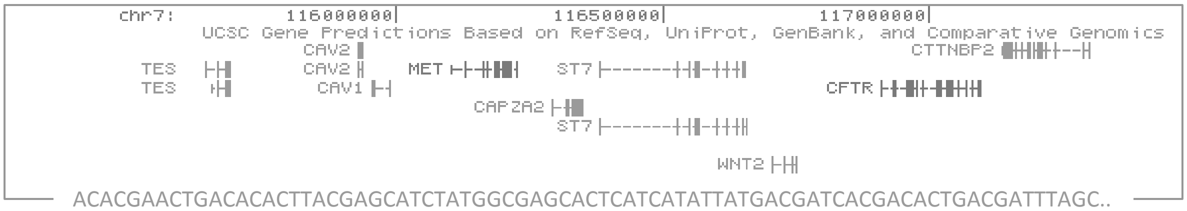
chr7:
116000000
116500000
117000000
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
CAV2
CTTNBP2
TES
CAV2
MET
ST7
TES
CAV1
CFTR
CAPZA2
ST7
WNT2
ACACGAACTGACACACTTACGAGCATCTATGGCGAGCACTCATCATATTATGACGATCACGACACTGACGATTTAGC..

CS 207 Scientific Databases and Knowledge Formation: Genomics

Spring 2008

Instructor: Liliana Florea

(W) 6:10-8:40 pm (Phillips 111)

Course objective: An overview of computational techniques related to the representation, storage, and information extraction from genomic data, including hands-on surveys of core genomics databases.

Topics (tentative):

I. Introduction to genomics

- Genomics: scope, data types, databases and knowledge bases
- Molecular biology basics

II. Genomic data representation and storage

- Sequence databases (GenBank)
- Sequence compression algorithms
- Sequence comparison (hashes, suffix trees, suffix arrays)
- Alignments

III. Knowledge extraction from biological sequences (‘annotation’)

- Information in biological sequences
- Genes: predictive (HMM-based) versus comparative (alignment-based) methods
- Regulatory regions: motif-finding versus motif-extraction
- Other features: repeats, CpG islands, structural RNAs
- Annotation environments: the UCSC Genome Browser and Database

IV. Biomedical applications

- Computational techniques for vaccine design, and/or
- Gene ontologies and protein functional annotation

Text: Lecture notes and copies of relevant articles describing current developments will be provided.

Grading: Short assignments/essays: 20%, midterm exam (take-home): 25%, class presentation (advanced topic): 25%, final project: 30%.

Who can attend: Open to graduate and upper-undergraduate students who have completed at least one algorithms and data structures course. You may contact the instructor (florea@gwu.edu) with questions.