

BME 110L / BIOL 181L
Computational Biology Tools
www.soe.ucsc.edu/classes/bme110/Winter09

February 19:

- In-class exercise: a phylogenetic tree for that protein family...
 - Leaving the world of proteins again...
- A few words on Lab-work relevant tools we've encountered

NEXT WEEK:

- Review of midterm Qs (for part/most of the week)

This evening: Homework 3, PART I
(Parts I and II due on Thu March 4 BEFORE class)

Accompanying Reading (B4D): Chp 5

Most likely in the lab, you'll be working with DNA sequence...

- **trying to figure out what a DNA-sequence you determined is about (we've talked about this plenty)**
- **OR you'll want to engineer (mutate, express, target) a region of DNA that you find particularly interesting**

Some of the tools you encountered in HW2 aim to help you in these tasks - "lab tools" are likely to be among those you'll use the most...

We assume that you have a grounding in these very basic topics of molecular biology - a quick refresher on the following slides, + a few gene-finding programs. (See also Ch 5, B4D)

Basic/Old ORF Finding in Bacteria

1. Look for “long” open reading frames (ORFs)
2. Scan sequence at nucleotide #1 (Frame #1),
begin ORF at first start codon: ATG (too simple - why?)
3. Continue scanning to first stop codon:
TAA, TGA, or TAG
4. That is your ORF!
5. Repeat, starting at nuc#2 (Frame #2) , then again,
starting at nuc#3 (Frame #3)
6. Take reverse complement of sequence ->
(i.e. 5'-CGAAC -> 5'-GTTTCG)
7. Scan sequence starting at nuc#1 (Frame -1),
nuc#2 (Frame -2), nuc#3 (Frame -3)

On-line ORF Finding @ NCBI

<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>

- Shows all ORFs of min. length 50, 100, or 300 nucleotides long
- Also shows all start (green) & stop (red) codons
- Allows “alternate” start codons (TTG, GTG, CTG)
- When you’ve selected an ORF you like, click on “Accept” button, then can display nucleotide or protein sequence for further analyses

Historical Note: Annotating the Yeast Genome

- Yeast is a eukaryote, but does not have many introns
- A strict cutoff for ORF length was used: minimum of 300 nucleotides ORF required to be considered a protein-encoding gene in original genome annotation (1996)
- Since then, *many* smaller ORFs have been found experimentally

More Sophisticated Methods

- Can analyze entire genome at once, use codon frequencies, not just one gene
- GeneMark
(<http://opal.biology.gatech.edu/GeneMark>)
- Also, Glimmer from TIGR
(<http://cbcb.umd.edu/software/glimmer/>)
- Not available on-line, one must download these programs and run them locally

Both methods build on Hidden Markov Model (HMM)-methodology, in which a computer program is trained to distinguish coding from non-coding DNA sequences (humans don't have to know the rules)

Eukaryotic Gene Finders

- Introns do not necessarily preserve reading frame
(e.g. Exon#1 uses frame +2
Exon#2 uses frame +1
Exon#3 uses frame +2)

- No start codon at beginning of internal exons, so we must “guess” where exon/intron junction is

Note that there are sequence preferences at pre-mRNA splice sites (e.g. most introns begin with GU, end with AG) but as always it's only that, a preference/trend...

- Very difficult by hand, so computer programs are required

Eukaryotic Gene Finding On-line: GenomeScan

(builds on GenScan(HMMs; splice sites etc) and checks up on probability of BLASTX matches against a protein DB)

Gene Scanner Tracks on the Human Genome at UCSC

- AceView, Twinscan, N-SCAN, GeneID, GeneID, Genscan, Pseudogenes
- All use codon frequencies, splice site prediction, and other sources of information (Blastxhits, cDNA information, comparative genomics)
- As more experimental data is collected, less reliance on purely computational predictions

Cutting DNA:

Restriction enzymes

- enzymes isolated from prokaryotes that break DNA at very specific sequence-specific positions
- in nature, act as host-defense against viruses

examples

Nla III: 5' ... CATG[^] ... 3'

BamH I: 5' ... G[^]GATCC ... 3'

Dra III: 5' ... CACNN[^]GTG... 3'

> 3,000 RE's found to date with >200 specificities!

Many RE's Create "Sticky" Ends

Before cutting:

5' -ATTGATGG[^]**AATTC**TTATGGATAG-3' 3' -
3' -TAACTAC**CTTAA**[^]GAATACCTATC-5'

After cutting, "sticky ends":

5' -ATTGATGG **AATTC**TTATGGATAG-3'
3' -TAACTACC**TTAA** + GAATACCTATC-5'

- Useful to increase efficiency & specificity of re-joining ends

Sample Agarose Gel



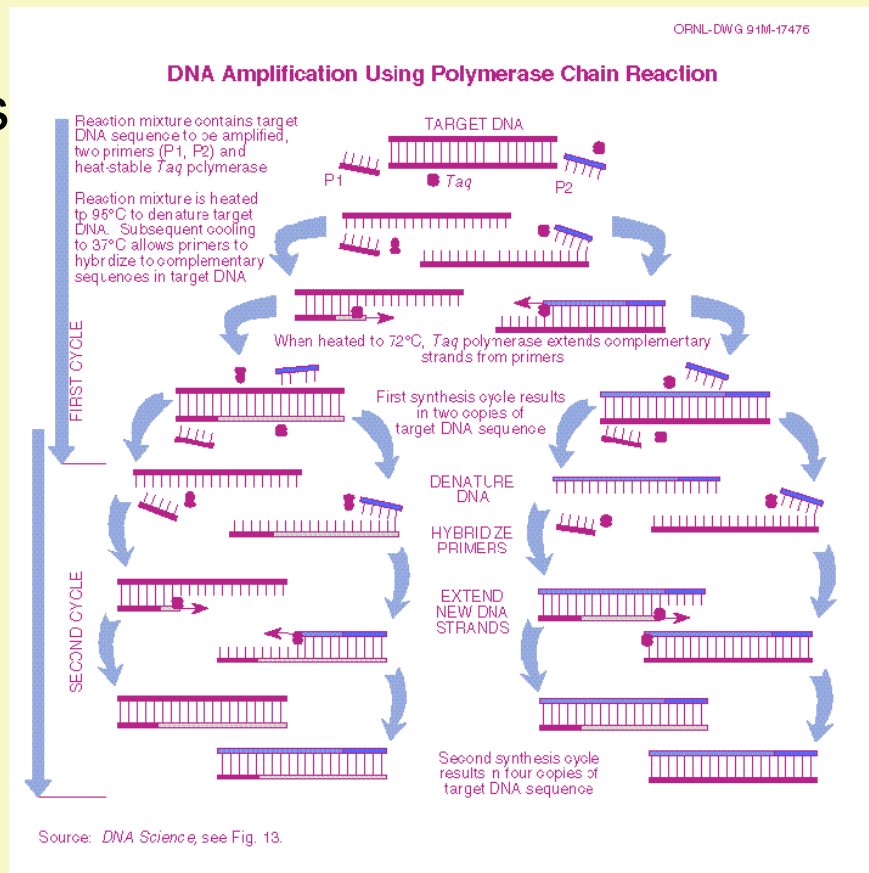
Same DNA cut by different restriction enzymes give different fragment lengths

New England BioLabsTools

<http://tools.neb.com/> (could be useful to try out)

NEBcutter—Display which restriction enzyme cut in your sequence, and where REBsites—Display a “virtual” digest of your DNA, showing how it would look on an agarose gel

Designing PCR Primers



Primer3 (as we used in HW2)

<http://frodo.wi.mit.edu/>

Key Input:

1. Sequence
 2. Targets (region to be amplified)
 3. Product Size Ranges
 4. Primer size (usually 18-22 bp)
 5. Primer T_m (annealing temperature)
- (Rest are usually OK as defaults)