Big Data, Genes, and Medicine
Video #1.5

Once the mRNA has been synthesized in the cell nucleus, it is transported into the cytoplasm and a ribosome for further processing. The next step in this process is called translation. Translation is the process of producing a protein from a messenger RNA (mRNA). Now I am going to explain in more detail what the proteome is about. Proteins are made of building blocks called aminoacids. Each aminoacid is coded by three nucleotides in the mRNA. The sequence of aminoacids composing a protein is synthesized by one by one and glued together as the translation progresses, in the order dictated by the mRNA strand. This process takes place in ribosomes, organelles located in the cytoplasm where the mRNA has been led after the transcription. A ribosome reads the mRNA strand in groups of three bases to assemble the protein.

This picture shows a ribosome translating a strand or mRNA and synthesizing new aminoacids. We see a transport RNA carrying a new aminoacid to be added to the growing chain of aminoacids composing the newly created protein.

There are 20 aminoacids in humans. A certain combination of 3 bases, also called a codon, always leads to the same aminoacid. There are 64 codons, out of which 61 code for aminoacids, plus 3 stop codons. Most aminoacids are therefore coded by several different codons. The wheel on the following slide can be used to find which one.

Start at the center with the first base, then move to the outside following the second then third base. For example AUG leads to aminoacid Methionine (Met).

What we see is that:

- DNA, RNA, and proteins can be transformed into one another through coding or decoding.
- Data representing DNA, RNA, or proteins can provide information about one another and can regulate one another.
- In fact, these molecular compounds are part of larger systems called gene regulatory networks where molecular regulators can interact with each other and with other substances in the cell to affect the transcription of genes and/or the creation of gene expressions.
- The regulators can be DNA, RNA, proteins, or other macromolecules.

In turn the change in protein composition in the cell can greatly affect its structure and behavior. There is a feedback mechanism such that existing proteins and mRNA signal the DNA to produce more or less mRNA and therefore proteins. The level and composition of proteins in a cell greatly affects its entire structure and functioning, impacting tissues, organs, systems, and in the end the whole human body. Therefore the analysis of the level of proteins and gene expressions in cells, as well as the integrity of the DNA are key to understanding the processes of human health or its disturbances. These types of data in other living organisms are also key because they affect human health, for example through viruses, bacteria, fungi, and nutrition.

This is taking advantage of Big Data methods and technologies in biomedicine is so important. We live in the era of Big Data biomedicine, as the following figures demonstrate:

- There are 3.2 billions bases in human genome, at least one copy in each cell of the human body (3.3 GB).
- There are 100,000 macromolecules in database of macromolecular structures.
- Sequence Read Archive (SRA) at NHI/NCBI stores over 3.6 petabases of raw sequence data (1 petabase, and 4 times more bytes) for 250,000 human individuals, 32,000 microbial genomes, and 5,000 animal and plant genomes.
- Electronic medical records (EMRs) store over 500 petabytes on a world scale (1 petabyte = 10^{15} bytes = 1 million GB).
- Predicted to grow very rapidly.

In the following modules of this course, you will learn how to explore and analyze such Big datasets.

I will end this lesson by showing a 3D animation of translation by the DNA Learning Center Institute.
https://www.dnalc.org/resources/3d/15-translation-basic.html