**Course: Advanced Bioinformatics**

**Module title: Proteins and Genes**

**Module no. : 8**

# Proteins

Proteins are polypeptides that have a three dimensional structure. They can be described through four different hierarchical levels:

* **Primary structure** – the sequence of amino acids constituting the polypeptide chain.
* **Secondary structure** – the local organization of the parts of the polypeptide chain into secondary structures such as α helices and β sheets.
* **Tertiary structure** – the three dimensional arrangements of the amino acids as they react to one another due to the polarity and resulting interactions between their side chains.
* **Quaternary structure** – if a protein consists of several protein subunits held together, then the protein can be described as well by the number and relative positions of the subunits.

|  |  |
| --- | --- |
| **Visualization of Protein Structures.** | |
| http://www.ebi.ac.uk/microarray/biology_intro_files/rasmol.gif | http://www.ebi.ac.uk/microarray/biology_intro_files/triphos.jpe |
| Magenta: alpha helix  Gold: Beta Sheets | Blue: Monomer A  Orange: Monomer B |

Image source:

<http://www.ebi.ac.uk/microarray/biology_intro.html>

Calculating the secondary and tertiary structure of a protein given its primary structure is not an easy task. Protein folding prediction will be covered at some point close to the end of the semester.

**Monomer** – Any small molecule that can be linked with others of the same type to form a polymer. For the purpose of this class, the molecules could be nucleic acids, amino acids, or proteins.

**Dimer** - Two small molecules of the same type linked together.

**Trimer** – Three small molecules of the same type linked together.

**Oligimer** – General term for a short polymer most commonly consisting of nucleic acids or amino acids.

**Polymer** – Any large molecule consisting of multiple identical or similar subunits linked by covalent bonds.

Putting it all together, we get the flow of genetic information. That is, DNA directs the synthesis of RNA, and RNA then in turn directs the synthesis of protein. This flow of genetic information from nucleic acids to protein has been called the Central Dogma of Molecular Biology.

|  |  |
| --- | --- |
| Central Dogma of Molecular Biology | |
| trinity  Image Source: http://www.people.virginia.edu/~rjh9u/dnaprot.html | DNA  ↓  RNA  ↓ PROTEIN |

# What is a Gene?

Aaah, the million dollar question. In short, a gene can be described as the physical and functional unit of heredity that carries information from one generation to the next. A gene can be thought of as the DNA sequence necessary for the synthesis of a functional protein or RNA molecule.

# Genome, Transcriptome, Proteome

Whenever the term ***genome*** is used, it typically refers to the chromosomal DNA of an organism, or as far as sequencing is concerned, the heterochromatic regions of the chromosomal DNA. The number of chromosomes and genome size varies quite significantly from one organism to another. An example list of genome sizes is given below. Don’t be fooled by this table that the size of the genome and the number of genes determines the complexity of an organism. In fact, many plant genomes are much greater in size than the human genome!

|  |  |  |  |
| --- | --- | --- | --- |
| **ORGANISM** | **CHROMOSOMES** | **GENOME SIZE** | **GENES** |
| [*Homo sapiens*](http://www.ncbi.nlm.nih.gov/genome/guide/human/) (Humans) | 23 | 3,200,000,000 | ~ 30,000 |
| [*Mus musculus*](http://www.ncbi.nlm.nih.gov/genome/guide/mouse/)  (Mouse) | 20 | 2,600,000,000 | ~30,000 |
| [*Drosophila melanogaster*](http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7227)(Fruit Fly) | 4 | 180,000,000 | ~18,000 |
| *Saccharomyces cerevisiae* (Yeast) | 16 | 14,000,000 | ~6,000 |
| *Zea mays (Corn)* | 10 | 2,400,000,000 | ??? |

The term ***transcriptome*** refers to the complete collection of all possible mRNAs (including splice variants) of an organism. This can be thought of as the regions of an organism’s genome that get *transcribed* into messenger RNA. In some cases, the transcriptome can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes.

The term ***proteome*** refers to the complete collection of proteins that can be produced by an organism. The proteome can be studied either as a static (sum of all proteins possible) or a dynamic (all proteins found at a specific time point) entity.