

Analysis tools

Ueng-Cheng Yang

National Yang-Ming University

Oct. 28, 2002

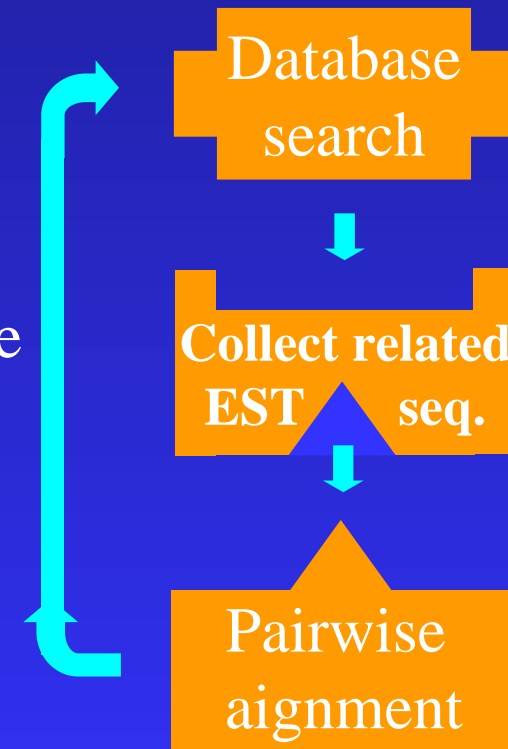
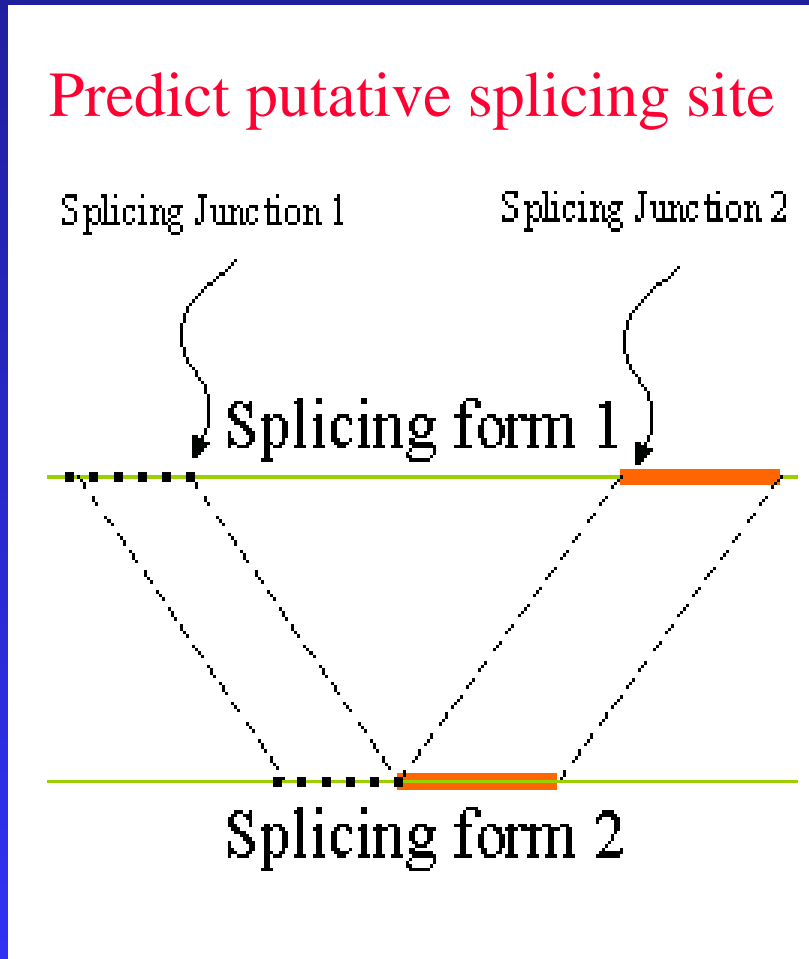
Command *vs* web interface

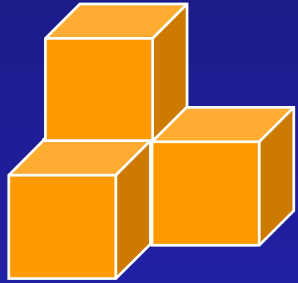
Web interface can cross-platforms.

vs

Command mode can create pipeline.

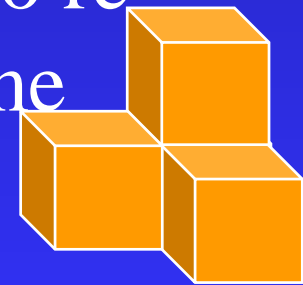
Splicing site can be predicted *in silico*





Lego-like “workflow” system can create a large amount of combinations

The users may spend less time in learning biocomputing, but they will be able to re-run these modules to save their time



Commonly used program packages

- GCG (commercial, need an account),
<http://gcg.nhri.org.tw/>
- BCM launcher,
<http://searchlauncher.bcm.tmc.edu/>
- Biology workbench,
<http://workbench.sdsc.edu/>
- **EMBOSS**,
<http://binfo.ym.edu.tw/emboss/Apps/>

Sequence Analysis Workbench

<http://saw.ym.edu.tw/>

- European Molecular Biology Open Software Suite (EMBOSS)
~ 140 programs
- Basic Local Alignment Search Tool (BLAST)
- profile Hidden Markov Models for biological sequence analysis
(HMMer)
- PHYLogeny Inference Package (PHYLIP)
- Vienna RNA package (V_RNA_Pac)
- Gene Locator and Interpolated Markov Modeler (Glimmer)

Jemboss can also cross platforms

It's easy to manage sequences in Jemboss,
but you need to have an account on server.

How to learn a program package?

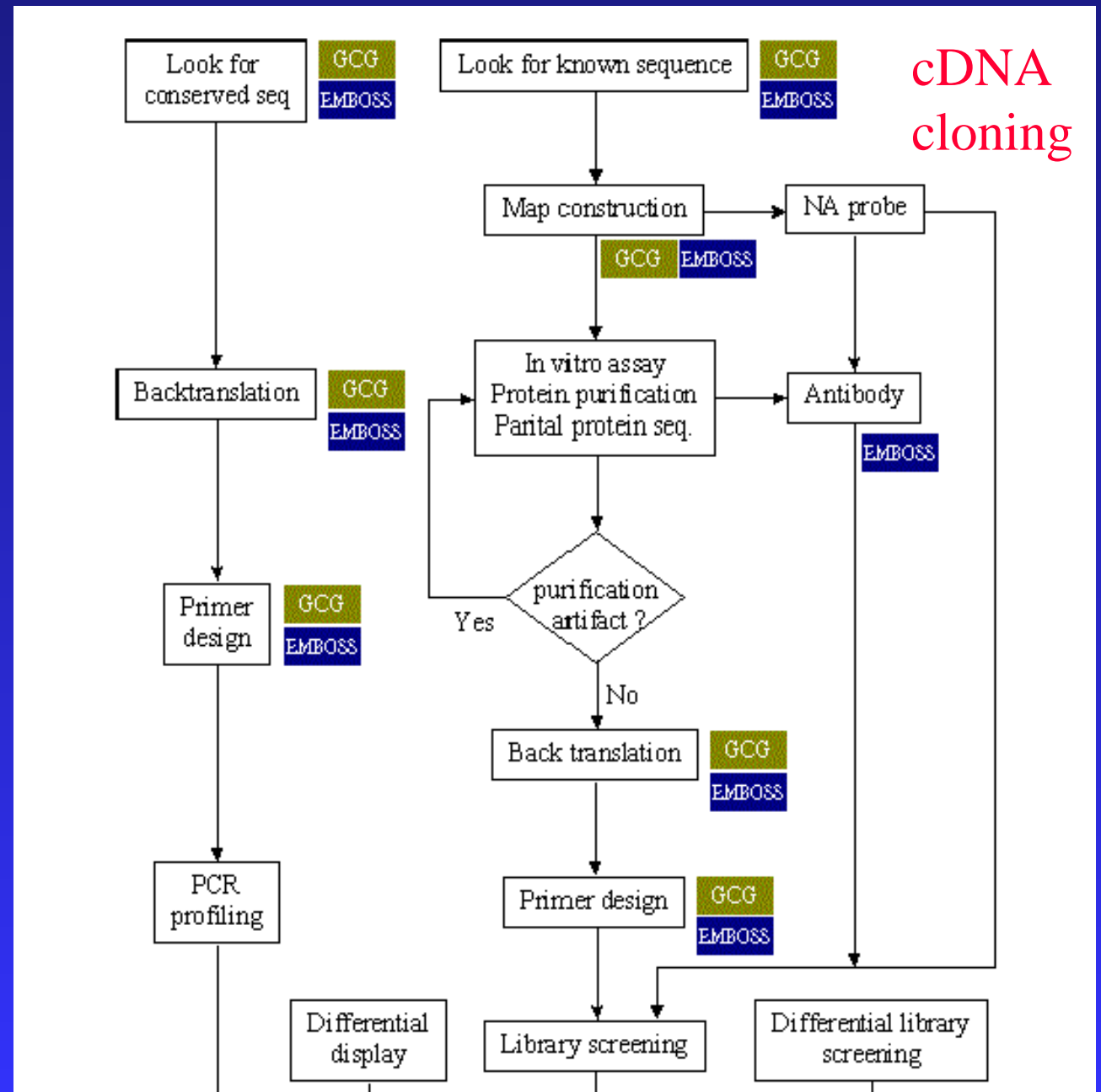
- Program selection
- Sequence management
- Parameter setting
- Data interpretation

The usual way to select a program

- List by function
 - Sequence format conversions
 - Sequence alignments and comparisons
 - Search genes and coding region
 - Pattern search and discovery
 - Protein sequence analysis
 - DNA sequence analysis
 - Structure analysis

Problem-Oriented Sequence analysis Tools (POST)

- 簡介
- 標準分析
- 個案分析
- 程式類別
- 參數設定
- 結果分析



Supported by “Program to promote academic excellence”

Program est2genome ([YMBC](#))

Function

Align EST and genomic DNA sequences

Description

est2genome is a software tool to aid the prediction of genes by sequence homology. The program will align a set of spliced nucleotide sequences (ESTs cDNAs or mRNAs) to an unspliced genomic DNA sequence, inserting introns of arbitrary length when needed. In addition, where feasible introns start and stop at the splice consensus dinucleotides GT and AG.

Unless instructed otherwise, the program makes three alignments: First it compares both stands of the spliced sequence against the forward strand of the genomic, assuming the splice consensus GT/AG (ie in the forward gene direction). The maximum-scoring orientation is then realigned assuming the splice consensus CT/AC (ie in the reversed gene direction). Only the overall maximum-scoring alignment is reported.

EST2GENOME : Align EST and genomic DNA sequences (EMBOSS)

your e-mail
(● = required, ● = conditionally required)

● EST sequence(s) (-est) : please enter [size\(s\)](#) :

1. the name of a file:

2. or the actual data here:

How to learn a program package?

- Program selection
- Sequence management
- Parameter setting
- Data interpretation

How to use a sequence?

Sequence search

Cut and paste / browse

Specify a sequence name

SHOWDB : Displays information on the currently available databases (EMBOSS)

Results:

[showdb.out](#)

[standard error file](#)

From now, this files will remain accessible for 10 days at:

<http://Pise/tmp/showdb/A15011103577940/>

You can save them individually by the **Save file** function if needed.

Job summary

default format

Unix exact command:

```
showdb -auto -stdout
```

Help

Databases installed

```
# Name          Type ID Qry All Comment
# =====
demo_p          P   -  -   OK  demo amino acid sequence fasta
                file direct access with no index
sprot           P   OK OK  OK  Swissprot native format with
                EMBL CD-ROM index
demo_n          N   -  -   OK  demo nucleotide sequence fasta
                file direct access with no index
refseq          N   OK -   -   Genbank ACs
```

SEQRET : Reads and writes (returns) a sequence (EMBOSS)

your e-mail

(= required, = conditionally required)

sequence [single sequence] (-sequence) : please enter either :

1. the name of a **file**:

```
sprot:fos_human
```

2. *or* the **actual data** here:

(sequence format)

outseq (-outseq)

Output format for: outseq

your e-mail

SEQRET : Reads sequence (EMBOSS)

Results:

[standard error file](#)

From now, this files will remain accessible
<http://> [/tmp/secret/3064](http://tmp/secret/3064)

You can save them individually by the Sa

Job summary

default format


Unix exact command:

```
-sequence=sequence.data -sforma  
osformat=fasta -auto -stdout
```

Your input data:

[sequence.data](#)

Help


Next page

secret job summary

Your e-mail: yang@ym.edu.tw

sequence [single sequence] (-sequence): [sequence.data](#)

Unix exact command:

```
-sequence=sequence.data -sformat=fastasecret -outseq=outseq.out -osformat=fasta -  
auto -stdout
```

Save job files

MIME attachments

This "Save job files" button allows you to save the input and output files of this job in the chosen format. By choosing MIME, you may also receive the results by email as attachments.

Save secret form with these default values

*This "Save secret form..." button allows you to save a form with the values that you have used for this job as default values (except input sequence or other input data). **Do not bookmark this form**, but instead **save it as a local file**.*

Register the whole procedure

Save all the commands you have done with their parameters for later reuse.

The fasta format

```
>FOS_HUMAN P01100 P55-C-FOS proto-oncogene protein (Cellular  
oncogene C-FOS) (G0S7 protein).  
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANF  
IPTVTAISTSPDLQWLVPALVSSVAPSQTRAPHPFGVPAPSAGAYSRAGVVKTMTGGRA  
QSIGRRGKVEQLSPEEEKRRIRRERNKMAAAKCRNRRRELDTLQAETDQLEDEKSALQ  
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVASLDLTGGLPEVATPESEEAFT  
LPLLNDPEPKPSVEPVKSISSMELKTEPFDDFLFPASSRPSGSETARSVPDMDLSGSFYA  
ADWEPLHSGSLGMGPMATELEPLCTPVVTCTPSCTAYTSSFVFTYPEADSFPSCAAHRK  
GSSSNEPSSDSLSSPTLLAL
```

RESTRICT : Finds restriction enzyme cleavage sites (EMBOSS)

yang@ym.edu.tw your e-mail

(= required, = conditionally required)

sequence -- DNA [sequences] (-sequence) : please enter either :

1. the name of a **file**:

refseq:NM_152736

2. *or* the **actual data** here:

(sequence format)

RESTRICT : Finds restriction enzyme cleavage sites (EMBOSS)

Results:

[outfile.out](#)



[standard error file](#)

From now, this files will remain

[http://](#) [/tmp](#)

You can save them individually

Unix exact command:

```
restrict -sequence=sequence.data -  
outfile=outfile.out -
```

Your input data:

[sequence.data](#)

```
#####  
# Program: restrict  
# Rundate: Tue Oct 29 00:54:26 2002  
# Report_file: outfile.out  
#####  
  
#-----  
#  
# Sequence: ZNF187      from: 1    to: 2298  
# HitCount: 488  
#  
# Minimum cuts per enzyme: 1  
# Maximum cuts per enzyme: 2000000000  
# Minimum length of recognition site: 4  
# Blunt ends allowed  
# Sticky ends allowed  
# DNA is linear  
# Ambiguities allowed  
#  
#-----
```

Start	End	Score	Enzyme_name	Restriction_site	5prime	3prime	5primerev	3primerev
3	8	0	BssSI	CACGAG	3	7	.	.
9	13	0	Fnu4HI	GCNGC	10	11	.	.
9	13	0	TauI	GCSGC	12	9	.	.
12	15	0	AcII	CCGC	9	11	.	.
18	21	0	AluI	AGCT	19	19	.	.
18	21	0	CviJI	RGCY	19	19	.	.
20	24	0	BstDEI	CTNAG	20	23	.	.

Data Warehouse: SRS6

TOP PAGE QUERY RESULTS SESSIONS VIEWS DATABANKS HELP

Reset Quick Search All Entries

show all + collapse all -

- Sequence
 - EMBL GENBANK SWISSPROT SWISSNEW SPTREMBL
 - REMTREMBL TREMBLNEW
- SeqRelated
 - PROSITE PROSITEDOC PRINTS PFAMA
 - PFAMB SWISSPFAM PFAMHMM PFAMSEED
 - PRODOM GENETICCODE UNIGENE UNISEQ
 - UNIEST
- TransFac
 - TFSITE TFCLASS TFMATRIX TFGENE
- User Owned Databanks
- Application Results
- Genome
 - LOCUSLINK
- Mutations
 - OMIM
- Metabolic Pathways
 - PATHWAY

Query forms

Standard

Extended

bookmark this link to return to your session

If you find problems or have suggestions please mail the [SRS administrator](#)

<http://srs.ym.edu.tw/srs6/>

↓

<http://srs.ym.edu.tw/srs7/>

How to learn a program package?

- Program selection
- Sequence management
- **Parameter setting**
- Data interpretation

RESTRICT : Finds restriction enzyme cleavage sites (EMBOSS)

your e-mail
(● = required, ● = conditionally required)

● sequence -- DNA [sequences] (-sequence) : please enter either :

1. the name of a **file**:

2. *or* the **actual data** here:

(sequence format)

(sequence [format](#))

<input type="text" value="1"/>	Minimum cuts per RE (-min)
<input type="text" value="2000000000"/>	Maximum cuts per RE (-max)
<input type="text" value="4"/>	Minimum recognition site length (-sitelen)
<input type="checkbox"/>	Force single site only cuts (-single)
<input checked="" type="checkbox"/>	Allow blunt end cutters (-blunt)
<input checked="" type="checkbox"/>	Allow sticky end cutters (-sticky)
<input checked="" type="checkbox"/>	Allow ambiguous matches (-ambiguity)
<input type="checkbox"/>	Allow circular DNA (-plasmid)
<input checked="" type="checkbox"/>	Only enzymes with suppliers (-commercial)
<input checked="" type="checkbox"/>	Limits reports to one isoschizomer (-limit)
<input type="checkbox"/>	Report preferred isoschizomers (-preferred)
<input type="checkbox"/>	Sort output alphabetically (-alphabetic)
<input type="checkbox"/>	Show fragment lengths (-fragments)
<input type="checkbox"/>	Show sequence name (-name)
<input type="text" value="all"/>	Comma separated enzyme list (-enzymes)
<input type="text"/>	Alternative RE data file (-datafile)
<input checked="" type="radio"/> <input type="text" value="outfile.out"/>	outfile (-outfile)

your e-mail

Some explanations about the options

Some explanations about the options

Main parameters

enter either the name of a file or the actual data

if you are using Netscape 2.x or later, you can select a file by typing its name, or better, by selecting it with the Netscape file browser (**Browse** button)

OR you can type your data in the next area, or cut and paste it from another application.

(but not both)

Comma separated enzyme list (-enzymes)

The name 'all' reads in all enzyme names from the REBASE database. You can specify enzymes by giving their names with commas between them, such as: 'HincII,hinfI,ppiI,hindiii'.

The case of the names is not important. You can specify a file of enzyme names to read in by giving the name of the file holding the enzyme names with a '@' character in front of it, for example, '@enz.list'.

Blank lines and lines starting with a hash character or '!' are ignored and all other lines are concatenated together with a comma character ',' and then treated as the list of enzymes to search for.

An example of a file of enzyme names is:

```
! my enzymes
HincII, ppiI
! other enzymes
hindiii
HinfI
PpiI
```

How to use “help”?

- Function
- Description
- Usage
- **Command line arguments**
- Input file format
- Output file format
- Data files
- Notes
- References
- Warnings
- Diagnostic error messages
- Exit status
- Known bugs
- See also
- Author(s)
- History
- Target users
- Comments

Command line arguments

Mandatory qualifiers

Optional qualifiers

Advanced qualifiers

Mandatory qualifiers		Allowed values	Default
[-sequence] (Parameter 1)	Sequence database USA	Readable sequence(s)	Required
-sitelen	Minimum recognition site length	Integer from 2 to 20	4
-enzymes	The name 'all' reads in all enzyme names from the REBASE database. You can specify enzymes by giving their names with commas between them, such as: 'HincII,hinfI,ppiI,hindiii'. The case of the names is not important. You can specify a file of enzyme names to read in by giving the name of the file holding the enzyme names with a '@' character in front of it, for example, '@enz.list'. Blank lines and lines starting with a hash character or '!' are ignored and all other lines are concatenated together with a comma character ',' and then treated as the list of enzymes to search for. An example of a file of enzyme names is: ! my enzymes HincII, ppiI ! other enzymes hindiii HinfI PpiI	Any string is accepted	all
[-outfile] (Parameter 2)	(no help text) report value	Report file	

(sequence [format](#))

Minimum cuts per RE (-min)
 Maximum cuts per RE (-max)
 Minimum recognition site length (-sitelen)

- Force single site only cuts (-single)
- Allow blunt end cutters (-blunt)
- Allow sticky end cutters (-sticky)
- Allow ambiguous matches (-ambiguity)
- Allow circular DNA (-plasmid)
- Only enzymes with suppliers (-commercial)
- Limits reports to one isoschizomer (-limit)
- Report preferred isoschizomers (-preferred)
- Sort output alphabetically (-alphabetic)
- Show fragment lengths (-fragments)
- Show sequence name (-name)

Comma separated enzyme list ([-enzymes](#))

Alternative RE data file (-datafile)

outfile (-outfile)



your e

Advanced qualifiers

-min	Minimum cuts per RE
-max	Maximum cuts per RE
-single	Force single site only cuts
-[no]blunt	Allow blunt end cutters
-[no]sticky	Allow sticky end cutters
-[no]ambiguity	Allow ambiguous matches
-plasmid	Allow circular DNA
-[no]commercial	Only enzymes with suppliers
-[no]limit	Limits reports to one isoschizomer
-preferred	Report preferred isoschizomers
-alphabetic	Sort output alphabetically
-fragments	Show fragment lengths
-name	Show sequence name
-datafile	Alternative RE data file

Some explanations about the options

Advanced qualifiers		Allowed values	Default
-min	Minimum cuts per RE	Integer from 1 to 1000	1
-max	Maximum cuts per RE	Integer up to 2000000000	2000000000
-single	Force single site only cuts	Yes/No	No
-[no]blunt	Allow blunt end cutters	Yes/No	Yes
-[no]sticky	Allow sticky end cutters	Yes/No	Yes
-[no]ambiguity	Allow ambiguous matches	Yes/No	Yes
-plasmid	Allow circular DNA	Yes/No	No
-[no]commercial	Only enzymes with suppliers	Yes/No	Yes
-[no]limit	Limits reports to one isoschizomer	Yes/No	Yes
-preferred	Report preferred isoschizomers	Yes/No	No
-alphabetic	Sort output alphabetically	Yes/No	No
-fragments	Show fragment lengths	Yes/No	No
-name	Show sequence name	Yes/No	No
-datafile	Alternative RE data file	Any string is accepted	<i>An empty string is accepted</i>

▼ datafile (-datafile)

EPAM60 [gappenalty](#) -- enter a number (-gappenalty)

EPAM290 [gaplength](#) -- enter a number (-gaplength)

EPAM470

EPAM110

EBLOSUM50 *[in part with your favorite browser's Back function]*

EPAM220

EBLOSUM62-12

EPAM400

EPAM150

EPAM330

EBLOSUM55

EPAM30 outfile (-outfile)

EPAM260

EBLOSUM90 format (-aformat) ? [default] default fasta MSF

EPAM440

EPAM190 *[in part with your favorite browser's Back function]*

EPAM370

EPAM70

EPAM480

EPAM120 your e-mail

EDNAMAT

EPAM300

EBLOSUM60

EPAM230

EBLOSUM62

EPAM410

EPAM160

EPAM340

EBLOSUM65

Information about the options

網際網路

Data files

For protein sequences EBLOSUM62 is used for the substitution matrix. For nucleotide sequence, EDNAMAT is used. Others can be specified.

EMBOSS data files are distributed with the application and stored in the standard EMBOSS data directory, which is defined by EMBOSS environment variable EMBOSS_DATA.

Users can provide their own data files in their own directories. Project specific files can be put in the current directory, or for tidier directory listings in a subdirectory called ".embosdata". Files for all EMBOSS runs can be put in the user's home directory, or again in a subdirectory called ".embosdata".

The directories are searched in the following order:

- . (your current directory)
- .embosdata (under your current directory)
- ~/ (your home directory)
- ~/.embosdata

How to learn a program package?

- Program selection
- Sequence management
- Parameter setting
- Data interpretation

How to use “help”?

- Function
- Description
- Usage
- Command line arguments
- Input file format
- **Output file format**
- Data files
- Notes
- References
- Warnings
- Diagnostic error messages
- Exit status
- Known bugs
- See also
- Author(s)
- History
- Target users
- Comments

Graphic files

DOTMATCHER : Displays a thresholded dotplot of two sequences (EMBOSS)

Results:

[dotmatcher.ps](#) (22.27 Ko)

[dotmatcher.out](#)

[standard error file](#)

From now, this files will remain accessible for 10 days at:

<http://140.129.151.16/Pise/tmp/dotmatcher/A15049103578182/>

You can save them individually by the **Save file** function if needed.

Job summary

default format



Unix exact command:

```
dotmatcher -sequencea=sequencea.data -sformat=fasta -sequenceb=sequenceb.data  
-sformat=fasta -graph=postscript -auto -stdout -outfile=dotmatcher
```

Output file format

The output is a standard EMBOSS alignment file.

The results can be output in one of several styles by using the command-line qualifier **-aformat xxx**, where 'xxx' is replaced by the name of the required format. Some of the alignment formats can cope with an unlimited number of sequences, while others are only for pairs of sequences.

The available multiple alignment format names are: unknown, multiple, simple, fasta, msf, trace, srs

The available pairwise alignment format names are: pair, markx0, markx1, markx2, markx3, markx10, srspair, score

See: <http://www.uk.embnet.org/Software/EMBOSS/Themes/AlignFormats.html> for further information on alignment formats.

The default output format is 'markx0'.

Output files for usage example

File: hba_human.stretcher

```

#####
# Program:  stretcher
# Rundate:  Wed Sep 04 17:28:39 2002
# Align_format:  markx0
# Report_file:  hba_human.stretcher
#####
#=====
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 148
# Identity:      64/148 (43.2%)
# Similarity:    89/148 (60.1%)
# Gaps:          9/148 ( 6.1%)
# Score: 272
#
#
#=====

                10         20         30         40
HBA_HU V-LSPADKTNVKAAWGKVG AHAGEYGA EALERMFLSFPTTKTYFPHF-DL
      :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :
HBB_HU VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDL
                10         20         30         40

                50         60         70         80         90
HBA_HU SH----GSAQVKGHGKKVADALTNAVAHVDDMPNALSALS DLHAHKLRV
      :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :
HBB_HU STPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHC DKLHV
                50         60         70         80         90

```