

Introdução a Bioinformática - CEN0485

Diego Mauricio Riaño Pachón

8 de junho de 2022

Conteúdo

1 Bases de bioinformática	9
2 Ferramentas Unix úteis em bioinformática	11
2.1 Introdução ao sistema Unix	11
2.1.1 A linha de comando	11
2.1.2 Sua home e árvore diretórios	13
2.1.3 Organizando arquivos	15
2.1.4 Algumas operações básicas com arquivos	16
2.2 Formatos de sequência	18
2.2.1 Fasta	18
2.2.2 GenBank	18
2.2.3 Algumas operações básicas com sequências no formato Fasta	20
3 Buscas em banco de dados biológicos	21
3.1 NCBI – Bancos de dados e busca de informações	21
3.1.1 Vamos começar uma visita aos seus bancos de dados	21
3.1.2 Recuperação de Sequências no NCBI com buscas mais específicas	27
3.2 Recuperação de sequência usando SRS@EBI	28
4 Manipulación básica de secuencias	33
4.1 Limpieza de secuencias	33
4.2 Mapa de restriccion	35
4.3 Análisis de la composición del ADN	35
4.3.1 Contenido de G+C	36
4.3.2 Composición monomérica y palabras cortas	36
5 Creación de bases de datos relacionales	37
6 Búsquedas en base de datos biológicas - Segunda parte	41
6.1 PubMed	41

6.1.1	Entendiendo la información en los registros de PubMed	41
6.1.2	Realizando búsquedas	42
6.2	Descarga por lotes usando Entrez	43
6.3	Recuperar todas las secuencias de un organismo o taxon	43
6.4	Recuperar la información publicada sobre un gen	44
6.5	Bases de datos en el European Bioinformatics Institute (EBI)	44
6.5.1	SRS	44
6.5.2	EB-eye	44
6.6	Expasy	45
6.7	Mas ejercicios	45
7	Ontologías en bioinformática: Gene Ontology	47
7.1	Consultas en GO	47
8	Introducción al análisis de redes usando Cytoscape	53
9	Análisis de enriquecimiento de anotaciones de genes	55
10	Comparação de Sequência I - Matrizes de pontos	57
11	EMBOSS	61
11.1	Recuperando sequências de bancos de dados	61
11.2	Seleção de quadros de leitura aberta	63
11.3	Embaralhar/misturar Sequências	63
11.4	Previsão de regiões hidrofóbicas	63
11.5	Alinhamentos	64
12	Comparação de Sequência II - Alinhamentos emparelhados	65
12.1	Matrizes de substituição	65
12.2	Alinhamento Global	65
12.3	Alinhamentos locais	66
12.4	Significado dos alinhamentos	67
13	BLAST: BASIC LOCAL ALIGNMENT SEARCH TOOL	69
13.1	Encontrando la región genómica de un transcripto.	73
13.2	Blast+ en la línea de comandos	73
14	Alinhamientos múltiples	75
14.1	Alinhando as sequências de aminoácidos de TRIM5 de primatas	75
14.1.1	MUSCLE	75

14.1.2	Visualização e edição de Alinhamentos	76
15	PSSMs, Logo de Sequências e HMMs	77
15.1	PSSM	77
15.2	logo de sequências	78
15.3	Modelos ocultos de Markov: HMMs	78
15.3.1	Procurando os domínios de uma proteína	79
15.3.2	Visualização de HMMs	80
16	Diseño de primers para PCR	81
16.1	Diseño de primers usando Quantprime	83
16.2	Crear primers a partir de alineamientos de proteínas	86
17	Montagem de genomas	91
17.0.1	Limpar sequencias	92
17.0.2	Montagem de genoma usando dados Illumina	92
17.0.3	Montagem de genoma usando dados PacBio	93
17.0.4	Avaliando a qualidade das montagens com QUAST	95
17.0.5	Avaliando a qualidade das montagens com BUSCO	95
18	Anotação de Genomas	97
18.0.1	Usando prokka para anotar genomas bacterianos	97
18.0.2	Usando IGV para olhar os resultados de prokka	98
18.0.3	Rondando blast local	99
Apendices		103

Listas de Figuras

1.1	O que é bioinformática?	10
2.1	Ícone do programa da terminal	12
2.2	Terminal no Linux	12
2.3	Árvore de diretórios no Linux	14
2.4	Sistema de permissão no Linux	15
3.1	Página inicial do NCBI	22
3.2	Janela de busca do NCBI	22
3.3	Página inicial do Entrez	24
3.4	Página inicial do SRS	29
3.5	Opções SRS	29
3.6	Opções SRS	30
3.7	Formulário de pesquisa SRS	31
3.8	Critérios de pesquisa avançados	31
4.1	VecScreen: Herramienta para detectar contaminación de vectores.	33
5.1	SQLite Manger en Firefox	38
7.1	Consultas en “Gene Ontology”	48
7.2	Visualización del grafo acíclico dirigido de una sección de GO	49
7.3	Consultas en “Gene Ontology”	50
7.4	Resultados de la consulta en “Gene Ontology”, usando el nombre de gen ANAC092 .	50
7.5	Términos GO asociados al gen ANAC092	51
10.1	Dot Let @ SIB	57
10.2	Adicionar Sequências em Dot Let	58
10.3	botões de controle	58
10.4	Resultado	59
11.1	Recuperando sequências de bancos de dados	62

11.2 Recuperando link de sequências em uniprot	62
13.1 Tipos de BLAST disponíveis no NCBI	69
13.2 Interface web NCBI BLAST usando o programa blastx	70
13.3 Parámetros de búsqueda en BLAST	70
13.4 Resultados blast: gráfica	71
13.5 Resultados blast: hits	72
13.6 Resultados blast: alineamientos	72
14.1 Tela do JalView	76
15.1 Lexa junta-se ao logo da sequência do site	78
15.2 Resultados de Pfam	79
16.1 Creación un proyecto en QUANTPRIME	84
16.2 Adicionando transcritos al proyecto en QUANTPRIME	84
16.3 QUANTPRIME buscando primers para los genes solicitados	85
16.4 Listado de los mejores primers encontrados por QUANTPRIME	85
16.5 Página de información para un par de primers seleccionados	86
16.6 Página de inicio en iCODEHOP	87
16.7 Diseño de primer en iCODEHOP	87
16.8 Diseño de primer en iCODEHOP	88
16.9 Detección de BLOCKS en el alineamiento de secuencias de proteínas. Se diseñaran primers para cada BLOCK	89
16.10 Detección de BLOCKS en el alineamiento de secuencias de proteínas. Se diseñaran primers para cada BLOCK	90
17.1 Montagem de genomas	91
17.2 Tela FASTQC	93
17.3 Visualização da montagem do unicycler	94
17.4 Visualização da montagem do flye	94
18.1 Visualização do arquivo de anotações .gff na ensamblagem	98

Capítulo 1

Bases de bioinformática

A bioinformática é uma disciplina que surge da interação entre biologia, estatística e ciência da computação. (Figura 1.1. Seus principais objetivos são a gestão e análise de grandes volumes de dados, principalmente o produto de novas tecnologias em biologia molecular, como genômica, proteômica e metabolômica, especialmente hoje com o advento de novas tecnologias de sequenciamento de ácidos nucleicos que estão revolucionando a forma como estudamos os genomas. Outro aspecto importante inclui o desenvolvimento de novos métodos computacionais, algoritmos e/ou softwares para a análise desses dados.

De acordo com Philip Bourne (UCSD), “a bioinformática tornou-se a intérprete da linguagem genômica do DNA e está tentando decifrar linguagens mais complexas em que as proteínas são os substantivos, as interações são a sintaxe, as vias metabólicas são frases e os sistemas vivos são o volume completo” (BOURNE, 2004).

Portanto, semelhante à biologia molecular, a bioinformática hoje constitui uma caixa de ferramentas que todo pesquisador de biologia tem que lidar (STEIN, 2008 apresenta um ponto de vista muito interessante).

Neste curso nos concentraremos na análise de dados biológicos, utilizando, na maioria dos casos, ferramentas de livre acesso, a maioria das quais têm melhor desempenho em sistemas operacionais Unix¹.

¹Linux, MacOSX, BSD, etc. Se você quiser tentar ter uma cópia em sua home ou escritório de qualquer um desses sistemas operacionais, recomendo que você use o VirtualBox (ou outra tecnologia de virtualização), para instalar, por exemplo, o Linux dentro do sistema operacional existente, por exemplo, o Windows XP; Claro que é se você tem um computador com pelo menos dois Núcleos e 2GB de RAM, caso contrário é mais conveniente ter um sistema dual boot.

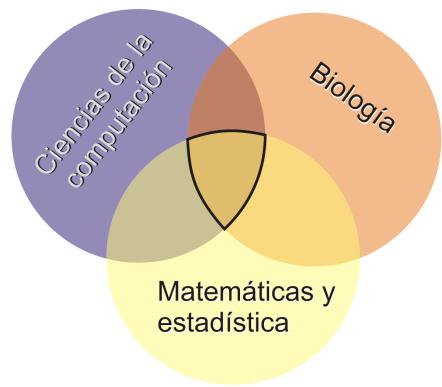


Figura 1.1: A bioinformática é a disciplina que surge da interação de três ciências básicas: Biologia, Matemática e Ciência da Computação. Quando alguns deles dominam o resto, outra disciplina diferente da bioinformática é obtida, por exemplo, se matemática e biologia são mais importantes, obtemos biomatemáticos. É importante que as três ciências-base sejam equilibradas para realizar projetos de bioinformática.

Capítulo 2

Ferramentas Unix úteis em bioinformática

2.1 Introdução ao sistema Unix

O sistema operativo¹ é o conjunto de programas (“software”) que serve como uma interface entre a máquina (‘hardware’) e o usuário, e que permite que este último execute aplicativos. Os sistemas operacionais mais comuns são: Windows (XP, Vista), Unix e MacOS X. Sistemas operacionais semelhantes ao Unix (por exemplo, Linux) são usados principalmente em servidores, mas seu uso em estações de trabalho e desktops está em ascensão. As principais características do Unix são: multitarefas, multi-usuário e portabilidade². A maioria dos Unixes hoje tem uma interface gráfica fácil de usar, a partir da qual você pode realizar quase todas as tarefas de uso diário, como criar documentos, imprimir e navegar na Internet. Além dessa interface gráfica, há uma interface de linha de comando que permite ao usuário executar tarefas muito mais complexas e poderosas. Em seguida, aprenderemos como usar a linha de comando e alguns comandos que facilitam o manuseio de arquivos grandes, usando o Linux como sistema operacional. orientação sobre o uso de vários desses comandos está disponível no apêndice 18.0.3³.

2.1.1 A linha de comando

A linha de comando é acessada através de um programa de interpretação chamado “shell”⁴. Existem vários tipos de “shell” em Unix. Na maioria das distribuições Linux o “shell” bash é instalado por padrão. Para usar o “shell” ou linha de comando do seu computador, inicie o programa **Terminal**,

¹Mais informações em http://en.wikipedia.org/wiki/Operating_system

²Refere-se a quais programas criados em diferentes Unixes podem ser executados em um ou outro geralmente sem problemas.

³Guias para outros programas comumente usados em bioinformática estão disponíveis em <http://www.embnet.org/en/QuickGuides>

⁴http://en.wikipedia.org/wiki/Unix_shell

que tem um ícone semelhante ao mostrado na Figura 2.1.



Figura 2.1: Ícone do programa da terminal

Clicando (uma ou duas vezes, dependendo da configuração) iniciará o programa **Terminal**, semelhante ao mostrado na Figura 2.2. Este aplicativo dá acesso à linha de comando Linux através de um *prompt*, que informa que o sistema está esperando suas instruções. Na Figura 2.2, o *prompt* consiste na string [user@server]\$, que consiste no nome do usuário que está usando o programa **Terminal**, seguido pelo nome da máquina e pelo símbolo do dólar, imediatamente após tem um cursor piscando esperando por seus comandos.

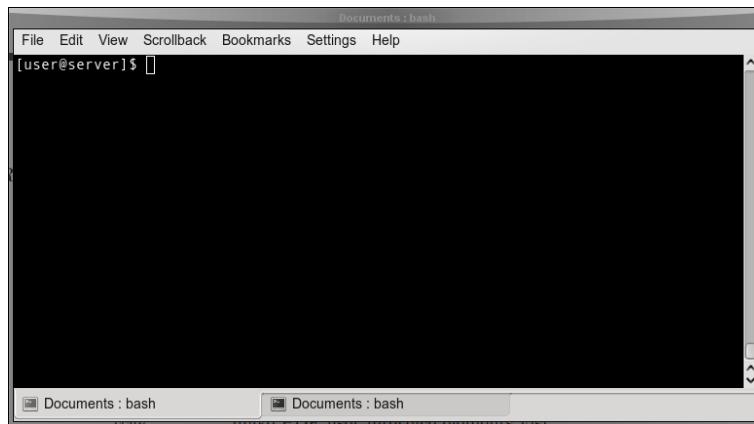


Figura 2.2: Terminal no Linux

O *prompt* pode ser modificado alterando a variável do sistema \$PS1⁵. Vamos alterar o *prompt* para ter certeza de que todos temos o mesmo.

Na sessão **Terminal** execute os comandos conforme mostrado na lista *Alterando prompt*. Na linha 4 salvamos o prompt na nova variável \$SAVE, caso precisemos recuperá-lo. Na linha 5 modificamos o *prompt* atual, \u⁶, indica a nossa “shell” mostrar o usuário atual, \h, mostra o nome da máquina e \w, mostra o diretório atual, o resto de caracteres são exibidos sem qualquer modificação⁷. Compare seu novo *prompt* (línea 6) com o antigo (línea 1), o símbolo ~ refere-se ao diretório da sua home ou ao diretório do usuário, no sistema (vea Sección 2.1.2)

Alterando o prompt

- 1 [user@server]\$
- 2 [user@server]\$ echo \$PS1

⁵<http://tldp.org/HOWTO/Bash-Prompt-HOWTO/c141.html>

⁶Lista de modificadores de *prompt* no bash: <http://tldp.org/HOWTO/Bash-Prompt-HOWTO/bash-prompt-escape-sequences.html>

⁷Exercício opcional: Como tornar permanente a alteração de *prompt*?

```

3  [\u0@h]$
4  [user@server]$ SAVE=$PS1
5  [user@server]$ PS1="[\u0@h:\w] $ "
6  [user@server:~]$
```

Vamos começar interagir com o sistema através de comandos. para começar a executar o comando mostrado na linha 7, wget é um programa para baixar arquivos da rede. A linha 8 ate 22 mostram a saida tipica deste comando, pode mudar levemente do que se amostra no seu **Terminal**. quando este comando termina executar o mostrado na linha 24, que abre o arquivo que você acabou de baixar.

```

____ Baixando arquivos _____
7  [user@server:~]$ wget https://github.com/labbcles/cen0485/raw/main/linux/praticas/file1.tar.gz
8  --2022-03-25 16:26:12-- https://github.com/labbcles/cen0485/raw/main/linux/praticas/file1.tar.gz
9  Resolving github.com (github.com)... 20.201.28.151
10 Connecting to github.com (github.com)|20.201.28.151|:443... connected.
11 HTTP request sent, awaiting response... 302 Found
12 Location: https://raw.githubusercontent.com/labbcles/cen0485/main/linux/praticas/file1.tar.gz [following]
13 --2022-03-25 16:26:18-- https://raw.githubusercontent.com/labbcles/cen0485/main/linux/praticas/file1.tar.gz
14 Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.111.
15 Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
16 HTTP request sent, awaiting response... 200 OK
17 Length: 73501 (72K) [application/octet-stream]
18 Saving to: 'file1.tar.gz'
19
20 file1.tar.gz                                     100%[=====]
21
22 2022-03-25 16:26:19 (141 KB/s) - 'file1.tar.gz' saved [73501/73501]
23
24 [user@server:~]$ tar xzf file1.tgz
```

2.1.2 Sua home e árvore diretórios

Cada usuário em um sistema Unix tem um espaço reservado, geralmente dentro do diretório “/home”, em um subdiretório que tem o mesmo nome do usuário, e.g., para o usuário ”diriano” seu diretório pessoal é “/home/diriano”, e é chamado de diretório ”home” ou diretório de usuário. A primeira vez que você faz login no Linux ou **Terminal**, está localizado em seu diretório home. se a qualquer momento você não sabe onde você está, você pode usar o comando mostrado na linha 25 para localizar o caminho dentro da árvore do diretório em que está localizada. é importante que você note que diretórios usam o caracter “/” para se referir a um caminho subdiretório aninhado como mostrado na linha 26 no listado *Navegando pela árvore de diretórios*.

A árvore diretório refere-se à organização aninhada de diretórios no sistema de arquivos (Figura 2.3), semelhante à organização de diretórios no Microsoft WindowsTMque pode ser visto com o **Windows Explorer**.

Com o comando “listar” (Línea 27) exibe os diretórios e arquivos que estão no diretório atual. Este comando recebe argumentos/opções que permitem obter mais informações sobre arquivos e

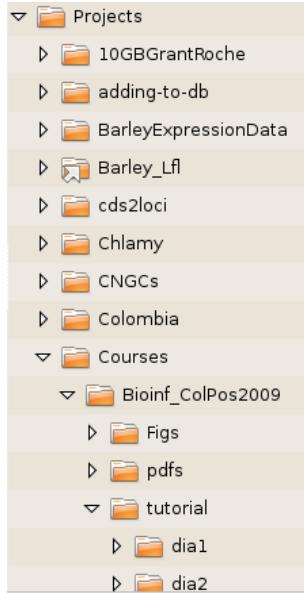


Figura 2.3: Árvore de diretórios no Linux

diretórios. Uma das opções mais utilizadas é ‘-l’ (“menos ele”; Linha29), cuja saída é exibida nas linhas 30 ate 32, onde a lista de diretórios no local atual é exibida, juntamente com permissões nesses diretórios, o número de subdiretórios, tamanho, data da última modificação e nome.

Navegando pela árvore de diretórios

```

25 [user@server:~]$ pwd
26 /home/user
27 [user@server:~]$ ls
28 dial dia2
29 [user@server:~]$ ls -l
30 total 0
31 drwxr-xr-x 2 user group 68 Aug  5 09:01 dial/
32 drwxr-xr-x 2 user group 68 Aug  5 09:02 dia2/
33 [user@server:~/dial]$ cd dial
34 [user@server:~]$ cd ..
35 [user@server:~]$ cd /home/user/dia2/

```

Como mencionado acima, os sistemas Unix são multi-usuários, o que implica que deve haver um sistema de permissão no sistema de arquivos, para evitar perdas accidentais de dados, e.g., que um usuário delete dados de outro. Na linha 43 as permissões de diretório são exibidas dia2 na primeira corda antes do primeiro espaço. O primeiro caractere indica se estamos em n diretório (d), um arquivo (-), ou um link (l). os 9 caracteres a seguir são divididos em 3 grupos de 3 caracteres cada, como mostrado na figura 2.4⁸.

Já sabemos como exibir informações sobre diretórios e arquivos na localização atual. Para alterar o diretório usamos o comando ‘cd nome_diretorio’, como se mostra na linha 33. se

⁸Exercício opcional: Como alterar as permissões de um arquivo ou diretório?

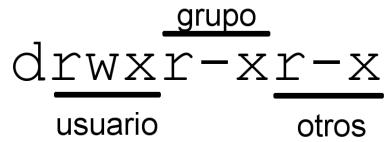


Figura 2.4: Sistema de permissão no Linux. r: permissão de leitura; w: permissão de escrita; x: permissão de Execução.

você quiser subir um nível na hierarquia do diretório executar o comando `cd ..`, outra opção é usar o caminho absoluto do diretório que você quer alcançar, como mostrado na linha 35. Retornar ao subdiretório `/home/usuario/dia1`.

Antes de continuar, eu gostaria de introduzir o comando mais importante de qualquer sistema Unix, é o comando “manual”, que mostra informações sobre o uso dos diferentes comandos, por favor, use-os sempre que você tiver alguma dúvida sobre as opções ou sintaxe de qualquer comando, e.g., `man ls`.

2.1.3 Organizando arquivos

As operações mais comuns com arquivos são: copiar, mover e excluir. A sintaxe dos comandos para mover ou copiar é a mesma: “comando fonte destino”. Por exemplo, suponha que você tem um arquivo chamado `“test1.txt”` em seu diretório `home` e você quer movê-lo para o diretório `“~/dia1/”`, você teria que executar o comando mostrado na linha 47. Você pode criar e remover diretórios (vazios) usando os comandos `mkdir` y `rmdir`, respectivamente.

Organizando arquivos e diretórios

```

36 [user@server:~]$ cd
37 [user@server:~]$ ls -l
38 total 0
39 drwxr-xr-x 2 user group 68 Aug  5 09:01 dia1/
40 drwxr-xr-x 2 user group 68 Aug  5 09:02 dia2/
41 [user@server:~]$ touch test1.txt
42 [user@server:~]$ ls -l
43 total 0
44 drwxr-xr-x 2 user group 68 Aug  5 09:01 dia1/
45 drwxr-xr-x 2 user group 68 Aug  5 09:02 dia2/
46 -rw-r--r-- 1 user group  0 Aug 18 20:42 test1.txt
47 [user@server:~]$ mv test1.txt dia1/
48 [user@server:~]$ ls -l dia1/
49 total 0
50 -rw-r--r-- 1 user group  0 Aug 18 20:42 test1.txt
51 [user@server:~]$ ls -l
52 drwxr-xr-x 2 user group 68 Aug  5 09:01 dia1/
53 drwxr-xr-x 2 user group 68 Aug  5 09:02 dia2/
54 [user@server:~]$
```

2.1.4 Algumas operações básicas com arquivos

Usando alguns comandos UNIX podemos obter informações sobre arquivos, e as informações que eles contêm, de forma rápida e eficiente, muitas vezes não é necessário abrir o arquivo, que pode ter vários megabytes, para obter essas informações.

No subdiretório “~/dial/”, encontra o arquivo “TAIR9_pep_20090619”, que corresponde ao banco de dados de sequências proteicas previstas no genoma da planta modelo *Arabidopsis thaliana*. para saber quantas linhas este arquivo tem execute o comando mostrado na linha 60.

Porque as diferenças nas saídas dos comandos executados nas linhas 60 e 62⁹?

Como mostrado na linha 58, o tamanho deste banco de dados é de 18.173.159 bytes. Para saber o quanto isso corresponde em uma unidade mais amigável use o comando mostrado na linha 64.

Na maioria dos casos é importante ver como o arquivo é, seja no seu início ou no final, mas devido ao grande tamanho dos arquivos com os quais você normalmente trabalha, não é conveniente abrir o arquivo com qualquer editor de texto, pois isso poderia reduzir o tempo de resposta do computador. Os comandos exibidos nas linhas 68 y 79, mostram a primeira e últimas 10 linhas no arquivo, respectivamente.

Usando o comando grep, como mostrado na linha 90, você pode obter uma lista das linhas no arquivo de interesse que contêm um determinado padrão, i.e., uma sequência de texto específica.

Operações básicas com arquivos

```
55 [user@server:~]$ cd dial/
56 [user@server:~/dial]$ ls -l
57 total 35496
58 -rw-r--r-- 1 user group 18173159 Aug 30 16:14 TAIR9_pep_20090619
59 -rw-r--r-- 1 user group 0 Aug 18 20:42 test1.txt
60 [user@server:~/dial]$ wc TAIR9_pep_20090619
61 274243 790613 18173159 TAIR9_pep_20090619
62 [user@server:~]$ wc -l TAIR9_pep_20090619
63 274243 TAIR9_pep_20090619
64 [user@server:~/dial]$ ls -lh
65 total 35496
66 -rw-r--r-- 1 user group 17M Aug 30 16:14 TAIR9_pep_20090619
67 -rw-r--r-- 1 user group 0B Aug 18 20:42 test1.txt
68 [user@server:~/dial]$ head TAIR9_pep_20090619
69 >AT1G51370.2 | Symbols: | F-box family protein
70 MVGGKKTKICDKVSHEEDRISQLPEPLISEILFHLSTKDSVRTSALSTKWRYLWQSVPG
71 LDLDPYASSNTNTIVSFVESFFDSHRDSWIRKLRLDLGYHHDKYDLMWSIDAATTRRIQH
72 LDVHCFHDNKNIPLSIYTCTTLVHLRLRWAVLTNPEFVSLPCLKIMHFENVSYPPNETLQK
73 LISGSPVLEELLIFSTMYPKGNVLQLRSDTLKRLDINEFIDVVVIYAPLLQCLRAKMYSTK
74 NFQIISSGFPAKLDIDFVNTPNGRYQKKKVIDEILIDISRVRDLVISSNTWKEFFLYSKSR
75 PLLQFRYISHLNARFYISDLEMPLTLESCPCKLESILVMSSFNPS*
76 >AT1G50920.1 | Symbols: | GTP-binding protein-related
77 MVQYNFKRITVVPNGKEFVDIILSRTQRQTPTVVKGYKINRLRQFYMRKVKYTQTNFHA
78 KLSAIIDEFPRLQIHPFYGDLLHVLYNKDHYKLALGQVNTARNLISKISKDYVKKLYG
79 [user@server:~/dial]$ tail TAIR9_pep_20090619
80 LLRYLTI*
```

⁹Revise a página do manual: man wc

```

81 >ATMG00070.1 | Symbols: NAD9 | NADH dehydrogenase subunit 9
82 MDNQFIFKYSWETLPKKWKKMERSEHGNRSNTDYLFLQLLCFLKLHTYTRVQVSIDIC
83 GVDHPSRKRRFEVVYNLLSTRYNSRIRVQTSADEVTRISPVVSLFFSAGRWEREWDMFG
84 VSFINHPDLRRISTDYGFEGHPLRKDLPLSGYVQVRYDDPEKRVVSEPIEMTQEFRYFDF
85 ASPWEQRSDG*
86 >ATMG00130.1 | Symbols: ORF121A | hypothetical protein
87 MASKIRKVTNQNMRRINSSLSKSSTFSTRLRITDSYLSSPSVTELAPLTLTGDDFTVTLS
88 VTPTMNSLESQVICPRAYDCKERIPPNQHIVSLELTYPHASIEPTATGSPETRDPPSAY
89 A*
90 [user@server:~/dial]$ grep ">" TAIR9_pep_20090619 | head -n 4
91 >AT1G51370.2 | Symbols: | F-box family protein
92 >AT1G50920.1 | Symbols: | GTP-binding protein-related
93 >AT1G36960.1 | Symbols: | unknown protein
94 >AT1G44020.1 | Symbols: | DC1 domain-containing protein

```

Nem sempre na bioinformática lidamos com sequências, em muitos casos temos dados em forma tabular, onde os campos são separados por algum caractere definido, por exemplo, Guias ou vírgulas. Na maioria dos casos, isso envolve armazenar e gerenciar os dados usando um sistema de banco de dados, como o MySQL. No entanto, é importante ter uma ideia dos resultados antes de integrá-los ao sistema de banco de dados, uma opção que apareceu recentemente, voltada para biólogos que trabalham com grandes quantidades de dados, é o Scriptome¹⁰, em que o autor oferece uma coleção de scripts PERL que podem ser executados na linha de comando. Nas linhas 95 ate 97 pode ver um exemplo onde todos os caracteres são alterados para maiúscula, o comando tem que ser executado em uma única linha, aqui é mostrado em linhas separadas apenas para facilitar sua visualização.

Exemplo do Scriptome

```

95 [user@server:~/dial]$ perl -e 'while(<>) {print lc($_);}' \
96 warn "Changed $. lines to lower case\n" \
97 TAIR9_pep_20090619 > TAIR9_pep_20090619.lc
98 changed 274243 lines to lower case
99 [user@server:~/dial]$ ls -l
100 total 70992
101 -rw-r--r-- 1 user group 18173159 Aug 30 16:14 TAIR9_pep_20090619
102 -rw-r--r-- 1 user group 18173159 Aug 30 19:54 TAIR9_pep_20090619.lc
103 -rw-r--r-- 1 user group 0 Aug 18 20:42 test1.txt
104 [user@server:~/dial]$ head -n 2 TAIR9_pep_20090619.lc
105 >at1g51370.2 | symbols: | f-box family protein
106 mvggkkktkicdkvshedrisqlpepliseilfhlstksalstkwrylwqsvpg
107 [user@server:~/dial]$

```

Na linha 90 foi usado o símbolo “|” ou “barra vertical”, ou “pipe” no UNIX, permite conectar comandos, de modo que a saída da esquerda da barra vertical sirva de entrada para o comando à direita da barra. Na linha 97 foi usado o símbolo “>” para redirecionar a saída padrão do comando para um arquivo.

¹⁰<http://sysbio.harvard.edu/csb/resources/computational/scriptome/UNIX/>

2.2 Formatos de sequência

Existem diferentes formatos para sequências, geralmente em texto simples. O que significa que eles podem ser vistos e editados com qualquer editor de texto, como vi o pico. alguns desses formatos são mais comuns do que outros e muitos programas de bioinformática aceitam vários dos formatos mais comuns. (LEONARD *et al.*, 2007).

Todos os formatos de sequência têm uma característica (campo) em comum: um identificador para cada sequência. Para que possa ser reconhecido inequivocamente.

2.2.1 Fasta

O formato mais simples é conhecido como Fasta¹¹. Em que uma entrada, sequência, pode ser dividida em duas partes: A linha de identificação, que deve começar com símbolo “>” e imediatamente seguido pelo identificador de sequência (Ver linha 108), qpode ser qualquer sequência de caracteres sem espaços. As linhas imediatamente após o identificador correspondem à sequência em si (Líneas 109-115).

Fasta é o formato de sequência mais usado em aplicações na bioinformática.

108 >gi|110742030|dbj|BAE98952.1| putative NAC domain protein [Arabidopsis thaliana]
109 MEDQVGFGRPNDEELVGHYLRNKIEGNTSRDVEVAISEVNICSYDPWNLRFQSKYKSRDAMWYFFSRRE
110 NNKGNRQSRTTVSGKWKTGESVEVKDQWGFCSEGFRKGKIGHKRVLAFLDGRYPDKTSWDVIHEFHDL
111 LPEHQRTYVICRLEYKGDDADILSAYAIDPTPAFVPNMTSSAGSVNVNQSRQRNSGSYNTSEYDSANHGQ
112 QFNENSNIMQQQPLQGSFNPLLEYDFANHGGQWLSDYIDLQQQVPYLAPYENESEMIWKHVIEENFEFLV
113 DERTSMQQHYSDRPKKPVSGLPDDSSDETGSMIFEDTSSSTDVGSSDEPGHTRIDDIPSLSNIIIEPL
114 HNYKAQEQPQKQSKEKVISSQKSECEWKMAEDSIKIPPSTNTVKQSWIVLENAQWNYLKNMIIGVLLFIS
115 VISWIILVG

2.2.2 GenBank

O formato GenBank¹²¹³ é usado pelo “National Center for Biotechnology Information” (NCBI¹⁴), o maior repositório de sequências, tanto ácidos nucleicos quanto proteínas, em todo o mundo. O NCBI juntamente com o EMBL¹⁵ e o DDBJ¹⁶, manter em conjunto “The International Nucleotide Sequence Database” (MIZRACHI, 2008).

Uma entrada neste formato é composta por duas partes. A primeira parte consiste em posições de 1 a 10, e geralmente contém o nome do campo, e.g., LOCUS, DEFINITION, ACCESSION ou SOURCE. A segunda parte de cada entrada contém as informações para o campo correspondente.

¹¹<http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>

¹²<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

¹³<ftp://ftp.ncbi.nih.gov/genbank/release.notes/gb172.release.notes>

¹⁴<http://www.ncbi.nlm.nih.gov/>

¹⁵<http://www.ebi.ac.uk/emb1/>

¹⁶<http://www.ddbj.nig.ac.jp/>

Cada entrada termina com o símbolo “\\” (Linha 178). Você pode encontrar mais informações sobre este tipo de arquivo seguindo o link <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

Sequência em formato GenBank

```

116 LOCUS      BAE98952          429 aa      linear    PLN 27-JUL-2006
117 DEFINITION putative NAC domain protein [Arabidopsis thaliana].
118 ACCESSION   BAE98952
119 VERSION    BAE98952.1  GI:110742030
120 DBSOURCE   accession AK226863.1
121 KEYWORDS   .
122 SOURCE     Arabidopsis thaliana (thale cress)
123 ORGANISM   Arabidopsis thaliana
124 Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
125 Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
126 rosids; eurosids II; Brassicales; Brassicaceae; Arabidopsis.
127 REFERENCE  1
128 AUTHORS    Totoki,Y., Seki,M., Ishida,J., Nakajima,M., Enju,A., Morosawa,T.,
129 Kamiya,A., Narusaka,M., Shin-i,T., Nakagawa,M., Sakamoto,N.,
130 Oishi,K., Kohara,Y., Kobayashi,M., Toyoda,A., Sakaki,Y.,
131 Sakurai,T., Iida,K., Akiyama,K., Satou,M., Toyoda,T., Konagaya,A.,
132 Carninci,P., Kawai,J., Hayashizaki,Y. and Shinozaki,K.
133 TITLE      Large-scale analysis of RIKEN Arabidopsis full-length (RAFL) cDNAs
134 JOURNAL   Unpublished
135 REFERENCE 2 (residues 1 to 429)
136 AUTHORS    Totoki,Y., Seki,M., Ishida,J., Nakajima,M., Enju,A., Morosawa,T.,
137 Kamiya,A., Narusaka,M., Shin-i,T., Nakagawa,M., Sakamoto,N.,
138 Oishi,K., Kohara,Y., Kobayashi,M., Toyoda,A., Sakaki,Y.,
139 Sakurai,T., Iida,K., Akiyama,K., Satou,M., Toyoda,T., Konagaya,A.,
140 Carninci,P., Kawai,J., Hayashizaki,Y. and Shinozaki,K.
141 TITLE      Direct Submission
142 JOURNAL   Submitted (26-JUL-2006) Motoaki Seki, RIKEN Plant Science Center;
143 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
144 (E-mail:mseki@psc.riken.jp, URL:http://rarge.gsc.riken.jp/,
145 Tel:81-45-503-9625, Fax:81-45-503-9586)
146 COMMENT   An Arabidopsis full-length cDNA library was constructed essentially
147 as reported previously (Seki et al. (1998) Plant J. 15:707-720;
148 Seki et al. (2002) Science 296:141-145).
149 This clone is in a modified pBluescript vector.
150 Please visit our web site (http://rarge.gsc.riken.jp/) for further
151 details.
152 FEATURES    Location/Qualifiers
153 source      1..429
154           /organism="Arabidopsis thaliana"
155           /db_xref="taxon:3702"
156           /chromosome="1"
157           /clone="RAFL08-19-M04"
158           /ecotype="Columbia"
159           /note="common name: thale cress"
160 Protein     1..429
161           /product="putative NAC domain protein"
162 Region      5..137
163           /region_name="NAM"
164           /note="No apical meristem (NAM) protein; pfam02365"
165           /db_xref="CDD:111274"
166 CDS         1..429
167           /gene="At1g01010"
168           /coded_by="AK226863.1:89..1378"
169 ORIGIN
170 1 medqvgfgfr pnndeelvghy lrnkiegnts rdvevaisev nicsydpwnl rfqskyksrd
171 61 amwyffsrre nnkgnrqsr tvsgkwkltg esvevkdwg fcsegfrki ghkrvlafld
172 121 grypdktsd wvihefhydl lpehqrtiyvi crleykgdda dilsayaipd tpafvpnmts
173 181 sagsvvnqsr qrnsqsynty seydsanhqq qfnensnimq qqplqgsfnp lleydfanhg
174 241 ggwlqsyidl qqgpvplapy enesemiwkh vileeneflv dertsmqhy sdhrpkpvs
175 301 gvlpdssdt etgsmifedt ssstdsvgss depghtridd ipslniepl hnykaqepl
176 361 qgskekviss qksecewma edsikippst ntvkqswivl enaqwnyln miigyllfis
177 421 viswiilvg
178 //

```

2.2.3 Algumas operações básicas com sequências no formato Fasta

Para o restante desta seção, e para a próxima, usaremos apenas sequências no formato Fasta. Por favor, verifique se as sequências de *A. thaliana* no arquivo TAIR9_pep_20090619 estão neste formato. Você pode usar o comando “head nome_arquivo”, ou o comando “less nome_arquivo”¹⁷.

Você já teve que contar o número de sequências ou alterar o identificador de sequência no formato Fasta? Se for uma dúzia de sequências, isso poderia facilmente ser feito em qualquer editor de texto, mas quando há milhares de sequências a opção do editor de texto deixa de ser viável. Felizmente, alguns comandos Unix nos permitem executar essas tarefas simples rapidamente.

Como viu na linha 90, o comando “grep” poderia nos ajudar a contar o número de sequências em um arquivo Fasta. O interruptor “-c” conta o número de linhas contendo um determinado padrão em um arquivo, e podemos tirar proveito do fato de que em um arquivo Fasta o símbolo “>” aparece apenas uma vez para cada sequência como mostrado na linha 185.

Usando comandos Unix com arquivos Fasta

```
179 [user@server:~]$ cd ~/dial/
180 [user@server:~/dial]$ ls -l
181 total 70992
182 -rw-r--r-- 1 user  group 18173159 Aug 30 16:14 TAIR9_pep_20090619
183 -rw-r--r-- 1 user  group 18173159 Aug 30 19:54 TAIR9_pep_20090619.lc
184 -rw-r--r-- 1 user  group          0 Aug 18 20:42 test1.txt
185 [user@server:~/dial]$ grep -c ">" TAIR9_pep_20090619
186 33410
187 [user@server:~/dial]$ sed 's/>/>ATH_/' TAIR9_pep_20090619 > TAIR9_pep_20090619.mod
188 [user@server:~/dial]$ head TAIR9_pep_20090619.mod
189 >ATH_AT1G51370.2 | Symbols:
190 MVGGKKKTKICDKVSHEEDRISQLPEPLISEIILFHLSTKDSVRTSALSTKWRYLWQSVPG
191 LDLDPYASSNTNTIVSFVESFFDSHRDWIRKLRLDLGYHHDKYDLMWSIDAATTRRIQH
192 [user@server:~/dial]$
```

Em outras ocasiões é importante modificar o identificador de cada sequência, de modo que inclua, por exemplo, uma abreviação que represente o nome da espécie a que a sequência pertence. Novamente Unix nos permite fazer essa mudança muito rapidamente usando o comando sed como mostrado na linha 187.

¹⁷Para sair de less pressione “q”

Capítulo 3

Buscas em banco de dados biológicos

Este capítulo corresponde a uma versão modificada de um guia original da professora Silvia Restrepo

3.1 NCBI – Bancos de dados e busca de informações

O National Center for Biotechnology Information, NCBI pela sigla em inglês, é uma instituição pública dos Estados Unidos da América, que guarda todas as informações sobre os genomas de várias espécies, bem como o maior banco de dados público sobre sequências de DNA e proteínas. Sua página principal da rede está localizada no seguinte link: <http://www.ncbi.nlm.nih.gov/>

Este site conecta todos os dados disponíveis em seus servidores (PubMed, TODOS os bancos de dados (Entrez), Blast, OMIM, Books, TaxBrowser, Structure), conforme mostrado na Figura 3.1. Embora o Entrez esteja listado como um dos serviços, na realidade quase todos eles dependem diretamente do Entrez. Por exemplo, PubMed e Taxonomy estão intimamente ligados ao Entrez.

3.1.1 Vamos começar uma visita aos seus bancos de dados

Como primeiro passo, vamos entrar no PubMed. Esta base de dados contém informações sobre publicações científicas, e seus registros foram compilados pela NLM (National Library of Medicine), com a colaboração dos editores. Lá você encontrará a maioria das referências necessárias, incluindo o resumo (Abstract) e em alguns casos a publicação gratuita.

Para obter ajuda sobre como realizar buscas consulte o seguinte link: <http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed>

As páginas possuem um menu de banco de dados em uma barra superior, as pesquisas devem ser colocadas na janela mostrada na Figura 3.2.

Uma busca deve ter um formato semelhante a este:

“palavrachave”[field] operador lógico “palavrachave”[field] ...

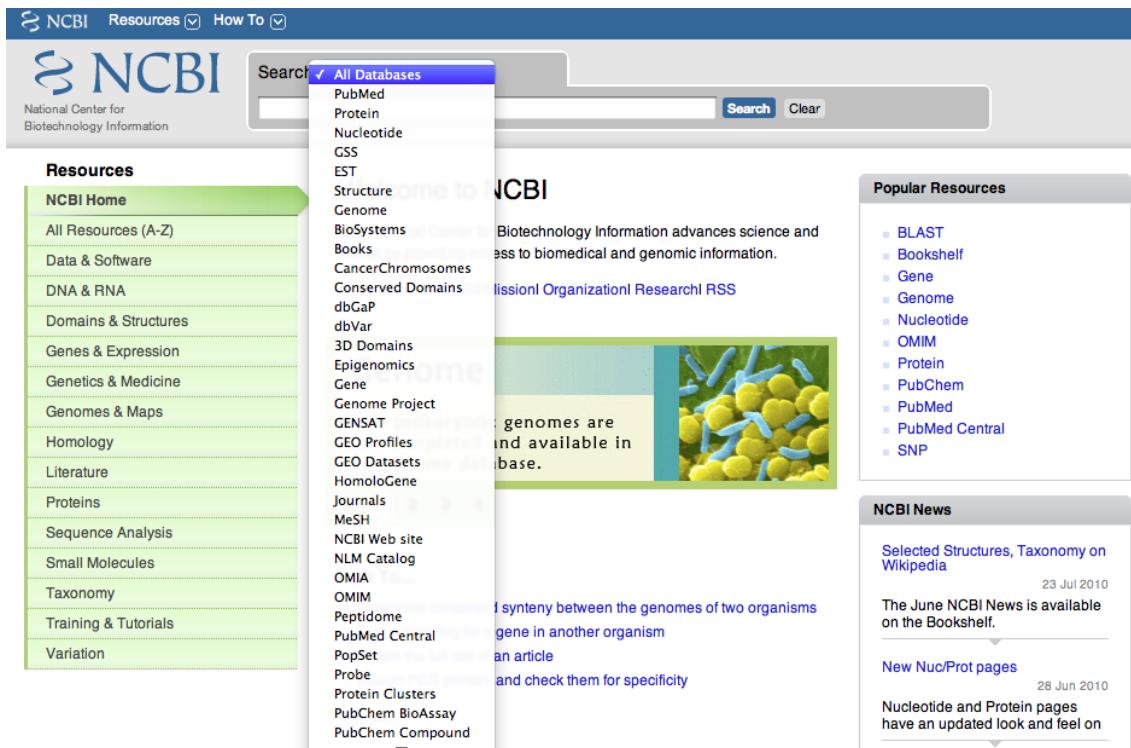


Figura 3.1: Página inicial do NCBI

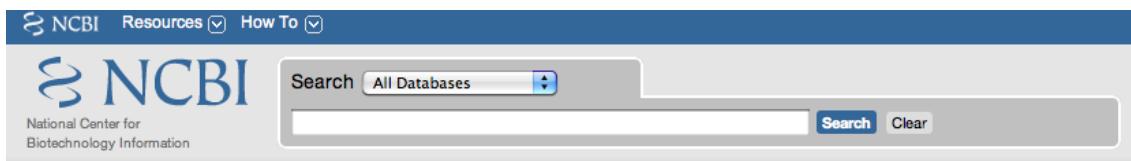


Figura 3.2: Janela de busca do NCBI

Onde **palavra chave** é a palavra utilizada para identificar um registro (record) de acordo com o campo (field) utilizado. Por exemplo, uma palavra chave pode ser "Silva" no campo "authors". **Operador lógico** é qualquer um destes operadores booleanos: AND, OR, NOT, BUT, etc. Ao substituir por suas próprias palavras-chave no formato acima, lembre-se de que os campos devem estar entre colchetes [], mas os operadores são independentes (sem os símbolos,), além disso, as aspas na palavra chave são opcionais, mas cumprem a função de forçar uma busca com a palavra exata ao invés de serem flexíveis.

Por exemplo, se eu quiser pesquisar todos os artigos de 1999 publicados por Silva et al na revista Science, eu uso o seguinte comando: "Silva"[AU] AND 1999[DP] AND "science"[TA]. Quanto mais informações forem inseridas na busca, mais restrita será a resposta (por exemplo, se eu incluir mais autores).

Os campos mais comuns que podem ser solicitados no PubMed são os seguintes:

All Fields [ALL] Inclui todos os campos pesquisáveis do PubMed. No entanto, apenas os termos em que não houver correspondência encontrada em uma das tabelas ou índices de tradução por meio do processo de Mapeamento Automático de Termos serão pesquisados em Todos os Campos. PubMed ignora palavras irrelevantes de consultas de pesquisa.

Author Name [AU] Vários limites no número de nomes de autores incluídos na citação MEDLINE existiram ao longo dos anos (consulte a política NLM sobre nomes de autores). MEDLINE não lista o nome completo. O formato para pesquisar o nome do autor é: sobrenome seguido de espaço e até as duas primeiras iniciais seguidas de espaço e abreviação do sufixo, se for o caso, tudo sem pontos ou vírgula após o sobrenome (por exemplo, fauci as ou o'brien jc jr). Iniciais e sufixos podem ser omitidos durante a pesquisa. O PubMed trunca automaticamente o nome de um autor para levar em conta as iniciais variadas, por exemplo, o'brien j [au] recuperará o'brien ja, o'brien jb, o'brien jc jr, bem como o'brien j. Para desativar esse truncamento automático, coloque o nome do autor entre aspas duplas e qualifique com [au] entre colchetes, por exemplo, "o'brien j"[au] para recuperar apenas o'brien j.

EC/RN Number [RN] Número atribuído pela Enzyme Commission para designar uma enzima específica ou pelo Chemical Abstracts Service (CAS) para números de registro.

Entrez Date [EDAT] Data em que a citação foi adicionada ao banco de dados PubMed. As citações são exibidas na ordem Entrez Date, que é o último a entrar, o primeiro a sair. As datas ou intervalos de datas devem ser inseridos usando o formato AAAA/MM/DD [edat], ex. 1998/04/06 [ed.] . O mês e o dia são opcionais (por exemplo, 1998 [edat] ou 1998/03 [edat]). Para inserir um intervalo de datas, insira dois pontos (:) entre cada data (por exemplo, 1996:1997 [edat] ou 1998/01:1998/04 [edat])

Issue [IP] O número do número da revista em que o artigo é publicado.

Journal Title [TA] A abreviatura do título do periódico, nome completo do periódico ou número ISSN .

Language [LA]

Publication Date [DP] A data em que o artigo foi publicado. As datas ou intervalos de datas devem ser pesquisados usando o formato AAAA/MM/DD [dp], ex. 1998/03/06 [dp] . O mês e o dia são opcionais (por exemplo, 1998 [dp] ou 1998/03 [dp]). Para inserir um intervalo de datas, insira dois pontos (:) entre cada data (por exemplo, 1996:1998 [dp] ou 1998/01:1998/04 [dp]). O nome de um produto químico discutido no artigo. Sinônimos para o Nome da Substância do Conceito Complementar serão mapeados automaticamente quando qualificados com [nm]. Este campo foi implementado em meados de 1980. Muitos nomes químicos são pesquisáveis como termos MeSH antes dessa data.

Text Words [TW] Inclui todas as palavras e números no título e resumo, e termos MeSH, subtítulos, nomes de substâncias químicas, nome pessoal como assunto e campo MEDLINE Secondary Source (SI). O campo Nome pessoal do assunto também pode ser pesquisado diretamente usando a tag do campo de pesquisa [ps], por exemplo, rouxinol f [ps].

Title Words [TI] Palavras e números incluídos no título de uma citação.

Title/Abstract Words [TIAB] Palavras e números incluídos no título e resumo de uma citação.

Unique Identifiers [UID]

Volume [VI] O número do volume da revista em que um artigo é publicado.

Agora vamos no site onde o ENTREZ está localizado. Para fazer isso, selecione TODOS OS BANCOS DE DADOS na janela do banco de dados na página principal. Entrez é um sistema de busca de sequências armazenadas em bancos de dados. Consultas sofisticadas podem ser solicitadas para obter um conjunto de sequências de seu interesse, por exemplo, posso pedir para exibir todas as sequências genômicas de Arabidopsis que foram incluídas no banco de dados entre os anos 97' e 99' que também contêm anotação (em a tabela "features") nas regiões promotoras. A Figura 3.3 mostra a página de login do servidor Entrez.

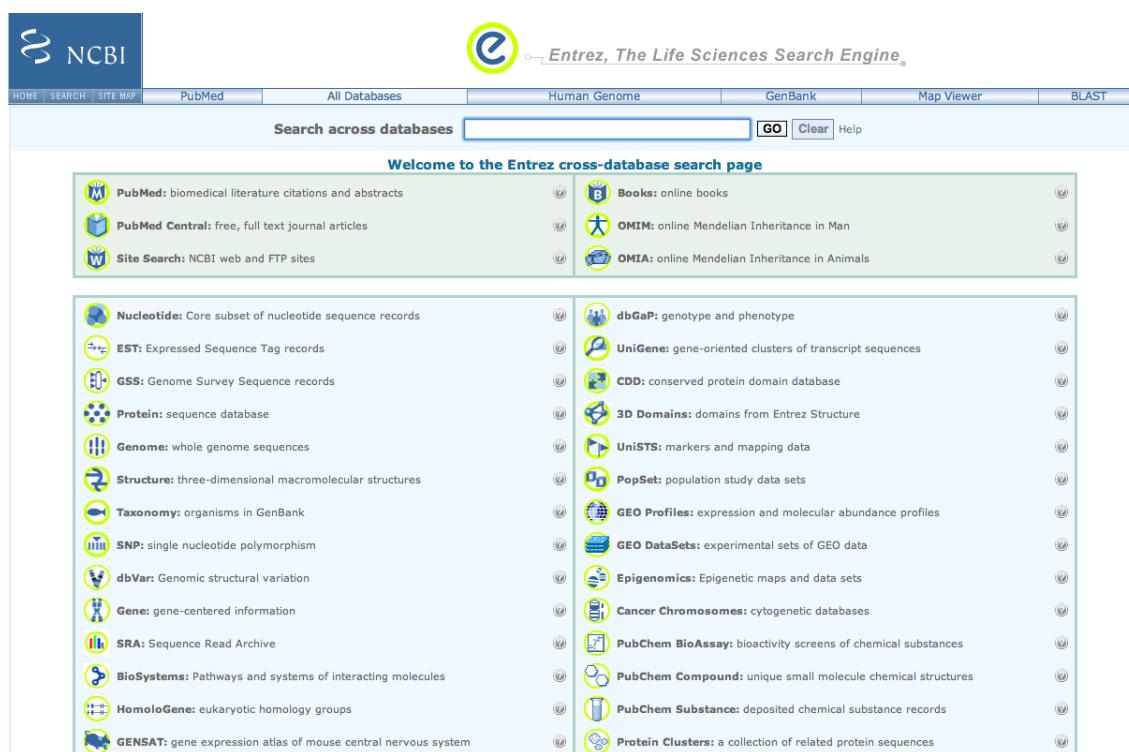


Figura 3.3: Página inicial do Entrez

Assim, em um único site podemos realizar buscas simultaneamente em todas as bases de dados ou selecionar uma única base de dados e realizar uma busca por base de dados.

Na caixa de busca, as sequências podem ser consultadas usando seus números identificadores (como o gi-number ou com o número de acesso). Questões mais complicadas também podem ser formuladas usando a sintaxe entrez, semelhante a como vimos o PubMed:

“palavrachave”[field] operador lógico “palavrachave”[field] ...

Para mais informações sobre o Entrez você pode seguir o link: <http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpentrez&part=EntreHelp>

Exercícios

1. Qual é a classificação taxonômica da alga *Chlamydomonas reinhardtii*?, e quais outras plantas estão próximas, para que possam ser usadas como fonte de marcadores?. Quantas sequências de proteínas estão presentes no GenBank para a espécie *Chlamydomonas reinhardtii*?
2. Acesse a página do PubMed e obtenha as referências que tratam da biologia molecular e/ou genética da mandioca (*Manihot esculenta*). Quantos foram publicados nos últimos dois anos e de quais laboratórios (ou regiões geográficas) são os autores? Explique como você pesquisou. Dica: GoPubMed <http://www.gopubmed.org/>
3. Use o Entrez para encontrar todas as sequências EST (Expressed Sequence Tag) de arroz que foram depositadas no banco de dados.

Revise a descrição dos principais formatos de sequências na seção 2.2.

Quais bancos de dados encontramos no NCBI?

O NCBI possui um grande número de bancos de dados. O mais conhecido no GenBank que contém todas as sequências de nucleotídeos. GenPept contém as sequências de proteínas. Outras bases de dados são Genome, Structure, PubMed

No GenBank as sequências estão organizadas em 17 divisões, 11 tradicionais e 6 Bulk Nas tradicionais, as sequências foram enviadas diretamente pelos pesquisadores, são caracterizadas e as divisões são:

PRI primatas

PLN plantas

BCT bactérias

INV invertebrados

ROD Roedores

VRL Viral

VRT outros vertebrados

MAM Mamíferos (Ej. ROD + PRI)

PHG Fagos

SYN Sintético (vetores de clonagem, etc)

UNA sem anotação

O Bulk consiste em sequências enviadas em grupos via email ou ftp, imprecisas e mal caracterizadas, são elas:

dbEST Banco de dados EST, tags de sequência expressa

dbSTS Sequence-tagged sites: são marcos genômicos curtos para os quais há informações de sequência e mapa.

dbGSS Genomic survey sequences. Contém: dados de sequência do genoma de etapa única, sequências terminais BAC, YAC e cosmídeos, sequências de éxon

dbHTGS High-Throughput Genomic Sequences. Ele foi criado para salvar informações de sequenciamento de genoma que não foram finalizadas ou curadas, mas para torná-las conhecidas da comunidade científica assim que estiverem disponíveis.

e também existem bancos de dados para:

HTC High Throughput cDNA

PAT Patent

RefSeq

Queremos colocar ênfase especial em um banco de dados NCBI chamado RefSeq. Este banco de dados foi criado para obter uma coleção biologicamente não redundante de sequências de DNA, RNA e proteínas. Cada RefSeq (sequência de referência) representa uma molécula única que ocorre naturalmente em um organismo. Esta base de dados é do tipo com curadoria de pesquisadores. Cada molécula não é um resultado de pesquisa, mas sim uma síntese de informações.

Vamos voltar para a página principal do NCBI e na janela de busca, deixando all databases, digite NC_001139¹. Vemos que em Nucleotide temos 1 hit, assim como em Genome e em Gene temos 631.

Vamos abrir Nucleotide: obtemos um flatfile de sequência que corresponde à sequência completa do cromossomo VII da levedura. Vamos dar uma olhada no arquivo flatfile, **quais informações ele contém?**

Observemos que os identificadores desta base de dados mudam e são do tipo 2+6 com duas letras e 6 números, a tabela a seguir nos mostra o que significam essas letras:

mRNA and Proteins	
NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA
Gene records	
NG_123456	Genomic Region
Chromosome	
NC_123456	Complete genomic molecule, Microbial replicons, organelle genomes
Assemblies	
NT_123456	Contig
NW_123456	WGS supercontig (assembly of WGS)

3.1.2 Recuperação de Sequências no NCBI com buscas mais específicas

CONHECEMOS O ORGANISMO. As pesquisas do NCBI podem ser mais direcionadas se conhece o organismo sobre o qual estamos procurando informações. Entramos na página inicial do NCBI, vamos para TaxBrowser, colocamos o nome do organismo que estamos procurando. Ao selecioná-lo, uma tabela do número de sequências por tipo de molécula ou projeto aparece à direita. Clicar em uma delas, por exemplo proteínas, nos leva diretamente às proteínas daquele organismo.

NÓS SABEMOS OS NÚMEROS DE ACESSO. Se você souber o número de acesso diretamente, pode colocá-lo na janela de pesquisa da página principal do NCBI. Para várias sequências os números são colocados com a palavra OR entre eles, por exemplo AJ487842 ou AJ487843. Por fim, para uma sequência de números de acesso, digite: AJ487842::AJ487851[ACCN]

¹certifique-se de incluir o símbolo de sublinhado

DIRECIONAMOS A PESQUISA COM LIMITES. Por exemplo, se eu quiser pesquisar as sequências de mRNA curadas relacionadas a um tipo de câncer em humanos, posso fazer a seguinte pesquisa: na janela de pesquisa, coloco COLON CANCER AND NONPOLYPOSIS , eu pesquiso o banco de dados de nucleotídeos. Então em LIMITS selecione a molécula de mRNA e em only from (banco de dados) selecione RefSeq. Então eu selecionei a outra janela Preview/index acima e lá em organismos eu escrevo humanos e selecionei AND

3.2 Recuperação de sequência usando SRS@EBI

Existe no entanto uma excelente alternativa para a busca de sequências biológicas, que nos permite controlar quase todos os aspectos da nossa busca, esta alternativa é o Sequence Retrieval System (SRS). Este sistema foi desenvolvido com esta tarefa de recuperar eficazmente sequências biológicas em mente, daí o seu design e capacidades.

Neste workshop trabalharemos com o SRS oferecido pelo European Bioinformatics Institute (EBI), <http://srs.ebi.ac.uk/>. Ou digitando EBI, (<http://www.ebi.ac.uk/>) , database → database browsing você chega ao SRS.

Uma forma simples de consultar o SRS é através da caixa Quick Text Search. Nesta caixa é possível pesquisar em vários bancos de dados disponíveis no menu suspenso, conforme mostrado na Figura ??

Por exemplo, selecionando a opção “Nucleotide Sequences”, realizaremos nossa busca no banco de dados EMBL DNA (homólogo ao genBank e DDBJ).

Faça uma pesquisa rápida pelo HIV-1 com diferentes opções no menu suspenso. Até este ponto, o SRS parece ser um pouco menos completo em comparação com o site do NCBI, mas agora começaremos a ver onde está todo o seu potencial.

Agora vamos realizar uma busca avançada. Selecione a guia Library Page localizada na parte superior da tela e mostrada na Figura ??

Você será então levado para a seção SRS onde estão descritos cada um dos bancos de dados que compõem o sistema (Figura 3.6). Como você pode ver, o SRS inclui muitos bancos de dados ao mesmo tempo e essa é uma de suas principais virtudes, por isso o SRS às vezes é conhecido como “banco de dados de bancos de dados”, pois através deste sistema podemos consultar vários bancos de dados ao mesmo tempo, de acordo com nossas necessidades particulares.

Como você pode ver, o SRS é semelhante ao sistema NCBI ENTREZ, no sentido de que nos permite consultar muitas bases de dados ao mesmo tempo, mas desta vez não se restringindo apenas àquelas que o NCBI possui, mas a praticamente qualquer base de dados. A quantidade de bancos de dados que o SRS possui depende de cada implementação, ou seja, o administrador do SRS determina quais bancos de dados deseja ou não incluir em seu sistema

Posicione o cursor do mouse sobre qualquer uma das entradas, após alguns segundos aparecerá uma caixa de texto explicativa. Que tipo de informação os bancos de dados EMBL (Contig

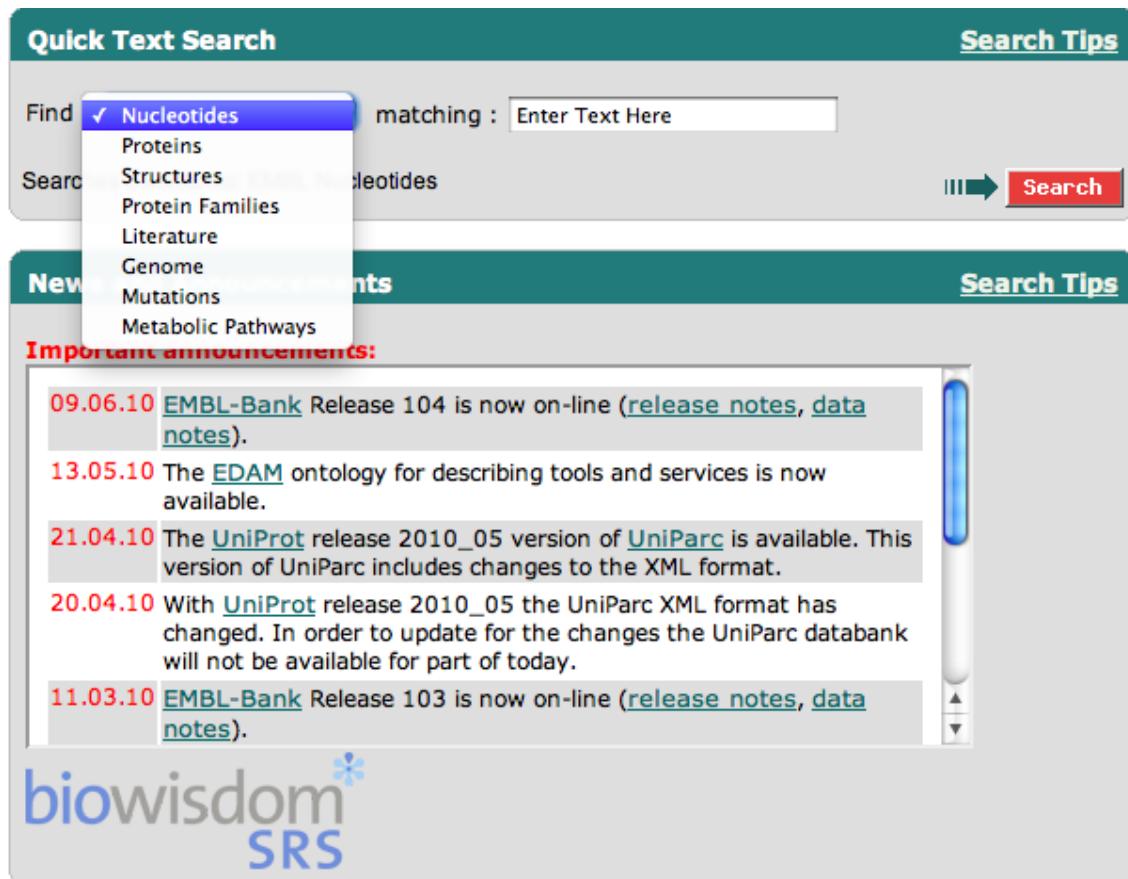


Figura 3.4: Página inicial do SRS

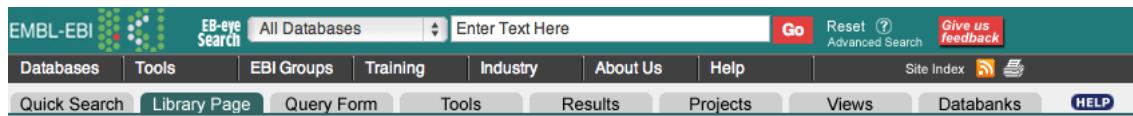


Figura 3.5: Opções SRS

Updates), UniprotKB/Swissprot fornecem?

Ao seguir o link para qualquer uma dessas bases de dados obteremos mais informações sobre ela, como o número de entradas presentes, data de atualização, etc. No entanto, por enquanto nosso interesse é selecionar algumas bases de dados para realizar nossas buscas. Marque as caixas para os bancos de dados “UniprotKB/Swissprot” e “UniprotKB/TrEMBL”. Certifique-se de que esses sejam os únicos bancos de dados selecionados.

À esquerda de sua tela você encontrará a caixa “Search Options” que nos permitirá selecionar o nível de profundidade de nossa busca. Como esta é a primeira vez que trabalhamos com este sistema, selecionaremos o formulário padrão de busca.

Pressione o botão “Standard query Form” na caixa “Search Options”

Esta ação o levará ao formulário de busca SRS padrão (Figura 3.7).

The screenshot shows the 'Available Databases' section of the SRS interface. It includes a toolbar with 'Expand all' and 'Collapse all' buttons, and a 'Show databases tooltips' checkbox. The main area lists several categories of databases:

- Literature, Bibliography and Reference Databases**: MEDLINE, Taxonomy, OMIM, OMIM Morbid Map, Patent Abstracts, Karyn's Genomes, Patent Equivalents.
- Literature, Bibliography and Reference Databases - subsections**: MEDLINE (Updates), MEDLINE (Main Release 2010), MED2PUB.
- Gene Dictionaries and Ontologies**.
- Nucleotide sequence databases**: EMBL, Patent DNA, EMBL (Contig), EMBL (Contigs expanded), EMBL (Coding Sequences), EMBL ID/Accession Mapping, EMBL MGA, IMGT/LIGM-DB, IMGT/HLA, IPD-KIR, Genome Reviews, GR Genes, GR Transcripts, RefSeq Genome, LiveLists, Patent DNA NRL1.
- Nucleotide sequence databases - subsections**: EMBL (Updates), EMBL (Release), EMBL (Whole Genome Shotgun), EMBL (Contig updates), EMBL (Release, Deleted), RefSeq Genome (Release).
- Nucleotide related databases**.
- UniProt Universal Protein Resource**: UniProtKB, UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, UniRef100, UniRef90, UniRef50, UniParc.

A tooltip for 'EMBL (Whole Genome Shotgun)' is displayed, stating: 'LiveLists - maps EMBL Accession numbers to NCBI gi identifiers' and 'To obtain comprehensive information on this databank, click the link'.

Figura 3.6: Opções SRS

Fields you can search Campos de busca, onde podemos inserir nossos termos de busca de acordo com qualquer uma das opções presentes nos respectivos menus suspensos.

Create View Criar vista, esta opção funciona em conjunto com a opção 3, e aqui podemos definir o tipo de campos que queremos ver na nossa página de resultados. Para o nosso exemplo, estamos interessados em selecionar todas as proteínas de superfície conhecidas de *Plasmodium falciparum* com atividade imunogênica, relacionadas ao merozoíto.

Result Display Options Opções para exibir os resultados, onde podemos definir o número de resultados que queremos por página, bem como o formato de saída, seja um dos definidos no menu suspenso ou criando uma visualização personalizada (opção “create view”).

Search Options Opções de busca, onde podemos definir, entre outras coisas, o tipo de conector lógico (Booleano) a ser usado para os termos definidos em 1.

Defina estes critérios na seção “Fields you can search” de acordo com a Figura 3.8.

Em seguida, pressione o botão “search” localizado no topo desta seção e aguarde alguns segundos.

Com certeza você já tem uma visão mais exata das possibilidades oferecidas pelo SRS e suas principais diferenças com o sistema Entrez. Primeiro, conseguimos definir exatamente não apenas o banco de dados que queríamos consultar, mas as seções específicas dele. Além disso, também conseguimos definir exatamente os termos de pesquisa em seções específicas dos posts, o que nos dá total controle sobre os resultados que queremos obter.

Brinque com as diferentes opções de formato que o SRS oferece na seção “Result Display

Search Options <p>Combine search terms with: & (AND)</p> <p>Use wildcards <input checked="" type="checkbox"/></p> <p>Get results of type: <input type="button" value="Entry"/></p>	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="background-color: #0070C0; color: white;">Fields you can search</th> <th style="background-color: #0070C0; color: white;">Your search terms</th> </tr> </thead> <tbody> <tr> <td colspan="2">In a single field, you can separate multiple values by: &, or !</td> </tr> <tr> <td><input type="button" value="i"/> AllText</td> <td><input type="text" value=""/></td> </tr> <tr> <td><input type="button" value="i"/> AllText</td> <td><input type="text" value=""/></td> </tr> <tr> <td><input type="button" value="i"/> AllText</td> <td><input type="text" value=""/></td> </tr> <tr> <td><input type="button" value="i"/> AllText</td> <td><input type="text" value=""/></td> </tr> </tbody> </table> <p style="margin-top: 10px;">Create a view</p> <p>Select the fields you want displayed in your view and choose the format</p> <p>Choose 1 or more fields: <input type="radio"/> ID <input type="radio"/> Display As: <input checked="" type="radio"/> Table <input type="radio"/> List</p> <p><input type="checkbox"/> EntryName <input type="checkbox"/> Data Class <input type="checkbox"/> AccessionNumber <input type="checkbox"/> Primary Accession Number <input type="checkbox"/> Sequence Version <input type="checkbox"/> Creation Date</p> <p style="margin-top: 10px;">Tips</p> <p>To do more advanced queries, use the Extended Query Form.</p>	Fields you can search	Your search terms	In a single field, you can separate multiple values by: &, or !		<input type="button" value="i"/> AllText	<input type="text" value=""/>	<input type="button" value="i"/> AllText	<input type="text" value=""/>	<input type="button" value="i"/> AllText	<input type="text" value=""/>	<input type="button" value="i"/> AllText	<input type="text" value=""/>
Fields you can search	Your search terms												
In a single field, you can separate multiple values by: &, or !													
<input type="button" value="i"/> AllText	<input type="text" value=""/>												
<input type="button" value="i"/> AllText	<input type="text" value=""/>												
<input type="button" value="i"/> AllText	<input type="text" value=""/>												
<input type="button" value="i"/> AllText	<input type="text" value=""/>												

Figura 3.7: Formulário de pesquisa SRS

Fields you can search <p>In a single field, you can separate multiple values by: &, or !</p>	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="background-color: #0070C0; color: white;">Your search terms</th> </tr> </thead> <tbody> <tr> <td><input type="button" value="i"/> Organism Name <input type="text" value="plasmodium falciparum"/></td> </tr> <tr> <td><input type="button" value="i"/> Keywords <input type="text" value="merozoite"/></td> </tr> <tr> <td><input type="button" value="i"/> Description <input type="text" value="surface antigen"/></td> </tr> <tr> <td><input type="button" value="i"/> AllText <input type="text" value=""/></td> </tr> </tbody> </table>	Your search terms	<input type="button" value="i"/> Organism Name <input type="text" value="plasmodium falciparum"/>	<input type="button" value="i"/> Keywords <input type="text" value="merozoite"/>	<input type="button" value="i"/> Description <input type="text" value="surface antigen"/>	<input type="button" value="i"/> AllText <input type="text" value=""/>
Your search terms						
<input type="button" value="i"/> Organism Name <input type="text" value="plasmodium falciparum"/>						
<input type="button" value="i"/> Keywords <input type="text" value="merozoite"/>						
<input type="button" value="i"/> Description <input type="text" value="surface antigen"/>						
<input type="button" value="i"/> AllText <input type="text" value=""/>						

Figura 3.8: Critérios de pesquisa avançados

“Options” do formulário de pesquisa. Tente também criar seu próprio formato de saída com a opção “Create view”.

Encontre todas as proteínas nucleares hipotéticas de *Saccharomyces cerevisiae* e exiba a informação em formato fasta.

Capítulo 4

Manipulación básica de secuencias

Este capítulo corresponde a una versión modificada de una guía original de la profesora Silvia Restrepo.

4.1 Limpieza de secuencias

Un Vector, es un agente que lleva fragmentos de ADN de interés a una célula específica. Si éste es utilizado para reproducir un fragmento de ADN, se le conoce como *Vector de Clonación*, si se utiliza para expresar cierto gen, se conoce como *Vector de Expresión*. Los vectores más usados son plásmidos, BACs, YACs, cósmidos y los bacteriófagos Lambda y P1. En cualquier caso que se utilice un vector, cuando se manda a secuenciar el fragmento de interés, se puede identificar las secuencias vector y eliminarlas. Para esto se puede emplear VecScreen siguiendo el enlace <http://www.ncbi.nlm.nih.gov/VecScreen/> (Figura 4.1).

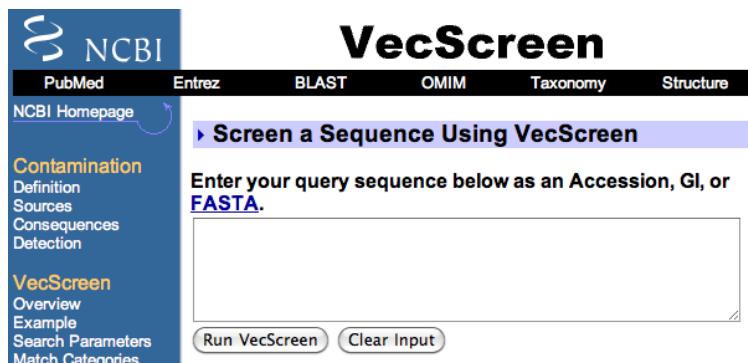


Figura 4.1: VecScreen: Herramienta para detectar contaminación de vectores.

En el campo de búsqueda que aparece en la página, pegue la Secuencia Problema 1 y ejecútela como “**Run VecScreen**”.

En la siguiente página, deje los campos que se encuentran por defecto (para ver los resultados

de manera gráfica) y dele click a “**View report**”.

Posibles resultados:

- Si la secuencia NO tiene secuencias de vector contaminantes: “Non-significant homology”.
- Si la secuencia SI tiene secuencias de vector contaminantes:
 - Sección gráfica: con diferentes colores muestra, sobre el mapa de la secuencia problema, donde se encuentran las secuencias contaminantes.
 - Alineamiento: se muestra el alineamiento entre la secuencia problema y las secuencias contaminantes homólogas de vectores que se encontraron.

Secuencia problema 1

```
>Secuencia_Problema_1
TCTATNGGCGATTGGGTACCGGGCCCCCTCGAGGTGCACGGTATCGATAAGCTTGATA
TCGAATTCATGGGATTCTAACAAACAATAGTTGCTTGTTCATTACCTTGCAATATTAA
TTCACTCATCCAAAAGCTCAAAACTCCCCCAAGATTATCTAACCTCACAATGCGCTC
GTAGACAAGTTGGTGTGGCCCCATGACATGGGACAATAGGCTAGCAGCCTATGCCAAA
ATTATGCCAATCAAAGAATTGGTGAATGGGGATGATCCACTCTCATGGCCCTTACGGCG
AAAACCTAGCCGCCCTCCCTCAACTTAACGCTGCTGGTGCTGAAAAATGTGGGTG
ATGAGAAGCGTTCTATGATTACAATTCTTGTAGGAGGAGTATGTGGACACT
ATACTCAGGTGGTGTGGCGTAACTCAGTACGTCTCGGTGTGCTAGGGTCAAGCAACA
ATGGTTGGTTTTTCATAACTTGCATTATGATCCACCAGGTAAATTATAGGACAACGTC
CCTTGGCGATCTTGAGGAGCAACCTTGATTCAAATTGAACTTCAACTGATGTCT
AAGAATTCTGCAGCCCCGGGGATCCACTAGTTCTAGAGCGGCCACCGCGGTGGAGC
TCCAGCTTGTGAGGAGCAACCTTGATTCAAATTGAACTTCAACTGATGTCTAG
CTGTTCTGTGAAATTGTTATCCGTCACAATTCCACACAATACGAGCCGGAAGC
ATAAAAGTGTAAAGCCTGGGTGCTAATGAGTGAGCTAACTCACATTATTGCGTTGCGC
TCACTGCCGCTTCACTGGGAAACCTGTCGNCCAGCTGCATTATGAATGCCAA
CGCGCGGGAAAAGGCGGGTTTGGCGTATGGGGCGCTTCCGCTTCCGCTACTGG
ACTCNGTTGCGCTCGGTGTCGGCTGCAGNGAGNNNAATCAGCCNCCCCAAAAGGN
GGNNAATCCGGTTANCCNGNAATCCGGGGAAAACNCNNNGAAAAACNTGGGANCAA
AAAGGNCCCCAAAAGGGCCCAGNAACCNNNNAAAAGGGCCNGNTGNNNGGGTTT
TNCCAAAGGGNCCCCCCCCCGNGAANANNNCAAAANTCCCCCTCAATCCAANGG
GGNGAAAACCCCCGGGNANTTAAAANANCGGGGTNTCCCNNGAAAACCCCCNGG
NCNNCCNGGTTCCNACCCGGCCCTAANGAAAATGNCNCNCNTT
```

En la Secuencia Problema 1, ¿Encuentra fragmentos similares a algún vector?

En el caso en que encuentre secuencias contaminantes de vectores, ¿Entre qué nucleótidos se encuentra el inserto de interés?

Elimine las secuencias contaminantes y vuelva a VecScreen con esta nueva secuencia. ¿Qué obtiene?

Se debe entonces proceder a limpiar la secuencia eliminando los fragmentos correspondientes a vector.

4.2 Mapa de restricción

Los mapas de restricción sirven para verificar que la secuencia que se recibió del centro de secuenciación en efecto corresponde a la secuencia que se mando. Igualmente se puede usar esta herramienta para verificar largas secuencias (como genomas bacterianos) que fueron ensambladas a partir de fragmentos mas cortos. El número y tamaño de los fragmentos predichos deben corresponder al mapa de restricción experimental.

Una herramienta que permite hacer este tipo de análisis se encuentra siguiendo el enlace <http://biotools.umassmed.edu>, seleccionando la opción “Restriction Mapping Tool”.

En la siguiente página pegue la secuencia “35_292648_.ab1” y seleccione la opción ‘entire linear map’ y ‘Submit Sequence to wwwtacg’. Note que tiene la opción de escoger que enzimas de restricción desea usar.

Secuencia 35_292648_.ab1

```
>35_292648_.ab1 ABIX Testing -- no comment RESTRICCION
CGGGCGTCACCGCATT TTTTTTTTTTTTTTTTTTTAAGGGATAATCTATTTC
NCTTATTCANANAATTAGTAATTACNCATAACNCNCAACTTGANGCCNCATTATAANG
ATTAGCAGGNCAATTATATAAGNGGGCANCTTTATTTCANACATTAACTTAAATTNN
GGGCAANCCANAAAANGGACAAGTCTAGAGTCNCATTACNGGGNACATATTGCCTNGGG
TTCATCACTCTCNCTTCACATACAAACTTCCATCTTACAAAANAANAGCAACCCTT
GNACCCGGGGCACANGGGGNACATCGGGGGGANAAATTAACGATTTCCCTGGGAACG
GGGACNTCTGAANAGGCAATTGGATCNCAATTAAANGGGCAAGCNTTGCCCTTT
NGGATCANATTNCCTCNCAAATAATTTCGAAAGAATTATAATAATTACNACCCCTT
ATAGCCGGAGCAACAATTGANGCATANGGGATTAGCGGACTTCCTCTGAACGGGGCA
TATCCCGAACCCAANATTACNCATTNCCTAGTACAAGCCTNGGCATCAACATATAGAAA
CNTCCAAGAACATTAGTAGGNAAGCGACAAAATTAACTCCTGGGAACNGCCNNNGAN
GGANAATTGATTACNAGTACCTNGGCTTTAATTNGGGNCGGGGGGGGGGGGGGGGGG
```

La página de resultados le muestra la información de su secuencia, las enzimas que no cortan su secuencia, el número de cortes para cada enzima y un mapa de sus secuencias con los sitios de corte.

¿La secuencia tiene sitios de corte para la enzima cfoI?

¿Si digiriera el fragmento con las enzimas EcoRI y BamH1 y corriera la digestión en un gel de agarosa, qué tamaños de bandas observaría?

4.3 Análisis de la composición del ADN

En esta sección vamos a usar algunos programas del paquete EMBOSS (“The European Molecular Biology Open Software Suite”; <http://emboss.sourceforge.net/>) para calcular algunas estadísticas sobre secuencias de ADN. Mas adelante nos volveremos a encontrar con EMBOSS para desarrollar tareas mas complicadas.

4.3.1 Contenido de G+C

El contenido en G+C de la secuencia de ADN es importante por varias razones. El apareamiento entre las bases G y C es más estable que entre las bases A y T. Así, el contenido en bases de la secuencia determinara el comportamiento de la secuencia en experimentos de laboratorio.

Siguiendo el enlace <http://mobyle.pasteur.fr/cgi-bin/portal.py?form=geecee> llegará a una interfaz web del programa geecee del paquete EMBOSS, que le permite calcular el contenido de G+C de una secuencias de ADN.

¿Cuál es el contenido de G+C de la secuencia 35_292648_.ab1?

4.3.2 Composición monomérica y palabras cortas

También podemos fácilmente calcular las frecuencias de k-meros, i.e., monómeros, dímeros, trímeros, tetrámeros, pentámeros, ...

Siga el enlace <http://mobyle.pasteur.fr/cgi-bin/portal.py?form=compseq> y calcule la proporción de monómeros, dímeros y trímeros de la secuencia 35_292648_.ab1. Presente los resultados en forma tabular..

Capítulo 5

Creación de bases de datos relacionales

En este capítulo vamos a crear bases de datos relacionales usando SQLite¹ como motor de base de datos y la extensión de Firefox SQLite Manager² como interfaz a la base de datos.

Primero tenemos que asegurarnos que el programa sqlite3 está instalado en nuestro computador. Para esto iniciemos el programa **Terminal**³. Una vez en **Terminal** podemos escribir sqlite3 en la línea de comandos, si se obtiene un salida similar a la mostrada en las líneas 15 a 18, sqlite3 está instalado y funcionando correctamente, de lo contrario es necesario descargarlo del sitio web referenciado en la nota al pie número 1

14 [user@server:~]\$ sqlite3 _____ Ejecutando sqlite3 _____
15 SQLite version 3.6.12
16 Enter ".help" for instructions
17 Enter SQL statements terminated with a ";"
18 sqlite>
19 [user@server:~]\$

Una vez hemos comprobado que el motor de bases de datos está instalado y funcionando correctamente, tenemos que asegurarnos que el complemento **SQLite Manager** de **Firefox** esta instalado, para esto, en el **Firefox**, vaya al menú **Herramientas → SQLite Manager**; si esta opción de menú no aparece entonces es necesario instalar la interfaz a SQLite desde el sitio referenciado en la nota al pie número 2.

Una vez hemos comprobado que el **SQLite Manager** está instalado podemos hacer click en el menú **Herramientas → SQLite Manager**, lo que iniciará una ventana como la que se muestra en la figura 5.1

¹<http://www.sqlite.org/>

²<https://addons.mozilla.org/en-US/firefox/addon/5817/>

³Como hacer esto depende del sistema operativo. En MacOSX puede usar spotlight, i.e., el ícono de lupa en la parte superior derecha de su escritorio, y escribir Terminal, luego darle click al ícono del programa.



Figura 5.1: SQLite Manager en Firefox

Aquí podemos empezar a manipular bases de datos relacionales usando el motor sqlite3.

Cree la base de datos PlnTFDB, haciendo click en el botón **New database**. Seleccione el directorio en donde desea guardarla, puede ser en el directorio **Documentos**.

Importe los archivos⁴ tf.csv, Species.csv, Domains.csv⁵ en las tablas TF, Species, y Domains respectivamente, usando la opción **import**, y asegurandose de seleccionar Tab como el separador de campos e indicar que la primera fila consiste en los nombres de los campos. En el siguiente cuadro de diálogo indique el tipo de datos de cada columna.

Cree los siguientes indices⁶:

Tabla TF: Sp_pepid sin duplicados, como una clave primaria, Sp_ID y family_id

Tabla Species: Sp_ID sin duplicados

Tabla Domains: Sp_pepid con duplicados⁷, domainid con duplicados.

Relaciones entre las tablas: El campo Sp_ID de la tabla TF está relacionado con el campo Sp_ID de la tabla Species. El campo Sp_pepid de la tabla Domains está relacionado con el campo

⁴Los archivos pueden ser descargados desde Sicua Plus

⁵Antes de importar abra cada uno de los archivos con un procesador de texto y defina que tipo de columnas aparecen, VARCHAR, NUMERIC, INTEGER, FLOAT

⁶¿Para qué sirven los indices?

⁷¿Por qué es necesario aceptar duplicados?

Sp_pepid de la tabla TF. ¿Qué tipo de relación hay entre los campos: uno-a-uno, uno-a-varios, varios-a-varios? En este ejercicio particular no los estamos usando, pero ¿qué son las claves externas (foreign keys)?

A manera de ejemplo vamos a realizar algunas consultas sencillas a la base de datos. Haga click en la pestaña **Execute SQL**, y en el cuadro **Enter SQL** escriba lo siguiente:

```
SELECT TF.Sp_pepid, TF.family_id, Species.Species_full_name  
FROM TF, Species  
WHERE TF.Sp_ID=Species.Sp_ID
```

Identifique las operaciones de **proyección, selección y conexión (JOIN)** en la anterior declaración SQL.

Vamos a hacer las siguientes consultas a la base de datos: **¿Cuáles son las familias de factores de transcripción presentes en las especies estudiadas?**

```
SELECT DISTINCT TF.family_id  
FROM TF
```

¿Cuántas familias son?

```
SELECT COUNT(DISTINCT TF.family_id)  
FROM TF
```

¿Cuántas familias hay en cada especie?

```
SELECT Species.Species_full_name, COUNT(DISTINCT TF.family_id)  
FROM TF, Species  
WHERE TF.Sp_ID=Species.Sp_ID  
GROUP BY Species.Sp_id
```

Responda las siguientes preguntas:

1. **¿Cuántos genes por familia y por especie hay?**
2. **¿Cuántos genes por especie hay?**
3. **¿Qué dominios están presentes en los genes de la familia MYB de la especie *Arabidopsis thaliana*?**
4. **¿Cuál es el número de dominios diferentes presentes en los genes de las diferentes especies?**
5. **¿Cuál es la especie con mayor número de dominios diferentes?**
6. **¿En que especie y gen se encuentra el dominio mas largo?**
7. **¿Para qué sirve la expresión `limit` en una declaración SQL en SQLite?**

Capítulo 6

Búsquedas en base de datos biológicas - Segunda parte

6.1 PubMed

Esta sección corresponde a una versión modificada del tutorial <http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/>. Esta guía consiste en seguir el tutorial disponible en el enlace anterior y resolver las preguntas que aparecen mas abajo en rojo.

PubMed es la base de datos de literatura mantenida por el NCBI, actualmente tiene alrededor de 19 millones de registros.

6.1.1 Entendiendo la información en los registros de PubMed

Una referencia bibliográfica en PubMed está compuesta de campos que ofrecen información específica (Título, autor, lenguaje, etc) sobre el artículo publicado. La siguiente lista es una muestra de los campos que aparecen generalmente:

- Título del artículo
- Nombres de los autores
- Resumen publicado con el artículo
- Vocabulario controlado de términos de búsqueda (Medical Subject Headings)
- Información sobre la revista
- Instituto o universidad a la que está afiliado el primer autor
- Lenguaje en que el artículo fue publicado

- Tipo de publicación (revisión, carta, nota pequeña, etc)
- Identificador único de PubMed (PubMed Unique Identifier, PMID)

Ejercicios:

- Haga una búsqueda en PubMed e identifique los campos que se mencionaron arriba.
- Realize una búsqueda en PubMed con el término “eye”. ¿Cuáles de los siguientes términos serán recuperados?
 - Eye, chin and forehead
 - Eye, eyelids, cornea, iris, y todos los demás términos que estén subordinados al término “eye” en MeSH.
 - Eye (únicamente)
- ¿Cuál fue la búsqueda exacta que realizó en el paso anterior? Pista: Ubique la caja de texto “Search details” en la página de resultados.
- Haga una búsqueda en la base de datos MeSH usando como palabras clave sus áreas de interés e identifique los términos MeSH asociados.

Preguntas

- ¿En qué consiste el “status” de una entrada en PubMed?
- ¿Cuál es la diferencia entre MEDLINE y PubMed?
- ¿Qué son y para qué sirven los términos MeSH?
- ¿Qué consiste “Automatic Term Mapping”?

6.1.2 Realizando búsquedas

Empleando la opción de búsqueda avanzada, usando la opción “**Search builder**”, recupere todos los artículos científicos publicados por las profesoras Silvia Restrepo y Adriana Bernal desde el 2008 hasta el 2009, responda:

- ¿Cuántos artículos encontró?
- ¿En qué revistas fueron publicados?
- ¿A qué tipo de publicación corresponden?
- ¿Qué términos MeSH hay en común?

- ¿Qué términos MeSH reflejan el tema principal de los artículos?
- Nombre tres referencias relacionadas al artículo mas reciente de la lista de resultados. ¿Cómo las identificó?
- Envíe los resultados de su búsqueda a su correo electrónico, usando la opción “Send to”

6.2 Descarga por lotes usando Entrez

En aquellos casos en que se tiene un colección de identificadores de alguna base de datos consultada por Entrez, el sistema cuenta con una aplicación de descarga por lotes: “Batch Entrez” (<http://www.ncbi.nlm.nih.gov/sites/batchentrez>)

Use el archivo ID_list.txt¹ para hacer una consulta Batch Entrez y responda:

- ¿Cuántos identificadores pueden ser recuperados por Entrez?
- ¿En qué base de datos se encuentran esos registros?
- ¿Existe algún aviso importante para cualquiera de los registros? En caso afirmativo, explique en qué consiste y por qué puede pasar.
- Enumere los pasos a seguir para cambiar la visualización de los registros y obtener las secuencias en formato Fasta y descargarlas en un archivo de texto.

6.3 Recuperar todas las secuencias de un organismo o taxón

En algunas ocasiones es necesario recuperar del NCBI todas las secuencias de ácidos nucleicos o de proteínas para una especie particular o para un grupo de organismos que pertenecen al mismo grupo taxonómico. Podemos empleada el “Taxonomy Browser” del NCBI para simplificar este proceso.

Haga una búsqueda en la base de datos de taxonomía usando *Ornithorhynchus anatinus* como especie de interés. Los nombres vulgares comunes pueden ser usados, pero siempre es preferible emplear el nombre científico. Al llegar al registro para la especie identifique su clasificación taxonómica. El número de registros en cada una de las bases de datos para la especie o grupo seleccionado, aparece como un enlace en la parte derecha de la página de resultados. Siguiendo esos enlaces puede descargar el conjunto completo de secuencias de la base de datos correspondiente.

Responda:

- ¿Cuántas proteínas se encuentran?

¹Disponible en Sicua Plus

- ¿Cuántas secuencias de ácidos nucleicos?
- ¿Qué otro tipo de información podría extraer?

6.4 Recuperar la información publicada sobre un gen

Haga una búsqueda en alguna de las base de datos usando como palabra clave el nombre del gen de interés y el organismo, Por ejemplo, usando la base de datos “Gene”:

```
tpo[sym] AND human[orgn]
```

En la página de resultados, siga el enlace al gen deseado. Si no existe un registro para el gen en las bases de datos seleccionadas, haga una nueva búsqueda en todas las bases de datos “All databases”.

Cuando encuentre el registro para el gen, identifique en la página de resultados el enlace “Link”. Haciendo clic en este enlace desplegará una lista con mas enlaces, seleccione PubMed. Allí encontrará los registros de la base de datos de literatura que hacen referencia a su gen de interés.

Haga una búsqueda en la base de datos de “Gene” usando el nombre de gen “ANAC092” en la especie *Arabidopsis thaliana*.

- Que artículos en pubmed hacen referencia a ese gen?
- Describa el gen usando la información encontrada en la base de datos “Gene”

6.5 Bases de datos en el European Bioinformatics Institute (EBI)

6.5.1 SRS

El “Sequence Retrieval System” (SRS) lo vimos en la Sección3.2. Siga el enlace <http://srs.ebi.ac.uk/srs/doc/index.html> y familiarícese con las opciones de búsqueda de este sistema.

6.5.2 EB-eye

Este es otro sistema de búsqueda en el EBI.

Haga una búsqueda en todas las bases de datos usando las palabras clave “glutathione s-transferase” en la página del EBI (<http://www.ebi.ac.uk/>).

Responda:

- Describa la página de resultados, ¿Cuántas bases de datos fueron consultadas? ¿En qué categorías están agrupadas esas bases de datos?

- ¿Cuántas y cuáles reacciones enzimáticas son mediadas por la enzima glutathione s-transferasa?
- ¿Qué ontologías tienen registros asociados para la enzima? Describalas.
- ¿Qué es una ontología? De ejemplos de algunas ontologías en biología

6.6 Expasy

El Expasy (<http://expasy.org/>) es el “Expert Protein Analysis System” mantenido por el Instituto Suizo de Bioinformática. Como su nombre lo indica está enfocado en el análisis de proteínas.

En esta sección vamos a usar algunas de las aplicaciones que se encuentran en el enlace <http://expasy.org/tools/>, principalmente aquellas con el logo del Expasy.

Use la secuencia de la proteína ANAC092 de *Arabidopsis thaliana* para desarrollar los ejercicios de esta sección.

Responda:

- ¿Cuál es el peso molecular y punto isoeléctrico de la proteína? ¿Qué herramienta usó para calcular esos parámetros?
- ¿Cuántos y cuáles fragmentos se generan luego de una digestión con tripsina? ¿Qué herramienta uso para hacer la predicción? Calcule el punto isoeléctrico y el peso molecular de cada fragmento.
- Identifique la composición de amino ácidos de ANAC092. ¿Qué aplicaciones empleó?

6.7 Mas ejercicios

Encuentren mas guías siguiendo los enlaces <http://www.ncbi.nlm.nih.gov/guide/all/howto/> y http://www.ebi.ac.uk/ind/help/search_help.html.

Capítulo 7

Ontologías en bioinformática: Gene Ontology

Para desarrollar este capítulo tiene que leer la documentación sobre “Gene Ontology” siguiendo el enlace: <http://www.geneontology.org/GO.doc.shtml> y posiblemente seguir otros enlaces que allí se encuentren.

Parte de esta guía es una versión modificada del tutorial encontrado en el enlace: http://www.geneontology.org/teaching_resources/tutorials/2007-10_GO-resources_jblake.doc, pueden seguir independiente ese tutorial para desarrollar mas habilidades usando GO.

- ¿Cuál es el objetivo del proyecto “Gene Ontology” (GO)?
- Describa las tres ontologías que hacen parte de GO.
- ¿En que consiste anotar un producto génico con términos GO?
- Describa en que consisten las versiones “Slim” de GO.
- ¿Cuál es la diferencia entre las ontologías y las anotaciones? Puede revisar los enlaces en la sección de descargas (“Downloads”).

Siga el enlace: <http://www.obofoundry.org/> de “The Open Biological and Biomedical Ontologies”. Seleccione tres ontologías (diferentes a GO) que puedan ser útiles en su investigación y descríbalas brevemente.

7.1 Consultas en GO

Vamos a usar “AmiGO” para hacer consultas a GO. AmiGO es un navegador basado en HTML que facilita la formulación de consultas tanto de las ontologías como de las asociaciones a los genes.

Haga una búsqueda de término usando “carbohydrate metabolism”.

El resultado de la consulta muestra todos los términos que incluyen la cadena de caracteres “carbohydrate metabolism”. Haga clic en el primer término “carbohydrate biosynthetic process”.

Lo primero que ve en cada línea es uno de los símbolos: +, -, o •, como se muestra en la Figura 7.1. El símbolo + puede ser usado para expandir un node, mostrando todos los hijos del término seleccionado. El símbolo – puede ser usado para cerrar el nodo seleccionado. Finalmente • significa que el término no tiene hijos. Luego de esos símbolos va a encontrar las letras P, I o R, que identifican el tipo de relación: “parte de” (“part of”), “es un” (“is a”), o “regula” (“regulates”), respectivamente. Enseguida encuentra el identificador del término y el término. Al término le sigue un número en paréntesis que le indica el numero de productos génicos que ha sido anotados con ese término o a términos mas específicos (hijos).

- all : all [446404 gene products]
 - + GO:0008150 : biological_process [340066 gene products]
 - + GO:0008152 : metabolic process [177489 gene products]
 - + GO:0009058 : biosynthetic process [81620 gene products]
 - **GO:0016051 : carbohydrate biosynthetic process [4904 gene products]**
 - + GO:0019578 : aldaric acid biosynthetic process [0 gene products]
 - + GO:0034637 : cellular carbohydrate biosynthetic process [3048 gene products]
 - + GO:0046399 : glucuronate biosynthetic process [6 gene products]
 - + GO:0009312 : oligosaccharide biosynthetic process [347 gene products]
 - + GO:0019685 : photosynthesis, dark reaction [277 gene products]
 - + GO:0000271 : polysaccharide biosynthetic process [3112 gene products]
 - + R GO:0043255 : regulation of carbohydrate biosynthetic process [213 gene products]
 - + GO:0044238 : primary metabolic process [139847 gene products]
 - + GO:0005975 : carbohydrate metabolic process [18086 gene products]
 - **GO:0016051 : carbohydrate biosynthetic process [4904 gene products]**
 - + GO:0019578 : aldaric acid biosynthetic process [0 gene products]
 - + GO:0034637 : cellular carbohydrate biosynthetic process [3048 gene products]
 - + GO:0046399 : glucuronate biosynthetic process [6 gene products]
 - + GO:0009312 : oligosaccharide biosynthetic process [347 gene products]
 - + GO:0019685 : photosynthesis, dark reaction [277 gene products]
 - + GO:0000271 : polysaccharide biosynthetic process [3112 gene products]
 - + R GO:0043255 : regulation of carbohydrate biosynthetic process [213 gene products]

Figura 7.1: Consultas en “Gene Ontology”

Busque la opción “Graphical View” para visualizar esta sección de GO como un grafo acíclico dirigido, como el que se muestra en el Figura 7.2.

Vamos a realizar otra consulta en GO usando como palabra clave el nombre de un gen (“ANAC092”), como se muestra en la Figura 7.3.

Ya que la búsqueda que realizamos fue muy específica, los resultados nos llevan directamente a la página de descripción de este gen en GO (Figura 7.4). El nombre del gen que usamos, ANAC092, no se encuentra en ninguna otra especie cubierta por GO. En esta página de resultados identifique



Figura 7.2: Visualización del grafo acíclico dirigido de una sección de GO

la sección “Term associations” y siga el enlace, allí encontraremos el conjunto de términos GO que han sido asignados a este gen en particular, ver Figura 7.5.

Haga clic sobre el término **GO:0007275 : multicellular organismal development**. Esto lo conducirá a la página de detalles del término, donde encuentra toda la información disponible sobre el término: nombre e identificador, sinónimos que pueda tener, definición, su posición en la estructura de GO, referencias a bases de datos externas, y los productos génicos asociados a ese término.

- Describa cada uno de los códigos de evidencia asociados a las anotaciones del gene ANAC092 que aparecen en la Figura 7.5
- Haga una lista de los términos GO asociados con este gen. ¿Qué está indicando el calificador “NOT”?
- Describa brevemente la función de este gen.
- Muestre el grafo acíclico dirigido para la sección que incluye el término **GO:0010150 : leaf senescence**

Figura 7.3: Consultas en “Gene Ontology”

Figura 7.4: Resultados de la consulta en “Gene Ontology”, usando el nombre de gen ANAC092

Accession, Term	Ontology	Qualifier	Evidence	Reference	Assigned by
<input type="checkbox"/> GO:0010150 : leaf senescence	33 gene products view in tree	biological process	IMP	TAIR:Publication:501729812	TAIR
			IEP	TAIR:Publication:501736296	TAIR
<input type="checkbox"/> GO:0007275 : multicellular organismal development	24103 gene products view in tree	biological process	ISS With Pfam:PF02365	TAIR:Communication:501714663	TIGR (via TAIR)
<input type="checkbox"/> GO:0010468 : regulation of gene expression	27494 gene products view in tree	biological process	IMP	TAIR:Publication:501736296	TAIR
<input type="checkbox"/> GO:0006979 : response to oxidative stress	2101 gene products view in tree	biological process	IMP	TAIR:Publication:501713092	TAIR
<input type="checkbox"/> GO:0009651 : response to salt stress	537 gene products view in tree	biological process	IEP	TAIR:Publication:501736296	TAIR
<input type="checkbox"/> GO:0010149 : senescence	104 gene products view in tree	biological process	IMP	TAIR:Publication:3011	TAIR
<input type="checkbox"/> GO:0005634 : nucleus	37778 gene products view in tree	cellular component	IDA	TAIR:Publication:501718231	TAIR
<input type="checkbox"/> GO:0046982 : protein heterodimerization activity	1028 gene products view in tree	molecular function	NOT	IPI With AGI LocusCode:AT3G29035	TAIR:Publication:501718231
<input type="checkbox"/> GO:0042803 : protein homodimerization activity	2165 gene products view in tree	molecular function		IPI	TAIR:Publication:501718231
<input type="checkbox"/> GO:0003700 : transcription factor activity	12404 gene products view in tree	molecular function		ISS	TAIR:Publication:1345963
				IPI	TAIR:Publication:501718231

Figura 7.5: Términos GO asociados al gen ANAC092

Capítulo 8

Introducción al análisis de redes usando Cytoscape

En este capítulo vamos a aprender a trabajar con Cytoscape¹. Sigan el tutorial básico que se encuentra en el enlace: <http://cytoscape.wodaklab.org/wiki/Presentations/Basic>. Van a encontrar Cytoscape instalado en sus computadores, así que no tienen que usar la opción de “Java Web Start”.

De la sección “Defining visual styles” del Tutorial 1: Getting Started, responda:

- En la subred que incluye los vecinos más cercano a TP53 ¿Cuál es el tipo mas común de lado/arista? ¿Cuál el menos común?
- ¿Cuántos nodos y aristas hay en la red “DNA replication” que cargó desde Reactome?

Despues de seguir el Tutorial 4: Expression Analysis, responda:

- ¿Cuáles son los valores de expresión en las condiciones (genes pertubados): Gal1, Gal4, and Gal80 para el gene de levadura: YOL051W?
- ¿Cuáles son los vecinos mas cercanos a ese gen (First Neighbors)?

¹<http://www.cytoscape.org/>

Capítulo 9

Análisis de enriquecimiento de anotaciones de genes

Siga el tutorial que esta disponible en el enlace:<http://www.psb.ugent.be/cbd/papers/BiNGO/Tutorial.html>

El archivo `SaltArabidopsis.txt` que está disponible en Sicua Plus, tienen una lista de genes de Arabidopsis que responden diferencialmente al tratamiento con sal, y que fueron identificados a traves de ensayos usando microarreglos de ADN.

Use BinGO para identificar los términos GO que aparecen sobre y sub representados para los genes que aparecen en el archivo `gene_list.txt`. Muestre el grafo de los términos e interprete los resultados.

Capítulo 10

Comparação de Sequência I - Matrizes de pontos

As matrizes de pontos (“Dot Plot”) são ferramentas exploratórias para comparar strings de texto, ou seja, sequências. Entre outros, eles nos permitem encontrar facilmente regiões repetidas em uma sequência comparando-a com ela mesma. Também podemos ter uma boa ideia da estrutura de um gene comparando a sequência de sua região de codificação com a sequência do locus onde se encontra.

Nesta seção, usaremos a implementação de matrizes de pontos do Instituto Suíço de Bioinformática, conhecida como Dot Let¹, que vemos na Figura 10.1.



Figura 10.1: Dot Let @ SIB

Faça uma comparação da sequência encontrada no arquivo aqc-MIR399 com ela mesma. A primeira coisa que você precisa fazer é clicar no botão “Input”, que abrirá a janela mostrada na

¹<http://myhits.isb-sib.ch/cgi-bin/dotlet>

Figura 10.2, dar um nome à sequência e colá-la na caixa correspondente.

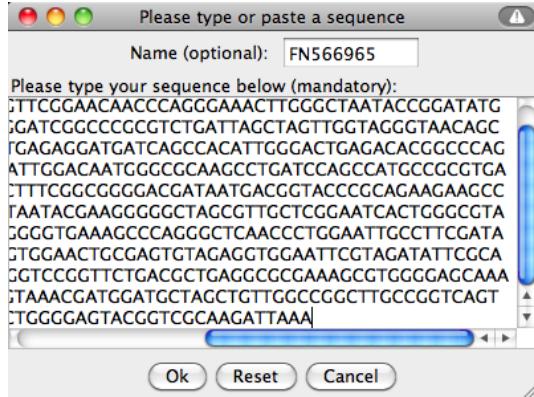


Figura 10.2: Adicionar Sequências em Dot Let

De volta à janela Dot Let vemos que encontramos dois botões habilitados (Figura ??), eles agora aparecem com o nome da sequência que você acabou de adicionar. Uma delas representa a sequência que aparece na direção horizontal, a outra a sequência que aparece na direção vertical.



Figura 10.3: botões de controle

À direita dos botões/listas que identificam as sequências, encontramos uma lista suspensa, atualmente desabilitada, que permite selecionar a matriz de substituição. Em seguida, encontramos uma lista suspensa com os tamanhos de janela que serão usados para a comparação das duas sequências. O próximo botão permite ampliar, ou seja, “Zoom”, e finalmente encontramos o botão “Calcular”, que preenche a matriz de pontos.

Uma vez que a matriz de pontos foi calculada, encontramos duas seções de resultados, semelhantes ao que aparece na Figura 10.4. A região da esquerda é a própria matriz, pixels escuros representam pontuações baixas, ou seja, ruins. À esquerda vemos um histograma da frequência de cada pontuação. Manipulando este histograma, com as barras de rolagem horizontais (para cima e para baixo) podemos modificar a exibição da matriz de pontos.

- Explique como o tamanho da janela afeta a exibição na matriz de pontos.
- Qual é o significado da linha rosa no histograma de pontuação?
- Que interpretação você pode fazer das repetições invertidas que podem ser detectadas na matriz de pontos?

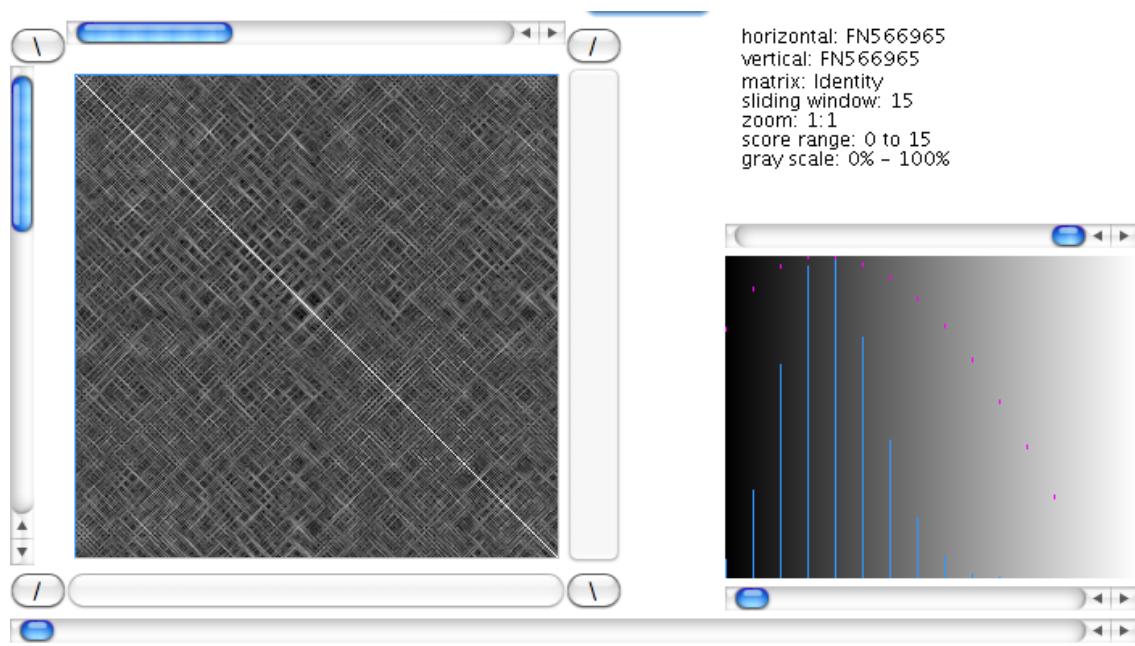


Figura 10.4: Resultado

- Compare a sequência de cDNA e sua contraparte genômica de ANAC092². Descreva os resultados.

²Disponível em e-disciplinas

Capítulo 11

EMBOSS

EMBOSS¹, “The European Molecular Biology Open Software Suite”, é um pacote de código aberto gratuito composto por centenas de apps² que foram desenvolvidos especificamente para atender às necessidades da comunidade de biologia molecular. O tutorial que você encontra abaixo faz parte dos tutoriais disponíveis em http://emboss.sourceforge.net/docs/emboss_tutorial/emboss_tutorial.html

Encontre uma descrição de cada um dos aplicativos presentes no EMBOSS seguindo o link: <http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/>

Algumas das áreas abrangidas pelas aplicações EMBOSS são:

- Alinhamento de sequência
- Pesquisar em bancos de dados usando padrões
- Identificação de motivos proteicos
- Análise de uso de códons

11.1 Recuperando sequências de bancos de dados

A recuperação de sequências de um banco de dados obviamente depende dos bancos de dados que temos disponíveis.

Vamos recuperar a sequência do gene ANAC092 do banco de dados de proteínas UniProt. Para fazer este exercício, você precisa ir ao site UniProt³ e encontrar o identificador apropriado para baixar a sequência em formato .txt Figura 11.2 usando wget (linha 1).

¹<http://emboss.sourceforge.net/>

²<http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/>

³<http://www.uniprot.org/>

obtendo sequência do uniprot

```

1 [user@server]$ wget https://www.uniprot.org/uniprot/D7MJK1.txt
2 --2022-04-06 14:23:06-- https://www.uniprot.org/uniprot/D7MJK1.txt
3 Resolving www.uniprot.org (www.uniprot.org)... 193.62.193.81
4 Connecting to www.uniprot.org (www.uniprot.org)|193.62.193.81|:443... connected.
5 HTTP request sent, awaiting response... 200
6 Length: 4570 (4,5K) [text/plain]
7 Saving to: 'D7MJK1.txt'
8 D7MJK1.txt      100%[=====] 4,46K --.-KB/s   in 0s
9 2022-04-06 14:23:08 (275 MB/s) - 'D7MJK1.txt' saved [4570/4570]
```

Figura 11.1: Recuperando sequências de bancos de dados

Figura 11.2: Recuperando link de sequências em uniprot

O programa que nos interessa chama-se “seqret”.

Agora, você pode usar o mesmo programa para converter a entrada recuperada do UniProt .txt para o formato fasta. **Como você faria isso?**

Já que você sabe como recuperar sequências do UniProt, vamos fazer alguns cálculos sobre essa

sequência, procure o programa `compseq`, que permite calcular a composição da palavra de uma sequência. **Calcular frequências mais fracas para ANAC092 extraídas do UniProt⁴**

11.2 Seleção de quadros de leitura aberta

Os programas `getorf` e `plotorf` buscam quadros de leitura abertos em sequências de nucleotídeos. Sendo um quadro de leitura aberto, uma sequência (subsequência) de um comprimento mínimo especificado flanqueado por dois códons de parada ou por um códon de partida e parada. Apesar da universalidade do código genético alguns grupos de organismos têm códons diferentes, por isso é importante especificar, seja o código genético que está sendo usado para traduzir a sequência, ou o início e parar códons permitidos.

Use esses dois programas para encontrar o quadro de leitura aberto correto da sequência "ANAC092_cDNA.fa"⁵. Você encontra alguma diferença nos resultados oferecidos pelos dois programas?

11.3 Embaralhar/misturar Sequências

Ao fazer certos tipos de análise, por exemplo, procurar sites de vinculação de fatores de transcrição em sequências de promotores ("TFBS"), é importante ter um grupo de sequências que servem como um controle negativo. Para que o TFBS não apareça com frequência neste controle negativo. Uma opção amplamente utilizada é gerar sequências aleatórias que contenham a mesma composição monomérica das sequências originais. O programa "shuffleseq" faz exatamente isso, pega uma sequência "real" e mistura, como se embaralhando um baralho de cartas, os monômeros constituintes, resultando em uma sequência aleatória. Ao usar este tipo de estratégia, 1000 sequências aleatórias são geradas para cada sequência original.

Use "shuffleseq" para gerar duas sequências aleatórias do rRNA encontradas no arquivo `FN566965.fasta`⁶.

11.4 Previsão de regiões hidrofóbicas

O programa `pepwwindow` prevê segmentos hidrofóbicos em uma proteína, seguindo a estratégia proposta por (KYTE and DOOLITTLE, 1982). O uso de janelas de 19 a 21 resíduos regiões transmembranas pode ser claramente detectado, com valores de índice de hidrofobidade de 1,6 na região central.

Pode detectar quaisquer regiões transmembranas no gene `NTM1`⁷?

⁴Ao enviar seu guia esses resultados devem ser enviados em um arquivo de texto plano (.txt).

⁵Arquivo `ANAC092_cDNA.fa` disponível no e-disciplinas

⁶Arquivo `FN566965.fasta` disponível no e-disciplinas

⁷Arquivo `NTM1.fasta` disponível no e-disciplinas

11.5 Alinhamentos

Descreva a função do programa distmat.

Capítulo 12

Comparação de Sequência II - Alinhamentos emparelhados

Algumas partes deste capítulo vêm do tutorial que está seguindo o link: http://emboss.sourceforge.net/docs/emboss_tutorial/emboss_tutorial.pdf

Para fazer os exercícios siga o link <http://mobyle.pasteur.fr/cgi-bin/portal.py>

12.1 Matrizes de substituição

no arquivo arquivo EPAM250.txt você vai encontrar a matriz de substituição PAM250.

- Quem, e como, criou a família PAM de matrizes de substituição?
- Onde estão as maiores pontuações? Explicar.
- Qual é a substituição com a maior pontuação?
- Por que as identidades não têm sempre a mesma pontuação?

12.2 Alinhamento Global

No alinhamento global, o objetivo é comparar as duas sequências ao longo de toda a sua duração, portanto é apropriado quando esperamos que a semelhança entre as duas sequências se estenda ao longo de toda a sequência.

No pacote EMBOSS você encontrará o aplicativo `needle` que implementa rigorosamente o algoritmo de Needleman e Wunsch (NEEDLEMAN and WUNSCH, 1970) para obter o alinhamento global ideal por programação dinâmica. Esta implementação pode levar algum tempo para obter o alinhamento quando as sequências são longas.

Quais outros aplicativos no EMBOSS permitem que você faça alinhamentos globais? O que os torna diferentes de `needle`?

- Faça um alinhamento global entre o cDNA e as sequências genômicas do gene ANAC092, que estão disponíveis no e-disciplinas
- Qual ação e a matriz de penalidades para abrir e estender GAPs que você usou? Explicar.
- Qual é a pontuação de alinhamento, seu comprimento e os percentuais de identidade e semelhança?
- Explique a diferença entre semelhança e identidade.
- O que significam os símbolos? :, . y | ?

Nos arquivos `ANAC092_pep.fasta` y `PpNAC_e_gw1.5.134.1.fasta` encontra as sequências de aminoácidos de dois genes da família NAC de fatores de transcrição em *Arabidopsis thaliana* e em musgo *Physcomitrella patens* respectivamente.

- Faça um alinhamento global entre as sequências de aminoácidos das proteínas NAC de *Arabidopsis thaliana* e em musgo *Physcomitrella patens*.
- Qual ação e a matriz de penalidades para abrir e estender GAPs que você usou? Explicar.
- Qual é a pontuação de alinhamento, seu comprimento e os percentuais de identidade e semelhança?
- Você pode melhorar o alinhamento escolhendo outros parâmetros?

12.3 Alinhamentos locais

Como mencionado na seção anterior, o alinhamento global alinha sequências ao longo de todo o seu comprimento. Você tem que decidir se essa estratégia é a mais apropriada em cada caso. **O que você acha que aconteceria se você comparar duas proteínas multi dominio que só compartilham um domínio entre elas?**

O objetivo do alinhamento local é encontrar regiões de similaridade local, e não é necessário incluir as sequências completas. Esse tipo de alinhamento é muito útil para pesquisar bancos de dados, ou quando você não tem uma ideia clara sobre a semelhança da sequência de interesse com sequências no banco de dados.

No pacote EMBOSS você encontrará o aplicativo `water` que implementa rigorosamente o algoritmo de smith y Waterman (SMITH and WATERMAN, 1981) para obter o alinhamento local ideal

por programação dinâmica. Esta implementação pode levar algum tempo para obter o alinhamento quando as sequências são longas.

Que outros aplicativos no EMBOSS permitem que você faça alinhamentos locais? O que os torna diferentes de water?

- Faça um alinhamento local entre as sequências de aminoácidos das proteínas NAC de *Arabidopsis thaliana* e no musgo *Physcomitrella patens*, que você usou na seção anterior.
- Qual ação e a matriz de penalidades para abrir e estender lacunas que você usou? Explicar.
- Qual é a pontuação de alinhamento, seu comprimento e os percentuais de identidade e semelhança?
- Você pode melhorar o alinhamento escolhendo outros parâmetros?
- Quais são as diferenças entre o alinhamento global e local dessas duas sequências?

12.4 Significado dos alinhamentos

Não importa quais sequências você dá aos programas de alinhamento, eles sempre criará um alinhamento.

Pegue as sequências de aminoácidos ANAC092 e use o programa shuffleseq, e criar duas sequências aleatórias com a mesma composição monomérica de ANAC092. Faça um alinhamento global e local com as duas sequências.

- Qual ação e a matriz de penalidades para abrir e estender lacunas que você usou? Explicar.
- Qual é a pontuação de alinhamento, seu comprimento e os percentuais de identidade e semelhança?
- Você pode melhorar o alinhamento escolhendo outros parâmetros?

Agora faça um alinhamento local e global entre a sequência de aminoácidos do ANAC092 e uma das versões aleatórias.

- Qual ação e a matriz de penalidades para abrir e estender lacunas que você usou? Explicar.
- Qual é a pontuação de alinhamento, seu comprimento e os percentuais de identidade e semelhança?
- Você pode melhorar o alinhamento escolhendo outros parâmetros?

Capítulo 13

BLAST: BASIC LOCAL ALIGNMENT SEARCH TOOL

Muitos de vocês conhecem a interface web BLAST no NCBI mostrado na Figura 13.2. Na primeira parte deste tutorial vamos fazer alguns exercícios usando esta interface.

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Figura 13.1: Tipos de BLAST disponíveis no NCBI

No e-disciplinas encontra `desconocido.nuc.fa`, que contém a sequência nucleotídea de uma transcrição que você descobriu analisando a expressão diferencial de *A. thaliana* em resposta à luz ultravioleta (UV-A), tratamento no qual esta transcrição foi induzida. Copie a sequência da transcrição e abra o site <http://blast.ncbi.nlm.nih.gov/> no navegador Firefox. Vamos realizar uma pesquisa blast básica, olhar na página para uma seção como a que aparece na Figura 13.1 e selecionar o `blastx`. **Por que usar `blastx`?**

No `blastx` cole sua sequência desconhecida no campo “**Enter query sequence**”, escreva *Viridiplanteae* no campo “**Organism**”, para estreitar a busca por sequências de plantas verdes (Figura 13.2). Certifique-se de que o banco de dados selecionado seja o banco de dados não redundante de sequências proteicas.

Em buscas envolvendo a tradução online de uma sequência de DNA você pode selecionar o código

The screenshot shows the NCBI BLAST web interface. In the top navigation bar, the 'blastx' tab is selected. The main area is titled 'Enter Query Sequence' and contains a text input field with the sequence: >desconocido [REDACTED]. Below this, there are fields for 'Or, upload file' (Choose File), 'Genetic code' (Standard (1)), 'Job Title' (desconocido), and 'Align two or more sequences' (checkbox). On the right, there is a 'Query subrange' section with 'From' and 'To' fields. Under 'Choose Search Set', the 'Database' is set to 'Non-redundant protein sequences (nr)' and the 'Organism' is set to 'Vireobalanus (taxid:33090)'. There are also 'Exclude' and 'Optional' sections.

Figura 13.2: Interface web NCBI BLAST usando o programa blastx

genético que será usado para fazer a tradução. Certifique-se de que o código genético selecionado neste caso é “Standard”.

This screenshot shows the 'Algorithm parameters' section of the BLAST search interface. It includes three main tabs: 'General Parameters', 'Scoring Parameters', and 'Filters and Masking'. In the 'General Parameters' tab, the 'Max target sequences' is set to 100, the 'Expect threshold' is set to 10, and the 'Word size' is set to 3. In the 'Scoring Parameters' tab, the 'Matrix' is set to BLOSUM62 and the 'Gap Costs' are set to Existence: 11 and Extension: 1. In the 'Filters and Masking' tab, the 'Filter' checkbox for 'Low complexity regions' is checked, while the 'Mask' checkboxes for 'Mask for lookup table only' and 'Mask lower case letters' are unchecked.

Figura 13.3: Parámetros de búsqueda en BLAST

Un poco más abajo, haga click en el vínculo “**Algorithm parameters**”, lo que le mostrará la serie de opciones que se ven en la Figura 13.3. En la sección de “**General parameters**”, encuentra el **Expected threshold** o **E value**. El E value es el número de alineamientos con un puntaje igual o mayor al obtenido que se espera que aparezcan por azar. En el momento de

seleccionar los alineamientos importantes este es el parámetro mas importante; como regla general alineamientos con E value menor que 1×10^{-5} representan secuencias homólogas. Sin embargo si está alineando secuencias muy cortas, e.g., 20 residuos, debe permitir alineamientos con un E value muy alto, alrededor de 100. En la sección “**Scoring parameters**”, puede seleccionar la matriz de sustitución (escoja BLOSUM80) y la penalización por introducir gaps en el alineamiento. Note que hay una diferencia entre el costo de introducir un gap y el de extenderlo **¿A qué se debe esa diferencia?** Las opciones de abrir y extender gaps dependen de la matriz de sustitución seleccionada. Por favor observe como cambiando de matriz estas opciones cambian¹.

Asegúrese que la opción **Filter** en la sección **Filters and Masking** esté seleccionada, con el fin de reducir el número de alineamientos con secuencias no relacionadas evolutivamente. **¿Qué programas usa BLAST para detectar regiones de baja complejidad? ¿Qué funciones cumplen las opciones “Mask for lookup table only” y “Mask lower case letters”?**

Ahora pinche el botón BLAST y espere sus resultados.

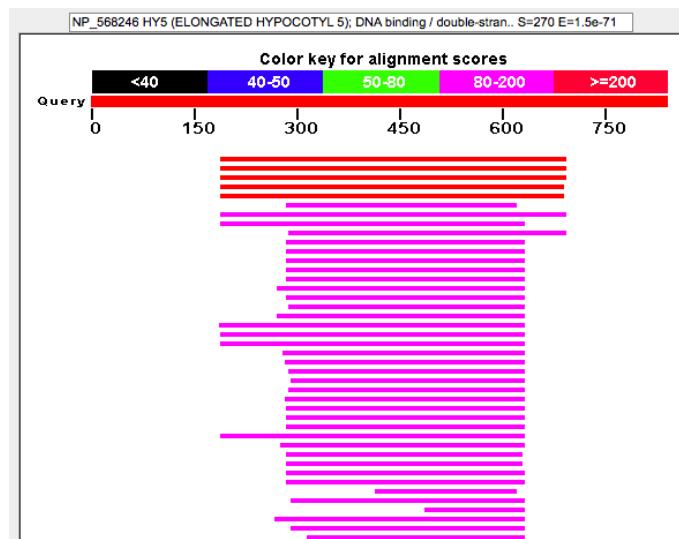


Figura 13.4: Representación gráfica de los mejores alineamientos obtenidos en la búsqueda con blastx

En la parte superior de la página de resultado encuentra una gráfica como la que se ve en la Figura 13.4. Consiste en una representación de los mejores alineamientos con un código de colores que representa la longitud del alineamiento.

Un poco mas abajo encuentra la tabla con los mejores hits, donde se muestra el identificador (Accesion number) de la secuencia hit, parte de su descripción, el puntaje del alineamiento entre su secuencia desconocida y la secuencia de la base de datos, el porcentaje de la secuencia “query” que está representada en el alineamiento, la identidad y el E value. Puede re-ordenar los datos en esta tabla pinchando en los nombres de las columnas.

¹En el enlace http://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html encontrará mayor información sobre la matriz de sustitución y la penalización de gaps.

Sequences producing significant alignments:	Score (Bits)	E Value	
ref NP_568246.1 HY5 (ELONGATED HYPOCOTYL 5); DNA binding / d...	270	1e-71	UG
gb ABY83460.1 elongated hypocotyl 5 protein [Brassica rapa s...	220	2e-56	
ref XP_002515537.1 transcription factor hy5, putative [Ricin...	201	9e-51	G
ref XP_002324289.1 predicted protein [Populus trichocarpa] >...	201	9e-51	UG
ref XP_002308656.1 predicted protein [Populus trichocarpa] >...	200	2e-50	UG
gb AAO2523.1 HY5 [Brassica rapa subsp. pekinensis]	192	6e-48	
emb CAN83322.1 hypothetical protein [Vitis vinifera]	190	3e-47	
sp Q9SM50_1 HY5 SOLIC RecName: Full=Transcription factor HY5;...	179	4e-44	G
emb CAO44204.1 unnamed protein product [Vitis vinifera]	173	3e-42	
gb ACU17915.1 unknown [Glycine max]	170	2e-41	
gb ACP28170.1 LONG1 [Pisum sativum] >gb ACP28171.1 LONG1 [P...	170	2e-41	
gb AAC05018.1 TGACG-motif-binding factor [Glycine max]	170	2e-41	G
gb AAC05017.1 TGACG-motif binding factor [Glycine max] >gb A...	170	2e-41	G
emb CAA66478.1 bZIP transcription factor [Vicia faba var. minor]	166	5e-40	
ref XP_002453510.1 hypothetical protein SORBIDRAFT_04g007060...	165	7e-40	UG
dbj BAC20318.1 bZIP with a Ring-finger motif [Lotus japonicu...	163	3e-39	G
ref XP_002437242.1 hypothetical protein SORBIDRAFT_10g023420...	159	5e-38	UG
ref NP_001152483.1 transcription factor HY5 [Zea mays] >gb A...	158	9e-38	UG
ref NP_001046236.1 Os02g0203000 [Oryza sativa (japonica cult...	158	1e-37	
gb ECC72704.1 hypothetical protein OsI_06291 [Oryza sativa I...	157	2e-37	
dbj BAD15505.1 putative bZIP protein HY5 [Oryza sativa Japoni...	157	2e-37	
gb ABK23948.1 unknown [Pinus sitchensis]	139	6e-32	
ref NP_001058004.1 Os06g0601500 [Oryza sativa (japonica cult...	137	1e-31	
gb ECC80926.1 hypothetical protein OsI_23604 [Oryza sativa I...	136	4e-31	
gb ABK26016.1 unknown [Pinus sitchensis]	128	1e-28	
ref NP_001147637.1 transcription factor HY5 [Zea mays] >gb A...	127	2e-28	UG
gb EAY72732.1 hypothetical protein OsI_00597 [Oryza sativa I...	127	2e-28	

Figura 13.5: Listado de “Hits”

La última parte de la sección de resultados esta compuesta por los alineamientos propiamente dichos (Figura 13.6). Aquí va a encontrar nuevamente el puntaje y el E value del alineamiento. Adicionalmente, además del alineamiento, encuentra el número de posiciones en que las dos secuencias eran idénticas y similares (de acuerdo a la matriz de sustitución) y el número de gaps.

¿Qué indican las regiones de los alineamientos que aparecen en gris y en minúscula?

```
>□ref|NP_568246.1| UG HY5 (ELONGATED HYPOCOTYL 5); DNA binding / double-stranded DNA
binding / transcription factor [Arabidopsis thaliana]
  sp|Q24646_1|HY5 ARATH G RecName: Full=Transcription factor HY5; AltName: Full=Protein
  LONG HYPOCOTYL 5; AltName: Full=bZIP transcription factor 56;
  Short=AtbZIP56
  dbj|BAA21116.1| G HY5 [Arabidopsis thaliana]
  dbj|BAA21327.1| G HY5 [Arabidopsis thaliana]
  emb|CAB96661.1| G HY5 [Arabidopsis thaliana]
  gb|ABF58937.1| G At5g11265 [Arabidopsis thaliana]
  dbj|BAF01225.1| G bzip transcription factor HY5 / AtbZip56 [Arabidopsis thaliana]
Length=168

  GENE ID: 830996 HY5 | HY5 (ELONGATED HYPOCOTYL 5); DNA binding /
  double-stranded DNA binding / transcription factor [Arabidopsis thaliana]
  (Over 10 PubMed links)

  Score = 216 bits (549), Expect = 3e-55
  Identities = 168/168 (100%), Positives = 168/168 (100%), Gaps = 0/168 (0%)
  Frame = +3

  Query 192 MQEQATSSLAASSLPSSSERSSSSAPHLIEIKEKGIESDEEIRVPFGEAEVGKETSGRES 371
  Sbjct  1 MQEQATSSLAASSLPSSSERSSSSAPHLIEIKEKGIESDEEIRVPFGEAEVGKETSGRES 60

  Query 372 GSATQERTQATVGEESQRKRGRTPAeknkrklrrlnrrVSQQARERKKAYLSELENRV 551
  Sbjct  61 GSATQERTQATVGEESQRKRGRTPAeknkrklrrlnrrVSQQARERKKAYLSELENRV 120

  Query 552 KDLENKNSELEERLSTLQNENQMLRHILkNtttgknrggggSNADASL 695
  KDLENKNSELEERLSTLQNENQMLRHILkNtttgknrggggSNADASL
  Sbjct 121 KDLENKNSELEERLSTLQNENQMLRHILkNtttgknrggggSNADASL 168

>□gb|ABY83460.1| elongated hypocotyl 5 protein [Brassica rapa subsp. rapa]
Length=167

  Score = 173 bits (439), Expect = 2e-42
```

Figura 13.6: Alineamientos resultantes de la búsqueda con blastx

¿Qué puede decir sobre la función de su transcripto?

La interfaz web de NCBI BLAST es muy amigable, pero tiene un par de problemas cuando trabajamos en genómica y proteómica, (i) no se pueden hacer búsquedas contra bases de datos personalizadas o privadas y (ii) el número de secuencias que puede usar como query en cada

búsqueda está restringido. La alternativa más poderosa para solucionar ambos problemas es instalar NCBI BLAST en un computador local y configurar las bases de datos sobre las cuales se quiere realizar búsquedas (ver sección 13.2).

13.1 Encontrando la región genómica de un transcripto.

Use la secuencia que se encuentra en el archivo `desconocido.nuc.fa` para hacer una búsqueda BLAST, usando `blastn` contra el genoma completo de *A. thaliana*. **¿Que opciones tiene que seleccionar para restringir su búsqueda a los cromosomas de *Arabidopsis thaliana*?** Ya que BLAST realiza la búsqueda usando alineamientos locales, este resultado solo le dará una idea muy preliminar de la ubicación del transcripto en el genoma. Pero puede usar esta información para refinar la predicción del locus del transcripto usando `est2genome` de EMBOSS.

¿Que opciones seleccionó para hacer la búsqueda en BLAST? ¿Por qué?

Describa los resultados de la búsqueda.

Los resultados de esta búsqueda nos permiten concluir que el locus del transcripto está en el cromosoma número 5 de *A. thaliana*. **¿Cuáles son las coordenadas aproximadas en el cromosoma? ¿Hay exones? Explique su respuesta.** Vamos a usar este resultado como entrada para `est2genome`. Primero extraiga de la secuencia del cromosoma 5, la región detectada por BLAST adicionándole 5000pb corriente arriba y corriente abajo. **¿Cómo puede hacer esto?** Use `est2genome` para refinar la predicción del locus. **¿Qué ventajas ofrece usar `est2genome` comparado con un simple BLAST?**

Para finalizar siga el enlace <http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=comgen&part=psibl> y desarrolle el tutorial de PSI-BLAST.

13.2 Blast+ en la línea de comandos

Los ejecutables mas recientes, para diferentes plataformas, de la suite Blast+ del NCBI los puede encontrar siguiendo el enlace <ftp://ftp.ncbi.nih.gov/blast/executables/blast+/LATEST>.

Para saber si la suite Blast+ está instalada en su computador ejecute el comando `blastp`, si la respuesta del sistema operativo es comando no encontrado tendrá que descargar e instalar la suite Blast+. De lo contrario ya está listo para empezar a usar Blast+ desde la línea de comandos.

Hay muchas opciones en los diferentes programas que componen la suite BLAST+, en este ejercicio solo tendremos tiempo de revisar unas pocas. Puede encontrar la documentación sobre estos en los siguientes enlaces:

- http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs
- http://www.ncbi.nlm.nih.gov/books/NBK1763/#CmdLineAppsManual.5_Cookbook

- http://www.ncbi.nlm.nih.gov/books/NBK1763/#CmdLineAppsManual.4_User_manual

Para desarrollar el ejercicio de hoy, descargue el archivo TAIR10_pep_20101214.gz y descomprimalo como se muestra en la línea 3. En este archivo encuentra todas las proteínas anotadas de *Arabidopsis thaliana* correspondientes a la versión 10 de la anotación del genoma. Descargue las secuencias, en formato FastA, con números de acceso BAK64065 y XP_002889081. Su objetivo es encontrar el mejor hit en la base de datos de proteínas de *A. thaliana*, usando BLAST desde la línea de comandos.

Ejecutando Blast+ en CLI

```

1 [user@server:~]$ mkdir ejercicio_blast
2 [user@server:~]$ cd ejercicio_blast
3 [user@server:~]$ wget http://biocomp-cms.uniandes.edu.co/exchange/TAIR10_pep_20101214.gz
4 [user@server:~]$ gunzip TAIR10_pep_20101214.gz
5 [user@server:~]$ makeblastdb -in TAIR10_pep_20101214 -dbtype prot -parse_seqids -taxid 3702
6 [user@server:~]$ blastp -query BAK64065.fasta -task blastp -db TAIR10_pep_20101214 \\
7 -out BAK64065.blastp.out.txt -evalue 1e-5 -matrix BLOSUM62 -num_descriptions 1 -num_alignments 1
8 [user@server:~]$ blastp -query BAK64065.fasta -task blastp -db TAIR10_pep_20101214 \\
9 -out BAK64065.blastp.out.xml -evalue 1e-5 -matrix BLOSUM62 -num_descriptions 1 -num_alignments 1 \\
10 -outfmt 7
11 [user@server:~]$
12 [user@server:~]$
13 [user@server:~]$
14 [user@server:~]$
15 [user@server:~]$

```

Antes de poder hacer búsquedas usando BLAST es necesario reformatear el archivo que nos va a servir como base de datos. El comando makeblastdb que se distribuye con la suite BLAST+ es el encargado de realizar esta tarea. En general para obtener información sobre como usar diferentes programas de la suite puede ejecutar nombre_programa -help. La línea 5, muestra el comando que debe ejecutar para crear la base de datos en formato blast. **¿Para que sirven cada uno de los argumentos que se pasan al programa makeblastdb?**

Teniendo la base de datos en el formato adecuado podemos hacer nuestra primera búsqueda. Use la secuencia proteínas BAK64065. Usaremos el programa blastp, para buscar el mejor hit de una proteína en una base de datos de proteínas (Línea 6). **¿Para que sirven cada uno de los argumentos que se pasan al programa blastp?**. Revise el archivo de salida, lo puede hacer con cualquiera de los siguientes comandos: pico, less. En la línea 8, encuentra básicamente el mismo comando, solo que esta vez pedimos que el formato de salida sea tabular con la opción -outfmt 7. Hay muchos otros formatos de salida que se pueden pedir durante la búsqueda con el parámetro -outfmt. **Describa los formatos de salida posibles.**

Haga la búsqueda con blastp para la secuencia con número de acceso XP_002889081, solo muestre los primeros 3 hits con e-value igual o menor que 10^{-10} . Asegúrese de solicitar un formato de salida tabular que incluya la longitud de las secuencias “Query” y “Subject”.

Capítulo 14

Alinhamientos múltiplos

Na teoria, os algoritmos de programação dinâmica descritos acima para o caso de alinhamentos emparelhados podem ser estendidos para o caso de um número arbitrário de sequências. Na prática, isso é muito computacionalmente caro, por isso outros algoritmos foram desenvolvidos que implementam atalhos na busca de alinhamentos ideais (heurística). O desenvolvimento de algoritmos para alinhamento de múltiplas sequências é uma das áreas mais dinâmicas da bioinformática. Atualmente existem dezenas de programas que implementam algoritmos diferentes (olha NOTREDAME, 2007 e LEMEY *et al.*, 2009 para uma revisão recente do tópico).

Nesta sessão vamos desenvolver a prática apresentada no capítulo 3 de LEMEY *et al.*, 2009.

Vamos alinhar as sequências dos genes TRIM5 de diferentes espécies de primatas. TRIM5 é um fator de restrição viral que protege a maioria dos macacos do velho mundo (*Cercopithecidae*) da infecção pelo HIV. Esses dados foram originalmente analisados por SAWYER *et al.*, 2005. Usaremos métodos de refinamento iterativo (MUSCLE) para criar vários alinhamentos proteicos e, em seguida, comparar os resultados usando JALVIEW. Vamos criar os alinhamentos das sequências proteicas, e vamos gerar o alinhamento correspondente no nível nucleotídeo, terminando com a inspeção manual e o refinamento do alinhamento.

14.1 Alinhando as sequências de aminoácidos de TRIM5 de primatas

14.1.1 MUSCLE

Para obter o alinhamento das sequências no arquivo `primatesAA.fasta` pelo método de refinamento iterativo, usaremos o programa MUSCLE (EDGAR, 2004). Este programa está instalado localmente em seu computador.

1 [user@server]\$ conda activate muscle_env
2 (muscle_env) [user@server]\$ muscle -h

```

----- MUSCLE -----
1  (muscle_env) [user@server]:~$ muscle -align primatesAA.fasta -output primates.afa
2
3  muscle 5.1.linux64 [] 4.0Gb RAM, 4 cores
4  Built Feb 24 2022 03:16:15
5  (C) Copyright 2004-2021 Robert C. Edgar.
6  https://drive5.com
7
8  Input: 22 seqs, avg length 504, max 551
9
10 00:00 17Mb  CPU has 4 cores, running 4 threads
11 00:04 244Mb  100.0% Calc posteriors
12 00:04 246Mb  100.0% Consistency (1/2)
13 00:04 246Mb  100.0% Consistency (2/2)
14 00:04 246Mb  100.0% UPGMA5
15 00:04 248Mb  100.0% Refining

```

14.1.2 Visualização e edição de Alinhamentos

Usaremos o aplicativo JalView instalado em seus computadores para exibir os alinhamentos. Pode utilizar o arquivo `primates.afa` gerado pelo MUSCLE.

Siga o link <http://www.jalview.org/examples/editing.html>, se desejar, poderá revisar a documentação completa no link <http://www.jalview.org/help.html>. (Figura 14.1)

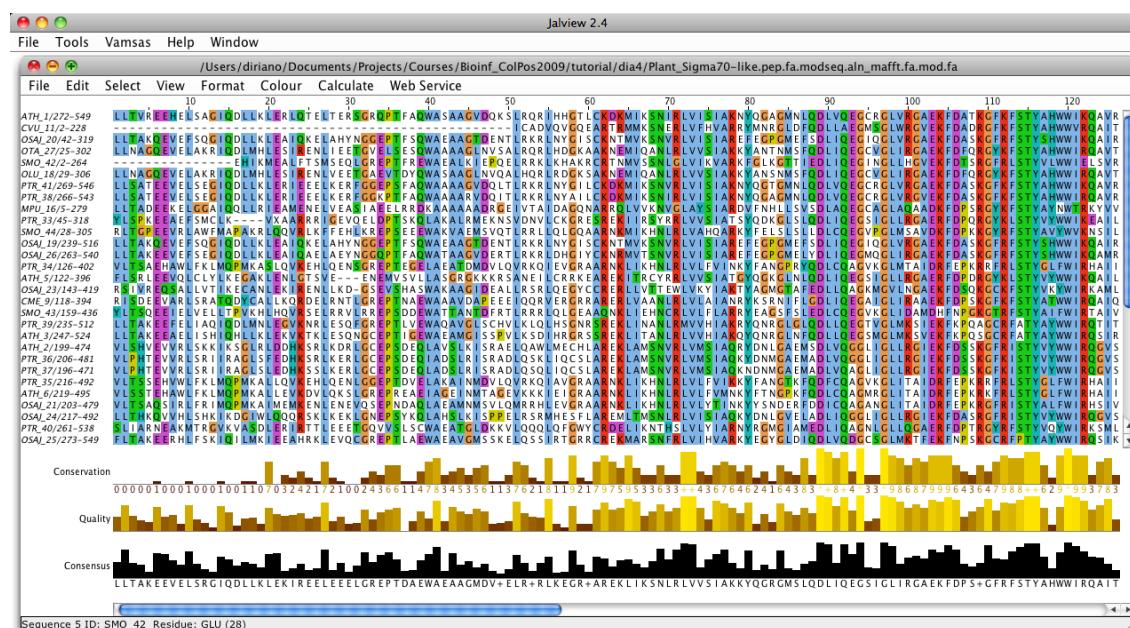


Figura 14.1: Tela do JalView

Capítulo 15

PSSMs, Logo de Sequências e HMMs

15.1 PSSM

As matrizes de pontuação específicas da posição “Position-specific scoring matrices” (PSSMs) oferecem uma maneira sensível de representar a variabilidade em um alinhamento. Os PSSMs são construídos com base no alinhamento múltiplo, por exemplo, dos sites de vinculação de fatores de transcrição.

Abaixo está uma matriz que foi obtida a partir da base de promotores de *Saccharomyces cerevisiae*¹ e construída usando um alinhamento de 12 sites de vinculação do fator de transcrição de levedura Pho4p.

A	3	2	0	12	0	0	0	0	1	3
C	5	2	12	0	12	0	1	0	2	1
G	3	7	0	0	0	12	0	7	5	4
T	1	1	0	0	0	0	11	5	4	4

Cada linha representa um resíduo (A, C, G ou T) e cada coluna uma posição no conjunto de sequências alinhadas. Algumas posições são perfeitamente preservadas em todas as sequências, enquanto outras apresentam algumas alternativas.

Ao utilizar esses tipos de matrizes para pesquisar, as posições mais conservadas impõem restrições mais fortes do que aquelas em que qualquer resíduo pode ser apresentado.

Siga o link <http://rulai.cshl.edu/cgi-bin/SCPD/getfactor?ABF1>, BAF1 e responde:

- Qual é o tamanho do alfabeto?
- Qual é a largura da matriz?
- Quantos sites de vinculação Abf1p estão armazenados no Banco de Dados de Promotores de Levedura (SCPD)?

¹<http://rulai.cshl.edu/SCPD/>

- Quais programas EMBOSS eu poderia usar para pesquisar com PSSMs?

15.2 logo de sequências

Logo de sequência são uma representação gráfica de alinhamentos múltiplos baseados na teoria das informações²

A Figura 15.1 corresponde ao logo da sequência do site de vinculação do fator de transcrição LexA de *Escherichia coli*

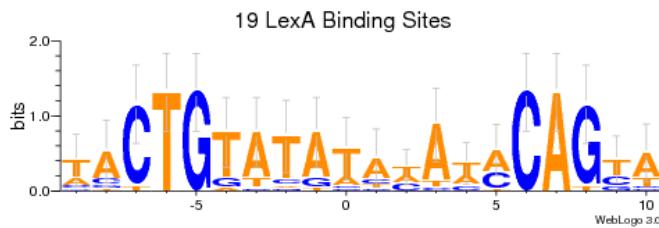


Figura 15.1: LexA junta-se ao logo da sequência do site

A altura do resíduo está correlacionada com sua frequência no alinhamento múltiplo Para obter mais informações, consulte: SCHNEIDER and STEPHENS, 1990.

Seguindo o link <http://weblogo.threethreeplusone.com/>, crie o logo das sequências dos sites de vinculação do fator de transcrição Abf1p que estudeu na seção anterior. Para realizar este exercício, você precisa recuperar todos os sites de ação do Abf1p disponíveis no SCPD.

O que o eixo y representa? ou seja, como o conteúdo das informações de cada posição é calculado?

15.3 Modelos ocultos de Markov: HMMs

Mesmo que você não encontre proteínas homólogas ao realizar uma busca com blast, você ainda tem outras opções.

A principal razão pela qual ele não encontra homologues triviais é que pesquisas com sequências usando ferramentas como BLAST têm baixa sensibilidade. O BLAST normalmente não encontra proteínas homólogas que tenham menos de 30% de identidade. No entanto, algumas proteínas podem ter a mesma estrutura tridimensional e ter apenas 10% de identidade.

Uma estratégia muito útil para encontrar contrapartes distantes é baseada no uso de Modelos Hidden Markov (HMMs). Um HMM nada mais é do que uma maneira de definir motivos ou domínios.

²https://en.wikipedia.org/wiki/Sequence_logo

Para criar um HMM tudo o que você precisa é de um alinhamento múltiplo, que será usado para criar uma representação probabilística, que pode então ser usada para procurar sequências relacionadas.

Os bancos de dados Pfam (FINN *et al.*, 2010) e SUPERFAMILY (WILSON *et al.*, 2009) são coleções de múltiplos alinhamentos para os quais os HMMs foram criados e são usados para anotar sequências proteicas. A maior parte do trabalho dos curadores dessas bases de dados é criar os alinhamentos múltiplos.

15.3.1 Procurando os domínios de uma proteína

Vamos usar a seguinte proteína para fazer uma pesquisa no Pfam:

```
proteína desconhecida
>seq
MEYWHYVETTSSGQPLLREGEKDIFIDQSVGLYHGKSKILQRQRGRIFLTSQRIIYIDDAKPTQ
NSLGLELDDLAYVNYYSSGFLTRSPRLILFFKDPSSKDELGKSAETASADVSTWVCPICMVSNETQGEFTKD
TLPTPICINCGVPADYELTKSSINCSNAIDPNANPRNQFGVNSENICPACTFANHPQIGNCEICGHRLPNAS
KVRSKLNRLNFHDSRVHIELEKNSLARNKSSHSALOSSSSTGSSTEFVQLSFRKSDGVLFQSATERALENIL
TEKNKHIFN
```

Vá para o site de Pfam³ e selecione "Search" depois "Sequence" e cole a sequência da proteína de interesse na caixa de texto para a busca e clique no botão "submmitt" para iniciar a busca. Figura 15.2 exibe a página de resultados.

Significant Pfam-A Matches														
<a>Show or hide all alignments.														
Family	Description	Entry type	Cian	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
<a>Vps36_ESCRT-II	Vacuolar protein sorting protein 36 Vps3	Domain	<a>CL0266	8	96	8	95	1	91	92	97.8	3.3e-28	n/a	<a>Show
<a>Vps36-NZF-N	Vacuolar protein sorting 36 NZF-N zinc-f	Domain	n/a	109	172	111	166	3	59	65	66.9	8.1e-19	n/a	<a>Show
Insignificant Pfam-A Matches														
<a>Show or hide all alignments.														
Family	Description	Entry type	Cian	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
<a>Vps36-NZF-N	Vacuolar protein sorting 36 NZF-N zinc-f	Domain	n/a	172	207	178	191	7	20	65	0.9	320	n/a	<a>Show
<a>zf-Sec23_Sec24	Sec23/Sec24 zinc finger	Domain	n/a	112	133	116	131	22	37	39	12.7	0.11	n/a	<a>Show
<a>zf-Sec23_Sec24	Sec23/Sec24 zinc finger	Domain	n/a	169	207	182	203	11	32	39	7.4	4.9	n/a	<a>Show
<a>DZR	Double zinc ribbon	Family	<a>CL0167	183	224	183	208	1	26	49	13.8	0.048	n/a	<a>Show
<a>Rubredoxin	Rubredoxin	Domain	<a>CL0045	95	130	106	127	23	44	47	12.7	0.11	n/a	<a>Show
<a>Rubredoxin	Rubredoxin	Domain	<a>CL0045	171	206	176	201	16	41	47	2.0	230	n/a	<a>Show
<a>DZR_2	Double zinc ribbon domain	Domain	<a>CL0167	112	165	138	153	34	49	56	4.3	41	n/a	<a>Show

Figura 15.2: Resultados de Pfam

- Qual é a diferença entre "Significant Pfam-A matches" e "Insignificant Pfam-A matches"?
- O que é Pfam-A?
- O que é Pfam-B?

³<https://pfam.xfam.org/>

A primeira seção de resultados “Significant Pfam-A matches”, informa-nos que há dois um hits, o modelo ”Vps36_ESCRT-II”, com uma pontuação de 97.8 e um e-value de 3.3^{-28} e o moduelo ”Vps36-NZF-N” com uma pontuação de 66.9 e um e-value de 8.1^{-19} . Encontramos também as coordenadas do domínio, em relação à proteína de consulta e em relação ao modelo.

Uma das principais características, e valores adicionados da Pfam, é que cada um dos modelos da Pfam-A tem sido estudado por um especialista que definiu uma série de limiares para definir hits significativos. A pontuação de limiar mais importante corresponde ao “gathering cutoff”. Qual é o “gathering cutoff” modelo ”Vps36_ESCRT-II”?

Clique no nome do modelo. Isso o levará a uma página com informações detalhadas sobre esse modelo em particular. Entre outros, você pode descobrir que outras espécies estão presentes nesse modelo (“Species”). Você pode baixar o modelo (‘Curation’) ou o alinhamento múltiplo (“Alignments”), entre outras informações.

[Use a sequência proteica ANAC092 e determine quais domínios estão presentes](#)

15.3.2 Visualização de HMMs

Você também pode exibir HMMs na forma de logo de sequências. Você pode usar o aplicativo LogoMat-M encontrado no link <http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi>. Ou no mesmo site do pfam

[Exibir o logo do domínio “zf-C2H2”](#)

Capítulo 16

Diseño de primers para PCR

La parte teórica que aquí se presenta consiste en un resumen del texto que se encuentra siguiendo el enlace: http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html.

La reacción en cadena de la polimerasa (PCR) inventada por Kary Mullis en la década de los 80s del siglo XX (MULLIS and FALOONA, 1987) es considerada uno de los inventos mas importantes en biología molecular. Mediante esta reacción pequeñas cantidades de material genético se pueden amplificar de tal forma que pueden ser identificadas y/o manipuladas.

La PCR involucra los siguientes pasos:

Denaturación El objetivo de este paso es convertir las moléculas de ADN de doble cadena en cadenas sencillas.

Anillamiento Durante este paso los primers hibridan con las hebras https://www.overleaf.com/project/624c8ebfae5 molde de cadena sencilla.

Extensión La ADN polimerasa extiende los primers.

Esos pasos dependen y son muy sensibles a la temperatura. Las temperaturas usadas comúnmente son 95°C, 60°C y 72°C, respectivamente.

Un buen diseño de primers es esencial para obtener reacciones exitosas. A continuación se describen las principales consideraciones a tener en cuenta durante el diseño.

Longitud de los primers: Normalmente se acepta que la longitud óptima para primers de PCR está entre 18 y 22 pb. Con esta longitud son lo suficientemente largos para asegurar especificidad y lo suficientemente pequeños para que se unan fácilmente al ADN molde a la temperatura de anillamiento.

Temperatura de fusión del primer (T_m): Se define como la temperatura a la cual la mitad de las moléculas de ADN de doble cadena se van a disociar y volverse de cadena sencilla. Es una forma de indicar la estabilidad del duplex. Primers con temperaturas de fusión entre 52°C y 58°C normalmente producen los mejores resultados. Primers con temperaturas de fusión superiores a

65°C tienen tendencia a formar anillamientos secundarios. El contenido de GC de la secuencia da una buena indicación de la temperatura de fusión del primer. Mayor precisión en su cálculo se alcanza empleando la teoría termodinámica de los vecinos más cercanos, según la cual:

$$T_m(^{\circ}C) = \{\Delta H / \Delta S + R \ln(C)\} - 273.15 \quad (16.1)$$

donde:

ΔH (kcal/mol) : H es la entalpía. La entalpía es la cantidad de energía calórica que poseen las sustancias. ΔH es el cambio en entalpía. En la fórmula 16.1, la ΔH se obtiene de sumar las entalpías de los pares de di-nucleótidos que son vecinos más cercanos.

ΔS (kcal/mol) : S es la cantidad de desorden de un sistema, recibe el nombre de entropía. ΔS es el cambio en la entropía. Se obtiene sumando los valores de entropía de pares de di-nucleótidos que son vecinos más cercanos. Normalmente se adiciona una corrección a los parámetros de vecinos más cercanos. Esta corrección representa el contenido de sales.

ΔS (corrección por sales) : $\Delta S(1MNaCl) + 0.368N \ln([Na+])$, donde N es el número de pares de nucleótidos en el primers, y [Na+] son los equivalentes de sal en mM.

Temperatura de anillamiento de los primers: La T_M es un estimador de la estabilidad del híbrido ADN-ADN y es importante para poder estimar la temperatura de anillamiento (T_a). T_a muy altas harán que se formen pocos híbridos primer - molde resultando en una reducción del producto de PCR. T_a muy bajas podrán causar anillamientos no específicos. La siguiente ecuación permite estimar la T_a a partir de la T_m

$$T_a = 0.3T_m(primer) + 0.7T_m(product) - 14.9 \quad (16.2)$$

donde,

$T_m(primer)$ Es la temperatura de fusión de los primers

$T_m(product)$ Es la temperatura de fusión del producto

Contenido de GC: La proporción de G+C en el primer debe ser de 40% a 60%.

Gancho de GC: La presencia de las bases G o C en las últimas 5 bases del extremo 3' del primer (GC clamp) ayuda a tener una unión más específica en ese extremo debido a la unión más fuerte entre G y C. Sin embargo, se deben evitar más de 3 Gs o Cs consecutivos en las últimas 5 bases del extremo 3'.

Estructuras secundarias de los primers: La presencia de estructuras secundarias producidas por interacciones intra o intermoleculares puede llevar a una disminución en la producción del

amplímero o no producción de este. Esas estructuras disminuyen la cantidad de primer disponible para la reacción.

Evitar hidridación cruzada: Los primers diseñados para una secuencias no deben amplificar otro gen en la mezcla. La opción mas común es tomar los primers candidatos y compararlos contra bases de datos de genes usando una herramienta como BLAST.

16.1 Diseño de primers usando Quantprime

QUANTPRIME¹ es una herramienta flexible para el diseño de primers a mediana y gran escala (ARVIDSSON *et al.*, 2008), principalmente para PCR en tiempo real usando SYBR GREEN. QUANTPRIME usa primer3 (ROZEN and SKALETSKY, 2000) como motor para la creación de primers y agrega diversas capaz de verificación contra distintas bases de datos y anotación de genomas para proponer primers con mayor probabilidad de funcionar en ensayos experimentales.

Una de las principales ventajas de QUANTPRIME es que aprovecha la anotación de genoma y colecciones de EST que estén disponibles al público. Por ejemplo, la anotación de genomas puede ser explotada para producir primers que anillen sobre border de exones, disminuyendo considerablemente la probabilidad de amplificar ADN genómico en ensayos de evaluación de la expresión de genes.

Vaya a la página de QUANTPRIME, <http://www.quantprime.de/>. Con el fin de prestar un mejor servicio a los usuarios finales, es necesario registrarse y activar la cuenta siguiendo las instrucciones que llegarán a su correo electrónico luego de registrarse.

El primer paso en el flujo de trabajo de QUANTPRIME es crear un nuevo proyecto, encontrará un botón New project en el menú de la izquierda. La Figura 16.1, muestra el formulario de creación de proyectos. Allí tendrá que dar un nombre a su proyecto, este le servirá para almacenar su información en el servidor de QUANTPRIME y mantener varios proyectos en paralelo si así lo desea. A continuación tiene que seleccionar el organismo de interés y la versión de la anotación de su genoma o de disponibilidad de ESTs, según sea el caso. Para este ejercicio seleccione *Arabidopsis thaliana* como organismo y **TAIR release 9**, como versión de la anotación. La última sección corresponde a la selección del protocolo de cuantificación, i.e., usando SYBR-GREEN, en tiempo real, o al final de la PCR usando geles de agarosa (end-point PCR). Para cada una de esas opciones puede seleccionar si desea que los primers tenga hibridación cruzada con diferentes variantes de splicing o no. En este ejercicio vamos a usar la segunda opción.

El siguiente paso, consiste en incluir los transcritos para los cuales se desea diseñar primers. La Figura 16.2, muestra el formulario que nos permite completar este paso. Tiene dos opciones: i) si conoce los identificadores de los genes de interés, solo los tiene que poner en la caja de texto y presionar el botón Add to project, de lo contrario puede hacer búsqueda BLAST dentro de

¹Hay un tutorial disponible en el sitio de QUANTPRIME que describe con mayor detalle cada paso y opción.

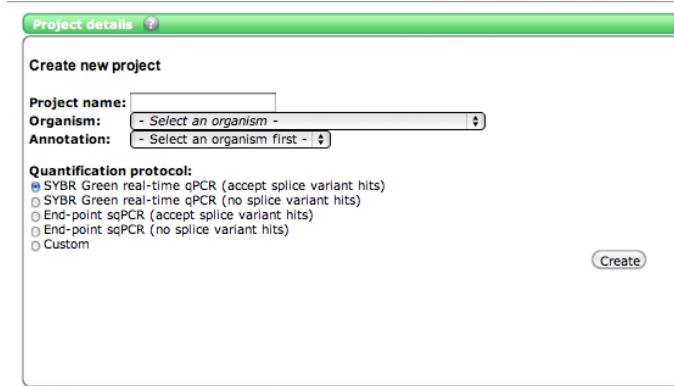


Figura 16.1: Creación un proyecto en QUANTPRIME

QUANTPRIME para encontrar los identificadores a partir de secuencias propias. En este caso vamos a diseñar primers para los genes: AT2G20825 y AT4G28190, que pertenecen a una familia pequeña de factores de transcripción conocida como ULT. Asegurese de adicionar esos identificadores a su proyecto y luego usar el botón Select all, seguido de find primers.

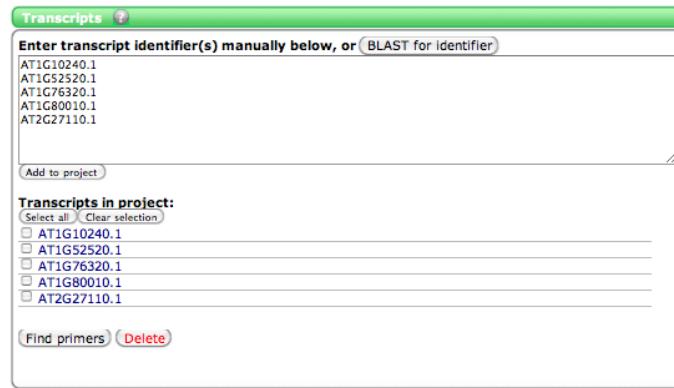


Figura 16.2: Adicionando transcritos al proyecto en QUANTPRIME

En este punto se iniciará el proceso de búsqueda de primers y su posterior verificación explotando la anotación del genoma de *Arabidopsis thaliana*. La Figura 16.3 muestra la ventana de progreso de la búsqueda. Si lo desea puede cerrar esta ventana y volver as tarde a recuperar sus resultados, esta es la ventaja de estar registrado en el sitio. En la Figura 16.3 se ve un indicador de progreso y una serie de cuatro casillas coloreadas por gen. La casilla de color verde oscuro indica el número de primers muy buenos que fueron encontrados, que cumplían con todos los criterios de búsqueda, i.e., específicos para el transcripto de interés, no amplifica ADN genómico, primers individuales no anillan con otros cDNAs. La casilla color verde claro indica el número de primers bueno, peor que podrían amplificar ADN genómico o alguno de los dos primers podría anillar con otro cDNA y por lo tanto reducir la eficiencia de la amplificación. En la casilla amarilla aparece el número de primers que se consideran adecuados, estos pueden amplificar ADN genómico, primers individuales pueden

anillar a otros cDNAs. La casilla roja indica el número de primers fallidos.

La casilla verde oscura solo va a estar desactivada en aquellos casos en que la especie de interés no tenga información de su genoma en la base de datos de QUANTPRIME.



Figura 16.3: QUANTPRIME buscando primers para los genes solicitados

Una vez la búsqueda ha terminado , presione el botón *List best primers* para obtener una lista detallada de los primers encontrados.

La lista de pares de primers está ordenada de acuerdo al color, como se explicó anteriormente, y en segundo lugar por el puntaje de rango de Primer3, la columna en el extremo derecho, el cual refleja la desviación de los criterios de diseño óptimo y el riesgo de formar estructuras secundarias y dímeros de primers. Los mejores primers son aquellos en que este número es más pequeño.

El botón *Select best* selecciona los mejores primers de los genes que se analizaron.

Favour primer pairs							
		Fw sequence	Rev sequence	Amplicon size	Spans exon border	Rank score	Test results
<input checked="" type="checkbox"/>	not amplifying genomic DNA (default)	TCTGGGGAGATGATTACCGTGAG	CAGGGGTCAATTGGTTTGGAG	149	Yes	3.7482	
<input type="checkbox"/>	with high single primer specificity (when genomic contamination is not a problem)	TCTGGGGAGATGATTACCGTGAG	TCTTCGGAGAAGGGAGGAGTTC	135	Yes	2.9346	
<input type="checkbox"/>	Select best	TCTTCGGAGAAGGGAGGAGTTC	TCTTCGGAGAAGGGAGGAGTTC	135	Yes	2.9346	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	ATATGCGGAGAGATTAATGTCAG	TCTTCGGAGAAGGGAGGAGTTC	146	Yes	2.6562	
<input type="checkbox"/>	<input type="checkbox"/>	ATCTTCAATTTGGGGAGCCTATCG	TCTTCGGAGAAGGGAGGAGTTC	136	Yes	2.6574	
<input type="checkbox"/>	<input type="checkbox"/>	CATTGGGGAGAACGATTAATGTCAG	TCTTCGGAGAAGGGAGGAGTTC	141	Yes	2.7032	
<input type="checkbox"/>	<input type="checkbox"/>	ACGATTAATGTTGGGGTTATGTTTG	TCTTCGGAGAAGGGAGGAGTTC	127	Yes	2.7039	
<input type="checkbox"/>	<input type="checkbox"/>	CGATTTACCTGGGGTTATVTTGTC	TCTTCGGAGAAGGGAGGAGTTC	126	Yes	2.8067	
<input type="checkbox"/>	<input type="checkbox"/>	GAAAGACAACTGGGGTTATVTTGTC	TCTTCGGAGAAGGGAGGAGTTC	126	Yes	2.8506	
<input type="checkbox"/>	<input type="checkbox"/>	GGATGAACTGGGGTTATVTTGTC	AGCTGGAACTTGGTTCTTCTATG	107	Yes	2.3025	
AT2G20825.1							
<input type="checkbox"/>	<input checked="" type="checkbox"/>	CACCTTGTTGTCGATGAGGAGAG	TGGCTTTTCTTCAGAAGATGC	68	Yes	4.6419	
<input type="checkbox"/>	<input type="checkbox"/>	TCATGGGGATGATGATGATGATGAT	TGGCTTTTCTTCAGAAGATGC	100	Yes+	2.8127	
<input type="checkbox"/>	<input type="checkbox"/>	ATAGGGGGATGGGGGGATGGGGGG	TGGCTTTTCTTCAGAAGATGC	100	Yes+	2.8195	
<input type="checkbox"/>	<input type="checkbox"/>	GGGATGGCTTGTGCTGAGTTGGG	TGGCTTTTCTTCAGAAGATGC	102	Yes+	4.2112	
<input type="checkbox"/>	<input type="checkbox"/>	GGGATGATGTTGTTGATGATTTAG	TGGCTTTTCTTCAGAAGATGC	102	Yes+	4.2112	
<input type="checkbox"/>	<input type="checkbox"/>	GATGATGATGTTGTTGATGATTTAG	TGGCTTTTCTTCAGAAGATGC	102	Yes+	4.2732	
<input type="checkbox"/>	<input type="checkbox"/>	GATACTGGGGATGTTGTTGATGAC	TGGCTTTTCTTCAGAAGATGC	107	Yes+	4.3257	
<input type="checkbox"/>	<input type="checkbox"/>	AGGTCATGTTGTTGTTGATCAGTC	TGGCTTTTCTTCAGAAGATGC	124	Yes+	5.2048	
<input type="checkbox"/>	<input type="checkbox"/>	CACACTTGGTTGGTTGAGGAGAG	TGGCTTTTCTTCAGAAGATGC	70	Yes+	3.8382	
<input type="checkbox"/>	<input type="checkbox"/>	CACACTTGGTTGGTTGAGGAGAG	ACTTTCTTCTTCTTCTTCTTCT	77	Yes+	4.2042	

Figura 16.4: Listado de los mejores primers encontrados por QUANTPRIME

Si desea ver información mas detallada sobre cada par de primers haga clic sobre el par de interés, esto lo conducirá a la página de información de ese par de primers en particular (Figura 16.5. En la parte superior de esta página encontrará información sobre el transcripto para el cual se diseñaron los primers.

Primer pair information

Transcript identifier: **AT2G20825.1** - ULT2, ULT2 (ULTRAPETALA 2); DNA binding, chr2:8965756-8966867 REVERSE

Forward primer
Sequence: TGTGGGAGACGATTAGTCGAG (22 b)
Melting temperature: 62.2 °C
G/C content: 54.5 %

Reverse primer
Sequence: CAGGCGTCAACTTGTCTTCGAG (reverse complement: CTCGAAGACAAGTTGACGCCCTG) (22 b)
Melting temperature: 62.1 °C
G/C content: 54.5 %

Amplicon
Size: 149 b
Melting temperature: 86.4 °C
G/C content: 52.3 %
Optimal annealing temperature: 64.2 °C

Alignment with transcript sequence
GATCCTATATAATATACCTCATAGGAAGACAAAGGAGCCCTTCGTCGTGGTTTACTTGCGCCGATCGAGATGGAGAG
AGAATGTGGTGTGAAAGGAGTTGTTAGCAGCAGAACAGGAGACTACAAAGAAATAAAGGGAGTCATGTGGAGAGCATTAGTCG
AGTCGATGTGGCTGACACAGCCACCTTACGGAGACCCGTTAGGAGGCTTAAGAATTTTCAGATGGAGAACTTCAA
ATCACCTGCCAATGCACTCTGTTCTGAG^AGACAGACTCTGCGCTGCTGCCTTCGAGAAAGCATTCAGAGAGAAAA
CCTCTAGAAACTGGAGAACAAAGTTGGTCTTATTGAAGGAGACAAGGTTCCGCTTCAAAAGACAGTGTGCTCAGA
TACTACAAACAAGCATTTGAGAAACTTAACGATTCATTGAAAGACTCATTCATCGGGAGGAGTTGGGGGCAGGACATGGG
GAAGGAGAGGAGTTCAAGATTGAGGAGCAGAGGGGAATGCCGGAGGCACCATGATGCAATTGCTGAGCCTAATTGGAAAGT
GTTGCGAATACCCATACGACAAAGAAACATGGCAGGGAGGAGGAAAGAAGGGAGCAGGAAACTGTTCAAGGGTTGCACT
CGCTCACCGCTCTGCAGGCTGCACTTCTTCGCTGGGCTGCAAGGCTTCTGCCTTCTGATTGTAACCTGCCA
GACTTGCTGATTTCACCAACCAAGTGCCTAAACCCATTGAACTTCAATTCTTAACTTAACTTAACTTCTGCAAAAGACTT
GTAATGTAATGCTCTGCAATTCTTAACTGAGCTTAGCTTACACCTGTT

Specificity test results
Overall score: **Excellent**
cDNA specificity: **Good**
Single primer specificity: **Good**
Amplifies genomic DNA: **No**

Figura 16.5: Página de información para un par de primers seleccionados

En la Figura 16.5, el amplicón aparece marcado con fondo gris, primers que anillan en límites entre exones aparecen en color verde, y el límite entre los exones se indica con el símbolo ^, los primers que aparecen en color azul no anillan sobre límites entre exones. En esta página también encuentra la T_m de cada primer, del amplicón y la temperatura óptima de anillamiento, así como otras características de los primers. Al final de la página encuentra información sobre los resultados de las pruebas de especificidad que se llevaron a cabo.

Revise la página de resultados del par de primer para un par bueno o muy bueno y para un para adecuado o malo, identifique las diferencias.

Vuelva a la página de resultados. En la parte inferior de la página encontrará el botón Export primer pairs, que le permite enviar los pares de primers seleccionados a un archivo de texto.

16.2 Crear primers a partir de alineamientos de proteínas

En esta sección vamos a usar el programa iCODEHOP² para diseñar primers a partir de un alineamiento de proteínas.

²[86](https://icodehop.cphi.washington.edu/i-codehop-context>Welcome</p>
</div>
<div data-bbox=)

Vamos a usar el alineamiento de las 22 secuencias de primates que usó hace alguans semanas. Asegurese que el alineamiento está en formato fasta o clustal.

Siga el enlace <https://icodehop.cphi.washington.edu/i-codehop-context>Welcome> que lo llevará al sitio web de iCODEHOP.

Inicie una sesión, esto lo llevará a una nueva página que luce como aparece en la Figura 16.6, seleccione la opción Design Primers

The screenshot shows a software interface titled "Analysis 1". At the top, there is a menu bar with "Print", "Quit", "New", "Delete", and "More...". Below the menu, a message reads "Please select from the following options:". There are three main buttons:

- Design Primers**: Description: "Design degenerate primers using the CODEHOP strategy. Clicking this button places you in the primer design workflow. You will be taken to a page where you can upload aligned and/or non-aligned sequences. iCODEHOP will know how to process your data so that it can design CODEHOPS. (?)".
- Clustal**: Description: "Use CLUSTAL to create a multiple alignment of your sequences. Clicking this button places you in the sequence multiple-alignment workflow. If you just want to design primers choose the 'Design Primers' button above. (?)".
- Predict T_m**: Description: "Clicking this button starts a program that will calculate the range of melting temperatures predicted one or more degenerate primer pools. (?)".

Figura 16.6: Página de inicio en iCODEHOP

The screenshot shows the "Analysis 1" page for primer design. At the top, there is a menu bar with "Print", "Quit", "New", "Delete", and "More...". Below the menu, a section titled "Upload and select sequences for designing CODEHOPs" contains the following instructions:

Upload sequences using the buttons at bottom of this page then make a selection (see [limitations on sequence selection](#)). When you are done, click the following button to proceed:

Proceed with Analysis

Below this, there is a note: "To upload your new sequences, first select which format they are in then, fill in and submit the form that is triggered by your selection: (?)".

Non-aligned amino acid sequences:

- Select NCBI accession number(s) [\(?\)](#)
- Select NCBI URL(s) [\(?\)](#)
- Select UniProt accession number(s) [\(?\)](#)
- Select Amino acid sequences stored in FASTA format [\(?\)](#)
- Select An amino acid sequence stored in a GENBANK formatted [\(?\)](#)

Aligned amino acid sequences:

- Select A CLUSTAL or FASTA formatted alignment [\(?\)](#)
- Select Ungapped multiple alignments in Blocks format [\(?\)](#)

Pre-loaded data for exploring iCODEHOP

- Select *E. coli* Non-aligned amino acid sequence data

Figura 16.7: Diseño de primer en iCODEHOP

En la página de diseño de primers puede seleccionar diferentes fuentes de datos de alineamiento de proteínas (Figura 16.7). En este ejercicio haga clic en el botón **Select** que se encuentra en frente de **A CLUSTAL or FASTA formated alignment**, seleccione el archivo de alineamiento de 22 secuencias de primates. La nueva página aparece como la Figura 16.8. Ahora puede proceder con el análisis haciendo clic en el botón **Proceed with Analysis**.

Sequence Groups	Description
↳ Gapped multiple-alignments	
↳ test	
✓ PMarmoset	MASRLVNIKEEVTCPLICLELLTEPLSLDCGHSGFCQACIT...
✓ AGM	MASGILLNVKEEVTCPLICLELLTEPLSLPCGHSGFCQACIT...
✓ Saki	MASRILMNKEEVTCPLICLELLTEPLSLDCGHSGFCQACIT...
✓ Gibbon	MASGILNVKEEVTCPLICLELLTQPLSLDCGHSGFCQACLT...
✓ Chimp	MASGILVNKEEVTCPLICLELLTQPLSLDCGHSGFCQACLT...
✓ Titi	MASRILVNKEEVTCPLICLELLTEPLSLDCGHSGFCQACIT...
✓ Baboon	MASGILLNVKEEVTCPLICLELLTEPLSLPCGHSGFCQACIT...
✓ Squirrel	MASRILGSIKEEVTCPLICLELLTEPLSLDCGHSGFCQACIT...
✓ Colobus	MASGILVNKEEVTCPLICLELLTEPLSLHGCGHSGFCQACIT...

To upload your new sequences, first select which format they are in then, fill in and submit the form that is triggered by your selection: [\(?\)](#)

Non-aligned amino acid sequences:

- Select** NCBI accession number(s) [\(?\)](#)
- Select** NCBI URL(s) [\(?\)](#)
- Select** UniProt accession number(s) [\(?\)](#)
- Select** Amino acid sequences stored in FASTA format [\(?\)](#)
- Select** An amino acid sequence stored in a GENBANK formatted [\(?\)](#)

Figura 16.8: Diseño de primer en iCODEHOP

El siguiente paso en el algoritmo CODEHOP es determinar los BLOCKS³, esto es hecho automáticamente por iCODEHOP (figura 16.9). En la siguiente página selecciones el código genético y la tabal de uso de codones que serán usada en el diseño de primers. Hay otros parámetros que puede variar antes de iniciar la búsqueda de primers. ¿Qué controla cada uno de esos parámetros?

Uno vez esté satisfech@ con su selección de parámetros, puede dar clic en el botón **Look for primers** para iniciar la búsqueda de primers en los BLOCKS detectados. Sea paciente la búsqueda de primers puede tomar bastante tiempo. Al finalizar la búsqueda los resultados se mostrarán en forma gráfica como aparece en la Figura 16.10, al hacer clic sobre los primers, encontrará información detallada.

Cada uno de los rectángulos que aparece en la imagen representa los BLOCKS originales, i.e., alineamientos múltiples sin gaps. El nombre del BLOCK aparece en la esquina superior izquierda del rectángulo.

³¿Qué son y como se determinan los BLOCKS?.

The screenshot shows the CODEHOP software interface with the following details:

- Analysis 1** tab is selected.
- Top menu bar includes Print, Quit, New, Delete, More... buttons.
- Middle section displays the message: "CODEHOP will suggest PCR primers based protein multiple sequence alignments".
- Block(s):** A list of blocks is shown, with the first one highlighted:
 - ID: x1982tu; BLOCK
 - AC: x1982tu; distance from previous block=0.0
 - DE: /home/bjorgar/i-codehop-context/tmp/x1982tu/test.aln
 - SL: UK motif: width=46; segs=22; 99.5%0; strength=0
 - Human: M~~A~~S~~G~~L~~V~~W~~K~~E~~V~~C~~T~~O~~L~~U~~T~~P~~L~~S~~D~~G~~H~~E~~F~~C~~O~~A~~G~~T~~A~~H~~M~~S 2.001157
 - Chimp: ()
 - Gorilla: ()
 - Orangutan: ()
- Genetic code:** Standard dropdown menu.
- Codon usage table:** Shows choices for Helicoverpa armigera, Heliothis virescens, Heterodera glycines, Hirudo medicinalis, and Homo sapiens.
- Advanced Settings:**
 - Re-weight the sequences that were used to create blocks: Alter sequence weights.
 - Core** (degenerate 3' region) settings: degeneracy [default=128]: 128, strictness [default=0.0]: 0.0.
 - Clamp** (non-degenerate 5' region) settings: temperature [default=60.0]: 60.0, poly-nuc [default=5]: 5.
 - Primer concentration [in nM, default=50nM] (K+)=50mM: 50.
 - Show the 3 least degenerate primers:
 - Show all overlapping primers:

Figura 16.9: Detección de BLOCKS en el alineamiento de secuencias de proteínas. Se diseñaran primers para cada BLOCK

Debajo del nombre del BLOCK encontrará una fila con información sobre el número de amino ácidos que constituyen el BLOCK y la distancia en amino ácidos al BLOCK anterior y al siguiente (esto último en paréntesis).

Enseguida encuentra el rectángulo que representa el BLOCK, aparece la secuencia consenso del alineamiento múltiple. El símbolo * aparece encima de los residuos completamente conservados. Amino ácidos en mayúscula representan sitios altamente conservados mientras que aquellos en minúscula representan sitios con un bajo nivel de conservación.

Debajo del rectángulo encuetrara los primers degenerados representados por flechas. Las flechas que se dirigen a la derecha, corresponden a los primer **forward**, las que se dirigen a la izquierda corresponden a los primers **reverse**. si una flecha es roja significa que iCODEHOP no pudo extender la región consenso del gancho en su longitud completa. Esto pasa cuando hay poco conservación en el extremo 5' de la región CORE degenerada de un primer.

Puede seleccionar un primer particular haciendo click sobre la flecha que lo representa y obtener información adicional usando el botón **Complete summary** en la parte superior de la página.

En la página **Compete summary** encuentra información detallada sobre el BLOCK que se usó para diseñar el primer seleccionado, así como ss temperaturas de anillamiento. Mas abajo encuentra una tabla con todos los primer potenciales para usar como compañeros del primer seleccionado, cada uno con infomración del nombre del BLOCK que se usó para su diseño, y sus temperaturas de anillamiento.

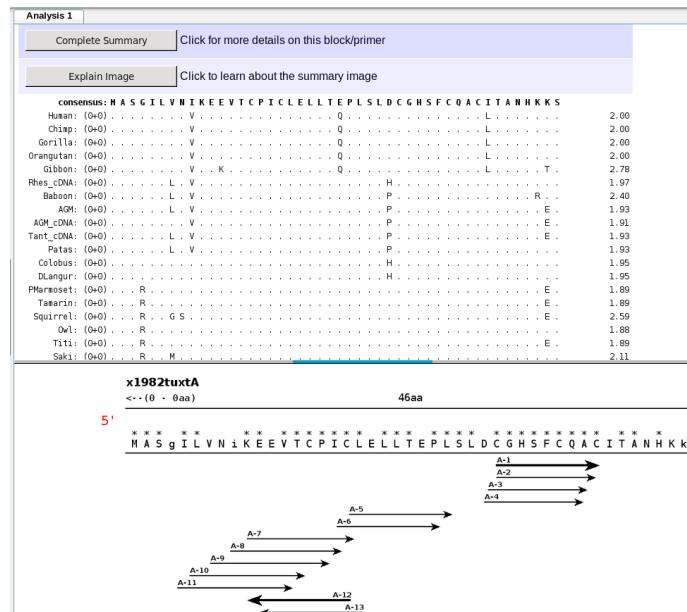


Figura 16.10: Detección de BLOCKS en el alineamiento de secuencias de proteínas. Se diseñarán primers para cada BLOCK

Capítulo 17

Montagem de genomas

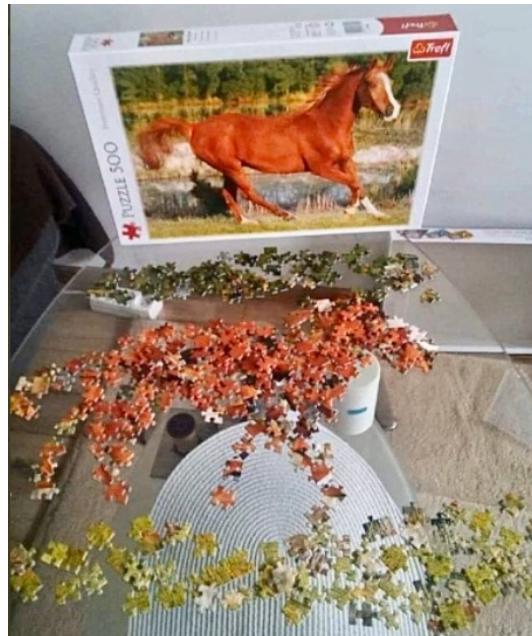


Figura 17.1: Montagem de genomas

O processo de montagem de genomas consiste na recuperação da sequência original a partir dos fragmentos de ADN produzidos pelo sequenciamento

Vamos montar o genoma de *Komagataeibacter rhaeticus*¹ a partir de leituras produzidas por duas tecnologias Illumina e PacBio que encontrara na pasta PRATICA_ENSAMBLAGEM_DE_GENOMAS onde teria que descomprimir um arquivo que contem as leituras.

1 _____ indo na pasta de trabalho _____
(base) user@server:\$ cd PRATICA_ENSAMBLAGEM_DE_GENOMAS

¹NCBI accession: NZ_CP050139

17.0.1 Limpar sequencias

As tecnologias de segunda geração como Illumina produzem leituras que devem ser filtradas por critérios de qualidade de leitura, cumprimento, presença de adaptadores, barcodes, contaminantes e artefatos. Por enquanto os sequenciadores de tecnologias de terceira geração como PacBio geralmente fazem esse processo por nós

Temos dois arquivos fastq com leituras pareadas de illumina²³

- O Que são leituras pareadas?
- Explique a organização do formato FASTQ

Limpamos eles usando BBduk⁴, BBduk pode fazer quality-trimming, filtrado, adapter-trimming, contaminant-filtering usando kmer matching.

```
1   _____ Limpando leituras Illumina com bbdruk _____
2   (base)user@server:$ conda activate bbdruk_env
3   (bbduk_env)user@server:$ bbdruk.sh in1=illumina_R1.fq in2=illumina_R2.fq out1=./bbduk/bbdruk.R1.fq \
4   out2=./bbduk/bbdruk.R2.fq minlength=75 qtrim=w trimq=20
4   (bbduk_env)user@server:$ conda deactivate
```

- Qual é o significado das opções qtrim=w e trimq=20

bbduk vai gerar dois arquivos filtrados bbdruk.R1.fq e bbdruk.R2.fq na pasta ./bbduk que usaremos na montagem do genoma

Agora vamos usar o programa FASTQC⁵ para visualizar o efeito da filtragem do bbdruk nas nossas leituras illumina

```
1   _____ Abrindo o FASTQC _____
2   (base)user@server:$ conda activate fastqc_env
2   (fastqc_env)user@server:$ fastqc
3   (fastqc_env)user@server:$ conda deactivate
```

- Carregar as leituras de Illumina antes e depois de filtrar e explique as diferenças no item "Per base quality" do reporte do fastqc

17.0.2 Montagem de genoma usando dados Illumina

Usaremos o software Unicycler⁶ um pipeline de varios softwares para ensamblagem de genomas bacterianos. Vamos ensamblar o genoma usando os arquivos arquivos bbdruk.R1.fq e bbdruk.R2.fq filtrados que gerou o bbdruk

²illumina_R1.fq

³illumina_R2.fq

⁴<https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbdruk-guide/>

⁵<https://github.com/s-andrews/FastQC>

⁶<https://github.com/rrwick/Unicycler>

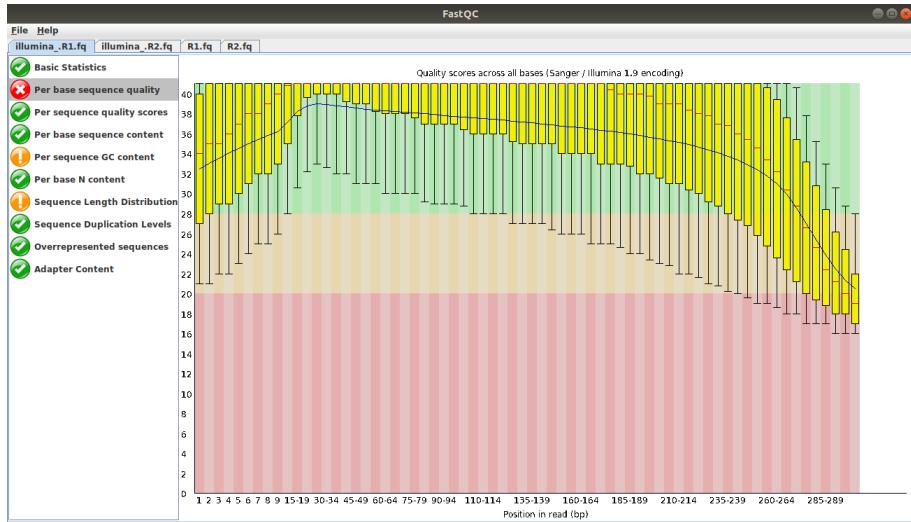


Figura 17.2: Tela FASTQC

```
Montando leituras com Unicycler
1 (base)user@server:$ conda activate unicycler_env
2 (unicycler_env)user@server:$ unicycler -1 ./bbduk/bbduk.R1.fq -2 ./bbduk/bbduk.R2.fq -o ./unicycler/
3 (unicycler_env)user@server:$ conda deactivate
```

Agora pode dar uma olhada nos arquivos gerados pelo Unicycler na pasta ./unicycler e usar o programa Bandage⁷⁸ para visualizar o arquivo assembly.gfa

```
Abrindo o Bandage
1 (base)user@server:$ conda activate bandage_env
2 (bandage_env)user@server:$ Bandage
3 (bandage_env)user@server:$ conda deactivate
```

- De que maneira estão representadas as regiões repetitivas do genoma na ensamblagem do unicycler?

17.0.3 Montagem de genoma usando dados PacBio

Usaremos o software Flye⁹. Flye é uma montadora de novo para leituras de sequenciamento de moléculas únicas, como as produzidas pela PacBio e Oxford Nanopore Technologies.

Usaremos o arquivo com leituras PacBio PacBio.fq

```
Montando leituras com Flye
1 (base)user@server:$ conda activate flye
2 (flye_env)user@server:$ flye --pacbio-hifi PacBio.fq -o ./flye
3 (flye_env)user@server:$ conda deactivate
```

⁷<https://rrwick.github.io/Bandage/>

⁸<https://pubmed.ncbi.nlm.nih.gov/26099265/>

⁹<https://github.com/fenderglass/Flye/>

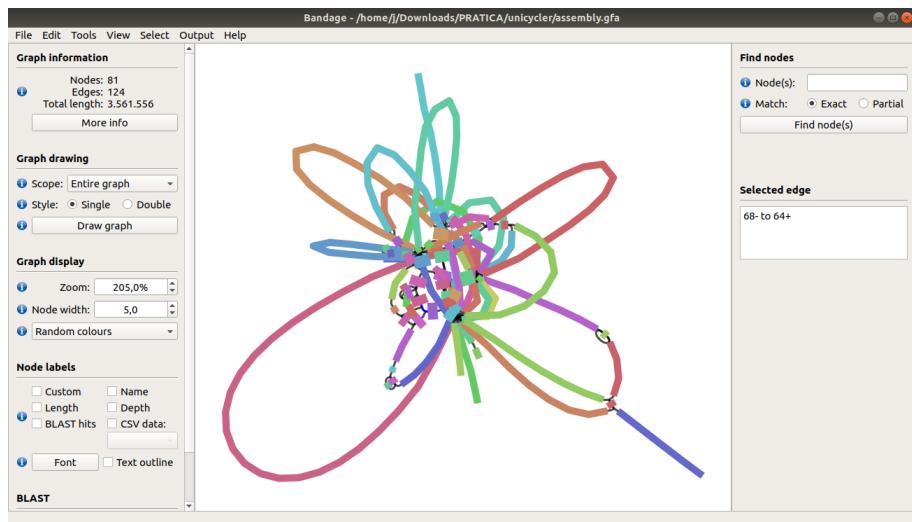


Figura 17.3: Montagem Illumina

Agora poder dar uma olhada nos aquivos gerados pelo Flye na pasta ./flye e usar o programa Bandage para visualizar o arquivo assembly_graph.gfa

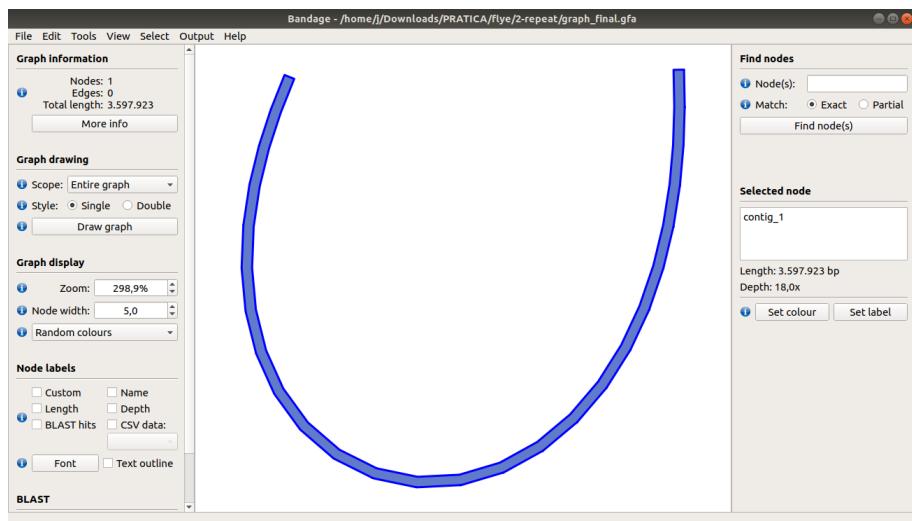


Figura 17.4: Montagem PacBio

- Explique porque as diferenças nas gráficas geradas pelas duas ensamblagens: Illumina e PacBio

17.0.4 Avaliando a qualidade das montagens com QUAST

Agora vamos utilizar QUAST¹⁰ para comparar a qualidade das nossas montagens. No primeiro lugar a montagem de Illumina (./unicycler/assembly.fasta) comparado com a montagem do genoma de referencia que se encontra em complete_genome.fasta.

```
1 (base)user@server:$ conda activate quast_env
2 (quast_env)user@server:$ quast -o ./quast/illumina -r complete_genome.fasta ./unicycler/assembly.fasta
3 (quast_env)user@server:$ conda deactivate
```

Olha os arquivos gerados pelo quast na pasta ./quast/illumina e analise o arquivo report.html. O quast também pode funcionar sem genoma de referencia no caso que seja necessário.

- Rodar o QUAST sem o genoma de referencia para os dados de Illumina. Compare os resultados da tabela no arquivo report.html e explique as diferenças (lembre que a melhor forma de facilitar as coisas é organizar os arquivos e pastas de uma maneira logica¹¹ e por exemplo ter uma pasta para cada analise)
- Repetir o processo com os dados PacBio utilizando o arquivo assembly.fasta que se encontra na pasta ./flye e compare os resultados de Illumina e PacBio usando a referencia explique as diferenças no report.html

17.0.5 Avaliando a qualidade das montagens com BUSCO

O BUSCO¹² informa da quantidade de genes ortologos de copia única que espera-se encontrar na nossa ensamblagem a comparação com o que se encontra em organismos do mesmo grupo taxonômico. Vamos avaliar a montagem de Illumina ./unicycler/assembly.fasta

```
1 (base)user@server:$ conda activate busco_env
2 (busco_env)user@server:$ busco -i ./unicycler/assembly.fasta -o illumina -m geno \
3 --lineage rhodospirillales_odb10 --out_path ./busco
4 (busco_env)user@server:$ conda deactivate
```

Olha os arquivos gerados pelo BUSCO na pasta ./busco/illumina

- Que significa a opção –lineage no comando do BUSCO
- Repetir o processo de avaliação de qualidade da montagem com BUSCO utilizando os dados PacBio no arquivo gerado pelo flye assembly.fasta que se encontra na pasta ./flye. Compare os resultados e explique as diferenças entre as montagens de leituras illumina e leituras PacBio segundo BUSCO
- Segundo os resultados do BANDAGE, BUSCO e QUAST qual é a melhor montagem e porque

¹⁰<http://quast.sourceforge.net/>

¹¹<https://zapier.com/blog/organize-files-folders/>

¹²<https://github.com/robsyme/busco>

Capítulo 18

Anotação de Genomas

Depois de ter montado as suas leituras na ensamblagem do genoma, é útil saber quais são as características genômicas dessa ensamblagem. O processo de identificar e rotular essas características é chamado de anotação do genoma.

Neste capítulo vamos anotar a montagem do genoma montado com leituras PacBio feito no capítulo 17

18.0.1 Usando prokka para anotar genomas bacterianos

Prokka¹², é uma ferramenta de software de linha de comando para anotação de genomas bacterianos

No primeiro lugar copiamos a ensamblagem feita no capítulo 17 correspondente aos dados de PacBio, esta montagem que você fez, deve-se encontrar em

~ /PRATICA_ENSAMBLAGEM_DE_GENOMAS/flye/assembly.fasta

A nossa pasta de trabalho de hoje é

~ /GENOME_ANNOTATION

Copiamos a ensamblagem na nossa pasta de trabalho

1 Copiando a ensamblagem para sua anotação _____
(base)user@server:\$ cp ~ /PRATICA_ENSAMBLAGEM_DE_GENOMAS/flye/assembly.fasta ~ /GENOME_ANNOTATION

Nos dirigimos na nossa pasta de trabalho

1 Indo na pasta de trabalho _____
(base)user@server:\$ cd ~ /GENOME_ANNOTATION

Agora ativamos o ambiente virtual que contém o prokka e rodamos ele no arquivo assembly.fasta

¹<https://pubmed.ncbi.nlm.nih.gov/24642063/>

²<https://github.com/tseemann/prokka>

```
1 (base)user@server:$ conda activate prokka
2 (prokka)user@server:$ prokka --outdir prokka --prefix assembly_pacbio assembly.fasta --force
3 (prokka)user@server:$ conda deactivate
```

O prokka gera varios arquivos de saída na pasta

~ /GENOME_ANNOTATION/prokka

Da uma olhada neles.

18.0.2 Usando IGV para olhar os resultados de prokka

. Agora podemos abrir o IGV para vizualizar as anotações no genoma

```
1 (base)user@server:$ igv.sh
```

Carregamos no IGV³ o arquivo

~ /GENOME_ANNOTATION/assembly.fasta

e o arquivo

~ /GENOME_ANNOTATION/prokka/assembly_pacbio.gff

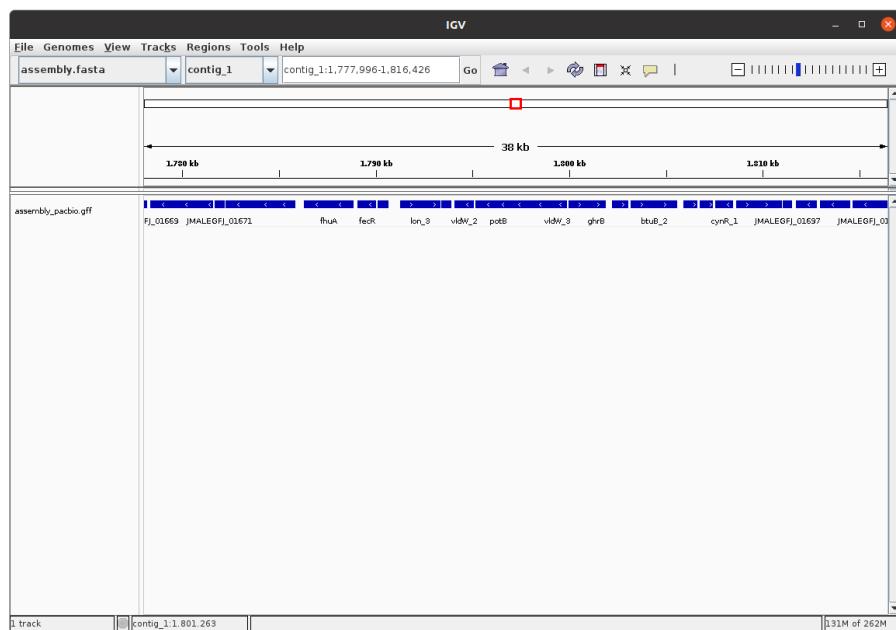


Figura 18.1: Tela IGV

³<https://igv.org/>

18.0.3 Rondando blast local

Agora vamos rodar um blast⁴ customizado para olhar coincidencias nas proteinas geradas pelos CDS detetados pelo prokka na nossa ensamblagem e numa base de dados que vamos gerar com sequencias conhecidas do género *Komagataeibacter* ao qual pertence a nossa bactéria.

o arquivo com as sequencias para construir a nossa base de dados encontra-se em
`~ /GENOME_ANNOTATION/sequences_for_db.fa`

```
1 _____ Criando base de dados customizada para o blast _____
2 (base)user@server:$ mkdir -p ~/GENOME_ANNOTATION/blastDB/
3 (base)user@server:$ export BLASTDB=~/GENOME_ANNOTATION/blastDB/
4 (base)user@server:$ cd ~/GENOME_ANNOTATION/blastDB/
5 (base)user@server:$ makeblastdb -in ../../sequences_for_db.fa -dbtype 'prot' -out myDB
6 (base)user@server:$ cd ..
```

- Na linha 1 criamos uma pasta chamada blastDB na nossa pasta de trabalho
- Na linha 2 fazemos saber ao BLAST que as bases de dados estão na pasta que acabamos de criar
- Na linha 3 nos dirigimos na pasta que acabamos de criar
- Na linha 4 criamos uma base de dados customizada com o nome myDB
- Na linha 5 voltamos na nossa pasta de trabalho

Agora estamos prontos para rodar o BLAST localmente na nossa base de dados!

```
1 _____ Rondando blast local _____
2 (base)user@server:$ mkdir -p blast
3 (base)user@server:$ blastp -db myDB -query ./prokka/assembly_pacbio.faa -out ./blast/out_blast.txt
```

- Na lina 1 criamos uma pasta chamada blast na nossa pasta de trabalho que vai conter a saída do BLAST
- Na linha 2 rodamos o BLAST das sequencias das proteínas geradas pelos CDS encontrados pelo prokka na nossa ensamblagem, na nossa base de dados com as proteínas do género *Komagataeibacter*

O BLAST vai gerar um arquivo de saida na em:

`~ /GENOME_ANNOTATION/blast/out_blast.txt`

o arquivo é o suficientemente grande para matar sua maquina se abre ele tudo de só uma vez da uma olhada nele usando o comando less

```
1 _____ Resultados do BLAST local _____
(base)user@server:$ less ./blast/out_blast.txt
```

⁴<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Bibliografia

- ARVIDSSON, S., M. KWASNIEWSKI, D. M. RIAÑO PACHÓN, and B. MUELLER-ROEBER, 2008 Quantprime—a flexible tool for reliable high-throughput primer design for quantitative pcr. *BMC Bioinformatics* **9**: 465.
- BOURNE, P. E., 2004 The future of bioinformatics. In *2nd Asia-Pacific Bioinformatics Conference (APBC2004)*.
- EDGAR, R. C., 2004 Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.
- FINN, R. D., J. MISTRY, J. TATE, P. COGGILL, A. HEGER, *et al.*, 2010 The pfam protein families database. *Nucleic Acids Res* **38**: D211–D222.
- KYTE, J., and R. F. DOOLITTLE, 1982 A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**: 105–132.
- LEMEY, P., M. SALEMI, and A.-M. VANDAMME, editors, 2009 *The Phylogenetic Handbook*. Cambridge University Press.
- LEONARD, S. A., T. G. LITTLEJOHN, and A. D. BAXEVANIS, 2007 Common file formats. *Curr Protoc Bioinformatics Appendix 1: Appendix 1B*.
- MIZRACHI, I. K., 2008 Managing sequence data. *Methods Mol Biol* **452**: 3–27.
- MULLIS, K. B., and F. A. FALOONA, 1987 Specific synthesis of dna in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **155**: 335–350.
- NEEDLEMAN, S. B., and C. D. WUNSCH, 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.
- NOTREDAME, C., 2007 Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol* **3**: e123.
- ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the www for general users and for biologist programmers. *Methods Mol Biol* **132**: 365–386.

SAWYER, S. L., L. I. WU, M. EMERMAN, and H. S. MALIK, 2005 Positive selection of primate trim δ alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A **102**: 2832–2837.

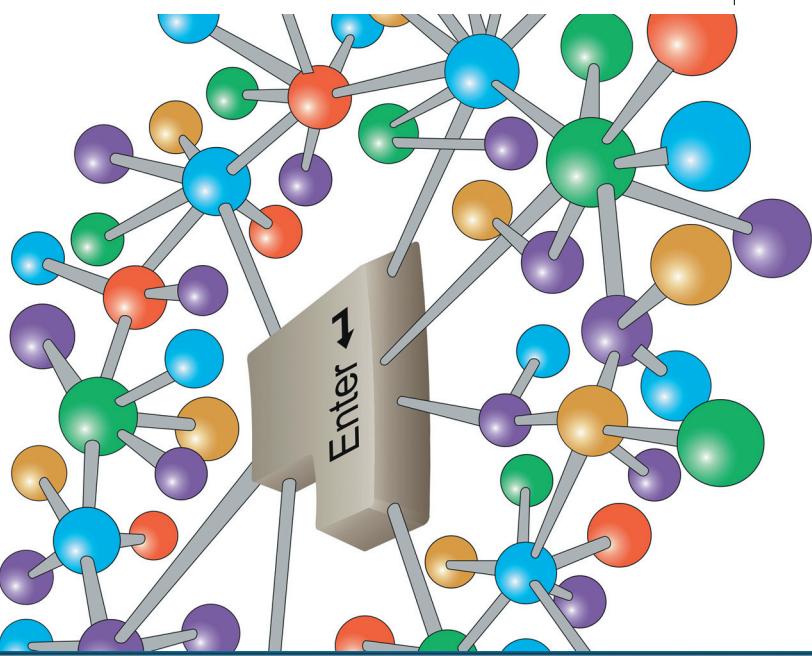
SCHNEIDER, T. D., and R. STEPHENS, 1990 Sequence logos: a new way to display consensus sequences. Nucleic Acids Res **18**: 6097–100.

SMITH, T. F., and M. S. WATERMAN, 1981 Identification of common molecular subsequences. J Mol Biol **147**: 195–197.

STEIN, L. D., 2008 Bioinformatics: alive and kicking. Genome Biol **9**: 114.

WILSON, D., R. PETHICA, Y. ZHOU, C. TALBOT, C. VOGEL, *et al.*, 2009 Superfamily—sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res **37**: D380–D386.

Appendices



EMBnet

A Quick Guide UNIX

PROCESSES	
<code>^c <ctrl>-c</code>	kills (definitely stops) current job
<code>^z <ctrl>-z</code>	suspends the current job. This can either be moved to the background or resumed in the foreground by using bg or fg
bg	moves the current process to the background
fg	moves a process to the foreground. (If there is more than one suspended job, use jobs to decide which you want to fg)
fg 2	moves process number 2, as listed by jobs , to the foreground
jobs	lists background and suspended processes (created with bg or <code>^z</code>)
jobs -1 ("el" not one)	includes the pid (process id number)
ps	lists all your processes
kill	stops a process (use ps or jobs to find your processes)
kill 2986	kills off the process with pid 2986
MISCELLANEOUS	
finger	tells you who is logged on (see also w)
w	shows information about logged in users
who	produces similar result (see finger)
tar	create (or extract) a tarball from (to) a list of files
tar -cvf tarball.tar subdir/*	
tar -xvf tarball.tar	the option -z compresses the files by gzip
wc	word count
wc long.file	prints the number of lines, words and characters in <i>longfile</i> . Options include -l to count lines only, and -c to count characters only
In	create a link or an alias for a file
In -s subdir/origfile alias.file	
history	displays last several commands used
!!	re-executes the last command
!51	executes command 51 in the history list use also <code><up></code> - and <code><down></code> - arrows to navigate in the history
date	displays current date and time
passwd	invokes a password changing program
exit	leaves the current shell (same as <code>^d</code> or <code><ctrl>-d</code>) usually = logout
GRAPHIC DISPLAY	
	To display graphics, most Unix require the configuration of the X-Window server.
	Commands on your local computer:
xhost +	set the list of allowed X-Window clients
	The "+" allows any remote computer to display on your local display
ifconfig	gives information about the network configuration (e.g., the current IP_address, usually similar to 123.145.167.189)
	Commands on the remote computer:
setenv	set up an environment variable (tc-shell)
setenv DISPLAY IP_address:0.0	required to tell the remote computer where it should display its graphics
xclock	starts a graphic clock (e.g., used to test the X-Window server or to get the current time... ;-)
	This document was originally written and designed by Aoife McLyagh and Andrew Lloyd© from the Irish EMBnet node, and modified by Laurent Falquet from the Swiss EMBnet node and distributed by the Publications Committee of EMBnet.
	EMBnet - European Molecular Biology network - is a network of bioinformatics support centres situated primarily in Europe. Most countries have a national node which can provide training courses and other forms of help for users of bioinformatics software.
	Further information about UNIX is available from your national node. You can find contact information about your national node from the EMBnet web site:
	http://www.embnet.org/
	If you have found this publication useful, please let us know. If you have ideas for similar documents we'd like to hear from you: emb-pi@embnet.org
	A Quick Guide To UNIX Revised edition 2003

A Quick Guide To UNIX

This is an introduction to the UNIX operating system. Unix may seem idiosyncratic, even impenetrable, to begin with but it has the virtue of minimising the number of keystrokes and so speeding up your access to the computer.

The commands listed here are common to different operating systems and shells. They include some of the most useful and frequently used commands in UNIX. The power and utility of most UNIX commands can be enhanced with switches or options preceded by a “-” sign.

More information on the options, the effects and how to use the commands is available by using the **man** command:

man gives manual information on a topic

man grep

displays the manual page about grep

apropos lists all the man(ual) entries relating to a topic
(same as **man -k**)

apropos print

Another useful source of information is the on-line EMBnet tutorial which includes a page on UNIX

<http://www.dk.embnet.org/Embnnet/Universl/unixcmds.html>

or equally

<http://www.uk.embnet.org/Embnnet/Universl/unixcmds.html>

The general format of this document is that anything in **bold** is a command you can enter. Anything in **italic** is a file or directory name you must change according to yours. Anything preceded by a hyphen “-” is an option which will modify the effects of a command. A general description of each command is followed by one or several examples of its use.

chmod modifies the read (**r**), write and delete (**w**), and execute (**x**) permissions of specified files and the search permissions of specified directories. The permission can be set for user (**u**), group (**g**) or other (**o**)

chmod go-w my.file
stops (-) anyone else (go) changing or deleting (w)
my.file

chmod g+rwx my.file
allows (+) anyone of my group (g) reading, changing, deleting or executing (**rwx**) *my/file*

cp copies files
cp orig.file copy.file
cp orig.file subdir/new.file
copies *orig/file* to *new/file* in *subdir* directory

cp subdir/orig.file .
copies *orig/file* from *subdir* to the current directory
(.) without changing its name

mv moves/renames a file (or directory)
mv oldname newname
mv my.file subdir/my.file
a move (mv) is equivalent to a copy (cp) followed by a remove (rm)

rm removes/deletes a file.
rm oldfile
rm -i *.*file
option -i (interactive) advised if wildcards (*) in use

diff compares two files and prints how they differ
diff file1 file2
prints differences to screen options include -b to ignore differences in blank space, and -i to ignore case

find searches the directory tree for a file
find . -name lostfile -print
will search your current directory (.) (and any subdirectories) for *lostfile*

grep searches a file for a string
grep word my.file
grep "two words" my.file
options include -i to ignore case and -n to print line numbers

FILES

ls lists files in a directory
ls -aF lists -a all files in -1 long format -F identifies directories /, executable files *, and symbolic links @, in the current directory

cat concatenates and displays files
cat my.file displays *my/file* on the screen

pico simple screen oriented text editor
pico myfile.txt

head prints the first few (default = 10) lines of a file
head oddfile
head -20 oddfile
displays first twenty lines of *oddfile*

tail displays last few lines of a file (see head)
more displays a file one screenful at a time
more longfile
hit <**Spacebar**> to see the next screen
Note: some people prefer **less**

vi simple screen oriented text editor

pico simple display oriented text editor
pico myfile.txt

head prints the first few (default = 10) lines of a file
head oddfile
head -20 oddfile
displays first twenty lines of *oddfile*

tail displays last few lines of a file (see head)
more displays a file one screenful at a time
more longfile
hit <**Spacebar**> to see the next screen
Note: some people prefer **less**

cp copies files
cp orig.file copy.file
cp orig.file subdir/new.file
copies *orig/file* to *new/file* in *subdir* directory

cp subdir/orig.file .
copies *orig/file* from *subdir* to the current directory
(.) without changing its name

mv moves/renames a file (or directory)
mv oldname newname
mv my.file subdir/my.file
a move (mv) is equivalent to a copy (cp) followed by a remove (rm)

rm removes/deletes a file.
rm oldfile
rm -i *.*file
option -i (interactive) advised if wildcards (*) in use

diff compares two files and prints how they differ
diff file1 file2
prints differences to screen options include -b to ignore differences in blank space, and -i to ignore case

find searches the directory tree for a file
find . -name lostfile -print
will search your current directory (.) (and any subdirectories) for *lostfile*

grep searches a file for a string
grep word my.file
grep "two words" my.file
options include -i to ignore case and -n to print line numbers

HEADERS

head redirects output of a command to a file
diff file1 file2 > newfile
puts differences into *new/file*

cat onefile twofile > bothfile
writes the output of the cat command into *both/file*
(overwrites *both/file*)

tail appends a file to the bottom of another
cat threefile >> bothfile
appends *three,file* to the bottom of *bothfile*

cat “pipe” - uses the output of the first command as the input of the second
grep string myfile | wc -1
finds how many lines on which “*String*” occurs (see **grep** and **wc**)

DIRECTORIES

cd changes current directory
cd /etc go to /etc directory
cd .. go up one level in directory tree
cd ./subdir2 go “sideways” to *subdir2*

mkdir creates a new subdirectory
rmkdir subdir removes a directory - you must delete all the files in it first

pwd print working directory, tells your current location (path)

e.g. `seqret "embl:hsfaul[-100:]"`

A part of the sequence can be specified by adding the range:
 e.g. `seqret "embl:hsfaul[1:57]"`

The last 100 bases of a sequence can be specified by a negative start:
 e.g. `seqret "embl:hsfaul[-100:]"`

List Files

A list file contains a list of USAs (one per line). The list file input is @listfile. A list file may be read in wherever a program can read multiple sequences. Blank lines and USAs starting with a '#' character are ignored. There is no limit on different sequence formats within one list file.

Format Conversion

The format of an output sequence file can be specified. `seqret` can read in sequences in one format and write them in the other format, for example to convert a sequence to GCG format:

```
seqret in.seq gcg::out.seq
```

The command line and parameters

EMBOSS programs are designed to be run from the command-line, as well as within scripts. To customise their behaviour, each has a distinct set of parameters, also known as options or flags. There are 3 classes of parameters: *standard*, *additional*, *advanced*. Information on allowable flags for each program is given in the help files.

If values for *standard* (mandatory) parameters are not specified, the programs will prompt for them.

If *additional* (optional) parameters are missed out, default values will be used unless you put options (or opt) on the command line.

EMBOSS programs never prompt for *advanced* parameters; these must be explicitly specified. They are defined in the program documentation.

General qualifiers

These can be used with any program:

- auto Turns off prompts and descriptions. Used when in running programs scripts
- stdout Writes to standard output (screen) by default
- filter Reads from standard input (keyboard), writes to standard output (screen) by default
- options Prompts for all required and additional values
- debug Writes debug output to the file *programname.debug*
- help Reports command line options. Or help verbose for more information on associated and general qualifiers
- warning Reports warnings
- error Reports errors

-fatal Reports fatal errors

-die Reports deaths

Each of these can be prefixed with "no" to negate the action.
 e.g. `-nowarning`

-sbegin States the first position of the sequence
 -send States the final position of the sequence

Some major programs

EMBOSS currently offers approximately 200 applications Use `wesename` to see them all together with below a selection of interesting tools:

TOOLS (examples)

<code>seqret</code>	Reads and writes (returns) sequences
<code>est2genome</code>	Aligns EST and genomic DNA sequences
<code>needle</code>	Needleman-Wunsh global alignment
<code>water</code>	Smith-Waterman local alignment
<code>dotmatcher</code>	Displays a thresholded dotplot of two sequences
<code>remap</code>	Displays a sequence with restriction cut sites, translation etc
<code>prettyplot</code>	Displays aligned sequences, with colouring and boxing
<code>extractseq</code>	Extracts regions from a sequence
<code>revseq</code>	Reverses and complements a sequence
<code>plotorf</code>	Plots potential open reading frames
	<i>and many other</i>

UTILS MISC

<code>embossdata</code>	Finds or fetches the data files read in by the EMBOSS programs
<code>embossversion</code>	Writes the current EMBOSS version number

This document was written and designed by Lisa Mullan from the UK EMBOSS node and being distributed by P&PR Publications Committee of EMBnet.

EMBnet - European Molecular Biology Network - is a bioinformatics support network of bioinformatics support centers situated primarily in Europe. Most countries have a national node which can provide training courses and other forms of help for users of bioinformatics software.

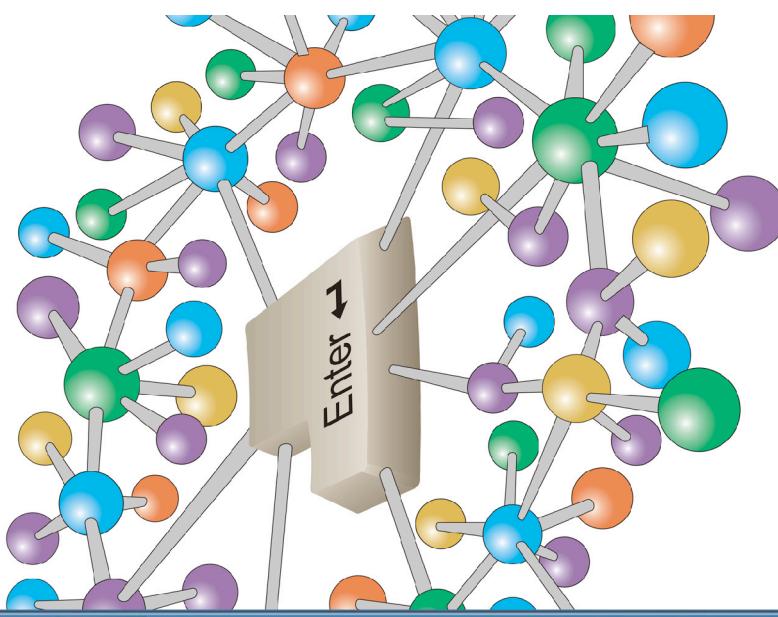
You can find information about your national node from the EMBnet site:

<http://www.embnet.org/>

A Quick Guide To EMBOSS
 First edition © 2004

A Quick Guide

EMBOSS



EMBnet

A Quick Guide To EMBOSS

<http://www.emboss.org>



This is a Quick reference Guide for EMBOSS version 2.8.0

Rice, P., Longden, I. and Bleasby, A. (2000)
“EMBOSS: The European Molecular Biology Open Software Suite” *Trends in Genetics* **16**(6):276-277.

Introduction

EMBOSS (European Molecular Biology Open Software Suite) is a freely available suite of programs and libraries for sequence analysis. It incorporates many tools originating from the EGCG package created in 1988. All EMBOSS programs are designed to run on a UNIX command-line or behind graphical interfaces (e.g., Jemboss, wEMBOSS).

Obtaining EMBOSS

To install EMBOSS download the current version from <ftp://ftp.uk.embnet.org/pub/EMBOSS/EMBOSS-2.8.0.tar.gz>, then follow the instructions at: <http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/download.html>

Graphical User Interfaces

There are a number of graphical interfaces to EMBOSS:
<http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/interfaces.html>

Jemboss is a java interface and is distributed with EMBOSS. If you are installing with the Jemboss interface you should use the installation script in the `EMBOSS-x.x/jemboss/utils` directory. Instructions for Jemboss installation are given at: <http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/Jemboss>

Support and Mailing lists

The mailing list `emboss@embnet.org` is used for discussions of user problems. To subscribe to this list, send a mail to `majordomo@embnet.org` with the message text: `subscribe emboss`. The mailing list archive is:
<http://www.rfcgr.mrc.ac.uk/Emboss/HYPERMAIL/emboss>

Please send bug reports to `emboss-bug@embnet.org`

Any program derived from Bill Pearson FASTA suite of programs has a markx default format.
-aformat Alters output format
-awidth Displays alignment width
-ausashow Displays the full USA (see below) in the alignment

Feature Formats	Description
gff	General Feature format defined by the Sanger Institute [default]
embl	Feature table used by EMBL database (em)
swissprot	Feature table used by SwissProt database (sw)
-ufo	UFO (uniform features object) features
-fformat	Opens features format

These flags can be applied to the output by using “o” as a prefix, e.g. -oufo

-begin	Specifies first position
-end	Specifies final position
-reverse	Reverses features (DNA only)

Graphic Formats

-graph	Static graphics using PLP plot. Output as X11 [default], PNG, ps, tektronics amongst others
--------	---

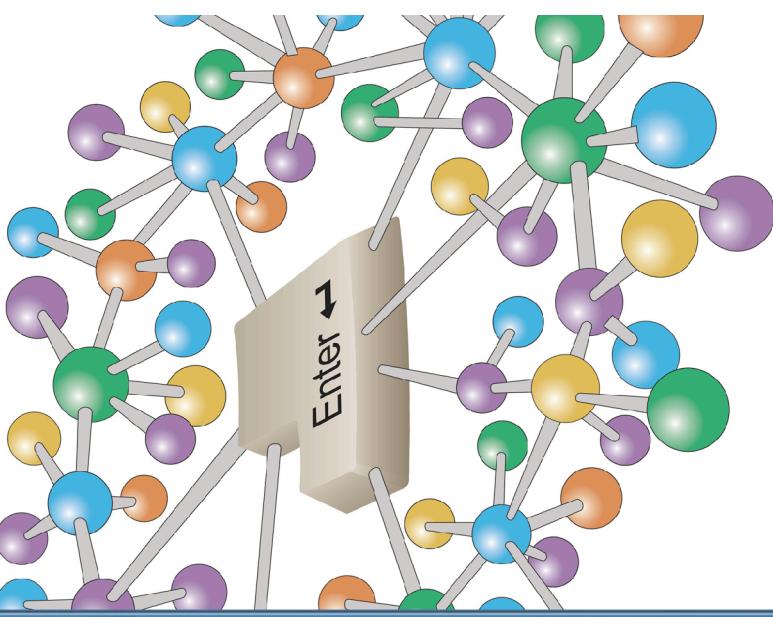
Your local EMBOSS installation may have many sequence databases set up. The program showdb will indicate the available databases.

Uniform Sequence Address (USA)

A USA is an unambiguous means of specifying sequences in EMBOSS. It has the following syntax:
`format::database:entry`
Only raw (text) or IntelliGenetics format need to be specified. EMBOSS identifies the rest automatically.

You may also use:	all sequences in a file
filename	an entry in a file
@listfilename	a list file (see below)
asis::ACGACTGACGG	a specific short sequence

The entry can include ‘*’ characters for wildcard matches of several entries and sequence may be specified by adding [start::end::rev] positions to the USA. The rev keyword will reverse complement a DNA sequence. Command lines using these characters must be encased in double quotes :



A Quick Guide

BLAST

EMBnet

Selected megablast arguments:

```
-D [integer] DB genetic code (def = 1).
-M [string] matrix (def = BLOSUM62),
produces HTML output (def = F).
-T [T/F] uses lower case filtering (def = T) Obs.: T =
any lower-case letter in input FASTA file
should be masked.
```

Position Specific Iterated BLAST

PSI-BLAST is a variant of blast that searches a query against a database using a position-specific scoring matrix created by PSI-BLAST. First run **blastpgp** to create and save a position-specific scoring matrix, then run **blastpgp** again to search iteratively with the previously saved matrix. e.g.,

```
blastpgp -i ff.chd -d yeast -C ff.chd.ckp
blastpgp -i ff.chd -d nr -j 3 -R ff.chd.ckp
```

Selected blastpgp arguments for PSI-BLAST:

```
-j [integer] maximum number of iterations (def = 1).
-h [number] E-value threshold for including sequences in
the score matrix model (def = 0.001).
-C [file out] stores the query and frequency count ratio
matrix in a file (opt).
-Q [file out] output file for PSI-BLAST matrix in ASCII
(opt).
-R [file in] restarts from a file stored previously with -C.
-B [file in] input alignment for restart.
```

Pattern-Hit Initiated BLAST

PHI-BLAST is a search program that combines the matching of regular expressions with local alignments surrounding the match. E.g.:

```
blastpgp -i query.file -k pattern.file -p patseedp
```

Selected blastpgp arguments for PHI-BLAST:

```
-i [file in] input sequence file in FASTA format.
-k [file in] pattern syntax follows the PROSITE
conventions).
-p [string] usage mode (def = blastpgp). Obs.: use
patseedp, if pattern occurs only once,
and 'seedp', if it occurs more than once per
protein.
```

Obs.: You can integrate a PSI-BLAST search after the PHI-BLAST search, using the argument ``-j''. E.g.,

```
blastpgp -i query -k pattern -p patseedp -j 2
```

Mega BLAST

Mega BLAST uses a greedy algorithm optimized for aligning sequences that differ slightly as a result of sequencing or other similar «errors». When a larger word size is used, it is up to 10 times faster than more common sequence similarity programs. It is also able to efficiently handle much longer DNA sequences than the blastn program.

Selected megablast arguments:

-D [integer]	DB genetic code (def = 0 = alignment endpoints and score; 1 = all ungapped segments endpoints; 2 = traditional BLAST output; 3 = tab-delimited one line format).
-M [integer]	maximal total length of queries for a single search (def = 20000000).
-f [T/F]	shows full IDs in the output (def = F, only GI or accessions).
-p [real]	identity percentage cut off (def = 0).
-s [integer]	minimal hit score to report (def = 0).

To compare two sequences

bl2seq performs a pairwise comparison between two sequences.

Selected bl2seq arguments:

-i [file in]	first sequence.
-j [file in]	second sequence.
-p [string]	program name (as in blastall; def = blastsp).
-o [T/F]	alignment output (def = stdout).
-G [integer]	cost to open a gap (def = 0; zero invokes default behavior).
-E [integer]	cost to extend a gap (def = 0; zero invokes default behavior).
-W [integer]	wordsize (def = 0; zero invokes default behavior).
-M [string]	matrix (def = BLOSUM62).
-F [string]	filters query sequence (def = T).
-e [real]	expectation value E (def = 10.0),
-T [T/F]	produces HTML (def = F).

This document was written and designed by Eduardo Fernandes Fornighieri with the help of Marcos Renato R. Araújo, Marcelo Falsarella Carazzolle and Gonçalo A. Guimaraes Pereira from the Brazilian EMBnet node and distributed by the P&PR Publications Committee of EMBnet.

EMBnet – European Molecular Biology network – is a network of bioinformatics support centers situated primarily in Europe. Most countries have a national node, which can provide training courses and other forms of help for users of bioinformatics software.

<http://www.embnet.org/>

A Quick Guide to NCBI Blast
First edition © 2004

A Quick Guide to the NCBI Blast

<http://www.ncbi.nlm.nih.gov/blast>

This guide doesn't replace the entire documentation for Blast. For beginners we suggest to first read the documentation of the Blast related to similarity searching (see link below). Other useful pages are available by following the links at the top of this page. E.g., the glossary and the tutorials:
<http://www.ncbi.nlm.nih.gov/Education/BLASTinfol/similarity.html>

Program selection – web interface options

BLASTN – used to search nucleotide databases with a nucleotide query sequence.
MEGABLAST – a version of BLAST specially designed to efficiently find very similar sequences in a database.
Discontiguous MEGABLAST – a version of MEGABLAST used to identify similar but not identical nucleotide sequences.

Search for short nearly exact matches – used to search for primer or short nucleotide motifs in nucleotide sequences or short peptides in protein sequences.
BLASTP – used to search protein databases with a protein query sequence.

PSI-BLAST (Position-Specific Iterated BLAST) – used to search protein databases with increased sensitivity potentially locating distant homologies. A position-specific scoring matrix is created after each iteration using the selected results from the previous search.

PHI-BLAST (Pattern-Hit Iterated BLAST) – a version similar to PSI-BLAST, but including a user-defined pattern limiting the output to sequences matching the pattern. The patterns must follow the pattern syntax conventions from PROSITE.

BLASTX – makes a six-frame nucleotide query search against a protein database, finding proteins similar to those encoded by the query. Useful when the reading frame of the query is unknown or when it contains errors that may lead to frame shifts.

TBLASTN – makes a protein query search against a dynamically translated nucleotide database. Useful when searching for a specific protein against an unannotated nucleotide database, like HTGs or ESTs databases.
TBLASTX – searches all six-frame query translations against all six-frame database translations. Effectively performs a more sensitive blastp search without doing manual translations.

CDD-Search (Conserved Domain Database Search)

– used to identify conserved protein domains.

CDART (Conserved Domain Architecture Retrieval Tool)

– explores the domain architectures of proteins.

Blast 2 sequences – direct comparison of two sequences.

VeeScreen – screens DNA sequence queries for vector contamination using a database of known vectors.

Main databases (available at NCBI)

Protein *nr* (non-redundant + PDB + SwissProt + PIR + PRF); *swissprot* (latest major release of the SWISS-PROT); *pat* (proteins from patent division of GenBank); *month* (new data released in the last 30 days); *pdb* (3-dimensional structure records from Protein Data Bank).

Nucleotide *nr* (GenBank + EMBL + DDBJ + some PDB); *est* (GenBank + EMBL + DDBJ from EST division); *pat* (nucleotides from patent division); *pdb* (3-dimensional structure records); *month* (new data released in the last 30 days); *chromosome* (complete genomes and chromosomes); *est human* (human subset of EST); *est mouse* (mouse subset of EST); *est others* (subset of EST other than human or mouse); *gss* (Genome Survey Sequence); *htgs* (Unfinished High Throughput Genomic Sequences); *aln repeats* (select Alu repeats from REBASE); *dbsts* (STS division + EMBL + DDBJ); *wgs* (assemblies of whole genome shotgun sequences).

LOCAL BLAST INSTRUCTIONS

Format source databases

formatdb formats protein or nucleotide source databases before they can be searched by blastall, blastp, blastpgp or megablast. The source database may be in either FASTA or AN.1 format.

Selected formatdb arguments:

-t [string] title for database (opt).
-i [file in] input file for formatting.
-l [file out] logfile name (opt, def = formatdb.log).
-Q [integer] query genetic code (def = 1).

type of file (opt; T = protein (def); F = nucleotide). parse options (opt; T = parse SeqID and create indexes; F = no parse, no indexes (def)). Obs.: the first word on the fasta definition line should be a unique identifier (SeqID).

size of the volume in millions of letters (opt; def = 0). Obs.: This option breaks up large FASTA files into ‘volumes’ (each with a maximum size of 2 billion characters). I.e.: -v 2000. base name for BLAST files (opt).

Fasta from databases

fastacmd retrieves FASTA formatted sequences from a BLAST database, if it was formatted using the ‘-o’ option.
Selected fastacmd arguments:

-d [string] database (def = nr).
-s [string] search string.
-i [string] input file with GIIs/accessions/locuses for batch retrieval (opt).
line length for sequence (def = 80, opt).

Stand-alone blast

performs all five flavors of blast comparison.
blastall performs all five flavors of blast comparison.

Selected blastall arguments:

-p [string]
-d [string]
-i [file in]
-e [real]
-o [file out]
-F [string]

To change SEG options, use: -F “S 10 1.0 1.5”, where 10 = window value, 1.0 = low cut and 1.5 = high cut.
For coiled-coil filter: -F “C 28 40.0 32”, where 28 = window, 40.0 = cut off and 32 = linker.
To use both SEG and coiled-coil: -F “C;S”.
number of alignments (def = 250).
number of one-line description (def = 500).
query genetic code (def = 1).