Analysis of DNA sequences

Sequence contamination

vector sequence: UniVec
(www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html)

verifying a restriction map:

REBASE database (rebase.neb.com)
Webcutter: www.firstmarket.com/cutter/cut2.html
commercial site: http://tools.neb.com/NEBcutter2/index.php

Designing PCR primers

biotools.umassmed.edu

Analyzing DNA composition (G+C content)

bioweb.pasteur.fr/seqanal/interfaces/geece.html

Counting words in DNA sequences

2- or 3-letter words

www.genomatix.de/cgi-bin/tools/tools.pl

Counting long words in DNA sequences

regulatory sequence motifs (for n-letters, 2^{2n} different words)

bioweb.pasteur.fr (English version)

DNA sequence analysis/codon usage, composition/wordcount

(e-mail reply)

http://www.bioinformatics.org/sms2/dna_stats.html

Finding internal repeats

tandem repeats, inverted repeats

finding repeats is a tricky business

Dot-plot approach

Molecular Toolkit (http://arbl.cvmbs.colostate.edu/molkit)

click Dot Plots

click Make Plots

how to identify inverted repeats (reverse complement)

How to assess the significance of repeats

9 ATGC repeats in 3000-bp DNA sequence

the random probability of observing ATGC: 1/256

the expected number of ATGC in 3000-bp: 3000/256=11.7

Finding protein coding regions

ORF(open reading frame) in microbial DNA sequences
or eukaryotic mRNA sequences
Start codon (ATG)
Stop codon (TAA, TAG, TGA)

ORF finder: www.ncbi.nlm.nih.gov/gorf/gorf.html
a more sophisticated: GeneMark (opal.biology.gatech.edu/GeneMark/)

Translation: http://web.expasy.org/translate/

Codon usage: http://www.bioinformatics.org/sms2/codon_usage.html

Finding internal coding exons
MZEF at Cold Spring Harbor (argon.cshl.org/genefinder)
GenomeScan at MIT (genes.mit.edu/genomescan)

Using nucleotide sequence databases

Gene-centric databases
Genome-centric resources
- Prokaryotic
- Eukaryotic

NCBI: www.ncbi.nlm.nih.gov

Ensembl: www.ensembl.org

TIGR: www.tigr.org
http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi

Microbial genome: http://mbgd.genome.ad.jp/

Genome level analysis

Chromosome localization & gene organization

Human

Human & other organisms: Ensembl: http://asia.ensembl.org/index.html
Exon-intron structure
Alignments
Synteny: the physical co-localization of genetic loci on the same chromosome within an individual or species
Neighboring sequences

Bacteria: neighboring genes

Bacteria map: http://wishart.biology.ualberta.ca/BacMap/
Microbial genome database: http://mbgd.genome.ad.jp/
Microbial genome annotation: https://www.genoscope.cns.fr/agc/microscope/home/index.php

Assignments

Analyze your DNA sequence

Human cDNA & bacterial DNA: composition (G+C content)

Human ORF sequence & bacteria ORF sequence: codon usage

Human cDNA & mouse cDNA: Dot-plot approach and assessment

Bacteria neighboring genes: compare 2 species

Gene organization

Human gene: chromosome localization
exon-intron structure

Differences between human, gorilla, mouse

Synteny: the physical co-localization of genetic loci on the same chromosome within an individual or species

Human gene: 5'-upstream sequence (1000 bp)