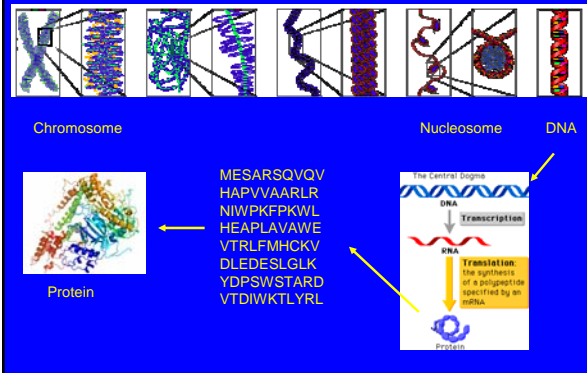


Making Sense of DNA and Protein Sequences

Making Sense of DNA and Protein Sequences



Objective

- Given a DNA sequence
 - Find potential genes
 - Find potential promoter sequences and transcription factor binding sites
- Given a protein sequence
 - Find conserved domains, patterns, or motifs

Predicting Genes

- Prokaryotes - simple, just find Open Reading Frames (ORFs) and conserved promoter sequences (-10 and -35 promoter elements)
- Eukaryotes - much more difficult due to complex promoters and splicing

Predicting Genes

- Problem: Given a new genomic DNA sequence, identify coding regions and their predicted RNA and protein sequences
- Steps:
 1. Search against protein/EST database
 2. Apply gene prediction programs
 3. Analyze regulatory regions

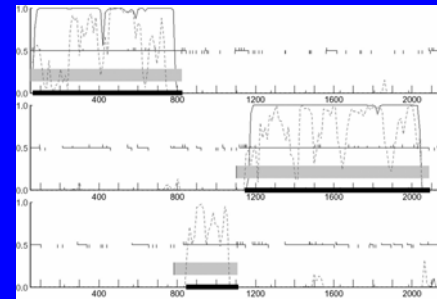
Important Lesson

- When predicting ANYTHING, the best approach is to use multiple methods and compare the results
- When different methods agree, we can be more confident of the prediction

Gene Prediction Programs

- There are lots out there
- We will use a simple ORF finder and GeneMark.hmm
- Others include Gene Sequer, Genscan, NetGene2, HMMGene, and many more

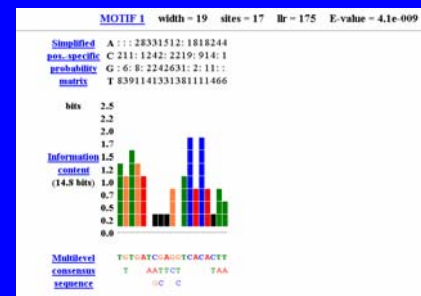
GeneMark Output



MEME

- Given a set of sequences, find conserved motifs

Reading the MEME Output



Reading the MEME Output

NAME	START	P-VALUE	SITES
ompa	51	3.07e-07	TTTCATATG CCTGACGGATTACACTT GTAAATTTTC
male	17	3.07e-07	CCGCCAATTC TGTAAACAGATCACACAA AGGGAAGGGG
deop2	10	4.90e-07	AGTGAATTA TTTGACCCAGATCCATTA CAGTGATGCA
ura	58	6.87e-07	ACATTGATTA TTTGCAGGCGTTCACACTT TGTATGCCA
lac	12	1.31e-06	ACGCAATTAA TTTGATTCCTCCTCAT TAGGCACGCG
tdc	79	2.03e-06	TGAAAGTTAA TTTGTGTTGTCCTCCTCAT ATCCCTTT
hgt1	79	2.89e-06	AGTTAATAG TTTGACGATGTCATATTT TTATCAAT
hna	74	3.18e-06	CCGCAAGAT TTTGATTCATTACACTT AAACAATTC
celg	64	3.18e-06	AGACTGTTT TTTGATCGTTTCACAAA ATGGAAGTCC
plr322	56	3.49e-06	CCATATGCGG TTTGAAATACCCACAGAT GOSTAAGGAG
crp	66	5.04e-06	ACTGCATGTA TGCAGAGACATTCACATTA CCGTGAGTA
gale	45	7.17e-06	ATTCCACTAA TTTATCCATTCACACTT TTGGATCCT
ucul	20	2.24e-05	GTGAAATGCT TTTGATTCCTTACCCCA TTGAAATTCG
mtl	44	3.26e-05	GATTTGGAAAT TTTGACAGTTCACACTT AGACACATAA
ey8	53	3.26e-05	ATCGCAAGG TTTTAAATGATCACCTT TAGCCATTT
mtk	64	5.00e-05	TAAAGGAATTT CTTGATTCCTTCACAAA ATCGGAGGCA
lv	42	5.36e-05	CAGTACAAA CTTGATCACCCCTCAAT TTCCCTTTC

TESS

- Searches your sequence for known transcription factor binding sites
- My favorite output is the Tabular Results section - just a list of the sites found

Protein Motifs and Domains

- Motif:
 - "Short" (~10 aa) conserved sequence pattern
 - Associated with a distinct function
- Domain:
 - "Longer" (~100) conserved sequence pattern
 - Defined as an independent functional and/or structural unit

Finding Protein Motifs and Domains

- We will use PROSITE and SuperFam
- Lots of other methods - PFP, Jafa, e-motif, PRINTS, BLOCKS, SMART, COGs, etc.

- <http://dobbslab.gdcb.iastate.edu/BCB590>