Bioinformatics

EMBOSS软件包总结

龚丽 吴萍

2007.1.13

Bioinformatics

主要内容

- EMBOSS简介
- 基本的显示和编辑命令
- 关于核酸和蛋白的命令
- ALIGNMENT中的一些命令

一、EMBOSS简介

- **Emboss（The European Molecular Biology Open Software Suite）**该软件包源于1988年开始开发的EGCG系统，是一个开放源代码的序列分析软件包。该软件包含160多个小型程序，涵盖序列比对、快速数据库搜寻序列、蛋白质模序及结构域分析、表达序列标签(EST)分析、核酸序列分析(例如CpG岛的识别)等多个领域。

- **Emboss**的主页在**http://emboss.sourceforge.net/**

特点

- 综合性 自动识别 互联网提取
- 高效性 同时对一组大规模的序列进行分析
- 兼容性 linux系统和Mac系统 、 Windows系统
- 开放性 完全公开且免费

功能

1、序列比对

2、快速搜索数据库

3、蛋白质模块分析

4、核苷酸序列分析，包括寻找CpG岛，寻找重复序列等。

5、小基因组的编码区分析，如寻找开放阅读框，寻找EST等。

6、序列模式识别与翻译

EMBOSS软件包程序分类

- **Display**
- **Edit**
- **Information**
- **Alignment**
- **Feature tables**
- **Protein**

  如*Motif、Composition、2D-structure、 Mutation*
- **Nuclear**

  如*Motif、Composition、2D-structure、 Mutation、Translation、Repeat、Primersearch*等。

部分EMBOSS显示命令

显示命令	**showalign：** Displays a multiple sequence alignment
	showfeat： Show features of a sequence.
	showseq： Display a sequence with features, translation etc..
	infoseq： Displays some simple information about sequences
	sixpack： Display a DNA sequence with 6-frame translation and ORFs
	remap： Display sequence with restriction sites, translation etc.

例一、infoseq程序

- **infoseq**的功能为显示序列的一些简单信息。
- 显示的信息包括有**USA (Uniform Sequence Address**，统一序列称谓)、序列名、登录号、类型(核酸或蛋白质)、长度、**G+C**含量和其它描述。

如:在**infoseq**中输入苹果红皮基因**|AB279598.1|**
回车后得到如下结果：

USA	Name	Accession	Type	Length	GC%	Description
118615: AB279598.1	AB279598.1	AB279598	N	859	40.28	Malus x domestica MdMYBA mRNA for myb Transcription factor, complete cds

参数	值	意义
－sequence	seqall	(有缺口) 序列文件名和有选择的格式，或参考文献
－outfile	输出文件名	把序列的详情输出到一个文件中
－html	Y／N	把输出格式化为 HTML 表格
-[no]columns		使用 columns 选项，可使输出文件中的序列信息按列排列整齐
-delimiter	字符串	用来界定输出文件中的单个记录。字符串通常是一单字符，如空格、Tab 符、通道符等，可以是任何字符串。
-only		选择只显示某项信息，如-only -length，只显示序列的长度。
-[no]heading		不显示／显示表头
-usa		显示序列的 USA
-name		显示“name”列。

部分EMBOSS编辑命令

编辑命令	**biosed：** Replace or delete sequence sections
	descseq： Alter the name or description of a sequence
	extractfeat： Extract features from a sequence
	pasteseq： Insert one sequence into another
	seqret： Reads and writes (returns) sequences
	splitter： Split a sequence into (overlapping) smaller sequences
	trimest： Trim poly-A tails off EST sequences
	union： Reads sequence fragments and builds one sequence

例二、Seqret程序

- 转换核酸或蛋白序列格式。

  EMBL, FASTA, GCG, GDE, GENBANK, IG, NBRF, PIR, RAW and SWISSPROT 之间的转换，CLUSTAL, FASTA, MSF, NEXUS, PHYLIP and STOCKHOLM 比对结果之间的转换。

apple

red

skin

```
LOCUS       AB279598                859 bp    mRNA    linear   PLN 22-SEP-2007
DEFINITION  Malus x domestica MdMYBA mRNA for myb transcription factor,
            complete cds.
ACCESSION   AB279598
VERSION     AB279598.1  GI:157679482
KEYWORDS    .
SOURCE      Malus x domestica
  ORGANISM  Malus x domestica
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Rosales; Rosaceae; Maloideae; Malus.
REFERENCE   1
  AUTHORS   Ban,Y., Honda,C., Hatsuyama,Y., Igarashi,M., Bessho,H. and
            Moriguchi,T.
  TITLE     Isolation and functional analysis of a MYB transcription factor
            gene that is a key regulator for the development of red coloration
            in apple skin
  JOURNAL   Plant Cell Physiol. 48 (7), 958-970 (2007)
   PUBMED   17526919
REFERENCE   2  (bases 1 to 859)
  AUTHORS   Honda,C., Ban,Y., Bessho,H. and Moriguchi,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-OCT-2006) Chikako Honda, National Institute of Fruit
            Tree Science; 2-1 Fujimoto, Tsukuba, Ibaraki 305-8605, Japan
            (E-mail:hondac@affrc.go.jp, Tel:81-29-838-6437)
FEATURES             Location/Qualifiers
```

Result: plotorf.fasta View: Raw View View Edit Download Friendly Print

Choose program: biosed (Replace or delete sequence sections)

Do analysis

>AB279598.1 AB279598.1 Malus x domestica MdMYBA mRNA for myb transcription factor, complete cds
AGTGGGTAGCAGGCAAAAGAATAGCTAAGCTTAGCTGCTAGCAGATAAGAGATGGAGGGA
TATAACGAAAACCTGAGTGTGAGAAAAGGTGCCTGGACTCGAGAGGAAGACAATCTTCTC
AGGCAGTGCGTTGAGATTCATGGAGAGGGAAAGTGGAACCAAGTTTCATACAAAGCAGGC
TTAAACAGGTGCAGGAAGAGCTGCAGACAAAGATGGTTAAACTATCTGAAGCCAAATATC
AAGAGAGGAGACTTTAAAGAGGATGAAGTAGATCTTATAATTAGACTTCACAGGCTTTTG
GGAAACAGGTGGTCATTGATTGCTAGAAGACTTCCAGGAAGAACAGCAAATGCTGTGAAA
AATTATTGGAACACTCGATTGCGGATCGATTCTCGCATGAAAACGGTGAAAAATAAATCT
CAAGAAATGAGAAAGACCAATGTGATAAGACCTCAGCCCCAAAAATTCAACAGAAGTTCA
TATTACTTAAGCAGTAAAGAACCAATTCTAGACCATATTCAATCAGCAGAAGATTTAAGT
ACGCCACCACAAACGTCGTCGTCAACAAAGAATGGAAATGATTGGTGGGAGACCTTGTTA
GAAGGCGAGGATACTTTTGAAAGAGCTGCATATCCCAGCATTGAGTTAGAGGAAGAACTC
TTCACAAGTTTTTGGTTTGATGATCGACTGTCGCCAAGATCATGCGCCAATTTTCCTGAA
GGACAAAGTAGAAGTGAATTCTCCTTTAGCACGGACCTTTGGAATCATTCAAAAGAAGAA
TAGCTAGAGAAAATGATTCTCACTTCTTTATTATCATCTAGCTTGTGTTCTATTATTTTC
CTTGCTTGTAAATGTGGCA

```
ID   AB279598   standard; DNA; UNC; 859 BP.
DE   Malus x domestica MdMYBA mRNA for myb transcription factor, complete cds
SQ   Sequence 859 BP; 302 A; 144 C; 202 G; 211 T; 0 other;
     agtgggtagc aggcaaaaga atagctaagc ttagctgcta gcagataaga gatggaggga        60
     tataacgaaa acctgagtgt gagaaaaggt gcctggactc gagaggaaga caatcttctc       120
     aggcagtgcg ttgagattca tggagaggga aagtggaacc aagtttcata caaagcaggc       180
     ttaaacaggt gcaggaagag ctgcagacaa agatggttaa actatctgaa gccaaatatc       240
     aagagaggag actttaaaga ggatgaagta gatcttataa ttagacttca caggcttttg       300
     ggaaacaggt ggtcattgat tgctagaaga cttccaggaa gaacagcaaa tgctgtgaaa       360
     aattattgga acactcgatt gcggatcgat tctcgcatga aaacggtgaa aaataaatct       420
     caagaaatga gaaagaccaa tgtgataaga cctcagcccc aaaaattcaa cagaagttca       480
     tattacttaa gcagtaaaga accaattcta gaccatattc aatcagcaga agatttaagt       540
     acgccaccac aaacgtcgtc gtcaacaaag aatggaaatg attggtggga gaccttgtta       600
     gaaggcgagg atacttttga aagagctgca tatcccagca ttgagttaga ggaagaactc       660
     ttcacaagtt tttggtttga tgatcgactg tcgccaagat catgcgccaa ttttcctgaa       720
     ggacaaagta gaagtgaatt ctcctttagc acggaccttt ggaatcattc aaaagaagaa       780
     tagctagaga aaatgattct cacttcttta ttatcatcta gcttgtgttc tattattttc       840
     cttgcttgta aatgtggca                                                   859
//
```


Motif
protein
nuclear
fuzzpro
preg
helixturnhelix
patmatmotifs
fuzztraa
fuzznuc
dreg
Composition
backtranseq
octanol
pepinfo
pepwindow
comseq
freak
banana
sirna
wordcount

例三、fuzztran程序

功能：采用特定的蛋白质序列（如一个短的MOTIF）寻找核苷酸序列上与之相匹配的序列。在寻找之前先将核苷酸序列翻译成蛋白质序列。

参数设置

参数	要求
［-sequence］	可读的核苷酸序列
[-pattern]	X：表示任意一个氨基酸均可以
	[ALT]：该位上的氨基酸可为A,L,或T
	{AM}：除了A或M, 其它氨基酸均可
	X(2)：表示2个任意氨基酸
	X(2,4)：表示2-4个任意氨基酸
	如：[DE](2)HS{P}X(2,4)C
[-outfile]	输出文件名为*.fuzztran

```
% fuzztran -opt
Protein pattern search after translation
Input nucleotide sequence(s): tembl:Z46957
Search pattern: RA
Translation frames
         1 : 1
         2 : 2
         3 : 3
         F : Forward three frames
        -1 : -1
        -2 : -2
        -3 : -3
         R : Reverse three frames
         6 : All six frames
Frame(s) to translate [1]: f
```

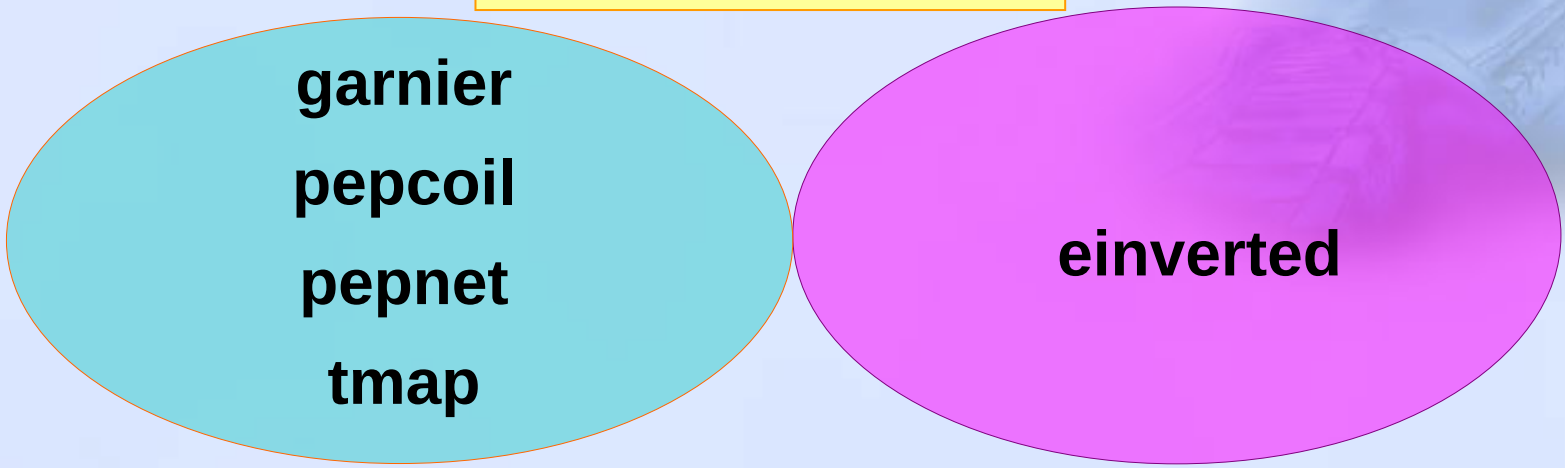
netic codes
0 : Standard
1 : Standard (with alternative initiation codons)
2 : Vertebrate Mitochondrial
3 : Yeast Mitochondrial
4 : Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma/Spiroplasma
5 : Invertebrate Mitochondrial
6 : Ciliate Macronuclear and Dasycladacean
9 : Echinoderm Mitochondrial
10 : Euplotid Nuclear
11 : Bacterial
12 : Alternative Yeast Nuclear
13 : Ascidian Mitochondrial
14 : Flatworm Mitochondrial
15 : Blepharisma Macronuclear
16 : Chlorophycean Mitochondrial
21 : Trematode Mitochondrial
22 : Scenedesmus obliquus
23 : Thraustochytrium Mitochondrial
de to use [0]: 回车
tput report [z46957.fuzztran]: 回车

```
#=====================================
#
# Sequence: Z46957     from: 1   to: 1493
# HitCount: 9
#
# Pattern_name Mismatch Pattern
# pattern1             0 RA
# TransTable: 0
# Frames: F
#
#=====================================
```

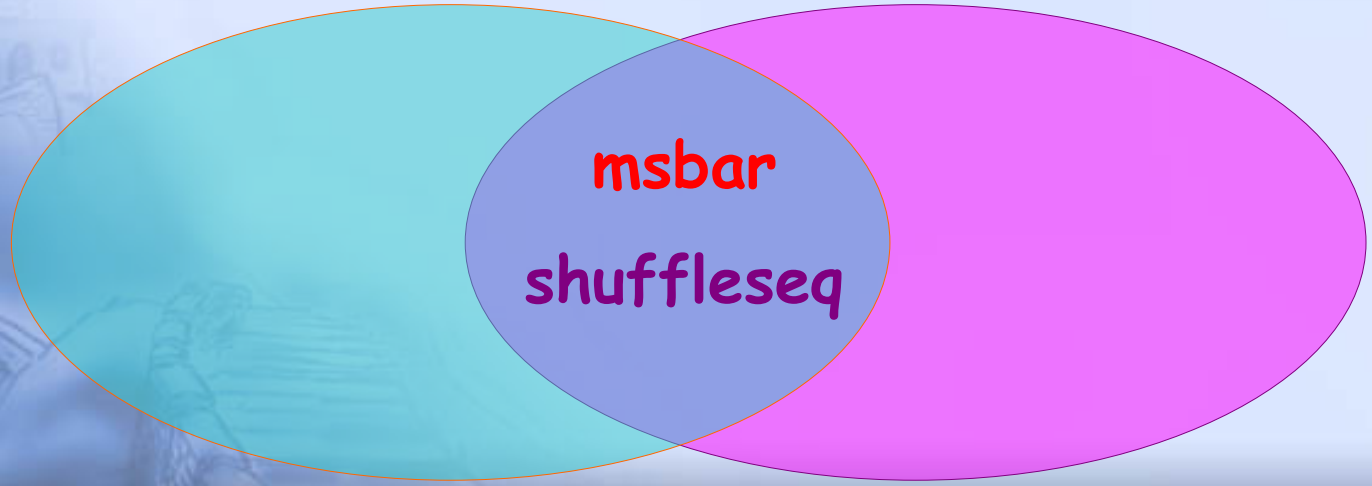
结果

Start	End	Score	Pattern_name	Mismatch	Frame	PStart	PEnd	Translation
97	102	2	pattern1	.	1	33	34	RA
133	138	2	pattern1	.	1	45	46	RA
421	426	2	pattern1	.	1	141	142	RA
625	630	2	pattern1	.	1	209	210	RA
835	840	2	pattern1	.	1	279	280	RA
919	924	2	pattern1	.	1	307	308	RA
227	232	2	pattern1	.	2	76	77	RA
752	757	2	pattern1	.	2	251	252	RA
72	77	2	pattern1	.	3	24	25	RA

2D-Structure



Mutation



例四、msbar程序

- 功能：对一段核苷酸序列设置突变

参数	要求
［-sequence］	可识别的序列
［-count］	突变的次数，需为整数
［-point］	点突变类型 **0: none 1:any of the following 2:insertions 3:deletions 4:changes 5:duplications 6:moves**
［-block］	批量突变的类型 **0: none 1:any of the following 2:insertions 3:deletions 4:changes 5:duplications 6:move**
［-codon］	密码子突变，仅适用于核苷酸序列 **0: none 1:any of the following 2:insertions 3:deletions 4:changes 5:duplications 6:move**
[-outfile]	输出文件名为***.format**

Input

* sequence: ----------

OR upload file from local disk: 浏览...

OR paste into window:

```
 cgatttggct acatgacatc aaccatatca gcaaaagtga tacgggtatt
atttttgccg
```

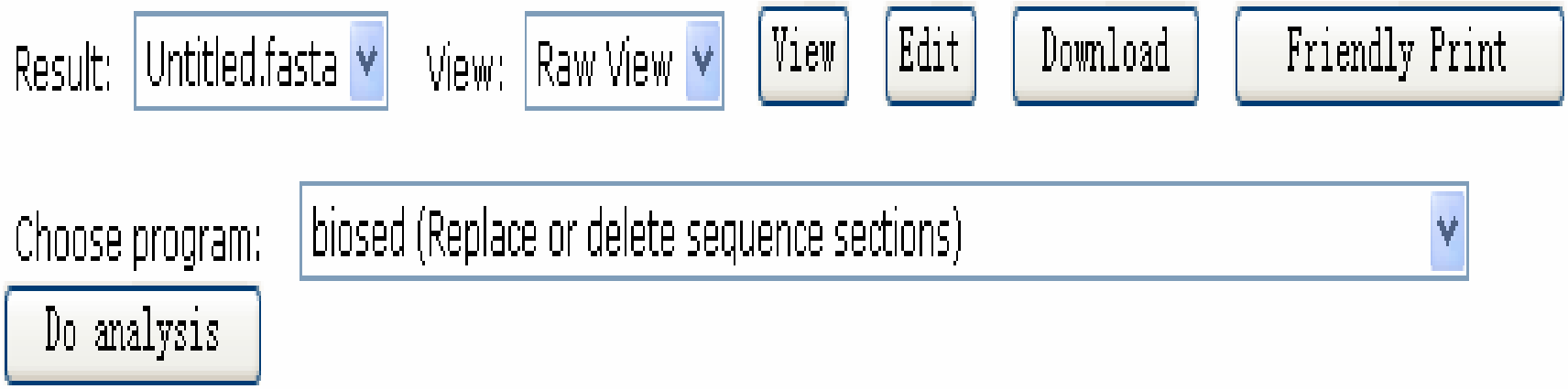
Other sequences that the mutated result should not match: ----------

Output section

* save result in directory:	Work Directory
* outseq: (bio:seq:fasta)	Untitled.fasta

Basic Options

* Number of times to perform the mutation operations:	2
* Types of point mutations to perform:	Insertions
* Types of block mutations to perform:	None
Types of codon mutations to perform:	None



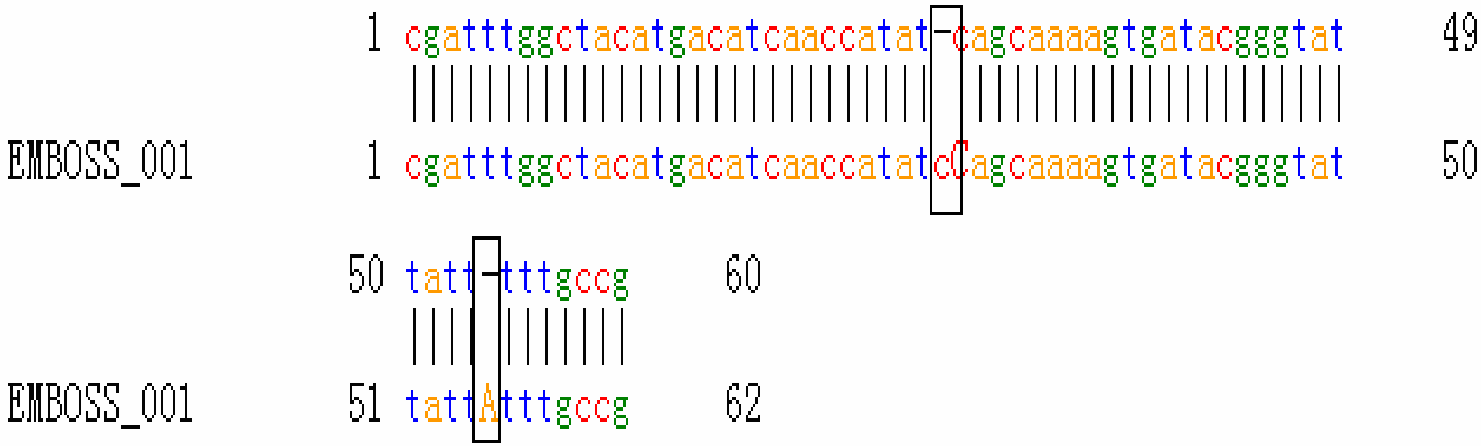
>EMBOSS_001
cgatttggctacatgacatcaaccatatcCagcaaaagtgatacgggtattatttAtttgc
cg

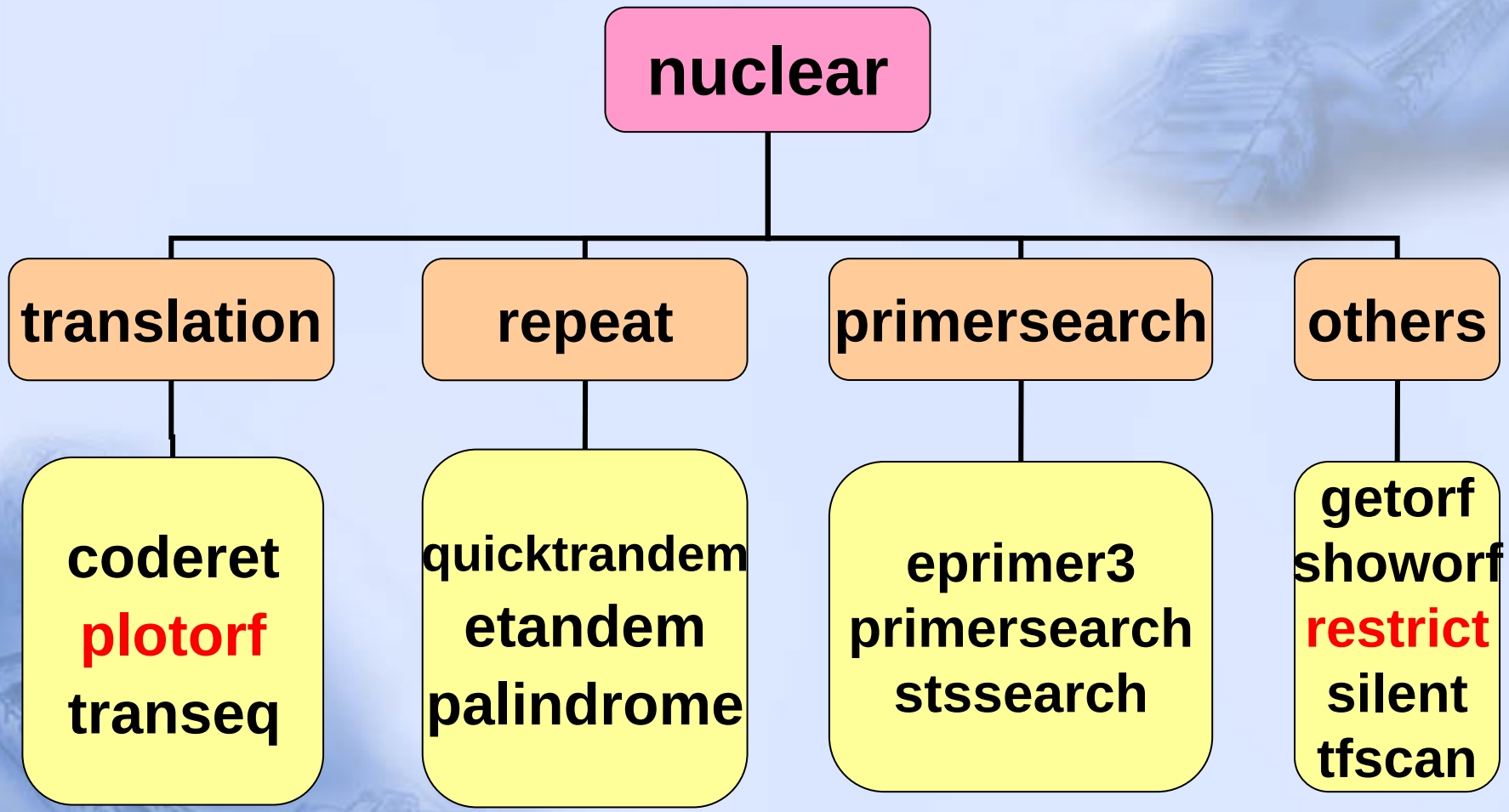
Result: Untitled.pair.2008-01-10 20:39 PM View: Pair-alignment view View Edit

Download Friendly Print

Pairwise Alignment Result

LENGTH	SCORE	IDENTITY	SIMILARITY	GAPS
62	280.0	60/62 (96.8%)	60/62 (96.8%)	2/62 (3.2%)




nuclear
translation
repeat
primersearch
others
coderet
plotorf
transeq
quicktrandem
etandem
palindrome
eprimer3
primersearch
stssearch
getorf
showorf
restrict
silent
tfscan

例五、plotorf程序

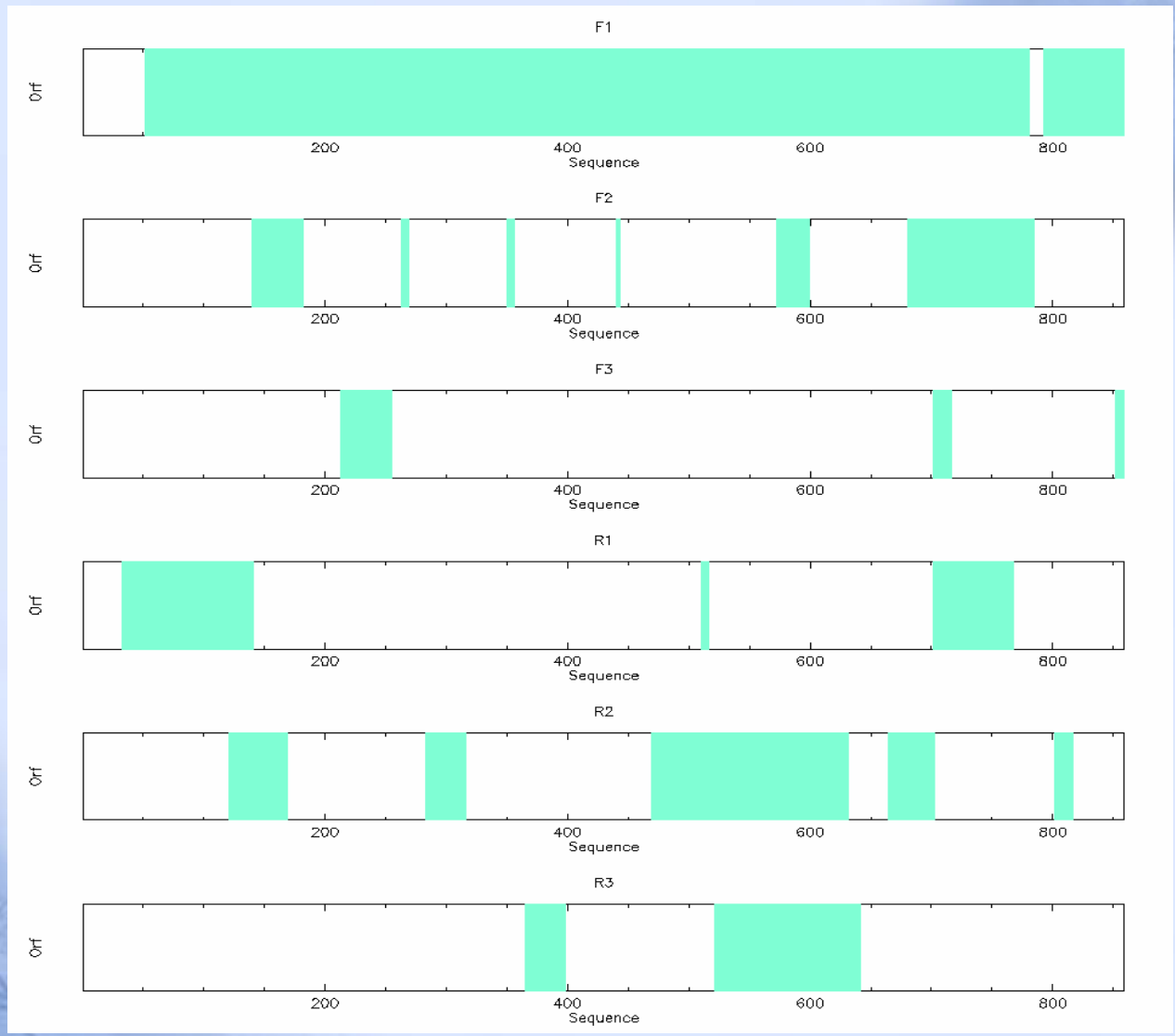
- 功能：寻找核酸序列中的读码框。

plotorf 命令之后可以使用的主要参数如下所示：

参数	值	意义
-sequence		核酸序列文件名和可选择的格式。
-graph	ps、pgl、p7470、ata、ng	输出的图像类型，包括有 ps，hpgl，hp7470，data，png 等
-start	字符串	启始密码子，可以接受任何字符串。
-stop	字符串	终止密码子，可以接受任何字符串。
-sbeginl	整数	使用的序列的开始位
-sendl	整数	使用的序列结束位

- 正反方向所有**6种**读码框都以蓝框显示出来。显示的是启始密码子和终止密码子之间的区域。不包含启始密码子的真核生物基因序列。所以只有在处理**原核序列或mRNA真核序列**时本程序才会真正有用。
- 程序可以处理的文件格式可以是**任何核酸序列格式**。
- **默认的启始密码子是：ATG。**
- **默认的终止密码子是：TAA、TAG和TGA。**
- **实际应用时，可以使用参数“-start”和“-stop”指定自己想使用的密码子。**

以apple red skin转录因子的一段cds序列|AB279598.1|为例



例六、restrict程序

- 功能：寻找限制性酶切位点。

参数	要求
[- sequence]	可读的核苷酸序列
[-sitelen]	限制性酶切位点识别的最小核苷酸长度。一般取2-20之间的整数，默认为4。
[-enzymes]	酶的名称，多个酶之间用逗号隔开。如： **'HincII,hinfI,ppiI,hindiii'**
[-outfile]	输出文件名为***.restrict**

```
% restrict
Finds restriction enzyme cleavage sites
Input nucleotide sequence(s): tembl:x65923
Minimum recognition site length [4]:
Comma separated enzyme list [all]:
Output report [x65923.restrict]:
```

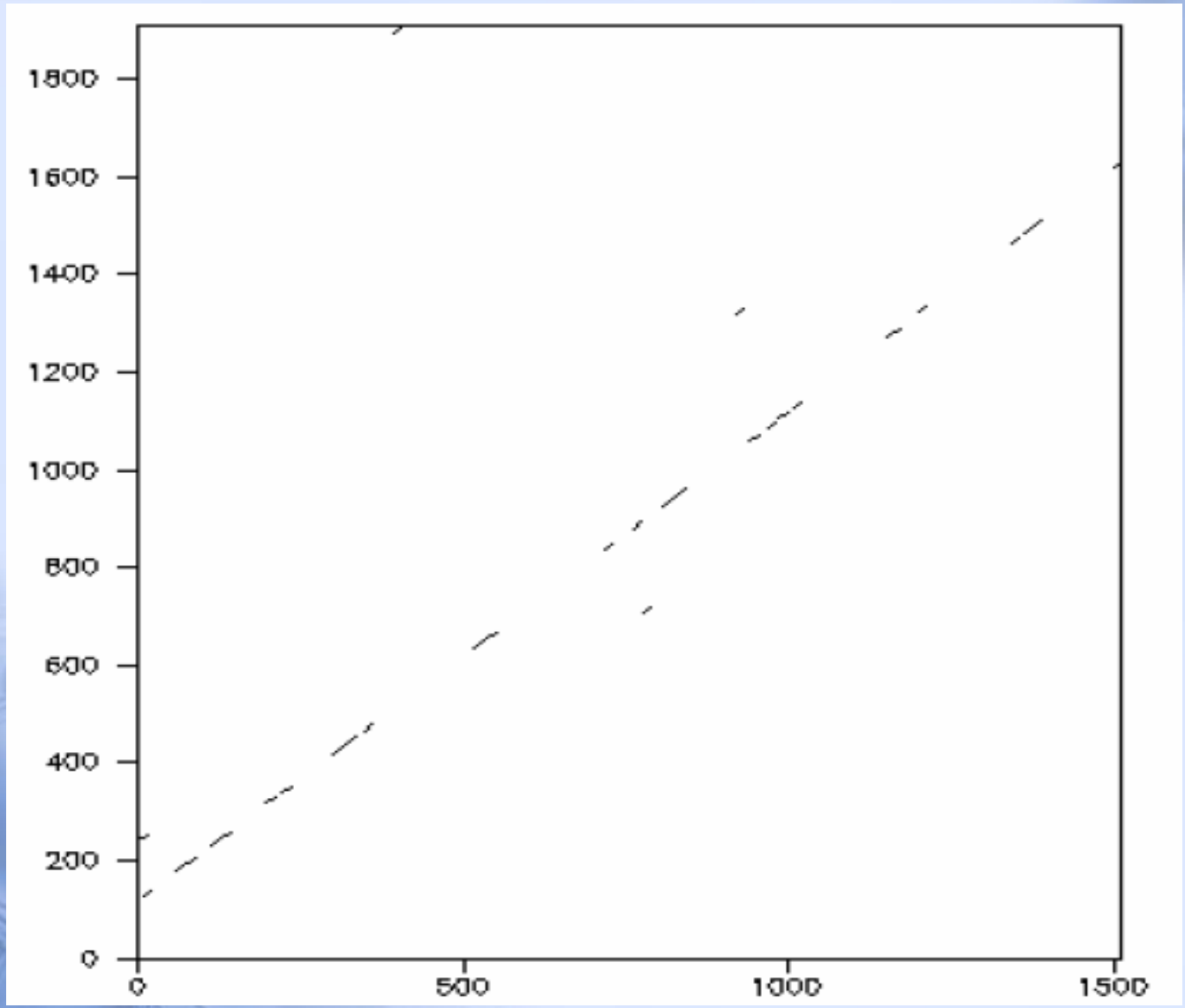
#
Sequence: X65923   from: 1   to: 518
HitCount: 54
#
Minimum cuts per enzyme: 1
Maximum cuts per enzyme: 2000000000
Minimum length of recognition site: 4
Blunt ends allowed
Sticky ends allowed
DNA is linear
Ambiguities allowed
#
#=====================================

Start	End	Enzyme_name	Restriction_site	5prime	3prime	5primerev	3primerev
11	14	TaqI	TCGA	11	13	.	.
28	25	AciI	CCGC	25	27	.	.
36	31	BseYI	CCCAGC	31	35	.	.
38	41	AciI	CCGC	38	40	.	.
44	40	BceAI	ACGGC	25	27	.	.
71	81	BsiYI	CCNNNNNNNGG	77	74	.	.
71	74	AciI	CCGC	71	73	.	.
73	76	Hin6I	GCGC	73	75	.	.
73	76	HhaI	GCGC	75	73	.	.
77	81	EcoRII	CCWGG	76	81	.	.
77	81	BssKI	CCNGG	76	81	.	.
94	97	TaqI	TCGA	94	96	.	.
103	106	HpaII	CCGG	103	105	.	.
105	108	HaeIII	GGCC	106	106	.	.
107	111	EcoRII	CCWGG	106	111	.	.
107	111	BssKI	CCNGG	106	111	.	.

Alignment

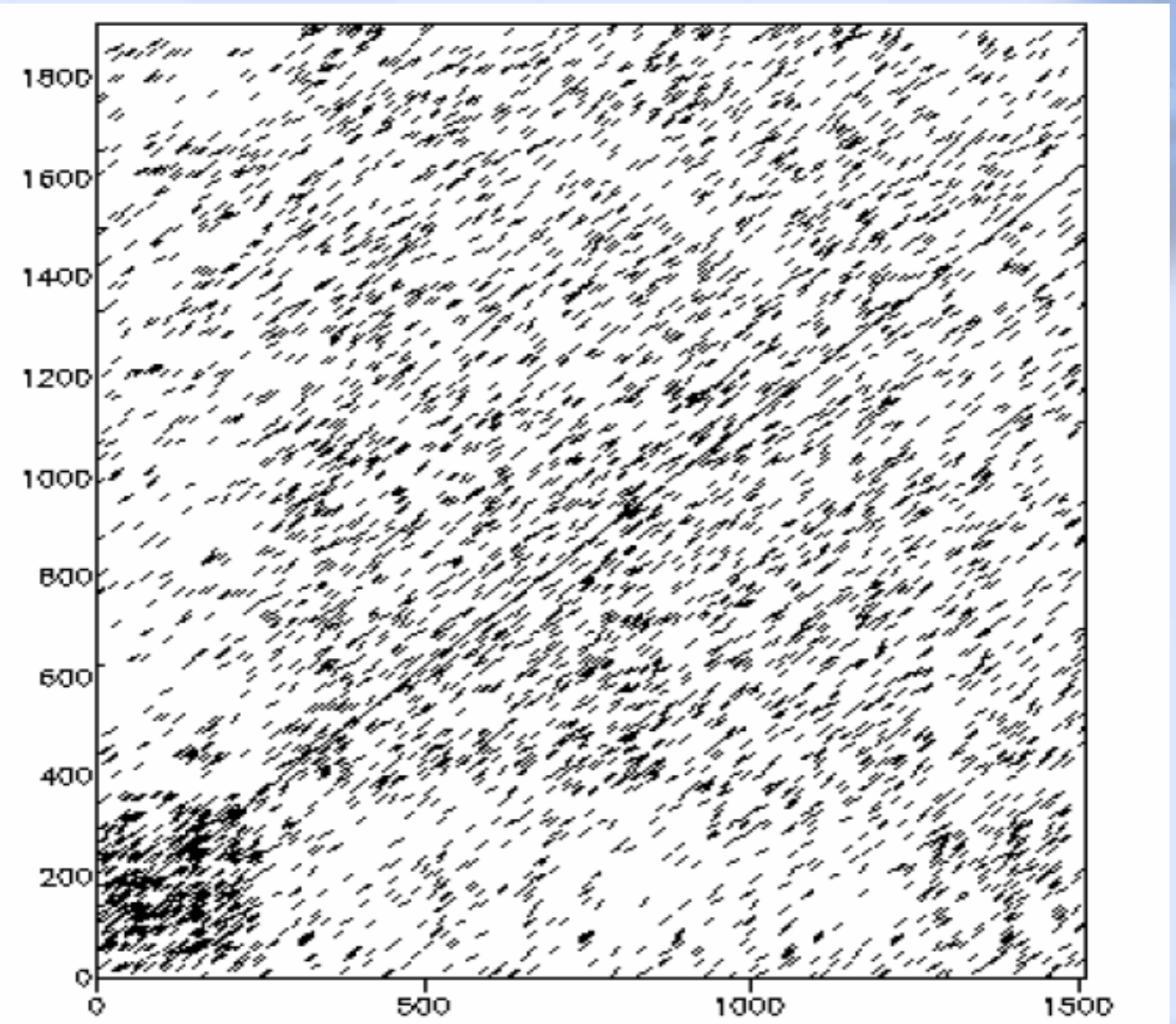
1、Dottup

- Dottup 程序执行的方式是在给定序列长度下(word size) 逐一比对，找到完全相同的序列片段(即：word size 为10，且有10个连续的 residues 完全相同)。其比对步骤为给定两条序列，设word=10，程序会依 序比对两条序列的第 1-10，2-11，3-12，一段一段交互比，只要在相对位置有相同的 residue 就以一个 "点" (dotplot) 来表示，找到两段序列有 10 个 residues完全相同的片段，就会在结果中显示出来。


1800
1600
1400
1200
1000
800
600
400
200
0
0
500
1000
1500

2、 Dotmatcher

Dotmatcher 执行的方式则是设定 threshold 的方式来图示序列的相似度。当两条序列在给定 的 word size 中找不到完全相同的序列（如 word size 为10，但 10个 residues 中只有 7个 residue 相同）时可以采用 dotmatcher 进行 dotplot 分析。其比对步骤为给定两条序列，设 word=6，Threshold=8， 程序会依序比对两条序列的第 1-6，2-7，3-8，一段一段交互比对，只要在相对位置有相同的 residue 就以一个"点"来表示，两段序列 residues 比对得分≧8，就会在结果中显示出来。


1800
1600
1400
1200
1000
800
600
400
200
0
0
500
1000
1500

needle

Pairwise Alignment Result

LENGTH	SCORE	IDENTITY	SIMILARITY	GAPS
432	1579.0	281/432 (65.0%)	281/432 (65.0%)	33/432 (7.6%)

```
x56325_cds_1       1 atggtgctctctgcagctgacaaaaccaacatcaagaactgctgggggaa      50
                     .|||||||.|||.|.|..||||||..|.||||.||||.....|||||.|||
j00153_cds_1       1 gtggtgctgtctcctgaggacaaggctaacaccaaggcggtctgggagaa      50

x56325_cds_1      51 gattggtggccatggtggtgaatatggcgaggaggccctacagaggatgt     100
                     ..|||||.|.|||...||.||..||||.|..|||||||||..||||.||||
j00153_cds_1      51 agttggcgaccacactgctggctatgccacggaggccctggagagcatgt     100

x56325_cds_1     101 tcgctgccttccccaccaccaagacctacttctctcacattgatgtaagc     150
                     |     |||               |||||.|  |||||||.|.|...|.||.
j00153_cds_1     101 t-----cct-------------gaccttc--ctctcacttggccctgagt     130

x56325_cds_1     151 cccggctctgcccaggtcaaggctcacggcaagaaggttgctgatgccct     200
                     ...||||||.||||||.||||||...||..||||||.||||.||.||..|.||
j00153_cds_1     131 tatggctcagcccagatcaagaaacaatgcaagtaggtggccgacacgct     180

x56325_cds_1     201 ggccaaagctgcagaccacgtcgaagacctgcctggtgccctgtccactc     250
                     |.||||.||.|..|.||||.|.||.|||.||||...||...||||.....
j00153_cds_1     181 gaccaatgccgtggtccacttagatgacatgcccaatgatgtgtctgagg     230

x56325_cds_1     251 tgagcgacctgcatgcccacaaactgcgtgtggatcctgtcaacttcaag     300
                     ||||..|.|||||||.||||.|.|||.|.|||||.||.|.||||.|||.|
j00153_cds_1     231 tgaggaagctgcatgtccacgagctgtgggtggacccaggcaacatcagg     280
```

water

Pairwise Alignment Result

LENGTH	SCORE	IDENTITY	SIMILARITY	GAPS
429	1579.0	281/429 (65.5%)	281/429 (65.5%)	31/429 (7.2%)

```
j00153_cds_1       2 tggtgctgtctcctgaggacaaggctaacaccaaggcggtctgggagaaa      51
                     |||||||.|||.|.|..||||||..|.||||.||||.....||||||.|||.
x56325_cds_1       2 tggtgctctctgcagctgacaaaaccaacatcaagaactgctgggggaag      51

j00153_cds_1      52 gttggcgaccacactgctggctatgccacggaggccctggagagcatgtt     101
                     .||||.|.|||...||.||..||||.|..|||||||||..||||.|||||
x56325_cds_1      52 attggtggccatggtggtgaatatggcgaggaggccctacagaggatgtt     101

j00153_cds_1     102 -----cct-------------gaccttc--ctctcacttggccctgagtt     131
                         |||               ||||||.|  |||||||.|.|...|.||..
x56325_cds_1     102 cgctgccttccccaccaccaagacctacttctctcacattgatgtaagcc     151

j00153_cds_1     132 atggctcagcccagatcaagaaacaatgcaagtaggtggccgacacgctg     181
                     ..||||||.||||||.||||||...||..||||||.||||.||.||..|.|||
x56325_cds_1     152 ccggctctgcccaggtcaaggctcacggcaagaaggttgctgatgccctg     201

j00153_cds_1     182 accaatgccgtggtccacttagatgacatgcccaatgatgtgtctgaggt     231
                     .||||.||.|..|.||||.|.||.|||.||||...||...||||.....|
x56325_cds_1     202 gccaaagctgcagaccacgtcgaagacctgcctggtgccctgtccactct     251

j00153_cds_1     232 gaggaagctgcatgtccacgagctgtgggtggacccaggcaacatcaggt     281
                     |||..|.|||||||.||||.|.|||.|.|||||.||.|.||||.|||.||
x56325_cds_1     252 gagcgacctgcatgcccacaaactgcgtgtggatcctgtcaacttcaagt     301
```

needle 和 water 的区别

- **needle**是用来寻找两条序列整体的最佳排列方式**(global alignment)**； 而**water**则是用来寻找两段序列间最佳排列区域**(local alignment)** 。
- **needle**是将两条序列从头到尾完全并列的分析，常会在序列中间插入许多的**gap**；**water** 则只列出最相似的区域两条序列，其余略去不列出来。

程序帮助命令

- **wossname keyword** 显示在描述信息中包含该关键字的所有程序列表
- **tfm programname** 显示该程序的用法
- **seealso programname** 显示与该程序相关的其它程序
- **programname –help** 给出该程序的可用参数

EMBOSS软件包，包罗万象真的好，
核酸蛋白都适用，功能强大效率高。
比对进化设引物，翻译酶切找重复，
程序名称不记得，wossname帮你找。
程序命令不会用，tfm 把你教。
参数设置技巧高，点点滴滴要记牢，
EMBOSS是法宝，活学活用不愁了。

Thanks !