CENG 465
Introduction to Bioinformatics

Spring 2009-2010

Tolga Can (Office: B-109)
e-mail: tcan@ceng.metu.edu.tr

Course Web Page:
http://www.ceng.metu.edu.tr/~tcan/ceng465_s0910/
Goals of the course

• Working at the interface of computer science and biology
  – New motivation
  – New data and new demands
  – Real impact
• Introduction to main problems in bioinformatics
• Opportunity to interact with algorithms, tools, data in current practice
High level overview of the course

• A way of thinking -- tackling “biological problems” computationally
  – how to look at a “biological problem” from a computational point of view?
  – how to formulate a computational problem to address a biological issue?
  – how to collect statistics from biological data?
  – how to build a “computational” model?
  – how to solve a computational modeling problem?
  – how to test and evaluate a computational algorithm?
Course outline

- Motivation and introduction to biology (1 week)
- Sequence analysis (4 weeks)
  - Analyze DNA and protein sequences for clues regarding function
  - Identification of homologues
    - Pairwise sequence alignment
  - Statistical significance of sequence alignments
  - Profile HMMS
  - Multiple sequence alignment
  - Efficient pattern search: suffix trees
- Phylogenetic trees (1 week)
Course outline

- Protein structures (4 weeks)
  - Structure prediction (secondary, tertiary)
  - Analyze protein structures for clues regarding function
    - Structure alignment

- Microarray data analysis (2 weeks)
  - Correlations, clustering

- Gene/Protein networks, pathways (2 weeks)
  - Protein-protein, protein/DNA interactions
  - Construction and analysis of large scale networks
Grading

- Midterm exam - 30%
- Final exam - 40%
- Assignments (written/programming) - 30%
Miscellaneous

• Course webpage
  – Lecture slides and reading materials
  – Assignments
  – Other relevant information

• Newsgroup
  – metu.ceng.course.465
  – You should follow the newsgroup for course related announcements
  – Students from other departments should get a CENG account for this semester (Room: A-210) in order to access the newsgroup
Bioinformatics: A simple view

Biological Data + Computer Calculations
What is Bioinformatics?

• *(Molecular)* **Bio-informatics**

• One idea for a definition?
  Bioinformatics is conceptualizing *biology in terms of molecules* (in the sense of physical-chemistry) and then applying “*informatics*” techniques (derived from disciplines such as applied math, CS, and statistics) to understand and *organize the information associated* with these molecules, *on a large-scale*.

• Bioinformatics is a practical discipline with many applications.
Computing versus Biology

• what computer science is to molecular biology is like what mathematics has been to physics ......
  -- Larry Hunter, ISMB’94

• molecular biology is (becoming) an information science ......
  -- Leroy Hood, RECOMB’00

• bioinformatics ... is the research domain focused on linking the behavior of biomolecules, biological pathways, cells, organisms, and populations to the information encoded in the genomes
  --Temple Smith, Current

Topics in Computational Molecular Biology
Computing versus Biology
looking into the future

• Like physics, where general rules and laws are taught at the start, biology will surely be presented to future generations of students as a set of basic systems duplicated and adapted to a very wide range of cellular and organismic functions, following basic evolutionary principles constrained by Earth’s geological history.

--Temple Smith, Current Topics in Computational Molecular Biology
DNA (Genotype) → Protein → Phenotype
Scales of life
Animal Cell

- Mitochondrion
- Nucleolus (rRNA synthesis)
- Nucleus
- Plasma membrane
- Cell coat
- Chromatin
- Cytoplasm
- Lots of other stuff/organelles/ribosome
Two kinds of Cells

• Prokaryotes – no nucleus (bacteria)
  – Their genomes are circular

• Eukaryotes – have nucleus (animal, plants)
  – Linear genomes with multiple chromosomes in pairs. When pairing up, they look like

/  
Middle: centromere
Top: p-arm
Bottom: q-arm
Molecular Biology Information - 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

atggcaat...
Molecular Biology Information: Protein Sequence

- 20 letter alphabet
  - ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria),
  ~200 aa in a domain
- ~1M known protein sequences
Molecular Biology Information: Macromolecular Structure

- DNA/RNA/Protein
  - Almost all protein
Structure summary

• 3-d structure determined by protein sequence
• Cooperative and progressive stabilization
• Prediction remains a challenge
  – ab-initio (energy minimization)
  – knowledge-based
    • Chou-Fasman and GOR methods for SSE prediction
    • Comparative modeling and protein threading for tertiary structure prediction
• Diseases caused by misfolded proteins
  – Mad cow disease
• Classification of protein structures
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. Good parts are called exons. This is the task of gene finding.
<table>
<thead>
<tr>
<th>Amino Acid (AA)</th>
<th>Codon(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine (GLY)</td>
<td>GG*</td>
</tr>
<tr>
<td>Alanine(ALA)</td>
<td>GC*</td>
</tr>
<tr>
<td>Valine (VAL)</td>
<td>GT*</td>
</tr>
<tr>
<td>Leucine (LEU)</td>
<td>CT*</td>
</tr>
<tr>
<td>Isoleucine (ILE)</td>
<td>AT(*-G)</td>
</tr>
<tr>
<td>Serine (SER)</td>
<td>AGT, AGC</td>
</tr>
<tr>
<td>Threonine (THR)</td>
<td>AC*</td>
</tr>
<tr>
<td>Aspartic Acid (ASP)</td>
<td>GAT,GAC</td>
</tr>
<tr>
<td>Glutamic Acid (GLU)</td>
<td>GAA,GAG</td>
</tr>
<tr>
<td>Lysine (LYS)</td>
<td>AAA, AAG</td>
</tr>
<tr>
<td>Start: ATG, CTG, GTG</td>
<td></td>
</tr>
<tr>
<td>Arginine (ARG)</td>
<td>CG*</td>
</tr>
<tr>
<td>Asparagine (ASN)</td>
<td>AAT, AAC</td>
</tr>
<tr>
<td>Glutamine (GLN)</td>
<td>CAA, CAG</td>
</tr>
<tr>
<td>Cysteine (CYS)</td>
<td>TGT, TGC</td>
</tr>
<tr>
<td>Methionine (MET)</td>
<td>ATG</td>
</tr>
<tr>
<td>Phenylalanine (PHE)</td>
<td>TTT,TTC</td>
</tr>
<tr>
<td>Tyrosine (TYR)</td>
<td>TAT, TAC</td>
</tr>
<tr>
<td>Tryptophan (TRP)</td>
<td>TGG</td>
</tr>
<tr>
<td>Histidine (HIS)</td>
<td>CAT, CAC</td>
</tr>
<tr>
<td>Proline (PRO)</td>
<td>CC*</td>
</tr>
<tr>
<td>Stop: TGA, TAA, TAG</td>
<td></td>
</tr>
</tbody>
</table>
Molecular Biology Information: Whole Genomes

Genome sequences now accumulate so quickly that, in less than a week, a single laboratory can produce more bits of data than Shakespeare managed in a lifetime, although the latter make better reading.

1995
Bacteria, 1.6 Mb, ~1600 genes [Science 269: 496]

1997
Eukaryote, 13 Mb, ~6K genes [Nature 387: 1]

1998
Animal, ~100 Mb, ~20K genes [Science 282: 1945]

2000?
Human, ~3 Gb, ~100K genes [???]

Genomes highlight the Finiteness of the “Parts” in Biology
Human Genome Project

Impacting many disciplines

Global Carbon Cycles
Industrial Resources • Bioremediation

Evolutionary Biology • Biofuels • Agriculture • Forensics

Molecular and Nuclear Medicine • Health Risks

Courtesy
U.S. Department of Energy
Human Genome Program
Gene Expression Datasets:

The Transcriptome

Young/Lander, Chips, Abs. Exp.

Also: SAGE; Samson and Church, Chips; Aebersold, Protein Expression

Brown, microarray, Rel. Exp. over Brown, microarray, Rel. Exp. over

Protein Expression

Timecourse

Snyder, Transposons, Protein Exp.

Young/Lander, Chips, Abs. Exp.

Protein Expression
Systematic Knockouts


Other Whole-Genome Experiments

Construction of a modular yeast two-hybrid cDNA library from human EST clones for the human genome protein linkage map


For yeast: 6000 x 6000 / 2 ~ 18M interactions
Molecular Biology Information: Other Integrative Data

- Information to understand genomes
  - Metabolic Pathways (glycolysis), traditional biochemistry
  - Regulatory Networks
  - Whole Organisms
    Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - The Literature (MEDLINE)
- The Future....
Organizing Molecular Biology Information: Redundancy and Multiplicity

- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code

**How do we find the similarities?**

**Integrative Genomics -** genes ↔ structures ↔ functions ↔ pathways ↔ expression levels ↔ regulatory systems ↔ ....
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

- Repetitive DNA
  - 420Mb
  - Non-coding tandem repeats
  - Genome-wide interspersed repeats

- Unique and low-copy number
  - 1680Mb

- Non-coding tandem repeats
  - Satellite DNA
  - Minisatellites
  - Microsatellites
  - DNA transposons
  - LTR elements
  - LINEs
  - SINEs
Where to get data?

• **GenBank**

• **Protein Databases**

• **And many others**
Data

• Diversity and size of information
  – Sequences, 3-D structures, microarrays, protein interaction networks, *in silico* models, bio-images

• Understand the relationship
  – Similar to complex software design
## Scalability challenges

  - **Sequence**
    - Genomes (more than 150), 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

• Programming
• Algorithms
• Machine learning/Pattern recognition/AI
• Statistics & probability
• Mathematics