CSCE 471/871 Lecture 0: Administrivia

Stephen Scott

sscott@cse.unl.edu

Welcome to 471/871!

- Check your name on the roster, or write your name if you're not listed
- Introduce yourself
  1. Who are you?
  2. What are you?
  3. Why are you here?
  4. What is one thing about you that few others know about?
- You should have the following handouts:
  1. Syllabus
  2. Copies of slides
- **Bring a laptop on Thursday!**

CSCE 471/871 Lecture 1: Introduction

Stephen Scott

(With thanks to Andy Benson and Jitender Deogun)

sscott@cse.unl.edu

Outline

- What is bioinformatics?
- Relevant biology background
- Fundamental questions in bioinformatics
- What we will (and will not) cover in this course

What is Bioinformatics?

- Bio = (molecular) biology
- Informatics = computer science
- Bioinformatics = using computer science tools and techniques for solving problems in (molecular) biology
- (Loose) synonym: Computational Biology

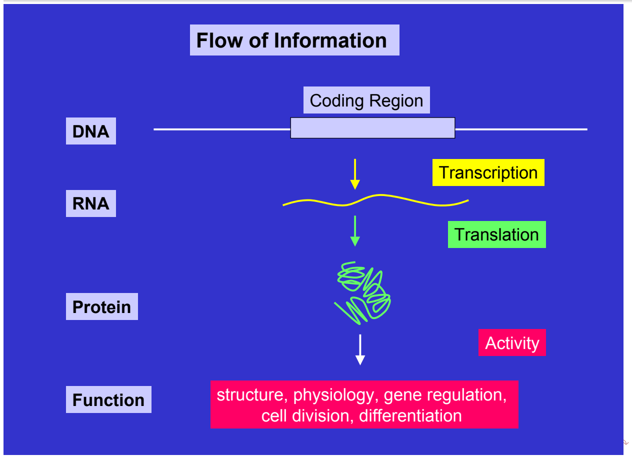
What is Bioinformatics? (cont'd)

- Original motivation comes from molecular biology
  - Sequence analysis
  - Most accurate analysis is via experimentation ("bench work"), but expensive and time-consuming (e.g., GenBank has $> 1.5 \times 10^{11}$ base pairs from $> 1.6 \times 10^{8}$ sequences)
- Bio problems suggest computational problems, which then suggest new biological experiments

Relevant Biology Background

- Basic idea: genes (chains of nucleotides) are converted into proteins (chains of amino acids)
- Proteins are the "workhorses" of biological systems, governing metabolic processes
  - E.g., blood clotting is a process that consists of a chain reaction of numerous protein interactions

Relevant Biology Background
Flow of Information



Relevant Biology Background
DNA and Genes

1. An organism's DNA is a (long) sequence of nucleotides (bases, residues), from {Adenine (A), Guanine (G), Cytosine (C), Thymine (T)}
2. Cellular machinery transcribes the coding regions of DNA into RNA
   - Has same alphabet, substituting U (uracil) for T
   - Non-coding regions are not transcribed

. . . ATTGATA $\boxed{\text{ATGCTGAACTACAAATTACGGCAGGCAACCGGAGCCTGGAAGTGA}}$ TAGGA . . .

$\Downarrow$

AUGCUGAACUACAAAUUACGGCAGGCAACCGGAGCCUGGAAGUGA

Relevant Biology Background
DNA and Genes (cont'd)

3. Then introns (non-coding subsequences) are removed, yielding mRNA
   - Adjacent triples are codons, each encoding an amino acid
4. mRNA is translated codon-by-codon into a polypeptide by ribosomes (organelles in cells' cytoplasm)
5. Proteins are comprised of one or more polypeptide chains

AUGCUG $\boxed{\text{AA}}$ CUA $\boxed{\text{C}}$ AAAUUACGGCAGGCAACCGGAGCCUGGAAGUGA

$\Downarrow$

AUG CUG CUA AAA UUA CGG CAG GCA ACC GGA GCC UGG AAG UGA

$\Downarrow$

M L L K L R Q A T G A W K [X]

Relevant Biology Background
Translation

First position 5' end	Second Position U	C	A	G	Third position 3' end
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

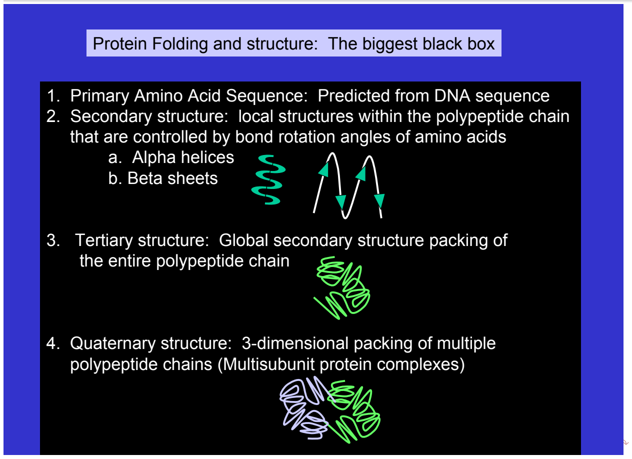
Genetic code is degenerate

64 codons
20 amino acids

Relevant Biology Background
Symbols for Amino Acids

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Apartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Relevant Biology Background
Protein Structure



Protein Folding and structure: The biggest black box

1. Primary Amino Acid Sequence: Predicted from DNA sequence
2. Secondary structure: local structures within the polypeptide chain that are controlled by bond rotation angles of amino acids
   a. Alpha helices
   b. Beta sheets
3. Tertiary structure: Global secondary structure packing of the entire polypeptide chain
4. Quaternary structure: 3-dimensional packing of multiple polypeptide chains (Multisubunit protein complexes)

Some Fundamental Questions

- Given an organism, what is its genetic sequence?
  ⇒ Sequence assembly
- Given a sequence, what genes does it encode?
  ⇒ Gene finding
- Given a protein:
  - What is its structure?
    ⇒ Structure prediction
  - What other proteins is it related to?
    ⇒ Homology prediction/phylogeny
  - What is its function?
    ⇒ Function prediction
- All this from (mainly) only sequences of letters!

What We Will Study

- Pairwise alignment of sequences
- Multiple alignment of sequences
- Profiling (modeling) a multiple alignment
- Building phylogenetic (evolutionary) trees (time permitting)
- Predicting secondary structure and/or function of RNA and proteins (time permitting)

What We Will *Not* Study
(but are still interesting problems)

- Gene finding
- Inferring metabolic pathways
- Predicting tertiary structure of proteins