情報基礎実験 動的計画法 1

05-175511 中林亮

1. 初期値

M_{ij} は $i-1, j-1$ 文字目までを使うアライメントで、最後がマッチ・ミスマッチで終わっているものの点数の最大値。 I^x_{ij}, I^y_{ij} は $i-1, j-1$ 文字目までを使うアライメントで、最後がギャップで終わっているものの点数の最大値。

$M_{00} = 0$ とするとき、

$M_{i,0} = -d - (i-1)e$ (if $i > 0$)
$M_{0,j} = -d - (j-1)e$ (if $j > 0$)
$I^x_{i,0} = -\infty, I^x_{0,j} = -\infty, I^y_{i,0} = -\infty, I^y_{0,j} = -\infty$

$M_{i,0}$ は、x の $i-1$ 文字目までと空文字列のアライメントと考えれば、ギャップのペナルティを適切に設定できる。
I^x, I^y は、直前がギャップではないのでありえない。なので $-\infty$ にしておく。

2. 対称性

更新式は

$$M_{i+1,j+1} = \max(M_{i,j}, I^x_{i,j}, I^y_{i,j}) + c_{s_i,s_j} I^x_{i+1,j} = \max(M_{i,j} - d, I^x_{i,j} - e, I^y_{i,j} - d) I^y_{i,j+1} = \max(M_{i,j} - d, I^y_{i,j} - e)$$

非対称になっている項は、I^x の 3 項目に対応する I^y の、$I^x_{i,j} - d$ となるはずの項。つまり、x がギャップになる直後の y のギャップは認めるが、y がギャップの直後の x のギャップは認めていない。例えば以下の例では case1 を認める一方 case2 を認めない:

```
case1:
x: - C
y: A -

case2:
x: C -
y: - A
```

x,y にまたがってギャップが連続する時、つまりアライメントできない領域が広がっている時に、そのギャップを x 側を左側に寄せたアライメントだけを代表として数える方針になっている。他のアライメントには影響がなく、またアフィンギャップを考えているならば点数が最大のものを取ることになるので、スコアにも影響はない。

便利だとすれば、確率の計算時などに「アライメントできない領域」をまとめて扱える点である。

3. 実装

NWG.py に実装しました。変数 method の中身によって、メモ化か表埋めかを切り替えるようにしました。
両方の実行結果は result にあります。

4. 確率計算

1:AC, 2:AG をアライメントする
可能なアライメントは以下。

```
AC
AG
score:0   prob:exp(0)/Z

AC-
-AG
score:-15 prob:exp(-15)/Z

-AC
AG-
score:-15 prob:exp(-15)/Z

AC-
A-G
score:-15 prob:exp(-15)/Z

AC--
--AG
score:-28 prob:exp(-28)/Z

A-C
-AG
score:-15 prob:exp(-15)/Z
```

課題 1-3 のアルゴリズムに従っているので、$(C,-),(-,G)$ と $(-,G),(C,-)$ を同一視し、片方のみをアライメントとして計算する。つまり、ギャップをのぞいた時の $(i,j) \in A$ の集合が合同かどうかで判定している。