

## Bio 595

### Computers in Biomedical Research Fall 2003 Slides for Monday, December 8

## Regular expressions

Example, looking for a six base motif, C T followed by a C, G or T but never an A, followed by A C G:

The regular expression looks like this:

```
C T [ C G T ] A C G
```

Test for the motif by pattern- matching:

```
if ( $dna =~ /CT[CGT]ACG/ ) {  
    print "I found the motif!!\n";  
}
```

Fundamental ideas about regular expressions:

Repetition: (asterisk \*) indicates 0 or more repetitions of the character just preceding it

Alternation: (a | b) "a or b" matches string a or string b

Concatenation: ab means character a followed by character b

Regular expressions usually enclosed within forward slashes

## Restriction enzymes and maps

Restriction enzymes: proteins that cut DNA at specific, short sequences..

Example: EcoRI cuts at GAATTC, between G and A, cutting both complementary strands, leaving an overhang at each end.

Sticky end makes possible insertion of DNA for cloning and sequencing.

HindIII cuts at AAGCTT, cutting between As.

Some restriction enzymes cut evenly with no overhang: "blunt ends".

There are at least 1,000 restriction enzymes.

Many restriction enzymes are "palindromes": the reverse complement is the same sequence (e.g. in EcoRI)

## Restriction enzyme recognition sites

Write a program to look for restriction enzyme sites in DNA and report back with a restriction map showing where in the DNA the enzymes would cut.

Restriction enzyme data located at REBASE (Restriction Enzyme Database) at [www.nsb.com/rebase/rebase.html](http://www.nsb.com/rebase/rebase.html)

In database, sites are represented in a specialized format.

Must be translated into regular expressions.

Generate a list of positions where sites are found.

Example of REBASE file data: (cut sites represented by caret)

Aal (XmaII)	C^GGCCG
AacI (BamHI)	GGATCC
AaeI (BamHI)	GGATCC
AagI (ClaI)	AT^CGAT
AarI	CACCTGCNNNN^

These data can be stored in a hash table; discard parentheses.

## Restriction maps

See examples ex29, ex30, ex31:

ex29: convert IUB ambiguity codes into regular expressions

ex30: parse a REBASE datafile

ex31: make a restriction map (see listing)

## GenBank

Genetic Sequence Data Bank: we need to know to extract DNA sequence, annotations, FEATURES table (containing information like location of regulatory regions, protein translation, exons, introns, important mutations), accession numbers, gene names, phylogenetic classification, literature citations.

Compare: databank and database; a database usually must be structured, queried; a databank like GenBank is a 'flat' file; an ASCII text file.

BLAST searches to look for sequences related to one you specify ("query" sequence).

Other search software available at NCBI, EBI, EMBL.

e.g. see [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

Databank repositories: NCBI, EMBL, DDBJ (DNA databank of Japan)

## GenBank

As of August 2001, 243 files of more than 200 MB each,  
13 million loci, about 14 billion bases in sequence from  
13 million reported sequences

Number of sequences in GenBank doubles  
approximately every 14 months.

Files in libraries are distributed in compressed format.  
Uncompressed files are > 50 GB.

Format specification of GenBank files is given in  
GenBank release notes, gbrel.txt at GenBank website,  
ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt

Programs to parse GenBank files: either put data into an  
array, or all of it into a scalar variable, using regular  
expressions to parse the data

## GenBank

GenBank libraries are organized into divisions based on phylogenetics  
or sequence details:

PRI primate sequences  
ROD rodent sequences  
MAM other mammalian sequences  
VRT other vertebrate sequences  
INV invertebrate sequences  
PLN plant, fungal, algal sequences  
BCT bacterial sequences  
VRL viral sequences  
PHG bacteriophage sequences  
SYN synthetic and chimeric sequences  
UNA unannotated sequences  
EST EST (expressed sequence tags) sequences  
PAT patent sequences  
GSS genome survey sequences  
HTG high throughput genomic sequencing data  
HTC high throughput cDNA sequence data

## sample GenBank data file annotation

```
LOCUS      AB031069      2487 bp      mRNA           PRI           27-MAY-2000
DEFINITION Homo sapiens PCCX1 mRNA for protein containing CXXC domain 1,
            complete cds.
ACCESSION  AB031069
VERSION    AB031069.1  GI:8100074
KEYWORDS   .
SOURCE     Homo sapiens embryo male lung fibroblast cell_line:HuS-L12 cDNA
            to mRNA.
ORGANISM   Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE  1 (sites)
AUTHORS    Fujino, T., Hasegawa, M., Shibata, S., Kishimoto, T., Imai, S. and
            Takano, T.
TITLE      PCCX1, a novel DNA-binding protein with PHD finger and CXXC
            domain,
            is regulated by proteolysis
JOURNAL    Biochem. Biophys. Res. Commun. 271 (2), 305-310 (2000)
MEDLINE    20261256
```

## GenBank annotation, continued

```
REFERENCE  2 (bases 1 to 2487)
AUTHORS    Fujino, T., Hasegawa, M., Shibata, S., Kishimoto, T., Imai, S. and
            Takano, T.
TITLE      Direct Submission
JOURNAL    Submitted (15-AUG-1999) to the DDBJ/EMBL/GenBank databases.
            Tadahiro Fujino, Keio University School of Medicine, Department of
            Microbiology; Shinanomachi 35, Shinjuku-ku, Tokyo 160-8582, Japan
            (E-mail: fujino@microb.med.keio.ac.jp,
            Tel:3-3353-1211(ex.62692), Fax:3-5360-1508)
FEATURES   Location/Qualifiers
            source          1..2487
                        /organism="Homo sapiens"
                        /db_xref="taxon:9606"
                        /sex="male"
                        /cell_line="HuS-L12"
                        /cell_type="Lung fibroblast"
                        /dev_stage="embryo"
            gene            229..2199
                        /gene="PCCX1"
            CDS             229..2199
                        /gene="PCCX1"
                        /note="a nuclear protein carrying a PHD finger and a
                        CXXC domain"
```

## GenBank annotation, continued

```
/codon_start=1
/product="protein containing CXXC domain 1"
/protein_id="BAA96307.1"
/db_xref="GI:8100075"
/translation="MEGDGSDPEPPDAGEDSKSENGENAPIYICIRKPDINCFMIGCD
NCNEWPHGDICIRITERMAKAIREWYCREKREDFKLEIRVRHKKSRERDGNRDSSEP
RDEGGRRKRVDPDILQRRAGSGTGVGAMLARGSAPHKSSPQLVATPSPQHQQQQQ
QIKRSARMCGECEACRRTEDCDCHGDFCDMKKFGGPNKIRKCRRLRQCQLRARESKY
FPSSLSPVTPSELPRRRLPTQQQPQPSQKLGRIREDGAVASVTYKPEEATATP
EPLSDEDLPLDPLDQFCAGAFDDHGLPMSDTEESPFLLPALRKRKRVKVKHKKRE
KXSEKKKEERYKRHRQKQKHDKWHPERADAKDFASLPQLGPGCVRPAQPSKYS
DCCGMKLANRIYEILPQRIQQWQSPCIAEHEGKLLERIRREQQSARTRLQEMERR
FHELEAILLRQQQAVRDEESNEGSDTDLQIFCVSCGHPINPRVALRHMECYAK
YESQTSFGSMYPTRIEGATRLFCVDVNPQSKTYCKRLQVLCPEHSRDPKPVADVCGC
PLVRDFELGDFCRIPRQCNRHYCWEKLRRAEVDLRRVWVKLDELFEQERNVRT
AMTNRGLLALLMHQTIQHDFLTDLRSSADR"
```

## GenBank data

```
BASE COUNT      564 a      715 c      768 g      440 t
ORIGIN
1 agatgggggc gctgaggggt cttgggggct ctaggccggc cactactggt ttgcaaggc
61 agacagcgca tggggcctgc gcaataggag tacgtgctct gggaggctg actagaagcg
121 gaagtatgtg tggggcctct tgaacccgct tgggaaggcg ccagatgttc tgtgaaggt
181 cgcgggtcgc tggcgggggt cgtgaggag tgcgcggcga gcgagatg agaggagat
241 ggttaagacc cagagcctcc agatggcggg gaggacagca agtcagaga tggggagat
301 ggcgccatct actgcatctg ccgcaaacgg gacataactg gottatgat cgggtgtgac
361 aactgcaatg agtggttcca tggggactgc atccgagaca ctgagagat ggcgaagcc
421 atcggggagt ggtactgtcg ggggtgcaga gagaagacc ccaagctaga gattcgtat
481 cggcaacaag agtcaacggg ggggatgac aatgagcggg acagcagtga gccocgggat
541 gaggtggag ggcgcaagag gccgtccct gatccagacc tgaagcggc ggcagggtca
601 gggacaaggg ttggggccat gctgtctcgg gctctgctct gcgcccaaa atctcttcg
661 cagcccttgg tggcccaacc cagcagcat caccagcagc agcaagagca gatcaaaagg
721 tcaagccgca tgtgtgtgta gttgaggca tgtggcgca ctgagactg tgtcaactg
781 gattctgtc ggaacatgaa gaagtctggg gcccccaaca agatccggca gaagtccgg
841 ctggccactg gccagctgag ggcocgggaa tgmtaactg atttccctc ctgctgtca
901 ccagtgagcg cctccagatc cctgcacagg ccccgccggc cactgcacc ccaacagcag
961 ccacagccat caacaagat aggggcatc cgtgaagtg aggggcaat ggcgtcaata
1021 acagtcaagg agcctcctga ggtcaagcc caactcagc cactctcaga tgaggacct
1081 cctctggatc ctgacctgta tcaagactc tgtcagggg ctttbtgta ccatgacctg
1141 ccttgatgta gcgacacaga agatgcccca tctctgacc ccgctctcg gaagaggcca
1201 gtgaagtga agcactgtaa ggtctgggag aagaagctg agaagaaga gggaggcca
1261 tacaagcggc atcggcaaga gcaagaacc aagataaat gaaacacc agagagcct
1321 gatgcaagg accctcctgc actgcccag tgcctggggc ccgctctgt ggcocccgc
1381 cagcccagct ccaagtattg ctcaatgac tgtggatga agctggcagc caaccgcatc
etc.....
```

