CS 4491/CS 7990
SPECIAL TOPICS IN BIOINFORMATICS

* Some contents are adapted from Dr. Jean Gao at UT Arlington

Mingon Kang, PhD
Computer Science, Kennesaw State University
Genetics

- **The discovery of DNA** (deoxyribonucleic acid)
  - In the early 1950s, Rosalind Franklin and Maurice Wilkins, studied DNA using **X-Rays**.
  - James Watson and Francis Crick studied the 3D structure of DNA.

X-ray diffraction image of DNA
(Omikron/Science Source)
The human genome has approximately 3 billion base pairs of DNA.

- Represented by a single letter: A (adenine), C (cytosine), G (guanine), and T (thymine).
- Arranged into 23 pairs of chromosomes in a cell.
The **world's largest collaborative biological project** mainly funded ($2.7 billion in 1991 dollars) by National Institutes of Health (NIH)

- **Identify and Map** the nucleotides contained in a **human haploid reference genome**

- Officially launched in 1990 and completed in 2003
  - **10 years** to sequence whole human genomes
DNA Sequencing Cost

- Sequencing cost was **below $1,000 in 2015**

Ref. https://www.genome.gov/sequencingcosts/
Human Genetic Disorders

- About 99% of DNA is same between humans and chimpanzees.
- About 99.9% of human genome (3 billions) are same
- Only 0.1% (3 hundred million) genetic variations makes the uniqueness of each individual.
Disease caused by a change, or mutation, in an individual’s DNA sequence

Many human diseases have genetic components.
- Autism
- Parkinson's disease
- Cancer (Colon\textsuperscript{2}, Breast\textsuperscript{3}, Prostate\textsuperscript{5}, Skin)

*: the *th leading cause of death in America
Core Research Question?

- What DNA variations are implicated in a human disease?
  - What are Genetic Biomarkers of human diseases?

- What are Genetic Mechanisms of a human disease?
  - Consist of complex interactions of multiple biological processes
  - Genetic, Epigenetic, and Transcriptional regulations
Biological Systems in Genetics

- **The Central Dogma** of molecular biology
  - **Transcription** and **Translation**
  - DNA ➔ RNA ➔ Protein

- **DNA**: Encodes hereditary information
- **RNA**: Intermediary between DNA and protein
- **Protein**: determines cell activities
image from the DOE Human Genome Program
http://www.ornl.gov/hgmis
DNA Genetic Code Dictates Amino Acid Identity and Order

DNA Sequence Is the Genetic Code.

Growing Protein Chain

image from the DOE Human Genome Program
http://www.ornl.gov/hgmis
DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code

**Gene A from Person 1**

<table>
<thead>
<tr>
<th>GCA</th>
<th>AGA</th>
<th>GAT</th>
<th>AAT</th>
<th>TGT...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Arg</td>
<td>Asp</td>
<td>Asn</td>
<td>Cys...</td>
</tr>
</tbody>
</table>

**Gene A from Person 2**

Codon change made no difference in amino acid sequence

<table>
<thead>
<tr>
<th>GCG</th>
<th>AGA</th>
<th>GAT</th>
<th>AAT</th>
<th>TGT...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Arg</td>
<td>Asp</td>
<td>Asn</td>
<td>Cys...</td>
</tr>
</tbody>
</table>

**Gene A from Person 3**

Codon change resulted in a different amino acid at position 2

<table>
<thead>
<tr>
<th>GCA</th>
<th>AAA</th>
<th>GAT</th>
<th>AAT</th>
<th>TGT...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Lys</td>
<td>Asp</td>
<td>Asn</td>
<td>Cys...</td>
</tr>
</tbody>
</table>

image from the DOE Human Genome Program
http://www.ornl.gov/hgmis
Bioinformatics Schematic of a Cell

- DNA
- Lipid membrane
- Proteins
- Nucleolus
- Nucleus
Chromosomes

- Chromatin: a fine network of threads composed of DNA in association with proteins.
- When a cell divides, the chromatin threads become highly condensed to form chromosomes.
- prokaryotes (single-celled organisms lacking nuclei) have a single circular chromosome
- eukaryotes (organisms with nuclei) have a species-specific number of linear chromosomes
DNA packs in the nucleus to form chromosome.
Eg: Human Chromosomes

Ref: https://www.guam.net/pub/ssh/s/depart/science/mancuso/apbiolecture/12_chromohg/humangen.htm
DNA

- can be thought of as the “blueprint” for an organism
- four different nucleotides distinguished by the four bases: adenine (A), cytosine (C), guanine (G) and thymine (T)
a single strand of DNA can be thought of as a string composed of the four letters: A, C, G, T

c tgt gg acc gg ggt gc ttag gac cct gac tgc c gc gg

gg cc gg ggg g g gtc ggg gc cgc gct gag
The Double Helix

- DNA molecules usually consist of two strands arranged in the famous Watson-Crick double helix
Watson-Crick Base Pairs

- in double-stranded DNA
  - A always bonds to T
  - C always bonds to G

- A-T, G-C: stabilized by hydrogen bonds. (2 between A-T pairing, 3 between G-C pairing.)
- Sequence = order of base pairs
The Double Helix

- each strand of DNA has a “direction”
  - at one end, the terminal carbon atom in the backbone is the 5’ carbon atom of the terminal sugar linked to a free phosphate group.
  - at the other end, the terminal carbon atom is the 3’ carbon atom of the terminal sugar attached to an unlinked hydroxy group.
- therefore we can talk about the 5’ and the 3’ ends of a DNA strand
- The direction from the 5’ end to the 3’ end is call FORWARD or POSITIVE.
- in a double helix, the strands are antiparallel (arrows drawn from the 5’ end to the 3’ end go in opposite directions)
RNA (Ribonucleic acid)

- RNA is like DNA except:
  - backbone is a little different (Not deoxy)
  - usually single stranded (folds up)
  - the base uracil (U) is used in place of thymine (T)
- a strand of RNA can be thought of as a string composed of the four letters: A, C, G, U
Genes

- Gene: a segment of DNA sequence corresponding to a single protein (or to a single catalytic or structural RNA molecule for those genes that produce RNA but not protein).
- A gene is said to encode a protein
- A single molecule of DNA contains many genes.
- Genes are the basic units of heredity

- Let’s check UCSC Genome Browser
Genes

- **Start codon**: genes begins with certain combinations of three DNA bases
  - In Eukaryotes, most start codon is ATG
  - In Prokaryotes, ATG (83%), GTG (14%), TTG (3%) and possibly ATT and CTG

- **Stop codon**: ends with TAG, TAA, and TGA.

- **Exon/Intron**: parts of genes
  - **Exons**: DNA converted into mRNA, coding regions
  - **Intron**: DNAs do not directly code for protein, non-coding regions.
    - Introns are only found in eukaryotic cells.
    - Prokaryotes have only exons
Genes

- **Codon**: a three-base sequence in mRNA that specifies one amino acid.
- Each codon is complimentary to a three-base code word in DNA.
Codons and Reading Frames
DNA encodes protein

- DNA uses an alphabet of 4 letters (ATCG), mostly commonly called bases.
- Long sequences of these 4 letters are linked together to create GENES and CONTROL INFORMATION.
- Each of the twenty protein amino acids can be specified by 3 consecutive DNA bases, which is a triplet code word.
- 61 of the 64 possible triplets are used to specify 20 amino acids.
- This means a given AA is usually specified by more than one code word.
- The 3 triplets that don’t specify amino acids are known as termination code words. (ACT, ATT, ATC)
Genomes

- Genome: a total collection of genetic information in an organism that specifies a particular organism’s (species’) structure and function.
- Can be thought of most directly as the base sequences of all the DNA molecules in a cell.
- The human genome consists of 46 chromosomes, containing between 50,000 and 100,000 genes. (old: 1995) (New discovery: less than 30,000 genes (year 2004))
- Every cell (except sex cells and mature red blood cells) contains the complete genome of an organism
Gene Density

- not all of the DNA in a genome encodes protein:

  microbes   90% coding   kb/gene
  human     3% coding    35kb/gene
The Central Dogma

The mechanism for expressing genetic information in all living organisms.

transcription

DNA → RNA → Protein

translation
Transcription

- Transcription: the genetic message is passed from DNA to mRNA in the nucleus.
- RNA that is transcribed from a gene is called messenger RNA (mRNA)
  - we’ll talk about other varieties of RNA later in the course
- RNA polymerase is the enzyme that builds an RNA strand from a gene
Transcription

**DNA**

```
ATGCGGTTAGACCCGTTAGCGGACCTGAC
TACGGCAATCTGGCAATCGCCTGGACTG
```

**mRNA synthesis**

```
AUUCGCGUUAGACCGUUAGCGGACCUCUGAC
```

**Top strand coding strand sense strand**

**Bottom strand template strand antisense strand**
RNA Splicing in eukaryotes

Gene structure:
- Chromosomal DNA
- Intron 1
- Exon 1
- Exon 2
- Exon 3
- Intron 2

Transcription:
- RNA synthesis

Nuclear RNA:
- Exon 1
- Exon 2
- Exon 3

RNA Splicing:
- Exon 1
- Exon 2
- Exon 3

Messenger RNA
Transcription

Question: Which of the two DNA strands is used as template for RNA synthesis?

Answer: This is determined by a specific sequence of nucleotides in DNA, called promoter, located at the beginning of each gene.

Promoter is present in only one of the two DNA strands.

Genes are transcribed only when RNA polymerase can bind to their promoter sites.
Translation

- Process that RNAs makes Proteins
- **Ribosomes** are the machines in the cytoplasm that synthesize proteins from mRNA
- The grouping of codons is called the reading frame
- Translation begins with the start codon
- Translation ends with the stop codon
Translation short summary

Protein synthesis

*Ribosome* assembles protein from AAs attached to *tRNAs* (transfer), according to message in mRNA.
Transfer RNA (tRNA)

- tRNA is synthesized in the nucleus by base-pairing with DNA nucleotides at specific tRNA genes, and then moves to the cytoplasm.

- Smallest of the three types of RNA (about 80 nucleotides long).

- Key role: can combine with both a specific amino acid and a codon in ribosome-bound mRNA specific for that AA.
Translation

Ribosome

Growing poly-peptide

This process repeats until reaching a stop codon
The 64 mappings of 3 bases to 1 amino acid is called the **GENETIC CODE** and is universal (on earth...).

**Ex:**

```
Gene A

One strand of DNA

Triplet code words
Met  Phe  Gly  Ala

Amino acids

Gene B

Termination code words

Protein B
Met  Ser  Val  Gly

Gene C

T-A-C  G-T-A  T-C-A  A-C-C  A-T-C

Protein C
Met  His  Ser  Trp
```
DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code

Gene A from Person 1:
- DNA sequence: GCA AGA GAT AAT TGT...
- Protein products: Ala Arg Asp Asn Cys...

Gene A from Person 2:
- DNA sequence: GCG AGA GAT AAT TGT...
- Codon change made no difference in amino acid sequence:
  - Protein products: Ala Arg Asp Asn Cys...

Gene A from Person 3:
- DNA sequence: GCA AAA GAT AAT TGT...
- Codon change resulted in a different amino acid at position 2:
  - Protein products: Ala Lys Asp Asn Cys...

Image from the DOE Human Genome Program
http://www.ornl.gov/hgmis
The Central Dogma

- https://youtu.be/J3HVVi2k2No
Protein Functions

- Structural support
- Storage of amino acids
- Transport of other substances
- Coordination of an organism’s activities
- Response of cell to chemical stimuli
- Movement
- Protection against disease
- Selective acceleration of chemical reactions
Proteins

- The subunit of protein structure is an amino acid.
- Cells build their proteins from 20 different amino acids.
- A peptide is a compound consisting of 2 or more amino acids.
- Polypeptides and proteins are chains of 10 or more amino acids, but peptides consisting of more than 50 amino acids are classified as proteins.
- Proteins account for about 50% of the organic material in the body.
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Abbreviation</th>
<th>One Letter Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>
If 3 RNA bases code for 1 AA, RNA could code for 64 AA. More than enough coding capacity for 20 AA.

Code is redundant for most AA.

Peptide Bond

- **Peptide bond**: the bond formed between the amino and carboxyl groups.

- **Polypeptide**: a sequence of amino acids linked by peptide bonds.

- **Peptide**: If the number of amino acids in a polypeptide is 50 or less, the molecule is known as peptide.

- **Protein**: If the sequence is more than 50 amino acid units long, the polypeptide is known as a protein.
Four levels of Protein Structure

Reference: http://www.slideshare.net/kaizettedetaza/central-dogma-12027441
Mutation

Reference: http://www.slideshare.net/kaizette/detaza/central-dogma-12027441
SNP

- Single-Nucleotide Polymorphisms (SNP)
  - DNA Variation occurring commonly within a population (e.g., 1%) in a single nucleotide.

There are two alleles, A and G, at this locus.
SNP

- Examine DNA variations at prespecified loci (about 10 millions)
- Bi-allele (major/minor allele)
  - Assumed that Minor allele may cause a specific rare phenotype (e.g., diseases) while Major allele is mainly observed in a population.
- Data Encoding

| SNP1 | SNP2 | SNP3 | ...
|------|------|------|-----
| G    | A    | A    | C   |
| G    | G    | A    | A   |
| G    | A    | C    | C   |
| G    | A    | C    | T   |

| SNP1 | SNP2 | SNP3 | ...
|------|------|------|-----
| 1    | 1    | 1    | 1   |
| 0    | 0    | 0    | 0   |
| 1    | 2    | 1    | ... |
GWAS

- **Genome-Wide Association Study**
  - Focus on associations between **SNPs** and a **phenotype** (e.g., cancer)
  - Conventionally, statistically compare SNPs of two groups (case and control) based on a linear regression model.

- **Pairwise univariate analysis**
  - **Perform** t-test between each SNP and a phenotype individually

\[
y = x_i \beta_i + \varepsilon
\]
eQTL

- expression Quantitative Trait Loci (eQTL) mapping study
  - Association between SNPs and mRNA
  - Find genomic loci that regulate expression levels of mRNAs (gene expression)
  - Capture the insight of the genetic architecture of gene expression.

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
<td>....</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td>....</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>....</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y</th>
<th>Gene1</th>
<th>Gene2</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.25</td>
<td>2.22</td>
<td>....</td>
<td></td>
</tr>
<tr>
<td>3.10</td>
<td>2.32</td>
<td>....</td>
<td></td>
</tr>
<tr>
<td>1.11</td>
<td>2.42</td>
<td>....</td>
<td></td>
</tr>
</tbody>
</table>