $P(E)$ $P(F)$ $P(G)$ $P(H)$ $P(I)$ $P(K)$ $P(L)$ $P(M)$ $P(N)$ $P(P)$ $P(Q)$ $P(R)$ $P(S)$ $P(T)$
 $P(V)$ $P(W)$ $P(Y)$ $P(gap_i)$

Where i is the MSA position, and c_{li} is the character for sequence 1 (for reference only). $P(A)$ etc is the probability of alanine, etc., and $P(gap_i)$ is the probability of a gap at i .

(2a) **Calculate the scoring matrix**. For each pair of positions ij , use the probability-weighted BLOSUM score. $S(i,j) = \sum_{aa_i=1,20} (\sum_{aa_j=1,20} (P(aa_i) * P(aa_j) * B(aa_i,aa_j)))$.

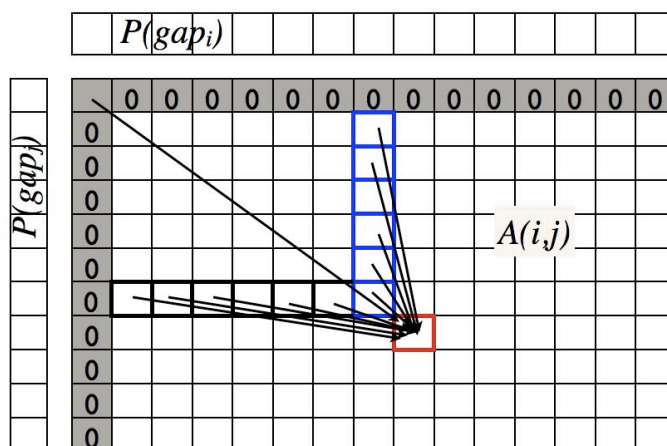
(2b) **Calculate the position-specific gap penalties**, $psgap$. For each i position in sequence 1, $psgap1(i) = gop * exp(-10 * P(gap_i))$. Do the same for each position i in sequence 2. Make gop a variable. Now you have two arrays of gap penalties, $psgap1$ and $psgap2$.

(3a) **Initialize the alignment matrix and traceback arrows**. Gap rows $A(0,j)$ and $A(i,0)$ should be set to zero for local DP alignment. Traceback values for gaps rows are all $(0,0)$, an arrow to the beginning of the alignment. (NOTE: The traceback "arrow" is an ij pair, **not** *down, across, diagonal* as it was in HW2.)

(3b) Calculate the rest of the alignment matrix A using local affine DP, as follows. To calculate $A(i,j)$, take the MAX over values in col 2 (for all values in column 1). Use the corresponding Traceback in col 2.

	$A(i,j)=\text{MAX}(\dots)$	Traceback(i,j)=
	$A(i-1,j-1) + S(i,j)$	(i-1,j-1)
for k=2,i-1	$A(i-k,j-1) + S(i,j) - \sum psgap1(g),g=i-k+1,i-1$	(i-k,j-1)
for k=2,j-1	$A(i-1,j-k) + S(i,j) - \sum psgap2(g),g=j-k+1,j-1$	(i-1,j-k)
	$S(i,j)$	(0,0) Beginning of alignment.

Note, this is a novel position-specific gap penalty. We'll see if it works! The following figure might help. It shows the arrows over which you are taking the maximum.



(4a) Find the maximum value $A(i,j)$ over the entire matrix. **This is the alignment score. Write it out.**

(4b) **Trace back** from the maximum to get the alignment.

Let $\text{Arrow}(1) = (i,j)$, the position of the maximum. [$\text{Arrow}()$ is an ordered pair. $\text{Arrow}(u)(1)$ refers to the first value of the pair, i , which is the position in sequence 1. $\text{Arrow}(u)(2)$ is the position j in sequence 2.] Let $(m,n) = \text{Traceback}(i,j)$. Let $\text{Arrow}(2) = (m,n)$. Let $i=m, j=n$. Repeat (Let $(m,n) = \text{Traceback}(i,j)$; Let $\text{Arrow}(u) = (m,n)$; Let $i=m, j=n$) until $i=0$ and $j=0$. Then reverse the order of $\text{Arrow}()$.

(5) **Output** the aligned MSAs as one MSA. For extra credit, use *block output*, 50 characters per block. Calculate the number of blocks n_{block} using the length of the local alignment starting from the first match and ending with the last match. These instructions will print only the local alignment, not overhanging sequences and end gaps.

```

For each block  $b=1, n_{\text{block}}$ ;
  For each MSA  $m=1,2$ ;
    For each sequence  $s$  in MSA  $m$ ;
      write label for sequence  $s$ ,
      initialize  $c = \text{Arrow}(1)(m)$ ,
      initialize  $x=0$ ,
      if ( $b==1$ ) write the character of sequence  $s$  at position  $c$ ,

```

```

For  $u=2, \text{length}(\text{Arrow})$ ;
  If ( $m==1$ );
    if ( $\text{Arrow}(u)(2) - \text{Arrow}(u-1)(2) > 1$ ); !! size of insertion in MSA1
      For  $1, \text{Arrow}(u)(2) - \text{Arrow}(u-1)(2) - 1$ ;
        increment the output character counter  $x$ ,
        if  $x$  is in block  $b$  write a gap character,
        while ( $c < \text{Arrow}(u)(1)$ )
          increment  $c$  and  $x$ ,
          if  $x$  is in block  $b$  write the character of sequence  $s$  at position  $c$ ,
        elseif ( $m==2$ );
          if ( $\text{Arrow}(u)(1) - \text{Arrow}(u-1)(1) > 1$ ); !! size of insertion in MSA2
            For  $1, \text{Arrow}(u)(1) - \text{Arrow}(u-1)(1) - 1$ ;
              increment the output character counter  $x$ ,
              if  $x$  is in block  $b$  write a gap character,
              while ( $c < \text{Arrow}(u)(2)$ )
                increment  $c$  and  $x$ ,
                if  $x$  is in block  $b$  write the character of sequence  $s$  at position  $c$ .

```

If not doing the extra credit, just ignore all references to blocks and b .

(6) Read the same two MSAs into UGENE and align them using MUSCLE profile option. Compare your program to UGENE. Where are there differences, if any? UGENE does not use position specific gap penalty. How is the new gap penalty improving (or ruining!) the alignment?

Please upload the program and the results for the data provided, using the course upload page.