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

Lesson 7 Database Searching by Keyword and Motif

Assignment

1. (Before class) Read GCG tutorial chapters 6 and 10.
2. (After class) Read the section in the GCG Program manual, "Defining patterns", which occurs in the Motifs and Findpatterns chapters.
3. http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html
   http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html#Short

Summary of Commands:

Note: In this document different fonts have different meanings:

Times 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.

`stringsearch`	Searches for text strings in annotation of sequence libraries.
`,`	Stringsearch AND operator.
`stringsearch -match=or`	Does Stringsearch search with "," standing for OR.
`lookup`	Also searches for text strings in annotation of sequence libraries, but is faster and somewhat more flexible, albeit somewhat less sensitive, than Stringsearch.
`@filename`	Specifies GCG list file.
Boolean operators used with Lookup:	
`&`	AND
`Chromosome & 13`	*Exempli gratia:* Chromosome 13.
`|`	OR
`13 | 13q`	F'rinstance: "13" or "13q".

To perform an Entrez search:
Click on http://www.ncbi.nlm.nih.gov/
Click on Entrez

Boolean operators used with Entrez:

AND	AND
Chromosome AND 13	*Exempli gratia:* Chromosome 13.
OR	OR
13 OR 13q	F'rinstance: "13" or "13q".
http://srs6.ebi.ac.uk/	Performs an SRS search.
motifs	Compares sequence to Prosite library of protein structure motifs. Excludes frequently occuring motifs.
motifs -freq	Compares sequence to Prosite library of protein structure motifs. Includes frequently occuring motifs.
findpatterns	Searches library for exact agreement with one or more sequence patterns.
findpatterns -nomon	Searches library for exact agreement with one or more sequence patterns but supresses screen output of sequences searched.
findpatterns -mism=1	Searches library for agreement with one or more sequence patterns, allowing one mismatch.

Pattern codes (GCG convention):

X	Any amino acid.
N	Any nucleotide.
R	Purines (A,G).
Y	Pyrimidines (T,C).
(symbol(s)){#}	Repeats symbols (#) times.
	Exempli gratia:
TAAC{5}GT	TAACCCCCGT
TAA(AT){3}GT	TAAATATATGT
(symbol(s)){#1,#2}	Repeats symbol(s) between #1 and #2 times.
	Exempli gratia:
TAA(AT){2,6}GT	"TAA" then "AT" repeated between 2 and 6 times, then "GT".
(symbol(s)){#, }	Repeats symbol(s) for at least # times.
	Exempli gratia:

`CATG{2, }A`	"CAT" then "G" repeated at least two times, then "A".
`(symbol(s)){, # }`	Repeats symbol(s) for between 0 and # times. F'rinstance:
`ACT{,10 }A`	"AC" then "T" repeated between 0 and 10 times, then "A".
`(symbol1, symbol2...)`	Either symbol1, symbol2, etc.. *Exempli gratia:*
`F(Q,A)S`	"F", then "Q" or "A", followed by "S".
`~(symbol(s))`	Any symbol but the one following. *Exempli gratia:*
`GC~ACG`	"GC", any symbol other than "A", then "CG".
`GC~(A,T)CG`	"GC", any symbol other than "A" or "T", CG.
`<symbols`	Symbols occuring at the beginning of a sequence only. *Exempli gratia:*
`<ATCG`	ATCG occuring at the beginning of a sequence only.
`symbols>`	Symbols occuring at the end of a sequence only. *Exempli gratia:*
`ATCG>`	ATCG occuring at the end of a sequence only.

PHIBLAST

The user specifies a protein query sequence and a protein motif. PHIBLAST (Pattern Hit Initiated BLAST) searches only those members of a protein database that contains the specified motif. This technique reduces the number of sequences being searched and hence raises the sensitivity of the search. To run PHIBLAST go to:

http://www.ncbi.nlm.nih.gov/BLAST/
PHI-BLAST search

and click on:

Pattern codes in the NCBI convention are the same as above except as illustrated by the following examples

`F[QA]S`	"F", then "Q" or "A", followed by "S".

Also in NCBI format, repeating symbols must be written out explicitly:, i.e. :

TAACCCCCGT

Web sites:

http://expasy.hcuge.ch/sprot/prosite.html
Prosite

http://www.ncbi.nlm.nih.gov/Entrez/
Entrez

http://srs6.ebi.ac.uk/
SRS

http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast?Jform=1
Phiblast

Lab or homework:

1. Work through the GCG tutorial chapter 6.1, using Stringsearch, Lookup, and Entrez. Suggestion: since Stringsearch takes a long time to run, Stringsearch in one session, and simulataneously, lookup in another.

2. (optional) Obtain the sequence file of aardvark myoglobin from Swissprot. Do your searching with:

A. Lookup.
B. Stringsearch.
C. Entrez.

3. Do the exercises in the GCG tutorial chapter 10.

4. Find the rare motifs in Chicken SRC protein. Find the common motifs in the same protein. Convince yourself that the uncommon motif found is indeed matched by the protein. If you have trouble doing this on the screen, print the output file out (either from cuccfa or from your own Mac or PC) and examine the match on paper. As there may be delays in getting a printed copy to work with in the lab, the pattern-match is reproduced below:

```
Protein_Kinase_Atp
(L,I,V)G~(P)G~(P)(F,Y,W,M,G,S,T,N,H)(S,G,A)~(P,W)(L,I,V,C,A,T)~(P,D)x(G,S,T,A,C,L,I,V,M,F,
Y)x{5,18}(L,I,V,M,F,Y,W,C,S,T,A,R)(A,I,V,P)(L,I,V,M,F,A,G,C,K,R)K
                        (L)G~PG~P(F)(G)~(P,W)(V)~(P,D)x(G)x{7}(V)(A)(I)K
           272: RLEVK   LGQGCFGEVWMGTWNGTTRVAIK   TLKPG
```

5. (optional) Do the exercises in the GCG tutorial chapter 6.2.

6. Translate the following into motif notation: “A peptide begins with an alanine, a glycine, and an arginine, followed by between 20 and 30 amino acids of any kind, followed by a glutamine and arginine pair repeated between 15 and 92 times and terminating in a tryptophan”.

7. Obtain the file ced4.tfa from /usr2/seq/seqclass/lab7 . Perform a Blastp and a Phiblast search with this sequence and the pattern (G,A)x4GK(S,T). What sequence has a marginal hit with Phiblast but not with Blastp?

Bibliography

1. GCG manual on Stringsearch, Lookup, Motifs, and Findpatterns.