Data Structures and Algorithms in Java

Assignment 2 (Global Sequence Alignment) Discussion

Introduction

Goal: find an optimal alignment for two DNA sequences x and y

We are permitted to insert gaps in either sequence to make them have the same length

We pay a penalty for each gap that we insert and also for each pair of characters that mismatch

Operation	Cost
Insert a gap	2
Align two characters that do not match	1
Align two characters that do match	0

Introduction

Edit distance is the cost of the best possible alignment between the two genetic sequences over all possible alignments

Two possible alignments of the sequences `x = "AACAGTTACC"` and `y = "TAAGGTCA"`

x	y	cost
A	T	1
A	A	0
C	A	1
A	G	1
G	G	0
T	T	0
T	C	1
A	A	0
C	-	2
C	-	2
		8

x	y	cost
A	T	1
A	A	0
C	-	2
A	A	0
G	G	0
T	G	1
T	T	0
A	-	2
C	C	0
C	A	1
		7

Edit distance for the two sequences is 7

Notation

`m` and `n` denote the lengths of `x` and `y`, respectively

`x[i]` denotes the `i`th character of the sequence `x`

`x[i..m]` denotes the suffix of `x` consisting of the characters `x[i], x[i + 1], ..., x[m - 1]`

`opt` is the `(m + 1) x (n + 1)` edit-distance matrix

`opt[i][j]` denotes the edit distance of `x[i..m]` and `y[j..n]`

Example: if `x = "AACAGTTACC"` and `y = "TAAGGTCA"`, then

- `m = 10` and `n = 8`
- `x[2]` is `'C'`
- `x[5..m]` is `"CAGTTACC"` and `y[8..n]` is `""`
- `opt` is a `11 x 9` matrix
- `opt[0][0]` is the edit distance of `x` and `y`

Recursive Solution

Case 1 (`x[i]` is matched with `y[j]`): `opt[i][j] = opt[i + 1][j + 1] + 0 or 1` depending on whether `x[i]` equals `y[j]`

Case 2 (`x[i]` is matched with a gap): `opt[i][j] = opt[i + 1][j] + 2`

Case 3 (`y[j]` is matched with a gap): `opt[i][j] = opt[i][j + 1] + 2`

We compute `opt[i][j]` by taking the minimum of the three quantities

```
opt[i][j] = min(opt[i + 1][j + 1] + 0 or 1, opt[i + 1][j] + 2, opt[i][j + 1] + 2)
```

Direct computation of this recursive scheme is spectacularly inefficient

We use dynamic programming

Key idea: break up a large problem into smaller subproblems, store the answers to those smaller subproblems, and use the stored answers to solve the original problem

Problem 1 (Compute Edit Distance)

Write a program called `EditDistance.java` that receives strings `x` and `y` as standard input; computes the edit-distance matrix `opt`; and outputs `x`, `y`, the dimensions of `opt`, and `opt`

```
× ~/workspace/global_sequence_alignment
$ javac -d out src/EditDistance.java
$ java EditDistance < data/example10.txt
AACAGTTACC
TAAGGTCA
11 9
  7   8  10  12  13  15  16  18  20
  6   6   8  10  11  13  14  16  18
  6   5   6   8   9  11  12  14  16
  7   5   4   6   7   9  11  12  14
  9   7   5   4   5   7   9  10  12
  8   8   6   4   4   5   7   8  10
  9   8   7   5   3   3   5   6   8
 11   9   7   6   4   2   3   4   6
 13  11   9   7   5   3   1   3   4
 14  12  10   8   6   4   2   1   2
 16  14  12  10   8   6   4   2   0
```

Problem 1 (Compute Edit Distance)

- Read sequences `x` (String) and `y` (String) from standard input
- Set `m` (int) and `n` (int) to the lenghts of `x` and `y`, respectively (use `GSA.length()`)
- Create an `(m + 1) x (n + 1)` array `opt` of ints
- Initialize the rightmost column of opt to `2(m - i)`, where `0 <= i <= m`
- Initialize the bottommost row of opt to `2(n - j)`, where `0 <= j <= n`

Problem 1 (Compute Edit Distance)

Fill in the rest of `opt`, starting at `opt[m - 1][n - 1]` and ending at `opt[0][0]`, as follows (use `GSA.charAt()` and `GSA.min()` where needed)

- If `x[i] = y[j]` then `opt[i][j] = min(opt[i + 1][j + 1], opt[i + 1][j] + 2, opt[i][j + 1] + 2)`
- Otherwise, `opt[i][j] = min(opt[i + 1][j + 1] + 1, opt[i + 1][j] + 2, opt[i][j + 1] + 2)`

Write the following output, each starting on a new line

- `x`
- `y`
- `m` and `n` separated by a space
- `opt` using the format string `"%3d "` for elements not in the last column, and `"%3d\n"` for the last-column elements

Problem 2 (Recover Alignment)

Write a program `Alignment.java` that receives as standard input the output produced by `EditDistance.java`; recovers an optimal alignment between `x` and `y`; and writes the edit distance and the alignment

```
× ~/workspace/global_sequence_alignment
$ javac -d out src/Alignment.java
$ java EditDistance < data/example10.txt | java Alignment
7
A T 1
A A 0
C - 2
A A 0
G G 0
T G 1
T T 0
A - 2
C C 0
C A 1
```

Problem 2 (Recover Alignment)

Read sequences `x` (String) and `y` (String) from standard input

Set `m` (int) and `n` (int) to the lenghts of `x` and `y`, respectively

Read the edit-distance matrix `opt` from standard input (use `StdArrayIO.readInt2D()`)

Write the edit distance between `x` and `y`, ie, the value of `opt[0][0]`

Problem 2 (Recover Alignment)

Set ints `i` and `j` both to 0

Recover and output the optimal alginment, starting at `opt[0][0]` and ending at `opt[m - 1][n - 1]`, as follows

- If `opt[i][j] = opt[i + 1][j] + 2`, then align `x[i]` with a gap and penalty of 2, and increment `i`
- Otherwise, if `opt[i][j] = opt[i][j + 1] + 2`, then align `y[j]` with a gap and penalty of 2, and increment `j`
- Otherwise, align `x[i]` with `y[j]` with a penalty of 0 or 1 depending on whether `x[i]` equals `y[j]`, and increment both `i` and `j`

If `y` is exhausted before `x` (ie, `i < m`), align the remaining `x` with gaps and penalty of 2

If `x` is exhausted before `y` (ie, `j < n`), align the remaining `y` with gaps and penalty of 2