

Trabalho 1

**Algoritmos para alinhamentos de sequências: Global e Local**

Unidade Curricular: **Algoritmos para a Bioinformática**

Prof. Vitor Manuel de Morais Santos Costa

8 de Abril de 2018

Carolina Ferreira Gomes Centeio Jorge - 201403090

Higor Araújo dos Anjos - 201711183

Maria Eduarda Marques de Almeida – 200906073

# Introdução

O alinhamento de sequências biológicas, como DNA, RNA e proteínas, permite a identificação de regiões semelhantes, i.e. regiões conservadas. Estas últimas podem representar uma relação funcional, estrutural ou evolutiva entre sequências semelhantes [1]. Desta forma, o alinhamento de sequências tornou-se uma ferramenta essencial na biologia molecular.

Abordagens computacionais permitiram agilizar o processo de alinhamento de sequências. Estas abordagens recaem na aplicação de algoritmos, como por exemplo algoritmos de programação dinâmica que apesar de serem processos lentos são altamente precisos, ou de algoritmos eurísticos ou probabilísticos que apesar de mais rápidos e permitirem a pesquisa de grandes bases de dados não garatem resultados tão precisos.

As abordagens aplicadas no alinhamento de sequências podem também ser dividas em alinhamentos globais e alinhamentos locais. Enquanto que os primeiros têm como objetivo alinhar todos os resíduos de todas as sequências abrangendo o comprimento total das sequências, por outro lado os alinhamentos locais identificam regiões de similaridade dentro de sequências longas e que podem ser no geral muito diferentes entre si.

Em suma, os alinhamentos globais são extraordinariamente úteis no caso de sequências extremamente semelhantes e com tamanhos idênticos. No caso de sequências com pouca similaridade, mas que são suspeitas de conter regiões internas similares, a melhor solução são os alinhamentos locais.

Um exemplo de um algoritmo de alinhamento global é *Needleman–Wunsch algorithm*, que é usado em bioinformática para alinhar sequências proteicas ou nucleotídicas. Este algoritmo foi desenvolvido em 1970 por Saul B. Needleman e Christian D. Wunsch e foi uma das primeiras aplicações da programação dinâmica na comparação de sequências biológicas [2]. Na sua raiz, este algoritmo consiste na divisão da sequência inicial inteira numa série de subsequências de menor tamanho, obtendo uma solução para cada uma destas últimas e reconstruindo uma solução para a sequência maior a partir destas soluções. Este algoritmo é correntemente amplamente usado, como por exemplo na plataforma do EBI-EMBL – o EMBOSS Needle.

O algoritmo Smith-Waterman é um exemplo de um método baseado em programação dinâmica que permite obter alinhamentos locais [2]. Este algoritmo é também correntemente amplamente usado, como por exemplo na plataforma do EBI-EMBL – o EMBOSS Water.

O presente trabalho teve como objetivo a aplicação de algoritmos para a realização de alinhamentos local e global, usando a linguagem Python.

# Desenvolvimento

O desenvolvimento do presente trabalho passou pela implementação de algoritmos que permitiram a obtenção dos alinhamentos global e local.

No caso do alinhamento global o algoritmo implementado foi o seguinte:

em que,

e,

.

Para a obtenção do alinhamento local, aplicou-se o seguinte algoritmo:

em que,

e,

.

Ambos os algoritmos de programação dinâmica, i.e., resolvendo os problemas um a um até chegar a uma solução final, foram implementados usando linguagem Python.

Tendo como objetivo a comparação do algoritmo global desenvolvido no presente trabalho com o implementado na plataforma EBI-EMBLE, o EMBOSS Needle, que é também um algoritmo de alinhamento global baseado no algoritmo de Needleman–Wunsch, foi também desenvolvido um pequeno programa que permite correr o EMBOSS Needle localmente.

A título de exemplo, usamos duas sequências proteicas humanas que codificam para *Plasticity Related Gene 3* (PRG\_3): B7Z8R9 e Q9Y2Y8. Estas sequências foram alinhadas usando o programa de alinhamento global desenvolvido no presente trabalho, bem como a plataforma on-line EMBOSS Needle do EBI-EMBL.

Por fim, uma interface *user-friendly* foi implementada com o intuito de correr estes algoritmos mais facilmente por qualquer tipo de utilizador.

# Resultados & Discussão

Ambos os algoritmos implementados com o presente trabalho (Local e Global), permitiram o alinhamento de sequências biológicas.

A comparação do programa dinâmico para o alinhamento global com a plataforma do EMBOSS Needle, usando 2 sequências diferentes do gene humano PRG\_3, demostrou um melhor desempenho do algoritmo global desenvolvido no presente trabalho. Isto porque usando as mesmas sequências proteicas para o alinhamento em ambos os programas se obteve um número maior de identidades completas (37 no programa desenvolvido no presente trabalho vs. 13 no EMBOSS Needle) e de similaridades (122 no programa desenvolvido no presente trabalho vs. 23 no EMBOSS Needle). Consequentemente, o número de *gaps* com o programa aqui desenvolvido foi menor (208 *gaps*) do que no EMBOSS Needle (414 *gaps*). Estes dados demostram um desempenho melhorado no alinhamento das sequências usando o programa de alinhamento global desenvolvido no presente trabalho.

No que respeita o algoritmo local desenvolvido no presente trabalho, fez-se também a comparação com a plataforma semelhante no EBI-EMBLE, o EMBOSS Water, usando as mesmas sequências anteriormente mencionadas. Esta comparação demostrou um desempenho equiparável de ambos os programas, isto porque o número de identidades completas foi o mesmo para ambas (30 identidades), bem como o número de *gaps* (30 *gaps*). Contudo, a nível de similaridades o programa aqui desenvolvido mostrou um melhor desempenho, uma vez que obteve 68 similaridades quando o EMBOSS Water obteve apenas 45 similaridades.

Em conclusão, ambos os programas desenvolvidos no presente trabalho demostraram uma melhor performance do que os seus correspondentes implementados na plataforma on-line EBI-EMBL.

# Referências

*[1]* Mount DM. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*0-87969-608-7*](https://en.wikipedia.org/wiki/Special:BookSources/0-87969-608-7).

[2] Polyanovsky, V. O.; Roytberg, M. A.; Tumanyan, V. G. (2011). "Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences". Algorithms for Molecular Biology. 6 (1): 25. doi:10.1186/1748-7188-6-25.