Computational Biology

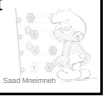
Lecture 1

Life

- In nature, we find living things and non living things.
- Living things can move, reproduce, … as opposed to non living things.
- Both are composed of the same atoms and conform to the same physical and chemical rules.
- What is the difference then?

Proteins and Nucleic Acids

- The main actors in the chemistry of life are molecules called proteins and nucleic acids.
- Proteins are responsible for what a living being is and does in a physical sense.
- Nucleic acids encode the information necessary to produce the proteins and are responsible for passing along this “recipe” to subsequent generations.



Proteins

- Most substances in our bodies are proteins
  - Structural proteins: act as tissue building blocks
  - Enzymes: act as catalyst of chemical reactions
  - Others: oxygen transport and antibody defense
- What exactly is a protein?
  - A chain of simpler molecules called amino acids

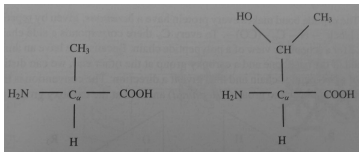
Saad Mneimneh

Amino Acid

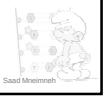
- An amino acid consists of:
  - Central carbon atom
  - Hydrogen atom
  - Amino group (NH_2)
  - Carboxy group (COOH)
  - Side chain
- The side chain distinguishes an amino acid from another
- In nature, we have 20 amino acids

Saad Mneimneh

Amino Acid



Examples of amino acids:
alanine (left) and threonine



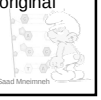
Saad Mneimneh

Peptide Bonds

- In a protein, amino acids are joined by peptide bonds.
- Peptide bond: the carbon atom in the carboxy group of amino acid A_i bonds to the nitrogen atom of amino acid A_{i+1}'s amino group.

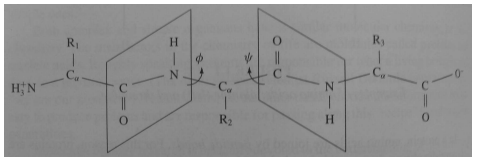
—**C**—(CO)—N—**C**—

- A water molecule is liberated in this bond, so what we really find in the protein chain is a residue of the original amino acid.



Saad Mneimneh

Poly Peptide Chain



- The protein folds on itself in 3D.
- The final 3D shape of the protein determines its function (why?).

Saad Mneimneh

Nucleic Acids

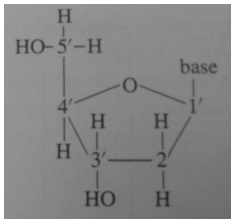
- How do we get our proteins?
- Amino acids of a protein are assembled one by one thanks to information contained in an important molecule called messenger ribonucleic acid.
- Two kinds of nucleic acids
  - RiboNucleic Acid: RNA
  - DeoxyriboNucleic Acid: DNA

Saad Mneimneh

DNA

- DNA is also a chain of simpler molecules.
- It is actually a double chain, each chain is called a strand.
- A strand consists of repetition of the same *nucleotide* unit. This unit is formed by a sugar molecule attached to a phosphate residue and a base.

Nucleotide



2′-deoxyribose molecule

4 bases:
- Adenine (A)
- Guanine (G)
- Cytosine (C)
- Thymine (T)

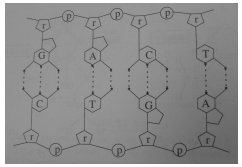
We use nucleotide and base interchangeably

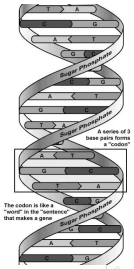
DNA Double Helix

- The two strands of a DNA are tied together in a helical structure.
- The famous double helix structure was discovered by James Watson and Francis Crick in 1953.
- The two strands hold together because each base in one strand bonds to a base in the other.

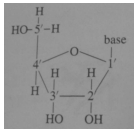
A ↔ T (complementary bases)
C ↔ G (complementary bases)

DNA Double Helix

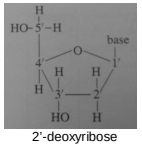




RNA



ribose



2'-deoxyribose

- Ribose instead of deoxyribose.
- RNA does not contain Thymine T, instead Uracil U is present (which also binds with A).
- RNA does not form a double helix.

Genes

- Each cell of an organism has a few very long DNA molecules, these are called chromosomes.
- Certain continuous stretches along the chromosomes encode information for building proteins.
- Such stretches are called genes.
- Each protein corresponds to one and only one gene.

Genetic Code

- To specify a protein we need just specify each amino acid it contains.
- This is what exactly a gene does, using triplets of bases to specify each amino acid.
- Each triplet is called a codon.
- Genetic code: table that gives correspondence between each possible triplet and each amino acid.
- Some different triplets code the same amino acid (why?).
- Some codons do not code amino acids but are used to signal the end of a gene.

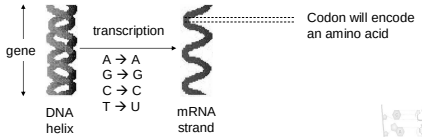
Genetic Code

The genetic code mapping codons to amino acids.

First position	Second position G	Second position A	Second position C	Second position U	Third position
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	STOP	Ser	Leu	G
	STOP	STOP	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U

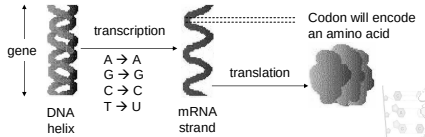
Transcription

The process by which a copy of the gene is made on an RNA molecule called messenger RNA, mRNA.

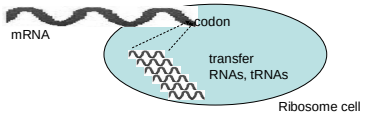


Translation

The process of implementing the genetic code and producing the protein. This happens inside a cellular structure called ribosome.



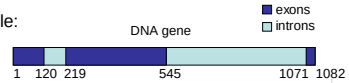
More on Translation



- Each tRNA has on one side high affinity for a specific codon, and on the other side high affinity for the corresponding amino acid.
- As mRNA passes through the ribosome, a tRNA matching the current codon binds to it, bringing along the corresponding amino acid.
- When a stop codon appears, no tRNA associates with it and the process stops.

Introns and Exons

- In complex organisms (e.g. humans), genes are composed of alternating parts called introns and exons.
- After transcription, all introns are spliced out from the mRNA.
- Example:



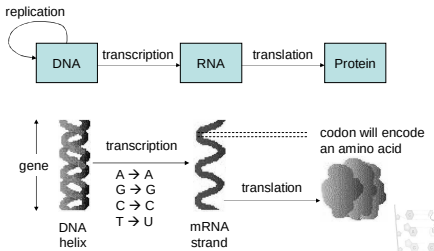
RNA will have 120 + 327 + 12 = 459 bases
Protein will have 153 residues

Junk DNA

- The DNA contains genes and regulatory regions around genes that play a role in controlling gene transcription and other related processes.
- Otherwise, intergenetic regions have no known function.
- They are called “Junk DNA”
- 90% of DNA in humans is JUNK.

Biology in ONE slide

the so-called central dogma of molecular biology



Chromosomes

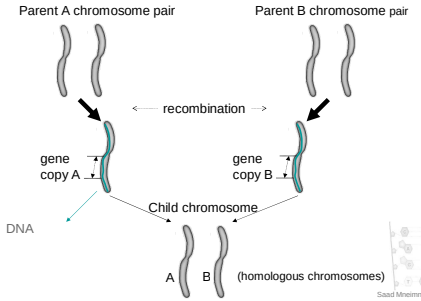
- Chromosomes are very long DNA molecules.
- The complete set of chromosomes is called the genome.
- Genetic information transmission occurs at the chromosome level (but genes are the units of heredity).
- Simple organisms, like bacteria, have one chromosome, which is sometimes a circular DNA molecule.
- In complex organisms, chromosomes appear in pairs. Humans have 23 pairs of chromosomes. The two chromosomes that form a pair are called homologous.

Gregor Mendel (1822 – 1884)

Mendel studied the characteristics of pea plants. He proposed two laws of genetics:

- (1st law) Each organism has two copies of a gene (one from each parent) on homologous chromosomes, and in turn, will contribute, with equal chance, only one of these two copies.
- (2nd law) genes are inherited independently (*not very accurate*).

Heredity



Looking for genes: 1980s

- *Cystic Fibrosis* is a fatal disease associated with recurrent respiratory infection
- *early 1980s*: the search for CF gene started
- *1985:* CF gene proved to reside on the 7th chromosome
- *1989*: the 1,480 bases long CF gene was found
- *Why all this*? Best cure of many hereditary diseases lies in finding the defective genes

Biological Problems

Genetics involve many computational problems:

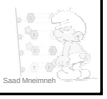
- Genetic Mapping
- Physical Mapping
- Sequencing
- Similarity Search
- Gene Prediction
- Proteomics

Genetic Mapping

- **Genetic Mapping**: Position genes on the various chromosomes to understand the genome's geography
- To understand the nature of the computational problem involved, we will consider an oversimplified model of genetic mapping, smurfs

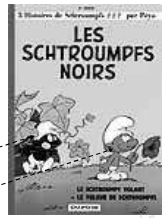
Smurfs

- Uni-chromosomal smurfs
- *n* genes (unknown order)
- Every gene can be in two states 0 or 1, resulting in two phenotypes (physical traits), e.g. black and blue



Example Smurfs

- Three genes, n=3
- The smurf's three genes define the color of its
  - Hair
  - Eyes
  - Nose
- 000 is all-black smurf
- 111 is all-blue smurf



Saad Mneimneh

Heredity

- Although we can observe the smurfs' phenotype (i.e. color of hair, eyes, nose), we don't know the order of genes in their genomes.
- Fortunately, smurfs like sex, and therefore may have children, and this helps us to construct the smurfs' genetic maps.

Saad Mneimneh

Smurfs Having Sex

X *cannot show picture*

Saad Mneimneh

Genetic Mapping Problem

- A child of smurf $m_1 \ldots m_n$ and $f_1 \ldots f_n$ is either a smurf $m_1 \ldots m_i f_{i+1} \ldots f_n$ or a smurf $f_1 \ldots f_i m_{i+1} \ldots m_n$ for some recombination position i.
- Every pair of smurfs may have $2(n+1)$ kinds of children (some of them maybe identical), with probability of recombination position at position i equal to $1/(n+1)$.
- **Genetic Mapping Problem**: Given the phenotypes of a large number of children of all-black and all-blue smurfs, find the gene order in the smurfs.

Frequencies of Pairs of Phenotypes

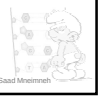
- Analysis of the frequencies of different pairs of phenotypes allows to determine gene order. How?
- Compute the probability p that a child of an all-black and an all-blue smurf has hair and eyes of different color.
- If the hair gene and the eye gene are consecutive in the genome, then $p=1/(n+1)$. In general $p=d/(n+1)$, where d is the distance between the two genes.

Reality

Reality is more complicated than the world of smurfs.

- Arbitrary number of recombination positions.
- Human genes come in pairs (not to mention they are distributed over 23 chromosomes).
  - Father: $F_1 \ldots F_n | F_1 \ldots F_n$
  - Mother: $M_1 \ldots M_n | M_1 \ldots M_n$
  - Child $f_1 \ldots f_n | m_1 \ldots m_n$, with $f_i = F_i$ or F_i and $m_i = M_i$ or M_i.

But same concept applies, if genes are close, recombination between them will be rare. This is where Mendel's 2nd law is wrong (genes on the same chromosome are not inherited independently).

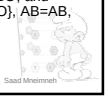


Difficulties

- Genes may not be consecutive on a single chromosome
  - Humans have 23 long chromosomes, it is very likely that genes are distant and distributed
- Very hard to discover the set of phenotypes to observe
  - If we are looking for the gene responsible for cystic fibrosis, which other phenotypes should we look for?

Variability of Phenotype

- Our ability to map genes in smurfs is based on the variability of phenotypes in different smurfs.
  - Example: If smurfs are either all-black or all-blue (which is the case by the way, ask Peyo), it would be impossible to map their genes.
- Different genotypes do not always lead to difference in phenotypes (i.e. difference not observable)
  - Example: *ABO* blood type gene has three states: *A*, *B*, and *O*. There exist six possible genotypes: *AA*, *AB*, *AO*, *BB*, *BO*, and *OO*, but only four phenotypes: A={AA, AO}, B={BB, BO}, AB=AB, O=OO



Observability of Phenotypes

There are a lot of variations in the human genome that are not directly expressed in phenotypes.

- Example: more than one variation is required to trigger a phenotype, for instance, some diseases are triggered by the presence of multiple mutations, but not by a single mutation.