

1 le problème de décision

J'ai deux séquences et je réalise un alignement global de type NWS. On calcule le pourcentage d'identité comme le nombre de symboles identique entre les deux séquences rapporté à la longueur de la plus courte séquence.

avec

$$l_{seq} = \min(length(seq1), length(seq2)) \quad (1)$$

$$id = \frac{païresidentiques}{l_{seq}} \quad (2)$$

le nombre de différences d est :

$$d = \lceil (1 - id) \times l_{seq} \rceil \quad (3)$$

Si les différences sont équitablement réparties on est sûr de partager au moins d mots de taille l_{max}

$$l_{max} = \lfloor \frac{l_{seq}}{d+1} - 1 \rfloor \quad (4)$$

et il restera un mot de taille l_{reste}

$$l_{reste} = l_{seq} - (l_{max} + 1) \times d - l_{max} \quad (5)$$

soit au minimum en commun $words$ mots de taille $l_{word} \leq l_{max}$

$$words = d \times (l_{max} - l_{word} + 1) + \max(l_{reste} - l_{word} + 1, 0) \quad (6)$$

Même si l'on oublie transitoirement le max il nous reste après simplification

$$words = l_{seq} - (d+1)l_{word} - \lfloor \frac{l_{seq}}{d+1} - 1 \rfloor + 1 \quad (7)$$

et là si j'essaye d'isoler d d'un côté

$$words - l_{seq} + l_{word} - 1 = -dl_{word} - \lfloor \frac{l_{seq}}{d+1} - 1 \rfloor \quad (8)$$

$$l_{seq} - words - l_{word} + 1 = dl_{word} + \lfloor \frac{l_{seq}}{d+1} - 1 \rfloor \quad (9)$$

je me heurte au floor dans lequel d apparait au dénominateur et donc je ne sais pas aller plus loin

$$l_{seq} - words - l_{word} + 2 = dl_{word} + \lfloor \frac{l_{seq}}{d+1} \rfloor \quad (10)$$