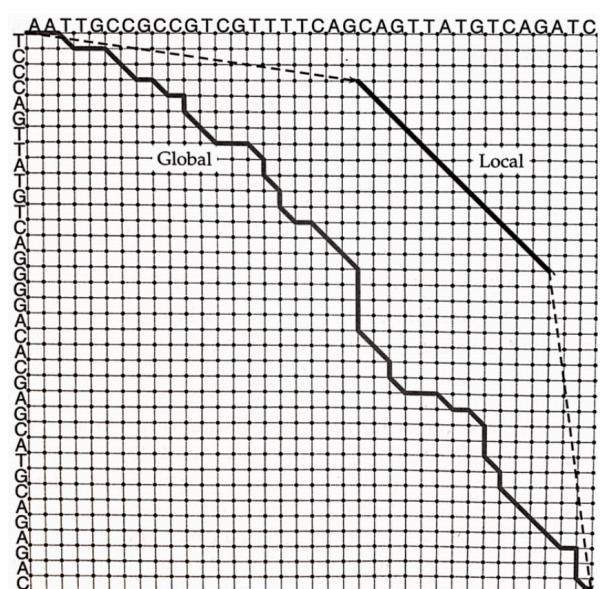


PROJETO 1

ALINHAMENTO DE SEQUÊNCIAS

Exemplo de um alinhamento múltiplo de sequências

Conforme vimos nas aulas do módulo sobre *alinhamento de sequências*, este tipo de técnica é provavelmente a mais utilizada em toda a Bioinformática em todos os tempos. O alinhamento de sequências tem as mais diversas aplicações que vão desde a genômica comparativa, os estudos sobre a evolução e filogenia, análises de variantes genéticas, montagem de genomas até a sobreposição estrutural e a modelagem comparativa de estruturas tridimensionais de proteínas.



Comparativo entre um alinhamento global e

sequências. Por outro lado, os alinhamentos múltiplos são obtidos por

Vimos também que existem vários tipos de algoritmos de alinhamento, com propósitos bastante diversos. Por um lado, há o *alinhamento global* que visa alinhar as sequências de proteínas em todo o seu comprimento. Por outro, há o *alinhamento local* que visa obter subsequências de alta similaridade. Em relação ao número de sequências sendo alinhadas, vimos que há o *alinhamento par-a-par*, em que se alinha um par de sequências e o *alinhamento múltiplo*, em que toda uma família de sequências é alinhada. Vimos que alinhamentos par-a-par são construídos através de algoritmos polinomiais de ordem quadrática em relação ao número de nucleotídeos / aminoácidos das

alinhamentos múltiplos são obtidos por meio de heurísticas as mais diversas, pois obter o alinhamento múltiplo ótimo é um problema não polinomial.

Tendo em vista a grande utilidade dos alinhamentos, vamos exercitar nesse curso a implementação de um dos algoritmos mais tradicionais e simples de alinhamento de sequências.

O objetivo deste trabalho prático é implementar o *algoritmo de Needleman-Wunsch* para construir alinhamentos globais par-a-par de sequências de proteínas. Sugerimos seguir o algoritmo proposto no livro de Jones e Pevzner [1] que utilizamos como referência no curso. O capítulo necessário foi disponibilizado no Moodle.

Para que se obtenha uma implementação com utilidade prática, sugerimos usar a matriz

Ala	4																		
Arg	-1 5																		
Asn	-2 0 6																		
Asp	-2 -2 1 6																		
Cys	0 -3 -3 -3 9																		
Gln	-1 1 0 0 -3 5																		
Glu	-1 0 0 2 -4 2 5																		
Gly	0 -2 0 -1 -3 -2 -2 6																		
His	-2 0 1 -1 -3 0 0 -2 8																		
Ile	-1 -3 -3 -3 -1 -3 -4 -3 4																		
Leu	-1 -2 -3 -4 -1 -2 -3 -4 -3 2 4																		
Lys	-1 2 0 -1 -3 1 1 -2 -1 -3 -2 5																		
Met	-1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5																		
Phe	-2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6																		
Pro	-1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7																		
Ser	1 -1 1 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4																		
Thr	0 -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5																		
Trp	-3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11																		
Tyr	-2 -2 -2 -3 -2 -1 -2 -3 -2 -1 -2 -1 3 -3 -2 -2 2 7																		
Val	0 -3 -3 -3 -1 -2 -3 -3 -3 1 -2 -1 -1 -2 -2 0 -3 1 4																		
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

Matriz BLOSUM62

- *matches*: pontuação para casamento de resíduos de aminoácidos idênticos
 - *mismatches*: pontuação para casamento de resíduos de aminoácidos diferentes (isso pode acontecer caso eles seja

- *indels*: inserções e deleções (gaps).

Vocês devem testar a implementação com diversas sequências. Sequências pequenas fictícias foram disponibilizadas, como exemplos, assim como sequências reais. Nos exemplos que apresentaremos abaixo usamos essa penalidade com valor 0. Sugerimos testar outros valores como -5, por exemplo.

Exemplos

1. Tente alinhar sequências idênticas e bem pequenas como DROT por exemplo.

	*	D	R	Q	T
*	0	0	0	0	0
D	0	6\	6_	6_	6_
R	0	6	11\	11_	11_
Q	0	6	11	16\	16_
T	0	6	11	16	21\

-DRQT

*	*	D	R	Q	T
*	0	0	0	0	0
D	0	6\	6_	6_	6_
R	0	6	11\	11_	11_
E	0	6	6	11	13\
T	0	6	6	11	18\

-DRQT
-DRET

2. Faça pequenas alterações na sequência substituindo uma *glutamina* (Q) por um *glutamato* (E).

3. Alinhe sequências um pouco maiores e de comprimentos diferentes como DRQTAQAAAGTTTIT e DRNTAQQLLGTDTT

*	D	R	Q	T	A	Q	A	A	G	T	T	T	I	T
*	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	6\	6_	6_	6_	6_	6_	6_	6_	6_	6_	6_	6_	6_
R	0	6	11\	11_	11_	11_	11_	11_	11_	11_	11_	11_	11_	11_
N	0	6	11	11\	11_	11_	11_	11_	11\	11\	11\	11_	11_	11\
T	0	6	11	11_	16\	16_	16_	16_	16_	16\	16\	16\	16_	16\
A	0	6	11	11_	16	20\	20_	20\	20_	20_	20_	20_	20_	20_
Q	0	6	11	16\	16_	20	25\	25_	25_	25_	25_	25_	25_	25_
L	0	6	11	16	16_	20	25	25_	25_	25_	25_	25_	25_	27\
L	0	6	11	16	16_	20	25	25_	25_	25_	25_	25_	25_	27_
G	0	6	11	16	16_	20	25	25_	25\	31\	31_	31_	31_	31_
T	0	6	11	16	21\	21_	25	25\	25\	31	36\	36\	36\	36\
D	0	6\	11	16	21	21_	25	25_	31	36	36_	36_	36_	36_
T	0	6	11	16	21\	21\	25	25\	31	36\	41\	41\	41_	41\
T	0	6	11	16	21\	21\	25	25\	31	36\	41\	46\	46\	46\

-DRQTAQ--AAGT-TTIT
-DRNTAQQLL--GTD-T-T

4. Aline pares de sequências reais como as fornecidas no arquivo suplementar a esse enunciado. Trata-se de diversas proteínas *Spike* do SARS-CoV-2 e de outros vírus. Se quiser saber mais sobre essa proteína, deixamos abaixo [4,5] os dois artigos mais importantes sobre a descoberta desta proteína e sua interação com a proteína ACE2 (Enzima conversora de angiotensina 2) através da qual entra na célula humana.

Implementação

Sugerimos que o algoritmo seja implementado em *Python*. Você deve implementar o algoritmo e não usar nenhuma biblioteca de terceiros (*BioPython*, por exemplo).

Grupos

O trabalho pode ser desenvolvido em grupos de até 5 pessoas. Sugerimos que vocês tentem criar grupos com integrantes de formações variadas por acreditar que isso enriquecerá muito o aprendizado e as discussões.

Entrega

Vocês devem entregar o código fonte em um arquivo .zip contendo também uma página README em pdf com todas as instruções para sua execução.

O trabalho deve ser entregue até o dia 15/08/2021 até 23:59 exclusivamente via Moodle.

Referências

- [1] Jones, Neil C., Pavel A. Pevzner, and Pavel Pevzner. *An introduction to bioinformatics algorithms*. MIT press, 2004.
- [2] Henikoff, Steven, and Jorja G. Henikoff. "Amino acid substitution matrices from protein blocks." *Proceedings of the National Academy of Sciences* 89.22 (1992): 10915-10919.
- [3] Eddy, Sean R. "Where did the BLOSUM62 alignment score matrix come from?." *Nature biotechnology* 22.8 (2004): 1035-1036.
- [4] Walls, Alexandra C., et al. "Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein." *Cell* (2020).
- [5] Lan, Jun, et al. "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor." *Nature* 581.7807 (2020): 215-220.