QBQ102 – 2021s1 – Turmas de Educação Física e Esportes

Tarefa C

Professor João Carlos Setubal – setubal@iq.usp.br

13 de maio de 2021

O objetivo desta tarefa é familiarizar o aluno com conceitos de genômica e bioinformática, e também avaliar aprendizado de conceitos de biologia molecular apresentados nas aulas.

Para realização da tarefa basta que cada aluno tenha acesso à internet e seja capaz de preparar as respostas (relatório) num computador pessoal que possua um processador de textos.

Formato do relatório

Os relatórios são individuais.

O relatório deve seguir o modelo apresentado ao final deste documento.

Relatórios devem estar em formato PDF. O nome do arquivo deve ser o nome do aluno. Por exemplo: joaosetubal.pdf

Os relatórios devem ser enviados ao professor por email (setubal@iq.usp.br) **usando obrigatoriamente no campo assunto as palavras**: entrega *tarefa C QBQ102* (os asteriscos são essenciais e o uso de maiúsculas e minúsculas deve ser seguido à risca!). Somente se você fizer isto você receberá um recibo de entrega (que será um email automático). Se você não receber o recibo automático de entrega, **não vou considerar seu projeto como entregue.**

prazo de entrega do relatório: até meio dia de **27 de julho de 2021**

Atividade 1: Conceitos básicos

Todo organismo tem um **genoma**, definido como sendo o conjunto de suas moléculas de DNA contendo informação genética. Mesmo virus, que não tem status de organismos independentes, possuem informação genética armazenada em moléculas de DNA ou de RNA.

Neste projeto, trabalharemos com genomas de bactérias e de arquéias. Estes dois conjuntos de organismos são coletivamente conhecidos como **procariotos**. Embora distintos, esses organismos se caracterizam por serem unicelulares. Além dos procariotos, encontramos na biosfera os eucariotos e os virus.

Iremos trabalhar com *representações computacionais* dos genomas de procariotos, o que é fruto de seu sequenciamento. Isto quer dizer que para fins deste projeto, um genoma é uma coleção de arquivos de computador, contendo informações diversas sobre o genoma enquanto

molécula. Um desses arquivos é a *sequência do genoma*. Por exemplo, a sequência do genoma da bactéria *Xanthomonas citri* cepa A306 tem esta cara:

```
>NC_003919.1 Xanthomonas axonopodis pv. citri str. 306, complete genome
TTTTCTGCGCGCCGCAGCCAGCAACAGATGATCATACTTTGATGGATGCTTGGCCCCGCT
GTCTGGAACGTCTCGAAGCCGAATTCCCGCCCGAAGATGTCCACACCTGGTTGAAACCCC
TGCAAGCCGAAGATCGCGGCGACAGCATCGTGCTGTACGCTCCGAACGCCTTCATCGTCG
```

Estas são apenas as 4 primeiras linhas do arquivo, de um total de 86261 linhas. Na primeira linha está o cabeçalho, que mostra a qual organismo pertence o genoma. NC_003919.1 é um identificador utilizado pelo *National Center for Biotechnology Information* (NCBI) para identificar esse genoma. Esse identificador recebe o nome de **Accession Number**. Nas demais linhas aparece a sequência de DNA, que pode ser entendida como uma longa cadeia de nucleotídeos, aqui representados pelas letras A, C, G, e T. A representação num arquivo de computador exige que a cadeia seja linear, mas na maioria das bactérias a molécula na verdade é circular; ou seja, para que possamos armazenar a sequência num arquivo, temos que linearizar a representação da molécula, "cortando-a" numa certa posição. A primeira base à direita desse corte é a primeira que aparece na linearização (T no exemplo acima); a base imediatamente à esquerda do ponto de corte será a última.

Outro conceito importante é o de **amplicons**. Para nós, um amplicon é um componente de um genoma que tem sua replicação independente dos demais amplicons. Os amplicons que nos interessam são os cromossomos; os demais, no caso de procariotos, se chamam plasmídeos. O exemplo acima é do cromossomo de *X. citri*. A maioria dos procariotos tem um único cromossomo.

Neste projeto iremos nos ocupar principalmente com os *genes* dos genomas, em particular os genes codificadores de proteínas (GCPs). No NCBI também é possível encontrar um arquivo com as sequências de todos os GCPs que se conhecem para um determinado genoma. Por exemplo, o primeiro gene de *X. citri*, na linearização apresentada, pode ser representado da seguinte forma:

```
>WP_011050021.1: chromosomal replication initiator protein DnaA
MDAWPRCLERLEAEFPPEDVHTWLKPLQAEDRGDSIVLYAPNAFIVDQVRERYLPRIRELLAYFVGNGDV
ALAVGSRPRAPEPAPAPVAVPSAPQAAPIVPFAGNLDSHYTFANFVEGRSNQLGLAAAIQAAQKPGDRAH
```

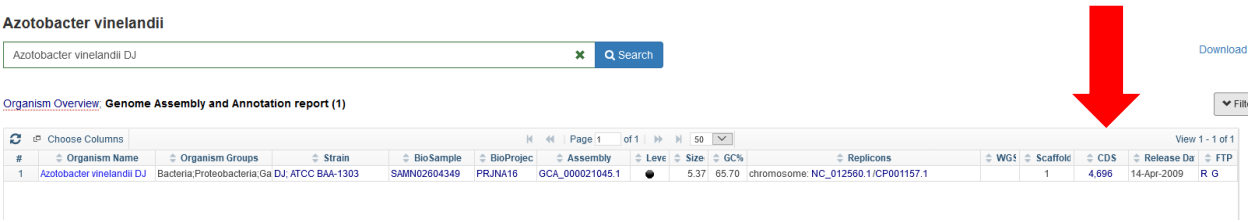
Veja que o formato é semelhante ao do genoma: mesmo tipo de cabeçalho, e nas linhas seguintes aparece a sequência, que é de aminoácidos a não de nucleotídeos (ou seja, a porção codificadora desse gene foi traduzida de nucleotídeos para aminoácidos utilizando o código genético). Estão mostradas apenas as primeiras 3 linhas do arquivo, de um total de 8.

Nos repositórios de dados a porção codificadora de um GCP recebe o nome de *Coding Sequence*, abreviado para **CDS**.

Roteiro para atividade 1

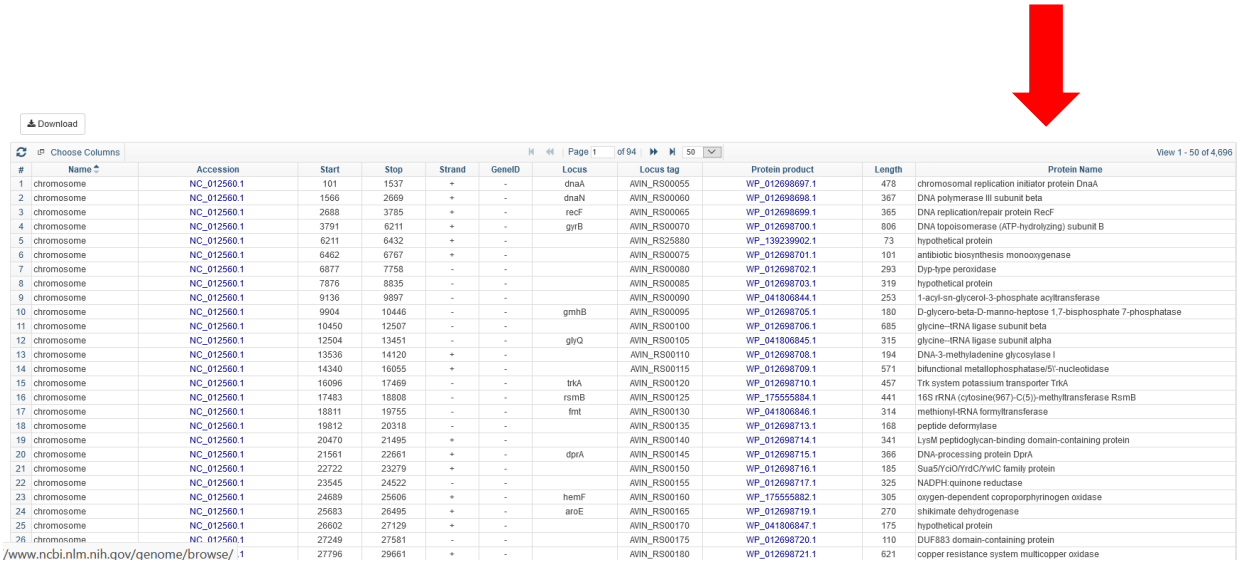
1. Cada aluno tem designado a si dois de genomas de procariotos, conforme a tabela tabelaGenomasQBQ102.pdf disponível no site da disciplina http://www.iq.usp.br/setubal/qbq102/2021/tabelaGenomasConsultaGenomasAlvo.pdf Os dois genomas são aqui chamados de *genoma-consulta* e *genoma-alvo*. Na tabela constam os **Accession Numbers** desses genomas. Cada "genoma" na verdade é apenas o cromossomo principal desse procarioto; estamos fazendo a simplificação de ignorar outros amplicons do organismo, caso existam.
2. Acesse a página https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/
3. Clique na aba "Prokaryotes"
4. Na caixa de busca, coloque o Accession Number do seu genoma-consulta.
5. Deve aparecer uma tabela com apenas 1 linha, que mostra dos dados do seu genoma-consulta (veja Figura 1).
6. Uma das colunas dessa tabela se chama **CDS**. Nesta coluna há um número (veja Figura 1). Clique nesse número. Deve aparecer a lista dos genes codificadores de proteína do seu genoma-consulta (veja Figura 2).

Figura 1



7. Nessa lista você deve procurar 3 genes de proteínas (enzimas) que foram mencionadas nas aulas de biologia molecular como enzimas ou proteínas importantes nos processos de replicação, transcrição, tradução, reparo de DNA, ou expressão gênica. Utilize a coluna **protein name** e/ou o recurso de busca (CTRL-F) para achar esses genes. Leve em conta que numa determinada tela só aparecem 50 genes; para achar seus genes você provavelmente terá que pesquisar várias páginas, usando o recurso de mudar de página que aparece na parte de baixo da tela).

Figura 2



Download

Choose Columns

Page 1 of 94 | 50 | View 1 - 50 of 4,696

#	Name	Accession	Start	Stop	Strand	GeneID	Locus	Locus tag	Protein product	Length	Protein Name
1	chromosome	NC_012560.1	101	1537	+	-	dnaA	AVIN_RS00055	WP_012698697.1	478	chromosomal replication initiator protein DnaA
2	chromosome	NC_012560.1	1566	2669	+	-	dnaN	AVIN_RS00060	WP_012698698.1	367	DNA polymerase III subunit beta
3	chromosome	NC_012560.1	2688	3785	+	-	recF	AVIN_RS00065	WP_012698699.1	365	DNA replication/repair protein RecF
4	chromosome	NC_012560.1	3791	6211	+	-	gyrB	AVIN_RS00070	WP_012698700.1	806	DNA topoisomerase (ATP-hydrolyzing) subunit B
5	chromosome	NC_012560.1	6211	6432	+	-		AVIN_RS25880	WP_139239902.1	73	hypothetical protein
6	chromosome	NC_012560.1	6462	6767	+	-		AVIN_RS00075	WP_012698701.1	101	antibiotic biosynthesis monooxygenase
7	chromosome	NC_012560.1	6877	7758	-	-		AVIN_RS00080	WP_012698702.1	293	Dyp-type peroxidase
8	chromosome	NC_012560.1	7876	8835	-	-		AVIN_RS00085	WP_012698703.1	319	hypothetical protein
9	chromosome	NC_012560.1	9136	9897	-	-		AVIN_RS00090	WP_041806844.1	253	1-acyl-sn-glycerol-3-phosphate acyltransferase
10	chromosome	NC_012560.1	9904	10446	-	-	gmhB	AVIN_RS00095	WP_012698705.1	180	D-glycero-beta-D-manno-heptose 1,7-bisphosphate 7-phosphatase
11	chromosome	NC_012560.1	10450	12507	-	-		AVIN_RS00100	WP_012698706.1	685	glycine--tRNA ligase subunit beta
12	chromosome	NC_012560.1	12504	13451	-	-	glyQ	AVIN_RS00105	WP_041806845.1	315	glycine--tRNA ligase subunit alpha
13	chromosome	NC_012560.1	13536	14120	+	-		AVIN_RS00110	WP_012698708.1	194	DNA-3-methyladenine glycosylase I
14	chromosome	NC_012560.1	14340	16055	+	-		AVIN_RS00115	WP_012698709.1	571	bifunctional metallophosphatase/5\'-nucleotidase
15	chromosome	NC_012560.1	16096	17469	-	-	trkA	AVIN_RS00120	WP_012698710.1	457	Trk system potassium transporter TrkA
16	chromosome	NC_012560.1	17483	18808	-	-	rsmB	AVIN_RS00125	WP_175555884.1	441	16S rRNA (cytosine(967)-C(5))-methyltransferase RsmB
17	chromosome	NC_012560.1	18811	19755	-	-	fmt	AVIN_RS00130	WP_041806846.1	314	methionyl-tRNA formyltransferase
18	chromosome	NC_012560.1	19812	20318	-	-		AVIN_RS00135	WP_012698713.1	168	peptide deformylase
19	chromosome	NC_012560.1	20470	21495	+	-		AVIN_RS00140	WP_012698714.1	341	LysM peptidoglycan-binding domain-containing protein
20	chromosome	NC_012560.1	21561	22661	+	-	dprA	AVIN_RS00145	WP_012698715.1	366	DNA-processing protein DprA
21	chromosome	NC_012560.1	22722	23279	+	-		AVIN_RS00150	WP_012698716.1	185	Sua5/YciO/YrdC/YwlC family protein
22	chromosome	NC_012560.1	23545	24522	-	-		AVIN_RS00155	WP_012698717.1	325	NADPH:quinone reductase
23	chromosome	NC_012560.1	24689	25606	+	-	hemF	AVIN_RS00160	WP_175555882.1	305	oxygen-dependent coproporphyrinogen oxidase
24	chromosome	NC_012560.1	25683	26495	+	-	aroE	AVIN_RS00165	WP_012698719.1	270	shikimate dehydrogenase
25	chromosome	NC_012560.1	26602	27129	+	-		AVIN_RS00170	WP_041806847.1	175	hypothetical protein
26	chromosome	NC_012560.1	27249	27581	-	-		AVIN_RS00175	WP_012698720.1	110	DUF883 domain-containing protein
27	chromosome	NC_012560.1	27796	29661	+	-		AVIN_RS00180	WP_012698721.1	621	copper resistance system multicopper oxidase

/www.ncbi.nlm.nih.gov/genome/browse/ :1

Como reportar resultados da atividade 1: apresente uma lista com 3 genes do seu genoma-consulta com as seguintes informações *para cada gene* (sua ficha):

nome do organismo / genoma utilizado; indique se o organismo é bactéria ou arquéia
locus tag (é uma coluna da lista de proteínas)
coordenadas genômicas (start e stop)
Comprimento em amino ácidos da proteína produzida pelo gene = (stop – start + 1)/3
fita (fita mais ou fita menos) (strand)
nome da proteína
sequência em aminoácidos, que você pode obter clicando no link da coluna *protein product*. Na página que aparecer, haverá um pequeno link à esquerda chamado GenPept: clique nele, e vai aparecer um menu. Nesse menu escolha a opção FASTA. Desta forma a sequência vai aparecer num formato fácil de copiar e colar.

Para cada um dos genes que você escolheu, explique a função da proteína codificada por esse gene no processo do qual ela participa. Para esta explicação você deve usar o material das aulas, eventualmente complementado com material que você mesmo pode achar na Internet. **Se você usar material da Internet, você precisa indicar de qual ou quais sites você retirou o material.**

Atividade 2: Conceitos básicos

Nesta atividade cada aluno irá comparar as sequências dos genes relatados na atividade anterior contra as sequências de genes do genoma-alvo, utilizando a ferramenta BLAST.

A ferramenta ou programa BLAST (*Basic Local Alignment Search Tool*) é capaz de comparar sequências de nucleotídeos e de proteínas entre si. Tal comparação tem muitas utilidades; uma delas é verificar a presença de um determinado gene *X* de organismo *A* em outro organismo *B*. Se *X* estiver presente em *B* como gene *Y*, então a ferramenta nos devolve um *bom alinhamento* entre *X* e *Y*. Um exemplo de alinhamento de duas sequências de proteínas é mostrado logo abaixo. A palavra 'bom' foi sublinhada porque este alinhamento precisar ser de *boa qualidade*, para podermos concluir que de fato *X* e *Y* representam o mesmo gene (ou seja, codificam proteínas que tem a mesma função nos respectivos organismos).

Como medir a qualidade de um alinhamento? Há várias formas, sendo que as principais são as seguintes:

O alinhamento deve incluir grande parte de *X* e de *Y*; idealmente deveria incluir *toda* a sequência *X* e *toda* a sequência *Y*. Mas é raro isto acontecer. Em geral, desejamos que ao menos **80%** de *X* e ao menos **80%** de *Y* participem do alinhamento. Vamos chamar esta medida de ***cobertura***. (Veja mais abaixo uma explicação mais detalhada do que é cobertura.)

Para medir a similaridade entre X e Y utilizaremos a medida *percentual de identidade*. Esta medida nos informa quantas posições do alinhamento tem aminoácidos que são idênticos, dividido pelo tamanho do alinhamento e multiplicado por 100.

Devemos também ter uma medida da *significância estatística* do alinhamento. Para entender este conceito, considere o alinhamento da sequência LLL contra a sequência LLL (onde L representa o aminoácido leucina). O percentual de similaridade é 100%, pois as sequências são idênticas. Mas tal alinhamento não é estatisticamente significativo; ele pode ser fruto do acaso (que é favorecido pelo fato de que leucinas são muito comuns em proteínas). Para medir significância estatística utilizaremos uma medida chamada **e-valor**, ou valor esperado (**e-value** em inglês).

E-valor nos dá o número esperado de alinhamentos ao acaso para a dada sequência de consulta e as dadas sequências do banco de sequências onde se fez a busca com BLAST. Assim, se um e-valor for igual a 3 para um dado alinhamento, podemos inferir que é alta a probabilidade de que esse alinhamento seja devido ao acaso. Um valor de 0,1 já indica que essa probabilidade é bem menor. Na prática, utilizaremos como limiar o valor $\mathbf{10^{-5}}$, que é 0,00001. Ou seja, se um e-valor for igual ou menor a 10^{-5}, diremos que o alinhamento é estatisticamente significativo; caso contrário, não. O fundamento teórico desta medida e deste limiar está além do escopo desta disciplina.

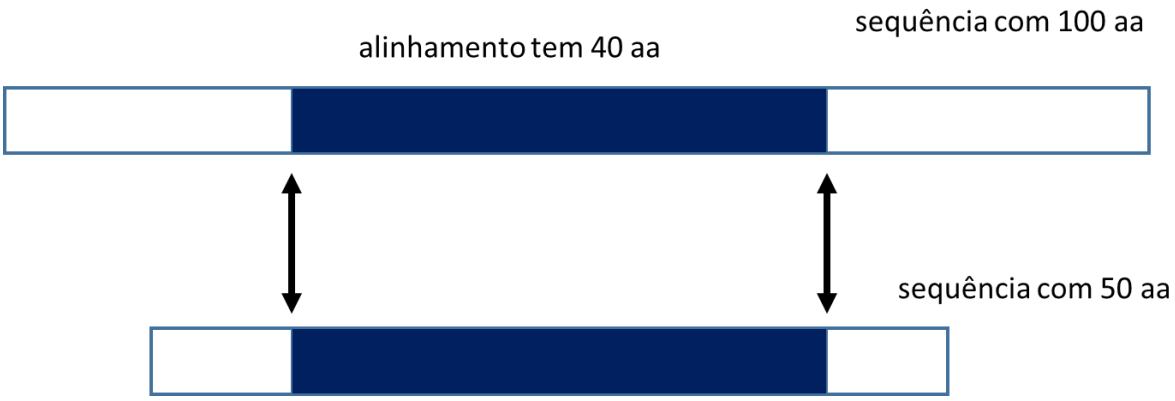
Note que comumente se utiliza a seguinte notação para representar e-valores menores do que 1: a letra `e`, seguida de um sinal negativo, seguida de um número, que é o valor do expoente. Por exemplo, o valor 10^{-5} nessa notação é `e-5`. Se fosse `e-47`, isto significaria 1 sobre 1 elevado a 47, ou 10^{-47}, portanto um número muito próximo de zero. Quanto mais próximo de zero for o e-valor, mais significativo é o alinhamento.

Exemplo de um alinhamento de BLAST:

```
>lcl|35099 t
Length=499

 Score =  604 bits (1558),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 301/499 (60%), Positives = 365/499 (73%), Gaps = 25/499 (5%)

Query  21   DAACSEAAGDKSAMMHDALFERFSARLKAQVGPEVYASWFARLKLHTVSKSVVRFTVPTT  80
            DA C E   ++       LF+  S++L+ QVG +VYASWF RLK  +VS ++V  +VPT
Sbjct  23   DARCLETTCEE-------LFKNVSSKLEDQVGSDVYASWFQRLKFRSVSHNIVYLSVPTN  75

Query  81   FLKSWINNRYMDLITSLVQSEDPDVLKVEILVRSASRPVRPAQTEERAQPVQEVGAAPRN  140
            FLK+WI NRY+D IT L Q     +  VEI+VRSA+  + P++T              +
Sbjct  76   FLKAWIKNRYIDTITKLFQESISSIQGVEIIVRSAA--LMPSETS-------------S  119

Query  141  KSFIPSQSATAPAAQPMAAQATLRQGGSGPLFGSPLDTRFTFDTFVEGSSNRVALAAAKT  200
             S I   +A  P            +    P+FGSPLD++F F  F+EG SNRVALAAA T
Sbjct  120  SSAIAHTTAKPPIINTGKISTIQGKQSINPVFGSPLDSKFVFSNFIEGPSNRVALAAAHT  179
```

A figura acima mostra apenas parte de um alinhamento. A primeira linha indica o nome da sequência-sujeito (`lcl|35099`) (a sequência que BLAST achou ao fazer a busca com base na sequência-consulta). A segunda linha indica que essa sequência tem 499 aminoácidos. Nas duas linhas seguintes há informações sobre o alinhamento. As partes que nos interessam são: `Expect` (é o e-valor); `Identities`, que é o **percentual de identidade**; e o denominador de `Identities` e de `Positives`, que é o tamanho do alinhamento em aminoácidos. Nas linhas seguintes aparece o alinhamento propriamente dito, em seções. Cada seção tem 3 linhas, representando um trecho do alinhamento. São necessárias diferentes seções porque o alinhamento não cabe numa única linha, então é necessário quebrar o alinhamento em várias seções. As 3 linhas de cada seção mostram a sequência-consulta na primeira linha (`query`), a sequência-sujeito na terceira linha (`subject`), e uma linha no meio que mostra as posições onde os aminoácidos são idênticos. O sinal de positivo (+) indica aminoácidos não-idênticos mas de propriedades físico-químicas semelhantes.

Neste parágrafo, explico melhor o conceito de **cobertura**. Considere o alinhamento ilustrado abaixo:



Neste alinhamento os retângulos representam sequências. A sequência de cima tem 100 aa e a sequência de baixo tem 50 aa. No entanto, o alinhamento entre elas resultou num tamanho de apenas 40 aa (a parte indicada pelas flechas). Neste caso, podemos dizer que o alinhamento cobre 40% da sequência de cima (40/100) e 80% da sequência de baixo (40/50). Note portanto que a sequência de cima tem uma cobertura diferente da sequência de baixo. Para calcular a cobertura é essencial saber o tamanho das sequências alinhadas.

Não confunda cobertura com percentual de identidade! esse percentual sempre aparece na saída do BLAST (veja `Identities`). As coberturas, você mesmo que terá que calcular com base na explicação acima.

Resumo dos critérios do que é um bom alinhamento

Todos os critérios abaixo devem ser satisfeitos:

cobertura: pelo menos 80% da sequência-consulta e pelo menos 80% da sequência-sujeito

e-value: no máximo 10^{-10}, que é a mesma coisa que `1e-10`. Veja que `1e-20 ou 3e-45`, etc, são bem menores do que `1e-10`, e portanto bem melhores.

percentual de identidade: quanto mais próximo de 100%, melhor. Se o % de identidade for menor do que 20%, é um alinhamento ruim. Afora essas duas pontas, a avaliação do % de identidade vai depender muito da distância filogenética entre os organismos que forem comparados. Se forem próximos (por exemplo, duas bactérias do mesmo gênero), esperamos que o % de identidade seja alto; se forem muito distantes (por exemplo, uma bactéria e uma arquéia), esse % de identidade será baixo, podendo se aproximar de 20% (e ainda assim o alinhamento ser bom por causa da boa cobertura e do baixo e-value).

Os genes que você deve ter escolhido na atividade 1 são genes fundamentais para qualquer organismo celular. Por esse motivo, espera-se que estejam presentes em qualquer bactéria e em qualquer arquéia. Isto quer dizer que todas as buscas que você fará na atividade 2 devem gerar alinhamentos (não necessariamente *bons* alinhamentos; não é uma exigência deste projeto que você necessariamente encontre bons alinhamentos).

Entretanto, a qualidade do alinhamento irá variar entre diferentes pares de genoma-consulta e genoma-alvo. Essa qualidade será tanto melhor quão mais próximos esses dois genomas forem filogeneticamente entre si. A distância filogenética é uma forma de quantificar a distância evolutiva entre dois quaisquer organismos. Por exemplo, a distância filogenética que separa humanos de chimpanzés é muito menor do que a distância que separa humanos de ratos, e menor ainda do que a distância que separa humanos de moscas. Com as bactérias e arquéias se passa o mesmo.

Roteiro para Atividade 2

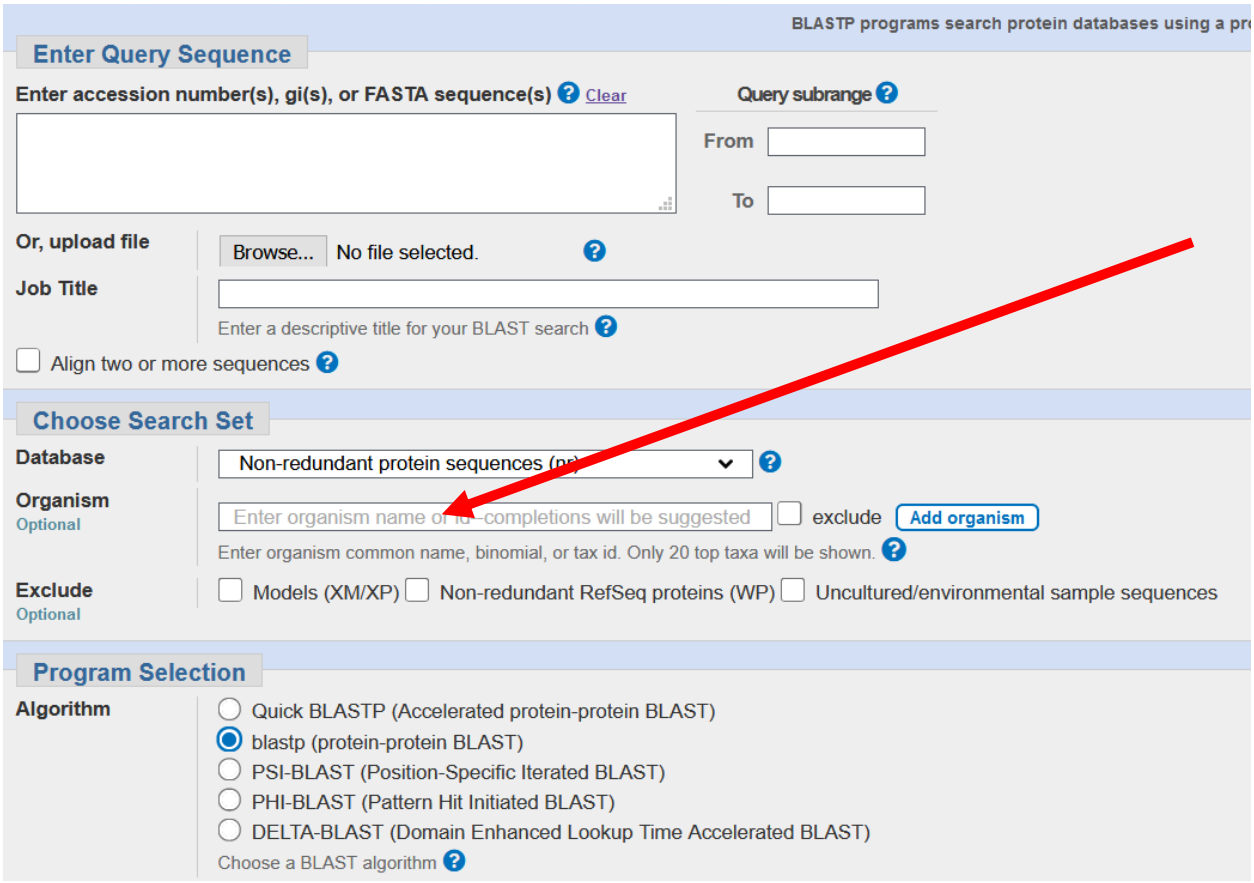
Para esta atividade, você deve ter as sequências em aminoácidos dos genes que descreveu na atividade 1. Estas serão as suas *sequências-consulta*. Além disso, na mesma tabela do website em que você pegou o genoma-consulta, você deve pegar o nome do seu *genoma-alvo*.

Para cada uma das sequências-consulta:

1. Acesse o programa BLAST em https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. Escolha a opção Protein BLAST (também conhecida por blastp)
3. Insira a sequência na caixa **Enter Query Sequence**
4. No quadro **Choose Search Set** e no subquadro **Organism** digite o nome do *genoma-alvo* (**não** digite o accession number). O sistema tem que reconhecer o nome que vc digitou (veja Figura 3). Note que para muitos procariotos, além de aparecer o nome da espécie também pode aparecer uma opção com designação da cepa (por exemplo: Escherichia coli K12). Você deve sempre escolher a opção genérica (sem cepa).
5. Rode BLAST apertando o botão azul BLAST.

Ao fazer isto, BLAST irá comparar a sua sequência-consulta contra *todas* as sequências de genes codificadores de proteínas do genoma-alvo (estas são as suas potenciais sequências-sujeito), e reportar todos os alinhamentos resultantes, começando pelo melhor (menor e-valor).

Figura 3



Como reportar resultados da Atividade 2

Informe o nome dos organismos do seu genoma-consulta e do seu genoma-alvo.

Seu relatório deve apresentar os alinhamentos que você obteve, para cada um dos seus 3 genes. Apresente *apenas o alinhamento do primeiro hit de cada um dos seus genes*. Você pode copiar e colar o alinhamento reportado por BLAST. O alinhamento apresentado deve ser **completo** (o mesmo que você vê na tela de resposta do BLAST; veja Figura 4).

Figura 4

DNA polymerase III subunit beta [Bacillus subtilis]

Sequence ID: CUB55951.1 Length: 379 Number of Matches: 1

Range 1: 1 to 377 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
166 bits(420)	6e-47	Compositional matrix adjust.	99/378(26%)	208/378(55%)	13/378(3%)

```
Query  1    MHFTIQREALLKPLQLVAGVVERRQTLPVLSNVLLVVEKQQLSLTGTDLEVELVGRVPLE  60
            M FTIQ++ L++ +Q V    V  R T+P+L+ + +V  ++ ++LTG+D ++ +   +P+E
Sbjct  1    MRFTIQKDYLVRSVQDVMKAVSSRTTIPILTGIKVVATEEGVTLTGSDADISIESFIPVE  60

Query  61   ENA-------EPGEITVPARKLMDICKSLPNDTLIDIRLDEQKLL-IKAGRSRFSLSTLP  112
            E+         + G I + A+   +I K LP +T ++I ++    +  I +G+S F+L+ L
Sbjct  61   EDGKEIVEVKQSGSIVLQAKYFSEIVKKLPKET-VEISVENHLMTKITSGKSEFNLNGLD  119

Query  113  ASDFPTAEEGLGSLTFSLGQSKLRRLIERTSFAMAQQDVRYYLNGMLLEMNGGVLRAVAT  172
            ++++P   +      F +    L+ +I +T FA++  + R  L G+   ++    L  +AT
Sbjct  120  SAEYPLLPQIEEHHVFKIPTDLLKHMIRQTVFAVSTSETRPILTGVNWKVYNSELTCIAT  179

Query  173  DGHRLALCSMQ-SGIEHADRHQVIVPRKGILELARLLTDQDGEVSIVLGQYHIRATTGEF  231
            D HRLAL   +  GI    +  V++P K + EL+++L + +  V IV+ +Y +   T
Sbjct  180  DSHRLALRKAKIEGIADEFQANVVIPGKSLNELSKILDESEEMVDIVITEYQVLFRTKHL  239

Query  232  TFTSKLVDGKFPDYERVLPRGGDKKVLGDRQLLREAFSRTAILSNE-KYRGIRLQ-LASG  289
             F S+L++G +PD  R++P      +  + +   +A  R ++L+ + +   ++L  L
Sbjct  240  LFFSRLLEGNYPDTTRLIPAESKTDIFVNTKEFLQAIDRASLLARDGRNNVVKLSTLEQA  299

Query  290  LLKIQANNPEQEEAEEEVAVD-YSGDALEIGFNVSYLLDVLGVMSAEQVCLTLSDSNSSA  348
            +L+I +N+PE  +  EEV  +   G+ L+I F+  Y++D L  + + ++ ++ + +
Sbjct  300  MLEISSNSPEIGKVVEEVQCEKVDGEELKISFSAKYMMDALKALDSTEIKISFTGAMRPF  359

Query  349  LLQEADNDDSAYVVMPMR  366
            L++  +++    +++P+R
Sbjct  360  LIRTVNDESIIQLILPVR  377
```

Inclua comentários sobre a *qualidade* desse alinhamento com base nos conceitos básicos explicados acima. As informações sobre a qualidade do alinhamento fazem parte do resultado do BLAST, conforme ilustrado acima. O alinhamento da Figura 4 é bom, pois seu e-value é baixo (e-47), a cobertura é boa (quase 100%, tanto para consulta quanto para alvo, e o percentual de identidade está acima de 20%).

Você deve no final comparar os 3 alinhamentos obtidos entre si. Qual foi o melhor? qual foi o pior? Em geral podemos dizer que um alinhamento X é melhor do que um alinhamento Y se X for melhor do que Y em todos os critérios listados acima (cobertura da consulta, cobertura do alvo, % de identidade, e e-value). Mas às vezes X pode ser melhor do que Y num critério e pior

em outro. Nesses casos, verifique qual alinhamento tem mais critérios melhores para fazer o desempate.

Nota bene 1: o resultado esperado de sua busca com BLAST é que você encontrará bons "hits". O motivo principal para isso é que os genes e suas proteínas que estudamos nas aulas, e que você deve ter escolhido para fazer suas buscas, são genes fundamentais para qualquer organismo. Assim sendo, a probabilidade é alta de que você vai encontrar bons hits. Entretanto, ocasionalmente poderá acontecer que uma particular dupla de genoma-consulta e genoma-alvo tenha como resultado "no hits". Este resultado poderia ocorrer por exemplo se seu genoma-consulta é uma bactéria, e seu genoma-alvo é uma arquéia. Esses 2 tipos de organismos são muito distantes filogeneticamente. Caso isto tenha acontecido com sua dupla de genomas, você deverá escolher outro genoma-alvo da tabela do website (ou seja, um genoma-alvo atribuído a outro aluno) e repetir a busca. De qualquer forma, no seu relatório você sempre deve indicar que genoma-consulta e que genoma-alvo você usou, mesmo que sejam os mesmos atribuídos a você na tabela. Se mesmo assim persistir seu resultado de "no hits", uma possível explicação é que você não fez uma boa escolha de genes!

Nota bene 2: Altere o parâmetro *word size* do BLASTP de 6 para 3. Esse parâmetro é encontrado na seção *Algorithm parameters* que aparece no final da página do BLAST, logo depois do botão de execução. Para ver os parâmetros disponíveis basta clicar no sinal de '+'.

Atividade 3: **conceitos básicos**

Toda CDS de um gene de proteína pode ser representada por uma sequência de nucleotídeos e por uma sequência de aminoácidos. Para ir de uma para a outra basta fazer a tradução usando o código genético (se a tabela de código genético tem U ao invés de T, basta interpretar U como sendo T). Na atividade 2 você usou a versão em aminoácidos dos genes que escolheu. Nesta atividade, você vai repetir a atividade 2, mas agora usando a versão em nucleotídeos.

Vamos então ver que cara tem um alinhamento de nucleotídeos, através do seguinte exemplo:

Score	Expect	Identities	Gaps	Strand
219 bits(242)	2e-54	646/985(66%)	22/985(2%)	Plus/Plus

```
Query  358      AACCCGAAATATACATTTGATACTTTCGTCATCGGCAAGGGGAACCAAATGGCCCATGCC  417
                || || ||||||||||||||||| || || ||||||   || ||||   | ||||||||
Sbjct  5157129  AATCCAAAATATACATTTGATACATTTGTTATCGGCTCTGGTAACCGTTTTGCCCATGCA  5157188

Query  418      GCAGCGTTAGTTGTGTCGGAAGAGCCCGGGACAATGTATAATCCGTTGTTTTTCTACGGG  477
                ||  | |||| |||  | || | ||| |  | |  |||||| ||  | ||| | ||||||
Sbjct  5157189  GCTTCATTAGCTGTAGCCGAGGCGCCAGCTAAAGCGTATAACCCACTCTTTATTTACGGG  5157248

Query  478      GGCGTTGGTCTGGGCAAAACCCACCTAATGCACGCTATCGGTAACAAATTGTTAGAAACC  537
                || ||||| || || || || ||  |||||||||| || ||| |  |  |  | |||
Sbjct  5157249  GGAGTTGGGCTTGGAAAGACGCATTTAATGCACGCAATTGGTCATTATGTAATTGAACAT  5157308
```

```
Query  538      GATCCGACTAGTAACATCAAATACGTCACTAGCGAATCCTTCACCAATGAATTGATCAAT  597
                 ||||  | |    ||   |    |||   |   |     |||    || || ||||||||  || ||
Sbjct  5157309  AATCCAAATGCAAAAGTTGTATATTTATCATCAGAAAAATTTACAAATGAATTTATTAAC  5157368

Query  598      GCCATCCAAACGAAAAAACAGGAAGCGTTTCGGGAAGAATACCGCAACGTCGACCTGTTA  657
                  | ||  |      ||  ||     |   |  |||||   |  |||| |||||||| ||   | |||
Sbjct  5157369  TCTATTCGTGATAATAAGGCTGTTGATTTTCGTAATAAATATCGCAACGTAGATGTTTTA  5157428
```

Estamos vendo apenas uma parte de um alinhamento mais longo (o alinhamento completo tem 985 posições, e acima estão mostradas apenas 300 posições, ou colunas). Este alinhamento é semelhante àquele visto anteriormente. Mas desta vez, a linha do meio apenas indica as posições onde a base de cima é igual à base de baixo, por meio de traços verticais. Os indicadores do alinhamento no cabeçalho indicam que o porcentual de identidade neste particular alinhamento é de 66% e o seu e-value é de 10^{-54}.

Roteiro da atividade 3:

Usaremos informações já descritas nos roteiros das atividades 1 e 2. Para cada gene que você escolheu:

1. Acesse a página que tem a lista dos genes codificadores de proteína do seu genoma-consulta. Encontre e clique no link da coluna **geneID** correspondente ao gene escolhido.
2. Se a coluna geneID não tiver links, siga para as instruções abaixo.
3. Na tela que aparecer, rolar para baixo até que apareça a seção *Reference assembly.* Nesta seção da página, o primeiro segmento se chama *Genomic* (depois vem outro segmento chamado mRNA and Protein(s)). No segmento *Genomic* clique no link chamado FASTA. Esse clique o levará para a sequência em nucleotídeos do gene escolhido, permitindo que você faça cópia da sequência.
4. De posse dessa sequência, repita os passos da atividade 2, mas desta vez escolha a opção **nucleotide blast**. Dentro desta opção existem sub-opções de algoritmo; você deve escolher a sub-opção **blastn** (veja a seção *Program selection*).
5. Reporte seu resultado mesmo que o resultado do blastn seja “no hits” (Tabela 3 do relatório)
6. Neste passo você vai escolher um genoma-alvo diferente. Escolha um genoma alvo que seja do mesmo gênero que seu genoma-consulta, mas de espécie diferente. Por exemplo, se seu genoma-consulta for *Xanthomonas citri*, você poderia escolher como genoma-alvo *Xanthomonas campestris.* Repita o processamento e coloque os resultados na Tabela 4 do relatório.

Caso a coluna geneID não tenha links:

Neste caso será necessário baixar um arquivo com as sequências nucleotídicas de todos genes do seu genoma-consulta (infelizmente – não tem outro jeito), e depois abrir esse arquivo e achar manualmente seu gene dentro do arquivo.

Para baixar o arquivo correto, coloque no seu browser o seguinte endereço:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&mode=text&rettype=fasta_cds_na&id= accession number do genoma-consulta

Exemplo:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&mode=text&rettype=fasta_cds_na&id=**NZ_CP026082.1**

Ao fazer isso, deverá se abrir uma janelinha de salvamento de um arquivo chamado **sequence.txt** em seu computador. Esse arquivo contém as sequências em nucleotídeos de todos os genes do seu genoma-consulta. Para quem usa windows, sugiro abrir o arquivo com o utilitário WordPad. Para achar seu gene nesse arquivo, busque pelo accession number da proteína (esse aparece na coluna *protein product* que você viu no finzinho do roteiro da atividade 1).

Nota bene 3: esses arquivos são relativamente grandes (alguns megabytes; varia conforme o tamanho do genoma).

Se por qualquer motivo você tiver dificuldade em recuperar essa sequência, escreva ao professor, indicando o accession number do seu genoma-consulta e o accession number da proteína de interesse.

Nota bene 4: Altere o parâmetro *word size* do BLASTN de 11 para 7. Esse parâmetro é encontrado na seção *Algorithm parameters* que aparece no final da página do BLAST, logo depois do botão de execução. Para ver os parâmetros disponíveis basta clicar no sinal de '+'.

Como reportar resultados da Atividade 3

Este relatório deve seguir basicamente as mesmas instruções do relatório da atividade 2.

Pergunta-chave desta atividade: Por que os alinhamentos da atividade 3 são piores que os alinhamentos da atividade 2 (para o mesmo par consulta-alvo)?

Critério de correção dos relatórios

Serão avaliados os seguintes 10 quesitos, com 1 ponto para cada quesito

1. A1: fez atividade 1
2. A1: escolheu genes relevantes em biologia molecular mencionados nas aulas
3. A1: apresentou a ficha completa de cada gene
4. A1: explicou corretamente a função de cada gene, mesmo que tenha escolhido genes em desacordo com o estipulado em (2)
5. A2: fez atividade 2
6. A2: apresentou 3 alinhamentos completos
7. A2: comentou qualidade de cada alinhamento e comparou os 3 alinhamentos entre si, apontando corretamente qual o melhor e qual o pior
8. A3: fez atividade 3
9. A3: apresentou 3 alinhamentos completos
10. A3: analisou os alinhamentos obtidos conforme (7) e respondeu à pergunta-chave.

Modelo de Relatório

disciplina QBQ102 – Introdução a Bioquímica e Biologia Molecular

2021s1

Projeto de Bioinformática

Nome completo do aluno

Número USP

Atividade 1

nome do organismo / genoma-consulta utilizado; indique se o organismo é bactéria ou arquéia

Tabela 1: Genes escolhidos

gene	locus tag	start	stop	fita	tamanho (AA)	nome da proteína
1						
2						
3						

sequência em aminoácidos do gene 1

Breve parágrafo descrevendo a função do gene 1, *com referência ao slide de alguma aula onde ele foi mencionado*

sequência em aminoácidos do gene 2

Breve parágrafo descrevendo a função do gene 2, *com referência ao slide de alguma aula onde ele foi mencionado*

sequência em aminoácidos do gene 3

Breve parágrafo descrevendo a função do gene 3, *com referência ao slide de alguma aula onde ele foi mencionado*

Atividade 2

nome do organismo / genoma-alvo utilizado; <u>indique se o organismo é bactéria ou arquéia</u>

Alinhamento do gene 1; não deixe de indicar o cabeçalho, mostrando qual foi a proteína hit

Alinhamento do gene 2; não deixe de indicar o cabeçalho, mostrando qual foi a proteína hit

Alinhamento do gene 3; não deixe de indicar o cabeçalho, mostrando qual foi a proteína hit

Tabela 2: descrição dos alinhamentos de proteínas

gene	tamanho do alinhamento (AA)	% identidade	e-value	cobertura do gene consulta (%)	cobertura do hit (%)
1					
2					
3					

Comentários sobre os alinhamentos obtidos

Atividade 3

nome do organismo / genoma-alvo designado; indique se o organismo é bactéria ou arquéia

Tabela 3: descrição dos alinhamentos em nucleotídeos para o genoma-alvo designado

gene	tamanho do alinhamento (nt)	% identidade	e-value	cobertura do gene consulta (%)
1				
2				
3				

Comentários sobre os alinhamentos obtidos

nome do organismo / genoma-alvo escolhido; indique se o organismo é bactéria ou arquéia

Tabela 4: descrição dos alinhamentos em nucleotídeos para o genoma-alvo escolhido (o genoma-alvo deve ser o mesmo para os 3 genes!)

gene	tamanho do alinhamento (nt)	% identidade	e-value	cobertura do gene consulta (%)
1				
2				
3				

Comparação entre Tabelas 2, 3 e 4.