

Guide EMBOSS

European Molecular Biology Open Software Suite

Site web de EMBOSS: <http://emboss.sourceforge.net/>

Utilisation avec interface graphique Java: **Jemboss**

Principalement utilisé en mode ligne de commande

Configuration personnalisée d'emboss avec les variables d'environnement

Dans le fichier *.bashrc* :

```
export EMBOSSRC=.embossrc
```

Par la suite, vous pouvez configurer votre environnement emboss à partir du fichier *.embossrc* dans \$HOME :

#Pour régler le format d'utilisation par défaut (en entrée) des programmes :

```
set emboss_FORMAT fasta
```

#le format de sortie des programmes est obligatoirement fixé sur fasta

```
set emboss_OUTFORMAT fasta
```

#Premières configurations de banques de séquences

```
DB entrezP [  
  type: P  
  format: genbank  
  method: entrez  
  fields: "id acc gi sv des org key"  
  url:    "http://www.ncbi.nlm.nih.gov/sites/gquery"  
]
```

```
DB entrezN [  
  type: N  
  format: genbank  
  method: entrez  
  fields: "id acc gi sv des org key"  
  url:    "http://www.ncbi.nlm.nih.gov/sites/gquery"  
]
```

Premier pas avec les programmes :

Pour connaître la version d'emboss installée sur votre ordinateur

```
embossversion
```

Pour obtenir la liste complète des programmes disponibles :

```
wosname -auto -out fichier
```

```
wosname -groups -auto -out fichier
```

```
wosname -alphabetic -auto -out fichier
```

Pour obtenir la documentation sur un programme

Mode simple

nom_du_prog -help

Mode « man page

tfm nom_du_prog

Vous pouvez vous déplacer avec la barre d'espace ou les flèches ↑ et ↓. Pour sortir de l'aide, entrez la touche « q »

Pour connaître la liste des banques disponibles ainsi que leur nom logique

showdb

Displays information on configured databases

#	Name	Type	ID	Qry	All	Comment
#	=====	=====	=====	=====	=====	=====
	entrezP	Protein	OK	OK	OK	
	entrezN	Nucleotide	OK	OK	OK	

Extraire une séquence

seqret entrezN:af000179

seqret entrezP:aakg1_mouse

Format d'annotation complet

entret entrezN:af000179

entret entrezP:aakg1_mouse

Introduction à la syntaxe USA (Uniform Sequence Adress). Les fichiers de séquences peuvent être utilisés selon une des configurations suivantes:

1. file
2. file:entry
3. format::file
4. format::file:entry
5. database:entry
6. database
7. @file

Exemple d'un fichier de liste (test_fosn) (FOSN : File Of Sequence Names) :

entrezN:m16801

entrezN:af000179

entrezP:002r_iiv3

entrezP:aakg1_mouse

seqret @test_fosn

Pour obtenir des fichiers individuels

seqret entrezN:xlu* -ossingle

Pour regrouper des fichiers individuels en un seul fichier (vous devez spécifier le format des fichiers en entrée):

seqret fasta::nm_* -outseq tous_nm.fasta

Pour changer le format d'un fichier:

```
secret genbank::af000179.entret fasta::af000179.fasta
```

Extraction de séquences à partir des champs indexés

```
entret entrezN-org:Carpinus
```

```
entret entrezN-key:mouse
```

EMBOSS contient aussi plusieurs fichiers de données nécessaires à l'analyse des séquences. Ils sont placés à des endroits spécifiques de l'ordinateur. Vous devez savoir où se trouvent ces fichiers pour en vérifier le contenu, les copier dans votre HOME et les corriger au besoin. Par défaut, ces fichiers sont entreposés dans le répertoire prg/emboss/current/share/EMBOSS/data:

1. Tables de traduction (codes génétiques): EGC.1.0
2. Matrices de pointage pour les substitution: EBLOSUM, EPAM, EDNAFUL, EDNA, ENUC
3. Sites de restriction
4. Facteurs de transcription
5. Table de fréquence d'utilisation des codons (Fichiers *.cut (codon usage table) dans le répertoire /cbi/prg/emboss/EMBOSS-2.7.1/share/EMBOSS/data/CODONS. Utilisés par le logiciels cai)
6. Enzymes protéolytiques et réactifs chimiques
7. Analyse protéique (isolation de peptides, structure secondaire, hydrophobicité, antigénicité)
8. Motifs protéiques
9. Volume des acides aminés
10. Profils

Pour connaître l'emplacement et le nom des répertoires et des fichiers de données (table de traduction, enzymes de restriction, etc) lus par les programmes de emboss:

```
embosdata
```

```
embosdata -showall | more
```

Pour extraire un de ces fichiers :

```
embosdata -filename EPAM60 -fetch
```

Indexation d'un fichier de séquences

Recherche des programmes avec wosname avec le mot 'index'

Avec dbiflat, vous pouvez utiliser l'option -field 'acc,des,key,org,sv'

Pour le fichier .embossrc lorsque vous utilisez dbxflat, dbxfasta (taille de fichiers > 2.0Gb) :

```
RES uniprot_sprot [  
  type: Index  
  idlen: 15  
  acclen: 15  
  svlen: 15  
  keylen: 35  
  deslen: 35  
  orglen: 35  
]  
DB uniprot_sprot [  
  type: P  
  method: emboss  
  format: swiss  
  dir: $bank_prot/uniprot/current  
  file: "uniprot_sprot.dat"  
  indexdir: $index_prot/uniprot/current  
  fields: "id acc sv des key org"  
  comment: "UniProt release 2011_02 08 Feb 2011"  
]
```