

Data structure project

TITLE: GLOBAL ALIGNMENT OF PROTEIN SEQUENCES USING NEEDLE WUNSCH ALGORITHM

DATA STRUCTURES USED

February 1, 2019

TEAM MEMBERS:

MUHAMMAD AFNAN 17K-3869

NOMAN MUQEEM 17K-3905

OMER SOHAIL 17K-3799

DYNAMIC SAFE ARRAY

&

STACK

# 

# **Global sequence alignment**

The best alignment over the entire sequences

length of two

Suitable when the two sequences are of similar length,

with a significant degree of similarity throughout.

Example:

SIMILARITY PI-LLAR---

# **The Needleman-Wunsch algorithm**

A smart way to reduce the massive number of possibilities

that need to be considered, yet still guarantees that the

best solution will be found (Saul Wunsch, 1970).

Needleman and Christian

The basic idea is to build up the best alignment by using

optimal alignments of smaller subsequences.

The Needleman-Wunsch algorithm dynamic programming, a discipline

is an example of invented by Richard

Bellman (an American mathematician) in 1953!

# 

# **The mathematics**

A matrix *D*(*i, j*) indexed by residues built recursively, such that

of each sequence is

*D*(*i, j*) = max

*D*(*i* − 1*, j* − 1) + *s*(*xi, yj*)

*D*(*i* − 1*, j*) + *g*

*D*(*i, j* − 1)

subject to a boundary conditions. *s*(*i,j*) is the substitution score for residues *i* and *j*, and *g* is the gap penalty.

# A walk-through: an overview

We consider all possible pairs of residue from two

sequences (this gives rise to a 2D matrix representation).

We will have two matrices: the score matrix and traceback matrix.

The Needleman-Wunsch algorithm consists of three steps:

* + 1. Initialisation of the score matrix
    2. Calculation of scores and filling
    3. Deducing the alignment from the traceback matrix

the traceback matrix

# **The score and traceback matrices**

The cells of the score matrix are labelled *C*(*i, j*) where

*i* = 1*,* 2*, ..., N* and *j* = 1*,* 2*, ..., M*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| done | left | left | left | left |
| up |  |  |  |  |
| up |  |  |  |  |
| up |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | −10 | −20 | −30 | −40 |
| −10 |  |  |  |  |
| −20 |  |  |  |  |
| −30 |  |  |  |  |

# **Scoring**

The score matrix cells are filled by row starting from the cell *C*(2*,* 2)

The score of any cell *C*(*i, j*) is the maximum of:

*qdiag* = *C*(*i* − 1*, j* − 1) + *S*(*i, j*)

*qup* = *C*(*i* − 1*, j*) + *g*

*qlef t* = *C*(*i, j* − 1) + *g*

Where *S*(*i, j*) is the substitution and *g* is the gap penalty.

# 

# Filling the score and traceback matrices

For the score matrix *C*(2*,* 2) = *qdiag* which is 1. The

corresponding cell of the traceback matrix is ”diag”:

S E N D S E N D

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | −10 | −20 | −30 | −40 |
| −10 | **1** |  |  |  |
| −20 |  |  |  |  |
| −30 |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| done | left | left | left | left |
| up | **diag** |  |  |  |
| up |  |  |  |  |
| up |  |  |  |  |

# **The traceback**

Traceback = the process of deduction alignment from the traceback matrix.

of the best

The traceback always begins with the last cell to be filled

with the score, i.e. the bottom right cell.

One moves according to the traceback cell.

value written in the

There are three possible moves: diagonally (toward the

top-left corner of the matrix), up, or left.

The traceback is completed when the first, top-left cell of

the matrix is reached (”done” cell).

# **The best alignment**

Sequences are aligned backwards.

The alignment is deduced from the values of cells along the traceback path, by taking into the cell in the traceback matrix:

*diag* – the letters from two sequences are aligned

*left* – a gap is introduced in the left sequence

*up* – a gap is introduced in the top sequence

# **A few observations**

As we consider longer sequences the situation quickly

turns against the exhaustive search:

Two 12 residue sequences would require considering

∼ 1 million alignments.

Two 150 residue sequences would require considering

88

78

∼ 10 alignments (∼ 10 is the estimated number of

atoms in the Universe).

For two 150 residue sequences the Needleman-Wunsch

algorithm requires filling a 150 × 150 matrix.

**SUMMARY**

The Needleman-Wunsch algorithm works in the same way

regardless of the length or complexity of sequences, and

*guarantees* to find the best alignment.

The Needleman-Wunsch algorithm is appropriate for

finding the best alignment of two sequences which are (*i*) of the similar length; (*ii*) similar across their entire lengths.