Estudo dirigido: Filogenia Molecular

AULA 24

Meta da aula

Apresentar a utilização da internet como ferramenta para o estudo da Filogenia Molecular.

objetivos

Esperamos que, após o estudo do conteúdo desta aula, você seja capaz de:

- Demonstrar como utilizar os bancos de dados genéticos da internet.
- Usar programas de busca, alinhamento e identificação de seqüências de ácidos nucléicos e proteínas.

Pré-requisito

Para acompanhar esta aula, é importante que você tenha claro o conceito de Filogenia Molecular, apresentado na Aula 23 desta disciplina.

INTRODUÇÃO

Repare no título desta aula: estudo dirigido; ela foi programada para ser um roteiro de aula prática e deve ser realizada em um dos pólos. Com ela vamos aprender, na internet, a acessar seqüências e a realizar uma pesquisa de busca e alinhamento de seqüências similares.

BIOINFORMÁTICA: COLEÇÃO E INTERPRETAÇÃO DE DADOS

Você viu, na aula passada, que uma das conseqüências do desenvolvimento das técnicas de Biologia Molecular foi a necessidade de utilizar computadores de alta velocidade, capazes de processar a análise de seqüências de nucleotídeos ou aminoácidos e de aplicar métodos de inferência filogenética.

Junto com os avanços da Biotecnologia ocorreu uma explosão na quantidade de informações sobre seqüências de genes e proteínas. Para que essas informações se tornassem úteis, fez-se necessário um acesso fácil a elas e uma maneira de compará-las com outros dados de seqüência.

Assim, em 1988, o governo americano criou um centro nacional de informação tecnológica, o NCBI (do inglês *National Center for Biotechnology Information*), como fonte de referência de informações em Biologia Molecular. O NCBI cria bancos de dados públicos e faz intercâmbio com outros centros internacionais de pesquisa, como o EMBnet (*The European Molecular Biology Network*) e o DDBJ (*DNA Database of Japan* – bancos europeu e japonês, respectivamente).

O conjunto de informações inclui o banco de dados de seqüências de DNA e de seqüência e estrutura de proteínas. Programas eficazes de pesquisa e comparação permitem a rápida comparação de seqüências.

A disponibilidade de toda essa informação molecular e a relativa facilidade de analisá-la levaram, de fato, ao desenvolvimento de uma nova área: a Bioinformática.

ACESSO AO BANCO DE DADOS GENÉTICOS

O primeiro passo será entrar na página eletrônica do NCBI, no endereço: http://www.ncbi.nlm.nih.gov (**Figura 24.1**).

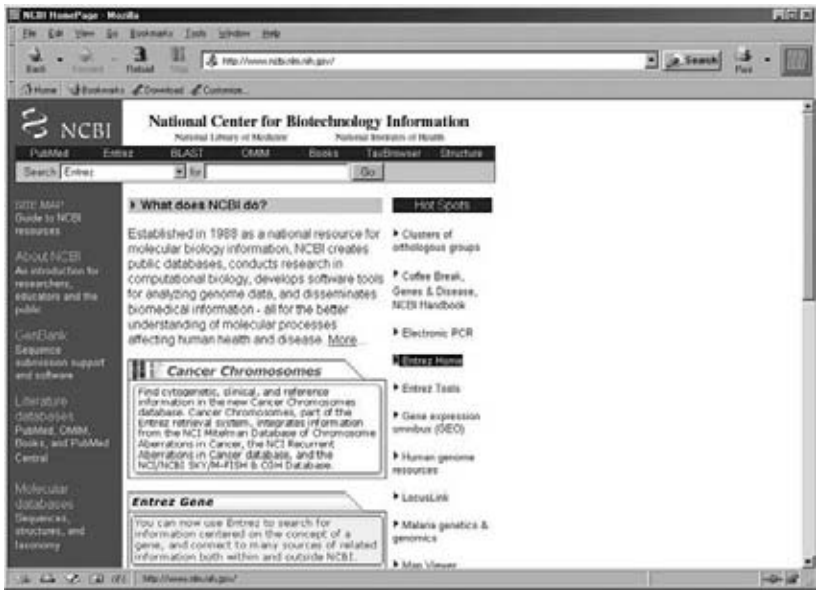


Figura 24.1: Sítio do NCBI na internet.

Em seguida, vamos alterar a busca, em inglês: *Search*, para nucleotídeos (em inglês *Nucleotide*), colocando o cursor na seta para baixo. Você verá uma tela como a apresentada na **Figura 24.2**. Clique no quadrinho *Go* para acessar a próxima página.

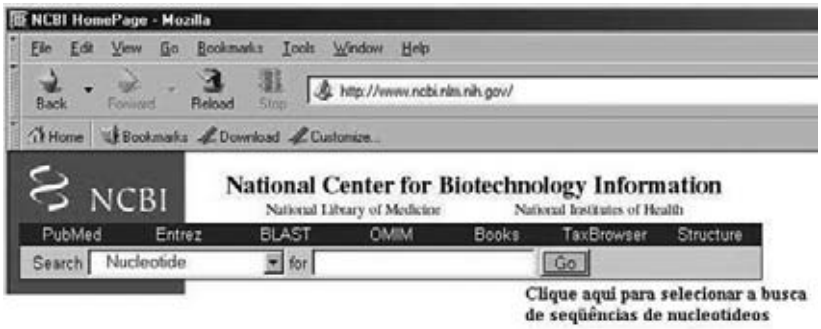


Figura 24.2: Tela com alteração da busca para *Nucleotide*.

Esta será a tela exibida (**Figura 24.3**):

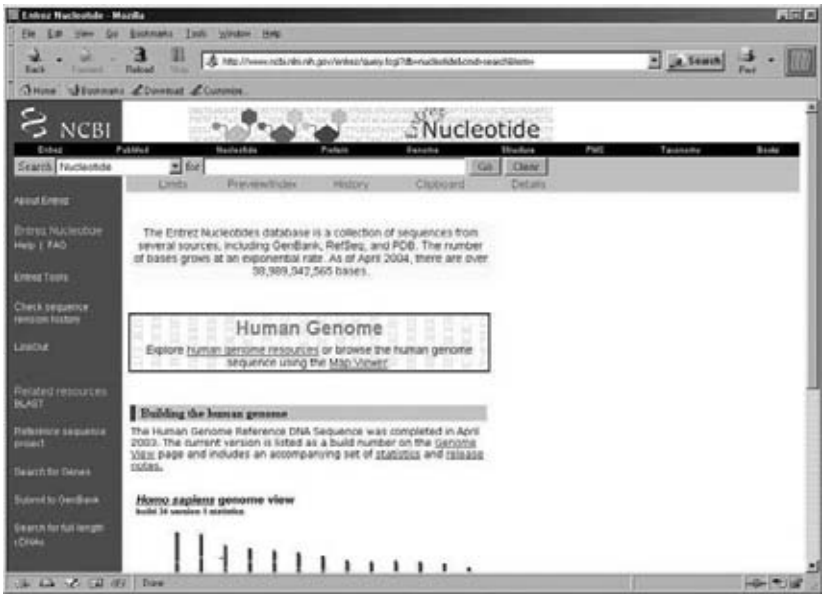


Figura 24.3: Tela apresentando a página de busca de nucleotídeos no banco de dados.

Bem, agora que estamos na página e já escolhemos a molécula (formada por nucleotídeos), que gene vamos estudar? Pode ser qualquer um, mas, como a página está em inglês, teremos de utilizar os nomes na mesma língua. Que tal o gene para a enzima amilase? Essa enzima atua na digestão de amido, um polímero de açúcar e um dos principais componentes da batata. Amilase em inglês se escreve da mesma forma que em português. Assim, vamos escrever amilase no quadro vazio após a palavra '*for*'.

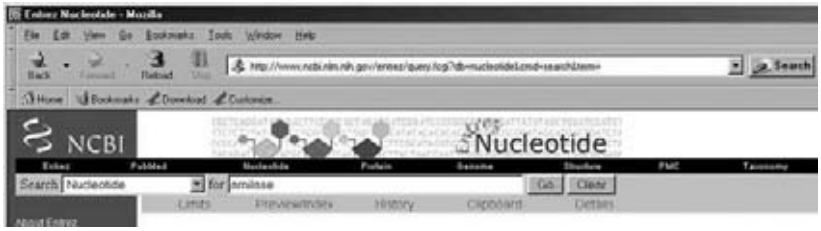


Figura 24.4: Tela apresentando a página de busca para o gene da amilase.

Clique no quadrinho *Go* para realizar a busca.

O resultado estará em forma de uma lista de ocorrências sempre precedidas de uma notação que corresponde ao código de acesso. Por exemplo, na **Figura 24.5**, a primeira seqüência tem código de acesso AW756751 e se refere ao RNA mensageiro de uma provável beta-amilase.

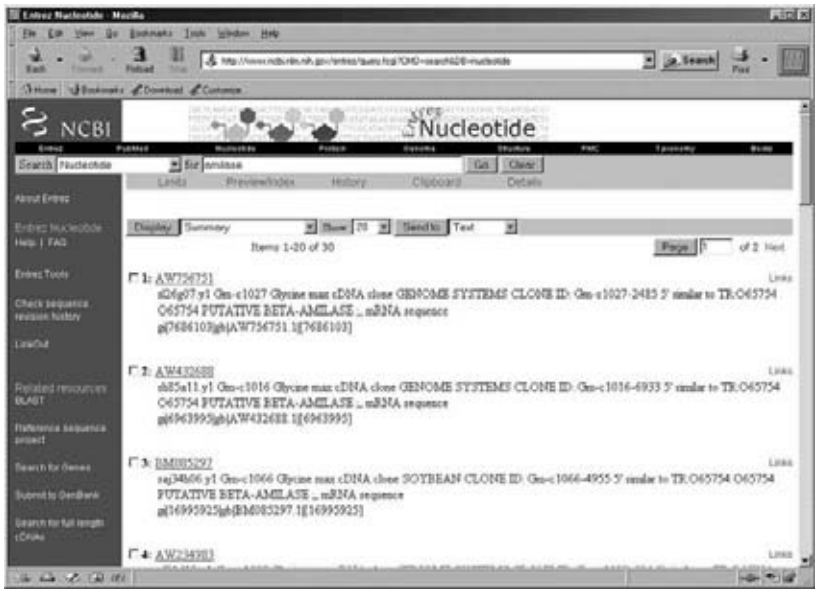


Figura 24.5: Tela apresentando os resultados da busca para o gene da amilase.

Nessa etapa de nossa atividade, se optarmos por utilizar a seqüência dessa beta-amilase para buscar no banco de dados de nucleotídeos outras seqüências homólogas, devemos tratar a seqüência escolhida e colocá-la em um formato que seja reconhecido pelos sistemas de busca. O primeiro passo é abrir a seqüência na lista, através de um duplo clique em cima do número de acesso. A **Figura 24.6** apresenta o que você verá na tela.

Figura 24.6: Tela apresentando a seqüência com número de acesso AW756751.

O **Quadro 24.1** apresenta outro exemplo de seqüência obtida do NCBI.

Quadro 24.1: Exemplo de seqüência retirada do NCBI

1. AF201671 . Megabalanus californicus...[gi:6694386]
LOCUS AF201671 837 bp DNA INV 16-JAN-2000
DEFINITION Megabalanus californicus 18S ribosomal RNA gene, partial seq.
ACCESSION AF201671
VERSION AF201671.1 GI:6694386
KEYWORDS .
SOURCE Megabalanus californicus.
Eukaryota; Metazoa; Arthropoda; Crustacea; Maxillopoda; Cirripedia;
Thoracica; Balanomorpha; Balanidae; Megabalanus.
REFERENCE 1 (bases 1 to 837)
AUTHORS Harris,D.J., Maxson,L.S. and Crandall,K.A.
TITLE Phylogeny of the Thoracican Barnacles based on 18S rDNA seq
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 837)
AUTHORS Harris,D.J., Maxson,L.S. and Crandall,K.A.
TITLE Direct Submission
JOURNAL Submitted (04-NOV-1999) Zoology, Brigham Young University, 574
Widtsoe Building, Provo, UT 84602, USA
FEATURES Location/Qualifiers
source 1..837
/organism="Megabalanus californicus"
/db_xref="taxon:110524"
rRNA <1..>837
/product="18S ribosomal RNA"
BASE COUNT 196 a 201 c 242 g 198 t

001 gaactactgc gaaagcattt gccgagaatg ttttcattag tcaagaacga aagttagagg
061 ttcgaaggcg atcagatacc gccctagttc taaccgtaaa cgatgtcgac cagcaatccg
121 caacggtcac tacacggact gtgcgggcag cttccccggg gaaaccagag tttttggact
181 ccgggggaag tatggttgca aagctgaaac ttaaaggaat tgacggaagg gcaccaccag
241 gagtggagct tgcggcttaa tttgactcaa cacgggacaa ctcaccaggc ccggacaccg
301 taaggattga cagactgata gctctctctt gattcagtgg gtggtggtgc atggccgttc
361 ttagttggtg gagtgatttg tctggtttat tccgataacg aacgagactc tggcctatta
421 aacttgacac tgtccgtctc ttgtgacggc ggtgcgcttc ttagagggat catcggcgtc
481 ccagccgaag gaaagggagc aataacaggt ctgtgatgcc cttagatgtt ctgggctgca
541 cgcgtgttac actgaagtgg tcagcgcgcc gttcaacacc cctctccgtg aggagcttgg
601 gcaaacgttt gaacccttttt cgtgatggga attgggggtt gcaattgtcc cccatgaacg
661 aggaattcca agtaagcgca ggtcactagc ctgcgttgat taagtccctg ccctttgtac
721 acaccgcccg tcgctactac cgatggatga tttggtgagg tcgcctagac tggctgctgg
781 cttcggccgt gcggccggga agacgcccaa acttggtcgt ctagaggaag taaaagt

Repare que a quantidade de informação é variável de seqüência para seqüência. No exemplo do **Quadro 24.1**, cada linha entre o número de acesso e o início da seqüência de nucleotídeos pode ser interpretada da seguinte forma:

- A 1ª linha apresenta o locos (número de acesso, identificador com até 10 caracteres), o comprimento em pares de base (bp) da seqüência, o tipo de molécula, o setor do GenBank onde ela está depositada (nesse exemplo, INV é a seção de invertebrados) e a data de submissão.
- A 2ª linha apresenta a definição resumida do organismo e seu gene.
- Na 3ª linha temos mais uma vez o número de acesso da seqüência no GenBank (este é freqüentemente citado em artigos, de forma que qualquer leitor pode localizar a seqüência utilizada pelos autores); esse não se altera com a modificação da seqüência.
- Na 4ª linha temos o identificador de versão ou identificador único da seqüência; esse número aponta para eventuais alterações na mesma.
- Na 5ª linha estão as palavras-chave.
- Da 6ª à 8ª linha está apresentada a classificação do organismo.
- Da 9ª à 17ª linha estão listadas as referências bibliográficas, o local de descrição da seqüência e a identificação no Medline e PubMed (banco de referências da mesma rede).
- Por fim, da 18ª à 24ª linha listam-se as *features*, que são características adicionais e informações úteis para pesquisa em Biologia.

Continuando a nossa atividade, agora precisamos colocar a seqüência da amilase em um formato próprio para comparações e alinhamentos. Esse formato chama-se FASTA e, para obtê-lo, basta posicionar o cursor na seta apontada para baixo na janela após o botão onde está escrita a palavra *Display*.

Note que, quando você abriu a seqüência, a janela *Display* estava no formato *default* (padrão) e agora deverá estar no formato FASTA. Para visualizar a seqüência nesse formato, basta clicar no botão *Display* (**Figura 24.7**).

Figura 24.7: Tela de entrada do Blast na internet.

Obtendo as seqüências no formato FASTA, estaremos prontos para utilizar uma ferramenta de busca do próprio sítio do NCBI, o BLAST, que pesca seqüências relacionadas à nossa seqüência-alvo.

ATIVIDADE 1

Que tal tentar, para outros genes, o que fizemos até agora para o gene da amilase?

COMENTÁRIO

*Esta atividade não tem uma "resposta certa"! Tente genes para proteínas e enzimas que conhecemos, tais como: actina, peroxidase, catalase, polimerase etc. Lembre-se de utilizar as palavras em inglês (**actin, peroxidase, catalase, polymerase**).*

OBTENÇÃO DE SEQÜÊNCIAS RELACIONADAS ATRAVÉS DE BUSCA POR BLAST (BLAST *SEARCH*)

Vamos, então, conectar com o NCBI Blast no endereço http: //www.ncbi.nlm.nih.gov/blast (**Figura 24.8**).

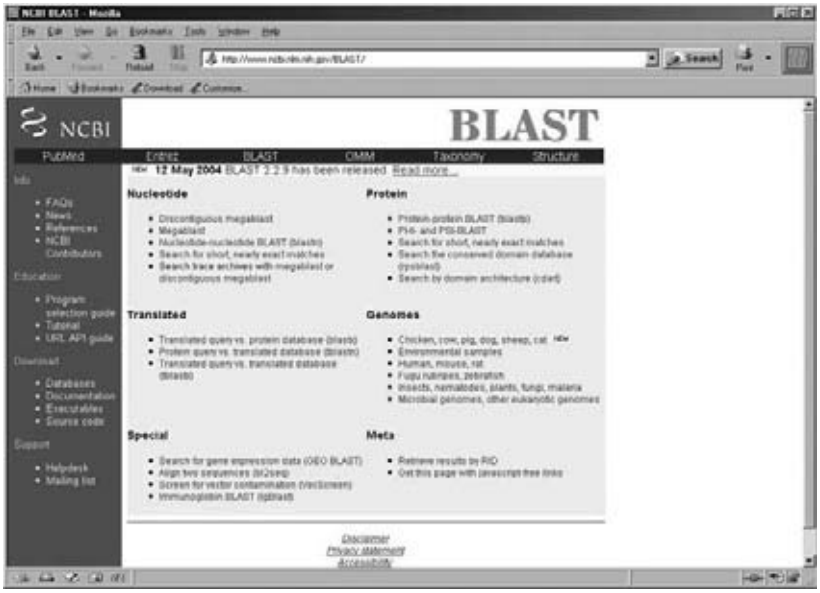


Figura 24.8: Seqüência AW756751 no formato FASTA.

Selecione, com seu cursor, a opção *Nucleotide-nucleotide BLAST (blastn)* na janela de início. Dessa forma, você está optando por utilizar uma seqüência de nucleotídeos para "pescar" outra seqüência de nucleotídeos que seja similar à utilizada como "isca".

Caso você queira utilizar seqüências de aminoácidos, opte pela função *Protein-protein BLAST (blastp)*.

Você verá uma janela como a mostrada na **Figura 24.9**.

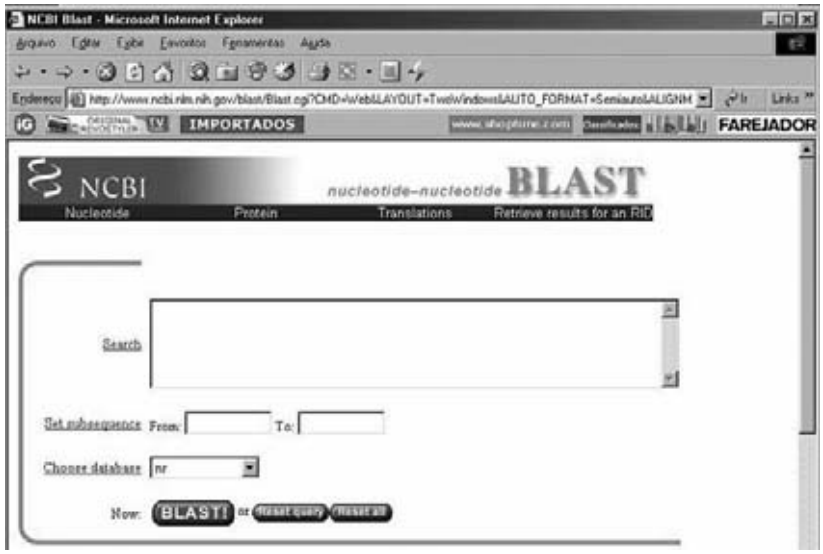


Figura 24.9: Busca de seqüências homólogas no BLAST.

Entrar na caixa de texto, após a palavra *Search*, e selecionar a seqüência de nucleotídeos no formato FASTA. Você pode manter o nome da seqüência, reduzindo a informação após o código de acesso ou, melhor ainda, colar apenas a seqüência de nucleotídeos na janela.

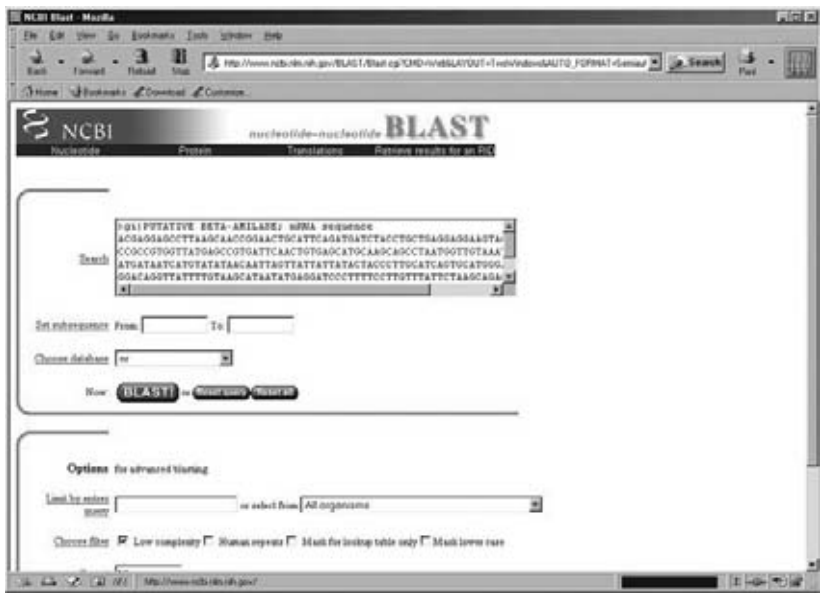


Figura 24.10: Busca de seqüências homólogas à AW756751 no BLAST.

Clique no botão BLAST! Aguarde a próxima tela.

Você verá uma janela informando que sua requisição foi submetida. Essa tela também informará, aproximadamente, quanto tempo levará a consulta. A **Figura 24.11** mostra a aparência desses resultados.

Figura 24.11: Tela de confirmação do requerimento de busca no BLAST.

A tela da **Figura 24.11** informa, em inglês, como visualizar a busca por seqüências relacionadas àquela inserida na caixa de texto; em outras palavras, indica que para checarmos os resultados da busca devemos apertar o botão *Format*!

Os resultados são apresentados primeiramente em uma forma gráfica, com uma série de barras horizontais coloridas. A cor da barra corresponde à qualidade do alinhamento, que é influenciada pelo tamanho da seqüência a ser comparada. Quanto maior a seqüência examinada, maior pode ser a qualidade do alinhamento. As cores dos grupos mostram a qualidade dos alinhamentos em ordem decrescente, com o melhor sendo mostrado primeiro.

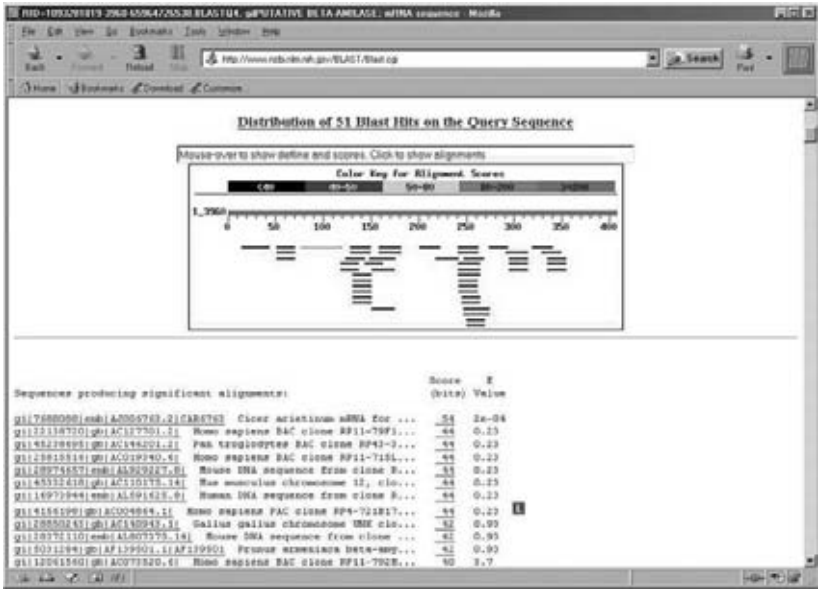


Figura 24.12: Parte da tela de resultados do BLAST.

Abaixo da série de barras horizontais está uma listagem das seqüências alinhadas correspondentes. O primeiro registro corresponde à primeira barra e assim por diante. Em uma coluna à direita de cada registro são colocados os valores E (E value). Esse número indica a probabilidade de os alinhamentos com essa qualidade poderem ser obtidos, ao acaso, em um banco de dados no mesmo tamanho do utilizado na pesquisa. Ou seja, quanto menor o E value, mais similar determinada seqüência será quando comparada à seqüência "isca".

O **Quadro 24.2** apresenta outro exemplo de busca BLAST, dessa vez para seqüências homólogas a uma glicoproteína de plaquetas humanas.

ATIVIDADE 2

Utilize as seqüências que você obteve na atividade anterior, sempre no formato FASTA, para fazer um BLAST.

COMENTÁRIO

Esta atividade também não tem resposta fechada! Dependerá das proteínas e enzimas utilizadas.

Quadro 24.2: Exemplo de resultado de busca BLAST

BLASTN 2.2.6 [Apr-09-2003]

RID: 1067808421-16791-2226431.BLASTQ3

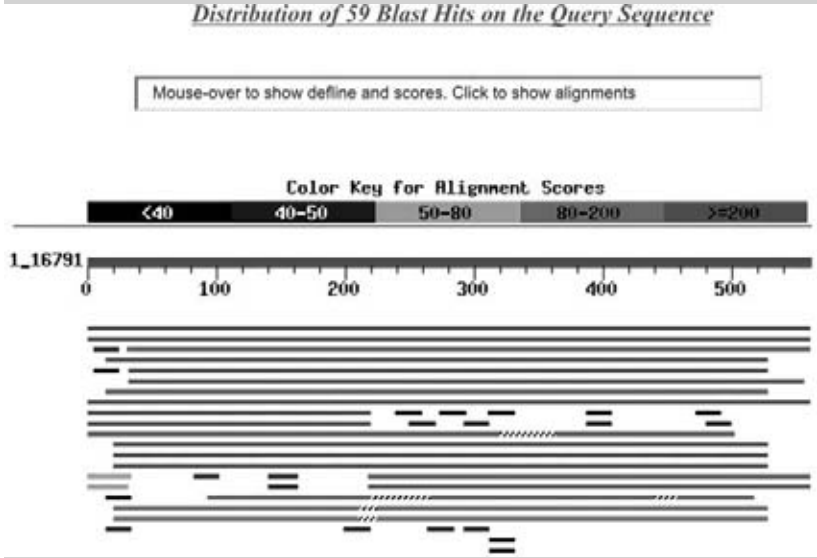
Query = (560 letters)

>gi|6006009|ref|NM_000419.2| Homo sapiens integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41B) (ITGA2B), mRNA

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences) 1,957,115 sequences; 9,384,639,548 total letters

Taxonomy reports

Distribution of 59 Blast Hits on the Query Sequence



Sequences producing significant alignments: Score E (bits) Value

gi|6006009|ref|NM_000419.2| Homo sapiens integrin, alpha 2b... 1110 0.0

gi|183510|gb|M34480.1|HUMGPIIBA Human platelet glycoprotein... 1102 0.0

gi|190067|gb|J02764.1|HUMPLG2B Human platelet membrane glyc... 1049 0.0

gi|5733733|gb|AF170524.1|AF170524 Canis familiaris glycopro... 636 e-179

gi|5932027|gb|AF153316.1|AF153316 Canis familiaris platelet... 603 e-169

gi|32481515|gb|AY322154.1| Equus caballus platelet glycopro... 571 e-160

gi|5805336|gb|AF170526.1|AF170526 Sus scrofa glycoprotein I... 517 e-144

gi|11693431|gb|AC007722.9|AC007722 Homo sapiens chromosome ... 436 e-119

gi|183505|gb|M33319.1|HUMGPIIB1 Human platelet glycoprotein... 436 e-119

gi|183448|gb|M22568.1|HUMGP2B1 Human platelet glycoprotein ... 436 e-119

gi|5805340|gb|AF170528.1|AF170528 Oryctolagus cuniculus gly... 381 e-102

gi|6754375|ref|NM_010575.1| Mus musculus integrin alpha 2b ... 278 1e-71

gi|5918985|gb|AF170316.1| Mus musculus glycoprotein IIb (GP... 278 1e-71

gi|7262858|gb|AF166384.1|AF166384 Mus musculus integrin cel... 278 1e-71

gi|2828776|gb|AC003043.1|AC003043 Homo sapiens chromosome 1... 254 1e-64

ALINHAMENTO DE SEQÜÊNCIAS UTILIZANDO O CLUSTALX OU W

Após obtermos as seqüências relacionadas à nossa seqüência alvo, através do BLAST, o próximo passo será proceder o alinhamento dessas seqüências. As seqüências alinhadas devem ser arrumadas umas em relação às outras, de forma que cada posição de um nucleotídeo corresponda a uma posição na molécula do ancestral comum, a partir do qual todas as seqüências evoluíram.

O alinhamento pode e deve ser feito manualmente. Outra opção é utilizar um programa próprio para esse fim. Nesta aula, vamos empregar o ClustalX (THOMPSON *et al.*, 1997), que pode ser utilizado pela internet ou pode mesmo ser instalado a partir dela.

Selecione cinco seqüências obtidas no BLAST, coloque-as no formato FASTA e salve-as em Word (em formato *.txt) ou Bloco de Notas.

Observe que o programa ClustalX só considera significativos os 30 caracteres iniciais, após o sinal de determinação de seqüência (“>”). Isto é, se houver uma semelhança muito grande entre os nomes de diferentes seqüências, o programa não as aceitará para processamento. Dessa forma, é importante editar o nome após o sinal “>”, atentando para que as cinco seqüências tenham denominações diferentes.

!

Outro detalhe importante do arquivo de seqüências: todos os caracteres devem estar na fonte “Courier New”. Toda a seqüência deve estar também em caracteres maiúsculos e, entre o “nome” de cada seqüência e a seqüência em si, deve haver um “parágrafo” (clique a tecla “enter”) – não há importância se o mesmo ocorrer entre cada linha da seqüência.

Após editar seu arquivo de seqüências, conecte o programa ClustalX em http://newfish.mbl.edu/Course/Software/ClustalX ou instale o programa ClustalX no seu drive "C:".

Quando você abrir o programa, aparecerá uma janela como a mostrada na **Figura 24.13**.

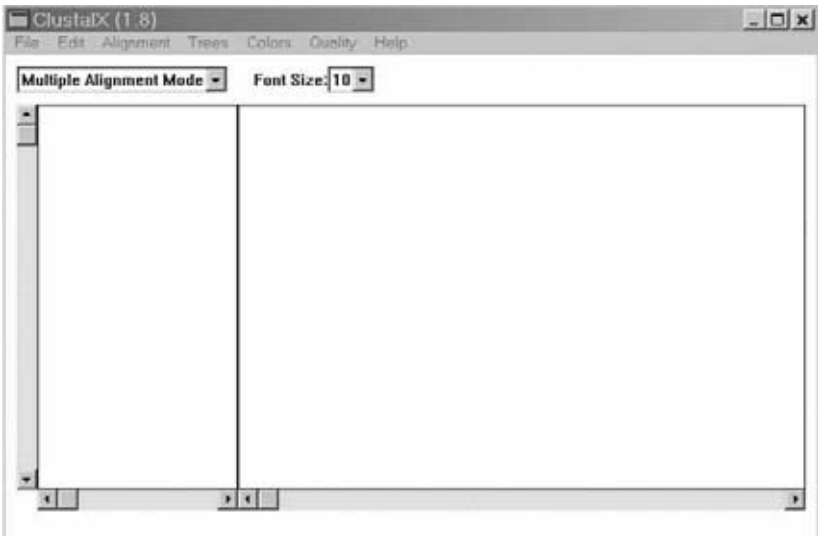


Figura 24.13: Tela de entrada do ClustalX.

O passo seguinte é copiar seqüências para a janela do programa ou carregar os arquivos *.txt (*File/Load sequences*) e mandar alinhar (*Alignment/Do complete alignment*). Veja as **Figuras 24.14 a 24.16.**

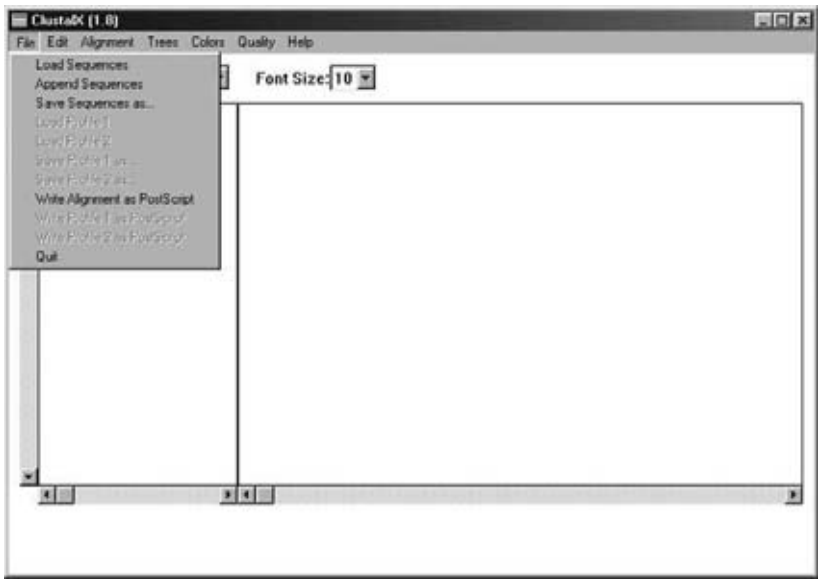


Figura 24.14: Tela de carregar arquivos do ClustalX.

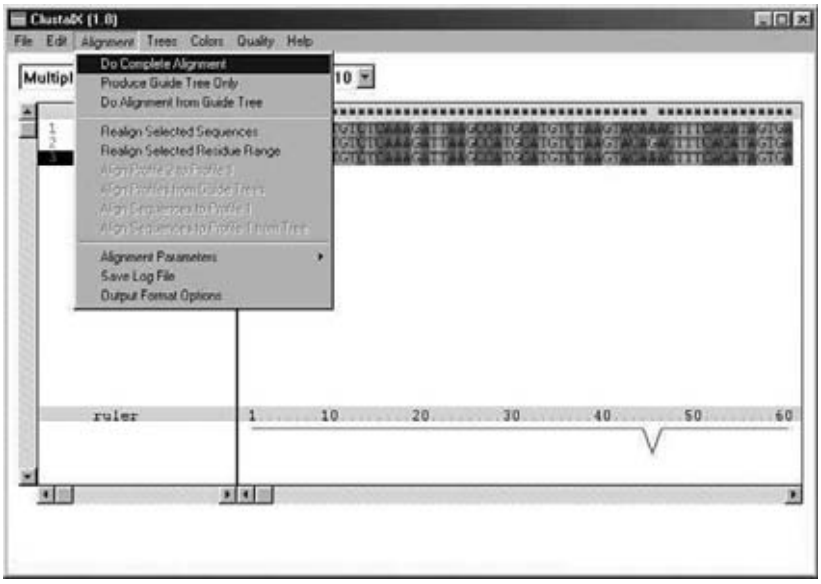


Figura 24.15: Tela de alinhamento do ClustalX.

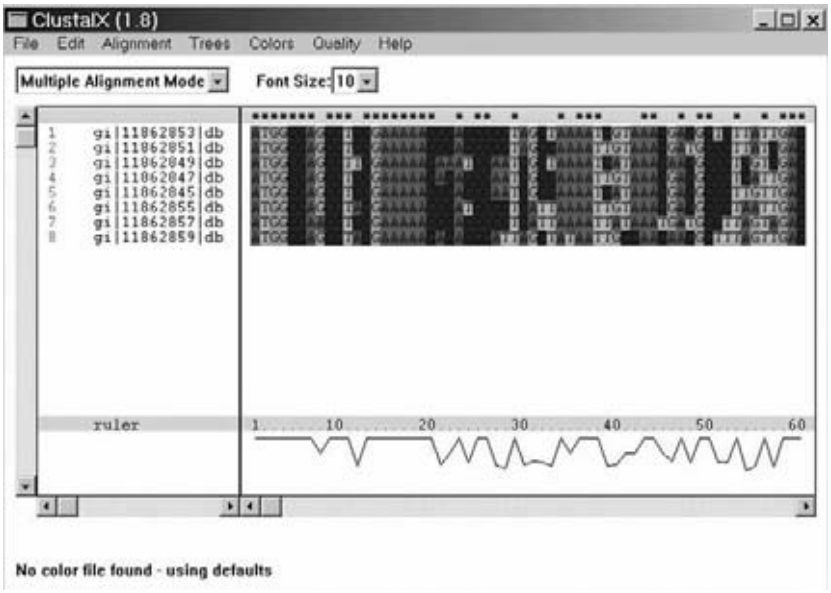
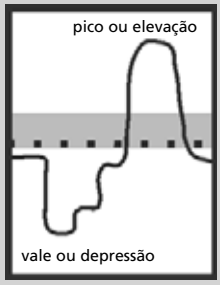


Figura 24.16: Formato de apresentação do alinhamento das seqüências pelo Clustal X.

Pronto! As seqüências estão alinhadas. Note que os nucleotídeos são numerados e que o programa mostra um tipo de régua, na parte inferior, com **vales** que correspondem às posições de diferença de seqüência. Na parte superior da tela, os asteriscos (*) correspondem à identidade total dos nucleotídeos daquela posição.

Vales, picos, depressões e elevações. Esses termos referem-se à posição em relação à linha base da régua. Se a posição está acima da linha, trata-se de um pico ou elevação; se está abaixo da linha, tem-se um vale ou depressão. Veja a ilustração que se segue:



ATIVIDADE 3

Alinhe manualmente as seguintes seqüências obtidas para três organismos:
GATCCTCGGATTGGTCCCGGGACGGCGGGAAACCATCGACCCGGCGTGCCGAG
CTCGGATTGGTCCCGGGACGGCGGGCAACCGCTGACCCGGCGTGCCGAGAAGA
CCTCGGATTGGCCCCGGGATGGTGGGCGACCGCCGACTCGGAGGCCGAGAAGA

RESPOSTA COMENTADA

As seqüências foram alinhadas com base nas regiões conservadas (invariáveis). Em negrito estão as regiões variáveis.

*GATCCTCGGATTGGTCCCGGGACGGCGGG**A**AACC**AT**CGACCCGGC*
GTGCCGAG----
*----CTCGGATTGGTCCCGGGACGGCGGGCAACCGC**T**GACCCGGCGT*
GCCGAGAAGA
*---CCTCGGATTGG**C**CCCGGGA**T**GG**T**GGGC**G**ACCGCCGAC**T**CGG**A**G-*
GCCGAGAAGA

CONCLUSÃO

Esta aula fornece os princípios básicos de busca e tratamento de seqüências de genes e proteínas utilizando ferramentas da internet. Com esse conhecimento, é possível obter resultados para uma monografia ou até mesmo para um artigo científico! Ainda são poucos os profissionais da área de Bioinformática, e um domínio das técnicas e da teoria por trás dessa área habilita o biólogo a ocupar um espaço vago no mercado de trabalho. Portanto, se você gostou dessa parte do curso, invista em aprimorar seu conhecimento em cursos de extensão. Boa sorte!

RESUMO

O centro americano de informação tecnológica, NCBI, criou bancos de dados públicos de seqüências de DNA, seqüência e estrutura de proteínas. Nesta aula, aprendemos a acessar seqüências, a realizar pesquisa de busca de seqüências similares no sítio do NCBI e, ainda, a alinhar seqüências com o programa ClustalX da internet.

ATIVIDADES FINAIS

1. O que é o NCBI?

_____

_____

_____

_____

_____

RESPOSTA

O National Center for Biotechnology Information *(NCBI), localizado nos Estados Unidos tornou-se a mais importante fonte de referência de informações em Biologia Molecular. O NCBI possui bancos de dados públicos de seqüências que podem ser acessados gratuitamente por qualquer pessoa. Essa instituição faz o intercâmbio desses dados com outros centros internacionais de pesquisa.*

2. Como fazer para procurar proteínas homólogas à amilase em diferentes organismos? Por exemplo: camundongo, gorila, besouro e caranguejo.

_____

_____

_____

_____

_____

RESPOSTA

Você tem duas opções básicas: 1) procure, entre todas as seqüências pescadas para amilase, alguma que seja originada dos organismos citados; ou 2) faça a busca no NCBI com duas palavras entre aspas; por exemplo, para a amilase de camundongo use "amilase and mouse".

3. No alinhamento manual de quatro seqüências, o que devemos fazer nas posições que não apresentam correspondência?

_____

_____

_____

_____

_____

RESPOSTA

É necessária a inserção de lacunas (simbolizadas por '–') quando uma seqüência não apresenta correspondência com outra. Essas lacunas são conseqüência de inserções ou deleções ocorridas de forma diferenciada em cada molécula.

AUTO-AVALIAÇÃO

Esta aula implica tempo para visitar o pólo e exercitar as buscas na internet. Contudo, você tem toda a liberdade de fazer o mínimo exigido pela aula ou investir bastante e até conseguir gerar um trabalho interessante e original. Portanto, mãos à obra! Ouse nas suas buscas, trabalhe com genes e organismos que sejam do seu interesse. Pode ser uma enzima que cause uma deficiência metabólica, como a lactase (enzima necessária à digestão de produtos derivados do leite), ou a queratina, que é um componente do seu cabelo e da carapaça de alguns invertebrados.

INFORMAÇÕES SOBRE A PRÓXIMA AULA

Na próxima aula, você vai estudar a evolução humana sob um enfoque molecular. Vai descobrir quais foram as contribuições da Biologia Molecular para o estudo da evolução humana e acompanhar os estudos de filogenias e rotas de colonização dos continentes, fundamentos em polimorfismos genéticos.