Introduction to Bioinformatics Lecture Notes

By: Dr. Hafeez Ur Rehman

Note: Majority of the contents are taken from Bioinformatics Computing by Bryan Bergeron.

Table of Contents

Introduction to Bioinformatics

Background:

Biology has traditionally been an observational rather than a deductive science. Although recent developments have not altered this basic orientation, the nature of the data has radically changed. It is arguable that until recently all biological observations were fundamentally anecdotal - admittedly with varying degrees of precision, some very high indeed. However, in the last generation the data have become not only much more quantitative and precise, but, in the case of nucleotide and amino acid sequences, they have become *discrete*. It is possible to determine the genome sequence of an individual organism or clone not only completely, but in principle *exactly*. Experimental error can never be avoided entirely, but for modern genomic sequencing it is extremely low. Not that this has converted biology into a deductive science. Life does obey principles of physics and chemistry, but for now life is too complex, and too dependent on historical contingency, for us to deduce its detailed properties from basic principles.

What is bioinformatics?

Definitions:

- Bioinformatics is an interdisciplinary scientific field that develops **methods** and **software tools** for **storing, retrieving, organizing** and **analyzing biological data**.
- "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."

Bioinformatics would not be possible without advances in computing hardware and software. Fast and high capacity storage media are essential even to maintain the archives. Information retrieval and analysis require programs; some fairly straightforward and others extremely sophisticated. Distribution of the information requires the facilities of computer networks and the World Wide Web. Computer science is a young and flourishing field with the goal of making most effective use of information technology hardware. Certain areas of theoretical computer science impinge most directly on bioinformatics.

Today, bioinformatics, like many sciences, deals with the storage, transport, and analysis of information. What distinguishes bioinformatics from other scientific endeavors is that it focuses on the information encoded in the genes and how this information affects the universe of biological processes. With this

in mind, consider how bioinformatics is reflected in the Central Dogma of molecular biology.

The Central Dogma

The Central Dogma of Molecular Biology, as originally defined by James Watson, is deceptively simple: DNA defines the synthesis of protein by way of an RNA intermediary. Documenting, controlling, and modifying this process, which is illustrated from a high-level structural perspective in Figure 1-5, is the focus of bioinformatics. It's also the basis for genetic engineering, mapping the human genome, and the diagnosis and treatment of genetic diseases. For example, genetic engineering involves modifying the process so that new proteins are synthesized; these new proteins in turn form the basis of everything from new drugs to new types of plants and animals.

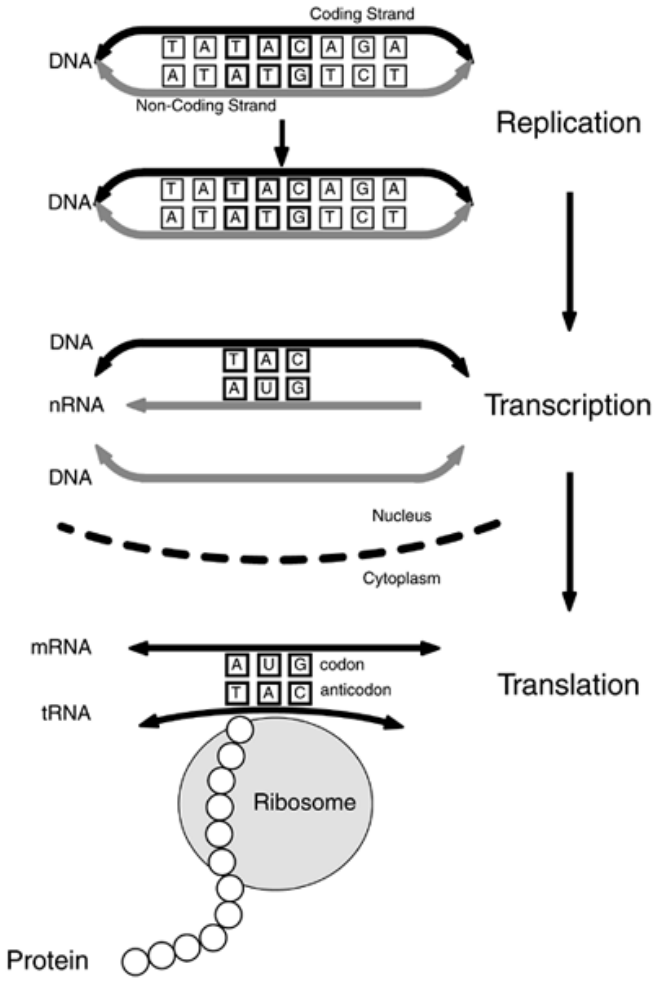

Figure 1. The Central Dogma of Molecular Biology. DNA is transcribed to messenger RNA in the cell nucleus, which is in turn translated to protein in the cytoplasm.

The simplified version of the Central Dogma, shown in Figure 1, in which DNA is duplicated through replication, transcribed to RNA, which is in turn translated to protein, only hints at the complexity of the information transfer process that is the driving force for bioinformatics. Consider that the archive of an individual's genetic information or genome is encoded in DNA as a sequence of four different nitrogenous bases on a sugar-phosphate backbone. This deoxyribonucleic acid can adopt a variety of conformations, including the infamous right-handed double helix first described by Watson and Crick in 1953. The sequence of four nitrogenous bases—some combination of Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)—in each strand of the double helix mirror each other in a predefined manner; Adenine on one strand always binds with Thymine on the other, and Cytosine always binds with Guanine.

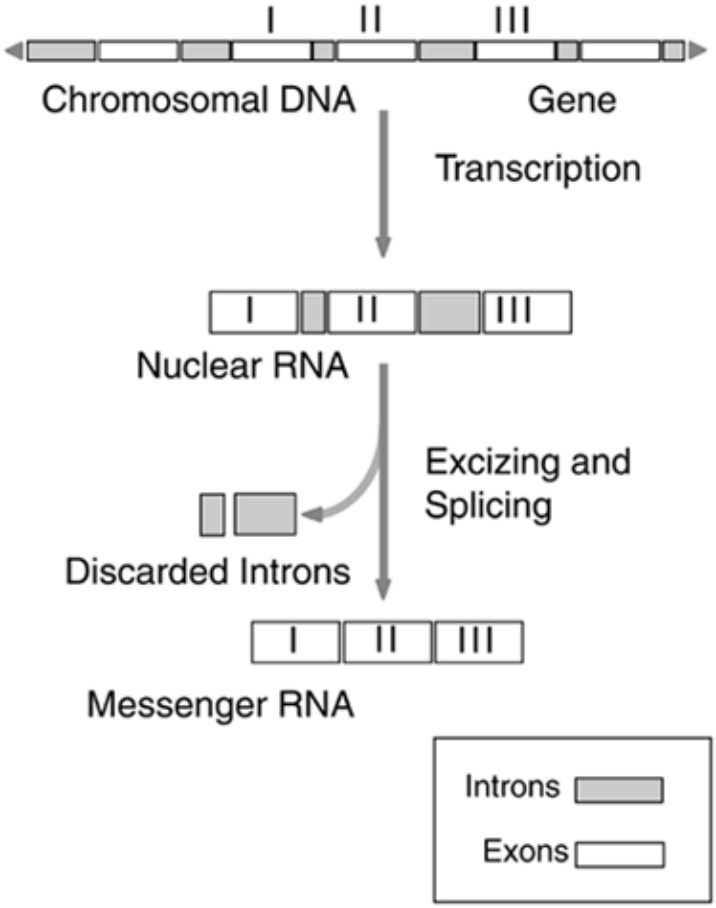

Figure 2. Messenger RNA (mRNA) Synthesis. DNA is transcribed to nuclear RNA (nRNA) this is in turn processed to mature mRNA in the nucleus. Maturation involves discarding the junk nucleotide sequences (introns) that interrupt the sequences that will eventually be involved in translation (exons).

In human cells, DNA is organized and compressed into 23 pairs of chromosomes, with one member of each pair inherited from each parent. Most of this DNA—on the order of 98.5 percent—is considered "junk," in that its

function is unknown. The remainder of the DNA is in the genes—the stretches of DNA involved in the transcription process.

Not only are there duplications in the remaining DNA, but there are additional non-coding nucleotide sequences. Interrupting the sequences of base pairs that will be expressed (the exons), there are interruptions in the sequence by segments that aren't expressed (the introns). Like the much larger expanses of "junk DNA" in the chromosome, these smaller interruptions in the DNA have unknown functions. Whether some of the non-coding DNA are remnants of provirus infections during hominoid evolution or somehow involved in compacting the DNA is open to conjecture.

In the process of RNA synthesis within the cell nucleus, DNA is transcribed to single-stranded nuclear RNA (nRNA), which is then processed to form mature messenger RNA (mRNA), as illustrated in Figure 1-6. Small nuclear RNA (snRNA) is involved in this maturation process, which includes excising the introns from the mRNA strands and concatenating the remaining exons according to their original order in the mRNA. As an information transport medium, RNA differs from DNA in that it's single stranded, much shorter, and the nitrogenous base Uracil (U) is substituted for Thymine.

Mature mRNA is transported through the nuclear membrane to the cytoplasm where the translation of mRNA to protein occurs with the aid of ribosomes. These ribosomes contain a variety of different proteins and an assortment of RNA molecules, collectively known as ribosomal RNA (rRNA). These short-lived but abundant rRNAs are involved in the binding of mRNA to the ribosome during the translation process.

The translation of mRNA to protein is facilitated by transfer RNA (tRNA), which associates with the 20 common amino acids (there are 22 genetically encoded amino acids) and controls the sequential binding of the amino acids according to the 3-letter base sequences (codons) on the mRNA. In this way, the tRNA is responsible for positioning the correct amino acid residue at the ribosome, as dictated by the base pair sequence in the mRNA.

Information is transmitted in transcription and translation processes through three-letter words and an alphabet of four letters. Because there are 3 base sequences, there are 43 or 64 possible codons (see Figure 3), 3 of which are used as stop codons, and 1 for the start codon, to mark the end and start of translation, respectively. The remaining codons are used as redundant representations of the amino acids. In most cases, the first two bases are fixed, and the redundancy occurs in the last base. For example, the codons CCC, CCT, CCA, and CCG all code for the amino acid Proline.

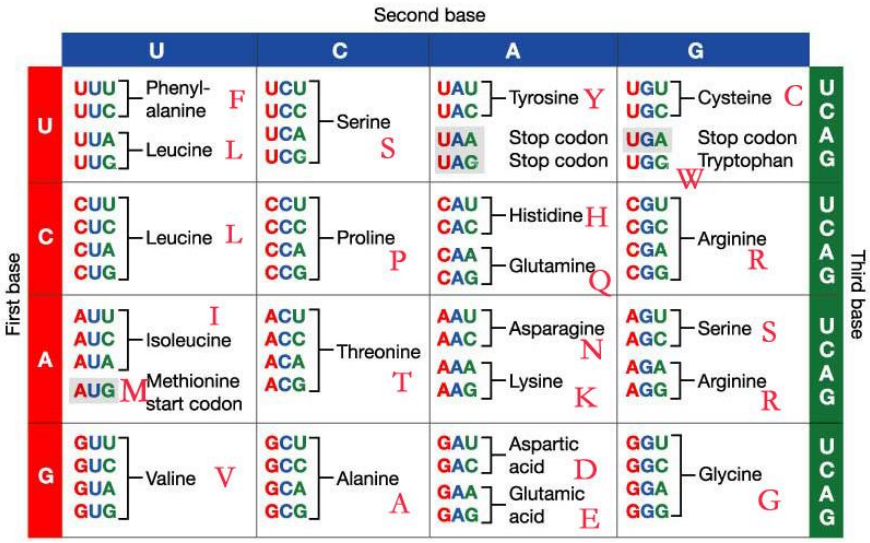

Figure 3. Codon Table. The 64 possible codons represent the 20 common amino acids, as well as one start (ATG) and three stop (TAG, TAA and TGA) markers. Redundancies normally occur in the last nucleotide of the three-letter alphabet.

During translation, amino acids are added to the growing protein sequence one at a time as the ribosome moves from codon to codon along the mRNA. The starting point for translation is marked by a start codon. At the end of the coding sequence, marked by a stop codon, the translation ends and the protein is released by the ribosome. Before the protein is transported outside of the cell to perform or promote a variety of tasks, it is usually modified by adding a sugar, for example, and it takes on a characteristic folded three-dimensional form—the focus of proteomics research.

The End