COMP 5970/6970-004
Computational Biology: Genomics and Transcriptomics
Lecture notes 18: 3/22/2022

Haynes Heaton

Spring, 2022

Lecture Objectives

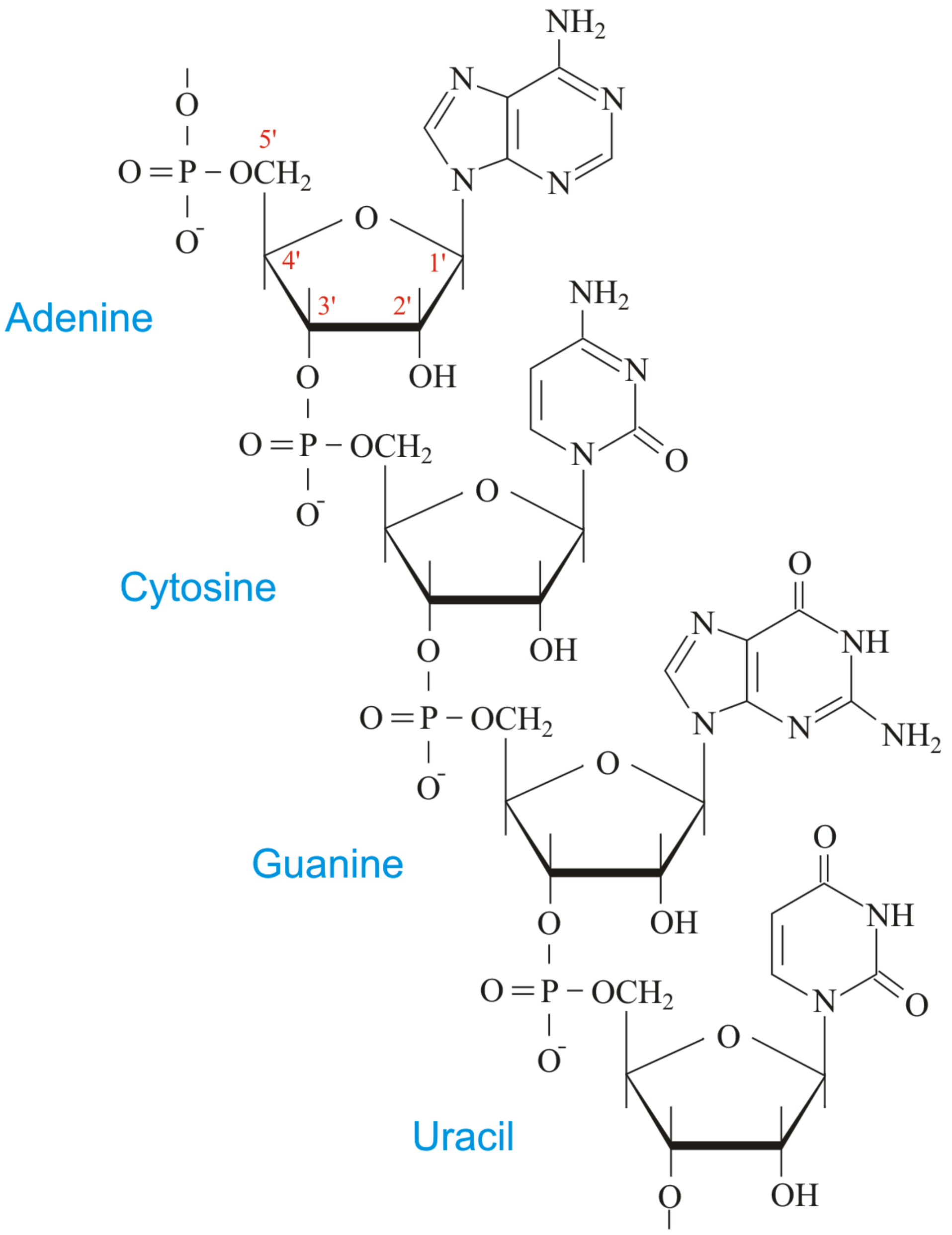
- Evolution - what exists either persists or replicates
- Storing data biologically for evolution
- Stable and redundant biological data storage and replication
- Central dogma of molecular biology

1 Evolution as observation bias

What exists either persists or replicates. And what replicates must exist long enough to replicate. Biological systems strive for both to different degrees. The cell maintains this low entropy state with its cell membrane and has developed many different mechanisms for self-repair. And more complex multicellular organisms have developed systems to fight off external attacks (the immune system) and heal injuries. In terms of replication, very simple systems might follow chemical signals to make more or less of each component and split into two when possible.

2 Storing data biologically for evolution

But eventually a blueprint is needed to map out what to do in different environments and for replication to be possible there, the blueprint must also be replicated. In computers, binary numbers are used to store information. In biology, **nucleotides** are used to store information. Initially, it is believed life used **ribonucleic acids**, or RNA to store information. RNA is a polymer of **ribonucleotides**. Each nucleotide consists of a purine (Adenine (A) or Guanine (G)) base or pyrimidine (Cytosine (C) or Uracil (U)) base and a ribose ring. These ribose rings can be bound to one another in series with a phosphate group creating a phosphate bond.

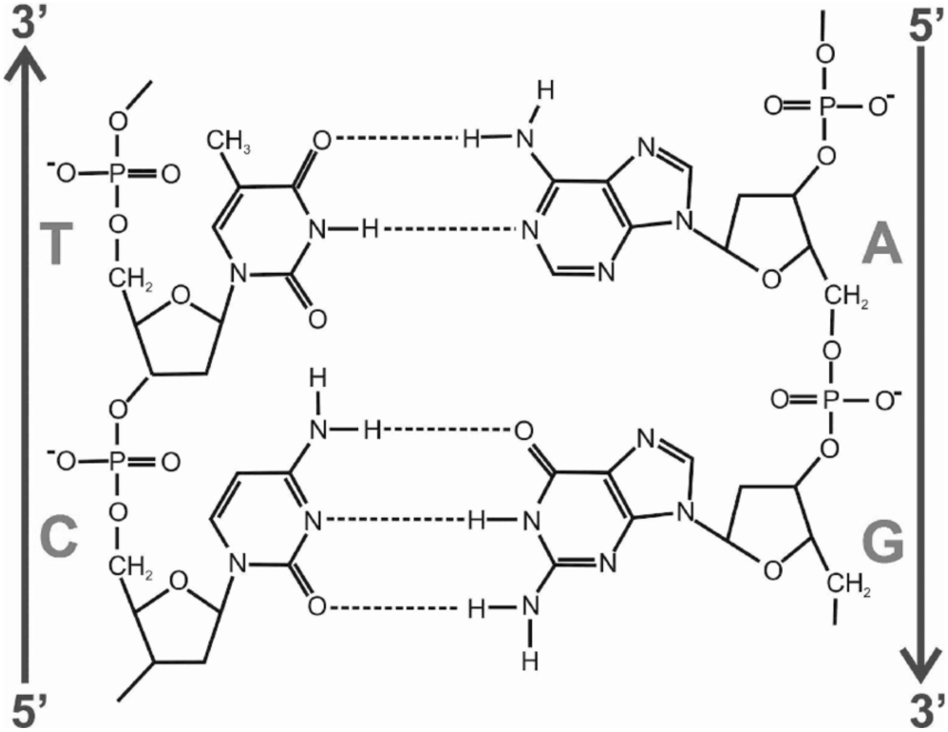


The carbon atoms in the ribose are numbered 1 to 5 as shown above with the 3rd and 5th carbon binding to either phosphate bond. With this we can describe a direction of the RNA sequence as either 5′ to 3′ or 3′ to 5′. This will be important later.

RNA is a very special molecule as it can act as information storage as well as fold into complex 3D shapes and carry out specific biochemical tasks similar to enzymes. These are called ribozymes. RNA is capable of enzymatically replicating other RNA. This is why it is believed that RNA was the first molecule used to transmit genetic information and that before DNA, life consisted of an RNA world.

3 Stable and redundant biological data storage and replication

RNA can form a double stranded molecule with its reverse complement molecule, but due to the hydroxyl group on the 2′ carbon of the ribose, the double stranded molecule is not stable and the strands will dissociate at moderate temperatures. Evolution eventually found that removing this hydroxyl group could produce a much more stable molecule, **deoxyribonucleic acid, or DNA**.



DNA is very similar to RNA, but the ribose is missing a hydroxyl group on the 2′ carbon giving less steric inhibition to the formation of the double stranded structure. And DNA has Thymine in place of Uracil which is simply a methylated version of Uracil. As you can see, Each purine can create hydrogen bonds with one pyrimidine. Adenines bind with Thymines and Guanines bind with Cytosines. Of note, Guanine-Cytosine bonds contain three hydrogen bonds whereas Adenine-Thymine bonds contain two. This makes the double strand structure stronger the more GC pairs in the region.

The angles of these bonds and other minor interactions create a helical structure which also strengthens the overall chemical stability. Because of the double stranded nature of this helical structure, it is termed a **double helix**.

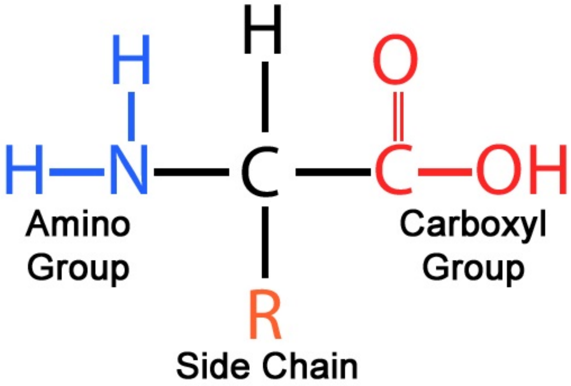
4 The central dogma of molecular biology

The central dogma of molecular biology states that in general, the flow of information in most organisms goes from DNA to RNA to protein. And proteins carry out the various functions of the cell. This idea is broken in a many ways, but is in general true.

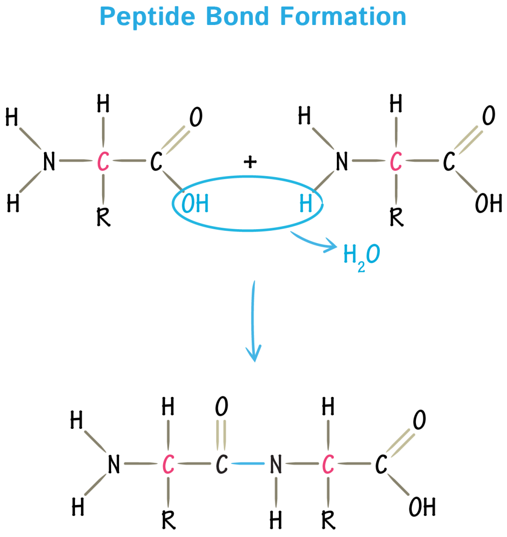
Some examples of how it is broken is that

But in general, messenger RNA (mRNA, yes, the same stuff in the covid vaccines) is **transcribed** from DNA and proteins, or polypeptides (explained further soon) are **translated** from mRNA. As a short aside, it was often stated that the vaccine was designed in two days. This was very surprising to many people, but not to people who know the technologies that exist. The secret sauce of the vaccine is actually some lipid surfactants that encapsulate the mRNA and then fuse with the cell membrane allowing the mRNA to enter the cell. Other than that, we just need the sequence of the spike protein and we have technologies for synthesizing a particular sequence of RNA or DNA up to some length. So really, a draft of the vaccine could have been made in, idk, maybe 15 minutes.

Okay back to the central dogma. The reason the words transcribed and translated are used is very simple. Because DNA and RNA share a four letter alphabet, there is a one to one correspondence of these letters. Three of the letters are identical. The only difference is that in RNA, Uracil take the place of Thymine in DNA. So this simple copying is termed transcription. But because there are 20 possible amino acids, a single DNA or RNA base cannot code for uniquely one amino acid. Nor can two DNA bases code for a unique amino acid as 4^2 is 16. It requires three bases to code for all of the amino acids uniquely. These three base segments which code for an amino acid are termed **codons**. But $4^3 = 64$, so there are more codons than there are amino acids. Now what is the structure of the amino acid.

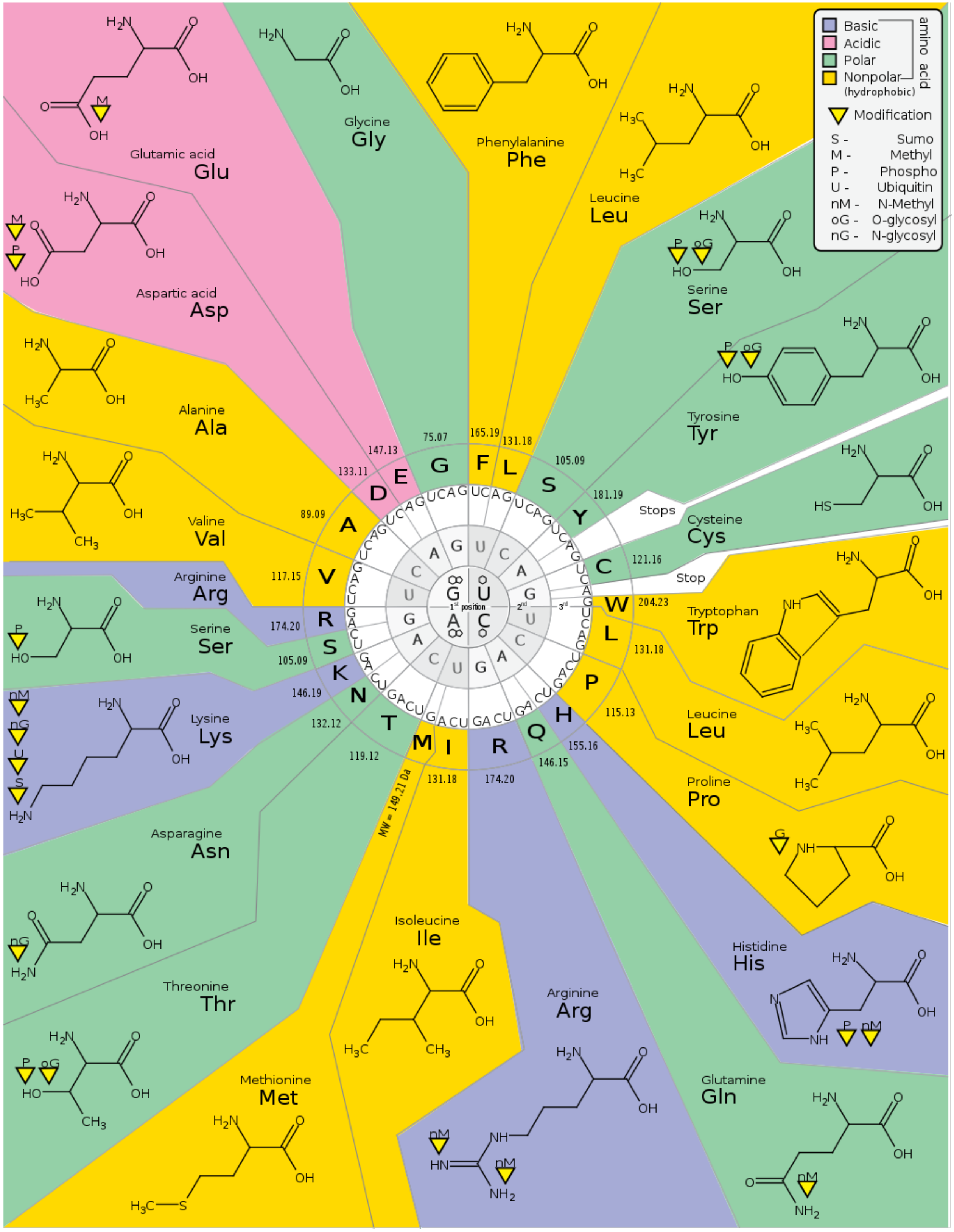


Proteins are chains of amino acids. These are alternatively call **polypeptides** because they are linked by peptide bonds. The amino group of one amino acid and the carboxyl group of another amino acid react to create a peptide bond and a water molecule.

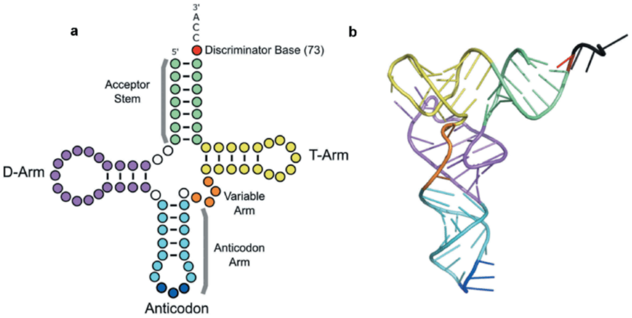


<div style="text-align:center">4</div>

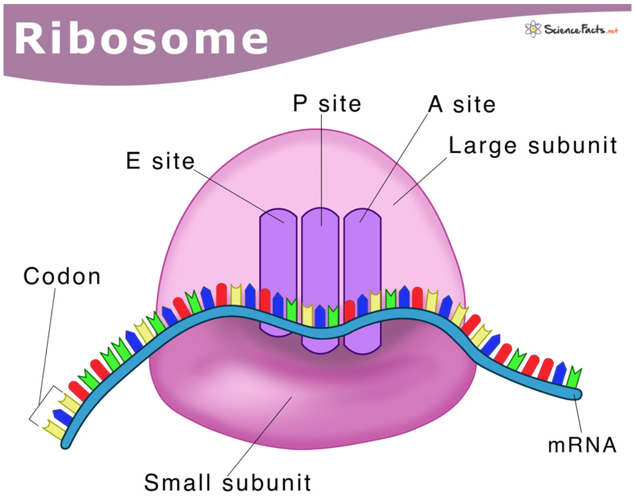
And what does this degenerate coding of codons for amino acids look like?



So how are polypeptides with a particular sequence of amino acids constructed from the messenger RNA information? Well, there are several additional types of RNA that help in this endeavor. First is the **transfer RNA or tRNA**. The tRNA is a piece of RNA that is folded on itself using the base pair binding like DNA does but in a way to create a particular shape in which one end binds to a particular amino acid and the other end has the complementary sequence of the codon that codes for that amino acid. And then there are two other arms which interact with our next type of RNA.



Ribosomes, or ribosomal RNA, or rRNA are functional macromolecules made from RNA and are similar to proteins in that they catalyze specific chemical reactions. Two different ribosomes together catalyze the translation of mRNA to protein.



The ribosome holds the mRNA and accepts tRNA with complementary tRNA anticodon sequences to the mRNA which are in **frame** (a frame is where the codon starts. the frame starts at the start codon of the mRNA and then proceeds in non overlapping 3 bases at a time).

5 Mutations

Because of the degeneracy of the codon coding system, different mutations can have different effects. If a mutation does not change the amino acid the gene codes for, this is called a **synonomous mutation**. These of course will not change the shape or function of the protein. If the mutation changes the amino acid sequence, but does not introduce a stop codon, this is termed a **missense or non-synonymous mutation**. These may or may not change the shape or function of the protein depending on what the change was and where in the protein it occurs. If the mutation puts an early stop codon in frame, this is called a **nonesense mutation** and almost always kills the function of the protein. In the case of insertions and deletions, if they are not a multiple of 3, this creates a **frame shift mutation** changing all amino acids downstream of the mutation and also gives a high probability of putting an early stop codon in frame thus killing the function of the protein.