DNA sequences

Not all DNA is coding for proteins

- Regulatory regions
- Introns
- Protein-coding region

One protein, many DNA entries

- the primary transcript
- the mature transcript
- the strict protein coding region
- numerous types of partial sequences (ESTs)

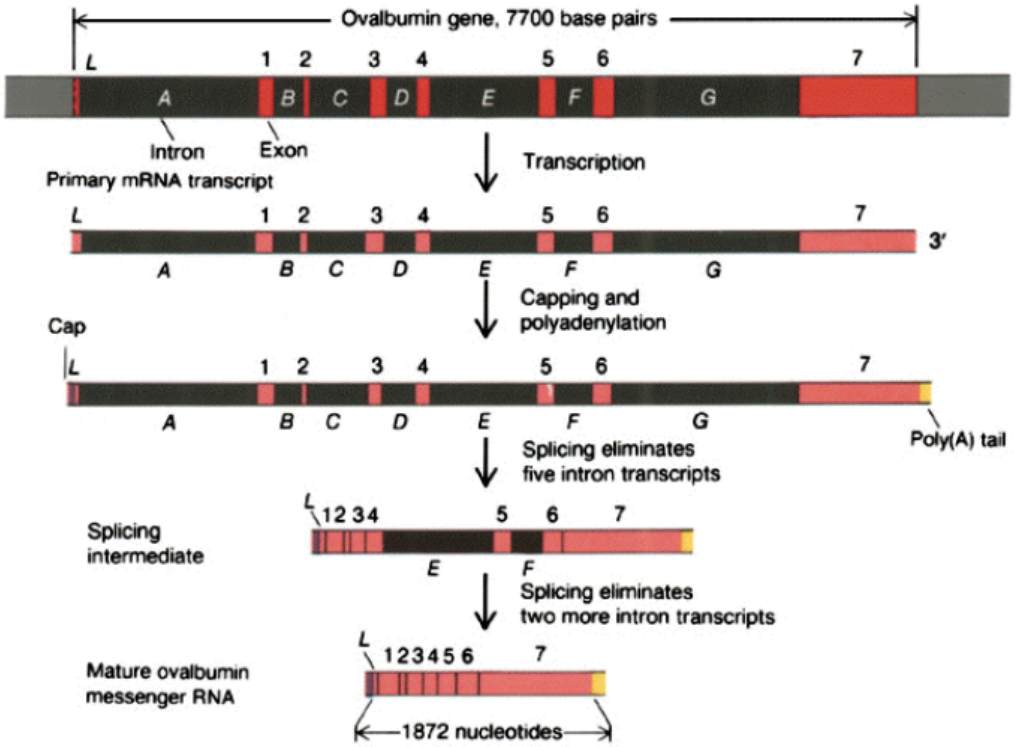
Retrieving the DNA sequences relevant to my protein

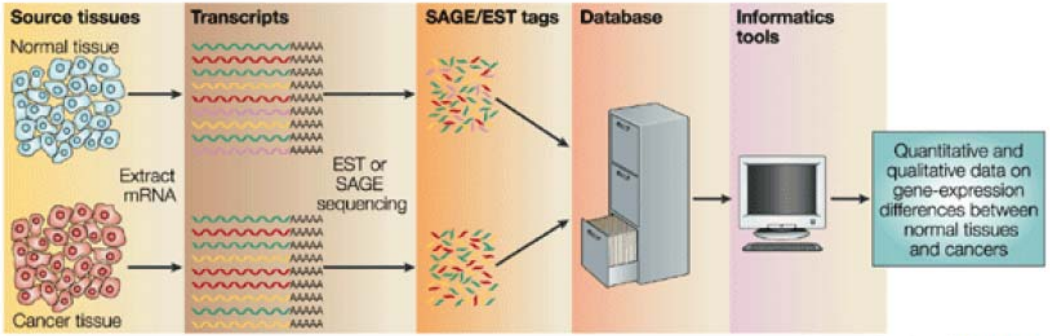
Go to www.expasy.org/sprot/ or http://www.ncbi.nlm.nih.gov/

Consists of 4 parts

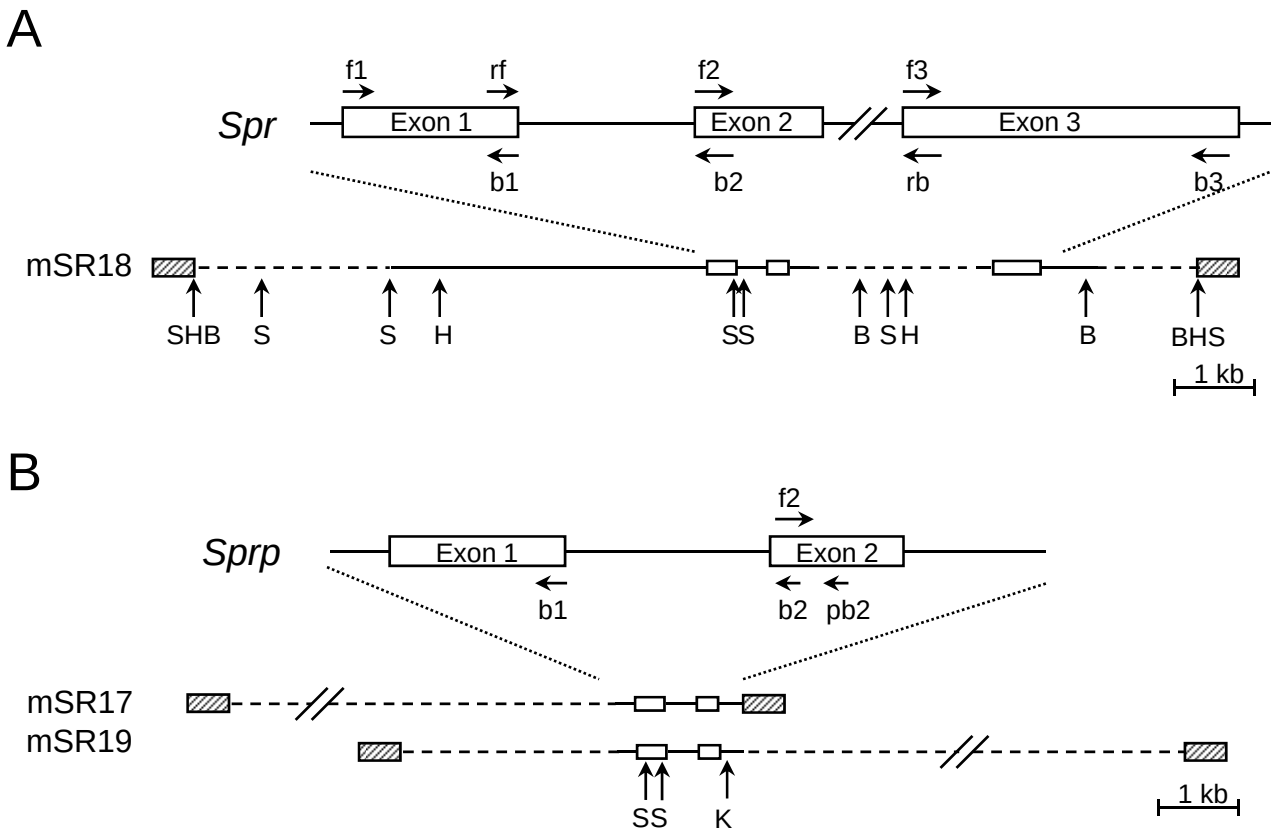
The locus name: write down the accession number
The reference section:
The features section
- promoter elements
- ribosome binding sites (RBS)
- protein coding segments (CDS)
- poly-A site

The sequence section
- actual nucleotide sequence submitted
- save FASTA format of the sequence as a text file





Nature Reviews | Genetics


A
f1
rf
f2
f3
Spr
Exon 1
Exon 2
Exon 3
b1
b2
rb
b3
mSR18
SHB
S
S
H
SS
B
S
H
B
BHS
1 kb
B
f2
Sprp
Exon 1
Exon 2
b1
b2
pb2
mSR17
mSR19
SS
K
1 kb

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/

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

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/

- searching MBGD:
- choose one
- homologous sequences: homolog list
- multiple genome map comparison: orthologous cluster

Assignments

1. Retrieve human and a bacterial DNAs and analyze the DNA sequences

Human cDNA sequence features

Human ORF sequence & bacteria ORF sequence: G+C ratio, dot-blot analysis, codon usage comparison

2. Gene organization

Human gene: exon-intron structure, chromosome localization

Differences between human and mouse

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

Retrieve 5’-upstream sequences (1000 bp) of human and mouse genes (ensemble/alignment-text)

3. Bacterial gene organization

compare 2 species