1 Introduction

To understand and fully appreciate the bioinformatics field, we must first step back and understand the basic concepts of what we know as life on Earth and the universal molecular components that constitutes all living cells on Earth. Despite the extraordinary diversity there are some underlying universalities. The first is in the molecules make up living systems. The important molecules of life are DNA, RNA and proteins. These are large and complex molecules, because they are made up of many atoms. They are really macromolecules because of their size.

The second universality is that of evolution. All life forms are related by common ancestry and can be traced back to what is also known as the LUCA (Last Universal Common Ancestor). Evolution of life is what allowed the vast diversity of life forms on earth to form despite the common ancestry of all organisms. Briefly, there are three phenomena in evolution: (a) self-replication: all organisms must be able to make copies of themselves, not exact copies though, offsprings must share some properties of their parents and this sharing is called inheritance, (b) variation: offsprings are not perfect replica of their parents, (c) selection: the process by which some organisms survive to successfully have offsprings while those who don't stand the test of evolution eventually perish in time. The reasons why some organisms survive and have progeny that will carry their genetic information while others don't is complex. In particular, it is because of combination of factors including an organisms genetic composition as well as the environment in which they are exposed to. The overall scope of evolution allows that the most successful individual traits that fits the environment will have higher probability of having healthy offsprings. This is called "survival of the fittest". Sep 4th's lecture will focus on two key molecules DNA and proteins, their sequences and problems associated with sequence analysis. However, we must first understand the central dogma of life and each of its component molecules that builds all forms of live on Earth.

2 Central Dogma of molecular biology

Before we describe each subcomponent of how living organism stores, replicates, and passes down the information of its life form, we should first discuss about the order of such events, called the Central Dogma. Information in biological systems flows from DNA to RNA to proteins. The process by which information stored in the DNA is turned into RNA is called transcription. This is similar to DNA replication, but uses a different enzyme called RNA polymerase. Further only one strand is used and it does not produce a new DNA but rather RNA. The same template is used over and over again to get multiple copies of the gene. Only one strand has the sequence for the protein, that is specifies what sequence of bases specifies the protein sequence. This is called the coding strand. The template is the other strand and is called the anti-sense strand or the non-coding strand. RNA polymerase always reads from 3' to 5' to make a "meaningful" protein sequence that goes from 5' to 3'. How the DNA specifies the protein sequence is through the genetic code.

3 DNA

DNA stands for de-oxyribonucleic acid. It is the blue-print of life and it stores the information that defines everything about an organism and is present in all individual living cells. In order for an organism to

thrive, DNA must be able to replicate itself with high fidelity. The structure of DNA is a double helix. This is inherently tied to its ability to replicate information. Every organism has a fixed number of DNA molecules.

DNA is very long and large molecule and it must be stored in special form where it is supercoiled into a form that is known as Chromosomes. The DNA molecule is a special type of molecule called a polymer which is made up smaller repeating units. These units are called nucleotides.The strand of DNA is built from these nucleotides, that is, the basic coding mechanism of life in quaternary sequence. Think of it as if the life programs itself in four different signals A, T ,C ,and G nucleotides where as computer consists of 0 and 1. These nucleotides have the same structure, and differ in one substructure called the "base". The structure is made up of a 5 carbon sugar molecule, and on the 1' carbon is the nitrogen containing base and the 5' position a phosphate group. The sugar in DNA is de-oxyribose and that's why the name de-oxyribonucleotide. DNA's nucleotides are of two types: purines (A and G) and pyrimidines (C and T). They differ in the base. Nucleotides are connected using the phosphate group of one nucleotide with the hydroxy group of the following. A sequence of DNA nucleotides has a free 5' end and a free 3' end. The sequence is defined by reading from 5' to 3'.

DNA is a double-stranded molecule and each strand of DNA is a string of these nucleotides. The sugar-phosphate groups are on the outside and the bases of each strand are on the inside. The way these two strands are connected is through a double helix where the two strands are wound around each other and held together by hydrogen bonds between the bases on each strand. The bases have a unique pair. A pairs with T and C pairs with G. Each strand is therefore a complement of the other strand. The strands run anti-parallel to each other. So if the sequence on one strand is AAG, then the sequence on the other strand is CTT and not TTC. Thus the two strands of DNA are anti-parallel copies of each other. Directionality is important. DNA is read always from 5' to 3'. The structure and the complementary base pairing of DNA sequences is one of the most profound discoveries in biology.

4 RNA

All RNA comes from copying of DNA and is another form of nucleic acid in cells that are directly transported and used for the function of information delivery and cellular signaling. RNA is similar to DNA in the sense that it is also a polymer made up of repeated nucleotides. However, it is single stranded. It is made up of also a different sugar. Its nucleotides are A, U, G, and C, where U is the analog of T in DNA. While most of the RNA gets translated into proteins there are some other types of RNA that do other important biological functions. These include

- tRNA: transfer RNAs, responsible for transferring specific amino acids into the ribosome. In fact for every codon there is a corresponding transfer RNA called the anti-codon.
- ribosomal RNA: They constitute major components of the ribosome, where the translation of mRNA to protein occurs.
- messenger RNA: mRNA these get turned into proteins

For all intents and purposes of our class, we will focus on the mRNA as the mRNA molecules deliver the copied version of the DNA message to be translated into a protein.

5 Proteins

Proteins the main workhorses in cells and are coded for in the DNA. Like DNA proteins are polymers too, meaning that they are made up of repeating units. These units are called amino acids. There are 20 amino acids. Specific segments of the DNA code for proteins. Such DNA segments are called genes. (To be precise, these are protein-coding genes because there are RNA genes that make RNA that does not get translated to proteins). Proteins have structure too. However this is much more complicated than DNA sequence. Specifically, proteins have the "primary sequence" which is the amino acid string, "secondary structure", which is composed "alpha" and "beta" helices, "tertiary structure", which is composed of multiple secondary structure units getting packaged and organized together, and finally, "quaternary structure" which is composed of multiple repeated units of tertiary structure components. The structure of the protein is very important for its function. The primary sequence specifies the structure of the protein.

6 The Genetic code

The genetic code dictates how the protein can be read out from a DNA sequence. Proteins are made up of 20 amino acids. So the key question is how many bases do you need to specify these 20 amino acids. You need only three. But there are 64 and the remaining are all redundant. The three bases are called codons. Translation is the process of going from the DNA sequence to protein sequence. DNA sequence is read non-overlapping sets of three. So the frame matters. Depending upon which position you started you might end with a different sequence. There are three frames on each strand.

7 A few definitions

- Cell: A fundamental unit of life. There are two types of cells eukaryotes and prokaryotes. The difference is eukaryotes have a nucleus, a membrane packaged cellular component in which DNA is packaged. Different cells have different functions. The functionality of each cell within a larger organism depends upon what genes are expressed in the cell despite the fact that all cells have the same DNA in the nucleus.
- Polymers: molecules made up of long strings of simpler (basic) repeating components.
- Nucleic acids: DNA (De-oxyribonucleic acids), and RNA (Ribonucleic acids). The basic components of nucleic acids are called nucleotides.
- Proteins: complex molecules that work together to carry out biological functions such as growth, differentiation, response to stress. Proteins constitute the basic machinery and building blocks within cells. Proteins generally do the same tasks in different organisms/species. For example the haemoglobin protein is responsible for carrying oxygen. The basic components of proteins are amino acids. The order in which the simpler units are linked together is called the sequence. Thus macromolecules can be thought of words, sentences or chapters from the book of living things.
- Gene: A single unit of inheritance. A unit of DNA sequence that codes for specific proteins.
- Genome: The complete complement of DNA of an organism.

- Structure and function: Two ways of studying the components and activities of living systems.Structure is the physical composition and physical relationships. Function describes the role a component plays.
- Central Dogma of molecular biology: DNA is transcribed (copied) to mRNA which is translated to proteins using amino acids.
- Genetic code: DNA specifies what proteins are present in a cell and the genetic code specifies this. Each substring of DNA must code for a protein. Since there are four nucleotides there must be at least three. In fact the codon which is the triplet of nucleotides that specifies which amino acid is to be made is redundant. This is because there are a total of 64 amino acids that can be specified, but only 20 are actually found in proteins. There are some special codons called the start and stop codons. The start codon specifies the start of a protein, and the stop codon species the end of a protein.
- Strand: This is used to refer to one of the strings of DNA that make up the double-helix of DNA. The two strands run anti-parallel to each other. Because of the sequences in each strand are "reverse-complements" of each other.
- Base or nucleotide: The unit of the DNA molecule. A DNA molecule is made of repeating bases of nucleotides. Since nucleotides differ in the "base" component the nucleotide and base terms are used interchangeably.
- Base-pairing: Pairing of A's with T's and G's with C's. Because of base-pairing, the sequence of one strand of DNA is the complement of the other strand.