Docente: Filipa Monteiro; Responsável pela UC: Maria Manuel Romeiras

Exercício: Identificação de espécies vegetais usando o BLAST

Contextualização

BLAST (**B** asic **Local** Alignment **Search Tool**) **é uma ferramenta** de busca online fornecida pelo NCBI (National Center for Biotechnology Information). Ela permite "encontrar regiões de similaridade entre sequências" (nucleotídeos ou proteína). O NCBI mantém um enorme banco de dados de sequências biológicas, que compara as sequências de consulta para encontrar as mais semelhantes com o disponível na base de dados. Usando o BLAST, pode-se inserir uma sequência de DNA de interesse e pesquisar bibliotecas moleculares inteiras para determinação de sequências idênticas ou semelhantes.

Tipo Blast	Sequência de consulta	Banco de dados	Alinhamento	Usar
Blastn	nucleotídeo	nucleotídeo	nucleotídeo	sequência identidade, útil para todas as categorias de taxa
blastx	nucleotídeo (traduzido para proteína)	proteína	proteína	Identificar proteínas codificadas, detecção de novos vírus
pblast	proteína	proteína	proteína	sequência EU IA e busca por similaridade
tblastx	nucleotídeo (traduzido para proteína)	nucleotídeo (traduzido para proteína	proteína	Sequências de nucleotídeos de ID com codificação regiões semelhantes à consulta
tblastn	proteína	nucleotídeo (traduzido para proteína)	proteína	Sequências de banco de dados de ID que codificam proteínas semelhante para a consulta

1. Escolher o tipo de BLAST baseado no objetivo. BLAST é geralmente suficiente para confirmar um taxon.

17.00	Like	bl translated nu	astx cleotide ≽ protein	000	
Nucleotide I	BLAST	tt protein > tra	anslated nucleotide	Protein BLAS	
	BLAST Genom	es			

2. Colar a sua sequência na caixa identificada caixa ou fazer upload de um ficheiro fsta com várias sequências.

			Standard Nucleotide BLA	1ST		
	blaste blasto blas	ts thiasta thiasts				
			ILASTN programs search nucleotide databases using	a nucleotide query. more	Reset page Bookma	R)
	Enter Query S	equence				-
	Enter accession n	umber(s), gi(s), or FASTA sequence(s) 🤢	Cheer Query subrange 😖			
			From	BL	AST results will be displayed	
or paste fasta file here			То	in a	new format by default	
or public fublic file fere				You	can always switch back to the	
	Or, upload file	Choose File No file chosen		Trad	Itional Results page.	
	Job Title					
		Enter a descriptive title for your BLAST search	0			
	Align two or me	re sequences 🤢				
	Choose Searc	h Set				
	Database	Standard databases (nr etc.): OrRNAT	TS databases Genomic + transcript databaser	s C Betacoronavirus		
		Nucleotide collection (nr/nt)	~ 0			
	Organism	Enter organism name or id-completions v	to successed			
	opene	Enter organism common name, binomial, or tax	at. Only 20 top taxa will be shown 😜			
	Exclude Optional	Models (XMXP) Uncultured/environm	nental sample sequences			
	Limit to	Sequences from type material				

FASTA formatar

O formato FASTA é usado para representar sequências de nucleotídeos ou peptídeos. A primeira linha é um comentário, começando com ">" e descrevendo a sequência. Todas as linhas seguintes são a sequência, em texto simples.

Exemplo sequência de DNA no formato FASTA :

Exemplo proteína sequência em FASTA formatar:

>MCHU - Calmodulina - Humano, coelho, bovino, rato, e frango

ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA DIDGDGQVNYEEFVQMMTAK* 3. Mais abaixo na página escolher o programa para otimizar a busca. Navegar para a Seção "Seleção de Programas". Para o blastn, pode escolher entre:

- 1. Megablast lata ser usado para encontrar o melhor correspondência sequência.
- 2. Descontíguo megablast usado para encontrar mais diferente sequências.
- 3. Blastn usado para encontrar relacionado sequências de outro organismos.

Megablast é geralmente usado para o blast.

4. Clique BLAST.

Program Sele	ection
Optimize for	 Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn) Choose a BLAST algorithm (g)
BLAST	Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

5. Interpretação dos resultados do BLAST Resultado.

Os resultados do BLAST mostram todos os taxa disponíveis na base de dados com sequência semelhante sequência de consulta. As métricas e os gráficos podem ajudar a determinar a qualidade do resultado obtido

O resultado da página vai mostrar o seguindo resumo de um BLAST:

- 1. O banco de dados usado para a busca.
- 2. O comprimento da nossa sequência.
- 3. Resultados (Descrições, Gráfico Resumo, Alinhamentos, Taxonomia)

BLAST * » blast	in suite » results for RID-RYR0M8CC016	Home Recent Results Saved Strategies Help
< Edit Search	Save Search Search Summary 🛩	How to read this report? BLAST Help Videos Back to Traditional Results Page
Job Title	sample_denv2	Filter Results
RID	RYR0M8CC016 Search expires on 10-10 02:36 am Download All V	
Program	BLASTN ? Citation ~	Organism only top 20 will appear exclude
Database	nt See details 🛩 🚽	Type common name, binomial, taxid or group name
Query ID	Icl Query_52797	+ Add organism
Description	sample_denv2	Percent Identity E value Query Coverage
Molecule type	dna	
Query Length	10703 <	
Other reports	Distance tree of results MSA viewer 🔞	Filter Reset
Descriptions	Graphic Summary Alignments Taxonomy	
📥 ences p	producing significant alignments	Download 🐣 Manage Columns 🐣 Show 🗐 Feedback

5.1. Navegar para o " **Descrições** " aba que tem métricas que pode ajuda você determinar a qualidade do golpe.

Max score: a maior pontuação de hits calculada a partir de correspondências e incompatibilidades de alinhamento. Quanto maior a pontuação, melhor o alinhamento.

Total score: a soma do alinhamento de todos os segmentos da sequência. Quanto maior a pontuação, melhor o alinhamento.

Query coverage: % do comprimento contíguo que alinha com o NCBI. Uma pequena porcentagem de cobertura de consulta significa que apenas uma pequena parte da sequência está alinhada. Se houver um alinhamento com 100% de identidade e 5% de cobertura de consulta, a sequência provavelmente não pertence a esse taxon.

Valor E : o número de sucessos esperado para ser visto por acaso. Quanto mais perto para 0, o melhorar. O sucessos são automaticamente classificados por E valor (melhor para pior). Essa métrica é extremamente útil para identificar acertos reais.

- a) E valor 1e⁻⁵⁰ pequeno E valor: baixo número de acessos, mas de alta qualidade.
- b) Valor E 0.01: BLAST com valor E menor que 0,01 ainda podem ser considerados bons acertos para correspondências de homologia.
- c) Valor E 10. Valor E grande: muitos acertos, alguns de baixa qualidade. Um valor E menor que dez incluirá acertos que não podem ser considerados tão significativos quanto um valor E baixo.

Percent identity: a porcentagem de bases que são idênticas ao genoma de referência

Accession [number]: um identificador exclusivo atribuído para registros na base de dados do NCBI.

Desc	riptions	Graphic Summary	Alignments	Taxonomy							
Seq	uences pr	roducing significant a	lignments		Download 🐣	Mana	age Col	umns	Y S⊦	iow 1	00 🗸 🔞
🗹 s	elect all 1	00 sequences selected				Gen	<u>Bank</u>	Graph	ics D	istance t	ree of results
			De	scription		Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
	Dengue virus	s 2 isolate CHRF_DenV002, par	rtial genome			4566	15967	81%	0.0	98.58%	MN328061.1
	Dengue virus	s 2 B17-1489 RNA, nearly comp	olete genome			4540	15849	81%	0.0	98.43%	LC436672.1
	Dengue virus	s 2 B17-1556 RNA, nearly comp	olete genome			4495	15780	81%	0.0	98.11%	LC436673.1
	Dengue virus	s 2 B17-1357 RNA, nearly comp	plete genome			4494	15740	81%	0.0	98.11%	LC436669.1
	Dengue virus	s 2 B17-1571 RNA, nearly comp	plete genome			4488	15729	81%	0.0	98.07%	LC436674.1
	Dengue virus	s 2 strain TM38 polyprotein gen	e, complete cds			4479	16098	81%	0.0	97.96%	KX452017.1
	Dengue virus	s 2 B17-1634 RNA, nearly comp	olete genome			4477	15706	81%	0.0	97.99%	LC436675.1
	Dengue virus	s 2 strain TM132 polyprotein ge	ne, complete cds			4468	15896	81%	0.0	97.88%	KX452030.1
	Dengue virus	s 2 strain TM39 polyprotein gen	e. complete cds			4468	15847	81%	0.0	97.88%	KX452018.1
	Dengue virus	s 2 isolate SG(EHI)D2/02862Y1	5. complete genome			4462	15655	81%	0.0	97.84%	MK513444.1
	Dengue virus	s 2 isolate SG(EHI)D2/09087Y1	5. complete genome			4462	15655	81%	0.0	97.84%	KY921905.1
	Dengue virus	s 2 strain QML16, complete gen	iome			4457	15638	81%	0.0	97	Eeedbac

5.2. Clique sobre a opção "Alinhamentos".

- a) Observação da sequência submetida à base de dados em azul entre o principal.
- b) A localização e o comprimento dos alinhamentos de sequência são representados abaixo. Cada linha representa um táxon.
- c) A cor do alinhamento representa a qualidade do alinhamento, baseado na pontuação do alinhamento. Vermelho representa os alinhamentos com a pontuação mais alta (melhor alinhamento), enquanto preto representa a pior pontuação e não é necessariamente confiável.
- d) As linhas horizontais cinzentas representam lacunas no alinhamento.

escription	s	Graphic Summary	Alignments	Taxonomy		
ignment v	iew F	Pairwise		▼ CDS fea	ature 🔞 Re	estore defaults
00 sequence	es selecte	ed 😮				
Ł Dow	nload 🗸	GenBank Graphic	Sort by: E value	~		
Dengu	e viru	s 2 isolate CHRF_De	nV002, partial ge	nome		
Sequen	ce ID: M	N328061.1 Length: 10	699 Number of Mate	hes: 11		
Range	1: 8153	to 10696 GenBank	Fraphics		Next Mate	h 🔺 Previous Mat
Score 4566 b	oits(247)	Expect Id 2) 0.0 25	entities 508/2544(99%)	Gaps 0/2544(0%)	Strand Plus/Plus	
Query	8149		AGGAGCTTTGGTGAGGA	ATCCACTCTCACGAAAT		8208
Sbjct	8153	CTACAAAGGAAATACGG	AGGAGCTTTGGTGAGGA	ATCCACTCTCACGAAAT	TCCACACAC	8212
Query	8209	GAGATGTACTGGGTATC		TAGTGTCATCAGTGAAC	ATGATTTCA	8268
Sbjct	8213	GAGATGTACTGGGTATC	CAATGCTTCCGGGAACA	TAGTGTCATCAGTGAAC	ATGATTTCA	8272
Query	8269	AGAATGTTGATTAACAG	ΑΤΤΟΑΟΑΑΤΘΑΘΑΟΑΟΑ	AGAAGGCCACATACGAG	CCGGATGTT	8328
Sbjct	8273	AGAATGTTGATTAACAG	ATTCACAATGAGACACA	AGAAGGCCACATACGAG	CCGGATGTT	8332
Jery	8329	GATCTCGGAAGTGGAAC	CCGCAACATCGGAATTG	AAAGTGAAGTACCAAAT		8388
Dict	8333	GATCTCGGAAGTGGAAC	CGCAACATCGGAATTG	ΔΔΔGTGΔΔGTΔCCΔΔΔΤ	CTAGACATA	8392

C 0 = 200 Alignment Scores < 40 40 - 50 50 - 80 80 - 200 Distribution of the top 200 Blast Hits on 100 subject sequences а 6000 8000 **4**000 2000 10000 b

Clique no alinhamento para visualizar o alinhamento da sequência.

5.3. Pode ser feito o **download do alinhamento ou das sequências** ao clicar download e escolhendo o arquivo de interesse no menu.

500 K	J1M.	1,500 K	2 M	2,500 K	I3 M	3,50	0K 4M	4,5	00 K 5,087,1
. v vá z i stan v		opuir num ur	• • • • • • •	. gan a se	yv vryg	74 ° 5 🥳		, mand	w
5 CP026491.1 - Find:	-	~ (¢¢)	0	-()	N CONT		🔀 Tools 🗸 🔯 1	Tracks 🕶 📩 Down	nload • 🥲 🤋 •
2,331 K	2,331,500	2,332 1	<	2,332,500		2,333 K	2,33	3,500	2,334 K
Sequence									0 2
Genes				040 (340					¥ 0 ¥
						C3K24_12145			
	2,925	2,600 2,400	2,200 2	rK 1,800	1,500	1,400 1,200	16 1	00 600	400 20
< r	RNA-55 ribosomal	R							
(U) BLAST Results for	: Nucleotide	Sequence		Query_45959					0 ×
(U) BLAST Results for	: Nucleotide	Sequence		Query_45959				1	0.5
(U) BLAST Results for	: Nucleotide	Sequence sults for: Nuc	leotide Se	Query_45959	11			1	• • •
(U) BLAST Results for (U) Cleaned Alignment	: Nucleotide s - BLAST Re	Sequence sults for: Nuc:	leotide Se	0uery_45959 < q					• • •

6. Registe as seguintes métricas:

- accession number of the best hit (usually the top hit).
- E value
- query coverage

7. <u>O taxon foi identificado!</u> Usar esses resultados para responder o questões abaixo.

Questões

1. Na secção Descrições, verificar o principal resultado, qual deve ser o resultado com a pontuação mais alta. Anote as informações sobre a melhor combinação:

Descrição

E value

Identity

Query cover

2. Na secção Alinhamentos, observe o alinhamento entre a sua sequência e a sequnência de referência.

Fazer você ver alguma incompatibilidade?

3. É possível identificar a sequência ao nível da espécie? Por favor elaborar.

4. Identificar o estado de conservação da espécie identificada.