Bioinformatics
Biology Review

• The genetic code is stored in DNA – Deoxyribonucleic acid.
• DNA molecules are chains of four nucleotide bases
  • Guanine, Thymine, Cytosine, Adenine
• DNA is normally found in pairs, held together by hydrogen bonds between the bases
  • A-T, T-A, G-C, C-G
• DNA is separated into chromosomes (humans have 23)
  • Around 3 billion base pairs in humans

http://www.llnl.gov/str/June03/gifs/Stubbs1.gif
RNA

• Ribonucleic acid – similar to DNA, but differs in several respects:
  • The sugar in RNA is ribose, not deoxyribose.
  • Instead of thymine, RNA contains uracil.
  • Ordinarily single-stranded.
  • Different function (more later).
• Both RNA and DNA are directional
  • The carbon atoms in sugar are numbered from 1’ to 5’
  • Each has a 5’ end and a 3’ end.

http://www.blc.arizona.edu/INTERACTIVE/DNA3/dnastruct.html
Proteins

- DNA encodes proteins – the actual building blocks of life.
- Proteins are chains of amino acids.
- 20 Different amino acids in all:
  - alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, valine
- The DNA that encodes a single protein is called a gene.
- Human DNA contains approximately 19,000 – 20,000 genes (as of Nov. 2014).
From DNA to Protein

• We will only look at the process in eucaryotes
  • Eucaryotes – cells have nuclei
  • Procaryotes – no nuclei
• DNA remains in the cell nucleus.
• Two main steps are required to create a protein:
  • Transcription – copy the genetic code
  • Translation – convert the code into a protein
Transcription

- An enzyme called RNA polymerase separates the DNA strands and makes a complementary RNA copy of the gene.
  - This RNA copy is called mRNA - messenger RNA.
  - Transcription moves in a 3' to 5' direction.
- Complication: genes are divided into introns and exons.
  - exons – regions that actually code proteins
  - introns – regions that do not.
- After the mRNA copy is created, introns are removed - splicing
  - The spliced mRNA then leaves the nucleus
Translation

• Ribosome reads mRNA and assembles the protein.
• 4 nucleotides, 20 amino acids. What is the code?
• Amino acids are encoded by triplets of nucleotides – codons.
• A single amino acid may be encoded by multiple codons.
• Three codons code for STOP – the end of an encoding region.
The Whole Process...

http://csbl.bmb.uga.edu/resources/bcmb3600.php
Bioinformatics

• Fredj Tekaia at the Institut Pasteur offers this definition of bioinformatics:
  “The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.”

  http://bioinformatics.org/faq/#definitions

• Two ways of thinking about bioinformatics:
  • Using mathematical, statistical, computing methods
  • Designing mathematical, statistical, computing methods.
Problems in Bioinformatics

- Sequencing
- Searching sequence data
- Finding genes
- Finding coding regions
- Predicting protein structure and function
- Analyzing gene expression data (microarray analysis)
- Comparing genomes to discover relationships between species.
Sequencing

- Complete draft of the human genome was released in Feb. 2001.
- There were two competing projects:
  - The human genome project – a large government funded consortium.
  - Celera Genomics – Private company held by Craig Venter
• 3 billion base pairs in human genome alone.
• All of this data is useless without tools to find what we are looking for.
• Consider the problem of searching for a 200bp sequence:
  • 3,000,000,000 possible locations for a match.
  • Each location requiring 200 comparisons.
  • Even if we can do 10,000,000 comparisons per second, search will take 16 hours.
• There are better ways....
BLAST – Basic Local Alignment Search Tool

NCBI – National Center for Biotechnology Information

- Vast repository of bioinformatics data.
- Provides BLAST access to many databases.
  - Nucleotide blast, protein blast, blastx.

- Entrez – The Life Sciences Search Engine:
  - Like Google for biological data.

- Let’s take a look….
19,000-20,000 Genes?

• Why the wide range? We have the whole genome. Here's the problem:

...atcg&tcatcgatcggaaatctagctaccaatgccatta
gacagtagccagatcatcaggggaatctctcgtcatcgat
cggaatctagctacccattagacagtagccagatgca
gatgggctagcgatagcatcaggggaatcatctgatc
gatcgggaatctagctaccaatgcccattagacagtagccag
gatcagatgggctagcgatagcatcaggggaatc...

• How can we tell where genes start and stop?
• How can we tell which regions encode proteins and which don't?
• Lots of “junk”- introns, complementary strands.
Reading Frames

• Any sequence of DNA has six “reading frames”
  • Three in one direction, three in the other
  • `aactgactagcagcagtcatggaactg`
  • `aac tga cta gca gca gtc atg gaa ctg`
  • `a act gac tag cag cag tca tgg aac tg`
  • `aa ctg act agc agc agt cat gga act g`

• Open reading frame – a reading frame that starts with the codon atg, and ends with one of the three stop codons (tag, tga, taa).
• Even if a sequence meets this criteria, it may not encode protein.
Finding Genes

- Certain sequences, *promoters*, tend to precede genes.
- Open reading frames correspond to potential genes.
- The only sure-fire way is to experimentally check if the sequence encodes a protein.
Finding Coding Regions

- Not exactly the same as finding genes – introns.
- Certain adjacent pairs of codons are more likely in coding regions – dimers, or hexmers.
- Some base sequences generally mark the boundary between introns and exons:
  - Donor site: coding region | GT
  - Acceptor: YAG | coding region
More About Proteins

- Protein function is determined by structure.
  - Primary: the sequence of amino acids
  - Secondary: local structure, like the DNA helix
  - Tertiary: folds in the secondary structure

- Proteins are organized into families with similar structure and function.
Comp. Issues Related to Proteins

- Clustering based on amino acid sequences.
  - Discovery of protein classes.
  - Finding class membership.

- Predicting protein structure.
  - Proteins are self-assembling – possible to predict structure.
Gene Microarrays

Programming Project 2...