

# Introdução a Bioinformática - CEN0485

Diego Mauricio Riaño Pachón

25 de Março de 2022



# Conteúdo

<b>1 Bases de bioinformática</b>	<b>9</b>
<b>2 Ferramentas Unix úteis em bioinformática</b>	<b>11</b>
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
<b>3 Búsquedas en bases de datos biológicas</b>	<b>21</b>
3.1 NCBI – Bases de datos y búsqueda de información . . . . .	21
3.1.1 Iniciemos una visita a sus bases de datos . . . . .	21
3.1.2 Recuperación de Secuencias en el NCBI con búsquedas más específicas . . . . .	27
3.2 Recuperación de secuencias usando SRS@EBI . . . . .	28
<b>4 Manipulación básica de secuencias</b>	<b>33</b>
4.1 Limpieza de secuencias . . . . .	33
4.2 Mapa de restricción . . . . .	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
<b>5 Creación de bases de datos relacionales</b>	<b>37</b>
<b>6 Búsquedas en base de datos biológicas - Segunda parte</b>	<b>41</b>
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
<b>7</b>	<b>Ontologías en bioinformática: Gene Ontology</b>	<b>47</b>
7.1	Consultas en GO . . . . .	47
<b>8</b>	<b>Introducción al análisis de redes usando Cytoscape</b>	<b>53</b>
<b>9</b>	<b>Análisis de enriquecimiento de anotaciones de genes</b>	<b>55</b>
<b>10</b>	<b>Comparación de secuencias I - Matrices de puntos</b>	<b>57</b>
<b>11</b>	<b>EMBOSS</b>	<b>61</b>
11.1	Recuperando secuencias de bases de datos . . . . .	61
11.2	Selección de marco de lectura abierto . . . . .	62
11.3	Barajar/mezclar secuencias . . . . .	63
11.4	Predicción de regiones hidrofóbicas . . . . .	63
11.5	Alineamientos . . . . .	63
<b>12</b>	<b>Comparación de secuencias II - Alineamientos pareados</b>	<b>65</b>
12.1	Matrices de sustitución . . . . .	65
12.2	Alineamiento Global . . . . .	65
12.3	Alineamientos locales . . . . .	66
12.4	Significancia de los alineamientos . . . . .	67
<b>13</b>	<b>BLAST: BASIC LOCAL ALIGNMENT SEARCH TOOL</b>	<b>69</b>
13.1	Encontrando la región genómica de un transcripto. . . . .	73
13.2	Blast+ en la línea de comandos . . . . .	73
<b>14</b>	<b>Alineamientos múltiples</b>	<b>75</b>
14.1	Alineando las secuencias de amino ácidos de TRIM5 $\alpha$ de primates . . . . .	75
14.1.1	CLUSTALX . . . . .	75

14.1.2 T-COFFEE . . . . .	76
14.1.3 MUSCLE . . . . .	77
14.1.4 Comparar los alineamientos usando la herramienta web ALTAVIST . . . . .	77
14.1.5 Del alineamiento de proteínas al de nucleótidos . . . . .	78
14.1.6 Editando y visualizando alineamientos . . . . .	79
14.1.7 Estimando distancias entre las secuencias . . . . .	79
<b>15 PSSMs, Logos de secuencias y HMMs</b>	<b>81</b>
15.1 PSSM . . . . .	81
15.2 Logos de secuencias . . . . .	82
15.3 Modelos Ocultos de Markov: HMMs . . . . .	82
15.3.1 Buscando los dominios de una proteína . . . . .	83
15.3.2 Visualización de HMMs . . . . .	84
<b>16 Diseño de primers para PCR</b>	<b>85</b>
16.1 Diseño de primers usando Quantprime . . . . .	87
16.2 Crear primers a partir de alineamientos de proteínas . . . . .	90
<b>Apendices</b>	<b>97</b>



# 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 de inicio NCBI . . . . .	22
3.2	Ventanilla de búsqueda en el NCBI . . . . .	22
3.3	Página principal de Entrez . . . . .	25
3.4	Página principal de SRS . . . . .	29
3.5	Opciones SRS . . . . .	29
3.6	Opciones SRS . . . . .	30
3.7	Formulario de búsqueda SRS . . . . .	31
3.8	Criterios de búsqueda avanzada . . . . .	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	Agregar secuencias en Dot Let . . . . .	58
10.3	Botones de control . . . . .	58
10.4	Resultado . . . . .	59
11.1	Recuperando secuencias de las bases de datos . . . . .	62

13.1 Tipos de BLAST disponibles en el NCBI . . . . .	69
13.2 Interfaz web de NCBI BLAST usando el 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 Resultados de la comparación de alineamientos con ALTAVIST . . . . .	77
14.2 Resultados de la comparación de alineamientos con ALTAVIST . . . . .	78
15.1 Logo de secuencias de los sitios de unión de LexA . . . . .	82
15.2 Resultados de una búsqueda en Pfam . . . . .	83
16.1 Creación un proyecto en QUANTPRIME . . . . .	88
16.2 Adicionando transcritos al proyecto en QUANTPRIME . . . . .	88
16.3 QUANTPRIME buscando primers para los genes solicitados . . . . .	89
16.4 Listado de los mejores primers encontrados por QUANTPRIME . . . . .	89
16.5 Página de información para un par de primers seleccionados . . . . .	90
16.6 Página de inicio en iCODEHOP . . . . .	91
16.7 Diseño de primer en iCODEHOP . . . . .	91
16.8 Diseño de primer en iCODEHOP . . . . .	92
16.9 Detección de BLOCKS en el alineamiento de secuencias de proteínas. Se diseñaran primers para cada BLOCK . . . . .	93
16.10 Detección de BLOCKS en el alineamiento de secuencias de proteínas. Se diseñaran primers para cada BLOCK . . . . .	94

# 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.

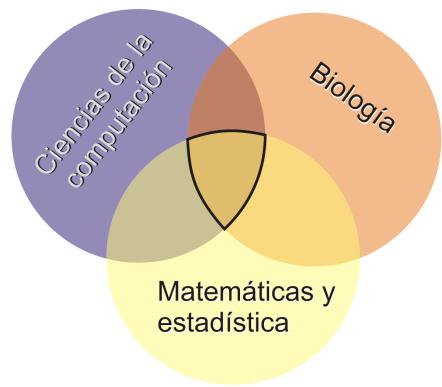
De acordo com Philip Bourne (UCSD), “a bioinformática tornou-se a intérprete da linguagem genômica do DNA e está tentando decifrar linguagens mais complexas em que as proteínas são os substantivos, as interações são a sintaxe, as vias metabólicas são frases e os sistemas vivos são o volume completo” (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 Unix<sup>1</sup>.

---

<sup>1</sup>Linux, 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 operativo<sup>1</sup> 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 portabilidade<sup>2</sup>. 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 16.2<sup>3</sup>.

#### 2.1.1 A linha de comando

A linha de comando é acessada através de um programa de interpretação chamado “shell”<sup>4</sup>. 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**,

---

<sup>1</sup>Mais informações em [http://en.wikipedia.org/wiki/Operating\\_system](http://en.wikipedia.org/wiki/Operating_system)

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

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

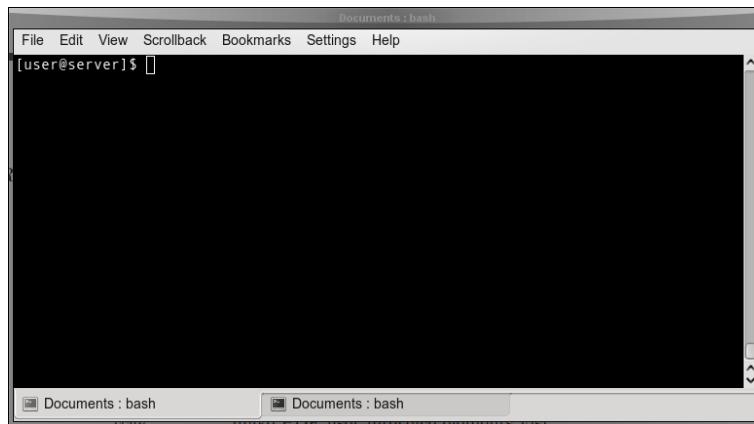
<sup>4</sup>[http://en.wikipedia.org/wiki/Unix\\_shell](http://en.wikipedia.org/wiki/Unix_shell)

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 \$PS1<sup>5</sup>. 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, \u<sup>6</sup>, 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 modificação<sup>7</sup>. 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

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

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

<sup>7</sup>Exercício 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.

```

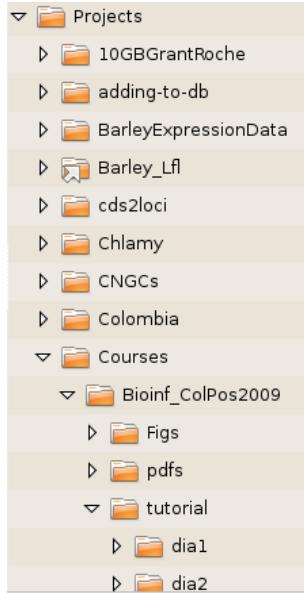
____ Baixando arquivos _____
7  [user@server:~]$ wget https://github.com/labbcles/cen0485/raw/main/linux/praticas/file1.tar.gz
8  --2022-03-25 16:26:12-- https://github.com/labbcles/cen0485/raw/main/linux/praticas/file1.tar.gz
9  Resolving github.com (github.com)... 20.201.28.151
10 Connecting to github.com (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 Windows<sup>TM</sup>que 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/

```

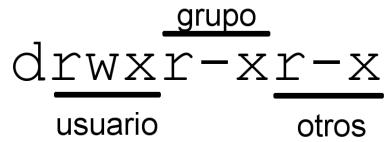
---

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 2.4<sup>8</sup>.

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

---

<sup>8</sup>Exercício 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 62<sup>9</sup>?

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*
```

---

<sup>9</sup>Revise 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 Scriptome<sup>10</sup>, 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 mvggkkktkicdkvshedrisqlpepliseifhlstksalstkwrylwqsvpg
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.

---

<sup>10</sup><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 Fasta<sup>11</sup>. 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), qode 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 GenBank<sup>1213</sup> é usado pelo “National Center for Biotechnology Information” (NCBI<sup>14</sup>), o maior repositório de sequências, tanto ácidos nucleicos quanto proteínas, em todo o mundo. O NCBI juntamente com o EMBL<sup>15</sup> e o DDBJ<sup>16</sup>, 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.

---

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

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

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

<sup>14</sup><http://www.ncbi.nlm.nih.gov/>

<sup>15</sup><http://www.ebi.ac.uk/emb1/>

<sup>16</sup><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”<sup>17</sup>.

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.

---

<sup>17</sup>Para sair de less pressione “q”

# Capítulo 3

## Búsquedas en bases de datos biológicas

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

### 3.1 NCBI – Bases de datos y búsqueda de información

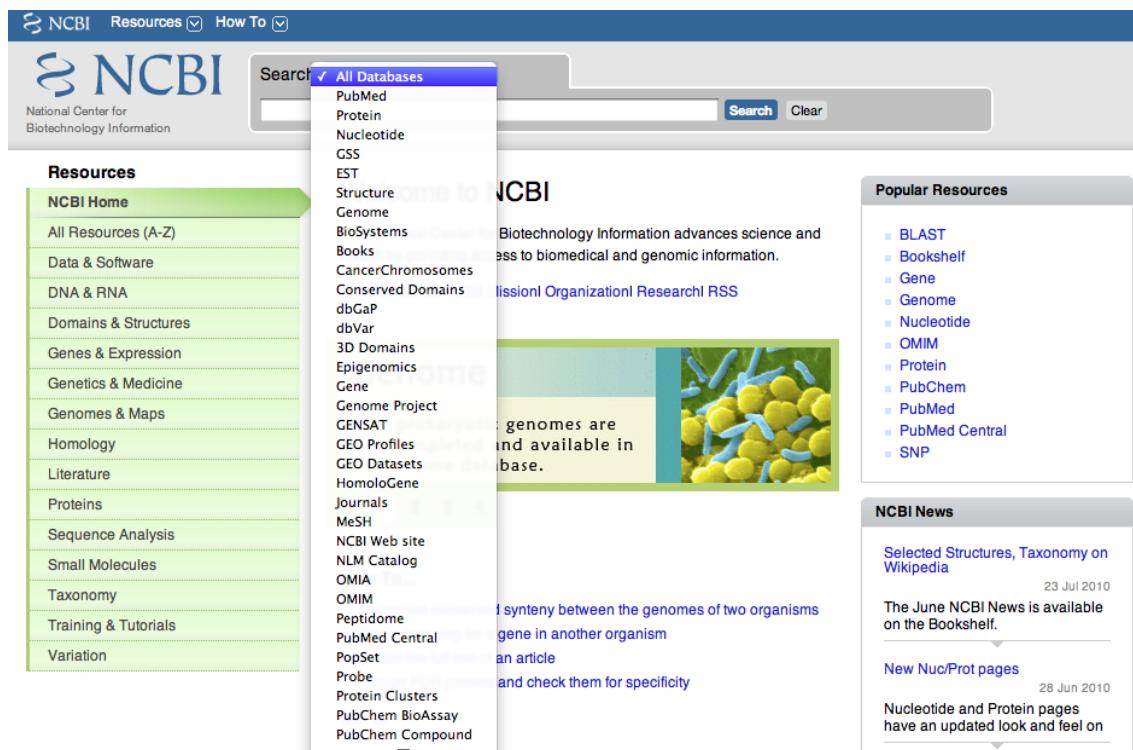
El Centro Nacional de Información en Biotecnología, NCBI por sus siglas en inglés, es una institución pública de los Estados Unidos de América, que salvaguarda toda la información sobre los genomas de varias especies, así como la mayor base de datos pública sobre secuencias de ADN y proteínas. Su página principal de red esta ubicada el siguiente enlace: <http://www.ncbi.nlm.nih.gov/>

Esta página conecta con todos los datos disponibles en sus servidores (PubMed, ALL Databases (Entrez), Blast, OMIM, Books, TaxBrowser, Structure), como se observa en la Figura 3.1. Aunque Entrez está listado como uno de los servicios, en realidad casi todos ellos dependen directamente de Entrez. Por ejemplo, PubMed y Taxonomy están íntimamente ligados al Entrez.

#### 3.1.1 Iniciemos una visita a sus bases de datos

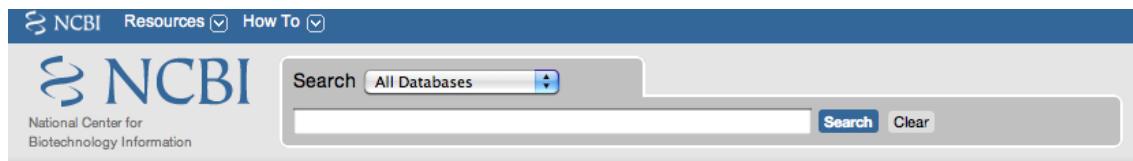
Como primera medida, entremos en PubMed. Esta base de datos contiene información sobre publicaciones científicas, y sus registros han sido compilados por el NLM (librería nacional de medicina), con colaboración de los editores. Allí encontrará la mayoría de referencias que necesite, incluyendo el resumen (Abstract) y en algunos casos la publicación de forma gratuita.

Para obtener ayuda sobre como efectuar búsquedas, refiérase al siguiente enlace: <http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed>



**Figura 3.1:** Página de inicio NCBI

Las páginas tienen un menú de bases de datos en una barra superior, las búsquedas deben colocarse en la ventanilla que se muestra en la Figura 3.2.



**Figura 3.2:** Ventanilla de búsqueda en el NCBI

Una búsqueda debe tomar un formato similar a este:

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

Donde **palabra clave** es la palabra que sirve para identificar un registro (record) según el campo (field) usado. Por ejemplo, una palabra clave puede ser ”garcía” en el campo de ”autores”. **Operador lógico** es cualquiera de estos operadores booleanos: AND, OR, NOT, BUT, etc. Cuando reemplace con sus propios términos claves en el formato de arriba, recuerde que los campos deben estar entre paréntesis cuadrados [ ], pero los operadores van solos (sin los símbolos, ), adicionalmente, las comillas en la palabra claves son opcionales, pero cumplen la función de forzar una búsqueda con la palabra exacta en vez de ser flexible.

Por ejemplo, si quiero buscar todos los artículos de 1999 publicados por García y demás en la revista Science, uso el siguiente comando: García[AU] AND 1999[DP] AND "science"[TA]. Entre más información entre en la búsqueda, mayor restringida será la respuesta, (por ejemplo, si incluyo más autores).

Los campos (fields) más comunes que se pueden preguntar en PubMed son los siguientes:

**All Fields [ALL ]** Includes all searchable PubMed fields. However, only terms where there is no match found in one of the Translation tables or Indexes via the Automatic Term Mapping process will be searched in All Fields. PubMed ignores stopwords from search queries.

**Author Name [AU ]** Various limits on the number of author names included in the MEDLINE citation have existed over the years (see NLM policy on author names). MEDLINE does not list the full name. The format to search for an author name is: last name followed by a space and up to the first two initials followed by a space and a suffix abbreviation, if applicable, all without periods or a comma after the last name (e.g., fauci as or o'brien jc jr). Initials and suffixes may be omitted when searching. PubMed automatically truncates on an author's name to account for varying initials, e.g., o'brien j [au] will retrieve o'brien ja, o'brien jb, o'brien jc jr, as well as o'brien j. To turn off this automatic truncation, enclose the author's name in double quotes and qualify with [au] in brackets, e.g., "o'brien j"[au] to retrieve just o'brien j.

**EC/RN Number [RN ]** Number assigned by the Enzyme Commission to designate a particular enzyme or by the Chemical Abstracts Service (CAS) for Registry Numbers.

**Entrez Date [EDAT ]** Date the citation was added to the PubMed database. Citations are displayed in Entrez Date order which is last in, first out. Dates or date ranges must be entered using the format YYYY/MM/DD [edat], e.g. 1998/04/06 [edat] . The month and day are optional (e.g., 1998 [edat] or 1998/03 [edat]). To enter a date range, insert a colon (:) between each date (e.g., 1996:1997 [edat] or 1998/01:1998/04 [edat]).

**Issue [IP ]** The number of the journal issue in which the article is published.

**Journal Title [TA ]** The journal title abbreviation, full journal name, or ISSN number.

**Language [LA ]**

**Publication Date [DP ]** The date that the article was published. Dates or date ranges must be searched using the format YYYY/MM/DD [dp], e.g. 1998/03/06 [dp] . The month and day are optional (e.g., 1998 [dp] or 1998/03 [dp]). To enter a date range, insert a colon (:) between each date (e.g., 1996:1998 [dp] or 1998/01:1998/04 [dp]).

**Substance Name [NM ]** The name of a chemical discussed in the article. Synonyms to the Supplementary Concept Substance Name will automatically map when qualified with [nm]. This field was implemented in mid-1980. Many chemical names are searchable as MeSH terms before that date.

**Text Words [TW ]** Includes all words and numbers in the title and abstract, and MeSH terms, subheadings, chemical substance names, personal name as subject, and MEDLINE Secondary Source (SI) field. The Personal Name of Subject field can also be searched directly using the search field tag [ps], e.g., nightingale f [ps].

**Title Words [TI ]** Words and numbers included in the title of a citation.

**Title/Abstract Words [TIAB ]** Words and numbers included in the title and abstract of a citation.

**Unique Identifiers [UID ]** PubMed Unique Identifier PMID and MEDLINE Unique Identifier UI .

**Volume [VI ]** The number of the journal volume in which an article is published.

Ahora vamos a la página donde se encuentra ENTREZ. Para ello seleccione ALL DATABASES en la ventanilla de bases de datos de la página principal. Entrez es un sistema de búsqueda de secuencias almacenadas en las bases de datos. Se pueden hacer preguntas sofisticadas para obtener un conjunto de secuencias que sean de interés propio, por ejemplo, puedo pedir que muestre todas las secuencias genómicas de Arabidopsis que fueron incluidas en la base de datos entre los años 97' y 99' que además contengan anotación (en la tabla de "features") sobre regiones promotoras. La Figura 3.3 muestra la página de entrada al servidor de Entrez.

Así en una sola página podemos realizar búsquedas simultáneamente en todas las bases de datos o seleccionar una sola base de datos y hacer una búsqueda por base de datos.

En la casilla de búsqueda, se pueden preguntar secuencias usando sus números identificadores (como el gi-number o con el número de accesión). También se pueden formular preguntas más complicadas utilizando la sintaxis de entrez, similar a como vimos PubMed:

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

Para obtener mayor información sobre Entrez puede seguir el enlace: <http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpentrez&part=EntrezHelp>

## Ejercicios

1. ¿Cuál es la clasificación taxonómica del alga *Chlamydomonas reinhardtii*, y qué otras plantas son cercanas, de tal manera que puedan usarse como fuente de marcadores. Cuantas secuencias de proteínas están presentes en GenBank para la especie *Chlamydomonas reinhardtii*?

**Welcome to the Entrez cross-database search page**

<b>PubMed:</b> biomedical literature citations and abstracts	<b>Books:</b> online books
<b>PubMed Central:</b> free, full text journal articles	<b>OMIM:</b> online Mendelian Inheritance in Man
<b>Site Search:</b> NCBI web and FTP sites	<b>OMIA:</b> online Mendelian Inheritance in Animals
<b>Nucleotide:</b> Core subset of nucleotide sequence records	<b>dbGaP:</b> genotype and phenotype
<b>EST:</b> Expressed Sequence Tag records	<b>UniGene:</b> gene-oriented clusters of transcript sequences
<b>GSS:</b> Genome Survey Sequence records	<b>CDD:</b> conserved protein domain database
<b>Protein:</b> sequence database	<b>3D Domains:</b> domains from Entrez Structure
<b>Genome:</b> whole genome sequences	<b>UniSTS:</b> markers and mapping data
<b>Structure:</b> three-dimensional macromolecular structures	<b>PopSet:</b> population study data sets
<b>Taxonomy:</b> organisms in GenBank	<b>GEO Profiles:</b> expression and molecular abundance profiles
<b>SNP:</b> single nucleotide polymorphism	<b>GEO DataSets:</b> experimental sets of GEO data
<b>dbVar:</b> Genomic structural variation	<b>Epigenomics:</b> Epigenetic maps and data sets
<b>Gene:</b> gene-centered information	<b>Cancer Chromosomes:</b> cytogenetic databases
<b>SRA:</b> Sequence Read Archive	<b>PubChem BioAssay:</b> bioactivity screens of chemical substances
<b>BioSystems:</b> Pathways and systems of interacting molecules	<b>PubChem Compound:</b> unique small molecule chemical structures
<b>HomoloGene:</b> eukaryotic homology groups	<b>PubChem Substance:</b> deposited chemical substance records
<b>GENSAT:</b> gene expression atlas of mouse central nervous system	<b>Protein Clusters:</b> a collection of related protein sequences

**Figura 3.3:** Página principal de Entrez

2. Vaya a la página de PubMed y consiga las referencias que traten sobre la biología molecular y/o genética de la yuca (*Manihot esculenta*). ¿Cuántas fueron publicadas en los últimos dos años y de qué laboratorios (o regiones geográficas) son sus autores? Explique cómo realizó la búsqueda. Hint: GoPubMed <http://www.gopubmed.org/>
3. Use Entrez para encontrar todas las secuencias tipo EST (Expressed Sequence Tag) de arroz que han sido depositados en la base de datos.

Revise la descripción de los principales formatos de secuencias en la sección 2.2.

### ¿Qué bases de datos encontramos en el NCBI?

El NCBI posee un gran número de bases de datos. La más conocida en GenBank que contiene todas las secuencias nucleotídicas. GenPept contiene las secuencias de proteínas. Otras bases de datos son Genome, Structure, PubMed

En GenBank las secuencias están organizadas en 17 divisiones, 11 tradicionales y 6 Bulk. En las tradicionales las secuencias han sido mandadas directamente por los investigadores, están caracterizadas y las divisiones son:

**PRI** primates

**PLN** plantas

**BCT** Bacterias

**INV** Invertebrados

**ROD** Roedores

**VRL** Virales

**VRT** otros vertebrados

**MAM** Mamíferos (Ej. ROD + PRI)

**PHG** Fagos

**SYN** Sintéticos (vectores de clonacion, etc)

**UNA** sin anotación

Las Bulk consisten en secuencias mandadas en grupo via email o ftp, inexactas y poco caracterizadas, son:

**dbEST** Base de datos de ESTs, Expressed Sequence Tags

**dbSTS** Sequence-tagged sites: son cortos landmarks genómicos para los cuales hay información de secuencia y de mapa

**dbGSS** Genomic survey sequences. Contiene: datos de secuencia paso único del genoma, secuencias terminales de BAC, YAC y cosmidos, secuencias de exones

**dbHTGS** High-Throughput Genomic Sequences. Fue creada para guardar información de la secuenciación de genomas que no estaba terminada ni curada pero para darla a conocer a la comunidad científica tan pronto como estuviera disponible

y además se tienen bases de datos para:

**HTC** High Throughput cDNA

**PAT** Patent

## RefSeq

Especial énfasis queremos hacer en una base de datos de NCBI llamada RefSeq. Esta base de datos fue creada para obtener una colección biológicamente no-redundante de secuencias de ADN, ARN y proteínas. Cada RefSeq (secuencia de referencia) representa una molécula única que ocurre naturalmente en un organismo. Esta base de datos es de tipo curada por investigadores. Cada molécula no es un resultado de investigación sino es una síntesis de información.

Volvamos a la página principal de NCBI y en la ventana de search, dejando all databases escribamos NC\_001139<sup>1</sup>. Vemos que en Nucleotide tenemos 1 hit, al igual que en Genome y en Gene tenemos 631.

Abramos la de Nucleotide: obtenemos un flatfile de secuencia que corresponde a la secuencia completa del cromosoma VII de la levadura. Examinemos un poco el flatfile, *¿qué información contiene?*

Notemos que los identificadores de esta base de datos cambian y son del tipo 2+6 con dos letras y 6 números, la siguiente tabla nos muestra que significan estas 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 Recuperación de Secuencias en el NCBI con búsquedas más específicas

**CONOCEMOS EL ORGANISMO.** Las búsquedas en NCBI se pueden hacer más dirigidas si conocemos el organismo del cual buscamos información. Entramos a la página inicial de NCBI, vamos a TaxBrowser, ponemos el nombre del organismo que estamos buscando. Cuando lo seleccionamos, a la derecha aparece una tabla de número de secuencias por tipo de molécula o proyecto. Al hacer clic en alguna de ellas, por ejemplo proteínas, nos lleva directamente a las proteínas de ese organismo.

<sup>1</sup>asegúrese de incluir el símbolo underscore

**CONOCEMOS EL O LOS NÚMEROS DE ACCESO.** Si conoce el número de acceso directamente lo puede poner en la ventana de búsqueda de la página principal de NCBI. Para varias secuencias se ponen los números con la palabra OR entre ellos, por ejemplo AJ487842 or AJ487843. Finalmente para una seguidilla de números de acceso se pone: AJ487842::AJ487851[ACCN]

**DIRIGIMOS LA BÚSQUEDA CON LIMITS.** Por ejemplo si quiero buscar las secuencias curadas de mRNA relacionadas con un tipo de cáncer en humanos puedo hacer la siguiente búsqueda: en la ventana de búsqueda pongo COLON CANCER AND NONPOLYPOSIS, busco la base de datos de nucleótidos. Luego en LIMITS selecciono la molécula mRNA y en only from (base de datos) selecciono RefSeq. Luego selecciono arriba la otra pestaña Preview/index y ahí en organismos escribo humans y selecciono AND

### 3.2 Recuperación de secuencias usando SRS@EBI

Existe sin embargo una alternativa excelente para la búsqueda de secuencias biológicas, que nos permite controlar casi todos los aspectos de nuestra búsqueda, esta alternativa es el Sistema de Recuperación de Secuencias (SRS). Este sistema fue desarrollado teniendo en mente precisamente esta labor de recuperar secuencias biológicas de una manera efectiva, de allí su diseño y sus capacidades.

En este taller trabajaremos con el SRS ofrecido por el Instituto Europeo de Bioinformática (EBI), <http://srs.ebi.ac.uk/>. O entrando al EBI, (<http://www.ebi.ac.uk/>) , database → database browsing se llega a SRS.

Una manera sencilla de consultar el SRS es mediante la casilla Quick Text Search. En dicha casilla es posible realizar búsquedas en diversas bases de datos disponibles en el menú desplegable, como se muestra en la Figura 3.4

Por ejemplo seleccionando la opción “Nucleotide Sequences” realizaremos nuestra búsqueda en la base de datos de DNA EMBL (homóloga al genBank y al DDBJ).

Realice una búsqueda rápida por HIV-1 con diferentes opciones del menú desplegable. Hasta este momento el SRS parece ser bastante menos completo comparado con el sitio web del NCBI, pero ahora empezaremos a comprobar donde radica todo su potencial.

Ahora realizaremos una búsqueda avanzada. Seleccione la pestaña Library Page ubicada en la parte superior de su pantalla y se muestra en la Figura 3.5

A continuación será llevado a la sección del SRS donde se describen cada una de las bases de datos que componen el sistema (Figura 3.6). Como puede ver el SRS comprende muchas bases de datos a la vez y esa es una de sus principales virtudes, por esta razón al SRS se le conoce algunas veces como una “base de datos de bases de datos”, pues a través de este sistema podemos consultar múltiples bases de datos al mismo tiempo, de acuerdo a nuestras necesidades particulares.

Como puede darse cuenta el SRS es similar al sistema ENTREZ del NCBI, en el sentido en que nos permite consultar muchas bases de datos al mismo tiempo, pero esta vez no restringidos

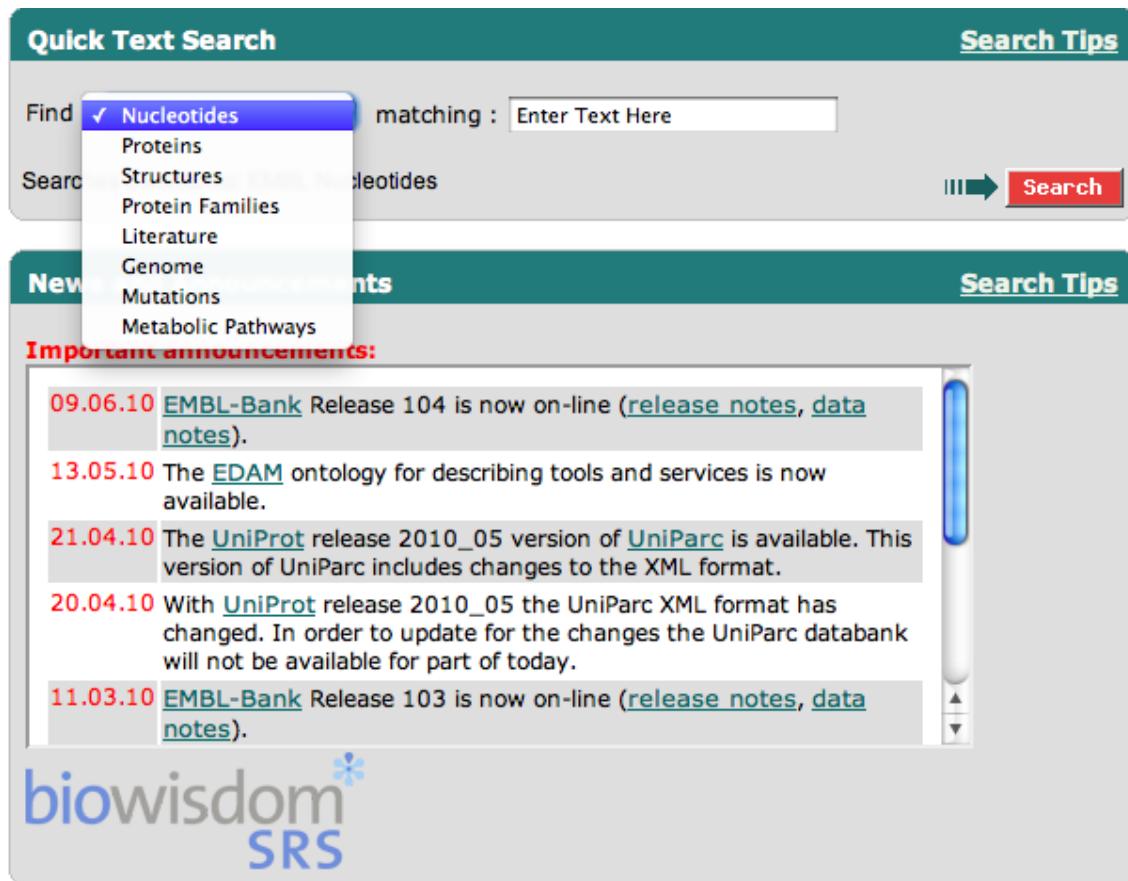


Figura 3.4: Página principal de SRS

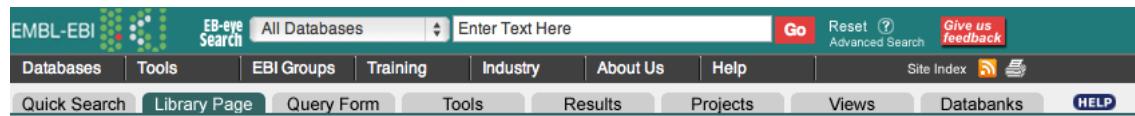


Figura 3.5: Opciones SRS

únicamente a aquellas con las que cuenta el NCBI sino a virtualmente cualquier base de datos. El número de bases de datos con las que cuenta el SRS depende de cada implementación, es decir el administrador del SRS determina qué bases de datos quiere o no incluir en su sistema

Ubique el cursor del ratón sobre alguna de las entradas, después de unos segundos una casilla de texto explicativo aparecerá. *¿Qué tipo de información proveen las bases de datos EMBL (Contig Updates), UniprotKB/Swissprot?*

Al seguir el enlace a cualquiera de estas bases de datos obtendremos mayor información acerca de esta, como el número de entradas presentes, fecha de actualización etc. Sin embargo, por ahora nuestro interés es el de seleccionar algunas bases de datos para realizar nuestras búsquedas. Seleccione las casillas pertenecientes a las bases de datos de “UniprotKB/Swissprot” y “UniprotKB/TrEMBL”. Cerciórese de que estas sean las únicas bases de datos seleccionadas.

**Available Databases**

[Expand all](#)  [Collapse all](#) Show databases tooltips:

- Literature, Bibliography and Reference Databases**
  - [all](#)  [MEDLINE](#)  [Taxonomy](#)  [OMIM](#)
    - [OMIM Morbid Map](#)  [Patent Abstracts](#)  [Karyn's Genomes](#)
    - [Patent Equivalents](#)
  - Literature, Bibliography and Reference Databases - subsections**
    - [all](#)  [MEDLINE \(Updates\)](#)  [MEDLINE \(Main Release 2010\)](#)  [MED2PUB](#)
- Gene Dictionaries and Ontologies**
- Nucleotide sequence databases**
  - [all](#)  [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](#)  [Patent DNA NRL2](#)
  - Nucleotide sequence databases - subsections**
    - [all](#)  [EMBL \(Updates\)](#)  [EMBL \(Release\)](#)  [EMBL \(Whole Genome Shotgun\)](#)
    - [EMBL \(Whole Genome Shotgun release\)](#)  [EMBL \(Whole Genome Shotgun\)](#)
    - [EMBL \(Contig updates\)](#)  [EMBL \(Contigs expanded\)](#)
    - [EMBL \(Release, Deleted\)](#)  [EMBL \(Whole Genome Shotgun\)](#)
    - [RefSeq Genome \(Release\)](#)  [RefSeq Genome \(Updates\)](#)
- Nucleotide related databases**
- UniProt Universal Protein Resource**
  - [all](#)  [UniProtKB](#)  [UniProtKB/Swiss-Prot](#)  [UniProtKB/TrEMBL](#)  [UniRef100](#)  [UniRef90](#)
    - [UniRef50](#)  [UniParc](#)

**Figura 3.6:** Opciones SRS

A la izquierda de su pantalla encontrará la casilla “Search Options” la cual nos permitirá seleccionar el nivel de profundidad de nuestra búsqueda, Por ser esta la primera vez que trabajamos con este sistema seleccionaremos la forma estándar de búsqueda.

Presione el botón “**Standard query Form**” de la casilla “**Search Options**”

Esta acción le llevará al formulario estándar de búsqueda en el SRS (Figura 3.7). El cual consta de 4 partes fundamentales.

**Fields you can search** Campos de búsqueda, donde podemos entrar nuestros términos de búsqueda de acuerdo a cualquiera de las opciones presentes en los respectivos menús desplegables.

**Create View** Crear vista, esta opción trabaja en conjunto con la opción 3, y acá podemos definir el tipo de campos que queremos ver en nuestra página de resultados. Para nuestro ejemplo, tenemos interés en seleccionar todas las proteínas de superficie conocidas de *Plasmodium falciparum* con actividad inmunogénica, relacionadas con el merozoito.

**Result Display Options** Opciones para mostrar los resultados, donde podemos definir el número de resultados que queremos por página, así como el formato de salida, ya sea alguno de los definidos en el menú desplegable o mediante la creación de una vista personalizada (opción “create view”).

**Search Options** Opciones de búsqueda, donde podemos definir, entre otras cosas, el tipo de conector lógico (booleano) a utilizar para los términos definidos en 1.

Defina estos criterios en la sección “Fields you can search” de acuerdo la Figura 3.8.

<b>Search Options</b> <p>Combine search terms with: &amp; (AND) <input type="checkbox"/></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: &amp;,   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;"><b>Create a view</b></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"/> UniprotView  <input type="radio"/> EntryName <input type="radio"/> Data Class  <input type="radio"/> AccessionNumber <input type="radio"/> Primary Accession Number  <input type="radio"/> Sequence Version <input type="radio"/> Creation Date</p> <p>Display As: <input type="radio"/> Table <input checked="" type="radio"/> List</p> <p>Sequence Format: <input type="button" value="swiss"/></p> <p style="text-align: right;"><input type="button" value="Search"/></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=""/>												
<b>Result Display Options</b> <p><input checked="" type="radio"/> View results using: <input type="button" value="UniprotView"/></p> <p>or</p> <p><input type="radio"/> Create a view</p> <p>Show <input type="button" value="30"/> results per page</p>													
<b>Tips</b> <p>To do more advanced queries, use the <a href="#">Extended Query Form</a>.</p>													

**Figura 3.7:** Formulario de búsqueda SRS

<b>Fields you can search</b> <p>In a single field, you can separate multiple values by: &amp;,   or !</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><input type="button" value="i"/> Organism Name</td> <td><input type="text" value="plasmodium falciparum"/></td> </tr> <tr> <td><input type="button" value="i"/> Keywords</td> <td><input type="text" value="merozoite"/></td> </tr> <tr> <td><input type="button" value="i"/> Description</td> <td><input type="text" value="surface antigen"/></td> </tr> <tr> <td><input type="button" value="i"/> AllText</td> <td><input type="text" value=""/></td> </tr> </tbody> </table>	Fields you can search	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=""/>
Fields you can search	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:** Criterios de búsqueda avanzada

A continuación presione el botón “**search**” ubicado en la parte superior de esta sección y espere unos segundos.

Seguramente en este momento ya tenga una visión más exacta de las posibilidades que ofrece el SRS y sus principales diferencias con el sistema Entrez. Primero, pudimos definir exactamente no solamente la base de datos que queríamos consultar, sino las secciones específicas de esta. Además de esto pudimos también definir exactamente los términos de búsqueda en secciones específicas de las entradas, lo cual nos da un completo control sobre los resultados que queremos obtener.

Juegue un poco con las diferentes opciones de formatos que ofrece el SRS en la sección “**Result Display Options**”, del formulario de búsqueda. Intente también creando su propio formato de salida con la opción “**Create view**”.

**Encuentre todas las proteínas nucleares hipotéticas de *Saccharomyces cerevisiae*, y muestre la información en 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).

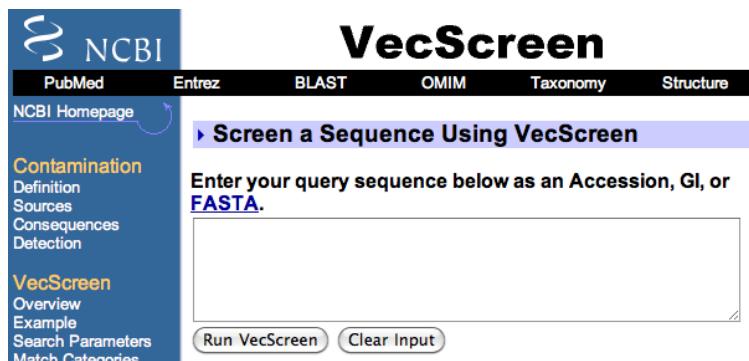


Figura 4.1: VecScreen: Herramienta para detectar contaminación de vectores.

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
ATGAGAAGCGTTCTATGATTACAATTCTTGTAGGAGGATATGTGGACACT
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 SQLite<sup>1</sup> como motor de base de datos y la extensión de Firefox SQLite Manager<sup>2</sup> 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**<sup>3</sup>. 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

---

<sup>1</sup><http://www.sqlite.org/>

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

<sup>3</sup>Como 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 archivos<sup>4</sup> tf.csv, Species.csv, Domains.csv<sup>5</sup> 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 indices<sup>6</sup>:

**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 duplicados<sup>7</sup>, 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

<sup>4</sup>Los archivos pueden ser descargados desde Sicua Plus

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

<sup>6</sup>¿Para qué sirven los indices?

<sup>7</sup>¿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. **¿Que 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.txt<sup>1</sup> 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?

---

<sup>1</sup>Disponible 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](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](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 <a href="#">view in tree</a>	<b>biological process</b>	<a href="#">IMP</a>	<a href="#">TAIR:Publication:501729812</a>	TAIR
			<a href="#">IEP</a>	<a href="#">TAIR:Publication:501736296</a>	TAIR
<input type="checkbox"/> GO:0007275 : multicellular organismal development	24103 gene products <a href="#">view in tree</a>	<b>biological process</b>	<a href="#">ISS</a> With Pfam:PF02365	<a href="#">TAIR:Communication:501714663</a>	TIGR (via TAIR)
<input type="checkbox"/> GO:0010468 : regulation of gene expression	27494 gene products <a href="#">view in tree</a>	<b>biological process</b>	<a href="#">IMP</a>	<a href="#">TAIR:Publication:501736296</a>	TAIR
<input type="checkbox"/> GO:0006979 : response to oxidative stress	2101 gene products <a href="#">view in tree</a>	<b>biological process</b>	<a href="#">IMP</a>	<a href="#">TAIR:Publication:501713092</a>	TAIR
<input type="checkbox"/> GO:0009651 : response to salt stress	537 gene products <a href="#">view in tree</a>	<b>biological process</b>	<a href="#">IEP</a>	<a href="#">TAIR:Publication:501736296</a>	TAIR
<input type="checkbox"/> GO:0010149 : senescence	104 gene products <a href="#">view in tree</a>	<b>biological process</b>	<a href="#">IMP</a>	<a href="#">TAIR:Publication:3011</a>	TAIR
<input type="checkbox"/> GO:0005634 : nucleus	37778 gene products <a href="#">view in tree</a>	<b>cellular component</b>	<a href="#">IDA</a>	<a href="#">TAIR:Publication:501718231</a>	TAIR
<input type="checkbox"/> GO:0046982 : protein heterodimerization activity	1028 gene products <a href="#">view in tree</a>	<b>molecular function</b>	<a href="#">NOT</a>	<a href="#">IPI</a> With AGI LocusCode:AT3G29035	<a href="#">TAIR:Publication:501718231</a>
<input type="checkbox"/> GO:0042803 : protein homodimerization activity	2165 gene products <a href="#">view in tree</a>	<b>molecular function</b>		<a href="#">IPI</a>	<a href="#">TAIR:Publication:501718231</a>
<input type="checkbox"/> GO:0003700 : transcription factor activity	12404 gene products <a href="#">view in tree</a>	<b>molecular function</b>		<a href="#">ISS</a>	<a href="#">TAIR:Publication:1345963</a>
				<a href="#">IPI</a>	<a href="#">TAIR:Publication:501718231</a>

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 Cytoscape<sup>1</sup>. 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)?

---

<sup>1</sup><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

# Comparación de secuencias I - Matrices de puntos

Las matrices de puntos (“Dot Plot”) son herramientas exploratorias para la comparación de cadenas de texto, i.e., secuencias. Entre otros, nos permiten fácilmente encontrar regiones repetidas en una secuencia al compararla contra si misma. También podemos hacernos una idea bastante clara de la estructura de un gen, al comparar la secuencia de su región codificante contra la secuencia del locus en donde se encuentra.

En esta sección usaremos la implementación de matrices de puntos del Instituto Suizo de Bioinformática, conocida como Dot Let<sup>1</sup>, que vemos en la Figura 10.1.

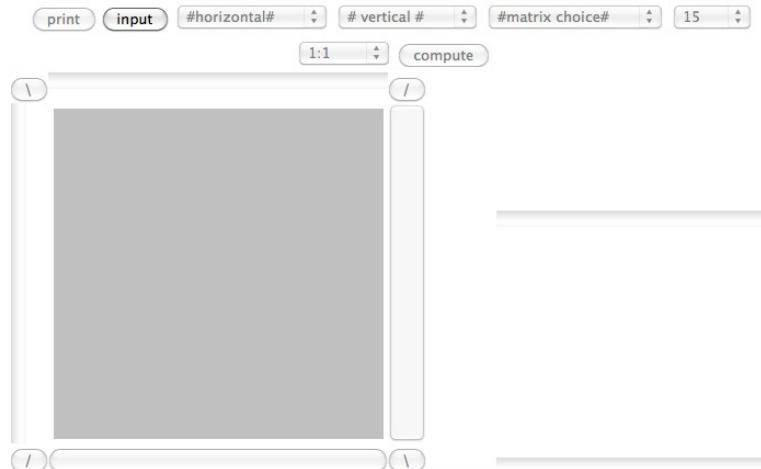
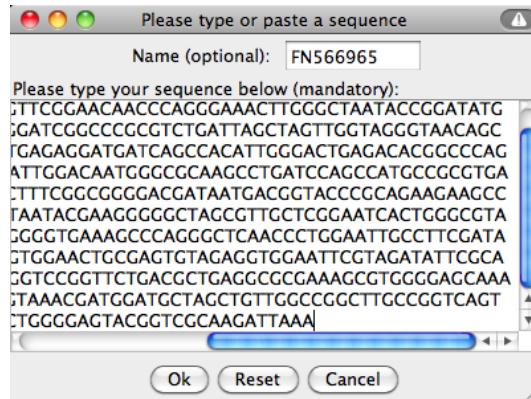


Figura 10.1: Dot Let @ SIB

Haga una comparación de la secuencia que se encuentra en el archivo aqc-MIR399 contra sí misma. Lo primero que tiene que hacer es pinchar el botón “Input”, eso abrirá la ventana que se

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

ve en la Figura 10.2, dele un nombre a la secuencia y pegue la misma en el cuadro correspondiente.



**Figura 10.2:** Agregar secuencias en Dot Let

De regreso en la ventana de Dot Let vemos que encontramos habilitados dos botones (Figura 10.3), aparecen ahora con el nombre de la secuencia que acabó de agregar. Uno de ellos representa a la secuencia que aparece en la dirección horizontal, el otro a la secuencia que aparece en la dirección vertical.

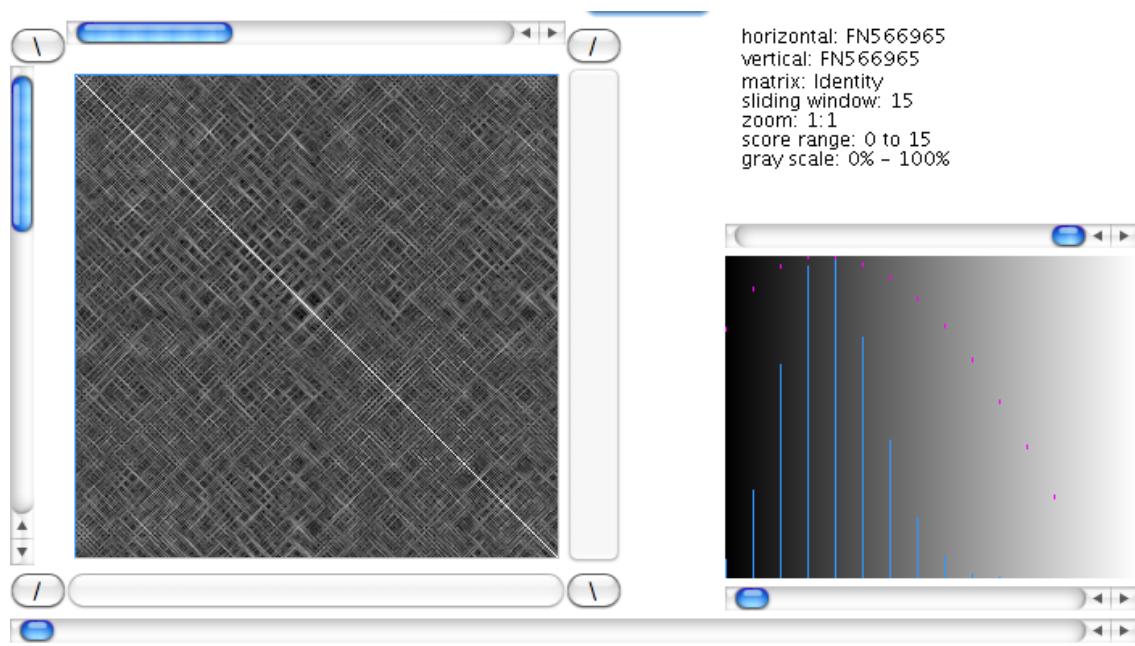


**Figura 10.3:** Botones de control

A la derecha de los botones/listas que identifican a las secuencias, encontramos un lista desplegable, desabilitada por el momento, que nos permite seleccionar la matriz de sustitución. A continuación encontramos una lista desplegable con los tamaños de ventana que se van a usar para la comparación de las dos secuencias. El siguiente botón nos permite hacer acercamientos, i.e., “Zoom”, y por último encontramos el botón “Calcular”, que llena la matriz de puntos.

Una vez se ha calculado la matriz de puntos, encontramos dos secciones de resultados, similar a como aparece en la Figura 10.4. La región de la izquierda es la matriz propiamente dicha, pixeles oscuros representan puntajes bajos, i.e., malos. A la izquierda vemos un histograma de la frecuencia de cada puntaje. Manipulando este histograma, con las barras de desplazamiento horizontal (arriba y abajo) podemos modificar la visualización de la matriz de puntos.

- Explique como afecta el tamaño de la ventana la visualización en la matriz de puntos.
- ¿Qué significado tiene la linea rosada en el histograma de puntajes?
- ¿Qué interpretación puede hacer de las repeticiones invertidas que se pueden detectar en la matriz de puntos?



**Figura 10.4:** Resultado

- Compare la secuencia de cDNA y su correspondiente genómica del ANAC092<sup>2</sup>. Describa los resultados.

---

<sup>2</sup>Disponible en Sicua Plus



# Capítulo 11

## EMBOSS

EMBOSS<sup>1</sup>, “The European Molecular Biology Open Software Suite”, es un paquete gratuito, con código fuente abierto, compuesto de cientos de aplicaciones<sup>2</sup> que se han desarrollado específicamente para resolver las necesidades de la comunidad de biología molecular. El tutorial que encuentra a continuación es una parte de los tutoriales disponibles en [http://emboss.sourceforge.net/docs/emboss\\_tutorial/emboss\\_tutorial.html](http://emboss.sourceforge.net/docs/emboss_tutorial/emboss_tutorial.html), adaptado al uso del la interfaz gráfica desarrollada por el Instituto Pasteur <http://mobyle.pasteur.fr/>.

Encuentra una descripción de cada una de las aplicaciones presentes en EMBOSS siguiendo el enlace: <http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/>

Algunas de las áreas cubiertas por aplicaciones de EMBOSS son:

- Alineamiento de secuencias
- Búsqueda en bases de datos usando patrones
- Identificación de motivos de proteínas
- Análisis de uso de codones.

### 11.1 Recuperando secuencias de bases de datos

La recuperación de secuencias desde una base de datos depende, obviamente, de las bases de datos que tengamos disponibles.

Vamos a recuperar la secuencia del gen ANC092 que esta en la base de datos de proteínas UniProt. Para poder realizar este ejercicio debe ir al sitio web de UniProt<sup>3</sup> y encontrar el identificador adecuado.

---

<sup>1</sup><http://emboss.sourceforge.net/>

<sup>2</sup><http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/>

<sup>3</sup><http://www.uniprot.org/>

Una vez a encontrado el identificador correspondiente en UniProt, puede ir a sitio de Mobyle@Pasteur que se citó anteriormente.

El programa que nos interesa se llama “seqret”. Llene los campos siguiendo las indicaciones que se muestran en el Figura 11.1

**Figura 11.1:** Recuperando secuencias de las bases de datos

Ahora, puede usar el mismo programa para convertir la entrada recuperada de UniProt en formato fasta. *¿Cómo lo haría?*

Ya que sabe recuperar secuencias de UniProt, hagamos algunos cálculos sobre esa secuencia, busque el programa compseq, que le permite calcular la composición de palabras de una secuencia. *Calcule las frecuencias de dímeros para ANAC092 extraído de UniProt*<sup>4</sup>

## 11.2 Selección de marco de lectura abierto

Los programas getorf y plotorf buscan marcos de lectura abiertos en secuencias de nucleótidos. Siendo un marco de lectura abierto, una cadena (subsecuencia) de una longitud mínima especificada flanqueada ya sea por dos codones de parada o por un codón de inicio y otro de parada. A pesar de la universalidad del código genético algunos grupos de organismo tienen codones de inicio y de parada diferentes, por esta razón es importante especificar, ya sea, el código genético que se está usando para traducir la secuencia, o los codones de inicio y parada permitidos.

Emplee estos dos programas para encontrar el marco de lectura abierto correcto de la secuencia “ANAC092\_cDNA.fa”. *¿Encuentra alguna diferencia en los resultados ofrecidos por los dos*

<sup>4</sup> Al entregar su guía estos resultados los debe enviar en un archivo de texto plano.

programas?

### 11.3 Barajar/mezclar secuencias

Al hacer ciertos tipos de análisis, e.g., búsqueda de sitios de unión a factores de transcripción en secuencias promotoras (“TFBS”), es importante contar con un grupo de secuencias que sirvan como control negativo. De forma tal que los TFBS no aparezcan de forma frecuente en ese control negativo. Una opción muy usada es la de generar secuencias aleatorias que contengan la misma composición monomérica que las secuencias originales. El programa “shuffleseq” hace precisamente esto, toma una secuencia “real”, y mezcla, como si estuviera barajando un mazo de cartas, los monómeros constituyentes, resultando en una secuencia al azar. Cuando se usa este tipo de estrategia, por cada secuencia original se generan 1000 secuencias aleatorias.

Use “shuffleseq” para generar dos secuencias aleatorias del ARNr que se encuentra en el archivo **FN566965.fasta**.

### 11.4 Predicción de regiones hidrofóbicas

El programa pepwindow predice segmentos hidrofóbicos en una proteína, siguiendo la estrategia planteada por (KYTE and DOOLITTLE, 1982). Usando ventanas de 19 a 21 residuos las regiones transmembranales se pueden detectar claramente, con valores de indice de hidrofobicidad de 1.6 en la región central.

¿Puede detectar alguna región transmembranal en el gen NTM1<sup>5</sup>?

### 11.5 Alineamientos

Describa la función del programa distmat.

---

<sup>5</sup> Archivo NTM1.fasta disponible en Sicua Plus



## Capítulo 12

# Comparación de secuencias II - Alineamientos pareados

Algunos apartes de este capítulo vienen del tutorial que se encuentra siguiendo el enlace: [http://emboss.sourceforge.net/docs/emboss\\_tutorial/emboss\\_tutorial.pdf](http://emboss.sourceforge.net/docs/emboss_tutorial/emboss_tutorial.pdf)

Para desarrollar los ejercicios siga el enlace <http://mobyle.pasteur.fr/cgi-bin/portal.py>

### 12.1 Matrices de sustitución

En el archivo EPAM250.txt encontrará la matriz de sustitución PAM250.

- ¿Quiénes, y cómo, crearon la familia de matrices de sustitución PAM?
- ¿Dónde se encuentran los puntajes mas altos? Explique.
- ¿Cuál es la sustitución con el mayor puntaje?
- ¿Por qué las identidades no tienen siempre el mismo puntaje?

### 12.2 Alineamiento Global

En el alineamiento global el objetivo es comparar las dos secuencias a lo largo de toda su longitud, por lo tanto, es apropiado cuando esperamos que la similitud entre las dos secuencias se extiende a lo largo de toda la secuencia.

En el paquete EMBOSS encontrará la aplicación needle que implementa rigurosamente el algoritmo de Needleman y Wunsch (NEEDLEMAN and WUNSCH, 1970) para obtener el alineamiento

global óptimo por programación dinámica. Esta implementación puede tomar bastante tiempo en obtener el alineamiento cuando las secuencias son largas.

¿Qué otras aplicaciones en EMBOSS le permiten hacer alineamientos globales? ¿Qué las hace diferentes de needle?

- Haga un alineamiento global entre las secuencias de cDNA y genómica del gen ANAC092, que están disponibles en el directorio de la semana pasada en Sicua Plus.
- ¿Qué matriz de sustitución y penalización para abrir y extender gaps usó? Explique.
- ¿Cuál es el puntaje del alineamiento, su longitud y los porcentajes de identidad y similitud?
- Explique la diferencia entre similitud e identidad.
- ¿Qué significan los símbolos :, . y +?

En los archivos ANAC092\_pep.fasta y PpNAC\_e\_gw1.5.134.1.fasta encuentra las secuencias de amino ácidos de dos genes de la familia NAC de factores de transcripción en *Arabidopsis thaliana* y en el musgo *Physcomitrella patens* respectivamente.

- Haga un alineamiento global entre las secuencias de amino ácidos de las proteínas NAC de *Arabidopsis thaliana* y en el musgo *Physcomitrella patens*.
- ¿Qué matriz de sustitución y penalización para abrir y extender gaps usó? Explique.
- ¿Cuál es el puntaje del alineamiento, su longitud y los porcentajes de identidad y similitud?
- ¿Puede mejorar el alineamiento escogiendo otros parámetros?

### 12.3 Alineamientos locales

Como se mencionó en la sección anterior el alineamiento global alinea a las secuencias a lo largo de toda su longitud. Usted tiene que decidir si esa estrategia es la mas adecuada en cada caso. ¿Qué cree que pasaría si compara dos proteínas multidominio que solo comparten un dominio entre ellas?

El objetivo del alineamiento local es encontrar regiones de similitud local, y no es necesario incluir las secuencias completas. Este tipo de alineamiento es muy útil para hacer búsquedas en bases de datos, o cuando no se tiene una idea clara sobre la similitud de la secuencia de interés con secuencias de la base de datos.

En el paquete EMBOSS encontrará la aplicación water que implementa rigurosamente el algoritmo de smith y Waterman (SMITH and WATERMAN, 1981) para obtener el alineamiento local óptimo por programación dinámica. Esta implementación puede tomar bastante tiempo en obtener el alineamiento cuando las secuencias son largas.

¿Qué otras aplicaciones en EMBOSS le permiten hacer alineamientos locales? ¿Qué las hace diferentes de water?

- Haga un alineamiento local entre las secuencias de amino ácidos de las proteínas NAC de *Arabidopsis thaliana* y en el musgo *Physcomitrella patens*, que usó en la sección anterior.
- ¿Qué matriz de sustitución y penalización para abrir y extender gaps usó? Explique.
- ¿Cuál es el puntaje del alineamiento, su longitud y los porcentajes de identidad y similitud?
- ¿Puede mejorar el alineamiento escogiendo otros parámetros?
- ¿Qué diferencias hay entre el alineamiento global y el local de estas dos secuencias?

## 12.4 Significancia de los alineamientos

No importa que secuencias le de a los programas de alineamiento, ellos siempre crearan un alineamiento.

Tome la secuencia de amino ácidos ANAC092 y use el programa shuffleseq, y cree dos secuencias aleatorias con la misma composición monomérica de ANAC092. Haga un alineamiento global y uno local con las dos secuencias.

- ¿Qué matriz de sustitución y penalización para abrir y extender gaps usó? Explique.
- ¿Cuál es el puntaje del alineamiento, su longitud y los porcentajes de identidad y similitud?
- ¿Puede mejorar el alineamiento escogiendo otros parámetros?

Ahora haga un alineamiento local y uno global entre la secuencia de amino ácidos de ANAC092 y una de las versiones al azar.

- ¿Qué matriz de sustitución y penalización para abrir y extender gaps usó? Explique.
- ¿Cuál es el puntaje del alineamiento, su longitud y los porcentajes de identidad y similitud?
- ¿Puede mejorar el alineamiento escogiendo otros parámetros?



# Capítulo 13

## BLAST: BASIC LOCAL ALIGNMENT SEARCH TOOL

Muchos de ustedes conocen la interfaz web de BLAST en el NCBI que se muestra en la Figura 13.2. En la primera parte de este tutorial vamos a hacer algunos ejercicios usando esta interfaz.

**Basic BLAST**

---

Choose a BLAST program to run.

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

**Figura 13.1:** Tipos de BLAST disponibles en el NCBI

En SicuaPlus encuentra el archivo `desconocido.nuc.fa`, que contiene la secuencia de nucleótidos de un transcripto que usted descubrió al analizar la expresión diferencial de genes de *A. thaliana* en respuesta a luz ultravioleta (UV-A), tratamiento en el cual este transcripto era inducido. Copie la secuencia del transcripto y abra la página <http://blast.ncbi.nlm.nih.gov/> en el navegador Firefox. Vamos a realizar una búsqueda básica de BLAST, busque en la página una sección como la que aparece en la Figura 13.1 y seleccione la opción `blastx`.  
*¿Por qué usar blastx?*

En la página de `blastx` pegue su secuencia desconocida en el campo “**Enter query sequence**”, escriba *Viridiplantae* en el campo “**Organism**”, para restringir la búsqueda a las secuencias de plantas verdes (Figura 13.2). Asegúrese de que la base de datos seleccionada sea la base de datos no redundante de secuencias de proteínas.

NCBI BLAST/blastx

Enter Query Sequence  
>desconocido  
CACACATCTACGGCCGTACCAAGACTAATCTATTCCCTCCAAAATGTCCTCCAAATTAGATTCTTCCAA  
GTTCTTCTGAAATCCCAACTCCCGCTTTCCTCTTTACCTTCAGCCTTGCTACTAACACAA  
CAAATCTTCTCTCTCTCTGCTGATCGACTTCAGACACTAACAGAACATCAAGAAAATGCAGGAAACAGCCAA  
GCTCTTCTGCTGCTGAGCTTACCATCAAGCAGCCAGAGGTCATCACGCTCTCCACATTGGAGAT  
CAAAGAACGAAATTAAAGCGATGAGAGATACGCCAGCTCCGCAACTGGAGAGACTGTCGGAAAA  
GAAAGCTCCGTAAGAAATCTGGATCCGGACCGTCAAGCCGAGACAGCGCAGCTGCGGAGAAA  
AAAGGAAGCCAGGCAAGACACCCGGCGAGAAGAACAGCCGCTAAAGGGTTGGAGAGACTGTCGGAGAAGCT  
TTCAGCTCAGCAAGCAGAGAGAAAAAGGCTTACTTGACCACTTCAAGACGAAACAGAGCTTGGAGACAT  
GAGACAAAAAAACTCTGAACCTGAAGACGCCACTTCAAGACGAAACAGAGCTTGGAGACAT  
TCTCTGAAAGAACACACAGAAACAGAGAGGTGTCGTGTTAATGCTGATGCAAGCCTTGTAC  
TCCCTCTCTCTCTGTTGTTATATTGTAATAATTACAGAAATTGTATATAAATTATTCATGTT  
AAAATTATATGGGATGTGACGCTTAATATCTAATTGACAATTCTTAAAAAAA  
AA

Or, upload file  
Genetic code  
Job Title  
desconocido  
Enter a descriptive title for your BLAST search

Choose Search Set  
Database: Non-redundant protein sequences (nr)  
Organism: Xenobiontae (taxid:33090)  
Exclude  
Optional  
Exclude  
Optional  
Entrez Query  
Default

**Figura 13.2:** Interfaz web de NCBI BLAST usando el programa blastx

En las búsquedas que involucran la traducción en línea de una secuencia de ADN de puede seleccionar el código genético que se usará para hacer la traducción. Asegúrese que el código genético seleccionado en este caso es “Standard”.

Algorithm parameters

General Parameters

Max target sequences: 100  
Select the maximum number of aligned sequences to display

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Filters and Masking

Filter: Low complexity regions (checked)

Mask: Mask for lookup table only (unchecked)

Mask: Mask lower case letters (unchecked)

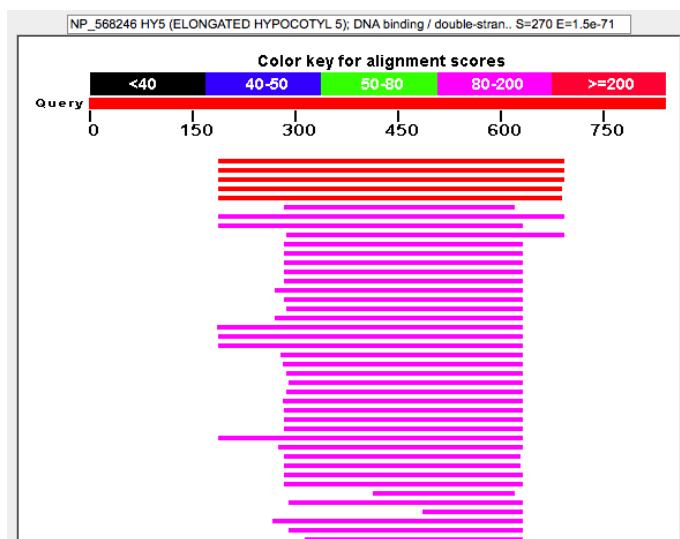
**Figura 13.3:** Parámetros de búsqueda en BLAST

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  $1 \times 10^{-5}$  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 cambian<sup>1</sup>.

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 (Accession 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

<sup>1</sup>En el enlace [http://www.ncbi.nlm.nih.gov/blast/html/sub\\_matrix.html](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	<b>UG</b>
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	<b>G</b>
ref XP_002324289.1  predicted protein [Populus trichocarpa] >...	201	9e-51	<b>UG</b>
ref XP_002308656.1  predicted protein [Populus trichocarpa] >...	200	2e-50	<b>UG</b>
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 <b>SOLIC</b> RecName: Full=Transcription factor HY5;...	179	4e-44	<b>G</b>
emb CAO44204.1  unnamed protein product [Vitis vinifera]	173	3e-42	
gb ACU17915.1  unknown [Glycine max]	170	2e-41	
gb ACF28170.1  LONG1 [ <i>Pisum sativum</i> ] >gb ACP28171.1  LONG1 [P...	170	2e-41	
gb AAC05018.1  TGACG-motif-binding factor [Glycine max]	170	2e-41	<b>G</b>
gb AAC05017.1  TGACG-motif binding factor [Glycine max] >gb A...	170	2e-41	<b>G</b>
emb CAA66478.1  bZIP transcription factor [Vicia faba var. minor]	166	5e-40	
ref XP_002453510.1  hypothetical protein SORBIDRAFT_04g007060...	165	7e-40	<b>UG</b>
dbj BAC20318.1  bZIP with a Ring-finger motif [Lotus japonicu...	163	3e-39	<b>G</b>
ref XP_002437242.1  hypothetical protein SORBIDRAFT_10g023420...	159	5e-38	<b>UG</b>
ref NP_001152483.1  transcription factor HY5 [Zea mays] >gb A...	158	9e-38	<b>UG</b>
ref NP_001046236.1  Os02g0203000 [Oryza sativa (japonica cult...	158	1e-37	
gb EFC72704.1  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 EBC80926.1  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	<b>UG</b>
gb EAY72732.1  hypothetical protein OsI_00597 [Oryza sativa I...	127	2e-28	

Figura 13.5: Listado de “Hits”

esta tabla pinchando en los nombres de las columnas.

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|024646.1| HY5 ARATH G RecName: Full=Transcription factor HY5; AltName: Full=Protein
LONG HYPOCOTYL 5; AltName: Full=bZIP transcription factor 56;
Short=AtbZip56
dbj|BAE21116.1| G HY5 [Arabidopsis thaliana]
dbj|BAE21327.1| G HY5 [Arabidopsis thaliana]
emb|CAB96661.1| G HY5 [Arabidopsis thaliana]
gb|ABF58937.1| G At5g11260 [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 MQEQTSSLAAASSLPPSSSERSSSSAPHLIEIKEGIESDEEIRRVPFEGGEAVKGETSGRES 371
Sbjct 1 MQEQTSSLAAASSLPPSSSERSSSSAPHLIEIKEGIESDEEIRRVPFEGGEAVKGETSGRES
Query 372 GSATQERTQATVGEESQRKRGRTPAEKENRKLKLLRNRRVSQAQCARERKKAYLSELENRV 551
Sbjct 61 GSATQERTQATVGEESQRKRGRTPAEKENRKLKLLRNRRVSQAQCARERKKAYLSELENRV 120
Query 552 KDLENKNSELEERLSTLQNENQMLRHILKNTTGNKrgggggSNADASL 695
KDLENKNSELEERLSTLQNENQMLRHILKNTTGNKRGGGGSNADASL
Sbjct 121 KDLENKNSELEERLSTLQNENQMLRHILKNTTGNKRGGGGSNADASL 168
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://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.5_Cookbook)

- [http://www.ncbi.nlm.nih.gov/books/NBK1763/#CmdLineAppsManual.4\\_User\\_manual](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

# Alineamientos múltiples

En teoría los algoritmos de programación dinámica que se describieron anteriormente para el caso de alineamientos pareados se pueden extender para el caso de un número arbitrario de secuencias. En la práctica, esto resulta muy costoso computacionalmente, por lo que se han desarrollado otros algoritmos que implementan atajos en la búsqueda de los alineamientos óptimos (heurísticas). El desarrollo de algoritmos para el alineamiento múltiple de secuencias es una de las áreas más dinámicas de la bioinformática. Actualmente existen decenas de programas que implementan diferentes algoritmos (vea NOTREDAME, 2007 y LEMEY *et al.*, 2009 para una revisión reciente del tema).

En esta sesión vamos a desarrollar la práctica que se presenta en el capítulo 3 de LEMEY *et al.*, 2009.

En este ejercicio vamos a alinear las secuencias de los genes TRIM5 $\alpha$  de diferentes especies de primates. TRIM5 $\alpha$  es un factor de restricción viral que protege a la mayoría de monos del viejo mundo (Cercopithecidae) de la infección con HIV. Estos datos fueron analizados originalmente por SAWYER *et al.*, 2005. Vamos a usar métodos de alineamiento progresivo (CLUSTALX), basado en consistencias (T-COFFEE) y de refinamiento iterativo (MUSCLE) para crear alineamientos múltiples de proteínas y luego vamos a comparar los resultados usando el servidor web ALTAVIST. Vamos a crear los alineamientos de las secuencias de las proteínas, vamos a comparar los diferentes métodos y vamos a generar el alineamiento correspondiente a nivel de nucleótidos, terminando con la inspección y refinamiento manual del alineamiento.

## 14.1 Alineando las secuencias de amino ácidos de TRIM5 $\alpha$ de primates

### 14.1.1 CLUSTALX

Decargue el archivo `primatesAA.fasta` del SicuaPlus, este archivo tiene 22 secuencias en formato fasta. Inicie el programa CLUSTALX que se encuentra instalado localmente en su computador (THOMPSON *et al.*, 1994). En CLUSTALX abra el archivo con las secuencias usando `File → Load sequences`. La interfaz gráfica le permite al usuario navegar sobre las secuencias. Seleccione `Alignment → Do complete alignment`. CLUSTALX lleva a cabo el alineamiento progresivo y crea como salidas el árbol guía (con extensión `dnd`) y el alineamiento en formato Clustal (con extensión `aln`). Es posible escoger un formato diferente de salida para el alineamiento, siguiendo el menú `Alignment → Output Format Options`, seleccionando por ejemplo el formato `PHYLIP` que puede ser leído por muchos paquetes para inferencia filogenética.

Recuerde que el árbol guía construido por CLUSTALX no debe ser usado para sacar conclusiones sobre las relaciones evolutivas entre los grupos que se están estudiando.

CLUSTALX también permite cambiar algunos de los parámetros del alineamiento (`Alignment → Alignment Parameters`). Desafortunadamente no existen reglas generales para escoger el mejor conjunto de parámetros en un caso particular. La mejor opción consiste en ensayo y error. Si un alineamiento tiene, por ejemplo, muchos gaps largos, el usuario podría intentar incrementar la penalización para abrir gaps y re-hacer el alineamiento. CLUSTALX indica el grado de conservación, revise la parte de abajo de la ventana del alineamiento,, ese nivel de conservación puede ser usado para evaluar el alineamiento. Seleccionando `Quality → Calculate Low Scoring Segment1` y `Quality → Show Low Scoring Segment`, es posible visualizar aquellas regiones del alineamiento que no son confiables, que podrían removese o refinarse manualmente. Salve el alineamiento usando `File → Save Sequences` y selecciones FASTA como formato de salida, asegurese de cambiar el nombre del archivo de salida para evitar sobre-escribir el archivo original.

### 14.1.2 T-COFFEE

Aunque el programa T-COFFEE se puede instalar localmente en sus computadores, para desarrollar este ejercicio vamos a usar un servidor web. Siga el enlace <http://www.tcoffee.org/>. Seleccione el formulario de envío regular para T-COFFEE, allí cargue su archivo, usando `Upload File` o pegue las secuencias en la caja de texto de envío. Asegurese que la opción `Computation mode` está en `regular`. Recuerde que los alineamientos con T-COFFEE son mas costosos desde el punto de vista computacional, ya que usa un método basado en consistencias, y por lo tanto puede tomar mas tiempo que un alineamiento con CLUSTALX (NOTREDAME *et al.*, 2000). Presiona el botón

---

<sup>1</sup>Siga el enlace <http://bips.u-strasbg.fr/fr/Documentation/ClustalX/> para entender como se calculan los segmentos de bajos puntaje.

Submit para ejecutar el programa. Cuando el procedimiento de alineamiento halla terminado sera direccionado a una nueva página con enlaces a los archivos de salida. Salve el alineamiento en formato fasta en su computador. El archivo `score_pdf` contiene una versión del alineamiento coloreada dependiendo de la calidad, este archivo también está disponible en formato HTML.

#### 14.1.3 MUSCLE

Para obtener el alineamiento por el método de refinamiento iterativo vamos a usar el programa MUSCLE (EDGAR, 2004). Este programa también puede ser instalado localmente en su computador, en esta ocasión vamos a usar el servidor web disponible en el EBI, siga el enlace <http://www.ebi.ac.uk/Tools/muscle/index.html>, allí puede cargar su archivo de secuencias usando la opción `Upload a file`. Asegurese que la salida aparecerá en formato Fasta. Guarde el archivo de salida en formato fasta en su computador.

#### 14.1.4 Comparar los alineamientos usando la herramienta web ALTAVIST

Para evaluar los diferentes alineamientos vamos a usar la herramienta ALTAVIST, siga el enlace <http://bibiserv.techfak.uni-bielefeld.de/altavist/>. ALTAVIST compara dos alineamientos alternativos y usa códigos de color para indicar concordancias y conflictos. Las regiones donde ambos alineamientos coinciden generalmente se consideran más confiables que las regiones en donde difieren. Esta misma forma de razonamiento se aplica de forma similar al análisis de grupos en árboles filogenéticos que han sido reconstruidos usando varios algoritmos. Grupos presentes en árboles obtenidos con diferentes métodos son más confiables que aquellos que no están presentes de forma consistente.

Human	MASGILVNKEEVTCPICLELLTQPLSLDCGHSPFCQACLTANHKKSMYDK-GESSCPVCR
Chimp	MASGILVNKEEVTCPICLELLTQPLSLDCGHSPFCQACLTANHKKSMYDK-GESSCPVCR
Gorilla	MASGILVNKEEVTCPICLELLTQPLSLDCGHSPFCQACLTANHKKSMYDK-GESSCPVCR
Orangutan	MASGILVNKEEVTCPICLELLTQPLSLDCGHSPFCQACLTANHKKSMYDK-GESSCPVCR
Gibbon	MASGILVNKEEVTCPICLELLTQPLSLDCGHSPFCQACLTANHKTSMYDE-GERSCPVCR
Rhes_cDNA	MASGILLVNKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMYKEERSCPVCR
Baboon	MASGILLVNKEEVTCPICLELLTEPLSLPCGHSPFCQACITANHRRSMYKEERSCPVCR
AGM	MASGILLVNKEEVTCPICLELLTEPLSLPCGHSPFCQACITANHKESMLYKEERSCPVCR
AGM_cDNA	MASGILVNKEEVTCPICLELLTEPLSLPCGHSPFCQACITANHKESMLYKEERSCPVCR
Tant_cDNA	MASGILLVNKEEVTCPICLELLTEPLSLPCGHSPFCQACITANHKESMLYKEERSCPVCR
Patas	MASGILLVNKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMYKEERSCPVCR
Colobus	MASGILVNKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMYKEERSCPVCR
DLangur	MASGILVNKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMYKEERSCPVCR
PMarmoset	MASRILVNKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKESTLHQ-GERSCPLCR
Tamarin	MASRILVNKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKESTLHQ-GERSCPLCR
Squirrel	MASRILGSNIKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKESTLHQ-GERSCPLCR
Owl	MASRILVNKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKKSMYHQ-GERSCPLCR
Titi	MASRILVNKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKESTLHQ-GERSCPLCR
Saki	MASRILMNKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKKSMYHQ-GERSCPLCR
Woolly	MASEILVNKEEVTCPICLELLTEPLSLDCGHSPFCQACITADHKESTLHQ-GERSCPLCR
Howler	MASKILVNKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKESTLHQ-GERSCPLCR
Spider	MASEILLNIKEEVTCPICLELLTEPLSLDCGHSPFCQACITANHKESTLHQ-GERSCPLCR

**Figura 14.1:** Resultados de la comparación de alineamientos con ALTAVIST

Para llevar a cabo el análisis siga el enlace anterior y seleccione `OPTION 2: Enter two different pre-calculated alignments of a multiple sequence set.` Cargue sus

dos alineamientos, empiece por el de CLUSTALX y el de MUSCLE, agregue los títulos correspondientes y presione el botón Submit. En la página de resultados aparecen los dos alineamientos con los residuos coloreados. La Figura 14.1 muestra una de las regiones comparadas. Cuando todos los residuos en una columna son del mismo color, como es en la mayoría de los casos en la Figura 14.1, los residuos fueron alineados de la misma forma en los dos alineamientos. si un residuo tiene un color diferente, e.g., la arginina "R" para "Howler" en la posición ±50, está alinado de forma diferente en el segundo alineamiento.

diferentes colores se usan para distinguir diferentes grupos de residuos donde los alineamientos coinciden dentro de los grupos pero no entre grupos. Un ejemplo de esto se puede ver en la Figura 14.2: los residuos GSLT estan alineados por tanto por CLUSTALX como por MUSCLE de la misma forma en las AGM, AGM\_cDNA y Tant\_cDNA, pero no con los mismos residuos en otros organismos. De forma similar ocurre para los residuos TFPS y SFPS en Rhes\_cDNA y Baboon, respectivamente; eso explica por que tienen color diferente.

Human	VDVTVPNISCAVISEDKRQVSSPKPQIIYGARGTRYQTFV-----
Chimp	VDVTVPNISCAVISEDMRQVSSPKPQIIYGAQGTRYQTFM-----
Gorilla	VDVTVPNISCAVISEDMRQVSSPKPQIIYGAQGTRYQTFM-----
Orangutan	VDVTVPNDISYAVISEDMRQVSCPEPQIIYGAQGTTYQTYV-----
Gibbon	VDVTVPNISYAVISEDMRQVSSPEPQIIFEAQGTISQTFV-----
Rhes_cDNA	VDVTLATNNISHAVIAEDKRQVSSRNPOIMYQAPGTLFTFPS-----
Baboon	VDVTLPNNISHAVIAEDKRQVSSRNPOQITYQAPGTLFSFPS-----
AGM	VDVTLPNNISHAVIAEDKRQVSYQNPQIMYQAPGSSFGSLTNFNYCTGVLGQSITSRK
AGM_cDNA	VDVTLPNNISHAVIAEDKRQVSYRNPOQIMYQSPGSFLGSLTNFNSYCTGVPGQSITSRK
Tant_cDNA	VDVTLPNNISHAVIAEDKRQVSYQNPQIMYQAPGSSFGSLTNFNYCTGVLGQSITSRK
Patas	VDVTLPNNISHVIAEDKRQVSSRNPOIMYWAQGKLFQSLK-----
Colobus	VDVTLPNNISHAVIAEDKRQVSSPNPOIMYRAQGTLFQSLK-----
DLangur	VDVTLPNNISHAVIAEDKRQVSPSPNPOIMYCRARGTLFQSLK-----
PMarmoset	AHVTLPSPSCTVISEDERQVRYQ-VPI-HQPLV-----
Tamarin	AHVTLPSPHSYAVISEDERQVRYQ-FQI-HQPSV-----
Squirrel	AHVTLPSPHSYTIISEDGRQVRYQ-KPI-RHLLV-----
Owl	AHVTLPSPHSCTVISEDERQVRYQ-KRI-YQPLF-----
Titi	AHVTLVASHPSRAVISEDERQVRYQ-EWI-HQSSG-----
Saki	VHVTLPSPHSACVISEDERQVRYQ-ERI-HQSF-----
Woolly	AHVTLPSPHSACVISEDQRQVRYQ-KQR-HRPSV-----
Howler	AHVTLPSPSCTVISEDKRREVRYQ-EQIHHHPSM-----
Spider	AHVTLPSPSCTVISEDERQVRYQ-EQI-HQPSV-----

**Figura 14.2:** Resultados de la comparación de alineamientos con ALTAVIST

Este tipo de análisis revela que algunos bloques de discrepancias en el alineamiento están concentrados cerca de las regiones con gaps. Por lo tanto, si las regiones con gaps se van a remover<sup>2</sup>, es recomendable también borrar las regiones vecinas que tienen alineamientos ambiguos. **Haga todas las comparaciones entre los tres alineamientos que obtuvo anteriormente y describa brevemente las diferencias que observa. ¿Qué métodos de alineamientos presentan las mayores diferencias entre si? ¿Qué métodos las menores? Explique su respuesta**

#### 14.1.5 Del alineamiento de proteínas al de nucleótidos

Los programas de alineamiento múltiple que se han mencionado generalmente rompen el marco de lectura cuando se alinean secuencias de ncleótidos. Esto invalidaría, por ejemplo, análisis posteriores

<sup>2</sup>Esto se suele hacer en los análisis filogenéticos

basados en codones, y necesitarían de una buena dosis de edición manual para restaurar los marcos de lectura. Para evitar esto, podemos usar el alineamiento basado en las secuencias de proteínas para generar el alineamiento correspondiente de nucleótidos. Este procedimiento también tiene la ventaja de que los alineamientos de proteínas son menos ambiguos y más rápidos de calcular.

Vamos a crear el alineamiento de las secuencias de ADN de TRIM5 $\alpha$  usando la herramienta RevTrans (WERNERSSON and PEDERSEN, 2003). Siga el enlace <http://www.cbs.dtu.dk/services/RevTrans/> y cargue el archivo con las secuencias de nucleótidos y uno de los alineamientos de proteínas. Revise que la opción Match DNA and peptide sequences by sea igual a Name.

¿Qué característica tienen los gaps en el alineamiento de las secuencias de ADN?<sup>3</sup>

#### 14.1.6 Editando y visualizando alineamientos

Vamos a usar la aplicación JalView que está instalada en sus computadores para visualizar los alineamientos.

Siga el enlace <http://www.jalview.org/examples/editing.html>, si lo desea puede revisar la documentación completa en el enlace <http://www.jalview.org/help.html>. Identifique las regiones ambiguas del alineamiento (use la información que obtuvo al usar ALTAVIST y eliminelas. Muestre pantallazos de dos regiones que identifique como ambiguas. Explique por qué las seleccionó y qué región(es) descartó.

#### 14.1.7 Estimando distancias entre las secuencias

Cuando el objetivo es inferir las relaciones evolutivas entre las secuencias (o los organismos representados por las secuencias) a partir del alineamiento múltiple y usando un método de distancia<sup>4</sup>, es necesario estimar una distancia evolutiva entre ellas. Hay muchos modelos para estimar las distancias evolutivas entre las secuencias, ya sean estas de ADN o de proteínas, por lo tanto es necesario escoger aquel modelo que mejor se ajuste a los datos (POSADA and CRANDALL, 2001; SULLIVAN and JOYCE, 2005). Se han desarrollado aplicaciones para automatizar y ayudar en la selección del mejor modelo siguiendo estrategias estadísticas; para secuencias de ADN existe MODELTEST (POSADA and CRANDALL, 1998; POSADA, 2006) y para secuencias de proteínas PROTTEST (ABASCAL *et al.*, 2005).

Para ahorrarr tiempo, para este ejercicio vamos solo a estimar la distancia  $p$  entre las secuencias que corresponde a la proporción de posiciones diferentes entre cada par de secuencia. En la realidad la distancia  $p$  sub-estima el número real de sustituciones, ya que ignora el fenómeno de hits múltiples, i.e., que una posición dada puede haber sido sustituida mas de una vez. Los modelos

<sup>3</sup>La característica se deriva de usar el alineamiento de proteínas como guía.

<sup>4</sup>Hay cuatro grandes familias (métodos) de inferencia filogenética: métodos de distancia, de máxima verosimilitud, bayesianos y de máxima parsimonia.

que se mencionaron anteriormente implementan diferentes tipos de correcciones para tener esto en cuenta.

Vamos a usar la aplicación `protdist` de `PHYLIP`, un paquete para la inferencia de relaciones evolutivas, que está disponible en <http://mobyle.pasteur.fr/>. Siga el enlace anterior y busque la aplicación `protdist`, seleccione la opción `File` y cargue el archivo con su alineamiento múltiple de proteínas. Seleccione `Similarity table` para el parámetro `Distance Model`, los demás parámetros los usamos con sus valores por defecto. Los valores en la matriz resultante indican la proporción de posiciones idénticas para cada par de secuencias. Para obtener la distancia, i.e., la proporción de posiciones diferentes entre cada par de secuencias, solo tenemos que restar el valor de la matriz a la unidad. Otra opción podría ser usar, por ejemplo, el programa `MEGA4` (TAMURA *et al.*, 2007)<sup>5</sup>, que puede calcular directamente la proporción de posiciones diferentes.

Identifique el par de primates más cercanos y el par más lejanos. ¿Cuáles son los valores en la matriz de distancia en los dos casos?

---

<sup>5</sup><http://www.megasoftware.net/>

# Capítulo 15

## PSSMs, Logos de secuencias y HMMs

### 15.1 PSSM

Esta sección es una versión modificada del tutorial que se encuentra siguiendo el enlace [http://rsat.ulb.ac.be/rsat/tutorials/tut\\_PSSM.html](http://rsat.ulb.ac.be/rsat/tutorials/tut_PSSM.html).

Las “Position-specific scoring matrices” (PSSM) ofrecen una forma sensible de representar la variabilidad en un alineamiento. Las PSSMs se construyen tomando como base el alineamiento múltiple, e.g., de sitios de unión de factores de transcripción.

A continuación se muestra una matriz que fue obtenida de la base de promotores de *Saccharomyces cerevisiae*<sup>1</sup> y se construyó usando un alineamiento de 12 sitios de unión del factor de transcripción Pho4p de levadura.

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 fila representa un residuo (A, C, G o T) y cada columna una posición en el conjunto de secuencias alineadas. Algunas posiciones están perfectamente conservadas en todas las secuencias, mientras que otras presentan algunas alternativas.

Cuando se usa este tipo de matrices para hacer búsquedas, las posiciones mas conservadas imponen restricciones mas fuertes que aquellas en que cualquier residuo se puede presentar.

Siga el enlace <http://rulai.cshl.edu/cgi-bin/SCPD/getfactor?ABF1,BAF1> y responda:

- ¿Cuál es el tamaño del alfabeto?
- ¿Cuál es el ancho de la matriz?

---

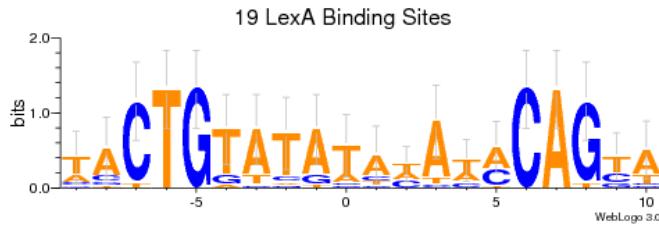
<sup>1</sup><http://rulai.cshl.edu/SCPD/>

- ¿Cuántos sitios de unión de Abf1p están almacenados en la base de datos de promotores de levadura (SCPD)?
- ¿Qué programa(s) de EMBOSS podría usar para realizar búsquedas con PSSMs?

## 15.2 Logos de secuencias

Los logos de secuencias son una representación gráfica de los alineamientos múltiples basados en la teoría de la información<sup>2</sup>

La Figura 15.1 corresponde al logo de secuencias de sitios de unión del factor de transcripción LexA de *Escherichia coli*



**Figura 15.1:** Logo de secuencias de los sitios de unión de LexA

La altura del residuo está correlacionada con su frecuencia en el alineamiento múltiple<sup>3</sup>.

Siguiendo el enlace <http://weblogo.threethreeplusone.com/>, cree el logo de secuencias de los sitios de unión del factor de transcripción Abf1p que estudió en la sección anterior. Para poder realizar este ejercicio necesita recuperar todos los sitios de unión de Abf1p disponibles en SCPD.

¿Qué representa el eje y? i.e., ¿Cómo se calcula el contenido de información de cada posición?

## 15.3 Modelos Ocultos de Markov: HMMs

Vamos a seguir algunos de los ejemplos/ejercicios que se encuentran siguiendo el enlace [http://www.mrc-lmb.cam.ac.uk/rbw/text/bioinfo\\_tuto/structure.html](http://www.mrc-lmb.cam.ac.uk/rbw/text/bioinfo_tuto/structure.html)

Incluso si no encuentra proteínas homólogas al realizar una búsqueda con BLAST, todavía tiene otras opciones.

La principal razón por la cual no encuentra homólogos triviales es que las búsquedas con secuencias usando herramientas como BLAST tienen poca sensibilidad. BLAST normalmente no encuentra proteínas homólogas que tengan menos de 30% de identidad. Sin embargo, algunas proteínas pueden tener la misma estructura tridimensional y tener solo 10% de identidad.

<sup>2</sup>[http://www.ccrnp.ncifcrf.gov/~toms/glossary.html#sequence\\_logo](http://www.ccrnp.ncifcrf.gov/~toms/glossary.html#sequence_logo)

<sup>3</sup>Para mayor información consulte: SCHNEIDER and STEPHENS, 1990

Una estrategia muy útil para encontrar homólogos distantes está basada en el uso de Modelos Ocultos de Markov (HMMs). Un HMM no es mas que una forma de definir motivos o dominios.

Para crear un HMM todo lo que necesita es un alineamiento múltiple, el cual será usado para crear una representación probabilistica, que puede ser luego usada para buscar secuencias relacionadas.

Las bases de datos Pfam (FINN *et al.*, 2010) y SUPERFAMILY (WILSON *et al.*, 2009) son colecciones de alineamientos múltiples para los cuales se han creado HMMs y son usadas para anotar secuencias de proteínas. La mayoría del trabajo de los curadores de esas bases de datos es crear los alineamientos múltiples.

### 15.3.1 Buscando los dominios de una proteína

Vamos a usar la siguiente proteína para hacer una búsqueda en Pfam:

```
proteína desconocida
>seq
MEYWHYVETTSSGQPLLREGEKDIFIDQSVGLYHGKSILQRQRGRIFLTSQRIIYIDDAKPTQ
NSLGLELDDLAYVNYYSSGFLTRSPRLILFKDPSSKDELGKSAETASADVSTWVCPICMVSNETQGEFTKD
TLPTPICINCGVPADYELTKSSINCSNAIDPNANPRNQFGVNSENICPACTFANHPQIGNCEICGHRLPNAS
KVRSKLNRLNFHDSRVHIELEKNSLARNKSSHSALENNNTGSSTEVQLSFRKSDGVLFQSATERALENIL
TEKNKHIFN
```

Vaya al sitio web de Pfam<sup>4</sup> y seleccione “Sequence search”, pega la secuencia de la proteína de interés en la caja de texto para la búsqueda, luego haga clic en el botón “go” para iniciar la búsqueda. La Figura 15.2 muestra la página de resultados.

Significant Pfam-A Matches													
Show or hide all alignments.													
Family	Description	Entry type	Clan	Envelope		Alignment		HMM	Bit score	E-value	Predicted active sites	Show/hide alignment	
<a href="#">Vps36_ESCRT-II</a>	Vacuolar protein sorting protein 36 Vps36	Family	<a href="#">CL0266</a>	8	96	8	95	1 <b>91</b>	106.7	3.1e-31	n/a	<a href="#">Show</a>	
Insignificant Pfam-A Matches													
Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From To	Bit score	E-value	Predicted active sites	Show/hide alignment	
<a href="#">Rubredoxin</a>	Rubredoxin	Domain	<a href="#">CL0045</a>	93	130	97	127	<b>13    44</b>	15.7	0.0077	n/a	<a href="#">Show</a>	
<a href="#">zf-Sec23_Sec24</a>	Sec23/Sec24 zinc finger	Domain	n/a	113	133	116	132	<b>23    39</b>	14.7	0.013	n/a	<a href="#">Show</a>	
<a href="#">zf-RanBP</a>	Zn-finger in Ran binding protein and others	Family	n/a	116	130	118	128	<b>5    15</b>	9.9	0.32	n/a	<a href="#">Show</a>	
<a href="#">Cas_CXXC_CXXC</a>	CRISPR-associated protein (Cas_CXXC_CXXC)	Domain	n/a	136	191	140	189	<b>5    60</b>	11.1	0.26	n/a	<a href="#">Show</a>	

Figura 15.2: Resultados de una búsqueda en Pfam

- ¿Cuál es la diferencia entre “Significant Pfam-A matches” e “Insignificant Pfam-A matches”?
- ¿Qué es Pfam-A?
- ¿Qué es Pfam-B?

<sup>4</sup><http://pfam.sanger.ac.uk/>

La primera sección de resultados “Significant Pfam-A matches”, nos informa que hay un solo hit, al modelo “Vps36\_ESCRT-II”, con un puntaje de 106.7 y un e-value de  $3.1e - 31$ . También encontramos las coordenadas del dominio, can relación a la proteína de consulta y con relación al modelo.

Una de las principales características, y valores agregados de Pfam, es que cada uno de los modelos en Pfam-A ha sido estudiado por un experto que ha definido una serie de puntajes límite para definir hits significativos. El puntaje umbral mas importante corresponde al “gathering cutoff”.  
**¿Cuál es el Gathering cutoff del modelo “Vps36\_ESCRT-II”?**

De clic en el nombre del modelo. Esto lo llevará a una página con información detallada sobre ese modelo particular. Entre otros, puede encontrar en que otras especies esta presente ese modelo (“Species”). Puede descargar el modelo (“Curation”) o el alineamiento múltiple (“Alignments”), entre otra información.

**Use la secuencia de la proteína ANAC092 y determine que dominios están presentes**

### 15.3.2 Visualización de HMMs

También puede visualizar los HMMs en forma de logos de secuencias. Puede usar la aplicación LogoMat-M que encuentra en el enlace <http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi>.

**Muestre el logo del dominio “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](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 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:

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

donde:

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

**$\Delta S$  (kcal/mol)** :  $S$  es la cantidad de desorden de un sistema, recibe el nombre de entropía.  $\Delta S$  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.

**$\Delta S$  (corrección por sales)** :  $\Delta S(1M\text{NaCl}) + 0.368N\ln([Na^+])$ , donde  $N$  es el número de pares de nucleótidos en el primer, 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  $T_m$

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

donde,

**$T_m(\text{primer})$**  Es la temperatura de fusión de los primers

**$T_m(\text{product})$**  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

QUANTPRIME<sup>1</sup> 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 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

---

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

The screenshot shows the 'Project details' window with the following fields:

- Create new project**
- Project name:** [Text input field]
- Organism:** [- Select an organism -] (dropdown menu)
- Annotation:** [- Select an organism first -] (dropdown menu)
- Quantification protocol:**
  - SYBR Green real-time qPCR (accept splice variant hits)
  - SYBR Green real-time qPCR (no splice variant hits)
  - End-point sqPCR (accept splice variant hits)
  - End-point sqPCR (no splice variant hits)
  - Custom
- Create** button

**Figura 16.1:** Creación un proyecto en QUANTPRIME

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**.

The screenshot shows the 'Transcripts' window with the following sections:

- Enter transcript identifier(s) manually below, or (BLAST for identifier)**
- List of identifiers: AT1G10240.1, AT1G52520.1, AT1G76320.1, AT1G80010.1, AT2G27110.1
- Add to project** button
- Transcripts in project:**
  - Select all
  - Clear selection
  - AT1G10240.1
  - AT1G52520.1
  - AT1G76320.1
  - AT1G80010.1
  - AT2G27110.1
- Find primers** and **Delete** buttons

**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.

Fw sequence	Rev sequence	Amplicon size	Spans exon border	Rank score	Test results
<b>AT2G20825.1</b>					
TGGGGGAAGGAAATTACCTTGAGG	TAAGGGCTTAACTTTCTTTGGAG	149	Yes	9.7442	
TCTGGGAAGGAAATTACCTTGAGG	TCTTCGAAAGAAAGGGAGCTGC	125	Yes	2.9246	
AAACTGAACTTCTATCGGGAGAG	TTTTTGAAAGAAAGGGAGCTGC	142	Yes	3.0588	
TAATGTGGAGTTCAATTCTGGAGAG	TCTTCGAAAGAAAGGGAGCTGC	148	Yes	9.1682	
ATUTGGGGAGGAGCATTCGTTGGAG	TCTTCGAAAGAAAGGGAGCTGC	136	Yes	3.6674	
AGTTCCATCTGGGGAGCAGTACG	TCTTCGAAAGAAAGGGAGCTGC	141	Yes	3.701	
CATTTGGGGAGAACGAACTTCGAGG	TCTTCGAAAGAAAGGGAGCTGC	137	Yes	3.7039	
ATCTTGTACGGTGGAGGTCTTGTG	TCTTCGAAAGAAAGGGAGCTGC	127	Yes	3.8067	
GGGGGGGGGGGGGGGGGGGGGGGG	TCTTCGAAAGAAAGGGAGCTGC	126	Yes	3.2586	
GAAAAGGGGGGGGGGGGGGGGGGG	AAGGGGGGGGGGGGGGGGGGGGG	107	Yes	9.3935	
<b>AT4G28190.1</b>					
CACCTGGCTGGATGAAAGAAG	TGGGGTTTTCTTCGAACTTC	68	Yes	4.5419	
TACGGCTGGATGCTGGTCTGAGC	TGGAGCTGGGGTCTGAACTTC	105	Yes+	2.5127	
ATAAGGGCTGGATGCTGGTCTGAGC	TGGAGCTGGGGTCTGAACTTC	106	Yes+	3.5905	
GGGATGGTCTGGTCTGAGCTTAA	TGGAGCTGGGGTCTGAACTTC	102	Yes+	4.212	
GGGGGGGGGGGGGGGGGGGGGGGG	TGGAGCTGGGGTCTGAACTTC	102	Yes+	4.121	
CGGGGGGGGGGGGGGGGGGGGGGG	TGGAGCTGGGGTCTGAACTTC	102	Yes+	4.1732	
CGGGGGGGGGGGGGGGGGGGGGGG	TGGAGCTGGGGTCTGAACTTC	101	Yes+	4.2257	
AGGGGGGGGGGGGGGGGGGGGGGG	TGGAGCTGGGGTCTGAACTTC	124	Yes+	5.1849	
CGACACTCTGGGGGGGGGGGGGG	TGGAGCTGGGGTCTGAACTTC	70	Yes+	2.8282	
ACCTACCTGGGGGGGGGGGGGG	ACTTCCTGGGGGGGGGGGGGG	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: TGTGGGAGACGATTACGTCGAG (22 b)  
Melting temperature: 62.2 °C  
G/C content: 54.5 %

**Reverse primer**  
Sequence: CAGCGTCAACTTGTCTTCGAG (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**  
GAT CCT AT AT A T A T C T C A T A G G A G G A C A A G G A G C C T T C G T C G T G G T T T A C T T G G T C C G A T C G G A G A T G G A G A G  
A G A A T T G G G T C G A A G G A G T T G T C A N C C A A G A G G A G C T A C A A G A A T T A G T G G A G T C C A T G T G G Q G R C G A T T A C G T C G  
R G T C A T G T G G C T G A C C A G C G C A C C G T T A C G G G A G A G C C G T T G C G A G G C T T A G A G A T T T T C A G A T G G G A G A C T T C A A  
A T C A C T T C C A A T G C A C T C T G T G A ^ A G R C A B A G T G A C G C C T G C T G C C T C G A G A G C A T T C T G A G A G A G A A A  
C C T C A G A A A C T C G A G G A A C A A T G T T G T T C A T T G A A G G A G C A A G G T C C G C T T C A A A B G A T A G T G T G C T C A G A  
T A C T A C A C A A A G C A T T G A B G A A C T T C A A C T T A C T G A A A G C T A T T C A T C G G G A T G A G T T G T G G G G T C A G C A C A T G C G G  
G A A G G A G G A G G T T C A G A T T G A G G A G C A G A G G G G A A T G C C G G A T G C A C C A T G A T G C A A T T G T G A G C C T A A T T G G A A G T  
G T T G C G A A T T A C C C A T A C G A C A A G A M A A C A T T G C G A G G A G G A G A A A G A G G G G A G C A G G G A A A G T G T T C A G A G G T T G C A C T  
C G G T C A C C G G T C T G C A A A G G A G C A G G A C T T T C T G C G T C G G G T C G C A A G G T C T G C G C T T T C T G A T T G T A A C T G C C A  
G A C T T G C C T G G A T T T C A C C A C C A M G C C A A A C C C A T T T G & T T C A T T T C T C A A C T T A T T T A A T T C T C T G C A A A A A C C T  
G T A A T G T A T G C T G C A T T T T C T T A A T C A G T A G C T T G T A C A C C T G T

**Specificity test results**  
Overall score: **Perfect**  
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

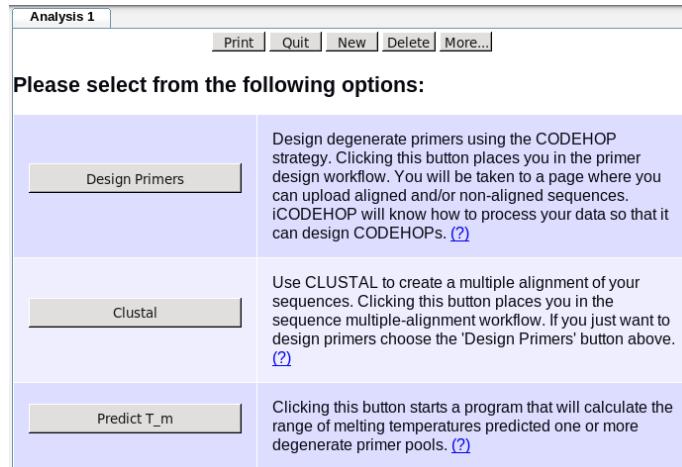
En esta sección vamos a usar el programa iCODEHOP<sup>2</sup> para diseñar primers a partir de un alineamiento de proteínas.

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

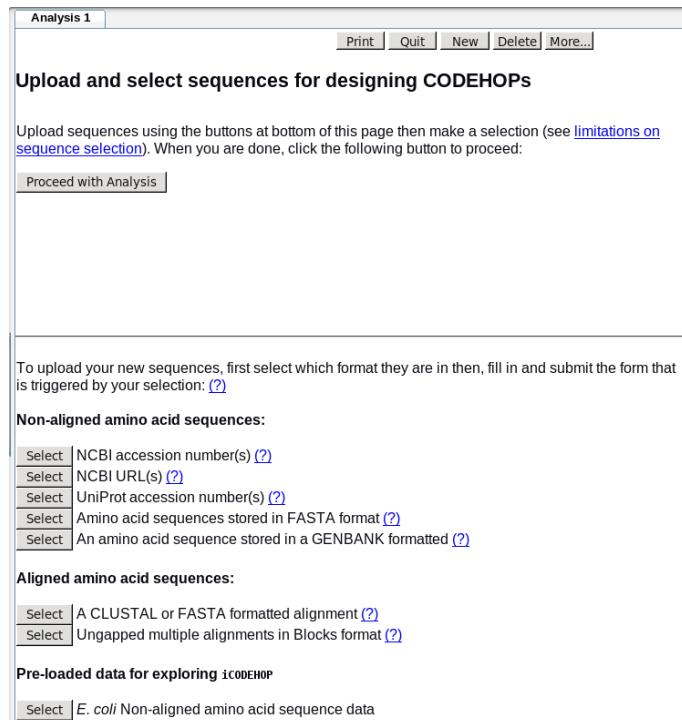
<sup>2</sup>[90](https://icodehop.cphi.washington.edu/i-codehop-context>Welcome</a></p>
</div>
<div data-bbox=)

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



**Figura 16.6:** Página de inicio en iCODEHOP



**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.

The screenshot shows the 'Analysis 1' window with the following details:

- Header:** Analysis 1, Print, Quit, New, Delete, More...
- Title:** Upload and select sequences for designing CODEHOPs
- Text:** 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
- Description:** Sequence Manager: Toggle the selection state of sequences in the sequence table below by clicking on their names or the sequence group name with your mouse. To expand a sequence group or family click on the '+' icon next to its name [\(?\)](#).
- Table:** Sequence Groups
 

Sequence Groups	Description
↳ Gapped multiple-alignments	
↳ test	
✓ PMarmoset	MASRLVNIKEEVTCPLICELLTEPLSLDCGHSHFCQACIT...
✓ AGM	MASGILNVKEEVTCPLICELLTEPLSLPCGHSHFCQACIT...
✓ Saki	MASRLVMNKEEVTCPLICELLTEPLSLDCGHSHFCQACIT...
✓ Gibbon	MASGILVNKEEVTCPLICELLTEPLSLDCGHSHFCQACIT...
✓ Chimp	MASGILVNKEEVTCPLICELLTEPLSLDCGHSHFCQACIT...
✓ Titi	MASGILVNKEEVTCPLICELLTEPLSLDCGHSHFCQACIT...
✓ Baboon	MASGILVNKEEVTCPLICELLTEPLSLDCGHSHFCQACIT...
✓ Squirrel	MASRILGSIKEEVTCPLICELLTEPLSLDCGHSHFCQACIT...
✓ Colobus	MASGILVNKEEVTCPLICELLTEPLSLHCGHSFCQACIT...
- Text:** 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: [\(?\)](#)
- Section:** 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 BLOCKS<sup>3</sup>, 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.

Debajo del nombre del BLOCK encontrará una fila con con información sobre el número de

<sup>3</sup>¿Qué son y como se determinan los BLOCKS?.

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

- Analysis 1** tab is selected.
- Print | Quit | New | Delete | More...** menu bar.
- CODEHOP will suggest PCR primers based protein multiple sequence alignments** message.
- Look for primers** button.
- Block(s):** A scrollable list of sequence blocks. One block is 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%; strength=0
Human: M A S G L I W K V E E V C T C T O L U T P I S L D C G H G F C O A L T A H H S 2.001157
Chimp: ( ) 0
Gorilla: ( ) 0
Orangutan: ( ) 0
```
- Genetic code:** Standard dropdown.
- Codon usage table:** A scrollable list of organisms:
 

```
29058 Helicoverpa armigera
7102 Heliothis virescens
51029 Heterodera glycines
6421 Hirudo medicinalis
9606 Homo sapiens
```
- Advanced Settings:**
  - Re-weight** 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

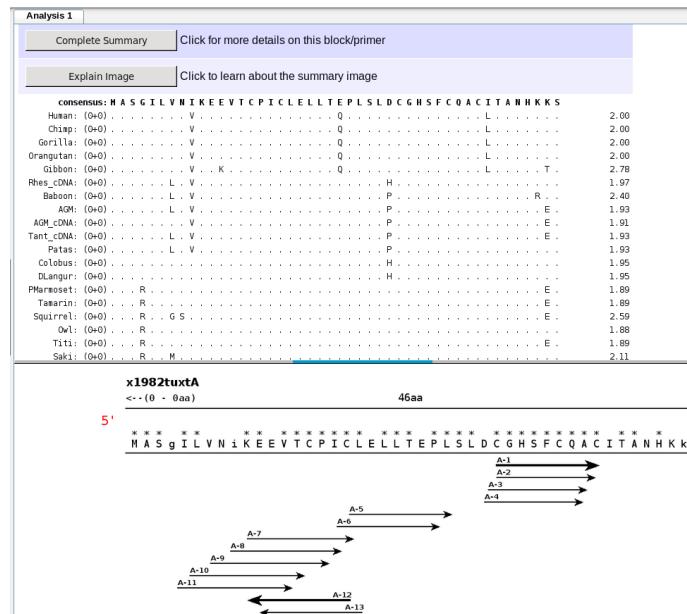
amino ácidos que constituyen el BLOCK y la distancia en amino ácidos al BLOCK anterior y al siguiente (esto último en paréntesis).

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.

Deabajo 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

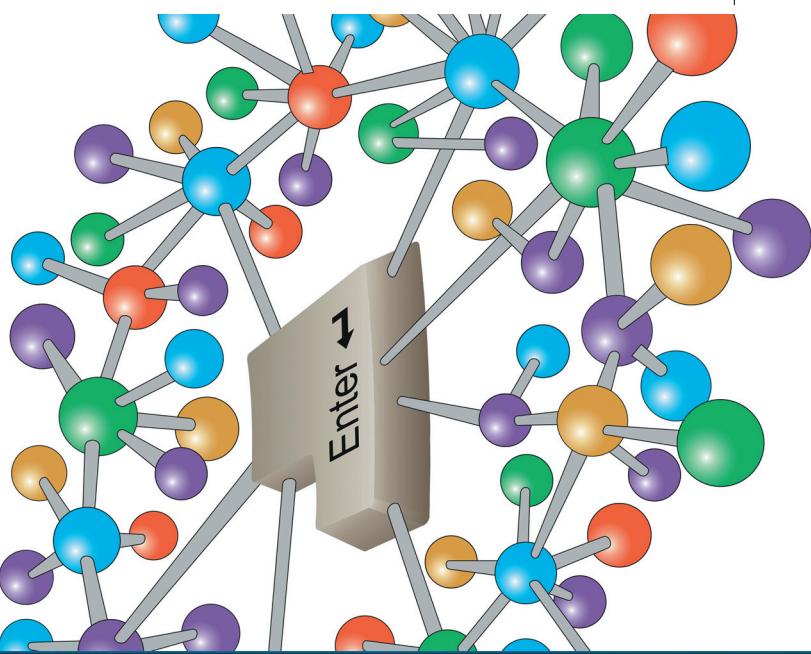
# Bibliografia

- ABASCAL, F., R. ZARDOYA, and D. POSADA, 2005 ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104–2105.
- 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.

- NOTREDAME, C., D. G. HIGGINS, and J. HERINGA, 2000 T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217.
- POSADA, D., 2006 Modeltest server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Res* **34**: W700–W703.
- POSADA, D., and K. CRANDALL, 1998 MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–8.
- POSADA, D., and K. CRANDALL, 2001 Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* **18**: 897–906.
- 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 trim5alpha 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.
- SULLIVAN, J., and P. JOYCE, 2005 MODEL SELECTION IN PHYLOGENETICS. *Annual Review of Ecology, Evolution, and Systematics* **36**: 445–466.
- TAMURA, K., J. DUDLEY, M. NEI, and S. KUMAR, 2007 MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.
- THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.
- WERNERSSON, R., and A. G. PEDERSEN, 2003 Revtrans: Multiple alignment of coding dna from aligned amino acid sequences. *Nucleic Acids Res* **31**: 3537–3539.
- 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

<b>PROCESSES</b>	
<code>^c &lt;ctrl&gt;-c</code>	kills (definitely stops) current job
<code>^z &lt;ctrl&gt;-z</code>	suspends the current job. This can either be moved to the background or resumed in the foreground by using <b>bg</b> or <b>fg</b>
<b>bg</b>	moves the current process to the background
<b>fg</b>	moves a process to the foreground. (If there is more than one suspended job, use <b>jobs</b> to decide which you want to <b>fg</b> )
<b>fg 2</b>	moves process number 2, as listed by <b>jobs</b> , to the foreground
<b>jobs</b>	lists background and suspended processes (created with <b>bg</b> or <code>^z</code> )
<b>jobs -1</b> ("el" not one)	includes the pid (process id number)
<b>ps</b>	lists all your processes
<b>kill</b>	stops a process (use <b>ps</b> or <b>jobs</b> to find your processes)
<b>kill 2986</b>	kills off the process with pid 2986
<b>MISCELLANEOUS</b>	
<b>finger</b>	tells you who is logged on (see also <b>w</b> )
<b>w</b>	shows information about logged in users
<b>who</b>	produces similar result (see <b>finger</b> )
<b>tar</b>	create (or extract) a tarball from (to) a list of files
<b>tar -cvf tarball.tar subdir/*</b>	
<b>tar -xvf tarball.tar</b>	the option <b>-z</b> compresses the files by <b>gzip</b>
<b>wc</b>	word count
<b>wc long.file</b>	prints the number of lines, words and characters in <i>longfile</i> . Options include <b>-l</b> to count lines only, and <b>-c</b> to count characters only
<b>In</b>	create a link or an alias for a file
<b>In -s subdir/origfile alias.file</b>	
<b>history</b>	displays last several commands used
<b>!!</b>	re-executes the last command
<b>!51</b>	executes command 51 in the history list use also <code>&lt;up&gt;</code> - and <code>&lt;down&gt;</code> - arrows to navigate in the history
<b>date</b>	displays current date and time
<b>passwd</b>	invokes a password changing program
<b>exit</b>	leaves the current shell (same as <code>^d</code> or <code>&lt;ctrl&gt;-d</code> ) usually = logout
<b>GRAPHIC DISPLAY</b>	
	To display graphics, most Unix require the configuration of the X-Window server.
	Commands on your local computer:
<b>xhost +</b>	set the list of allowed X-Window clients
	The "+" allows any remote computer to display on your local display
<b>ifconfig</b>	gives information about the network configuration (e.g., the current IP_address, usually similar to 123.145.167.189)
	Commands on the remote computer:
<b>setenv</b>	set up an environment variable (tc-shell)
<b>setenv DISPLAY IP_address:0.0</b>	required to tell the remote computer where it should display its graphics
<b>xclock</b>	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:
	<a href="http://www.embnet.org/">http://www.embnet.org/</a>
	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.

## 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

**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

**vi** simple screen oriented text editor

**pico** simple display oriented text editor

**pico myfile.txt**

prints the first few (default = 10) lines of a file

**head 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**

## OUTPUT REDIRECTION

>

**diff file1 file2 > newfile**

puts differences into *newfile*

**cat onefile twofile > bothfile**

writes the output of the cat command into *bothfile*  
(overwrites *bothfile*)

>>

**cat threefile >> bothfile**

appends *threefile* to the bottom of *bothfile*

|

**cat threefile > bothfile**

appends *threefile* to the bottom of another

**cat**

**grep string myfile | wc -1**

finds how many lines on which “*String*” occurs  
(see **grep** and **wc**)

## DIRECTORIES

**cd ..**

go up one level in directory tree

**cd ./subdir2**

go “sideways” to *subdir2*

**cd ..**

go to /etc directory

**cd ..**

creates a new subdirectory

**mkdir subdir**

removes a directory - you must delete all the files in it first

**rmdir subdir**

print working directory, tells your current location (path)

**pwd** simple screen oriented text editor

**vi** simple display oriented text editor

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.

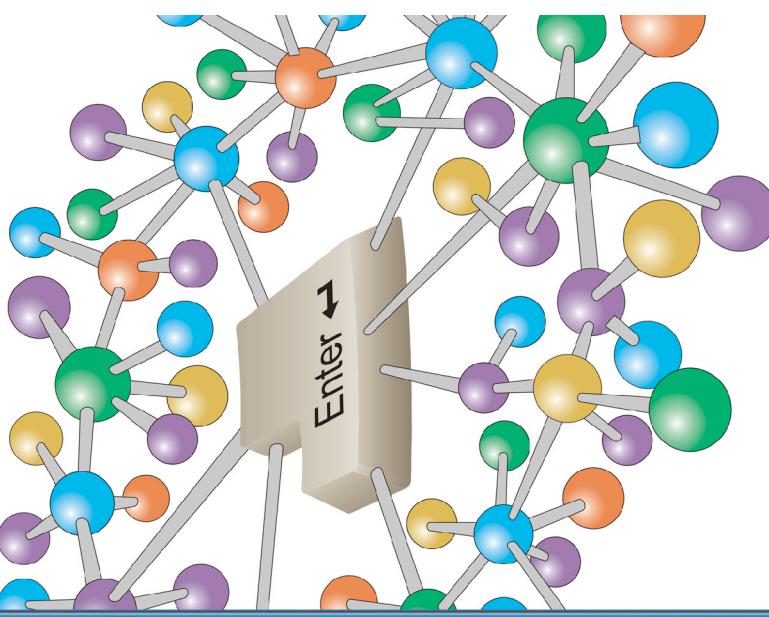
You can find information about your national node from the EMBnet site:

<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	
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
--------	---

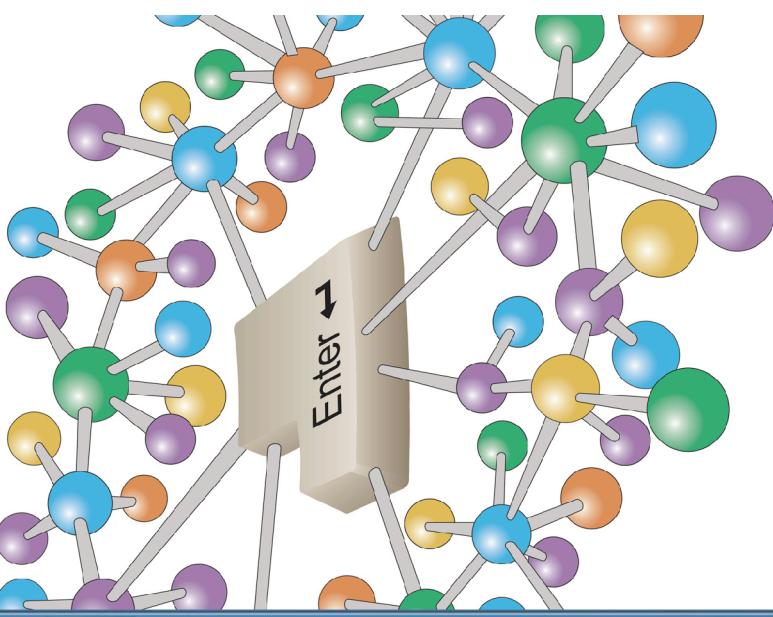
Your local EMBOSS installation may have many sequence databases set up. The program showdb will indicate the available databases.

## 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
filename:entry	a list file (see below)
@listfilename	a specific short sequence
asis::ACGACTGACGG	

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 bl2seq arguments:*

```
-D [integer] type of megablast output (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 GIs
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 = blastp).
-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),
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>

## Where to start?

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.

<b>BLASTX</b>	– 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.	-p [T/F] -o [T/F]
<b>TBLASTN</b>	– 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.	-v [integer]
<b>TBLASTX</b>	– searches all six-frame query translations against all six-frame database translations. Effectively performs a more sensitive blastp search without doing manual translations.	-n [string]
<b>CDD-Search</b> (Conserved Domain Database Search)		<b>Fasta from databases</b>
– used to identify conserved protein domains.		<b>fastacmd</b> retrieves FASTA formatted sequences from a BLAST database, if it was formatted using the ‘-o’ option.
<b>CDART</b> (Conserved Domain Architecture Retrieval Tool)		<i>Selected fastacmd arguments:</i>
– explores the domain architectures of proteins.		-d [string] -s [string] -i [string]
<b>Blast 2 sequences</b> – direct comparison of two sequences.		-l [integer]
<b>VeeScreen</b> – screens DNA sequence queries for vector contamination using a database of known vectors.		<b>Stand-alone blast</b>
		performs all five flavors of blast comparison.
		<b>blastall</b>
		<i>Selected blastall arguments:</i>
		-p [string] -d [string]
		-i [file in]
		-e [real] -o [file out] -F [string]
		<b>Main databases (available at NCBI)</b>
<b>Protein</b> <i>nr</i> (non-redundant + PDB + SwissProt + PIR + PRF); <i>swissprot</i> (latest major release of the SWISS-PROT); <i>pat</i> (proteins from patent division of GenBank); <i>month</i> (new data released in the last 30 days); <i>pdb</i> (3-dimensional structure records from Protein Data Bank).		<b>blastp</b>
<b>Nucleotide</b> <i>nr</i> (GenBank + EMBL + DDBJ + some PDB); <i>est</i> (GenBank + EMBL + DDBJ from EST division); <i>pat</i> (nucleotides from patent division); <i>pdb</i> (3-dimensional structure records); <i>month</i> (new data released in the last 30 days); <i>chromosome</i> (complete genomes and chromosomes); <i>est human</i> (human subset of EST); <i>est mouse</i> (mouse subset of EST); <i>est others</i> (subset of EST other than human or mouse); <i>gss</i> (Genome Survey Sequence); <i>htgs</i> (Unfinished High Throughput Genomic Sequences); <i>aln repeats</i> (select Alu repeats from REBASE); <i>dbsts</i> (STS division + EMBL + DDBJ); <i>wgs</i> (assemblies of whole genome shotgun sequences).		performs all five flavors of blast comparison.
		<b>blastn</b>
		<i>Selected blastn arguments:</i>
		-p [string] -d [string]
		-i [file in]
		-e [real] -o [file out] -F [string]
		<b>LOCAL BLAST INSTRUCTIONS</b>
<b>Format source databases</b>		<b>formatdb</b>
formats protein or nucleotide source databases before they can be searched by blastall, blastp, or megablast. The source database may be in either FASTA or ANS.1 format.		formats protein or nucleotide source databases before they can be searched by blastall, blastp or megablast.
<i>Selected formatdb arguments:</i>		<i>Selected formatdb arguments:</i>
-t [string]		-t [string]
-i [file in]		-i [file in]
-l [file out]		-l [file out]
		<b>To use both SEG and coiled-coil: -F “C;S”.</b>
		number of alignments (def = 250).
		number of one-line description (def = 500).
		query genetic code (def = 1).