Algorithms in Bioinformatics: Lecture 01 Introduction

Lucia Moura

Fall 2010

Introduction to the course

"**Bioinformatics** is the study of biology through computer modeling and analysis. It is a multi-discipline research involving biology, statistics, data-mining, machine learning and algorithms."

textbook: Wing-Kin SUNG, *Algorithms in Bioinformatics*, CRC Press, 2009.

This course will give an in-depth view of **algorithmic techniques** used in bioinformatics.

Course contents (tentative):

- Introduction to Molecular Biology (chapter 1)
- Sequence Similarity (chapter 2)
  global/local/semi-global alignment, gap penalty, scoring functions
- Suffix trees and related data structures (chapter 3)
  algorithms to build a suffix trees, applications
- Genome Alignment (chapter 4)
  methods use suffix tree and longest common subsequence algorithm
- Multiple sequence alignment (chapter 6)
  dynamic programming, approximation algorithms, heuristics
- Phylogeny Reconstruction (chapter 7)
  constructing a phylogenetic tree given different types of data
- Genome Rearrangement (chapter 9)
  reversals, transpositions, etc, various distances considered
- Other topics: RNA secondary structure prediction (guest lecture); other topics/guest lectures TBA

Course Administration

Please refer to the course outline:
http://www.site.uottawa.ca/ lucia/courses/5126-10/outline.html

Intro to Molecular biology: DNA, RNA, Protein

Our body has **organs** formed by **tissues** which are collections of similar **cells** that perform **specialized functions**.
A cell is the minimal self-reproducing unit in all living species. It performs two functions:

1. **stores and passes genetic information for preserving life from generation to generation.**
   This is done via **DNA** molecules.
2. **Performs chemical reactions necessary to maintain our life.**
   To do this portions of DNA called **genes** are transcribed into **RNA** molecules, which in turn guide the synthesis of **proteins**. Proteins are the main catalysts for chemical reactions in the cell.

Next we discuss these **macromolecules** (molecules formed from a collection of smaller molecules): **protein**, **DNA** and **RNA**.

Proteins

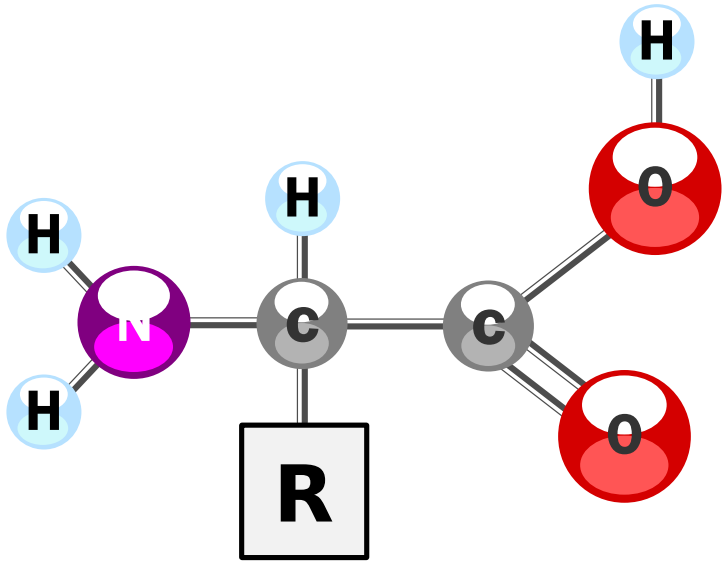
Proteins are the building blocks of cells; they execute nearly all cell functions.

Understanding proteins is essential to understanding how the body functions and other biological processes.

A **protein** (also called polypeptide) is a chain of **amino acids** (on average around 350 amino-acids form a protein), each bonding to its neighbour through a covalent peptide bond. The protein's **primary structure** is given by its sequence of amino-acids.
There are **20** different common amino acids.

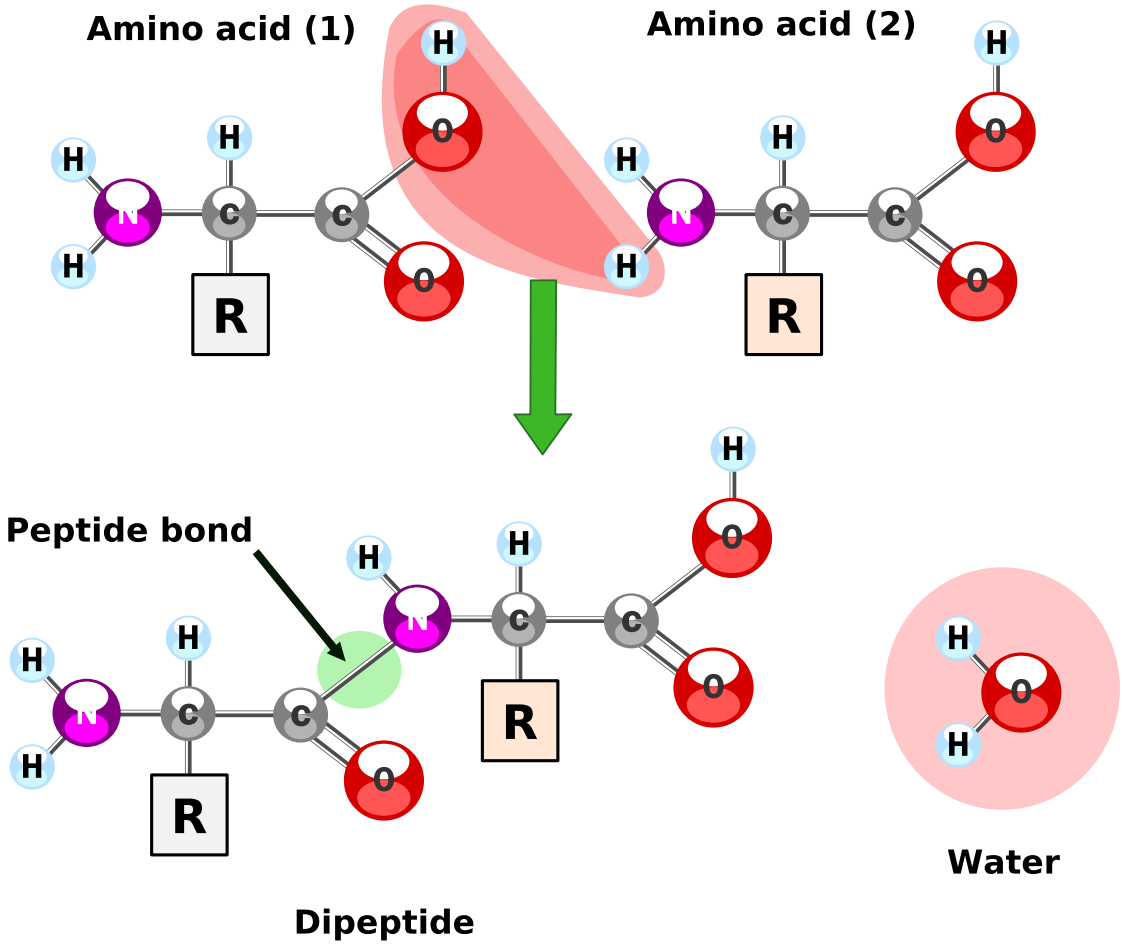
Computer science language: a protein's primary structure corresponds to a string (of length in average 350 symbols) over an alphabet of size 20.

Amino acid structure



1. Amino group (NH_2)
2. Carboxyl Group (COOH)
3. R-group (side chain):
   Different R-groups (side chain) characterize each of the 20 common amino acids.

Amino acids join together via a peptide bond



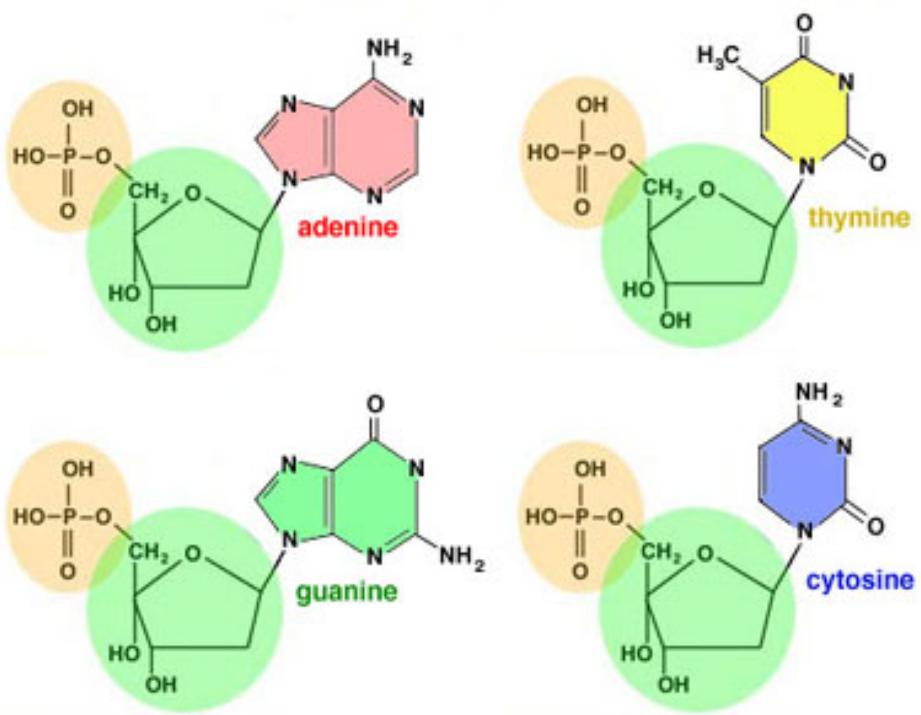
DNA

Deoxyribonucleic Acid (DNA) is the genetic material in all living organisms. It stores the instructions for the cell to perform its functions.
"DNA can be thought of as a large cookbook with recipes for making every protein in a cell. (...)
The information in the genes is read, perhaps millions of times in the life of an organism, but the DNA itself is never used up."

DNA consists of 2 strands of **nucleotides** forming a double helix structure.
DNA nucleotides vary depending on 4 possible **nitrogenous bases**: adenine (A), guanine (G), cytosine (C), thymine (T).
One strand is a polynucleotide (a sequence of nucleotides of 4 types); the second strand has their complementary base pairs ($A = T$, $C \equiv G$).

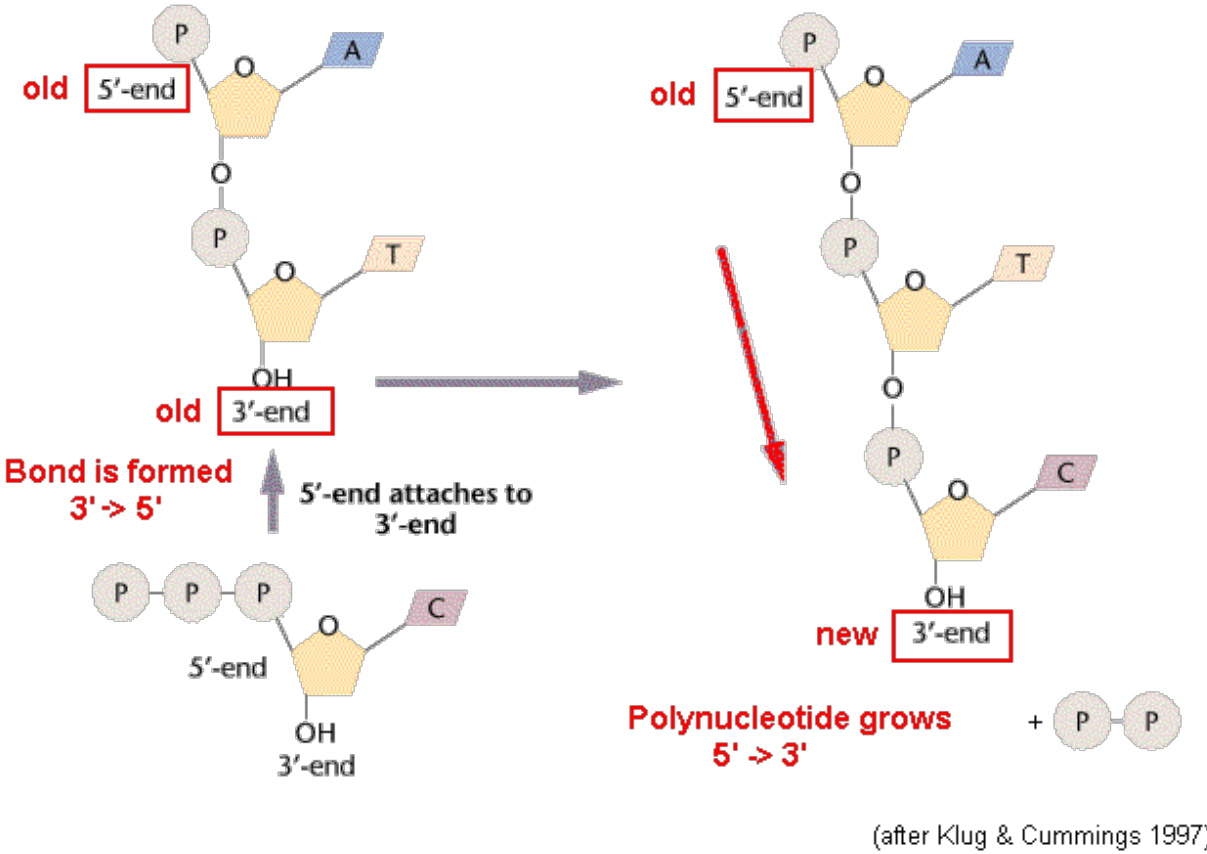
Computer science language: a DNA's primary structure corresponds to a string over the alphabet A, C, T, G (the second strand is determined by the first).

DNA Nucleotides



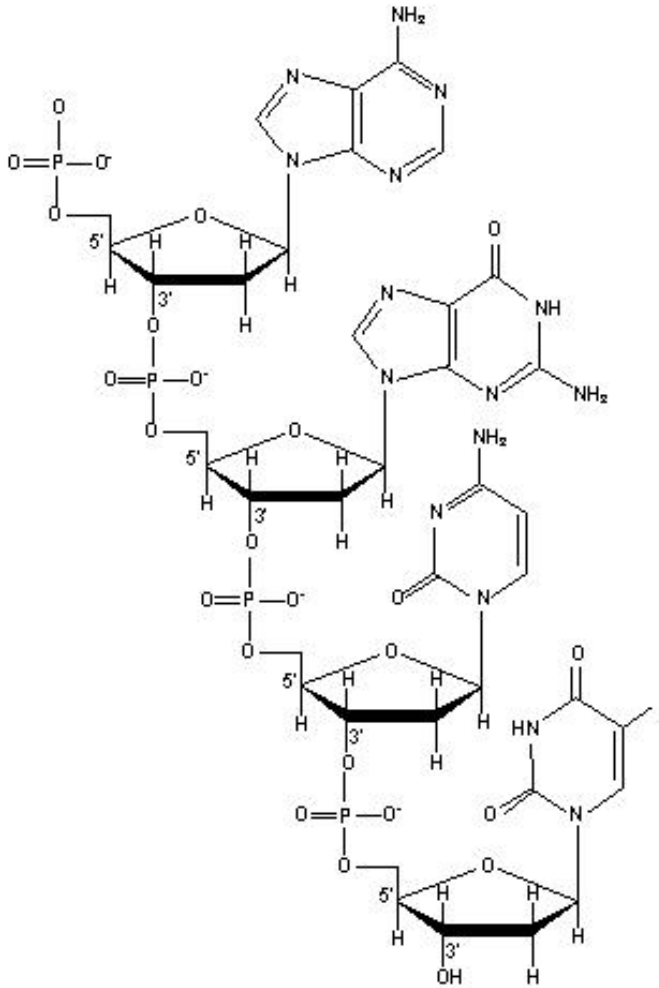
1. A pentose sugar deoxyribose
2. Phosphate group (bound to the 5' carbon)
3. Nitrogenous base (bound to the 1' carbon): A, C, T, G

DNA formed by chaining nucleotides I

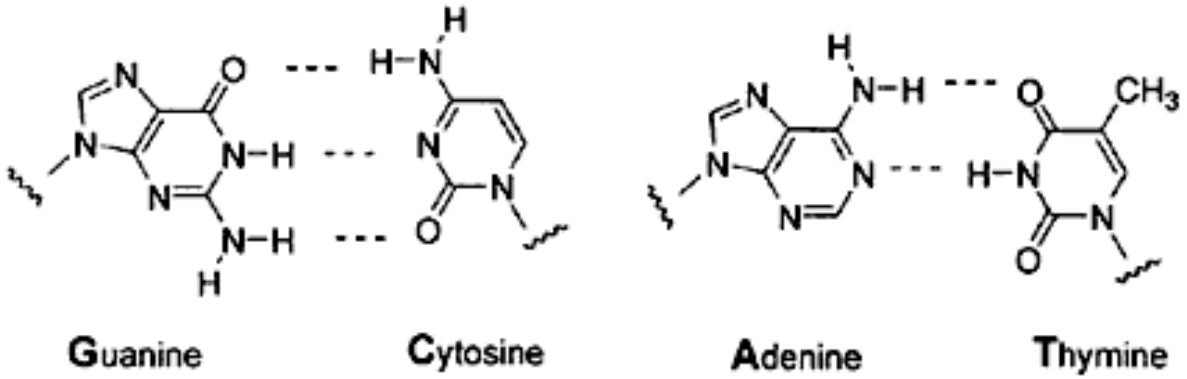


(after Klug & Cummings 1997)

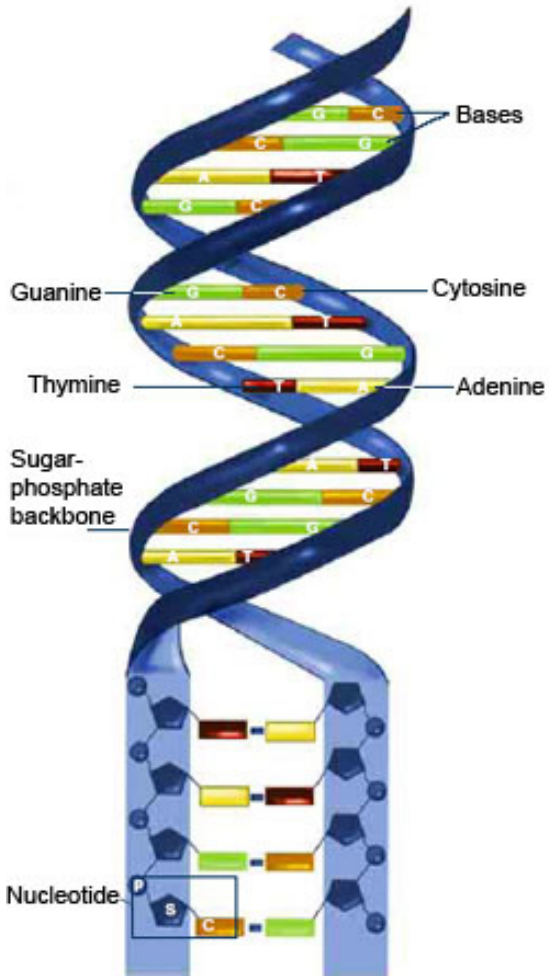
DNA formed by chaining nucleotides II



Watson-Crick base paring



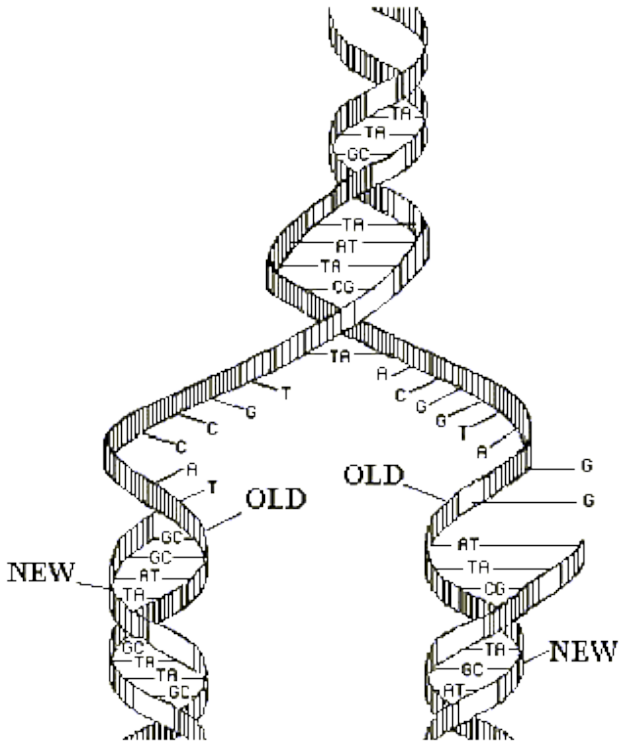
DNA double helix structure (Watson and Crick 1958)



DNA replication

Cell duplicates and passes DNA to two daughter cells.

1. double strand separated
2. each strand forms a template for a complementary new strand



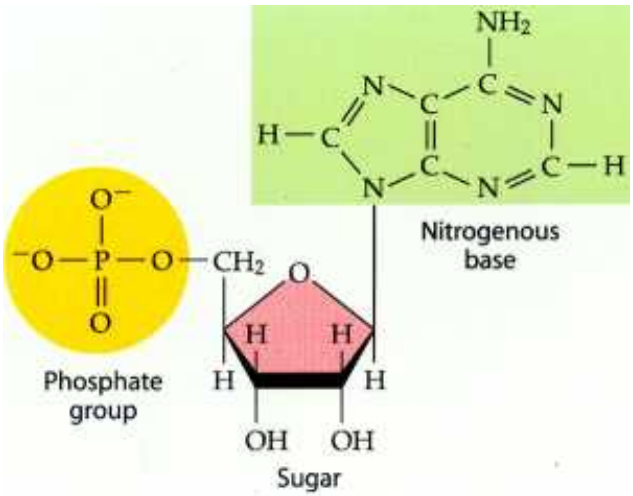
RNA

Ribonucleic Acid (RNA) is the nucleic acid produced during the transcription process (from DNA to RNA).
The nucleotide structure for RNA is similar to the one for DNA.
Differences:

1. Ribose Sugar in place of Deoxyribose;
2. Nitrogenous bases are (A, U), (C, G); Uracyl instead of Thymine.

RNA is single stranded.
RNA can form more complex 3D structures (than DNA) to perform more functions.
Proteins can perform even more functions than RNA.
DNA is more stable to store information than RNA.

Nucleotide structure for RNA



1. A pentose sugar ribose
2. Phosphate group (bound to the 5' carbon)
3. Nitrogenous base (bound to the 1' carbon): A,U,C,G

Different types of RNA

- mRNA: messenger RNA
  carry encoded information needed to make proteins
- ncRNA: non-coding RNA, which includes:
  - ribosomal RNA (rRNA):
    are parts of ribosomes, help translate mRNA into proteins
  - transfer RNA (tRNA):
    are like molecular diccionaries that translate the nucleic acid code into the amino acid sequence of proteins.
  - short ncRNA:
    regulate the process for generating proteins from genes.
  - long ncRNA:
    diverse functions, unknown functions.

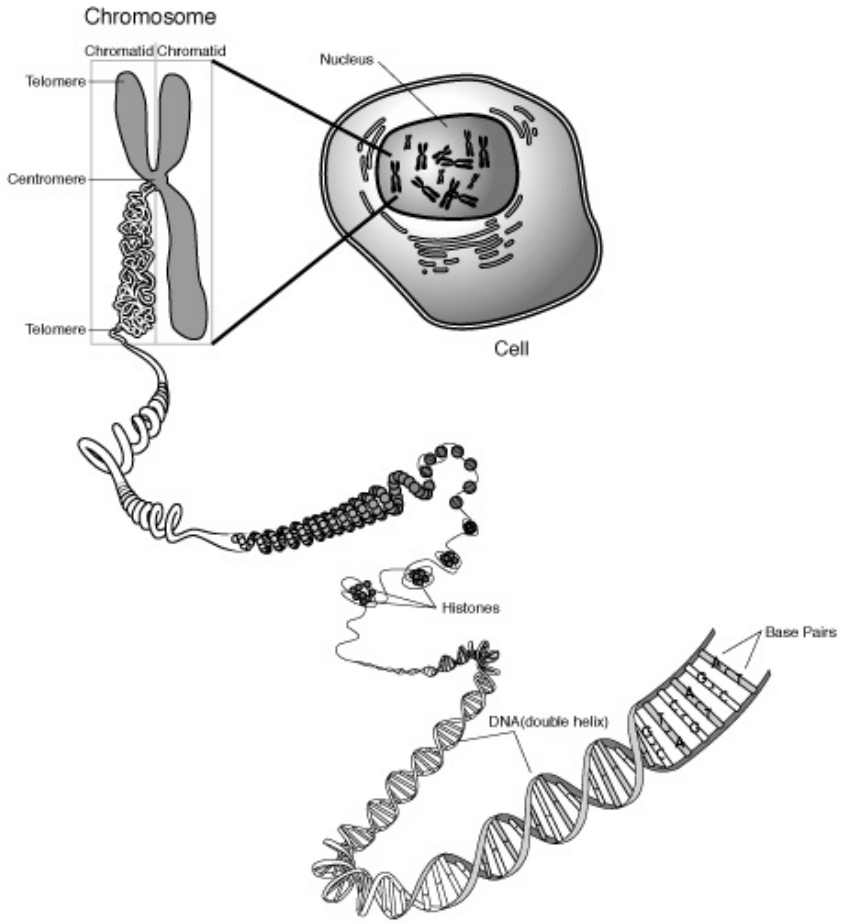
Genome, Chromosome and Gene

genome: the set of all DNA in an organism.
genome size varies; size doesn't necessarily correspond to complexity:
bacteria *Mycoplasma genitalium* genome has $\sim$ 600,000 base pairs;
human and mouse genomes have $\sim$ 3 billion base pairs;
the single cell organism *Amoeba dubia* has $\sim$ 670 billion base pairs!

the genome is partitioned into **chromosomes**; each chromosome is s double-stranded DNA chain wrapped around histones. Humans have 23 pairs of chromosomes (e.g. males have 22 pairs of autosomes, one X and one Y chromosome).

a **gene** is a "substring" of DNA that encodes a protein or an RNA molecule. Each chromosome contains many genes. In the human genome there are $\sim$ 30,000 genes.

Chromosomes



Processes to be understood in more detail next class:

- **DNA replication and DNA mutation.**
- **Central Dogma (proposed by Crick in 1958)**
  process of transfering information from DNA to RNA to protein.
  - **Transcription (transfer of genetic information from the DNA to the mRNA):**
    DNA is transcribed to mRNA, i.e., during the transcription process, an mRNA is synthesized from a DNA template.
  - **Translation (mRNA is translated to protein):**
    The mRNA is translated into an amino acid sequence. Here the genetic code is used: each codon (3 consecutive symbols) is translated into their corresponding amino acid.

Brief History of Bioinformatics

- 1866: Mendel discovered genetics (hybridization of peas, genes)
- 1869: DNA was discovered.
- 1944: Avery & McCarty show DNA is the carrier of genetic info.
- 1953: Watson and Crick deduced the double helix structure of DNA.
- 1970's and beyond: several biotechnology techniques were developed. E.g. DNA sequencing using any tissue ; polymerase chain-reaction.
- 1986: RNA splicing in eukaryotes is discovered (introns/extrons)
- 1998: Fire and Mello discovered RNA interference
- 1980-1990: genome sequencing of various organisms (e.g. E. coli)
- 1990: the **human genome project** is lauched
- 2003: sequencing of the human genome (first draft 2000)
- 2006-now second generation sequencing technology is available
- Other projects: Genomes to Life (understand the detailed mechanism of cells), ENCODE (annotating: all the genes & functional elements), HAPMAP(study differences in genetic data among people)