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
- Elements of informatics (managing existing information) and science (creating new information)
  - Example: algorithm locates hypothetical new gene; researchers look for similar genes in a sequence database; use information about known sequences to infer structure, function of the protein coded by the new gene
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)
- Central Dogma:
  
  DNA → RNA → Protein

  (see next slide)
Central Dogma

- Cell nucleus contains chromosomes (DNA)
- DNA is *transcribed* (mRNA)
- RNA is *translated* by a ribosome to create a protein molecule
Bioinformatics Areas

- The next set of slides will highlight 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,...)
- **CS/IT:**
  - “shotgun” sequencing
  - clone, contig databases
  - sequence assembly
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 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
  - ATGCCCTGA 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)
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)
- parallel processing
Protein Structure

- Proteins are polymers: chains of smaller molecules (amino acids)
- 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. $O_2$ molecules attach to surface of hemoglobin
- Assumption: structure determines function
- CS/IT:
  - modeling and simulation
  - knowledge representation (ontology)
  - databases