Sequence Analysis Spring 2000 Dr. Richard Friedman
(212)305-6901 (76901) friedman@cuccfa.ccc.columbia.edu 130BB

Lesson 10 Mapping and Primer Design

In this class we are going to learn how to generate restriction maps of DNA and design primers.

Reading:

1. GCG tutorial Chapter 8.
2. GCG Program manual, Prime.
3. http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen_docs.html

Summary of Commands:

Note: In this document different fonts have different meanings:

New York is used to explain commands.

`Courier is used to indicate commands and command options.`

Courier italics are used to indicate command parameters, for example, filenames.

Courier bold is used to indicate commands that are not displayed.

Courier bold italics are used to indicate computer- generated output.

Helvetica is used to indicate menu items and web sites.

tk	Sets GCG so that graphics are output to the screen.
lw	Sets GCG so that graphics are output to a postscript file.
lz	Sets GCG so that graphics are output to cuccfa's laser printer.
map	Generates a restriction map of a linear DNA sequence.

Selected options in map in response to the ***Enzyme(* * *):*** prompt:

<return> or *	Generates the cut sites for all enzymes in the database with one enzyme for each kind of cut site (i.e. not including isoschizomers).
enzymename	Generates the cut sites corresponding to that enzyme (for example "ecorii").

Command	Description
*partialenzymename**	Generates the cut sites for enzyme whose names begin with *partialenzymename*. enzyme (for example "ecor*").
**	Generates the cut sites for all enzymes in the database with more than one enzyme for each kind of cut site (i.e. including isoschizomers).
?	Lists options and all enzymes.
??	Lists options and all enzymes and cut sites.
'	Cut site top strand (in enzyme specification)
-	Cut site bottom strand (in enzyme specification).
<space-bar>	Gives map w/o cut sites.
`fetch` *enzyme.dat*	Copies Rebase, the restriction enzyme database, to your directory.
`mapplot`	Displays restriction sites of a linear DNA sequence graphically.
`mapsort`	Sorts the restriction fragments of a linear piece of DNA generated by one enzyme at a time by size.
`mapsort -digest`	Sorts the restriction fragments generated by a multiple enzyme size.
`mapsort -plasmid`	Sorts the restriction fragments generated by a single or multiple enzyme digest of a Plasmid or other circular DNA by size. Generates a *.tick file, which can be turned into a graphics file using `plasmidmap`.
`plasmidmap`	Plots a restriction map of a plasmid from a tick file.
`plasmidmap -mmth=5.33 -minth=3.2 -font=3`	Good settings for tick sizes and fonts for publication quality restriction maps of Plasmids.
http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html	Filters out contaminating vector sequences.
`prime`	Designs primers for a piece of DNA according to specifications in a file called *prime.init* .

Example of *prime.init* file:

```
Input for primer of pET11b sequence
courtesy of Clay Bracken of the Palmer lab

                                      ..
-in=tenascin.seq
-begin1=1
-end1=670
-cla=G
-minpri=18
-maxpri=22
-dna=50
-gcminpri=40.
-gcmaxpri=55.
-tmminpri=50
-tmmaxpri=60
-tmminpro=80
-tmmaxpro=95
-salt=50.
-minpro=1
-maxpro=670
-out=tenascin.prime
```

Meaning of the parameters in the example *prime.init* file:

`-in=`*tenascin.seq*	The sequence for which the primers are designed is in *tenascin.seq* .
`-begin1=1`	Begin the sequence range for which primers are searched at nucleotide #1.
`-end1=670`	End the sequence range for which primers are searched at nucleotide #670.
`-cla=G`	A G at the end of the primer. You can specify the nucleotide according to ambiguity codes discussed in Lesson 7, in more detail in the ambiguity codes in Appendix III of the GCG program manual).
`-minpri=18`	The primer should be a minimum of 18 bases in length. (18 is the minimum number of bases necessary to specify a unique sequence in human DNA if the sequences are distributed randomly.)
`-maxpri=22`	The primer should be a maximum of 22 bases in length.
`-dna=50`	The DNA concentration should be 50 nanomolar.
`-gcminpri=40.`	The primer should have a minimum of 40% GC content.

`-gcmaxpri=55.`	The primer should have a minimum of 55% GC content.
`-tmminpri=50`	The minimum melting temperature of the primer should be 50C.
`-tmmaxpri=60`	The maximum melting temperature of the primer should be 60C
`-tmminpro=80`	The minimum melting temperature of the product should be 80C.
`-tmmaxpro=95`	The maximum melting temperature of the product should be 95C.
`-salt=80.`	The salt concentration should be 80 millimolar.
`-minpro=600`	The minimum PCR product length should be 600 bases.
`-maxpro=670`	The maximum PCR product length should be 670 bases.
`-out=`*`tenascin.prime`*	The ouput file should be called *`tenascin.prime`* .
`-for`	The forward strand only should be searched (not used in this example).
`-rev`	The reverse strand only should be searched (not used in this example).
`findpatterns -nomon`	Searches library for exact agreement with one or more sequence patterns but supresses screen output of sequences searched. Can be used to see if the primer occurs in a known but undesired sequence in a library. Should be done both with the primer and its reverse complement. Search the computational library corrresponding to the actual DNA library that you are searching

DNA Melting Theory

$$K = \frac{[\sin gle - strand]^2}{[double - strand]}$$

K= Equilibrium constant in 1 molar salt.

$G^0 = -RT \ln K$

$G^0 =$ standard free energy of melting in 1 molar salt.

$G_{37}{}^0 =$ standard free energy of melting in 1 molar salt at 37C.

$$G^0 = {}_{ends}\; G_{init}{}^0 + {}_{steps}\; G_{step}{}^0$$

$G_{init}{}^{0}$ = Free energy to form a base-pair at the end of the double helix.

$G_{step}{}^{0}$ = Free energy to add a base-pair in the middle of the double helix.

for example, for the oligomer:

5' C-G-T-T-G-A 3'
3' G-C-A-A-C-T 5'

$$G_{37}{}^{0} = \quad G_{37}{}^{0}(CG/GC) + \quad G_{37}{}^{0}(GT/CA) + \quad G_{37}{}^{0}(TT/AA) + \quad G_{37}{}^{0}(GA/CT) + \quad G_{37}{}^{0}(C-G) + \quad G_{37}{}^{0}(A-T)$$

T_M = Melting temperature

At equilibrium (at melting):

$$G^0 = 0$$

$$G^0 = \quad H^0 - T \quad S^0$$

$$S^0 = \frac{H^0}{T_M}$$

$$S^0 = \quad S - a\ln[K^+]$$

Where a is a constant.

Lab or homework:

1. Mapping

   A. Work through GCG Tutorial Chapter 8.

2. Primer Design

   A. Find primers for the above example. Verify that some of them are not already part of a known sequence.

3. Vector filtration,
   Filter the sequence vec2.seq (Courtesy of Tatiana Kisselva of Chris Schindler's lab) of the vector sequence.

Seqlab Users:

Repeat exercises using `prime`, and `findpatterns` using `Seqlab.`

Bibliography:

1. References cited in the program documentation.
2. Biochemistry, L. Stryer, 5th Ed. , W.H. Freeman, 1995.

Web sites:

Vecscreen:
http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html

Mapping:
http://www.firstmarket.com/cutter/cut2.html

Answer key lab 9

(Where not noted, results were as demonstrated in lecture, or are obtainable with relative ease.)

2C. (I used sw:*human as my search set).The highest scoring false positive that I found was

SW:RNBP_HUMAN

In the *pfs file it occured:

SW:STA5_HUMAN + 5.93 34.67 794 ! P42229 homo sapiens (human). signal transducer and activator of transcription 5a. 7
SW:RNBP_HUMAN + 5.47 30.24 417 ! P51606 homo sapiens (human). renin-binding protein (rnbp). 10/1996

I would therefore make z=5.92 the cutoff. Note that there are many SH2 containing proteins with scores less than 5.47.

3F. Comparison of searches. I got:

Profilesearch: 48
Ssearch3: 41
PSIBLAST: 52.