Implementation of the Needleman–Wunsch algorithm in R

Author: Kyrylo Bessonov  
kbessonov@ulg.ac.be  
+32 4366 9544 (office phone)  
Office number: 1/16 (section D)  
Institute Montefiore (Building B37)  
http://www.montefiore.ulg.ac.be/~kbessonov/  

Help: Please contact me preferential via email or via my office phone. I would be glad to help you out  

Due date: 30th of October 2012 (at the end of the Tue class)  

Instructions  

Global alignment is an essential topic in Bioinformatics. This computer based assignment is designed to allow you to gain in-depth understanding of the algorithm and also allow you perfect your programming skills in R. This assignment will be a significant effort for person that just started his/her programming career. Therefore please choose wisely.  

Introduction  

In this assignment you will implement global alignment and test it. You will be provided with the pseudo-code representing inner logic of the algorithm. In essence pseudo-code is an abstract programming code that describes particular logic/algorithm. Thus pseudo-code is general enough to be implemented in any language. You are free to use any of your favorite programming languages, but R is suggested if you do not have any experience in other languages.  

You can either implement the algorithm for DNA/RNA or protein sequences. In either case show the scoring matrix used to calculate the scores.  

Your submission should include  

- a brief description showing that you implementation works. Please provide couple of example sequences and the generated alignment with the total score and the corresponding alignment matrix F  
- the source code (for testing purposes)  
- Short description of scoring matrix that you had used  

Algorithm logic and pseudo code  

The algorithm consists of two major steps: 1) generate of the alignment F-matrix 2) The trace-back calls generating final alignment sequence.  

The pseudo code was extracted from Wikipedia and is of excellent quality (source: http://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm)
A) First part of the pseudo-code concerns construction of the alignment F-matrix of sequence A and B

```
for i=0 to length(A)
  F(i,0) ← d*i
for j=0 to length(B)
  F(0,j) ← d*j
for i=1 to length(A){
  for j=1 to length(B){
    Match ← F(i-1,j-1) + S(Ai, Bj)
    Delete ← F(i-1, j) + d
    Insert ← F(i, j-1) + d
    F(i,j) ← max(Match, Insert, Delete)
  }
}
```

B) The second part of the pseudo-code describes the trace-back logic that allow generation of the final global alignment

```
AlignmentA ← ""
AlignmentB ← ""
i ← length(A)
j ← length(B)
while (i > 0 and j > 0)
{
  Score ← F(i,j)
  ScoreDiag ← F(i - 1, j - 1)
  ScoreUp ← F(i, j - 1)
  ScoreLeft ← F(i - 1, j)
  if (Score == ScoreDiag + S(Ai, Bj))
  {
    AlignmentA ← Ai + AlignmentA
    AlignmentB ← Bj + AlignmentB
    i ← i - 1
    j ← j - 1
  }
  else if (Score == ScoreLeft + d)
  {
    AlignmentA ← Ai + AlignmentA
    AlignmentB ← "-" + AlignmentB
    i ← i - 1
  }
  otherwise (Score == ScoreUp + d)
  {
    AlignmentA ← "-" + AlignmentA
    AlignmentB ← Bj + AlignmentB
    j ← j - 1
  }
}
while (i > 0)
{
  AlignmentA ← Ai + AlignmentA
  AlignmentB ← "-" + AlignmentB
  i ← i - 1
}
while (j > 0)
{
  AlignmentA ← "-" + AlignmentA
  AlignmentB ← Bj + AlignmentB
  j ← j - 1
}