CENG 734
Advanced Topics in Bioinformatics

Spring 2014-2015
Instructor Info

- Tolga Can
  - e-mail: tcan@ceng.metu.edu.tr
  - Office: B-109
  - e-mail me to schedule an appointment or to ask any course related question
  - Past courses:
    - http://www.ceng.metu.edu.tr/~tcan/
Course Web Page

http://www.ceng.metu.edu.tr/~tcan/ceng734_s1415/

- Lecture slides
- Syllabus
- Reading material (papers and additional optional readings)
- Announcements

I will use ODTU-CLASS for announcement of your grades and for sending urgent announcements via e-mail. So, check that CENG 734 is in your course list in ODTU-CLASS.
Prerequisites

- No formal prerequisites. However, some familiarity with Bioinformatics will help the students get the most benefit out of the course.
- Programming: required for the project
- Algorithms and Complexity Analysis
- Basic probability and statistics
- Some molecular and cellular biology terminology is required
- If you are new to Bioinformatics, I encourage you to take CENG 465 offered in Fall semesters
Prerequisites

- Motivation is the most important prerequisite.
- This is a research oriented course. Take it if your thesis is going to be on Bioinformatics.
Course Objectives

- The primary objectives of this course are to expose students to recent developments in the field of bioinformatics and to enable students initiate research in this area. Upon completion of this course the students will:
  - be aware of the current challenges in Bioinformatics,
  - have learnt the state-of-the-art methods to tackle important biological problems,
  - and be able to initiate and conduct research in the area of Bioinformatics.
Reading Material

- Reading material will be provided online on the course web site
- Mostly papers from recent conferences or journals
Grading

- Reading: 40%
  - 8-10 quizzes about the reading material
- Term project: 40%
- Final exam: 20%
Project

- May be related to your current research or what you may want to do for research
- Groups of 1-4 students
- You are free to choose project topics but will discuss details/goals/work plan with the instructor before starting to work on the project
- Project topic examples:
  - Small improvements on techniques/algorithms discussed in class
  - Application of a technique on a different data set.
Outline of the course

● This week: Introduction and characteristics of biological data. Who is working on what?

● Challenges in sequence analysis: next generation sequencing
  – Genome assembly, RNA-Seq, Sequence Variations

● Whole genome analysis, genome annotation

● Evolution and phylogeny

● Protein structure, functional classification
Gene regulation and transcriptomics
Text mining in bioinformatics
Protein interactions and molecular networks
Challenges in heterogeneous data integration
Bioimage informatics
Biological Data

- Comes in many different forms
  - Sequence Databases:
    - Nucleotide (GenBank), SWISS-PROT, Whole genome databases
  - Structure databases
    - Protein Data Bank
  - Expression data
    - NCBI GEO dataset
  - Interaction data, Pathways
  - Taxonomy data
  - Publication data (PubMed)
  - Domain, annotation information
# NCBI Entrez

Welcome to the Entrez cross-database search page

## Databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PubMed</strong></td>
<td>biomedical literature citations and abstracts</td>
</tr>
<tr>
<td><strong>PubMed Central</strong></td>
<td>free, full text journal articles</td>
</tr>
<tr>
<td><strong>Site Search</strong></td>
<td>NCBI web and FTP sites</td>
</tr>
<tr>
<td><strong>Books</strong></td>
<td>online books</td>
</tr>
<tr>
<td><strong>OMIM</strong></td>
<td>online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td><strong>OMIA</strong></td>
<td>online Mendelian Inheritance in Animals</td>
</tr>
<tr>
<td><strong>Nucleotide</strong></td>
<td>sequence database (includes GenBank)</td>
</tr>
<tr>
<td><strong>Protein</strong></td>
<td>sequence database</td>
</tr>
<tr>
<td><strong>Genome</strong></td>
<td>whole genome sequences</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>three-dimensional macromolecular structures</td>
</tr>
<tr>
<td><strong>Taxonomy</strong></td>
<td>organisms in GenBank</td>
</tr>
<tr>
<td><strong>SNP</strong></td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td><strong>Gene</strong></td>
<td>gene-centered information</td>
</tr>
<tr>
<td><strong>HomoloGene</strong></td>
<td>eukaryotic homology groups</td>
</tr>
<tr>
<td><strong>PubChem Compound</strong></td>
<td>unique small molecule chemical structures</td>
</tr>
<tr>
<td><strong>PubChem Substance</strong></td>
<td>deposited chemical substance records</td>
</tr>
<tr>
<td><strong>Genome Project</strong></td>
<td>genome project information</td>
</tr>
<tr>
<td><strong>UniGene</strong></td>
<td>gene-oriented clusters of transcript sequences</td>
</tr>
<tr>
<td><strong>CDD</strong></td>
<td>conserved protein domain database</td>
</tr>
<tr>
<td><strong>3D Domains</strong></td>
<td>domains from Entrez Structure</td>
</tr>
<tr>
<td><strong>UniSTS</strong></td>
<td>markers and mapping data</td>
</tr>
<tr>
<td><strong>PopSet</strong></td>
<td>population study data sets</td>
</tr>
<tr>
<td><strong>GEO Profiles</strong></td>
<td>expression and molecular abundance profiles</td>
</tr>
<tr>
<td><strong>GEO DataSets</strong></td>
<td>experimental sets of GEO data</td>
</tr>
<tr>
<td><strong>Cancer Chromosomes</strong></td>
<td>cytogenetic databases</td>
</tr>
<tr>
<td><strong>PubChem BioAssay</strong></td>
<td>bioactivity screens of chemical substances</td>
</tr>
<tr>
<td><strong>GENSAT</strong></td>
<td>gene expression atlas of mouse central nervous system</td>
</tr>
<tr>
<td><strong>Probe</strong></td>
<td>sequence-specific reagents</td>
</tr>
<tr>
<td><strong>Journals</strong></td>
<td>detailed information about the journals indexed in PubMed and other Entrez databases</td>
</tr>
<tr>
<td><strong>NLM Catalog</strong></td>
<td>catalog of books, journals, and audiovisuals in the NLM collections</td>
</tr>
</tbody>
</table>
Introductory Biology

DNA (Genotype) → Protein → Phenotype
AN ANIMAL CELL

- Nucleus
- Nucleolus
- Mitochondrion
- Cytoskeleton
- Ribosomes
- Golgi apparatus
- Centrioles
- Smooth endoplasmic reticulum
- Rough endoplasmic reticulum
- Peroxisome
- Plasma membrane
DNA

- Raw DNA Sequence
  - Coding or Not?
  - Parse into genes?
  - 4 bases: AGCT
  - ~1 Kb in a gene,
  - ~2 Mb in genome
  - ~3 Gb Human
Protein Sequence

• 20 letter alphabet
  – ACDEFGHIKLMNPQRSTVWY but not BJOUXZ

• Strings of ~300 aa in an average protein (in bacteria),
  ~200 aa in a domain

• >3M known protein sequences

• Uniprot
  – UniProtKB/Swiss-Prot: proteins with high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.) 519348 entries in August 10 release.
  – UniProtKB/TrEMBL: a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot. 11636205 entries in August 10 release.
Structures

- DNA/RNA/Protein
  - Mostly protein structures at PDB
Genes and Proteins

- One gene encodes one* protein.
- Like a program, it starts with start codon (e.g. ATG), then each three code one amino acid. Then a stop codon (e.g. TGA) signifies end of the gene.
- Sometimes, in the middle of a (eukaryotic) gene, there are introns that are spliced out (as junk) during transcription. Used parts are called exons. This is the task of gene finding.
Complete Genomes

- NCBI Entrez Genome Database
- Sequences from 6597 species
  - Achaea: 102
  - Bacteria: 1534
  - Eukaryote: 2455
  - Viruses: 2426
  - Viroids: 41
  - Plasmids: 39
Human genome

Genes and gene-related sequences 900Mb

- Noncoding DNA 810Mb
  - Pseudogenes
  - Gene fragments
    - Introns, leaders, trailers
  - Single-copy genes
  - Multi-gene families
  - Regulatory sequences

Coding DNA 90Mb

- Single-copy genes
- Multi-gene families

Extragenic DNA 2100Mb

- Repetitive DNA 420Mb
  - Non-coding tandem repeats
  - Genome-wide interspersed repeats
- Unique and low-copy number 1680Mb
  - Satellites DNA
    - Minisatellites
    - Microsatellites
  - DNA transposons
    - LTR elements
    - LINEs
    - SINEs

- Noncoding DNA 810Mb
  - Pseudogenes
  - Gene fragments
    - Introns, leaders, trailers

Coding DNA 90Mb

- Single-copy genes
- Multi-gene families

Regulatory sequences
Gene expression data

Figure 1. Schematic diagram of the relationships between GEO Platform, Sample, DataSet and Profiles. For each gene on a Platform (e.g. Gene A), multiple Sample measurement values are generated (Sample1–Sample3). Related Samples make up a DataSet, from which multiple, individual gene profile entities are generated.

as of September 26

from NCBI GEO NAR 2005 paper
Protein Network Data

from STRING database

The database covers 2,590,259 proteins from 630 organisms as of September 26.
Bioimages
Other Types of Data

- Information to understand genomes
  - Metabolic Pathways (glycolysis), traditional biochemistry
  - Regulatory Networks
  - Whole Organisms
    Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - The Literature
    (PubMed)
Data sources

- NAR (Nucleic Acids Research) journal maintains a list of data collections
  - Sequence
    - Genomes, ESTs, Promoters, transcription factor binding sites, repeats, ..
  - Structure
    - Domains, motifs, classifications, ..
  - Others
    - Microarrays, subcellular localization, ontologies, pathways, SNPs, ..
Challenges of working in bioinformatics

- Need to feel comfortable in an interdisciplinary area
- Depend on others for primary data
- Need to address important biological and computer science problems
Skill set

- Artificial intelligence
- Machine learning
- Statistics & probability
- Algorithms
- Databases
- Programming
- Molecular and Cellular Biology
- More?
Challenging sequence related problems

- More sensitive pairwise alignment
  - Dynamic programming is $O(mn)$
    - $m$ is the length of the query
    - $n$ is the length of the database

- Scalable multiple alignment
  - Dynamic programming is exponential in number of sequences
  - Currently feasible for around 10 protein sequences of length around 1000

- Shotgun alignment
  - Current techniques will take over 200 days on a single machine to align the mouse genome
Challenging structure related problems

- Alignment against a database
  - Single comparison usually takes seconds.
  - Comparison against a database takes hours.
  - All-against-all comparison takes weeks.

- Multiple structure alignment and motifs

- Combined sequence and structure comparison

- Secondary and tertiary structure prediction
• And many more other challenging problems in other areas of bioinformatics....
Top journals

- Science
- Nature (Nature Genetics, Nature Biotechnology)
- PNAS (Proceedings of the National Academy of Sciences)
- NAR (Nucleic Acids Research)
- Bioinformatics
- JCB (Journal of Computational Biology)
- BMC Bioinformatics
- Genome Research
- Proteins: Structure and Function, and Bioinformatics
- PLoS Computational Biology
- PLoS One
- IEEE/ACM Transactions on Computational Biology and Bioinf.
Top conferences

- RECOMB: Research in Computational Molecular Biology
- ISMB: Intelligent Systems for Molecular Biology
- ECCB: European Conference on Computational Biology
- PSB: Pacific Symposium on Biocomputing
- CSB: Computational Systems Bioinformatics
- CIBCB: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology
- BIBE: IEEE International Conference on Bioinformatics and Bioengineering
Current research?

- Bioinformatics recent issue
- BMC Bioinformatics web site
- PLoS Computational Biology recent issue
- RECOMB 2014 accepted papers
- ISMB 2014 accepted papers
- ECCB 2014 accepted papers
- PSB 2015 accepted papers
Next week

- An integrative approach to predicting the functional effects of non-coding and coding sequence variation by Shihab et al.
  - Bioinformatics Advance Access published February 11, 2015

- BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-seq profiles by Videm et al.