Computational Biology (MAT4930/MAT5932)

Lecture 1: Molecular Biology Primer

Abdulmelik Mohammed

University of South Florida

01/12/2021

Biology at the microscopic level

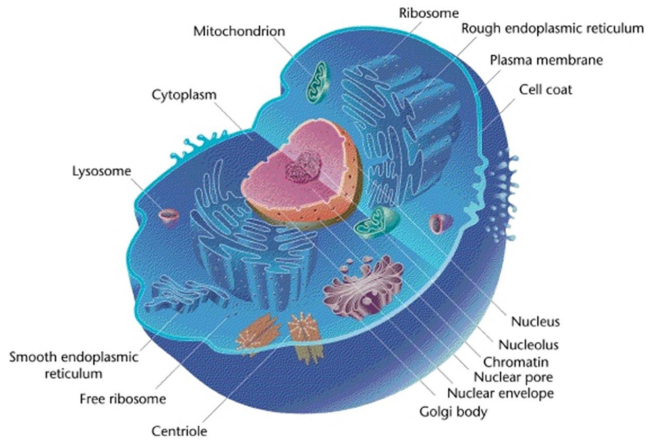

Figure: Eukaryotic cell. Image courtesy OpenStax CNX 2016.

- Cells have a life cycle: are born, eat, replicate and die.
- Cells constantly "compute": copy, translate, regulate, ...
- Much of the complexity of cells that is conserved across many organisms organized through three string-like molecules: DNA, RNA and proteins.

Traits and inheritance

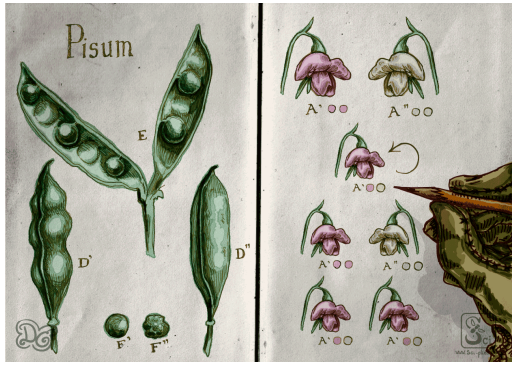

Figure: Mendel's experiments. Image courtesy Galagovsky 2021.

- Variety in numbers of chromosomes in different organisms suggests chromosomes carry information related to species.
- Mendel's work on garden pea plants showed dominant and recessive traits.
- Evidence for *genes* in chromosomes being trait carrying- and hereditary units given by Thomas Morgan in his fruit fly experiments.

Genes

- In 1941, Beadle and Tatum experimented with X-ray irradiated *Neurospora*, a bread mold, to conclude that genes code for proteins. Irradiated *Neurospora* did not grow on medium lacking vitamin B_6, showing that x-rays damaged a gene responsible for the production of a protein/enzyme that catalyzed the production of B_6.
- Genes are segments of *Deoxyribonucleic acid (DNA)*, a polymer composed of *(deoxyribo) nucleotides.* A nucleotide is a compound composed of a sugar, a phosphate group and one of four possible nitrogenous bases: adenine (A), cytosine (C), guanine (G), thymine (T).

Molecular composition of DNA

- One end of a (linear) single-stranded DNA has a free phosphate group and is called the 5′-*end*.
- The other end is the 3′-*end*.
- Hence, there is a 5′-to-3′ direction to a DNA strand.

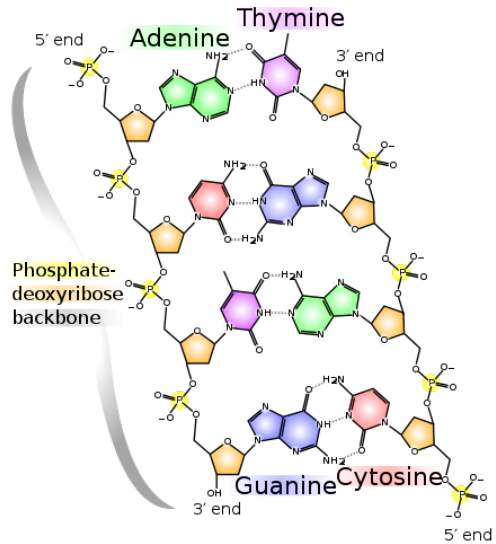

Figure: Components of dsDNA: phosphate group in yellow, sugar in orange and nitrogenous bases in other colors. Image by M. P. Ball 2016.

3D structure of DNA

- Watson and Crick showed the double helical 3D structure of DNA in 1953.
- The two strands run in opposite directions.
- Adenine binds thymine and cytosine binds guanine in double helical DNA. Thus A-T and C-G are said to be *Watson-Crick complements*.

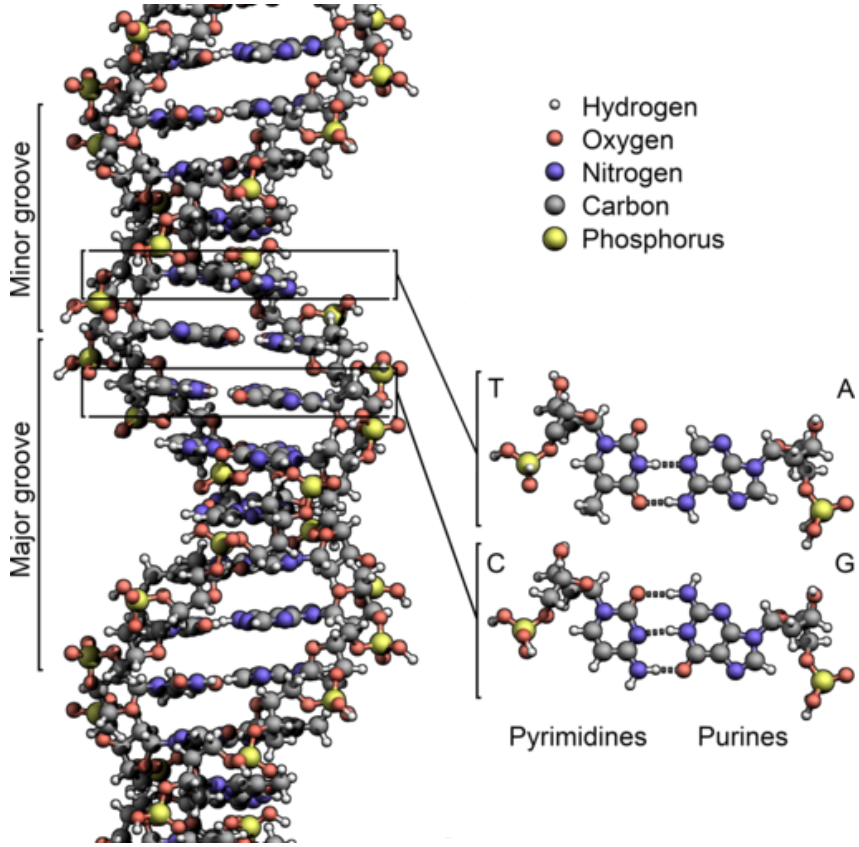

Figure: Right-handed double helical DNA. Image by Zephryis 2011.

Structure of DNA, continued

- The information in a DNA strand is completely described by the nitrogenous base sequence from the $5'$-end to the $3'$-end.
- Watson-Crick complementarity naturally extends to DNA sequences. For e.g., $5'$-ATGACC-$3'$ is complementary to $3'$-TACTGG-$5'$.
- Watson and Crick: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

Brief overview of DNA replication

- Two copies of DNA are obtained from a single copy through *DNA replication*.
- Carried out by an enzyme called *DNA polymerase* in the 5′-to-3′ direction.
- A number of other proteins (molecular machines) involved: helicase, topoisomerase, ligase, ...

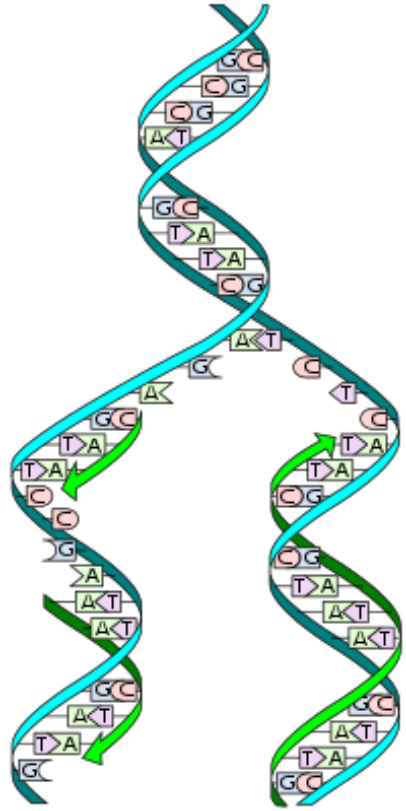

Figure: DNA replication sketch. Image by M. P. Ball, 2007

Proteins

- Proteins are large biomolecules with diverse set of functional roles in the cell, such as catalyzing reactions, DNA replication, structural functions and molecular transport.
- Proteins are biopolymers made up of long chain of *amino acids* that are connected through peptide bonds. Amino acids come in various types which differ in their residue group R.

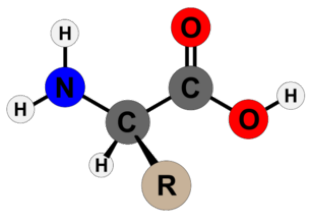

Figure: An amino acid. Image reprinted from Wikipedia, 2020.

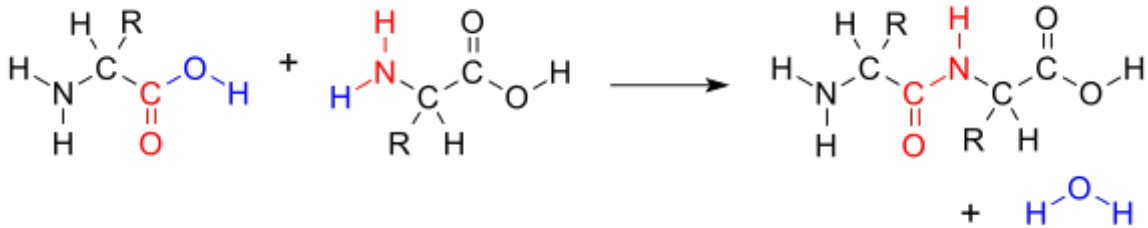

Figure: Peptide bond. Image reprinted from Wikipedia 2009.

There are 21 different amino acids in eukaryotes

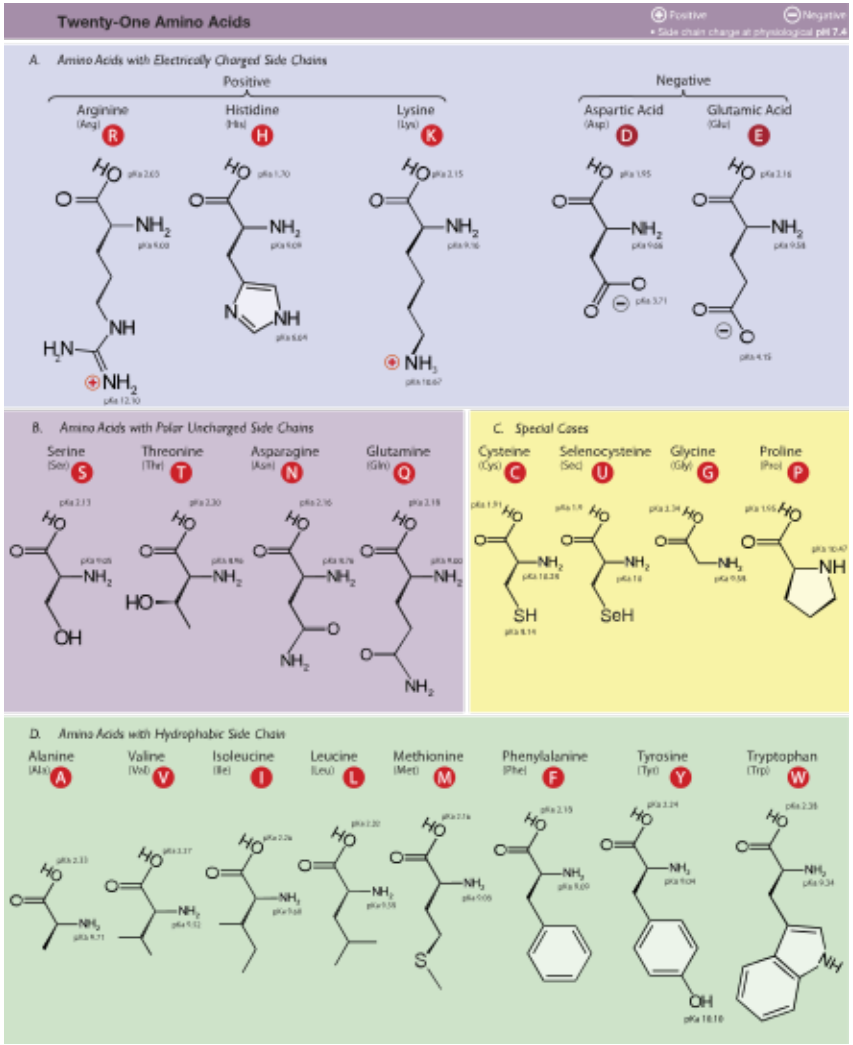

Figure: Image by D. Cojocari, 2010.

Levels of protein structure

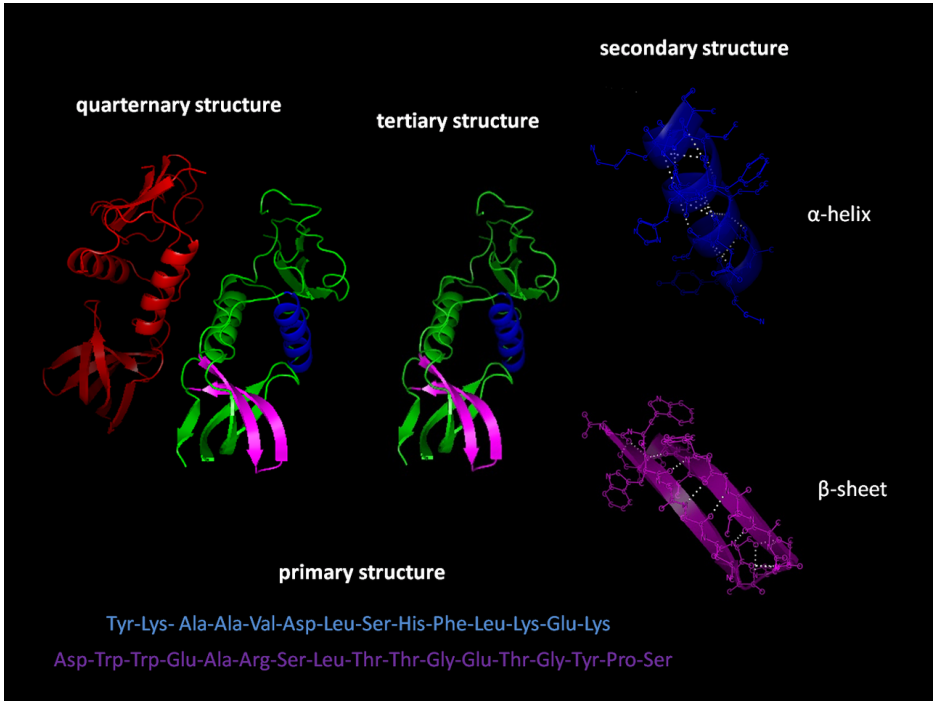

Figure: Image reprinted from Wikipedia 2012.

- Protein sequences are reported from amino terminal to carboxyl terminal by convention.
- The 3D structure of a protein determines its function.

RNA

- *Ribonucleic acid (RNA)* is a nucleic acid polymer like DNA. Its sugar is a ribose sugar (includes an hydroxyl group at the 2nd carbon).
- RNA is composed of the same A, C, G bases, but thymine is replaced with its unmethylated version uracil (U).
- RNA is typically found in single-stranded form, which usually folds into 3D structure through intra-strand base pairing.

Messenger RNA and the central dogma of molecular biology

- *Messenger RNA (mRNA)* is responsible for the *translation* of DNA sequences in genes into amino-acid sequences in proteins.

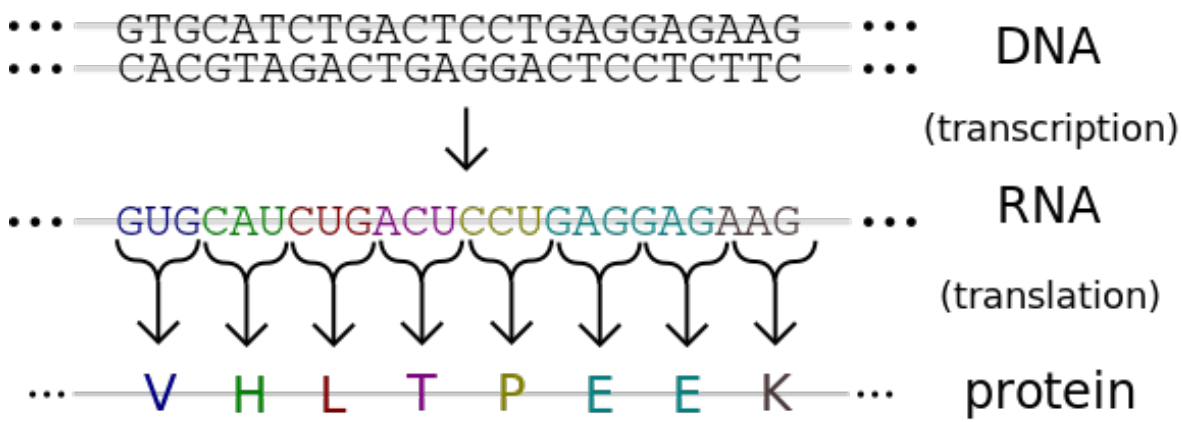

Figure: From DNA to proteins. Image by M. P. Ball 2013.

The genetic code

- Each triplet of bases of DNA called a *codon* corresponds to a specific amino acid.
- There are $4^3 = 64$ possible codons, but only 21 amino acids.
- Different triples may code for the same amino acid. For instance, both TTT and TTC code for phenylalanine.
- The full translation table, i.e. *genetic code*, is available in Table 3.1 of the textbook.

Transcription

- A protein complex called *RNA polymerase* is responsible for *transcribing* a gene to its messenger RNA sequence. The transcription rules are: A-U, C-G, G-C and T-A.

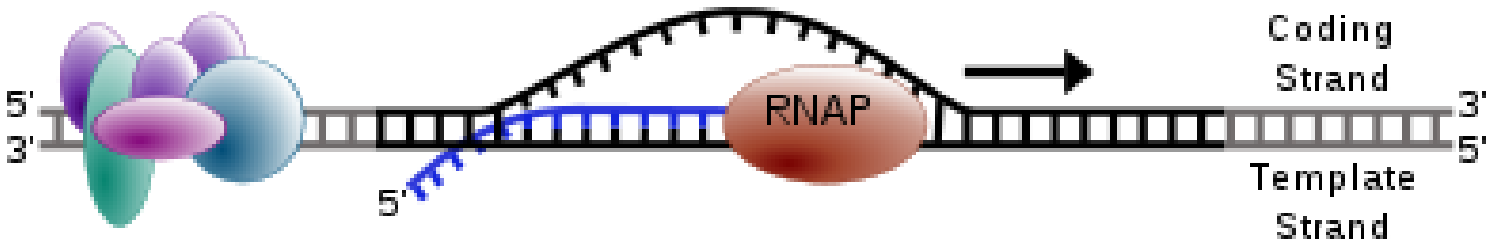

Figure: Schematic of transcription. Reprinted from Wikipedia, 2007.

In eukaryotes, a gene can code for multiple proteins

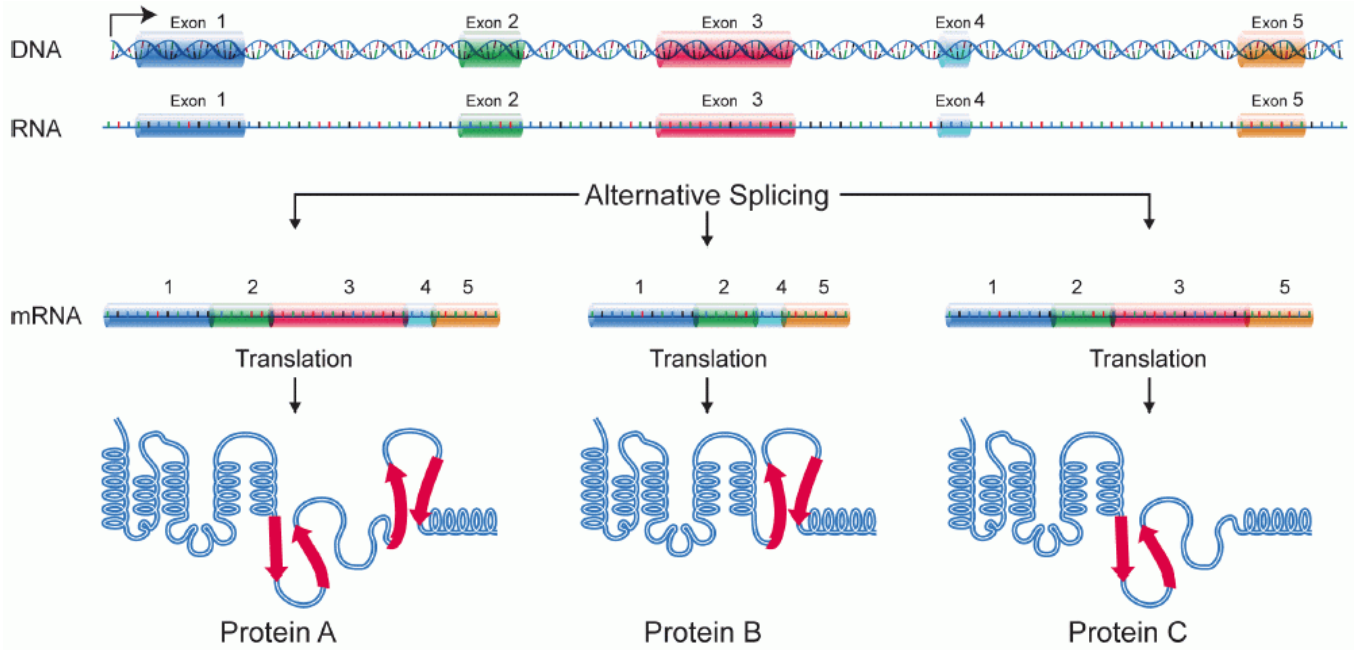

Figure: Regions called *introns* are spliced out from a pre-mRNA. Different selection of *exons* in a gene result in different proteins. Image by NHGRI, 2007.

Translation: From mRNA to proteins

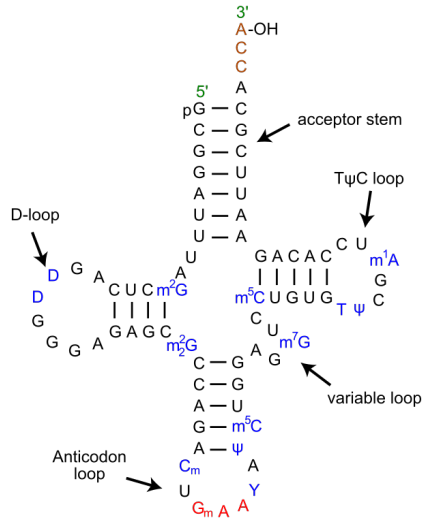

Figure: Transfer RNA (tRNA) of a yeast. The amino acid is attached to the 3′ tail downstream the acceptor stem. During translation, the tRNA attaches to the mRNA at the *anti-codon* loop. Image by Yikrazuul, 2010.

- mRNA is translated into an amino-acid sequence in the *ribosome* with the aid of transfer RNA.
- For each codon in mRNA, there is a tRNA that carries the corresponding amino acid to the ribosome.

Translation in the ribosonme

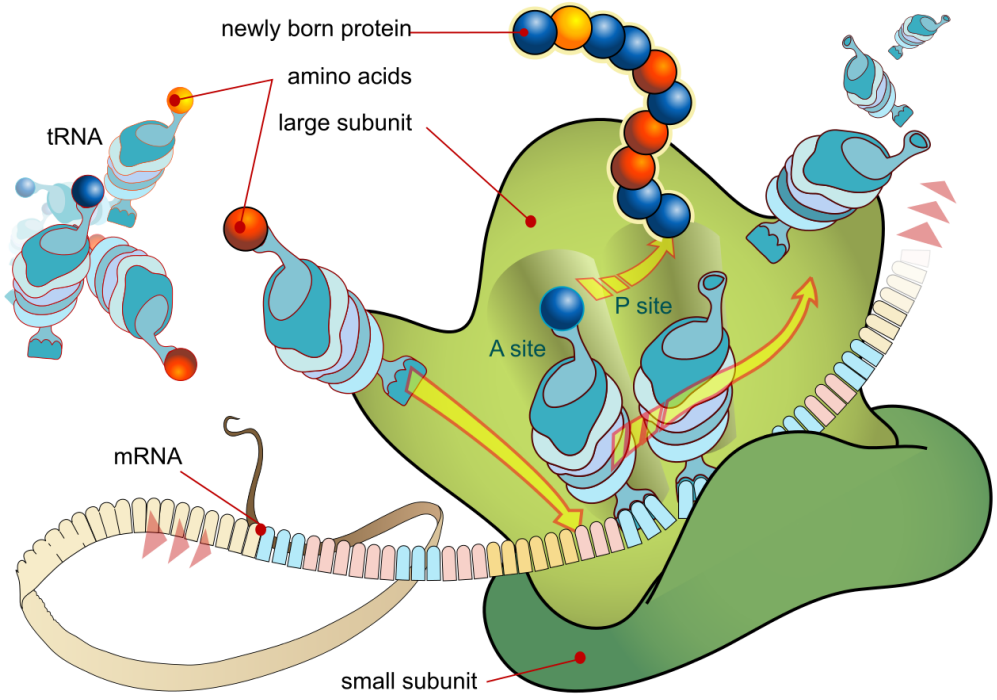

Figure: A schematic of translation in the ribosome. Image by M. R. Villarreal 2008.

- The ribosome itself is composed of *ribosomal RNA (rRNA)* and ribosomal proteins!

The genome and genetic diversity

- The *genome* of an individual consists of all the genetic material in the chromosomes of the individual.
- The haploid human genome consists of roughly 3 billion base pairs in 23 chromosomes.
- Only approximately 20,000 protein-coding genes have been found in the genome. Less than 2 percent of the genome codes for a protein. The study of non-coding DNA is an ongoing effort.
- Variations in genes leads to distinct traits in individuals such as hair color. Roughly 0.1% of the human genome varies between individuals.
- Up to 99% of human genes are conserved across all mammals. This is possible because of evolution from common ancestry.

Bioinformatics

- *Sequencing*: are sets of related technologies for reading out the DNA sequence of a target genome.
- *Bioinformatics* uses computational and mathematical techniques to decipher the information in genomes, transcriptomes, proteomes.
- For e.g. we can study the similarities and differences between the genomes of two individuals or species through alignment algorithms.

Manipulating DNA

- *Polymerase chain reaction (PCR)*: amplifying DNA, similar to DNA replication, but as a chain reaction to gain exponential copies.
- *Synthesis*: producing short DNA strands of a desired sequence.
- *Gel electrophoresis*: filtering or selecting DNA by size.
- Other enzymatic actions: cutting by restriction enzymes, ligation to join broken segments.
- *Probes*: use DNA hybridization to detect existence of DNA fragments.