Introdução ao EMBOSS

Gary Williams

[tradução Carlos E. Winter – 2006]

O que é EMBOSS?

- Pacote Wisconsin, GCG
- Amplamente utilizado, fontes disponíveis para inspeção
- 1988 - EGCG - início de "add-on" acadêmico
- GCG comercial - fontes **não** disponíveis livremente!
- 1999 - EGCG se separa do GCG para se tornar o **EMBOSS**

O que é EMBOSS?

- Um novo conjunto de programas
- Programa de código aberto - códigos disponíveis
- Domínio público (GNU Public Licence)
- Escrito pelo HGMP/Sanger/EBI/Dinamarca... etc

O que ele pretende fazer

- Um conjunto de programas úteis e integrados
- Eles compartilham um aspecto comum
- Incorpora muitos programas grandes e pequenos
- Fácil de executar a partir da linha de comando
- Fácil de chamar a partir de outros programas (p. ex. perl)
- Fácil de instalar atrás de GUIs e interfaces Web

Campos de aplicação

- Há muitos programas EMBOSS (>140)
- Ver:

http://www.uk.embnet.org/Software/EMBOSS/Apps/

- Muitos programas para apresentação e análise de seqüências.
- Predição de estrutura 3D de proteínas está sendo desenvolvido.
- Outros programas diversos, p. ex.: cinética enzimática.

Um exemplo de programa EMBOSS

- É fácil esquecer o nome de um programa.
- Para encontrar os programas EMBOSS, use **wossname**
- **wossname** acha programas procurando por palavras-chave na descrição ou no nome do programa.

Executando na linha de comando

- Digite wossname no promp Unix %

```
Unix % wossname
```

- Mostra descrição de uma linha.
- Pede informação para voce:

```
Finds programs by keywords in their one-line documentation
Keyword to search for: restrict

SEARCH FOR 'RESTRICT'
recode          Remove restriction sites but maintain the
   same translation
remap           Display a sequence with restriction cut
   sites, translation
```

etc.....

Parâmetros opcionais

```
Unix % wossname -opt
Finds programs by keywords in their one-line
  documentation
Keyword to search for: protein
Output program details to a file [stdout]: myfile
Format the output for HTML [N]: Y
String to form the first half of an HTML link:
String to form the second half of an HTML link:
Output only the group names [N]:
Output an alphabetic list of programs [N]:
Use the expanded group name [N]:
```

Help

```
Unix % wossname -help
 Mandatory qualifiers:
  [-search]   string     Enter a word or words here.

 Optional qualifiers (* if not always prompted):
   -outfile   outfile this program will write the program names

 Advanced qualifiers:
   -[no]emboss bool      EMBOSS program
                         documentation will be searched.
```

- Obrigatório - necessário, são parâmetros em ‘[]’
- Opcional - use **-opt** para que eles sejam solicitados.
- Avançado - coisas que não são geralmente utilizadas!

Escrevendo no monitor

- Observe que o arquivo de saída para o wossname era:

  stdout (Standard output)
- Use isso sempre que for solicitado para um arquivo de saída.
- Este é um nome "mágico" de arquivo.
- O resultado será apresentado no monitor, não num arquivo.

Prática

- Tente rodar **wossname**
- Voce consegue achar um programa para:

- Mostrar alinhamentos múltiplos?
- Encontrar ORFs (Open Reading Frames)?
- Traduzir uma seqüência?
- Encontrar sítios de enzimas de restrição?
- Calcular o ponto isoelétrico de uma proteína?
- Fazer alinhamentos globais?

Trabalhando com seqüências

- EMBOSS le seqüências de **arquivos** ou de **bancos de dados**.
- Ele automaticamente reconhece o formato da seqüência de entrada.
- Voce pode facilmente especificar muitos formatos de saída.

Pegando seqüências de bancos de dados

- Registro individual do banco de dados (ID)
  - **database:entry**
  - Por exemplo embl:hsfau
- Wildcarded (Query)
  - **database:hs***
- Todos os registros
  - **database:***
- Muitos bancos de dados aceitam os três métodos - alguns não.

showdb

Unix % **showdb**

Displays information on the currently available databases

```
#Name          Type ID  Qry All Comment
#====          ==== ==  === === =======
pir             P    OK  OK  OK  PIR/NBRF
remtrembl       P    OK  OK  OK  REMTREMBL sequences
sptrembl        P    OK  OK  OK  SPTREMBL sequences
swissprot       P    OK  OK  OK  SWISSPROT sequences
embl            N    OK  OK  OK  EMBL sequences
emblnew         N    OK  OK  OK  New EMBL sequences
est             N    OK  OK  OK  EMBL EST sequences
```

seqret

- Lê uma seqüência e cria um arquivo.

```
Unix % seqret
Reads and writes (returns) a sequence

Input sequence: embl:xlrhodop
Output sequence [xlrhodop.fasta]:

unix % more xlrhodop.fasta

>XLRHODOP L07770 Xenopus laevis rhodopsin
ggtagaacagcttcagttgggatcacaggcttctagggatcctttgggcaaaaaagaaac
acagaaggcattctttctatacaagaaaggactttatagagctgctaccatgaacggaac
        .
        .
```

seqret da linha de comando

- De a seqret todos os dados na linha de comando.
- Não necessita solicitar nada mais.

Unix % **seqret embl:xlrhodop -outseq xlrhodop.fasta**

- O '**-outseq**' pode ser abreviado para '**-out**'
- Qualquer abreviatura tem que ser única.

- Ainda mais curto, não ponha o qualificador:

Unix % **seqret embl:xlrhodop xlrhodop.fasta**

Mudando os formatos de saída (reformatando)

- **seqret** pode reformatar seqüências se o formato de saída for especificado:

Unix % **seqret embl:xlrhodop xlrhodop.fasta -osformat gcg**

Unix % **more xlrhodop.gcg**

```
!!NA_SEQUENCE 1.0
Xenopus laevis rhodopsin mRNA, complete cds.
XLRHODOP  Length: 1684  Type: N  Check: 9453 ..
    1 ggtagaacag cttcagttgg gatcacaggc ttctagggat cctttgggca
   51 aaaaagaaac acagaaggca ttctttctat acaagaaagg actttataga
        .
        .
```

Lendo seqüências a partir de arquivos

- De só o nome do arquivo:

Unix % **seqret myclone.seq gcg::myclone.gcg**

- Voce pode especificar o formato de entrada (não necessário):

Unix % **seqret gcg::myclone.gcg clone2.seq**

- Uma seqüência de um arquivo com muitas seqüências:

Unix % **seqret allclones.seq:52H12 52H12.seq**

Liste arquivos (arquivos ou nomes de arquivos)

- Um método rápido de agrupar seqüência para trabalhar, como um banco de dados privados.
- Qualquer especificação válida de seqüência deve ser utilizada, não simples nomes de arquivos.
- Uma entrada por linha num arquivo.
- Linhas de comentário começam com um ‘#’
- Indique que isto é uma lista começando por um ‘@’:

  Unix % **infoseq @mylist**
- Muitos programas (infoseq, fuzznuc, fuzzpro) podem fazer listas de arquivos com uma pesquisa (use a opção '**-usa**')

Seqüências múltiplas, arquivo único

- **EMBOSS** coloca muitas seqüências num único arquivo.
- A maioria dos formatos é compatível com isso:
  - Fasta, EMBL, PIR, MSF, Clustal, Phylip, etc.
- MAS NÃO: Plain, Staden and GCG
- **EMBOSS** lê muitas seqüências de um único arquivo.
- Use **arquivo:entrada** se voce quiser especificar uma única seqüência.
- If there is only one sequence, or you wish to read all entries, use just the **filename.**

Seqüências múltiplas, arquivo único

- Se voce quer colocar uma seqüência por arquivo, use:

  ‘-ossingle’

Unix % **seqret “embl:hsf*” dummy –ossingle**

- Os nomes dos arquivos de saída serão baseados no nome das seqüências.

- O programa **seretsplit** separará um arquivo múltiplo existente em diversos arquivos.

Asterisco na linha de comando

- Voce não pode usar um '*' na linha de comando do UNIX
- UNIX tenta encontrar arquivos com este mesmo nome.
- Use-o entre aspas ou com uma barra invertida:

  "embl:*"

  embl:*

- Por exemplo:

Unix % seqret “embl:hsf*” hsf.seq

Prática

- Tente executar **showdb**, **seqret** e **infoseq:**

- Mostre somente os bancos de dados de ácidos nucléicos
- Pegue a seqüência '**hsfau**' do banco de dados do EMBL e coloque no arquivo '**this.seq'**.
- O mesmo, mas no arquivo '**this.gcg**' no formato GCG.
- Mostre a informação sobre a seqüência em '**this.seq**'.
- Mostre a informação sobre todas as seqüências cujo nome começa com '**10**' no banco de dados do SwissProt.

GUIs

- Há muitas interfaces disponíveis ou em preparação::
- **emnu** - menu alegre baseado em caracteres
- **w2h** – interface para W3
- **spin** – do time do Staden
- Other web interfaces:

  http://userpage.fu-berlin.de/~sgmd/

  http://corba.ebi.ac.uk/cgi-bin/alweb2/alweb.start?CFG=Emboss

  http://bioinfo.pbi.nrc.ca:8090/emboss/index.html

  http://ubigcg.mdh4.mdc-berlin.de:8080/

  http://www-alt.pasteur.fr/~letondal/Pise/

Conclusão - help

- If in doubt, use:

  wossname

  program -help

  program -opt

  tfm program

Conclusão – dados de seqüência

- Para informação sobre bancos de dados, use
  showdb
- Uniform Sequence Addresses (USAs):
  - **database**
  - **database:entry_name** or **database:accession_number**
  - **database:wildcard**
  - **filename**
  - **filename:entry**
  - **format::filename**
  - **@list**

Conclusão – outros qualificadores

- -**sbegin** posição de início da seqüência
- **-send** posição do fim da seqüência
- **-sreverse** reverso complementar da seqüência
- **-slower** mude a seqüência para minúsculas
- **-supper** mude a seqüência para maiúsculas
- **-osformat** formato da seqüência de saída
- **-help** mostre o ‘help’
- **-options** peça parâmetros adicionais
- **-auto** execute silenciosamente (para uso em ‘scripts’, p. ex. perl)