AAGB - TD1 - 2020

1. Introduction

For each of the biological problems seen during the course describe:

1. The problematic.
2. The type of input data .
3. Where this data can be found (databases, experiment, ...).
4. The actual state-of-the-art and/or models.

2. Sequence comparison

1. What is the point of comparing sequences ?
2. What is the definition of homology ?
3. How can one compare sequences?

2.1 Dot Plot (similarity matrix)

Two sequences can easily be compared by placing one on the abciss and the other on the ordinate, and by marking the pair of identical letters. In order to better discriminate the conserved regions, it is also possible to use a n-sized window, so as to highlight only the identical words of length > n.

1. Given the two following sequences A = (GCACTAGACC) and B = (GCATCGAC), build the corresponding dot-plot, for a window of length 3.
2. What would be the interpretation if we had a long diagonal in the dot-plot? A diagonal cut in two?

2.2 Sequence alignment: the Needleman & Wunsch algorithm

Sequence alignment is by definition their comparison in order to show their similarity. The goal is to maximize the number of matches between nucleotides/amino acids, and to minimize the number of differences (mismatches).

Definition

Let :

- $A = (a_1, a_2, ..a_n)$ et $B = (b_1, b_2, ..b_n)$ be two sequences,
- $S_{i,j}$ the maximum score of the alignment between the subsequences $(a_1, a_2, ..a_i)$ and $(b_1, b_2, ..b_j)$,
- g the score given to a gap,
- $\sigma(a_i, b_j)$ the mismatch or match score.

Algorithm

- Initialization : $\begin{cases} S_{i,0} = i \times g \\ S_{0,j} = j \times g \end{cases}$

- Induction : $S_{i,j} = max \begin{cases} S_{i-1,j-1} + \sigma(a_i, b_j) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{cases}$

- Traceback : At each step, the best scoring alignment is stored. This will give the path corresponding to the best global alignment(s).

We would like to align sequences A = (CATGAC) and B = (TCTGAAC), with the following scores:

- $g = -1$ (gap)

- mismatch = -2

- match = 1

1. Fill the table, following the Needleman Wunsch algorithm.

2. Use the traceback to find the corresponding alignment.

3. How can this algorithm be adapted if we would rather get a local alignment instead of a global one?

		C	**A**	**T**	**G**	**A**	**C**
	0	-1	-2	-3	-4	-5	-6
T	-1						
C	-2						
T	-3						
G	-4						
A	-5						
A	-6						
C	-7						