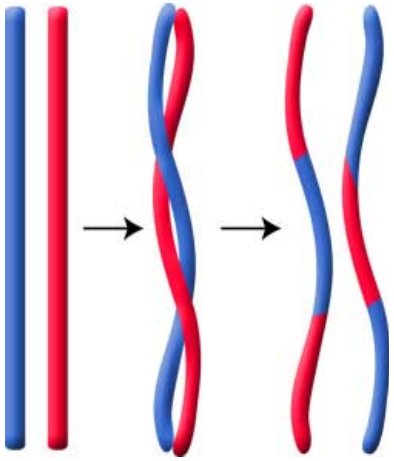
http://petang.cgu.edu.tw/bioinformatics/index.htm



Bioinformatics

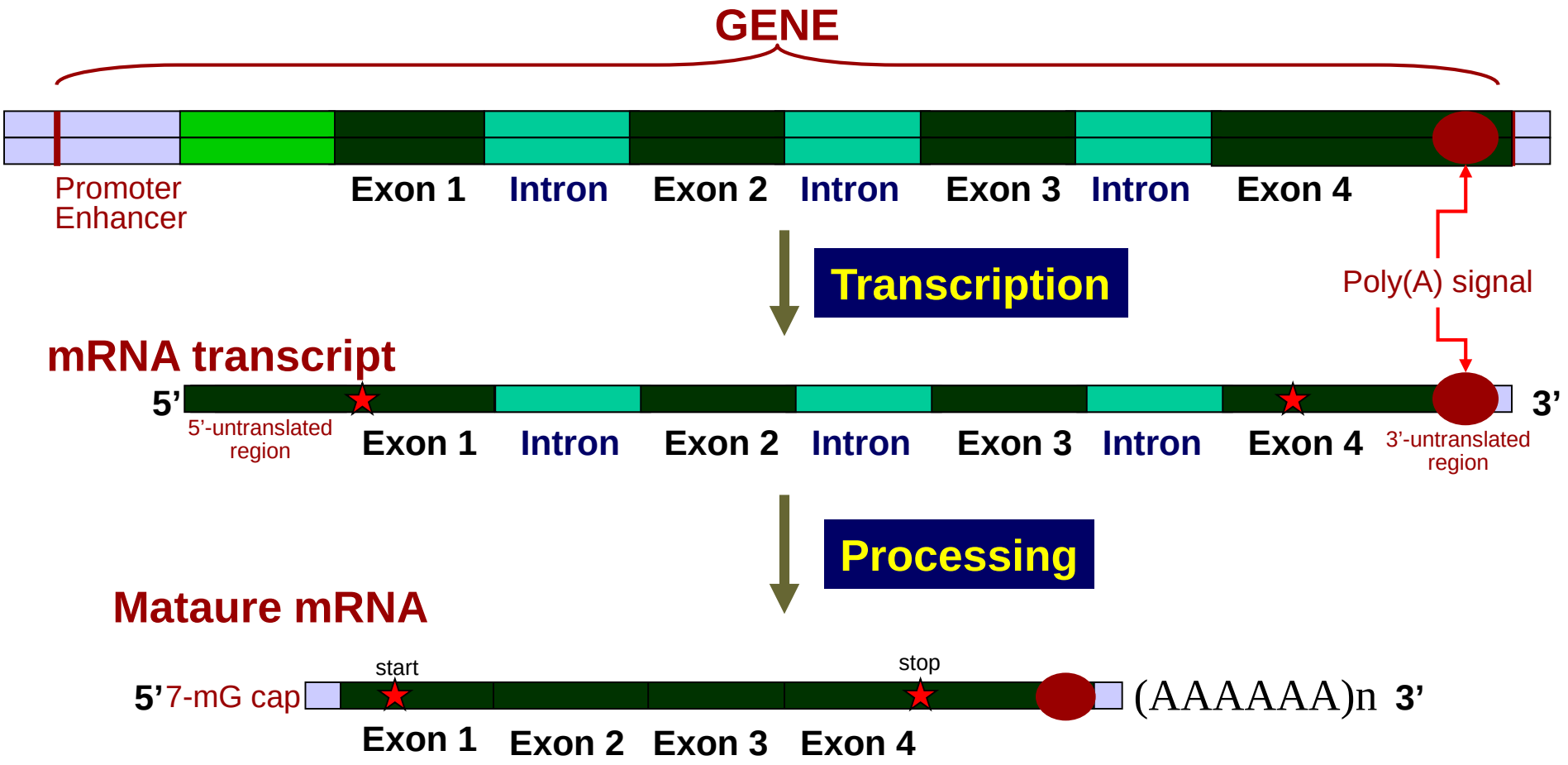
Lecture 6 – Nucleic Acids Sequence Analysis

AGTCCGCGAATACAGGCTCGGT AGTCCGCGAATACAGGCTCGGT

Petrus Tang, Ph.D. (鄧致剛)
Graduate Institute of Basic Medical Sciences
and
Bioinformatics Center, Chang Gung University.
petang@mail.cgu.edu.tw
EXT: 5136

助教：
蔡智宇(分機5690)

The Organization of an Eukaryotic Gene



Gene identification involves 4 main stages

Find the putative coding region(s) in the sequence

Open reading frame

Find non-coding features of interest in the sequence

CpG islands
Tandemly and dispersed repeats
Promoter regions (TATA box, cap signal, CCAAT-box)
Transcription factors, Poly-A sites

Determine the exon-intron organization

Branch point signal
CT(G,A)A(C,T)

5' and 3' splice sites:
AG/GUAAGU----------PyPyPyPyPyPyPyPy-CAG/G

Identify the gene

motif, signal and pattern
Blast, FASTA
Functional studies

Function	Command	GCG	SeqWEB
Sequence manipulation	Reverse 	+	+
ORF Searching	Frames 	+	+
	Map 	+	+
	Translate 	+	+
Mapping (restriction sites)	Map 	+	+
	(-minc)	+	+
	(-maxc)	+	+
	Mapsort 	+	+
	(-exclude)	+	-
	(-digest)	+	-
	Mapplot 	+	+
Mapping (transcription factors)	Map tfsites 	+	-

REVERSE

EXAMPLE

[Previous | Top | Next]

The file test.seq contains all of the legitimate GCG sequence characters. Here is a session using Reverse to show how they would look on the opposite strand:

```
% reverse

  REVERSE of what sequence ?  test.seq

                  Begin (* 1 *) ?
                End (*   389 *) ?

  Do you want to:

       1) reverse only
       2) complement only
       3) reverse and complement

  Please choose one (* 3 *):

  What should I call the output file (* test.rev *)

%
```

OUTPUT

[Previous | Top | Next]

Here is the output file test.rev:

```
!!NA_SEQUENCE 1.0
  REVERSE-COMPLEMENT of: test.seq  check: 3365  from: 1  to: 389

This sequence contains every symbol in the alphabet of
legitimate GCG sequence characters (Appendix III).

test.rev  Length: 389  October 21, 1996 10:15  Type: N  Check: 8096

       1  .~....rxwb aasy...nk. m..dc..hgv t...~..RXW BAASY...NK

      51  .M..DC..HG VTYYATYAYT ATCACTATTA YTSCTGTTGC TCTTCRTAGT

     101  AATAXXXCCA XACCACTACG ACAACXGTCG TTGTGGTAGT XSWRCTXGAG

     151  CTACTCGATG AGGAAGAXCK YCKCCKTCKY CTXCGCCTTC TCCGCCGTCG

     201  GCGACGYTGC TGTTGXGGCG GTGGGGGAGG RTTGTTATTC ATXARYARCA

     251  RTARXAGYAA CAGTAGGAGA AGTAACAAYT TCTTTTTDAT TATGATAATR

     301  TGGTGATGXC CCCCTCCGCC ACCRAAGAAA AAYTCCTCTT CRTCGTCATC

     351  RCAGCAACAR TYGTTATTGT CATCXGCCGC TGCGGCAGC
```

FRAMES

```
% frames -RARe

 Plot FRAMES for what sequence?  Bacterial:EcoOmpa

             Begin (* 1 *) ?
           End (*  2270 *) ?

 Mark rare codons at or below what threshold (* 0.0 *) ?  0.1

 When your LaserWriter attached to tty07 is ready, press <Return>.

%


Minimal Syntax: % frames [-INfile1=]Bacteria:EcoOmpa -Default

Prompted Parameters:

-BEGin=1 -END=2270        sets the range of interest

Local Data Files:

-TRANSlate=translate.txt  defines start and stop codons
-MARk1=ecoompa.mrk        marks regions of interest on the plot

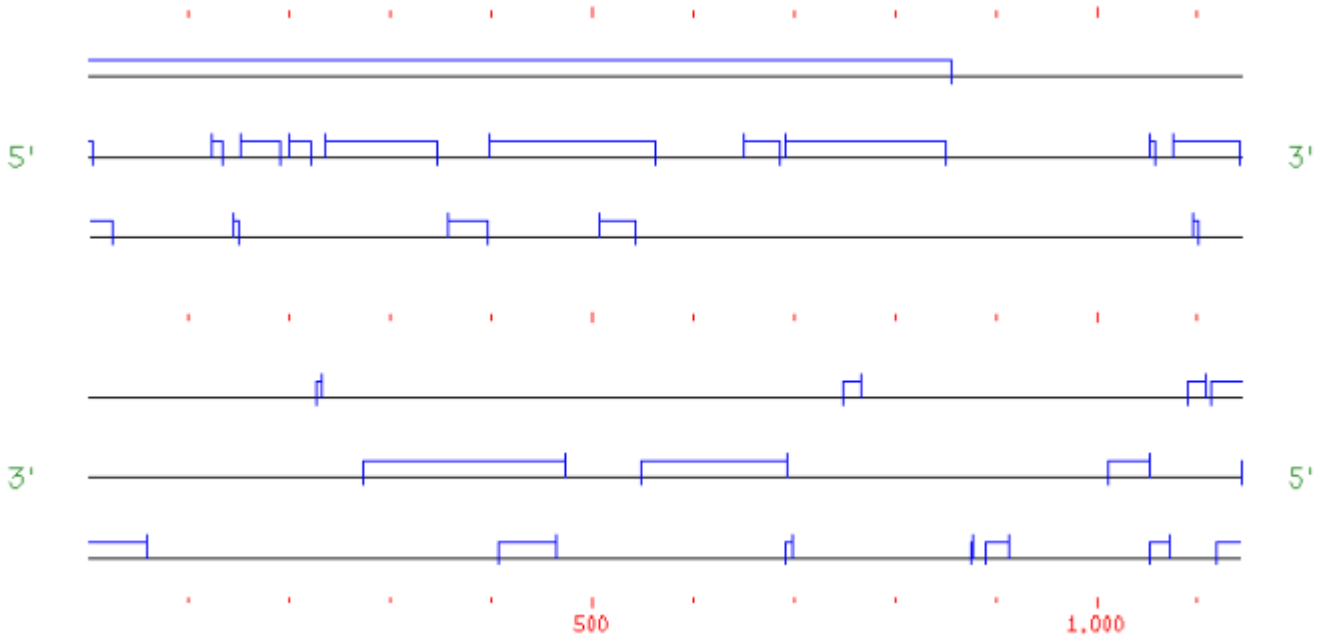
Optional Parameters:

-ALL               shows all start and stop signals, not just open frames
-RARe=ecohigh.cod  marks rare codons according to codon frequency table
-THReshold=0.0     sets threshold at or below which rare codons are shown
-DENsity=2270      sets the number of bases per 100 platen units

All GCG graphics programs accept these and other switches. See the Using
Graphics chapter of the USERS GUIDE for descriptions.

-FIGure[=FileName]  stores plot in a file for later input to FIGURE
-FONT=3             draws all text on the plot using font 3
-COLor=1            draws entire plot with pen in stall 1
-SCAle=1.2          enlarges the plot by 20 percent (zoom in)
-XPAN=10.0          moves plot to the right 10 platen units (pan right)
-YPAN=10.0          moves plot up 10 platen units (pan up)
-PORtrait           rotates plot 90 degrees
```

FRAMES output



MAP

```
% map

(Linear) MAP of what sequence ?  gamma.seq

                Begin (* 1 *) ?  2161
              End (* 11375 *) ?  2600

Select the enzymes:  Type nothing or "*" to get all enzymes. Type "?"
for help on which enzymes are available and how to select them.


                                        Enzyme(* * *):

What protein translations do you want:

     a) frame 1   b) frame 2   c) frame 3
     d) frame 4   e) frame 5   f) frame 6

     t)hree forward frames   s)ix frames   o)pen frames only

     n)o protein translation   q)uit

Please select (capitalize for 3-letter) (* t *):  s

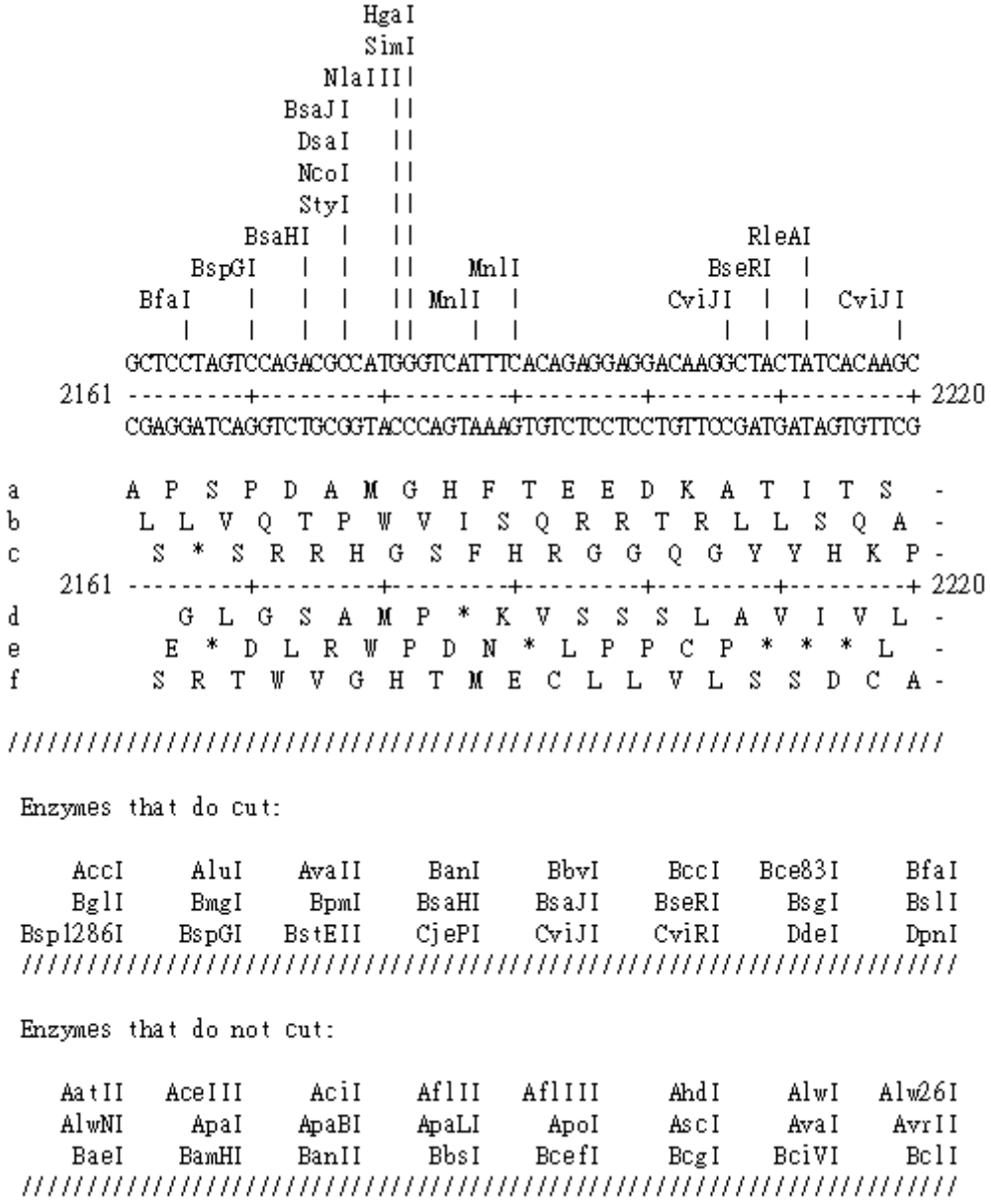
What should I call the output file (* gamma.map *) ?

Mapping ........................

Writing ........ ..
MAP complete with:

  Sequence Length:     440
   Enzymes Chosen:     216
   Cutsites found:     116
         CPU time:   00.91
```

MAP output



//

Enzymes that do cut:

AccI	AluI	AvaII	BanI	BbvI	BccI	Bce83I	BfaI
BglI	BmgI	BpmI	BsaHI	BsaJI	BseRI	BsgI	BslI
Bsp1286I	BspGI	BstEII	CjePI	CviJI	CviRI	DdeI	DpnI

//

Enzymes that do not cut:

AatII	AceIII	AciI	AflII	AflIII	AhdI	AlwI	Alw26I
AlwNI	ApaI	ApaBI	ApaLI	ApoI	AscI	AvaI	AvrII
BaeI	BamHI	BanII	BbsI	BcefI	BcgI	BciVI	BclI

//

Select enzymes for MAP:

Type "*" to select all enzymes.
Type "**" to select all enzymes including isoschizomers.
Type individual names like "AluI" to select specific enzymes.
Type "?" to see this message and all available enzymes.
Type "??" to see the available enzymes AND their recognition sites.
Type "?A*" to see what enzymes start with "A."
Type "A*" to select all enzymes starting with "A."
Type parts of names like "Al*" to select all enzymes starting with "AL."
Type "~A*" to unselect all selected enzymes starting with "A."
Type "/*" to see what enzymes you have selected so far.
Type "#" to select no enzymes at all.

Press <Return> after each selection.
Press <Return> and nothing else to end your selections.
Spaces are allowed; upper and lower case are equivalent.

MAP optional parameters

```
Optional Parameters:
-RSF[=map.rsf]   names the RSF output file
-OPEn[=20]       translates only in open reading frames [minimum ORF length]
-CIRcular        treats the sequence as circular
-LINear          treats the sequence as linear (default)
-PAGe[=62]       adds form-feeds to keep clusters on a single page
-WIDth=100       sets display width to something other than 60 bp/line
-THReeletter     uses three-letter amino acid codes to show translation
-MISmatch=1      finds restriction sites with one or fewer mismatches
-SILent          finds translationally silent potential restriction sites
-PERFect         finds only perfect symbol matches between site and sequence
-ALL             finds "overlapping-set" matches
-APPend          appends enzyme data file to your output
-CUTters[=fn]    writes enzyme data file with enzymes that did cut
-NONCUTters[=fn] writes enzyme data file with enzymes that did not cut
-EXCUTters[=fn]  writes enzyme data file with enzymes that were excluded
-MINSitelen=6    selects enzymes with 6 (or more) bases in recognition site
-OVErhang=0      selects only blunt-end cutters ("5" for 5', "3" for 3')
-MINCuts=2       shows only enzymes that cut at least 2 times
-MAXCuts=2       shows only enzymes that cut no more than 2 times
-ONCe            shows only enzymes that cut once
-EXCLude=n1,n2   doesn't show enzymes that cut between bases n1 and n2
-BOTtom          shows both forward and reverse strand cut points
-VERtical        displays enzyme names vertically over cut points
-NOCUTline       suppresses line of '|' characters showing cut points
-NOSEQline       suppresses the sequence display
-NOSCALeline     suppresses the scale line
-NOCOMPline      suppresses the complement sequence display
-TABle           shows the map as a list of cut positions, sorted by position
  -SORtbyenzyme    sorts table output first by enzyme, then by cut position
-NOMONitor       suppresses the screen monitor
-NOSUMmary       suppresses the screen summary
```

TRANSLATE

```
% translate

 TRANSLATE from what sequence ?  gamma.seq

                  Begin (* 1 *) ?  2179
                End (* 11375 *) ?  2270
               Reverse (* No *) ?

 Range begins ATGGG and ends GGAAG.  Is this correct (* Yes *) ?

 That is done, now would you like to:

   A) Add another exon from this sequence
   B) Add another exon from a new sequence

   C) Translate and then add more genes from this sequence
   D) Translate and then add more genes from a new sequence

   W) Translate assembly and write everything into a file

 Please choose one (* W *):  a

                  Begin (* 1 *) ?  2393
                End (* 11375 *) ?  2615
               Reverse (* No *) ?

 Range begins GCTCC and ends TCAAG.  Is this correct (* Yes *) ?

 That is done, now would you like to:

   A) Add another exon from this sequence
   B) Add another exon from a new sequence

   C) Translate and then add more genes from this sequence
   D) Translate and then add more genes from a new sequence

   W) Translate assembly and write everything into a file

 Please choose one (* W *):  a

                  Begin (* 1 *) ?  3502
                End (* 11375 *) ?  3630
               Reverse (* No *) ?

 Range begins CTCCT and ends ACTGA.  Is this correct (* Yes *) ?

 That is done, now would you like to:

   A) Add another exon from this sequence
   B) Add another exon from a new sequence

   C) Translate and then add more genes from this sequence
   D) Translate and then add more genes from a new sequence

   W) Translate assembly and write everything into a file

 Please choose one (* W *):

 What should I call the output file (* gamma.pep *) ?  ggamma.pep

%
```

TRANSLATE output

OUTPUT

[Previous | Top | Next]

Here is the output file ggamma.pep:

```
!!AA_SEQUENCE 1.0
TRANSLATE of: gamma.seq check: 6474 from: 2179 to: 2270
      and of: gamma.seq check: 6474 from: 2393 to: 2615
      and of: gamma.seq check: 6474 from: 3502 to: 3630
generated symbols 1 to: 148.

Human fetal beta globins G and A gamma
from Shen, Slightom and Smithies,  Cell 26; 191-203.
Analyzed by Smithies et al. Cell 26; 345-353.

ggamma.pep  Length: 148  October 5, 1998 12:59  Type: P  Check: 6924

       1  MGHFTEEDKA TITSLWGKVN VEDAGGETLG RLLVVYPWTQ RFFDSFGNLS

      51  SASAIMGNPK VKAHGKKVLT SLGDAIKHLD DLKGTFAQLS ELHCDKLHVD

     101  PENFKLLGNV LVTVLAIHFG KEFTPEVQAS WQKMVTGVAS ALSSRYH*
```

MAPSORT

MapSort finds the coordinates of the restriction enzyme cuts in a DNA sequence and sorts the fragments of the resulting digest by size. MapSort can sort the fragments from single or multiple enzyme digests.

EXAMPLE

[Previous | Top | Next]

Here is a session using MapSort to identify all of the restriction sites and restriction patterns for all the enzyme digests of pbr322:

```
% mapsort -CIRcular

 (Circular) MAPSORT of what sequence ?  GenBank:SympBR322

                  Begin (* 1 *) ?
                End (*  4361 *) ?

 Select the enzymes:  Type nothing or "*" to get all enzymes. Type "?"
 for help on which enzymes are available and how to select them.

                                        Enzyme(* * *):

 What should I call the output file (* synpbr322.mapsort *) ?

 Mapping ..........................

%
```

-DIGest	sorts the cuts for all the enzymes together in one digest
-LINear	treats chosen range as linear (default)
-PLAsmid	makes output suitable for display by PLASMIDMAP
-FRAGments	puts "blocks" not "ticks" into the PLASMIDMAP label file
-NOSIZe	suppresses the report of fragment sizes in your output
-MISmatch=1	finds potential sites with one or fewer mismatches.
-SILent	finds translationally silent potential restriction sites
-PERFect	looks only for perfect matches
-ALL	does an "overlapping-set" map
-APPend	appends enzyme file and translation table to your output
-CUTters[=fn]	writes enzyme data file with enzymes that did cut
-NONCUTters[=fn]	writes enzyme data file with enzymes that did not cut
-EXCUTters[=fn]	writes enzyme data file with enzymes that were excluded
-MINSitelen=6	selects enzymes with 6 (or more) bases in recognition site
-OVErhang=0	selects only blunt-end cutters ("5" for 5', "3" for 3')
-MINCuts=2	shows only enzymes that cut at least 2 times
-MAXCuts=2	shows only enzymes that cut no more than 2 times
-ONCe	shows enzymes that cut only once
-EXCLude=n1,n2	suppresses enzymes that cut between bases n1 and n2

What to do next?

The predictions by these programs is just that: ***a prediction*****.**



NEVER TRUST A COMPUTER!

Exercise 105-06-1

Programs used in this exercise:

(1) Sequence manipulation – reverse

(3) ORF Searching – frames, map, translate

(4) Mapping (restriction sites)

- map (-minc, -maxc), mapsort(-exclude, -digest), mapplot, plasmidmap

(5) Mapping (transcription factor)

- map(tfsites).

Sequences used in this exercise:

fetch z18853 (*C.elegans* mRNA for capping protein alpha subunit.) cds:10-858

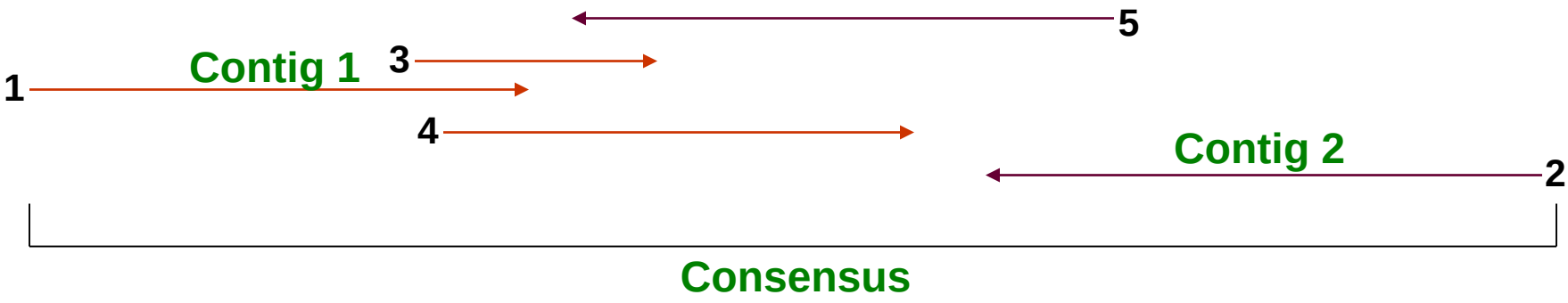
fetch x03795 (Human mRNA for platelet derived growth factor A-chain, PDGF-A) cds:388-1020.

BioEdit

Copyright © 1997-2013
Tom Hall
Ibis Biosciences
Carlsbad, CA 92008

Biological sequence alignment editor for Win95/98/NT/2K/XP/7

(1) Store fragment sequences;
(2) Recognize overlapping sequences and create aligned assemblies, called contigs;
(3) Display, edit and output the contigs for further analysis.



Assemble overlapping fragment sequences from a sequencing project.

http://www.mbio.ncsu.edu/bioedit/bioedit.html



Copyright © 1997-2013
Tom Hall
Ibis Biosciences
Carlsbad, CA 92008

Biological sequence alignment editor for Win95/98/NT/2K/XP/7

BioEdit is a biological sequence alignment editor written for Windows 95/98/NT/2000/XP/7. An intuitive multiple document interface with convenient features makes alignment and manipulation of sequences relatively easy on your desktop computer. Several sequence manipulation and analysis options and links to external analysis programs facilitate a working environment which allows you to view and manipulate sequences with simple point-and-click operations.

BioEdit.zip (Full install)
Bug fixes / changes
BioEdit General information
BioDoc.pdf (pdf format help doc)
View Screenshots

BioEdit's features include:

- Several modes of hand alignment
- Automated ClustalW alignment
- Automated Blast searches (local and WWW)
- Plasmid drawing and annotation
- Accessory application configuration
- Restriction mapping
- RNA comparative analysis tools
- Graphical matrix data viewing tools
- Shaded alignment figures
- Translation-based nucleic acid alignment
- ABI trace viewing, editing and printing
- Customizable ... other features

Note: BioEdit is no longer being regularly maintained, and the documentation is out of date and no longer maintained. BioEdit is free for use by any and all interested parties, but is supplied as is. Users must agree with the terms of the license.

Citeline.com

VALUABLE Citeline Internet Research Software RESOURCE

Enter disease, therapy or topic and click Search: Tips

Search

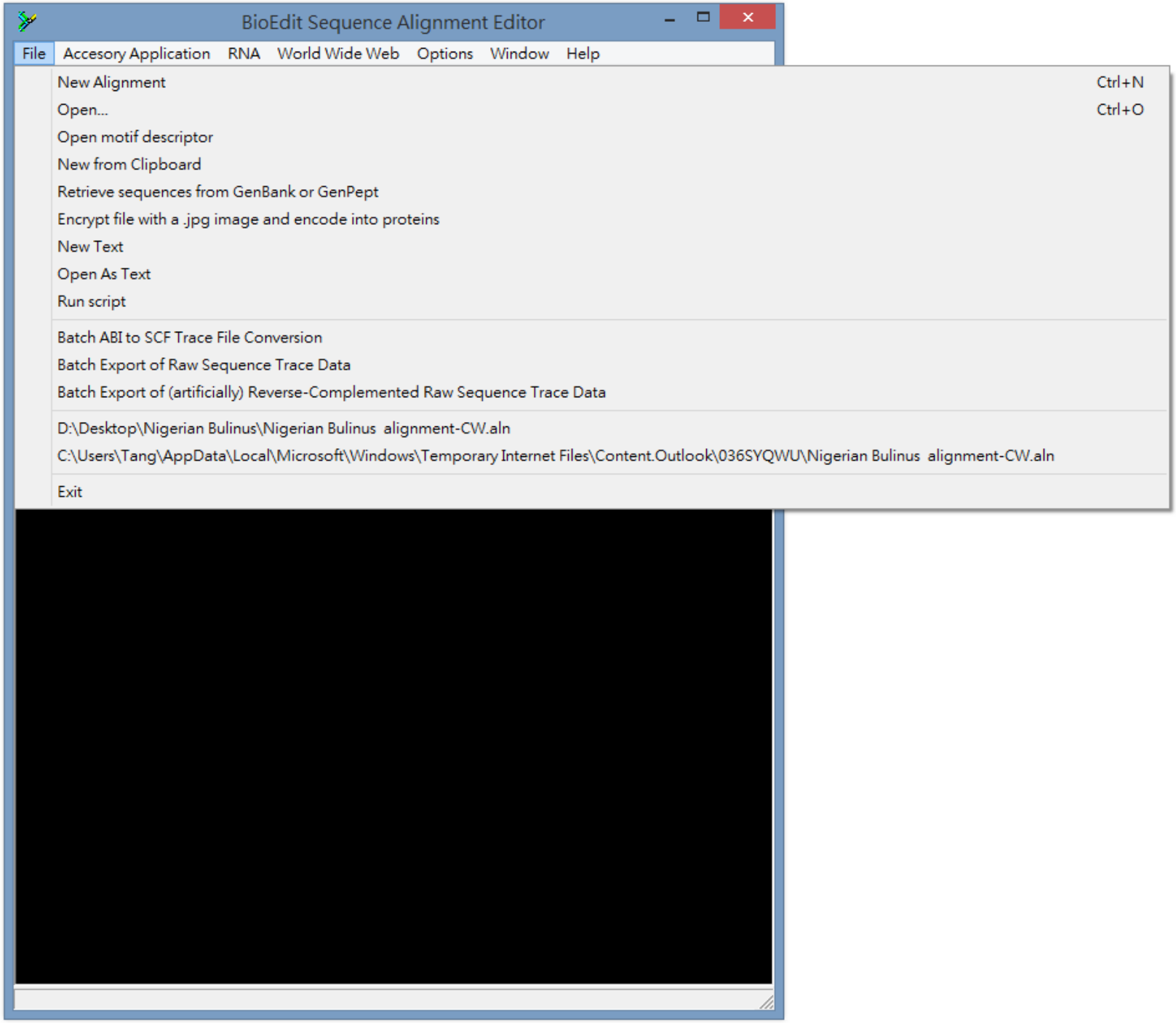
Copyright © 1999 Caredata.com, Inc.

Be updated on the software you use!
software.informer

BioEdit Software Informer

Current version 7.2.5
Last updated 12/11/2013
Tom Hall
Thomas.Hall@abbott.com

RNaseP databasee | ReadMe file | See screen shots | Accessory applications | Multi-file download | email Tom


BioEdit Sequence Alignment Editor
File
Accesory Application
RNA
World Wide Web
Options
Window
Help
New Alignment Ctrl+N
Open... Ctrl+O
Open motif descriptor
New from Clipboard
Retrieve sequences from GenBank or GenPept
Encrypt file with a .jpg image and encode into proteins
New Text
Open As Text
Run script
Batch ABI to SCF Trace File Conversion
Batch Export of Raw Sequence Trace Data
Batch Export of (artificially) Reverse-Complemented Raw Sequence Trace Data
D:\Desktop\Nigerian Bulinus\Nigerian Bulinus alignment-CW.aln
C:\Users\Tang\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.Outlook\036SYQWU\Nigerian Bulinus alignment-CW.aln
Exit

ASS02

Download EX105_6Seq.exe
Decompress the file
-> Seq01.txt - seq10.txt

Use BioEdit to assembly the sequences.
Q1: What is the correct order of the assembled sequence?
Transfer the assembled sequence to GCG
Q2: Identity putative protein coding region
Q3: Is there any potential regulatory elements upstream of the gene?
Q4: Export the amino acid sequence
Q5: What is the identity with the human homologues protein?

E-mail the ANSWER as attached files to
--petang@mail.cgu.edu.tw. before 12:00 26 Oct 2017
****郵件主旨： ASS02 bioinfo – (學號)