CENG 734
Advanced Topics in Bioinformatics

Fall 2006-2007

Instructor Info

- Tolga Can
  - e-mail: tcan@ceng.metu.edu.tr
  - Office: B-109
  - you can e-mail me to schedule an appointment or to ask any course related question
Class Web Page & E-mail list

http://www.ceng.metu.edu.tr/~tcan/ceng734/
- Lecture slides
- Syllabus
- Reading material (next week’s reading material is going to be posted today)
- Announcements

I will also maintain an e-mail list for announcements
- Sign-up for the e-mail list

Prerequisites

- No formal prerequisites. However, some familiarity with Bioinformatics will help the students get the most benefit out of the course.
- Programming: may be 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 Spring
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 web or by e-mail
- Mostly papers from recent conferences or journals
Grading

- Reading: 20%
  - Writing critics on papers or maybe small quizzes
- Term project: 50%
- Final exam: 30%

Project

- May be related to your current research or what you may want to do for research
- Groups of 2-3 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
  - Implementation of a paper and application on a different data set.
Outline

- This week: Introduction and characteristics of biological data. Who is working on what?
- Challenges in sequence analysis
- Challenges in structural bioinformatics
- Protein functional classification, genome annotation
- State of the art statistical techniques for solving bioinformatics problems
- Dealing with multiple heterogeneous data
- Challenges in protein/gene networks

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

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>Biomedical literature citations and abstracts</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>Free, full-text journal articles</td>
</tr>
<tr>
<td>Site Search</td>
<td>NCBI web and FTP sites</td>
</tr>
</tbody>
</table>

### NCBI Entrez Options

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide</td>
<td>Sequence database (GenBank)</td>
</tr>
<tr>
<td>Protein</td>
<td>Sequence database</td>
</tr>
<tr>
<td>Genomes</td>
<td>Whole genome sequences</td>
</tr>
<tr>
<td>Structure</td>
<td>Three-dimensional macromolecular structures</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Organisms in GenBank</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphisms</td>
</tr>
<tr>
<td>Gene</td>
<td>Gene-centered information</td>
</tr>
<tr>
<td>Homologene</td>
<td>Askaryotic homology groups</td>
</tr>
<tr>
<td>PubChem Compound</td>
<td>Unique small molecule chemical structures</td>
</tr>
<tr>
<td>PubChem Substance</td>
<td>Deposited chemical substance records</td>
</tr>
<tr>
<td>Genome Project</td>
<td>Genome project information</td>
</tr>
</tbody>
</table>

### Additional Resources

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniGene</td>
<td>Gene-oriented clusters of transcript sequences</td>
</tr>
<tr>
<td>CDD</td>
<td>Conserved domain database</td>
</tr>
<tr>
<td>3D Domains</td>
<td>Domains from Entrez Structure</td>
</tr>
<tr>
<td>UniSTS</td>
<td>Markers and mapping data</td>
</tr>
<tr>
<td>PopSet</td>
<td>Population study data sets</td>
</tr>
<tr>
<td>GEO Profiles</td>
<td>Expression and molecular abundance profiles</td>
</tr>
<tr>
<td>GEO Datasets</td>
<td>Experimental sets of GEO data</td>
</tr>
<tr>
<td>Cancer Chromosomes</td>
<td>Chromosomal databases</td>
</tr>
<tr>
<td>PubChem BioAssay</td>
<td>Bioassay screens of chemical substances</td>
</tr>
<tr>
<td>GDS</td>
<td>Gene expression atlas of mouse central nervous system</td>
</tr>
<tr>
<td>Probes</td>
<td>Sequence-specific reagents</td>
</tr>
<tr>
<td>Medline</td>
<td>Detailed information about NLM's controlled vocabulary</td>
</tr>
<tr>
<td>NLM Catalog</td>
<td>Catalog of books, journals, and audiovisuals in the NLM collections</td>
</tr>
</tbody>
</table>

---

### Introductory Biology

DNA (Genotype) → Protein → Phenotype
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.) 232345 entries as of September 19.
  – 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. 3189332 entries as of September 19.

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
- >800 organisms sequenced
  - Achaea: 34
  - Bacteria: 468
  - Eukaryote: 112
  - Viral genomes: 213
Human genome

Genes and gene-related sequences 900Mb
- Noncoding DNA 810Mb
  - Pseudogenes
  - Gene fragments
  - Introns, leaders, trailers
- Coding DNA 90Mb
  - Single-copy genes
  - Multi-gene families
  - Regulatory sequences
- Extragenic DNA 2100Mb
  - Regulatory sequences
  - Repetitive DNA 420Mb
    - Non-coding tandem repeats
    - Genome-wide interspersed repeats
    - Unique and low-copy number 1680Mb
  - Noncoding tandem repeats
    - Satellite DNA
      - Minisatellites
      - Microsatellites
    - DNA transposons
      - LTR elements
      - LINEs
      - SINEs

Coding DNA 90Mb
- Single-copy genes
  - Multi-gene families
  - Regulatory sequences
- Extragenic DNA 2100Mb
  - Regulatory sequences
  - Repetitive DNA 420Mb
    - Non-coding tandem repeats
    - Genome-wide interspersed repeats
    - Unique and low-copy number 1680Mb

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.

from NCBI GEO NAR 2005 paper
Protein Network Data

from STRING database

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 (MEDLINE)
Data sources

- NAR (Nucleic Acids Research) journal maintains a list of data collections (858 databases with the 2006 Database Issue update)
  - Sequence
    - Genomes (more than 800), 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 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

Top conferences

- RECOMB: Research in COnputational Molecular Biology
- ISMB: Intelligent Systems for Molecular Biology
- ECCB: European Conference on Computational Biology
- PSB: Pacific Symposium on Biocomputing
- CSB: Computational Systems Bioinformatics
- IEEE CIBCB 2006: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology
- BIRD (new conference): International Conference on Bioinformatics Research and Development
Who is doing what recently?

- Bioinformatics recent issue
- Journal of Computational Biology recent issue
- BMC Bioinformatics recent issue
- Proteins: Structure, Function and Bioinformatics recent issue
- RECOMB 2006 accepted papers
- ISMB accepted papers
- ECCB accepted papers
- CSB accepted papers

Next week

- Multiple sequence alignment problem
- Challenges
- Recent methods for MSA
- Reading:
  - One paper from ECCB 2006 (not yet published!)
  - One paper from NAR 2004
  - MSA tutorial: read the Wikipedia entry at: http://en.wikipedia.org/wiki/Multiple_sequence_alignment
  - See if you can find any other papers on MSA dated 2005 or 2006.