Introduction
http://users.rowan.edu/~polikar/classes/ece504

Instructor: Dr. Robi Polikar  
Office & Phone: 214 Rowan, 256-5372 (voice-mail available)  
Office Hours: Open door policy – You are welcome in anytime when my office door is open  
E-mail: polikar@rowan.edu  
Class Meeting: Tuesdays & Thursdays 15:15 – 16:45 at Rowan 202  

Reference Texts:  
2. Introduction to Machine Learning and Bioinformatics, Mitra et al., CRC, 2008  
5. Algorithms in Bioinformatics, Sung, CRC, 2010  


Reference for Web Lectures: http://videolectures.net
Introduction: What is this class about?

Part I - Fundamentals

- The code of life: DNA, RNA, genes, proteins and all that; Central dogma of molecular biology

Part II - Processes, Tools & Technologies

- Bioinformatics databases; Sequencing; Sequence alignment; Gene expression; Phylogenetics

Part III - Computational Intelligence for Bioinformatics

- Bayesian analysis; Hidden Markov models; Neural networks & support vector machines

Contemporary Issues

- The Human Genome Project; Personalized medicine; Creating Synthetic life

Photo / diagram credits

Courtesy: National Human Genome Research Institute

M. Zvelebil & J.O. Baum, Understanding Bioinformatics, Garland Sci., © 2008

Robi Polikar – All Rights Reserved © 2011.
SEMESTER READING ASSIGNMENT

How Craig Venter Tried to Capture the Code of Life and Save the World

THE GENOME WAR

James Shreeve

THE COMMON THREAD

A Story of Science, Politics, Ethics, and the Human Genome

John Sulston, Georgina Ferry

Bioinformatics, © 2011 Robi Polikar, Rowan University
What is Bioinformatics?

- NIH / National Human Genome Research Institute
  - The study of collecting, sorting, and analyzing of DNA and protein sequence information using computers and statistical techniques.

- NCBI (National Center for Biotechnology Information):
  - Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

- Wikipedia:
  - Application of computer science and information technology to the field of biology and medicine. Bioinformatics deals with algorithms, databases, artificial intelligence and soft computing, information and computation theory, software engineering, data mining, image processing, modeling and simulation, signal processing, discrete mathematics, control and system theory, circuit theory, and statistics, for generating new knowledge of biology and medicine, and improving & discovering new models of computation.

**Origin of Bioinformatics**

- Dramatic increase in the amount biological – but specifically genomic – information thanks in part due to advances in computational and molecular biology coupled with those in bioinstrumentation.
  - Need computational approaches to store, organize, access, analyze, and visualize data

- Bioinformatics started with cataloguing the vast amount of data being created with such applications as genome sequencing, the process of obtaining the genetic code of an organism
  - With growing data, came the need to align the sequences, finding patterns (motifs) in these sequences, grouping the sequences of different species with respect to their similarities, predicting the functions of proteins encoded by a given gene, computing distances between the genes, creating phylogenetic trees (family trees, so to speak) of organisms, etc.
We all do, because of immense and vastly expanding applications of bioinformatics:

- **Molecular & personalized medicine**
  - Custom designed treatment based on knowing to what drugs you will respond
  - Knowing your risk factors for all diseases that have a genetic component

- **Gene therapy**
  - Treat genetic disorders caused by mutations

- **Drug development**
  - Design drugs that specifically target misfolded proteins

- **Forensic analysis**

- **Genetically modified food**
  - Crops that are resistant to disease, pesticides, drought, etc.
  - Modify / improve nutritional content of food
WHAT IS IN THIS COURSE?

The course consists of three primary parts:

- Fundamentals – including basic background on genetics and bioinformatics
- Processes, Tools & Technologies – used in bioinformatics research
- Computational Intelligence – the underlying algorithms and approaches that make some of the processes and tools possible.

Generally follow these parts in order, but we often will move around the individual topics.

We will also discuss contemporary and controversial issues, such as the Human Genome Project (should the genes / genome be patentable?), creation of synthetic life (ethical issues? should we do it?) and personalized medicine.
There are three set of molecules that are essential for all known forms of life.

- Two forms of nucleic acids (DNA and RNA) and the proteins they encode. Together, these three sets of molecules are the raw material of life.

**Deoxyribonucleic acid (DNA):** The molecule that stores and carries genetic instructions of all living things. All information necessary to make and maintain an organism is stored in DNA.

- Double stranded molecule that wind around each other ➔ double helix
- Each strand has a sequence of four bases attached

**Ribonucleic acid (RNA):** Similar to DNA, but in a single strand, carries genetic instructions from DNA to the ribosome, where proteins are synthesized.

**Proteins** are the molecules that actually direct and control all life processes.

- They facilitate and catalyze all biochemical reactions as enzymes
- They form all muscles (structural proteins)
- They participate in immune system as antibodies
- They control all body functions as hormones (regulatory proteins): e.g. insulin
- They are the carriers of molecules:, e.g., hemoglobin for transporting oxygen
The entire information necessary to make and maintain an organism is stored in a few DNA molecules, collectively called the **genome**.

- In humans, there are 23 pairs of these molecules, each of which is called a **chromosome**, and contains a very tightly packaged single DNA molecule, wound around proteins called **histones**.
- The packing allows such a large molecule, 6 feet in length if unwound, to fit inside the nucleus!

**Where are they?**

- In each and every one of your cell (except blood cells) is a nucleus, in which lies the chromosomes, which consists of an exact replica of your genome!
Note that, the exact replica of this DNA lies in each and every one of your cells!
DNA is a linear polymer, made of nucleotides, which consists of three chemical parts:

- A phosphate group, a sugar group (deoxyribose) and one of four nitrogen bases (the nucleic group); hence the name deoxyribonucleic acid).
- The phosphate and the sugar group makes up the backbone or the support structure of the DNA.
- The information content of the DNA lies in this long sequence of the nitrogen bases: adenosine (A), guanine (G), cytosine (C) and thymine (T).
- In RNA, thymine is replaced by uracil (U).
Note the numbers on the sugar molecule:

- The carbons of the sugar molecule are numbered 1 through 5. This is important: the bases always attach at location 1 (called 1').
- The phosphate backbone attach to the preceding sugar at location 3 and the next sugar at location 5.
- This leaves one location 3 and one location 5 free at the ends of the chain, which are referred to as 3' end and 5' end.

The base sequence, or the nucleotide sequence, or simply the DNA sequence is – by convention – read from the 5’ end to the 3’ end.
DNA actually consists of two strands of such long linear molecules of repeating nucleotides, with each strand running in *opposite directions* and forming a helical structure, perhaps the best known molecular structure, *the double helix*.

The two strands form base-pairings that follow a specific rule (*Watson-Crick base pairing*): \( A \leftrightarrow T \) and \( C \leftrightarrow G \), forming *complementary base sequences*.

The two strands run in opposite (anti-parallel) directions, so the complementary sequence of \( \text{CATTGCCAGT} \) is not \( \text{GTAACCGTCA} \), but rather \( \text{ACTGGCAATG} \)
James Watson and Francis Crick are widely credited with the discovery of the double helix structure in 1953.

It is also rumored that they stormed into Eagle pub in Cambridge, U.K., and announced “We have found the secret of life.”

They share the 1962 Nobel Prize in Physiology or Medicine, along with Maurice Wilkins.
The controversy behind this however is that another scientist, Rosalind Franklin, had taken an X-ray diffraction image (Photograph 51) of the DNA molecule that showed an X (hinting the double helix structure). She then knew that the DNA had to have a helical structure.

However, Wilkins had shown the photo to Watson, based on which they announced the double helix structure before she could.

Franklin died in 1958. Since Nobel Prize is not given posthumously, she was left out of the Nobel Prize.
The RNA, the second nucleic acid, while similar to DNA in many respects, have three important differences:

1. RNA is single stranded, as a result of which RNA has a much more flexible structure that can take many and complex helical, hinge, and folding shapes. DNA, on the other hand, has a rigid double helix structure;
2. Complementary base to adenine (A) is not thymine (T) – as in DNA – but rather uracil (U); and
3. The sugar in RNA is ribose, as opposed the deoxyribose in DNA (hence the name)

While DNA is the primary storage of genetic information, RNA is the carrier of this information, so that it can be used to make (synthesize / encode) proteins.

There are several types of RNA involved, and the process goes as follows:

1. The genetic information from the DNA is transcribed to messenger RNA (mRNA) to be carried to ribosomes, the protein building factories of the cell...
2. where transfer RNA (tRNA) molecules are used to deliver the amino acids (the building blocks of proteins) ...
3. and the ribosomal RNA (rRNA) links the amino acids to form the proteins, a process called translation.
This process of transcription of DNA to RNA and then translation of protein is called the central dogma of molecular biology (credited to F. Crick).
The central dogma of molecular biology simply states that there is only single direction of transfer of information: DNA to RNA to protein:

- The DNA stores the genetic information, which is then *transcribed* to the RNA to carry this information to be *translated* into proteins, which are the work horse of all living organisms.
- This also means that the information cannot move in any other direction: for example a DNA cannot be synthesized from a protein, and hence the process of producing proteins cannot be reversed.
- DNA, however, can (of course) self-replicate, the essence of reproduction.
A gene is simply a specific subsequence of the DNA. Genes are the basic physical and functional units of heredity, and carry specific information to make proteins.

- In humans, there are 23 pairs of chromosomes that each pack a total of about 3 billion base pairs of DNA (slightly different in males and females). Within these 3 billion consecutive letters of A, T, G and C, are about 23000 genes that encode proteins.

- About 1.5% of the human genome code proteins, the rest are non-coding DNA, regulatory sequences, introns, etc., previously – and incorrectly – called junk DNA.

Of the 23 pairs of human chromosomes, 22 pairs are called autosomes, common to both males and females. The 23rd pair is the sex chromosomes, which differ between males and females.

Females have two copies of the X chromosome, whereas males have an X and a Y chromosome.
Note that different chromosomes include different number of genes:

- Chromosome 1 has the most number of genes, and hence is the longest chromosome; about 8% of the total human genome.
- There are some 890 known diseases related to Chromosome 1, the most of any chromosome, including AD, breast cancer, deafness and glaucoma.
- Also note that the Y chromosome is noticeably shorter, and hence with fewer genes. Therefore, the male genome is in fact shorter than the female genome. In fact, the number of genes in human Y chromosome is now estimated to be between 70 – 200. [http://ghr.nlm.nih.gov/chromosome/Y](http://ghr.nlm.nih.gov/chromosome/Y) • Takes more genetic material to make a female!

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Genes</th>
<th>Total base pairs</th>
<th>Sequenced bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4,220</td>
<td>247,199,719</td>
<td>224,599,719</td>
</tr>
<tr>
<td>2</td>
<td>1,491</td>
<td>242,751,149</td>
<td>237,712,649</td>
</tr>
<tr>
<td>3</td>
<td>1,550</td>
<td>199,446,827</td>
<td>194,704,827</td>
</tr>
<tr>
<td>4</td>
<td>446</td>
<td>191,263,063</td>
<td>187,297,063</td>
</tr>
<tr>
<td>5</td>
<td>609</td>
<td>180,837,866</td>
<td>177,702,766</td>
</tr>
<tr>
<td>6</td>
<td>2,281</td>
<td>170,696,993</td>
<td>167,273,993</td>
</tr>
<tr>
<td>7</td>
<td>2,135</td>
<td>168,821,424</td>
<td>154,952,424</td>
</tr>
<tr>
<td>8</td>
<td>1,106</td>
<td>146,274,826</td>
<td>142,612,826</td>
</tr>
<tr>
<td>9</td>
<td>1,920</td>
<td>140,442,298</td>
<td>120,312,298</td>
</tr>
<tr>
<td>10</td>
<td>1,793</td>
<td>135,374,737</td>
<td>131,624,737</td>
</tr>
<tr>
<td>11</td>
<td>379</td>
<td>134,452,384</td>
<td>131,130,653</td>
</tr>
<tr>
<td>12</td>
<td>1,430</td>
<td>132,289,534</td>
<td>130,303,534</td>
</tr>
<tr>
<td>13</td>
<td>924</td>
<td>114,127,980</td>
<td>95,559,980</td>
</tr>
<tr>
<td>14</td>
<td>1,347</td>
<td>106,360,585</td>
<td>88,290,585</td>
</tr>
<tr>
<td>15</td>
<td>921</td>
<td>100,339,915</td>
<td>81,341,915</td>
</tr>
<tr>
<td>16</td>
<td>909</td>
<td>88,822,264</td>
<td>78,884,764</td>
</tr>
<tr>
<td>17</td>
<td>1,672</td>
<td>78,654,742</td>
<td>77,800,220</td>
</tr>
<tr>
<td>18</td>
<td>519</td>
<td>76,117,153</td>
<td>74,656,155</td>
</tr>
<tr>
<td>19</td>
<td>1,556</td>
<td>63,606,651</td>
<td>55,785,651</td>
</tr>
<tr>
<td>20</td>
<td>1,008</td>
<td>62,435,965</td>
<td>59,606,264</td>
</tr>
<tr>
<td>21</td>
<td>578</td>
<td>46,944,323</td>
<td>34,171,998</td>
</tr>
<tr>
<td>22</td>
<td>1,092</td>
<td>49,528,953</td>
<td>34,893,953</td>
</tr>
<tr>
<td>X (sex chromosome)</td>
<td>1,846</td>
<td>154,913,754</td>
<td>151,058,754</td>
</tr>
<tr>
<td>Y (sex chromosome)</td>
<td>454</td>
<td>57,741,652</td>
<td>25,121,652</td>
</tr>
<tr>
<td>Total</td>
<td>32,185</td>
<td>3,079,843,747</td>
<td>2,857,598,560</td>
</tr>
</tbody>
</table>
The so-called coding region of a gene that contains the code for producing proteins is called an **exon**. Each exon codes for a specific portion of the complete protein.

In some species, including humans, the exons are separated by long regions of DNA, called **introns**. The function of the introns are unknown, which resulted these regions to be also called junk DNA.
# The Human Genome vs. Others

<table>
<thead>
<tr>
<th>Organism type</th>
<th>Organism</th>
<th>Genome size (base pairs)</th>
<th>Genome size (in human-readable format)</th>
<th>mass in pg</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virus</td>
<td>Bacteriophage MS2</td>
<td>3,569</td>
<td>3.5kb</td>
<td>0.000002</td>
<td>First sequenced RNA-genome[^8]</td>
</tr>
<tr>
<td>Virus</td>
<td>SV40</td>
<td>5,224</td>
<td>5.2kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virus</td>
<td>Phage d-X174</td>
<td>5,386</td>
<td>5.3kb</td>
<td></td>
<td>First sequenced DNA-genome[^10]</td>
</tr>
<tr>
<td>Virus</td>
<td>Phage λ</td>
<td>48,502</td>
<td>48kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virus</td>
<td>Mimivirus</td>
<td>1,161,404</td>
<td>1.1Mb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterium</td>
<td><em>Haemophilus influenzae</em></td>
<td>1,830,000</td>
<td>1.8Mb</td>
<td></td>
<td>First genome of a living organism sequenced, July 1996[^12]</td>
</tr>
<tr>
<td>Bacterium</td>
<td>Carsonella ruddii</td>
<td>159,662</td>
<td>159kb</td>
<td></td>
<td>Smallest non-viral genome[^13]</td>
</tr>
<tr>
<td>Bacterium</td>
<td>Buchnera aphidicola</td>
<td>600,000</td>
<td>600kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterium</td>
<td>Wigglesworthia glossinidica</td>
<td>700,000</td>
<td>700kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterium</td>
<td>Escherichia coli</td>
<td>4,660,000</td>
<td>4.6Mb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterium</td>
<td><em>Solfacter usitatus</em> (strain Ellin 5076)</td>
<td>9,970,000</td>
<td>9.9Mb</td>
<td></td>
<td>Largest known Bacterial genome</td>
</tr>
<tr>
<td>Amoeboid</td>
<td>Polythecos dubium (<em>Amoeba</em> dubia)</td>
<td>670,000,000,000</td>
<td>670Gb (Disputed[^15])</td>
<td>737</td>
<td>Largest known genome[^16] (Disputed[^17])</td>
</tr>
<tr>
<td>Plant</td>
<td>Arabidopsis thaliana</td>
<td>157,000,000</td>
<td>157Mb</td>
<td></td>
<td>First plant genome sequenced, December 2000[^18]</td>
</tr>
<tr>
<td>Plant</td>
<td>Geraniaceae margaretae</td>
<td>63,400,000</td>
<td>63Mb</td>
<td></td>
<td>Smallest recorded flowering plant genome, 2006[^19]</td>
</tr>
<tr>
<td>Plant</td>
<td><em>Fruit</em> <em>assynca</em></td>
<td>130,000,000,000</td>
<td>130Gb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plant</td>
<td>Populus trichocarpa</td>
<td>490,000,000</td>
<td>490Mb</td>
<td></td>
<td>First tree genome sequenced, September 2006</td>
</tr>
<tr>
<td>Plant</td>
<td><em>Pleiis japonica</em> (Japanese-native, pale-petal)</td>
<td>150,000,000,000</td>
<td>150Gb</td>
<td>152.23</td>
<td>Largest plant genome known</td>
</tr>
<tr>
<td>Moss</td>
<td>Physcomitrella patens</td>
<td>480,000,000</td>
<td>480Mb</td>
<td></td>
<td>First genome of a bryophyte sequenced, January 2008[^19]</td>
</tr>
<tr>
<td>Yeast</td>
<td>Saccharomyces cerevisiae</td>
<td>12,100,000</td>
<td>12.1Mb</td>
<td></td>
<td>First eukaryotic genome sequenced, 1996[^20]</td>
</tr>
<tr>
<td>Fungus</td>
<td>Aspergillus nidulans</td>
<td>30,000,000</td>
<td>30Mb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nematode</td>
<td>Caenorhabditis elegans</td>
<td>100,300,000</td>
<td>100Mb</td>
<td></td>
<td>First multicellular animal genome sequenced, December 1998[^21]</td>
</tr>
<tr>
<td>Nematode</td>
<td>Pratylenchus coffeae</td>
<td>20,000,000</td>
<td>20Mb</td>
<td></td>
<td>Smallest animal genome known[^22]</td>
</tr>
<tr>
<td>Insect</td>
<td>Drosophila melanogaster (fruit fly)</td>
<td>130,000,000,000</td>
<td>130Mb</td>
<td></td>
<td>[^23]</td>
</tr>
<tr>
<td>Insect</td>
<td>Bombyx mori (silkworm)</td>
<td>530,000,000</td>
<td>530Mb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insect</td>
<td>Apis mellifera (honey bee)</td>
<td>236,000,000</td>
<td>236Mb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insect</td>
<td>Solenopsis invicta (fire ant)</td>
<td>480,000,000</td>
<td>480Mb</td>
<td></td>
<td>[^24]</td>
</tr>
<tr>
<td>Fish</td>
<td>Tetraodon nigroviridis (type of puffer fish)</td>
<td>385,000,000</td>
<td>3.9Mb</td>
<td></td>
<td>Smallest vertebrate genome known</td>
</tr>
<tr>
<td>Mammal</td>
<td>Homo sapiens</td>
<td>3,200,000,000</td>
<td>3.2Gb</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Fish</td>
<td>Protcophorus aethiopicus (marbled lungfish)</td>
<td>130,000,000,000</td>
<td>130Gb</td>
<td>143</td>
<td>Largest vertebrate genome known</td>
</tr>
</tbody>
</table>

[^8]: [First sequenced RNA-genome](http://en.wikipedia.org/wiki/Genome) | [^9]: |
[^10]: [First sequenced DNA-genome](http://en.wikipedia.org/wiki/Genome) | [^11]: |
[^12]: [First genome of a living organism sequenced, July 1996](http://en.wikipedia.org/wiki/Genome) | [^13]: |
[^14]: [Smallest non-viral genome](http://en.wikipedia.org/wiki/Genome) | [^15]: |
[^16]: [Largest known genome](http://en.wikipedia.org/wiki/Genome) | [^17]: |
[^18]: [First plant genome sequenced, December 2000](http://en.wikipedia.org/wiki/Genome) | [^19]: |
[^20]: [First eukaryotic genome sequenced, 1996](http://en.wikipedia.org/wiki/Genome) | [^21]: |
[^22]: [First multicellular animal genome sequenced, December 1998](http://en.wikipedia.org/wiki/Genome) | [^23]: |
[^24]: [Smallest vertebrate genome known](http://en.wikipedia.org/wiki/Genome) | [^25]: |
Transcription is the process of copying a protein-coding gene from DNA to mRNA. mRNA is much shorter since it contains the information to make a single protein, whereas the DNA includes the information to make all proteins.

Non-coding exons are then removed by a process called RNA splicing.

mRNA then directs the process to synthesize the appropriate protein, a process called translation. The translation itself is carried out by the ribosome.

A gene transcribed into RNA to be used in protein synthesis is said to be expressed.

While only one segment of DNA is transcribed for a given gene, genes may overlap, so the same location may encode parts of different proteins.
The information in the genes tell the ribosomes how to make proteins. However, the main building blocks of proteins are amino acids.

- There are just 20 amino acids, each of which is encoded by a different three letter combination of A, G, C and U, called a **codon**.
- Since there are 4 letters, one can have $4^3=64$ different combinations of three letter words, i.e., codons, whereas there are only 20 amino acids.
- Hence there is some redundancy, a built-in error-correction mechanism.
- Of these 64 codons, three - **UAA**, **UGA** and **UAG** – are used as **stop codons** that tell the ribosomes when to stop translating, whereas one (in general) - **AUG** is used as a start code*.
  - Unlike the stop codons, the start of translation requires more than just a start codon (usually AUG); certain control signals are also involved.

*The rules for associating the base sequences in the DNA / RNA to the amino acid sequence of a protein is called the **genetic code**.
Depending on which base we start, there are three possible ways to translate a given nucleotide sequence into a protein sequence. These are called reading frames.

- Given two strands of DNA, there are in fact six ways to encode proteins.
- The control signals (determined by nearby sequences and other initiation factors) along with the start codon, determine the actual reading frame to be used by the ribosome.
- In predicting protein coding sequencing, we do not have the luxury of knowing these control signals, and hence all six scenarios need to be analyzed, only one of which will actually encode a functional protein.
TRANSCRIPTION & TRANSLATION
The redundancy in the genetic code is called **degeneracy**.

- Multiple codons may encode the same amino acid, e.g., leucine (Leu) is specified by UUA, UUG, CUU, CUC, CUA, CUG codons; despite the redundancy, there is no ambiguity however: UUA will always encode Leu.

The degeneracy provides some built-in error correction:

- Consider Glycine (Gly): GGA, GGG, GGC, GGU. Any mutation in the third place makes no difference, as they all encode the same amino acid, Gly.

- Consider Glutamic Acid (Glu): GAA and GAG. The third position in this case is a two-fold degenerate site, as a mutation of A → G or G → A results in the same amino acid. Other mutations, however, cause a change in the amino acid, which then results the appropriate protein not being synthesized.
A variation in a DNA segment that includes only a single nucleotide is called a *single nucleotide polymorphism* (SNP).

The (often only) two variations of gene caused by a SNP are referred to two *alleles* of that gene.

- If both alleles produce the same peptide sequence (resulting in the same protein), we have a synonymous polymorphism – also called a silent mutation.
- Otherwise, we have a replacement polymorphism, which either results in a different amino acid sequence (called *missense*) or an interrupted sequence due to a premature stop codon (called *nonsense*).

In bioinformatics research, SNPs are commonly analyzed for:

- Comparing subsequences of different cohorts (e.g., people with a particular disease vs. healthy controls)
- Personalized medicine, in determining whether a given SNP makes a person more or less responsive to a particular treatment
- There are several databases just for SNPs; such as NCBI's *dbSNP* database.
### Notable Mutations

**Examples of notable Mutations**

<table>
<thead>
<tr>
<th>1st base</th>
<th>2nd base</th>
<th>3rd base in each row</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>UUU (Phe/F) Phenylalanine</td>
<td>UCU (Ser/S) Serine</td>
<td>UAU (Tyr/Y) Tyrosine</td>
</tr>
<tr>
<td>UUC (Phe/F) Phenylalanine</td>
<td>UCC (Ser/S) Serine</td>
<td>UAC (Tyr/Y) Tyrosine</td>
</tr>
<tr>
<td>UUA (Leu/L) Leucine</td>
<td>UCA (Ser/S) Serine</td>
<td>UAA Ochre (Stop)</td>
</tr>
<tr>
<td>UUG (Leu/L) Leucine</td>
<td>UCG (Ser/S) Serine</td>
<td>UAG Amber (Stop)</td>
</tr>
<tr>
<td>CUU (Leu/L) Leucine</td>
<td>CCA (Pro/P) Proline</td>
<td>CAA (Gln/Q) Glutamine</td>
</tr>
<tr>
<td>CUC (Leu/L) Leucine</td>
<td>CCC (Pro/P) Proline</td>
<td>CAC (His/H) Histidine</td>
</tr>
<tr>
<td>CUA (Leu/L) Leucine</td>
<td>CUG (Leu/L) Leucine</td>
<td>CAU (His/H) Histidine</td>
</tr>
<tr>
<td>AUU (Ile/I) Isoleucine</td>
<td>ACU (Thr/T) Threonine</td>
<td>AAU (Asn/N) Asparagine</td>
</tr>
<tr>
<td>AUC (Ile/I) Isoleucine</td>
<td>ACC (Thr/T) Threonine</td>
<td>AAC (Asn/N) Asparagine</td>
</tr>
<tr>
<td>AUA (Ile/I) Isoleucine</td>
<td>ACA (Thr/T) Threonine</td>
<td>AAA (Lys/K) Lysine</td>
</tr>
<tr>
<td>AUG (Met/M) Methionine</td>
<td>ACG (Thr/T) Threonine</td>
<td>AAG (Lys/K) Lysine</td>
</tr>
<tr>
<td>GGU (Val/V) Valine</td>
<td>GCC (Ala/A) Alanine</td>
<td>GAA (Glu/E) Glutamic acid</td>
</tr>
<tr>
<td>GUC (Val/V) Valine</td>
<td>GCA (Ala/A) Alanine</td>
<td>GAG (Glu/E) Glutamic acid</td>
</tr>
<tr>
<td>GUA (Val/V) Valine</td>
<td>GCG (Ala/A) Alanine</td>
<td>GGG (Gly/G) Glycine</td>
</tr>
<tr>
<td>GUG (Val/V) Valine</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Selection of notable mutations**: ordered in a standard table of the genetic code of amino acids.

Clinically important missense mutations generally change the properties of the coded amino acid residue between being basic, acidic, polar or nonpolar, while nonsense mutations result in a stop codon.

**Amino acids**
- Basic
- Acidic
- Polar
- Nonpolar (hydrophobic)

**Polyglutamine (PolyQ) Diseases**
- Huntington’s disease
- Spinocerebellar ataxia (SCA) (most types)
- Spinobulbar muscular atrophy (Kennedy disease)
- Dentatorubral-pallidoluysian atrophy

**Mutation type**
- Trinucleotide repeat
- Deletion
- Missense
- Nonsense

So far we talked about how proteins are synthesized, what can go wrong, etc. But what is a protein, and why is it so important?

By definition, proteins are large compounds of polypeptides, folded into globular (circular) or fibrous (linear) forms.

- The role of a protein is to facilitate a biological function.
- The specific folding of a protein is closely related to its function. In order to fully understand the biological function of a protein, it is therefore necessary to determine its structure.
- The structure of a protein is usually obtained via X-ray crystallography (90%) or NMR spectroscopy (~10%).

Protein structures can be visualized in several different ways, which can be obtained by such protein visualization software as PyMol (download at [http://www.pymol.org/](http://www.pymol.org/))

If nucleic acid is the raw material for life, the protein is the stuff from which life is made and maintained! They do it all:

- Regulate all gene activity
- Provide cellular / muscular structure
- Serve as enzymes, regulating and controlling all biochemical activities
- Serve as hormones in cellular signaling
- Serve as antibodies in the immune systems
- Play a role in inter / intra cellular transportation of ions

In fact, they are so important that, we have special names:

- The total content of proteins present in a cell or cell type is called its proteome,
- The study of such large-scale data sets that include proteomes is the field of proteomics.

In bioinformatics, a particular problem is predicting the protein structure (so called tertiary or quaternary structure) from its sequence (also called primary structure).

- If the sequence and function of Protein\textsubscript{A} are known, it is reasonable that the function of unknown Protein\textsubscript{B} will be similar to that of Protein\textsubscript{A}, if they have similar sequences (i.e., they are homologous).
- This similarity can be quantified using machine learning algorithms.
While the integrity of genetic information in DNA is protected and the number of errors in replication is small, it is nonzero. Some errors cause no apparent change, whereas other errors (mutations), allow the species to change, and over long periods of time, evolve into new species.

It is now widely accepted that all life has evolved from a single, common and evolutionarily speaking, a very distant ancestor.

The study of evolutionary relationship of species with each is called phylogenetics, and is often represented using a tree of life structure, called the phylogenetic tree.

Based on the analysis of molecular sequences, there are three primary groups, also called domains or superkingdoms:

- **Bacteria** – Single cell organisms with no nucleus. They have a single ribosomal RNA.
- **Archaea** – Single cell organisms with no nucleus, whose genes and enzymes often behave like those of eukaryotes. They have three ribosomal RNAs and are often extremophiles/
- **Eukaryotes** – Usually multi-cell organisms, whose cells include a nucleus, as well as other complex organelles.

Bacteria and Archaea are collectively referred to as prokaryotes.
Note that eukaryotes, which include all complex organisms – including all animals and plant – constitute only a small part of the tree of life.

- It is estimated that there are ~ 40 million bacterial cells in a gram of soil and a million bacterial cells in a milliliter of fresh water.
- There are approximately five nonillion \((5 \times 10^{30})\) bacteria on Earth\(^1\), forming a biomass on Earth, which exceeds that of all plants and animals\(^2\).

---

http://en.wikipedia.org/wiki/Bacteria

Phylogenetics: evolutionary relationships between organisms

- Requires genetic data which is only completely available for living organisms, as fossil records include only partial information.
- The primary method for classifying organisms and hence generating phylogenetic trees is **cladistics**
  - A clade is a single branch on the tree of life, representing one ancestor and all of its descendants – and nothing else.
  - The tree generates using this approach is called a **cladogram**.
- Statistical machine learning algorithms, such as maximum likelihood, Bayesian inference, and Markov chains are commonly used for predicting phylogenetic trees.
Phylogenetic tree of life of species whose genomes have been sequenced.

This particular figure is a **cladogram**, whose branch lengths do not represent time or relative amount of character change.
Taxonomy: classification, organization and naming of organisms.

Biological classification of life today recognizes eight major taxonomical ranks:

- The lowest rank is the species – it is understood that the genomes of different individuals of given species differ very little from those of each other.
- For example, humans are thought to be genetically 99.9% identical, regardless of their race.
  - Note that at 3 billion nucleotides, this means that, on average, humans may have 300,000 nucleotides that differ from each other!
- As we go up in the ranks, the genomes of the individual members in that taxa become more and more divergent.

Machine learning algorithms also play an important role in biological classification:

- Given a partially sequenced genome of a previously unknown species, what rank does the species belong to?
- How similar are the genomes of different species within a given taxa?
NUCLEIC ACID WORLD
What Is Next?

- We will look at the following topics:

- What computational problems can be addressed by bioinformatics?
- What tools & databases are there for analyzing genetic data?
- What technologies are there for sequencing genome and analyzing gene expression?