CIS 454/554 Bioinformatics

Instructor: John Conery
    conery@cs.uoregon.edu
Text: Introduction to Bioinformatics, by Arthur Lesk
Web page: http://www.cs.uoregon.edu/classes/W10/cis454

My Background

- Professor of Computer Science
  - PhD UC Irvine, 1983
  - 25th year at UO
- Early research interests
  - parallel processing
  - implementation of logic programming languages
- Computational science interests
  - population models
  - gene duplication
  - (both with Mike Lynch, UO and IU)

Current Projects

- A catalog of human genes and their zebrafish co-orthologs (with John Postlethwait and lab)
- Automated construction of gene family databases, software for phylogentic reconstruction (Joe Thornton)
- Software for designing, managing analysis pipelines
- Sequence alignment using information theory

What is Bioinformatics?

- Informatics: European term for “computer science”
- In US, “information science” has included business systems (e.g. MIS), library science, other fields that use large amounts of data
- Emphasis: managing existing information
- CS issues:
  - databases
  - knowledge representation
  - human-computer interaction
  - networks
Informatics in Academia

- Several universities have established schools or colleges of informatics
  - Georgia Tech, Indiana, UC Irvine, …
- NSF ITR (Information Technology Research)
  - major interdisciplinary research initiative
- At UO: undergraduate minor in Computer Information Technology (CIT)

Computational Biology

- Computational Science: “third leg”
  - (theoretical, experimental, computational) science
- Emphasis: creation of new information or knowledge
- Computational Biology:
  - molecular dynamics
  - systems biology
  - population modeling

Bioinformatics

- Combination of information technology and computational molecular biology
- Computational approach to studies of genes, gene expression, molecular evolution, and other areas of biology
- Makes use of very large distributed databases

Bioinformatics (cont’d)

- Combines elements of informatics (managing existing information) and science (creating new information)
- Example:
  - “genefinder” algorithm locates hypothetical new gene X
  - researchers look for similar genes in a sequence database (“remote homology search”)
  - use information about known sequences to infer structure and function of the protein coded by X
Molecular Biology

- Core discipline for most of bioinformatics
- Study of biological systems at the molecular level
  - DNA
    - structure, replication, expression (e.g. translation to proteins)
  - Proteins
    - structure, function (e.g. hemoglobin transports oxygen, transcription factors used in DNA expression)
- Next Lecture: Molecular Biology for Computer Scientists

Bioinformatics Areas

- The next set of slides will highlight a few active areas of bioinformatics research
- For each area there is a brief description of the role of math, computer science, and informatics

DNA Sequencing

- Goal: uncover the molecular structure of complete genomes
  - human genome
  - model organisms (mouse, zebrafish, fruitfly, yeast, rice, bacteria, ...)
- DNA: long strand of nucleotides
  - sequencing: determining the exact sequence
  - uses a combination of lab techniques and computer science
- CS/IT:
  - “shotgun” sequencing, sequence assembly
  - DNA sequence databases

Genome Annotation

- Genes are short sequences of DNA interspersed between long segments of “junk DNA”
- Gene mapping: where are the genes?
  - genetic (linkage) maps
  - physical maps
- Gene finders
- Identifying, locating other genomic “features”
- CS/IT:
  - data mining
  - machine learning
Sequence Similarity

- DNA sequences are represented as strings
- Mutations -- replacements, insertions, and deletions -- change DNA over time
- Alignment can help quantify sequence “similarity”
  - ATGACCTGA       original sequence
  - ATGCCTGA        4 mismatches: not similar
  - ATG–CCTGA       1 insertion, 0 mismatch: similar?
- CS/IT:
  - pattern matching
  - statistics

Homology Search

- Two sequences are homologs if they are descended from a common ancestor
  - paralog: related gene in same organism, e.g. “gene family”
  - ortholog: related gene in separate organism
- Goal: given a sequence, search for highly similar sequences in a database
- CS/IT:
  - pattern matching
  - statistics
  - numerical analysis

Molecular Evolution

- If mutations occur at a constant rate, the number of changes can be used as a “molecular clock”
- Given two homologous sequences, can we determine how long ago they diverged from their common ancestor?
- CS/IT:
  - statistics
  - algorithms (e.g. numeric optimization)

Phylogenetics

- Given a set of several sequences, can we infer evolutionary relationships (phylogenetic tree)?
  - If A,B more similar than A,C or B,C:
- CS/IT:
  - algorithms (e.g. multiple alignment)
  - statistics
  - parallel processing
Protein Structure

- Proteins are polymers: chains of smaller molecules (amino acids)
- the nucleotide sequence in a gene is the “blueprint” for a protein
- genes are transcribed, translated into amino acid sequences
- The chains fold up into complex 3D structures
- Assumption: sequence determines structure
- CS/IT:
  - modeling and simulation (e.g. molecular dynamics)
  - databases

Protein Function

- “Binding sites” on the surface of a protein are critical to its function
  - e.g. O2 molecules attach to surface of hemoglobin
- Assumption: structure determines function
- CS/IT:
  - modeling and simulation
  - knowledge representation (ontology)
  - databases

Course Goals

- Introduction to main areas of bioinformatics
  - Existing software tools
  - Web resources
- In-depth study of a few areas
  - Overview of algorithms
  - Related informatics technology (HTML, XML, Perl, Ruby, Python, PHP, MySQL)
  - Software projects to implement, explore basic methods

Course Outline

- Background: Molecular Biology
- Genome and sequence databases
- Homology
  - define “sequence similarity”
  - algorithms for sequence alignment (pairwise, multiple)
  - homology search (e.g. BLAST)
- Phylogenetic inference
- Other topics, if time and/or interest:
  - sequence assembly
  - gene finding algorithms
  - protein structure and function

Term Project

Help out on a current research project from UO CEEB or MolBio
Learn software, algorithms, etc by applying them in real-world project

Help topics driven by term projects...
Course Activities

- Construct a local gene database (using existing software)
- Pairwise alignment (C++ programming project)
- Phylogenetic inference
- Term project
  - Group project
  - Define your own project, or build on existing projects
- Term paper
- No exams

Project #0

- Due ASAP
- Send e-mail to conery@cs.uoregon.edu:
  - your name (full name, nickname)
    (call me “John”)
  - your background
    programming skills
    biology background

Project #1

- Due Monday, Jan 11
- Locate, download *E. coli* genome
  - look for *Escherichia coli K12*
  - best bet: ncbi.nih.gov
- Look for NCBI format genome annotation file (“Genbank report”, or .gbk file)
  - text file with descriptions of features and the complete genomic sequence
- See project web page for details
  - goal is to familiarize yourself with file format