1 Pairwise Alignments – gap costs

Generelt

~~Høj~~ Ensartethed i sekvenser har høj sandsynlighed for at have ensartet funktion eller struktur.

Edit Distance

Number of operations needed to transform A into B

$a \to -$ deletion
$a \to b$ substitution
$- \to b$ insertion

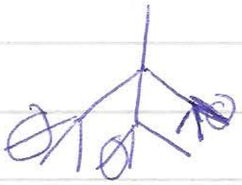
In biointormatics $a \to g$ might be more alike then $a \to c$. Therefore a score matrix is defined along with a gapcost.

Linear gapcost

Given A, B, S, gapcost find OPT alignment of A, B.
Define $cost(i,j) \equiv$ cost of $OPT(A[1 \dots i], B[1 \dots j])$

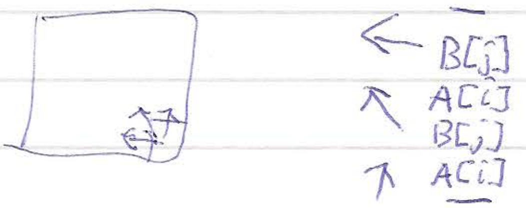
$$cost(i,j) = \max \begin{cases} 0 \\ cost(i-1, j-1) + S(A[i], B[j]) & \begin{bmatrix} \sim A[i] \\ \sim B[j] \end{bmatrix} \\ cost(i-1, j) + gapcost & \begin{bmatrix} \sim A[i] \\ \sim - \end{bmatrix} \\ cost(i, j-1) + gapcost & \begin{bmatrix} \sim - \\ \sim B[j] \end{bmatrix} \end{cases}$$

rekursion vil beregne samme resultat mange gange



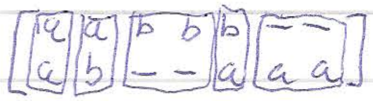
bengt i stedet dynamisk programmering.

Backtracking

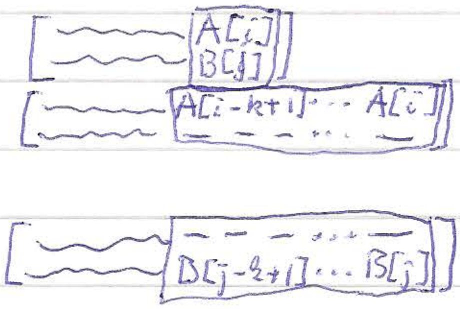


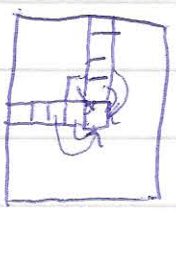
Generel gapcost

$$\text{cost} = \sum_{\text{blocks}} \text{"cost of block"}$$



$$S(i,j) = \max \begin{cases} 0 \\ S(i-1,j-1) + s(A[i],B[i]) \\ \max\limits_{0<k\le i} \left(S(i-k,j) + g(k)\right) \\ \max\limits_{0<k\le j} \left(S(i,j-k) + g(k)\right) \end{cases}$$





Each cell: $O(n)$
Cells: $O(n^2)$

Total: $O(n^3)$

Affine gapcost

$g(k) = a \cdot k + b$

Delition

1) New block

$$\overbrace{}^{m(i-1,j)-(a+b)}$$
$$\left[\begin{array}{cc}\sim\sim\ A[i-1] & A[i] \\ \sim\sim\ B[j] & - \end{array}\right]$$

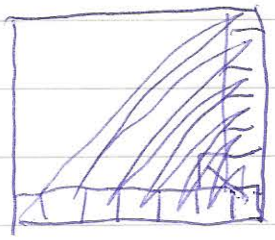
$$\overbrace{}^{I(i-1,j)-(a+b)}$$
$$\left[\begin{array}{cc}\sim\sim\ \cancel{A[} & A[i] \\ \sim\sim\ B[j] & - \end{array}\right]$$

2) Extension of block

$$\overbrace{}^{D(i-1,j)-a}$$
$$\left[\begin{array}{ccc}\sim\sim\ A[i-k] & \cdots & A[i] \\ \sim\sim\ - & - \cdots & - \end{array}\right]$$

$$D(i,j) = \max \begin{cases} m(i-1,j) - (a+b) \\ I(i-1,j) - (a+b) \\ D(i-1,j) - a \end{cases} = \max \begin{cases} S(i-1,j) - (a+b) \\ D(i-1,j) - a \end{cases}$$

Backtracking



$O(k)$ for k steps

$O(n)$ total.