

# CMSC 838T: High Performance Computing & Bioinformatics

---



**Chau-Wen Tseng**

Department of Computer Science  
University of Maryland, College Park

## CMSC 838T

- ◆ **Bioinformatics**
  - The creation and development of advanced information and computational techniques for solving problems in biology
- ◆ **High Performance Computing (HPC)**
  - Hardware and software techniques for building computer systems to quickly perform large amounts of computation

## CMSC 838T

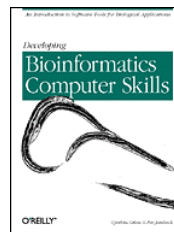
### Course goals

1. Learn algorithms and characteristics of bioinformatic applications
2. Examine software techniques used in high-performance computing
3. Study how to apply high-performance computing to bioinformatic applications

## CMSC 838T

### ◆ Textbook

- “Developing Bioinformatic Computer Skills”
- Gibas & Jambeck
- O’Reilly
- High-level overview
- Supplement with papers



## CMSC 838T

- ◆ **This course will not**
  - Train you to be a programmer
  - Train you to be a biochemist
  
- ◆ **This course will (hopefully)**
  - Teach basic concepts in bioinformatics
  - Allow you to work with researchers in bioinformatics
  - Begin training you to be a bioinformatics researcher
  
- ◆ **To do (relevant) research in bioinformatics**
  - Need to learn some biochemistry
  - Need to work with molecular biologists, biochemists

## CMSC 838T

- ◆ **My background**
  - High performance computing
  - Parallelizing compilers
  - Programming environments
  
- ◆ **Your background (hopefully)**
  - Computer science
    - Programming
    - Compilers
  - Bioinformatics
    - Basic biology
    - Basic chemistry

## CMSC 838T

- ◆ **Course organization**
  - I will present some lectures on bioinformatics, computing
  - Students will read & present some papers on bioinformatics
- ◆ **Projects (tentative)**
  - Access web-based bioinformatic tools & databases
  - Install, evaluate, and modify bioinformatic software
- ◆ **Grading (tentative)**
  - 50% Exams
  - 20% Presentations
  - 30% Projects

## Premise of Bioinformatics

- ◆ **Gene sequences determine biological function**
  - Genomic DNA → Amino acids → Proteins → Function
- ◆ **Similar composition → similar function?**
  - DNA sequences
  - Amino acid sequences
  - Protein 3D structure
- ◆ **Predicting protein function**
  - Designer drugs
  - Personalized treatments

## Bioinformatics

### ◆ Determining protein function

- Hard way
  - Biological / chemical analyses
  - Determine 3D structure w/ x-ray crystallography, NMR
- Easy way?
  - Sequence protein / DNA → find close match in database
  - Guess function based on match
  - Validate guess in lab

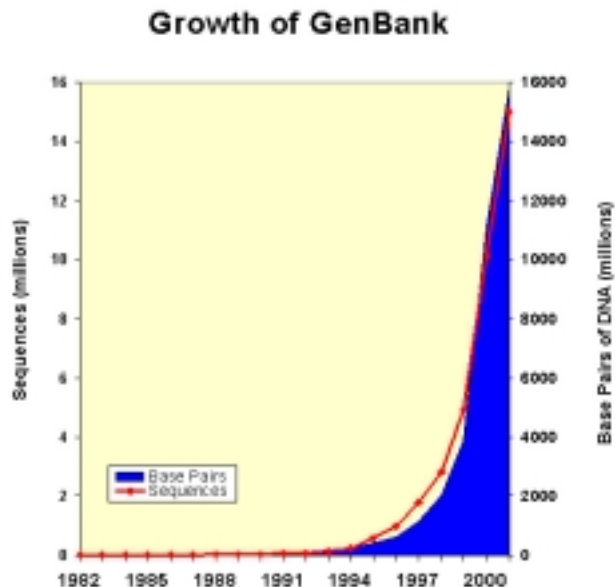
### ◆ Bioinformatics is imprecise

- Similar to data-mining
- Usually only suggest possible relationships
- Must validate correlation → causation

## Computers and Bioinformatics

- ◆ Amount of biological information quickly increasing

- ◆ Computers & software are needed to organize & analyze data



## Growth of Bioinformatics

- ◆ **1970's**
  - DNA sequencing
  - Alignment w/ Smith-Waterman (dynamic programming)
- ◆ **1980's**
  - Sequence databases (EMBL, GenBank)
  - Alignment w/ FASTA (linked lists, hashing)
- ◆ **1990's**
  - Automatic DNA sequencing
  - Alignment w/ BLAST (neighborhood words, probabilities)
  - Internet & WWW
- ◆ **Now**
  - Genomics, proteomics

## Bioinformatics Topics

- ◆ **Sequence alignments**
  - Find similarity between DNA / protein (amino acid) sequences
- ◆ **Genome assembly**
  - Combining genomic fragments to form whole genome
- ◆ **Gene identification & annotation**
  - Identify and classify genes on the genome
- ◆ **Microarrays & gene expression analysis**
  - Use DNA microarray (gene chip) to measure mRNA
- ◆ **Protein folding**
  - Compute 3D protein structure ↔ protein sequence
- ◆ **Phylogenetic analysis**
  - Find genetic relationships between sequences / species

## Open Problems in Bioinformatics

- ◆ Find genomes of all organisms
- ◆ Identify and annotate all genes
- ◆ Compute sequence ↔ 3D structure for all proteins
- ◆ Compare DNA / protein sequences for similarity
- ◆ Compare families of DNA / protein sequences
  
- ◆ Reason to be optimistic
  - Biology is finite...
    - ~30,000 human genes
    - ~1000 protein superfamilies
  - ...but computers keep improving!

## High Performance Computing (HPC)

- ◆ Increase available computation power
- ◆ Exploit parallelism
  - Custom supercomputers becoming too expensive
  - Use multiple processors in parallel
  - Application must be parallelized
- ◆ Exploit locality
  - processors faster than memory, network
  - in cache → avoid memory latency
  - on processor → avoid network latency

## High Performance Computing Topics

### ◆ Architectures

- Shared-memory multiprocessors
- Cluster & distributed processors

### ◆ Software

- Parallel programming languages / paradigms
- **Compilers**
  - Program analysis
  - Program transformations
  - Locality optimizations
  - Parallelism optimizations
- Run-time systems