Implementação do algoritmo Needleman-Wunsch

O algoritmo Needleman-Wunsch é utilizado para alinhamento global de sequências de DNA ou aminoácidos. Esse algoritmo retorna o melhor alinhamento possível, porém, o resultado obtido não necessariamente tem uma significância biológica, nesse sentido, é muito importante a escolha de um sistema de pontuação e/ou matriz de substituição adequados ao problema.

Implementou-se o algoritmo, em Python 3.8, para o alinhamento de duas sequencias. O código fonte está disponível no *Github* (https://github.com/glenjasper/needleman-wunsch.git). O algoritmo aceita um arquivo fasta contendo as duas sequencias biológicas, para o qual é requerido escolher a matriz de substituição (BLOSUM ou PAM) e o *gap penalty* para sequencias proteicas, entretanto para sequencias de DNA podem-se configurar o *match* e o *mismatch*. Além do alinhamento feito, o algoritmo retorna a matriz de pontuação em um arquivo TXT. Também, é possível visualizar o alinhamento através do *framework* Dash (https://dash.plotly.com), que oferece uma visualização simples é amigável em HTML.

Pré-requisitos

pip install dash
pip install dash-bio

Uso básico

```
$ python needleman_wunsch.py --help
usage: needleman_wunsch.py [-h] -t {nt,aa} [-sm {BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80,
BLOSUM90, PAM30, PAM70, PAM250}] -f FILE [-m MATCH] [-mi MISMATCH_PENALTY] [-gap
GAP_PENALTY] [-o FOLDER] [--version]

Implementation of the Needleman-Wunsch algorithm

optional arguments:
 -h, --help          show this help message and exit
 -t {nt,aa}, --type {nt,aa}
                     nt: Nucleotide sequence | aa: Amino acid sequence
 -sm {BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, BLOSUM90, PAM30, PAM70, PAM250}, --
substitution_matrix {BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, BLOSUM90, PAM30, PAM70,
PAM250}
                     Substitution Matrix type (Only for amino acid sequence) [default: BLOSUM62].
 -f FILE, --fasta FILE
                     Fasta file
 -m MATCH, --match MATCH
                     Match value (Only for nucleotide sequence) [default: 1].
 -mi MISMATCH_PENALTY, --mismatch_penalty MISMATCH_PENALTY
                     Mismatch penalty value (Only for nucleotide sequence) [default: 0].
 -gap GAP_PENALTY, --gap_penalty GAP_PENALTY
                     Gap penalty value [default: 0].
 -o FOLDER, --output FOLDER
                     Output folder
 --version           show program's version number and exit
```

Parâmetros

Parâmetro	Descrição	Possíveis valores	*Default*
-t | --type	Tipo de sequências a alinhar. Podem ser de aminoácidos ou nucleotídeos.	nt, aa	
-sm | --substitution_matrix	Matriz de substituição BLOSUM ou PAM, usado quando as sequencias forem proteicas.	BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, BLOSUM90, PAM30, PAM70, PAM250	BLOSUM62
-f | --fasta	Arquivo fasta que contem as sequências biológicas a serem alinhadas.		
-m | --match	Valor de *match*. Usado quando as sequencias forem de DNA.		1
-mi | --mismatch_penalty	Valor de *mismatch*. Usado quando as sequencias forem de DNA.		0
-gap | --gap_penalty	Valor de *gap penalty*.		0
-o | --output	Pasta de saída.		

Exemplos

1. Alinhar as sequências proteicas DRQTAQAAGTTTIT e DRNTAQLLGTDTT (contidas no arquivo **file.fa**), com a matriz de substituição BLOSUM80 e *gap penalty* -1.

   $ python needleman_wunsch.py -t aa -f file.fa -gap -1 -o out_align1

   Output:

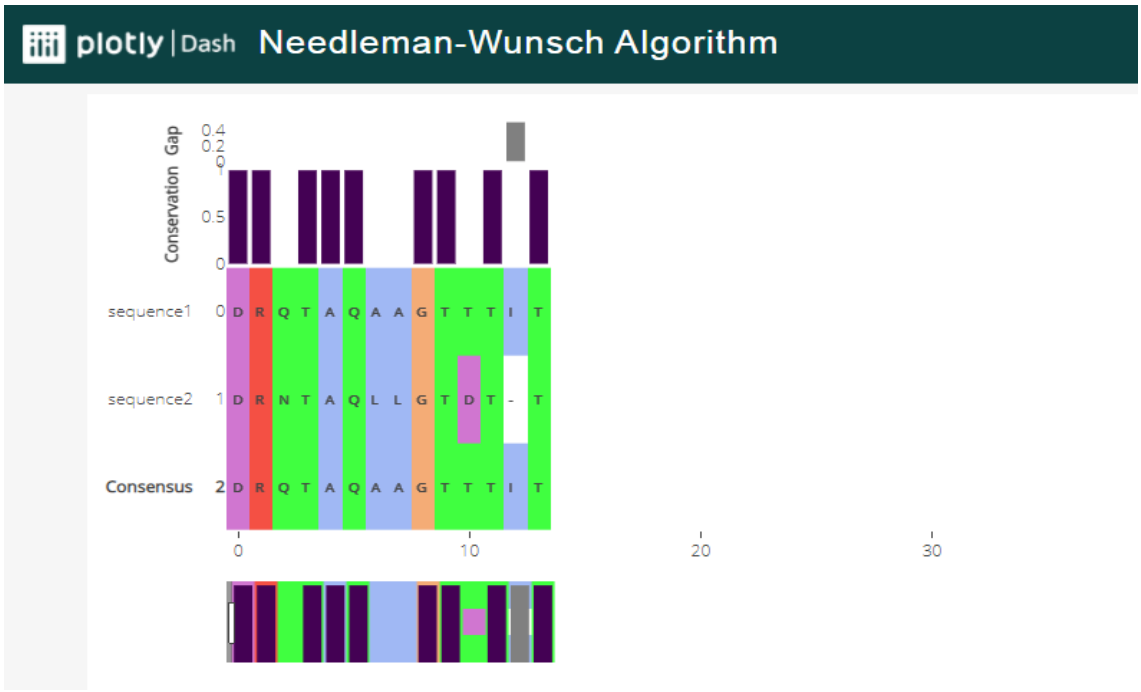
```
2021-02-05 00:14:21 ############################################################
2021-02-05 00:14:21 ################# Needleman-Wunsch Algorithm #################
2021-02-05 00:14:21 ############################################################
2021-02-05 00:14:21 Input:
2021-02-05 00:14:21   Fasta file: C:\Users\Glen\Dropbox\UFMG\Disciplinas\Bioinformática\Atividades\TP1\script\needleman-wunsch\file.fa
2021-02-05 00:14:21
2021-02-05 00:14:21 Parameters:
2021-02-05 00:14:21   Matrix: BLOSUM62
2021-02-05 00:14:21   Gap penalty: -1
2021-02-05 00:14:21
2021-02-05 00:14:21 Alignment:
2021-02-05 00:14:21   Score: 42
2021-02-05 00:14:21
2021-02-05 00:14:21   sequence1    DRQTAQAAGTTTIT
2021-02-05 00:14:21                ||·|||··||·| |
2021-02-05 00:14:21   sequence2    DRNTAQLLGTDT-T
2021-02-05 00:14:21
2021-02-05 00:14:21 Matrix file: C:\Users\Glen\Dropbox\UFMG\Disciplinas\Bioinformática\Atividades\TP1\script\needleman-wunsch\out_align1\alignment_matrix.txt
2021-02-05 00:14:21
2021-02-05 00:14:21 Running Dash's alignment viewer
2021-02-05 00:14:21 Run the address http://127.0.0.1:8050 in your browser
Dash is running on http://127.0.0.1:8050/

 * Serving Flask app "needleman_wunsch" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
```

   Matriz de substituição (ver arquivo **alignment_matrix.txt**):

```
-    -    D    R    Q    T    A    Q    A    A    G    T    T    T    I    T
-    0   -1   -2   -3   -4   -5   -6   -7   -8   -9  -10  -11  -12  -13  -14
D   -1  6|d  5|l  4|l  3|l  2|l  1|l  0|l -1|l -2|l -3|l -4|l -5|l -6|l -7|l
R   -2  5|u 11|d 10|l  9|l  8|l  7|l  6|l  5|l  4|l  3|l  2|l  1|l  0|l -1|l
N   -3  4|u 10|u 11|d 10|d  9|l  8|d  7|l  6|l  5|d  4|d  3|d  2|d  1|l  0|d
T   -4  3|u  9|u 10|u 16|d 15|l 14|l 13|l 12|l 11|l 10|d  9|d  8|d  7|l  6|d
A   -5  2|u  8|u  9|u 15|u 20|d 19|l 18|d 17|d 16|l 15|l 14|l 13|l 12|l 11|l
Q   -6  1|u  7|u 13|d 14|u 19|u 25|d 24|l 23|l 22|l 21|l 20|l 19|l 18|l 17|l
L   -7  0|u  6|u 12|u 13|u 18|u 24|u 24|d 23|d 22|l 21|d 20|d 19|d 21|d 20|l
L   -8 -1|u  5|u 11|u 12|u 17|u 23|u 23|d 23|d 22|l 21|d 20|d 19|d 21|d 20|d
G   -9 -2|u  4|u 10|u 11|u 16|u 22|u 23|d 23|d 29|d 28|l 27|l 26|l 25|l 24|l
T  -10 -3|u  3|u  9|u 15|d 15|u 21|u 22|d 23|d 28|u 34|d 33|d 32|d 31|l 30|d
D  -11 -4|d  2|u  8|u 14|u 14|u 20|u 21|u 22|u 27|u 33|u 33|d 32|d 31|l 30|d
T  -12 -5|u  1|u  7|u 13|d 14|d 19|u 20|d 21|d 26|u 32|d 38|d 38|d 37|l 36|d
T  -13 -6|u  0|u  6|u 12|d 13|d 18|u 19|d 20|d 25|u 31|d 37|d 43|d 42|l 42|d
```

Visualização do alinhamento com o *framework* Dash (HTML+CSS), através do endereço http://127.0.0.1:8050:



2. Alinhar as glicoproteínas Spike sp|P11223|SPIKE_IBVB e sp|P12651|SPIKE_IBVM (contidas no arquivo **file2.fa**), com a matriz de substituição BLOSUM62 e *gap penalty* -2.

```
$ python needleman_wunsch.py -t aa -f file2.fa -gap -2 -o out_align2
```

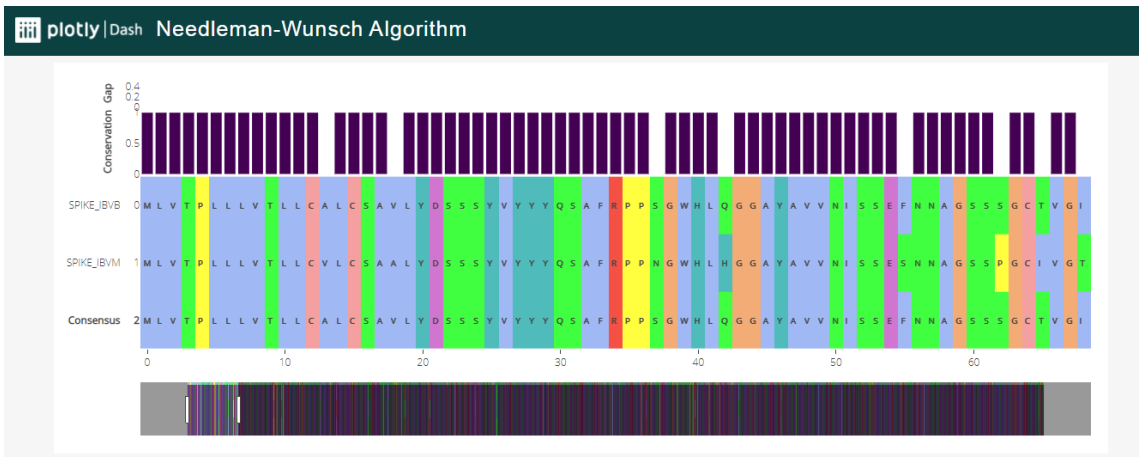
Output:

```
2021-02-05 00:25:56 ############################################################
2021-02-05 00:25:56 ################ Needleman-Wunsch Algorithm #################
2021-02-05 00:25:56 ############################################################
2021-02-05 00:25:56 Input:
2021-02-05 00:25:56   Fasta file: C:\Users\Glen\Dropbox\UFMG\Disciplinas\Bioinformática\Atividades\TP1\script\needleman-wunsch\file2.fa
2021-02-05 00:25:56
2021-02-05 00:25:56 Parameters:
2021-02-05 00:25:56   Matrix: BLOSUM62
2021-02-05 00:25:56   Gap penalty: -2
2021-02-05 00:25:56
2021-02-05 00:25:56 Alignment:
2021-02-05 00:26:13   Score: 5869
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    MLVTPLLLVTLLCALCSAVLYDSSSYVYYYQSAFRPPSGWHLQGGAYAVVNISSEFNNAG
2021-02-05 00:26:13                           ||||||||||||||·||||·||||||||||||||||·||||·||||||||||||·||||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    MLVTPLLLVTLLCVLCSAALYDSSSYVYYYQSAFRPPNGWHLHGGAYAVVNISSESNNAG
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    SSSGCTVGIIHGGRVVNASSIAMTAPSSGMAWSSSQFCTAHCNFSDTTVFVTHCYKHGGC
2021-02-05 00:26:13                           ||·||·||·||||||||||||||||||||||||||||||||||||||||||||||||··||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    SSPGCIVGTIHGGRVVNASSIAMTAPSSGMAWSSSQFCTAHCNFSDTTVFVTHCYKYDGC
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    PLTGMLQQNLIRVSAMKNGQLFYNLTVSVAKYPTFRSFQCVNNLTSVYLNGDLVYTSNET
2021-02-05 00:26:13                           |·|||||·|···|||||||||||||||||||||·|||||||||||||||||||||||||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    PITGMLQKNFLRVSAMKNGQLFYNLTVSVAKYPTFKSFQCVNNLTSVYLNGDLVYTSNET
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    IDVTSAGVYFKAGGPITYKVMREVKALAYFVNGTAQDVILCDGSPRGLLACQYNTGNFSD
2021-02-05 00:26:13                           ·|||||||||||||||||||||·|||||||||||||||||||||||||||||||||||||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    TDVTSAGVYFKAGGPITYKVMRKVKALAYFVNGTAQDVILCDGSPRGLLACQYNTGNFSD
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    GFYPFTNSSLVKQKFIVYRENSVNTTCTLHNFIFHNETGANPNPSGVQNIQTYQTKTAQS
2021-02-05 00:26:13                           |||||·||||||||||||||||·|||||·|||||||||||||||·||||·||||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    GFYPFINSSLVKQKFIVYRENSVNTTFTLHNFTFHNETGANPNPSGVQNILTYQTQTAQS
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    GYYNFNFSFLSSFVYKESNFMYGSYHPSCKFRLETINNGLWFNSLSVSIAYGPLQGGCKQ
2021-02-05 00:26:13                           ||||||||||||||||||||||||||||||·||||||||||||||||||||||||||||||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    GYYNFNFSFLSSFVYKESNFMYGSYHPSCNFRLETINNGLWFNSLSVSIAYGPLQGGCKQ
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    SVFKGRATCCYAYSYGGPSLCKGVYSGELDHNFECGLLVYVTKSGGSRIQTATEPPVITQ
2021-02-05 00:26:13                           |||·||||||||||||||||||||||||||·|||||||||||||||||||||||||||·
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    SVFSGRATCCYAYSYGGPSLCKGVYSGELDLNFECGLLVYVTKSGGSRIQTATEPPVITR
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    NNYNNITLNTCVDYNIYGRTGQGFITNVTDSAVSYNYLADAGLAILDTSGSIDIFVVQGE
2021-02-05 00:26:13                           ·||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    HNYNNITLNTCVDYNIYGRTGQGFITNVTDSAVSYNYLADAGLAILDTSGSIDIFVVQGE
2021-02-05 00:26:13
2021-02-05 00:26:13   sp|P11223|SPIKE_IBVB    YGLNYYKVNPCEDVNQQFVVSGGKLVGILTSRNETGSQLLENQFYIKITNGTRRFRRSIT
2021-02-05 00:26:13                           |||·||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2021-02-05 00:26:13   sp|P12651|SPIKE_IBVM    YGLTYYKVNPCEDVNQQFVVSGGKLVGILTSRNETGSQLLENQFYIKITNGTRRFRRSIT
```

Matriz de substituição (ver arquivo **alignment_matrix.txt**). Apenas uma parte é mostrada porque a matriz é muito extensa:

-	-	M	L	V	T	P	L	L	L	V	T	L	L	C	A	L	C	S
-	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24	-26	-28	-30	-32	-34
M	-2	5|d	3|l	1|l	-1|l	-3|l	-5|l	-7|l	-9|l	-11|l	-13|l	-15|l	-17|l	-19|l	-21|l	-23|l	-25|l	-27|l
L	-4	3|u	9|d	7|l	5|l	3|l	1|d	-1|d	-3|d	-5|l	-7|l	-9|d	-11|d	-13|l	-15|l	-17|d	-19|l	-21|l
V	-6	1|u	7|u	13|d	11|l	9|l	7|l	5|l	3|l	1|d	-1|l	-3|l	-5|l	-7|l	-9|l	-11|l	-13|l	-15|l
T	-8	-1|u	5|u	11|u	18|d	16|l	14|l	12|l	10|l	8|l	6|d	4|l	2|l	0|l	-2|l	-4|l	-6|l	-8|l
P	-10	-3|u	3|u	9|u	16|u	25|d	23|l	21|l	19|l	17|l	15|l	13|l	11|l	9|l	7|l	5|l	3|l	1|l
L	-12	-5|u	1|d	7|u	14|u	23|u	29|d	27|d	25|d	23|l	21|l	19|d	17|d	15|l	13|l	11|d	9|l	7|l
L	-14	-7|u	-1|d	5|u	12|u	21|u	27|d	33|d	31|d	29|l	27|l	25|d	23|d	21|l	19|l	17|d	15|l	13|l
L	-16	-9|u	-3|d	3|u	10|u	19|u	25|d	31|d	37|d	35|l	33|l	31|d	29|d	27|l	25|l	23|d	21|l	19|l
V	-18	-11|u	-5|u	1|d	8|u	17|u	23|u	29|u	35|u	41|d	39|l	37|l	35|l	33|l	31|l	29|l	27|l	25|l
T	-20	-13|u	-7|u	-1|u	6|d	15|u	21|u	27|u	33|u	39|u	46|d	44|l	42|l	40|l	38|l	36|l	34|l	32|l
L	-22	-15|u	-9|d	-3|u	4|u	13|u	19|d	25|d	31|d	37|u	44|u	50|d	48|d	46|l	44|l	42|d	40|l	38|l
L	-24	-17|u	-11|d	-5|u	2|u	11|u	17|d	23|d	29|d	35|u	42|u	48|d	54|d	52|l	50|l	48|d	46|l	44|l
C	-26	-19|u	-13|u	-7|u	0|u	9|u	15|u	21|u	27|u	33|u	40|u	46|u	52|u	63|d	61|l	59|l	57|d	55|l
V	-28	-21|u	-15|u	-9|d	-2|u	7|u	13|u	19|u	25|u	31|d	38|u	44|u	50|u	61|u	63|d	62|d	60|l	58|l
L	-30	-23|u	-17|d	-11|u	-4|u	5|u	11|d	17|d	23|d	29|u	36|u	42|d	48|d	59|u	61|u	67|d	65|l	63|l
C	-32	-25|u	-19|u	-13|u	-6|u	3|u	9|u	15|u	21|u	27|u	34|u	40|u	46|u	57|d	59|d	65|u	76|d	74|l
S	-34	-27|u	-21|u	-15|u	-8|u	1|u	7|u	13|u	19|u	25|u	32|u	38|u	44|u	55|u	58|d	63|u	74|u	80|d
A	-36	-29|u	-23|u	-17|u	-10|u	-1|u	5|u	11|u	17|u	23|u	30|u	36|u	42|u	53|u	59|d	61|u	72|u	78|u
A	-38	-31|u	-25|u	-19|u	-12|u	-3|u	3|u	9|u	15|u	21|u	28|u	34|u	40|u	51|u	57|d	59|u	70|u	76|u
L	-40	-33|u	-27|d	-21|u	-14|u	-5|u	1|d	7|d	13|d	19|u	26|u	32|d	38|d	49|u	55|u	61|d	68|u	74|u
Y	-42	-35|u	-29|u	-23|u	-16|u	-7|u	-1|u	5|u	11|u	17|u	24|u	30|u	36|u	47|u	53|u	59|u	66|u	72|u
D	-44	-37|u	-31|u	-25|u	-18|u	-9|u	-3|u	3|u	9|u	15|u	22|u	28|u	34|u	45|u	51|u	57|u	64|u	70|u
S	-46	-39|u	-33|u	-27|u	-20|u	-11|u	-5|u	1|u	7|u	13|u	20|u	26|u	32|u	43|u	49|u	55|u	62|u	68|d
S	-48	-41|u	-35|u	-29|u	-22|u	-13|u	-7|u	-1|u	5|u	11|u	18|u	24|u	30|u	41|u	47|u	53|u	60|u	66|d
S	-50	-43|u	-37|u	-31|u	-24|u	-15|u	-9|u	-3|u	3|u	9|u	16|u	22|u	28|u	39|u	45|u	51|u	58|u	64|d

Visualização do alinhamento com o *framework* Dash (HTML+CSS), através do endereço http://127.0.0.1:8050:



3. Alinhar as sequências de DNA AATTTACGCGGCATTATAGATACAATCGTGTCT e GCAATTGGCCGGAATTTAATTGATACAGCGC (contidas no arquivo **file3.fa**), com valores de *match* +2, *mismatch* -1 e *gap penalty* -2.

   $ python needleman_wunsch.py -t nt -f file3.fa -m 2 -mi -1 -gap -2 -o out_align3

   Output:

   Visualização do alinhamento com o *framework* Dash (HTML+CSS), através do endereço http://127.0.0.1:8050: