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

Lesson 11 Genomic Analysis

In this class we learn how to identify repeat sequences, exons, promoters, and transcription factor binding sites in genomic DNA.

Reading:

1. http://ftp.genome.washington.edu/RM/webhelp.html .
2. */usr2/seq/doc/genscan.doc* or http://CCR-081.mit.edu/GENSCAN.html .

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.

http://ftp.genome.washington.edu/cgi-bin/RepeatMasker

Filters out common, non-coding repeat sequences from genomic DNA sequence. Use before identifying genes.
On web-site, only check:

Skip simple, low complexity region masking

`genscan`

Identifies exons in genomic DNA and translates the sequence.
Typing `genscan` alone will give you the genscan command syntax.

genscan /usr2/seq/genscan/HumanIso.smat humrash.rep.tfa -v > humrash.genscan

Identifies the exons in *humrash.rep.tfa*. .

Genscan parameter files (all of which are in */usr2/seq/genscan/)*:

HumanIso.smat	Parameters derived from human sequences, which have been found to work for vertebrates in general and also drosophila.
Maize.smat	Parameters derived from maize sequences, and will probably work well for other monocots (rice, etc).
Arabidopsis.smat	Parameters derived from Aradopsis. Should be used for plants other than Monocots. Also the best of the three parameter sets for C. elegans.
http://genomic.sanger.ac.uk/gf/gf.html	Fgene family of gene identification programs.Is your organism closer to humans or drosophila, on the one hand, or to yeast, C. elegans, or plants on the other? In the former case use Fgenesh. In the latter case use Fgenes.
http://genome.cs.mtu.edu/aat/aat.html	Finds matches between genomic DNA and cDNA and protein databases. The resulting alignments display gene structure.
bestfit -pena=12	Aligns genomic and cDNA sequences scoring all gaps longer than 12 basepairs as if they were 12 basepairs long. This is useful for bridging long introns., Limited to sequences 32Kb long,
http://genome.cs.mtu.edu/align/align.html	Aligns Genomic DNA with cDNA (GAP2) or protein (NAP) taking splice-sites into account. I have a local version of these programs, but they are relatvelydifficult to run.

/usr2/seq/transfac/site.dat	File containing Transfac transcription factor binding sites.
/usr2/seq/transfac/factor.dat	File containing Transfac transcription factors.
/usr2/seq/transfac/matrix.dat	File containing Transfac matrices.
/usr2/seq/doc/tfdoc34.doc	Transfac documentation.
matinspector	Searches a query sequence for profile matches to matrices in the transfac library.
http://genomatix.gsf.de/cgi-bin/matinspector/matinspector.pl	Web version of matinspector.
http://genomatix.gsf.de/cgi-bin/matinspector_prof/mat_fam.pl	Web version of Matinspector Pro, a far more accurate version of Matinspector. Unfortunately this site is useable for free only until May 15, 1999. If you want to use it after that, and are able to contribute towards an online account, please see me.
findpatterns -dat=*/usr2/seq/transfacgcg/tfsites.vert*	Searches your sequence for the patterns in the Transfac database for vertebrates. Other options include: */usr2/seq/transfacgcg/tfsites.insect* */usr2/seq/transfacgcg/tfsites.plant* */usr2/seq/transfacgcg/tfsites.fungi* */usr2/seq/transfacgcg/tfsites.other* */usr2/seq/transfacgcg/tfsites.vert*
http://www.fruitfly.org/seq_tools/promoter.html	Identifies promoters.

Lab or homework:

A. Obtain a copies of humrash.n.tfa and rash_human.swissprot from /usr2/seq/seqclass/lab11.

B. Filter it of low complexity regions and repeats.

C. Identify the exons in the sequence with both Genscan and Fgenesh. Which of the two programs predicts the exon structure of rash_human better?

D. Identify the promoter sites.

E. Compare the putative promoter sites and putative exons. Can they both be correct? Which program corresponds to the experimental results?

F. Identify the transcription factor binding sites.

Additional web sites:

The web version of Genscan is available at:

http://CCR-081.mit.edu/GENSCAN.html

A useful key website for transcription factor sites is:

http://www.isbi.net/

Seqlab Users:

Repeat exercises using `findpatterns` using `Seqlab.`

Bibliography:

1. References cited in the program documentation.
2. Biochemistry, L. Stryer, 5th Ed. , W.H. Freeman, 1995.
3. Eukaryotic Transcription Factors, D. Latchman, Academic Press, 1995.