Course Info

- 4 problem sets, 4-5 problems each, 3 late days total
- Can work in groups of 2-3 or individually

Course Intro

- Computational genomics is interdisciplinary
- Molecular biology was born in ~1954, when Watson & Crick published the classic paper on the structure of the DNA helix
    - Watson & Crick were not biologists - physicist, ornithologist
- Biology is becoming more and more a computational science
- 20th century laid the foundation for 2 key paradigms in genetics
    - (1) the Central Dogma of Molecular Biology - how is information passed across generations? How is does it store information, and how do these instructions become a physical reality in an organism?
        - DNA is transcribed into RNA; RNA is translated into protein sequences; protein sequences fold into protein structures; these protein structures form the building blocks of life
    - (2) Evolution (started in the 19th century) -- tells us how life arose, at least once
    - Note that these discoveries are both descriptive, rather than predictive
- This class is about biology shifting from a descriptive science to a very precise, engineering discipline
- 1990 - was clear we had the technology to sequence a fair amount of DNA; **Human Genome Project** started, with budget of $3 billion
    - only 1.5% of the genome consists of coding patterns
    - 1997: Webers & Meyers - computational simulation suggests that we should sequence using the whole-genome shotgun method (Craig Venter hired one of them and announced they'd finish sequencing the human genome by 2000)
    - This induced a race toward sequencing the whole human genome by 2000 for the NIH, announced success in 2000
- 2004: $30 million for human genome → 2008: $100,000 → 2010: $10,000 → 2016: $1,250 -- progressing at a rate faster than Moore's Law! (Life Technologies, Illumina, etc. were competing, helping to push the price → plateaued off because Illumina had a monopoly for some time
- At the current rate, it seems like in the future, everyone will take advantage of sequencing (eg. prenatal applications, liquid biopsies)
    - The applications of sequencing are vast
    - Sequencing is really just a basic molecular biology assay
    - Several companies are emerging in this field
    - Now the questions is when, rather than if, we will all be sequenced
    - Optimistic view: could help us beat cancer soon, big data could help us understand diet, etc.

Intro to Molecular Biology

- Biology comes from the Greek word bios, meaning “life”
- Cell - basic building block of life
    - Organisms can be divided into two broad categories - prokaryotes, eukaryotes
        - Prokaryotes - unicellular, cells do not have a nucleus, very early or primitive organisms -- but also the most successful forms of life, in that they are ubiquitous (we have more prokaryotic cells than eukaryotic cells in our body by a factor of 10!)
        - Eukaryotes - contains:
            - Nucleus - stores DNA
            - Other organelles (see figure)
- Structure of DNA
    - Nucleotides: 4 bases: Adenine, Thymine, Cytosine, Guanine
    - Phosphate backbone
    - A single strand is composed of a double helix where A pairs with T and C pairs with G; allows one strand to pair with its reverse complement
    - DNA is always written 5’ to 3’
    - Strands wrap around 8 roughly identical histone proteins; about 146bp wrap around one set of histones
    - These are folded into higher level helical structures, all the way into the chromosome
- Structure of RNA
    - Chemical similar to DNA
    - Bases: adenine, guanine, uracil, cytosine
- The Human Genome
    - ~3 billion nucleotides in human genome
    - But the length of the genome is not necessarily proportional to the complexity of the organism
    - Genes are shorter subsequences of the DNA (in the thousands or tens of thousands of bps)
    - ~20,000 genes in the human genome
- Gene Transcription
    - Transcription factors recognize transcription factor binding sites (specific subsequences) and bind them, forming a complex -- this is a regulatory mechanism, because this process depends on the concentration of the TF in question)
    - RNA polymerase binds the complex; unwinds the helix and transcribes the strand → creates pre-RNA
    - RNA processing: add poly-A tail and 5’ cap, splice out introns and retain exons...this forms the mature mRNA
    - mRNA is then translated into a protein sequence -- ATG triplet (methionine) is what starts the transcription process
- Gene Structure
    - 5’ untranslated region (UTR)

    - then introns & exons -- usu 8 exons per gene (some have >100); usu 170 nucleotides per exon and 5,500 nucleotides per intron
    - 3' UTR
- Protein Structures
    - There are 20 amino acids (there are few non-standard amino acids)
    - Amino acids are the basic building block of proteins
    - Backbone + side chain (R group); R group lends the chemical property to the amino acid
- Translation
    - Ribosome synthesizes a protein by reading the mRNA in triplets (codons)
    - Each codon is translated to an amino acid
    - Ribosome sits on mRNA, translates each triplet of nucleotides into an amino acid
    - There is some redundancy here -- 3rd base wobble → eg. GUX (where X is A, U, C, G) is the same thing
    - happens with the aid of tRNA - have a cross shape
    - tRNA is attached to amino acid corresponding to a specific triplet of nucleotides
- Errors
    - What happens if something goes wrong?
    - Missense mutation -- single nucleotide change causes it to code for a different protein
    - Nonsense mutation -- change a codon into a STOP codon
    - Frameshift mutation -- delete or add a known multiple of 3 nucleotides -- reading frame shifts and garbles entire sequence
- Gene Expression Regulation
    - When should each gene be expressed?
    - Recall TFs discussed above -- transcription factor specificity
        - master regulators versus other types of TFs
    - One form of gene expression regulation is duplication of genes
- Chromatin
    - A lot of secondary epigenetic code exists on top of chromatin
    - eg. marks on histones

Sequence Alignment -- how do we algorithmically compare sequences and assess their similarities?

- Over 100 complete genomes have been sequenced
- Evolution occurs at the DNA level: sequence edits, rearrangements
- Changes that happen to an important part of the DNA tend to reduce fitness → hence important functional sequences also tend to be more conserved
- Alignment is key to
    - finding important regions