



# CIIMAR

## Curso de formação

# Análise de dados provenientes de técnicas moleculares

Formadores: Filipe Pereira e Filipe Lopes

Manual do Curso

# Índice

Objetivo Geral do Curso .....	3
Público-alvo .....	3
Objetivos Específicos do Curso .....	4
Métodos de Ensino e Atividades de Aprendizagem .....	5
Língua de Trabalho .....	5
Componentes de Avaliação .....	5
Aula 1 - Como Identificar uma Sequência de DNA Desconhecida.....	6
Aula 2 – Pesquisas no Blast Utilizando o Geneious .....	9
Aula 3 – Como Comparar Sequências ?.....	13
Aula 4 – Alinhar Sequências com o Geneious.....	17
Aula 5 – Análise de Eletroferogramas.....	20
Formadores .....	24

## Objetivo Geral do Curso

O curso tem como objetivo ensinar de forma prática e intuitiva conceitos básicos da utilização de sequências de DNA, RNA e proteínas no âmbito de estudos de biologia molecular, genética, genómica, ecologia, biodiversidade e taxonomia.

Os alunos irão aprender a utilizar vários programas de computador, sites da internet e bases de dados amplamente utilizados na investigação científica, no mundo empresarial e em entidades prestadoras de serviços e diagnósticos.

O curso terá uma forte componente prática, sendo que as análises são baseadas em casos de investigação reais. No entanto, serão fornecidos conceitos básicos e avançados de biologia molecular aplicados a diferentes áreas de estudo.

## Público-alvo

Estudantes, licenciados, MSc, PhD, pós-docs, professores ou profissionais de áreas relevantes relacionadas com a implementação da DQEM ou na gestão sustentável do meio marinho.

## Objetivos Específicos do Curso

Adquirir a capacidade de mobilizar recursos bioinformáticos na resposta a um determinado problema, sendo capaz de escolher, aplicar e avaliar criticamente as técnicas mais adequadas.

Melhorar a capacidade de recolha e análise crítica da informação científica relacionada com técnicas moleculares.

Aprender a utilizar as principais bases de dados públicas com informação resultante de técnicas moleculares.

Conhecer os principais programas de computador utilizados na organização e armazenamento de dados moleculares.

Adquirir conhecimentos básicos na interpretação e análise dos dados genéticos obtidos em espécies marinhas.

Obter competências no alinhamento e anotação de sequências de DNA, RNA e proteínas.

Aprender a construir árvores filogenéticas tendo em vista a identificação de espécies e a determinação de índices de biodiversidade.

Integrar os conhecimentos adquiridos numa perspetiva global da gestão sustentável dos recursos biológicos marinhos utilizando dados moleculares.

Compreender os objetivos gerais da Diretiva Quadro Estratégia Marinha e a aplicação das técnicas moleculares no âmbito da sua implementação.

# Métodos de Ensino e Atividades de Aprendizagem

Curso lecionado em formato e-learning. O curso incluirá vários exercícios com tutoria (sem aulas virtuais), onde os participantes trabalharão com diferentes programas de análise de dados e bases de dados disponíveis na internet de forma gratuita.

(caso deseje realizar o curso, contacte [fpereirapt@gmail.com](mailto:fpereirapt@gmail.com))

## Língua de Trabalho

O curso será ministrado em português. Os programas e páginas da internet utilizados estão em inglês, mas as definições e opções mais relevantes serão traduzidas.

## Componentes de Avaliação

Exame online com questões de escolha múltipla no final de cada aula. O aluno tem acesso ao resultado após terminar cada exame.

Fórmula de cálculo da classificação final: Somatório das classificações obtidas em cada aula, ponderados para uma escala de 0 a 20.

# AULA 1 - COMO IDENTIFICAR UMA SEQUÊNCIA DE DNA DESCONHECIDA

**Objetivo:** Aprender a identificar uma sequência de DNA por comparação com sequências disponíveis em bases de dados públicas. Utilizar o sítio da Internet 'Sequence Manipulation Suite' para obter estatísticas básicas de uma sequência de DNA. Adquirir noções básicas da utilização do programa Basic Local Alignment Search Tool (BLAST). Saber interpretar os resultados obtidos com este programa e as possíveis implicações biológicas.

**Tempo previsto de duração da aula:** 2 a 3 horas

**Língua:** A aula é ministrada em português. Os programas e páginas da internet utilizados estão em inglês, mas as definições e opções mais relevantes serão traduzidas.

**Pré-requisitos académicos:** Noções básicas de biologia molecular.

**Pré-requisitos informáticos:** Conhecimentos básicos de navegação na internet e informática na ótica do utilizador. Versões recentes do sistema operativo e navegador da internet.

**Componentes de avaliação:** Exame online com questões de escolha múltipla. O aluno tem acesso ao resultado após terminar cada exame.

## Como identificar uma sequência de DNA desconhecida?

Imagine que no decorrer de uma investigação científica, obtém uma sequência de DNA sobre a qual não tem informação nenhuma.

- Como podemos saber do que se trata?
- Que tipo de informação poderemos obter?
- Esta sequência é nova ou já foi identificada por outro investigador?
- Em que local do genoma se encontra?
- É um gene?
- Que informação biológica posso obter?
- Existem estudos científicos relacionados com esta sequência?



Estas são algumas das questões que irá aprender a responder.

A sequência de DNA que obteve no seu estudo é a seguinte:

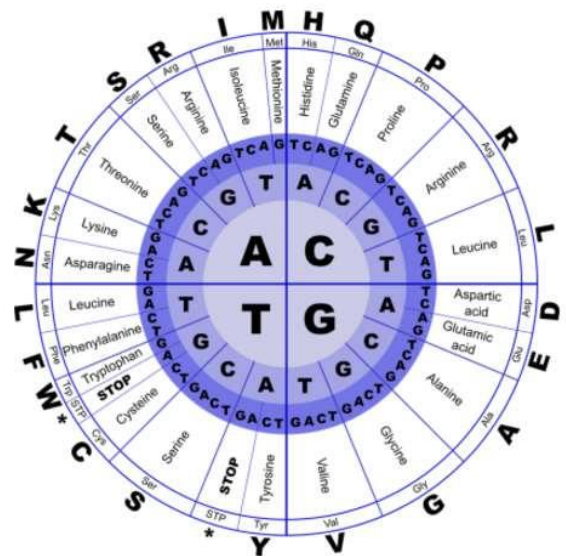
```
TGTTTCTGCTGATGCCAGTCTTACTGGTCAGTTGTTTTCTGAGCCAAGGGGCAGCGATG  
GAAACCAACGGCTCTTCAACATCGCGGTCAACCGGGTGCAACATCTCCACCTAATGGC  
TCAGAAGATGTTCAATGACTTT
```

Podemos começar por obter algumas informações muito simples acerca desta sequência. Como sabe, o DNA é formado por quatro tipos de nucleótidos (ou bases) identificados pelas letras

- **A** (adenina)
- **T** (timina)
- **G** (guanina)
- **C** (citosina)

Vamos contabilizar o número total de nucleótidos (e de cada tipo) que esta sequência tem. Pode contar à mão, mas para evitar erros e poupar tempo (e porque poderá ter que trabalhar com sequências muito maiores) vamos utilizar o seguinte sítio da internet:

[Sequence Manipulation Suite](#).



Seria interessante poder **comparar** a nossa sequência com outras obtidas noutros estudos. Talvez desta forma a conseguíssemos identificar. Talvez esta sequência já tenha sido identificada por outros investigadores.

Na verdade, este tipo de **comparação é feito regularmente pelos investigadores**.

Vamos utilizar a ferramenta **Basic Local Alignment Search Tool (BLAST)**.

## O que é o BLAST?

O **BLAST** encontra regiões de similaridade entre sequências.

O programa **compara sequências** de nucleótidos ou proteínas contra bases de dados de sequências e calcula a **significância estatística** dos resultados.

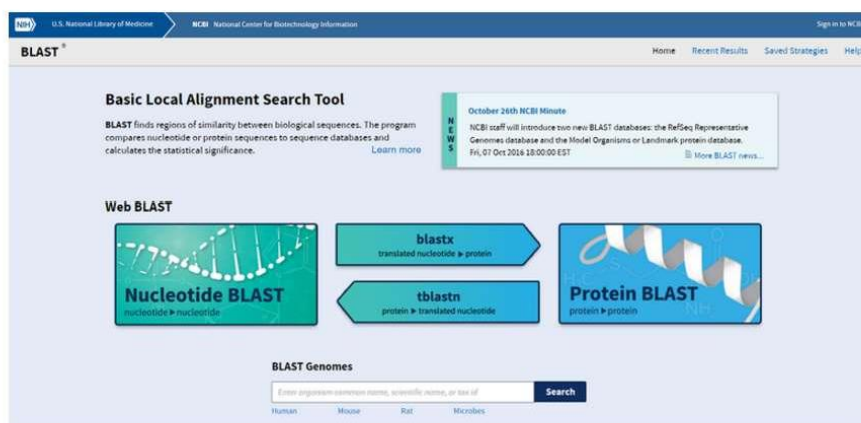
O **BLAST** pode ser utilizado para:

- identificar sequências desconhecidas ou novas
- inferir relações funcionais e evolutivas entre sequências
- identificar membros de famílias de genes.

O resultado de uma pesquisa **BLAST** será um **conjunto alinhado de sequências potencialmente relacionadas**, classificadas de acordo com a sua **semelhança**.

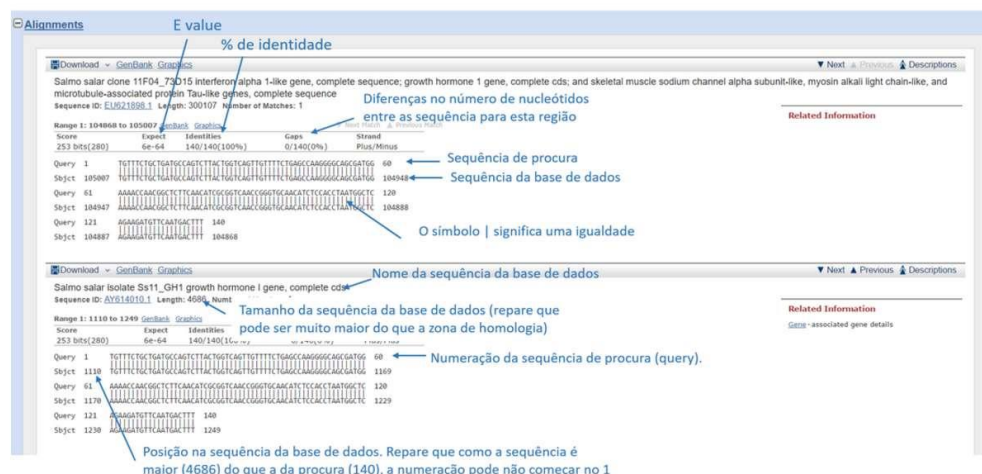
Dito por outras palavras, o **BLAST** funciona no mundo das sequências de DNA como o Google funciona nas páginas da internet. Utilizando um texto de procura ('query'), identificamos os textos mais parecidos com a nossa busca.

- Comece por abrir a página da internet do programa **BLAST**



## Alinhamentos e Resultados

Analise atentamente a figura e visualize o vídeo abaixo.





## AULA 2 - PESQUISAS NO BLAST UTILIZANDO O GENEIOUS

**Objetivo:** Nesta aula irá ter aprofundar os conhecimentos sobre o programa BLAST, aprender como realizar pesquisas BLAST individuais e de grupo com o programa Geneious e conhecer algumas opções do programa.

**Tempo previsto de duração da aula:** 1 a 2 horas.

**Língua:** A aula é ministrada em português. Os programas e páginas da internet utilizados estão em inglês, mas as definições e opções mais relevantes serão traduzidas.

**Pré-requisitos acadêmicos:** Noções básicas de biologia molecular.

**Pré-requisitos informáticos:** Conhecimentos básicos de navegação na internet e informática na óptica do utilizador. Versões recentes do sistema operativo e versão Pro do Geneious. Veja a informação sobre a utilização do Geneious aqui.

**Componentes de avaliação:** Exame online com questões de escolha múltipla. O aluno tem acesso ao resultado após terminar cada exame.

Imagine que obteve uma sequência de uma proteína como resultado de uma experiência e gostaria de descobrir do que se trata. Na aula anterior aprendeu a utilizar a ferramenta Basic Local Alignment Search Tool (BLAST) para identificar uma sequência de DNA desconhecida, utilizando diretamente o site do programa. Poderá fazer uma procura semelhante com uma sequência de proteína. No entanto, é possível aceder ao BLAST utilizando outros programas que servem de intermediários. Nesta aula vamos aprender a utilizar um destes programas - o Geneious.

Para executar uma pesquisa BLAST com uma única sequência, **baixe o arquivo "p00656.geneious"**, guarde-a no seu computador e importe-a para o Geneious (arraste-a para o programa ou vá a File, Import, From file...).



Depois clique no botão **BLAST** na barra de ferramentas. Esta operação irá abrir a caixa de diálogo do BLAST.



#### DNA query:

- **Megablast** - rápido, mas só origina sequências muito semelhantes.
- **Discontiguous megablast** - mais sensível, permite obter sequências menos semelhantes, e pode ser configurado para ignorar determinados tipos de bases.
- **blastn** - mais lento, mas mais sensível, permite obter sequências pouco semelhantes. É a melhor opção para as espécies mais distantes.
- **blastx** - traduz a sequência de procura para proteína e procura na base de dados de aminoácidos.

#### Protein query:

- **blastp** - compara sequências de proteínas com a base de dados de proteínas
- **tblastn** - compara sequências de proteínas com as 6 traduções possíveis das sequências da base de dados de nucleótidos.

Como a nossa sequência 'P00656' é uma sequência de proteína, o programa seleciona a base de dados de proteínas do **NCBI nr** e a opção **blastp** que irá realizar uma procura rápida de sequências proteicas.

Confirme que a opção '**Results**' está selecionada para **Hit Table** e '**Retrieve**' para **Matching Region**.

Esta configuração irá originar uma lista dos resultados (hits) mais significativos e um alinhamento de cada um deles, para além de um alinhamento do tipo 'query-centric'.

Deixe as outras configurações inalteradas e, em seguida, clique no botão **Search**.

O **Geneious** irá enviar a sua informação de consulta para o **NCBI** e criar uma nova pasta de pesquisa, que será exibida como uma subpasta da pasta que contém a sequência utilizada na procura.



O nome da pasta será o da sequência utilizada para a procura, a base de dados pesquisada, o programa utilizado para realizar a pesquisa e o número de resultados obtidos entre parênteses.

Quando a procura for concluída, o Geneious irá recolher todos os resultados do NCBI e colocá-los na pasta recém-criada.

Por predefinição, os resultados da pesquisa devem ser ordenados pelo valor **E ('E Value')** que indica a frequência esperada de ocorrência de cada alinhamento por acaso. Lembre-se que **quanto menor o valor E, melhor**.

Se os resultados não estiverem ordenados pelo valor E, clique no cabeçalho dessa coluna.

O resultado da procura deve ficar parecido com a tabela descrita em baixo, mas os resultados podem variar ligeiramente à medida que novas sequências são adicionadas ao Genbank.

Hit Table									
Bit-Score	E Value	% Pairwise Id...	Grade	Name	Description	Sequence L...	Hit start	Hit end	Info
311.612	1.18e-107	100.0%	100.0%	NP_001014408	ribonuclease pancreatic precursor [Bos t...	150	1	150	
313.153	2.63e-107	100.0%	100.0%	CDG32088	TPA: ribonuclease A C2 [Bos taurus]	150	66	215	
310.071	3.83e-107	99.3%	99.7%	XP_005901936	PREDICTED: ribonuclease pancreatic [Bo...	150	1	150	
270.396	9.61e-92	100.0%	92.7%	1C0B_A	Chain A, Bovine Pancreatic Ribonuclease ...	128	1	128	
269.24	7.28e-91	86.1%	91.0%	AAA72757	RNase A [synthetic construct]	151	6	156	
266.929	4.03e-90	91.3%	95.7%	XP_005960934	PREDICTED: ribonuclease pancreatic-like...	150	1	146	
261.922	2.22e-88	100.0%	91.3%	P61824	RecName: Full=Ribonuclease pancreatic...	124	1	124	
261.922	2.40e-88	100.0%	91.3%	CAB37066	artificial [synthetic construct]	124	5	128	
261.536	2.65e-88	100.0%	91.3%	CAA33801	unnamed protein product [Bos taurus]	124	2	125	
260.381	8.76e-88	99.2%	90.9%	1EIE_A	Chain A, Crystal Structure Of F120w Muta...	124	1	124	
260.381	9.25e-88	99.2%	90.9%	3DH6_A	Chain A, Crystal Structure Of Bovine Panc...	124	1	124	
260.381	9.25e-88	99.2%	90.9%	3DI7_A	Chain A, Crystal Structure Of Bovine Panc...	124	1	124	
260.381	9.25e-88	99.2%	90.9%	3DI8_A	Chain A, Crystal Structure Of Bovine Panc...	124	1	124	
260.381	9.25e-88	99.2%	90.9%	3DIC_A	Chain A, Crystal Structure Of Bovine Panc...	124	1	124	
260.381	9.89e-88	99.2%	90.9%	4WYN_A	Chain A, The Crystal Structure Of The A1...	124	2	125	
265.002	1.11e-87	100.0%	91.7%	3MVR_A	Chain A, Crystal Structure Of Ribonuclease	125	120	254	

Tendo em conta o que aprendeu na aula anterior sobre os resultados do BLAST, pode verificar que a sua sequência pertence a uma ribonuclease pancreática de bovino. Caso esteja interessado, pode obter mais informações sobre esta enzima no site da [UniProt](#).

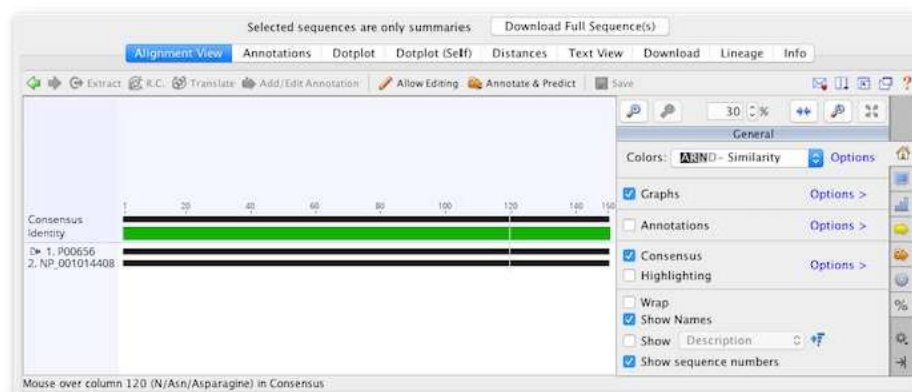
Para além do valor E, há também uma coluna chamada **% Pairwise Identity**. Esta coluna indica o grau de semelhança entre a sequência encontrada na base de dados e a sequência utilizada na procura. É possível ver que muitos dos resultados obtidos neste exemplo são 100% idênticos à sequência utilizada na procura, mas têm diferentes **Tamanhos de sequência ('Sequence Lengths')**. Este valor resulta do facto do alinhamento ser do tipo **'local similarity alignment'** que alinhou a região de tamanho maior existente entre duas sequências. A identidade refere-se apenas à região alinhada, sendo que é possível ter alinhamentos muito curtos com elevada percentagem de identidade. É por este motivo que os alinhamentos tendem a ser classificados pelo seu valor de E, em vez do valor de identidade ('% Pairwise Identity').

O programa Geneious também fornece um valor de **Grade**, que combina os valores de **'query coverage'** (a região da sequência utilizada onde se encontrou uma identidade), o **valor E** e o valor de identidade ('% Pairwise Identity') para cada resultado com pesos de 0,5, 0,25 e 0,25, respectivamente.

**Este valor permite identificar os resultados com maior homologia.**

Agora que se obteve um conjunto de resultados, será possível analisar alguns alinhamentos.

Clique no resultado obtido para **NP\_001014408** de forma a ver algo deste tipo:



Pode ver no gráfico de identidade por cima do alinhamento que as duas sequências são **idênticas (cor verde)**.

Como qualquer outro alinhamento no Geneious, pode aplicar zoom para ver a sequência, alterar as configurações de cor, e destacar as regiões iguais ou diferentes do consenso utilizando o menu no lado direito do visualizador.

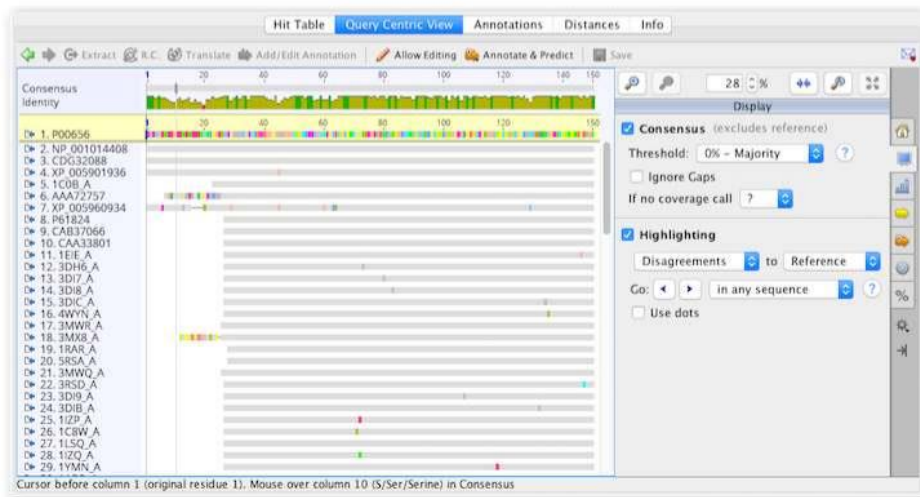
Esta vista de alinhamento só mostra a região de alinhamento entre a sequência utilizada na procura e a obtida no BLAST. O documento BLAST obtido é um resumo e não contém o registo completo do Genbank para essa sequência. Para obter a sequência completa e respetivas anotações, clique em **Download Full Sequence(s)**. Quando a sequência completa for obtida, um separador **Sequence View** é adicionado ao programa.

Este procedimento mostra a sequência completa e anotada, com uma nova anotação "BLAST Hit" que indica qual a região da sequência obtida que coincide com a da sequência utilizada na procura.

## Visualização centrada na sequência de procura ('Query-centric view')

A 'Query-centric view' é útil para visualizar todos os resultados obtidos contra a sua sequência de procura numa única janela, permitindo identificar as regiões mais conservadas.

- Clique no separador **Query Centric View** no topo da 'Hit table'.
- Retire as anotações no separador 'Annotations and Tracks'
- Escolha highlight 'Disagreements to Reference' no separador 'Display'
- Deverá ver algo deste tipo:



## AULA 3 - COMO COMPARAR SEQUÊNCIAS?

**Objetivo:** Aprender a alinhar duas sequências de DNA, RNA ou proteínas. Identificar regiões conservadas e variáveis no alinhamento. Aprender a traduzir sequências de RNA para proteína. Ficar familiarizado com o programa T-Coffee e a ferramenta de tradução do ExPASy. Identificar diferentes tipos de polimorfismo. Saber interpretar os resultados obtidos com este programa e as possíveis implicações biológicas.

**Tempo previsto de duração da aula:** 2 a 3 horas

**Língua:** A aula é ministrada em português. Os programas e páginas da internet utilizados estão em inglês, mas as definições e opções mais relevantes serão traduzidas.

**Pré-requisitos académicos:** Noções básicas de biologia molecular.

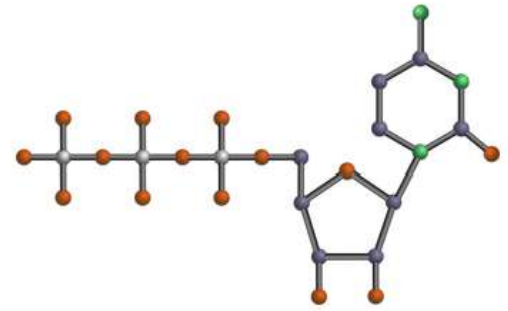
**Pré-requisitos informáticos:** Conhecimentos básicos de navegação na internet e informática na ótica do utilizador. Versões recentes do sistema operativo e navegador da internet.

**Componentes de avaliação:** Exame online com questões de escolha múltipla. O aluno tem acesso ao resultado após terminar cada exame.



Imagine que trabalha com um gene humano e um colega investigador envia-lhe uma **sequência de RNA** de talvez pertença ao mesmo gene.

- Serão as sequências muito parecidas?
- Podem ser consideradas do mesmo gene?
- Que tipo de diferenças possuem?
- Qual o grau de conservação entre elas?



Estas são algumas das questões que irá aprender a responder.

Vamos começar por fazer um **alinhamento** entre as duas sequências.

## Mas o que é um alinhamento?

Um alinhamento serve para **organizar sequências** (DNA, RNA ou proteína) de forma a identificar **regiões de homologia** que possam estar relacionadas do ponto de vista funcional, estrutural e evolutivo. O alinhamento é feito comparando cada 'letra' de uma sequência com cada 'letra' da outra. Desta forma são identificadas letras iguais e diferentes entre as sequências. Num alinhamento, cada linha representa uma sequência. As diferentes sequências estão sobrepostas, sendo que as regiões com 'letras' iguais com suposta relação de homologia estão colocadas na mesma coluna.

Por exemplo:

```
TCTGGCCTTAAA  
TCTAGCCTCAAA
```

O algoritmo de alinhamento também considera **inserções ou deleções** em ambas as sequências, ou seja a existência de 'letras' a mais ou a menos numa das sequências.

Neste caso, as duas sequências estarão alinhadas inserindo interrupções ('gaps') indicados por um traço '-':

**Antes do alinhamento**

```
TCTGGCCTTAAA  
TCTCCAAA
```

**Depois do alinhamento**

```
TCTGGCCTTAAA  
TCT--CC--AAA
```

Vamos então alinharmos as nossas duas sequências chamadas “RNA1” e “RNA2”.

Existem vários programas de alinhamento de sequências, muitos deles com características diferentes. É recomendado utilizar vários tipos de alinhamentos e comparar os resultados obtidos.

## Programa T-Coffee

O objetivo do programa T-Coffee é construir um alinhamento que tenha um alto nível de consistência. O programa constrói uma **biblioteca de alinhamentos** com pares de sequências. Esta biblioteca pode incluir vários alinhamentos alternativos das mesmas sequências feitos com diferentes parâmetros. No geral, a biblioteca é uma **coleção de alinhamentos** que se acreditam estarem corretos. Dentro desta biblioteca, cada alinhamento recebe um peso que é uma estimativa da sua probabilidade biológica, ou seja, até que ponto temos confiança que este alinhamento possa estar correto.

O algoritmo T-Coffee utiliza esta biblioteca para estimar um valor para cada dois resíduos alinhados num alinhamento. Este valor tem em consideração a compatibilidade do alinhamento dos dois resíduos com o resto dos alinhamentos observados dentro da biblioteca. Ou seja, este valor tem em conta a frequência do alinhamento desses dois resíduos em toda a biblioteca. **O alinhamento de dois resíduos é tanto mais relevante quanto mais frequente for na biblioteca do T-Coffee.**

O programa T-Coffee estima ainda o chamado ‘**índice CORE**’ que indica a média dos valores de consistência na biblioteca para os pares de resíduos que envolvem cada resíduo numa coluna do alinhamento. O programa atribui uma numeração e cor a cada resíduo, como descrito mais à frente na aula.

Vamos abrir a página da internet chamada ‘[T-Coffee](#)’.

Conforme explicado anteriormente, o programa T-Coffee atribui uma cor aos resíduos de acordo com um valor de consistência dos resíduos alinhados (**índice CORE**). A cada resíduo é atribuído um valor arredondado entre 0 e 9, que surgem coloridos nalguns ficheiros de resultados (pdf, postscript ou html) sendo:

- azul / verde para resíduos com **baixa pontuação**;
- laranja / vermelho para os que apresentam **maior pontuação**.

**Quanto mais elevado o valor do índice CORE, maior a probabilidade de ele estar bem alinhado tendo em conta a biblioteca gerada pelo programa.**

**BAD AVG GOOD**

Neste caso, com apenas duas sequências essa informação não é muito relevante. De qualquer forma, as posições com as duas bases iguais estão coloridas a vermelho, enquanto as posições onde existem diferenças têm outras cores.

Pode ver que as sequências diferem nalgumas posições no tipo de nucleótido (por exemplo, uma sequência tem um A e a outra um C). Em alguns contextos, esta diferença é chamada **polimorfismo de nucleótido único (single nucleotide polymorphism - SNP)**.

RNA1	:	98
RNA2	:	97
cons	:	96
RNA1	A - - UGCAAUCAUAUGCUUCUGCUAUGGUAAGCGUAUUAACAGC	
RNA2	AUGGUGCAGUCAUAUGCUUCUGCUAUGGUCAGAGUAUUGAAAGCG	
cons	* **** *	
RNA1	GAUGAUUACAGUCCAGCUGUGCAA - - GAGAAUAUUCGCCGUCUC	
RNA2	GACGAUACAGCCAGCGGCACAGCAGCAAAUAUUCUGGCCUUG	
cons	** *	
RNA1	CGGAGAAGCUCUCCUCCUUGCACUGAAAGCUGUAACUCUAAG	
RNA2	GGGAAAGGCUCUCACUAUUUCCGACGGACAUAUAGCUCUAAAC	
cons	*** *	
RNA1	UAUCAGUGUGAAACGGGAGAAAAAGUAAGGCAACGUCCAGGAU	
RNA2	GAUGGACGUGAAACUAGAGGAAGUGGUAGAGAGAGUGGCCAGGAU	

Noutras posições do alinhamento, as sequências diferem pela ausência ou presença de nucleótidos (por exemplo, uma sequência tem um G e a outra não tem nenhum resíduo nesse local). Em alguns contextos, esta diferença é chamada **polimorfismo de inserção/deleção** (**insertion/deletion polymorphism – indel**).

The diagram shows a 3D representation of an RNA structure. A blue box highlights a specific region of the structure, which corresponds to the sequence GAUAAUUAUUCGCGCUCUC GACGAUUACAGCCACAGCGGACAGCAGC in the sequence below. The structure is composed of various colored blocks representing different regions of the RNA molecule.

RNA1 : 98  
RNA2 : 97  
cons : 96

RNA1 A - - UGCAAUCAUAGCUUCUGCUAUGUUAAGCGUAUUCAACAGC  
RNA2 AUGGUGCAGUCAUAGCUUCUGCUAUGUUCAGAGUAUUGAAAGCG

cons \*

RNA1 GAUGAUUACAGUCCAGCUGUGCAA - - GAGAAUUAUUCGCGCUCUC  
RNA2 GACGAUUACAGCCACAGCGGACAGCAGC AAAAUUUCUGGCCUUG

cons \*

RNA1 CGGAGAAGGCUCUUCUUCUUUCCACUGAAAGCUGUAACUCUAA  
RNA2 GGGAAAGGCUCUCACUAUUUCCAGCGGACAAUCAUAGCUCAAAC

cons \*

RNA1 UAUCAGUGUGAAACGGGAGAAAACAGUAAAGGCACGUCACAGGAU  
RNA2 GAUGGACGUGAAACUAGAGGAAGUGGUAGAGAGAGUGGCCAGGAU



## AULA 4 - ALINHAR SEQUÊNCIAS COM O GENEIOUS

**Objetivo:** Nesta aula terá uma breve introdução aos algoritmos de alinhamento de sequências, irá desenvolver as capacidades necessárias para alinhar sequências de DNA e proteínas com o Geneious e aprender a utilizar 'dotplots' para explorar relações entre sequências. Saber interpretar os resultados obtidos com este programa e as possíveis implicações biológicas.

**Tempo previsto de duração da aula:** 2 a 3 horas.

**Língua:** A aula é ministrada em português. Os programas e páginas da internet utilizados estão em inglês, mas as definições e opções mais relevantes serão traduzidas.

**Pré-requisitos académicos:** Noções básicas de biologia molecular.

**Pré-requisitos informáticos:** Conhecimentos básicos de navegação na internet e informática na óptica do utilizador. Versões recentes do sistema operativo e versão Pro do Geneious. Veja a informação sobre a utilização do Geneious aqui.

**Componentes de avaliação:** Exame online com questões de escolha múltipla. O aluno tem acesso ao resultado após terminar cada exame.

O objetivo deste módulo é desenvolver as capacidades necessárias para alinhar pares de sequências com o Geneious.

O alinhamento de pares de sequências ('Pairwise sequence alignment') permite encontrar regiões idênticas em sequências para identificar prováveis semelhanças estruturais e funcionais.



Neste módulo, vamos alinhar sequências de nucleotídeos (DNA) e sequências de polipéptidos (proteína) usando um método de alinhamento global (Needleman e Wunsch) e local (Smith e Waterman).

## Algoritmos de alinhamento

Ao alinhar duas sequências, o algoritmo identifica a relação ótima entre elas. Isto é feito através da comparação de cada 'letra' de uma sequência com cada 'letra' de outra. O algoritmo tem em consideração igualdades ('match') e diferenças ('mismatch') e calcula o melhor caminho matemático através destes emparelhamentos. O algoritmo também considera inserções ou deleções em ambas as sequências. As duas sequências seguintes podem ser alinhadas inserindo interrupções ou hiatos ('gaps', indicados por '-') para alinhar resíduos idênticos:

### Antes do alinhamento

AFGIVHKLIVS

AFGIHKIVS

### Depois do alinhamento

A F G I V H K L I V S

A F G I - H K - I V S

O alinhamento de pares de sequências ('Pairwise sequence alignment') utilizado no Geneious é baseado em [programação dinâmica](#) utilizando os algoritmos [Needleman & Wunsch \(1970\)](#) ou [Smith & Waterman \(1981\)](#). Estão disponíveis em variantes **global** e **local**. Um alinhamento global garante que todas as regiões de duas sequências são alinhadas. Um alinhamento local irá alinhar as áreas de melhor similaridade, como se verifica quando apenas uma parte das duas sequências está relacionada (por exemplo, em sequências de proteínas multi-domínio). Forçar um alinhamento global de uma sequência de múltiplos domínios não seria razoável, visto que o alinhamento implica que exista uma semelhança entre as sequências em todo o seu comprimento, sendo que neste caso partes das sequências não estariam relacionadas.

**Nota:** Um alinhamento pode ser matematicamente ótimo, mas não necessariamente ser biologicamente relevante, como será demonstrado nos exercícios seguintes.

## Exercício 1: Utilizar 'dotplots' para explorar relações entre sequências

Os dotplots são uma excelente forma de visualizar as regiões de similaridade entre pares de sequências que não podem ser identificados apenas pelo alinhamento de sequências. Em especial, permite identificar regiões repetidas, inversões e translocações que não são analisadas no alinhamento de sequências.

O dotplot identifica as igualdades ('match') entre duas sequências numa grelha bidimensional. Zonas contínuas de igualdade entre sequências originam linhas diagonais nesta grelha. Para ilustrar isto, são fornecidas duas sequências. Uma sequência original do chimpanzé pigmeu e a mesma sequência editada para originar um dotplot interessante.



pygmy\_chimpanzee.geneious  
[Baixar o arquivo](#)



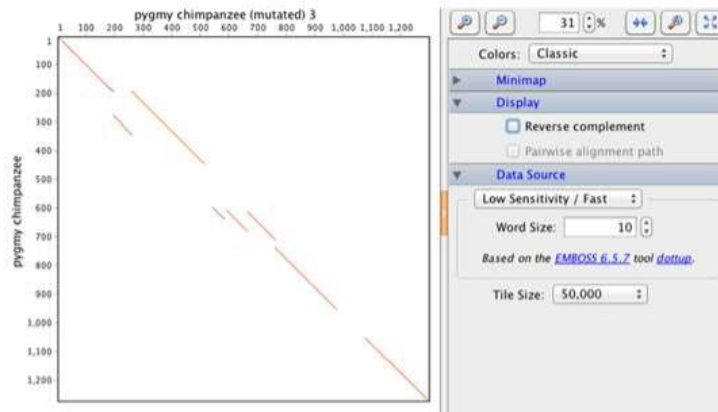
pygmy\_chimpanzee\_\_mutated\_\_3.geneious  
[Baixar o arquivo](#)

Guarde-as no seu computador e importe-as para a mesma pasta do Geneious (arraste-a para o programa ou vá a File, Import, From file...).

Pode ter que seleccionar o separador que diz **Dotplot**. Pode também abrir o gráfico uma nova janela clicando no botão



Deve conseguir ver uma grelha 2D com linhas diagonais. Uma sequência aparece na parte superior no eixo dos X e a segunda ao longo do lado esquerdo, no eixo dos Y. A tendência geral é termos uma diagonal do canto superior esquerdo para o canto inferior direito.



Pode fazer zoom in e out no dotplot, ajustando o nível de zoom. Se ampliar o suficiente, será capaz de ver as letras individuais das duas sequências em cada eixo. Além disso, observe como há uma cruz que segue o ponteiro do rato que indica a posição nas duas sequências do ponteiro. Pode usar isso para anotar a localização no dotplot das características em que está interessado, algo que o vai ajudar mais tarde quando for fazer alinhamentos.

Tente mudar os parâmetros na opção **Data Source** e veja como isso afeta a dotplot. Observe como alterar a sensibilidade aumenta ou reduz o número de linhas curtas no dotplot. Aumentar o tamanho da janela ('window size') tenderá a preencher os gaps entre as diagonais menores de forma a originar outras maiores. Reduzir a janela irá encurtar as linhas diagonais. Pode haver algum detalhe interessante a ser detectado neste caso, mas geralmente reduzir este parâmetro apenas aumenta ruído ao plot. Por isso, aumentando este valor permite ver apenas as diagonais maiores e diminuir o ruído de fundo.

As diagonais paralelas num dotplot indicam que existe um padrão repetitivo na sequência. Uma diagonal invertida indica uma inversão na sequência. Isto significa que uma sequência tem uma região que corresponde ao complemento inverso de uma parte da outra sequência.

Agora que está familiarizado com a forma como estas duas sequências estão relacionados, podemos começar a alinhá-las.

## AULA 5 - ANÁLISE DE ELETROFEROGRAMAS

**Objetivo:** Nesta aula irá utilizar sequências Sanger e aprenderá como editar e alinhar eletroferogramas. A aula abrange o corte de sequências de baixa qualidade, a edição de sequências a partir de alinhamentos, a identificação de heterozigóticos e de bases incorretas, e a construção de sequências de consenso a partir de sequências diretas (forward) e reversas (reverse) do mesmo gene.

**Tempo previsto de duração da aula:** 1 a 2 horas.

**Língua:** A aula é ministrada em português. Os programas e páginas da internet utilizados estão em inglês, mas as definições e opções mais relevantes serão traduzidas.

**Pré-requisitos acadêmicos:** Noções básicas de biologia molecular.

**Pré-requisitos informáticos:** Conhecimentos básicos de navegação na internet e informática na óptica do utilizador. Versões recentes do sistema operativo e versão Pro do Geneious. Veja a informação sobre a utilização do Geneious [aqui](#).

**Componentes de avaliação:** Exame online com questões de escolha múltipla. O aluno tem acesso ao resultado após terminar cada exame.



Neste tutorial, irá utilizar dados em bruto de sequenciação Sanger e aprenderá como editar e alinhar eletroferogramas para análises posteriores, como por exemplo, para construir uma árvore filogenética ou calcular a diversidade nucleotídica. O tutorial abrange o corte de sequências de baixa qualidade, a edição de sequências a partir de alinhamentos, a identificação de heterozigóticos e de bases incorretas, e a construção de sequências de consenso a partir de sequências diretas (forward) e reversas (reverse) do mesmo gene.



**No Exercício 1**, irá editar e alinhar um conjunto de sequências de DNA mitocondrial do chapim azul *Cyanistes teneriffae*.

**No Exercício 2**, irá editar e juntar as sequências diretas e reversas de um gene nuclear de três espécies de rouxinol.

Este tutorial requer que instale o plugin Heterozygotes no Geneious. Para o instalar, vá a **Tools->Plugins**, encontre-o na lista de plugins disponíveis e clique em **Install**.

## Sequências de DNA mitocondrial - Introdução

O complexo de espécies do chapim azul inclui *C. caeruleus*, presente por toda a Europa, *C. teneriffae*, encontrado no Norte de África e nas Ilhas Canárias, e *C. cyanus*, presente na Ásia e Europa de leste. Os dados do DNA mitocondrial podem ser utilizados para investigar a filogeografia e a estrutura populacional destas espécies.

O conjunto de dados aqui proporcionado compreende 34 sequências da região de controlo do DNA mitocondrial de *C. caeruleus* e *C. teneriffae*. A sequência do chapim-real *Parus major* também é incluída, uma vez que este servirá de outgroup para a análise filogenética.

A tabela abaixo fornece a localização da amostragem e os códigos para as sequências deste tutorial

Código	Espécie	Origem
CEH	<i>C. teneriffae</i>	Ilhas Canárias - El Hierro
CFU	<i>C. teneriffae</i>	Ilhas Canárias - Fuerteventura
CGC	<i>C. teneriffae</i>	Ilhas Canárias - Gran Canaria
CLG	<i>C. teneriffae</i>	Ilhas Canárias - La Gomera
CLP	<i>C. teneriffae</i>	Ilhas Canárias - La Palma
CLA	<i>C. teneriffae</i>	Ilhas Canárias - Lanzarote

## Editar sequências de DNA mitocondrial



Lista de Sequências  
Baixar o arquivo

- Selecione a **lista de sequências** contendo os dados em bruto da sequenciação da região de controlo do DNA mitocondrial.
- Guarde-a no seu computador e importe-a para o Geneious (arraste-a para o programa ou vá a File, Import, From file...). Clique duas vezes na lista para abri-la numa nova janela.
- No **General tab** no lado direito da zona de visualização da sequência, **escolha o display 'Colors' de acordo com 'Quality'**. Isto irá realçar as bases de acordo com a qualidade da sequência - quanto mais escuro for o azul, menor será a qualidade.

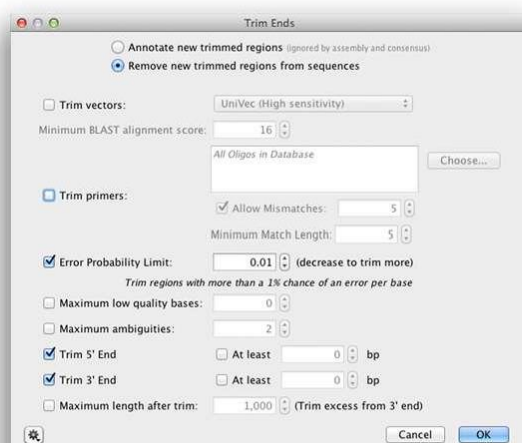
Quando o zoom não estiver selecionado, **não verá as bases individuais ou os picos do eletroferograma**, mas verá um gráfico com uma indicação da qualidade da sequência. Se percorrer as sequências verá que a qualidade diminui drasticamente no final de cada sequência.

- **Faça zoom** de pelo menos 50% para ver como os eletroferogramas são em regiões de qualidade boa versus pobre.

Uma das amostras (**CLG3**) não tem sequência, indicando que a reação de sequencição falhou, por isso **apague-a desta lista**.

A amostra **SRE1** tem apenas uma curta região de sequência de boa qualidade antes da se tornar ilegível, por isso **apague também esta amostra da lista**.

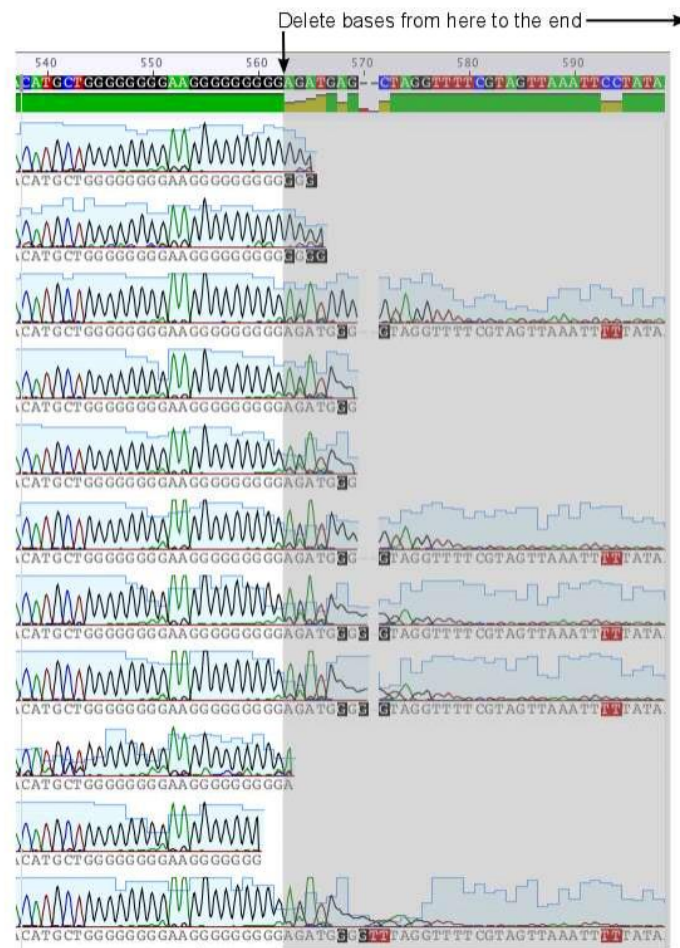
- Clique em **Save** para guardar as alterações à lista de sequências e feche a janela.
- **Corte ('trim')** as regiões terminais das sequências com má qualidade clicando em **Annotate and Predict**→**Trim Ends**.
- Escolha "**Remove new trimmed regions from sequences**"
- Selecione a 'Error probability limit' para 0.01, como na imagem em baixo.
- Clique **OK** e depois em **Save** quando o trimming tiver terminado.



A partir daqui, é mais eficiente limpar e editar as sequências depois de alinhadas.

- **Selecione** a lista de sequências novamente e clique **Align/Assemble**→**Multiple Align**.
- **Escolha** o algoritmo de alinhamento **MUSCLE** e utilize-o com os parâmetros predefinidos.
- **Dê um clique duplo** no alinhamento para abri-lo e faça zoom a cerca de 50% para que possa ver as bases e o eletroferograma.
- Pode precisar de escolher **Show Graphs** no separador **Graphs** para ver os eletroferogramas.
- **Percorra as bases até à região a 3' (lado direito)** e vai verificar que os picos do eletroferograma ficam mais fracos depois do motivo GGGGGGGAAGGGGGGGG (ver imagem em baixo). Em muitas das sequências, a região que se segue a este motivo já está cortada.
- **Corte** as sequências restantes clicando **Allow Editing** e selecionando as base a partir da posição 563 da sequência **consensus** e clicando no botão **delete**.

A edição da sequência de consenso irá ser aplicada a todas as sequências do alinhamento. Deve também **excluir as primeiras 20 bases no início do alinhamento** para que as sequências fiquem todas com o mesmo comprimento, uma vez que esta região já foi cortada em várias sequências.



- Clique **Save** e escolha **Yes** quando é perguntado se pretende que estas alterações sejam aplicadas a todas as sequências ('apply the changes to the original sequences'). Repare que por vezes é preferível **não** aplicar as alterações a todas as sequências, caso pretenda manter o ficheiro original sem alterações.

Este alinhamento pode agora ser usado para construir uma árvore filogenética usando a opção **Tree** no Geneious.

Para obter mais informações sobre a construção e interpretação de árvores filogenéticas, consulte os tutoriais do Geneious sobre este tema disponíveis [aqui](#).

### **Formador Responsável**

#### **Filipe Pereira**

Licenciado em Biologia pela Faculdade de Ciências da Universidade do Porto (FCUP). Desenvolveu o seu trabalho de doutoramento e pós-doutoramento no Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP) e no Instituto Smurfit, Trinity College, Dublin, Irlanda. Desde 2013 é Investigador FCT (Fundação Para a Ciência e a Tecnologia) no Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR). A sua investigação foca-se no estudo da molécula de DNA com recurso a métodos computacionais e bioquímicos, abrangendo as áreas da genética, biologia molecular, ecologia e biotecnologia. Uma descrição detalhada da sua investigação pode ser encontrada em <http://fpereira.portugene.com/>.

### **Desenvolvimento**

#### **Filipe Lopes**

Licenciado em Biologia (ramo educacional) pela Faculdade de Ciências da Universidade do Porto (FCUP). Efetuou o mestrado em educação no Instituto de Educação da Universidade do Minho sendo a dissertação sobre a utilização dos media no ensino Das ciências naturais e na abordagem de temáticas ambientais em contexto de sala de aula. Atualmente é bolseiro de gestão de ciência e tecnologia no ICVS | Escola de Medicina da Universidade do Minho. Portfolio: [filipelopes.weebly.com](http://filipelopes.weebly.com)