

Single Nucleotide Polymorphism

Thesis of the course of Biomedical Informatics
A.A. 2013-2014

Deoxyribonucleic acid (en-us: Deoxyribonucleic acid.ogg /dɒiˈnʊkliːk ˈæsɪd/ (help·info)) (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints, or a recipe, or a code, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules. The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

Chemically, DNA consists of two long polymers of simple units called nucleotides, with bases made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are therefore anti-parallel. Attached to each sugar is one of four types of molecules called bases. It is the sequence of these four bases along the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related messenger RNA, in a process called transcription.

random][plasmid

Within cells, DNA is organized into long structures called chromosomes. These chromosomes are duplicated before cells divide, in a process called DNA replication. Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA inside the cell nucleus and some of their DNA in organelles, such as mitochondria or chloroplasts. In contrast, as prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm. Within the chromosomes, chromatin proteins such as histones compact and organize DNA. These compact structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

DNA is a self-complementary molecule. The two strands are called the template strands. The sequence of bases on one strand determines the sequence of bases on the other strand. The sequence of bases on one strand is complementary to the sequence of bases on the other strand. The sequence of bases on one strand is complementary to the sequence of bases on the other strand. The sequence of bases on one strand is complementary to the sequence of bases on the other strand.

The first published model of the DNA double helix was proposed by James Watson and Francis Crick in 1953. The model was based on the work of Rosalind Franklin and Maurice Wilkins, who had discovered the structure of DNA using X-ray diffraction. The model was a major breakthrough in the understanding of DNA and its role in heredity.

Although the B-DNA model was the first to be published, it was not the only model. Other models were proposed, including the A-DNA model and the Z-DNA model. The B-DNA model was the most widely accepted and is the most common form of DNA.

Compared to B-DNA, the A-DNA model is a compact, wide, and narrow, and the Z-DNA model is a narrow, zig-zag, and the C-DNA model is a compact, wide, and narrow. The A-DNA model is the most common form of DNA in the cell nucleus, and the Z-DNA model is the most common form of DNA in the cell cytoplasm.

DNA exists in many forms. The most common form is B-DNA, but there are also A-DNA, Z-DNA, and C-DNA. The A-DNA model is a compact, wide, and narrow, and the Z-DNA model is a narrow, zig-zag, and the C-DNA model is a compact, wide, and narrow.

The first published model of the DNA double helix was proposed by James Watson and Francis Crick in 1953. The model was based on the work of Rosalind Franklin and Maurice Wilkins, who had discovered the structure of DNA using X-ray diffraction. The model was a major breakthrough in the understanding of DNA and its role in heredity.

Although the B-DNA model was the first to be published, it was not the only model. Other models were proposed, including the A-DNA model and the Z-DNA model. The B-DNA model was the most widely accepted and is the most common form of DNA.

Compared to B-DNA, the A-DNA model is a compact, wide, and narrow, and the Z-DNA model is a narrow, zig-zag, and the C-DNA model is a compact, wide, and narrow. The A-DNA model is the most common form of DNA in the cell nucleus, and the Z-DNA model is the most common form of DNA in the cell cytoplasm.

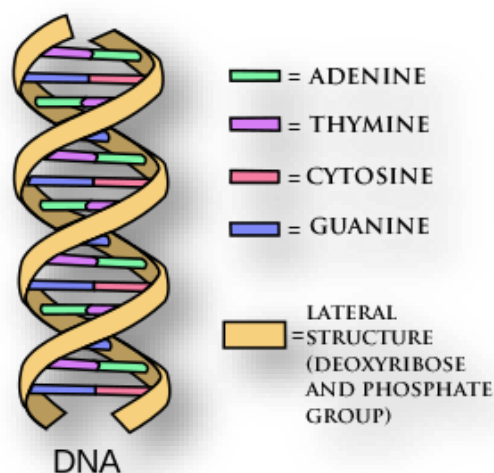
INTRODUCTION

In 1870, the Swiss chemist *Miescher* discovered inside the nucleus of a cell a giant molecule: **deoxyribonucleic acid**.

In 1953, two biochemists, the American *James Watson* and the English *Francis Crick* show that the structure of the DNA molecule is comparable to that of a spiral staircase; a sort of spiral-shaped double helix.

Deoxyribonucleic acid (DNA) is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms and many viruses. DNA is a nucleic acid; together with proteins and carbohydrates, nucleic acids compose the three major macromolecules essential for all known forms of life.

Most DNA molecules consist of *two biopolymer strands coiled around each other to form a double helix*. The two DNA strands are known as polynucleotides since they are composed of simpler units called *nucleotides*. Each nucleotide is composed of a **nitrogen-containing nucleobase**—either **guanine (G)**, **adenine (A)**, **thymine (T)**, or **cytosine (C)**—as well as a monosaccharide sugar called **deoxyribose** and a **phosphate group**. The nucleotides are joined to one another in a chain by *covalent bonds between the sugar of one nucleotide and the phosphate of the next*, resulting in an *alternating sugar-phosphate backbone*. According to base pairing rules (A with T and C with G), hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double-stranded DNA.



Picture 1 – DNA structure

DNA is well-suited for biological information storage. The DNA backbone is resistant to cleavage, and both strands of the double-stranded structure store the same biological information. Biological information is **replicated** as the two strands are separated. A significant portion of DNA (more than 98% for humans) is non-coding, meaning that these sections do not serve a function of encoding proteins.

That said, it is easy to understand how DNA is important for life. For this reason, even a small mutation (a change of the nucleotide sequence of the genome of an organism) can be decisive and cause diseases.

In this essay we will discuss a particular case of genomic mutation, the Single Nucleotide Polymorphism.

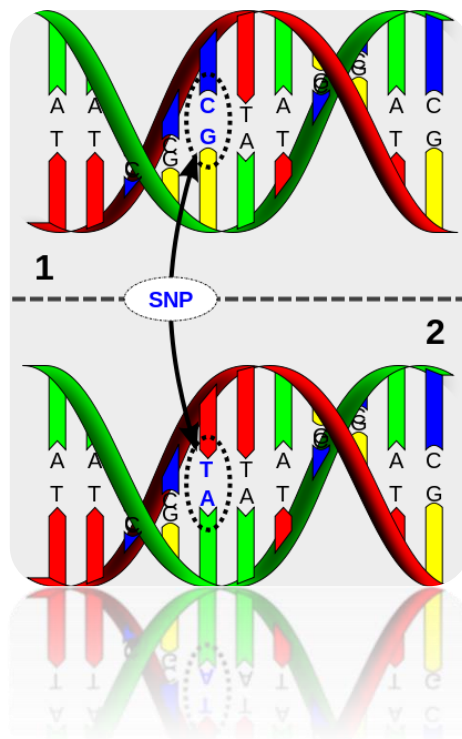
Single Nucleotide Polymorphism: what is it?

A **Single Nucleotide Polymorphism (SNP)** is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a Single Nucleotide — A, T, C or G — in the genome differs between members of a biological species or paired chromosomes.

For example, if we have two sequenced DNA fragments from different individuals (see *Picture 2*):

- AAGCCTA
- AAGCTTA

The second one contains a difference in a single nucleotide. In this case we say that there are two alleles. Almost all common SNPs have only two alleles.



Picture 2 – SNP example

The genomic distribution of SNPs is not homogenous; SNPs occur in non-coding regions more frequently than in coding regions.

The main causes of a SNP are:

1. natural selection, acting and fixating the allele of the SNP that constitutes the most favorable genetic adaptation
2. like genetic recombination
3. mutation rate

The different possible types of SNPs

As previously seen, Single Nucleotide Polymorphisms may fall within *coding* sequences of genes, *non-coding* regions of genes, as well as in the *intergenic* regions (regions between genes).

What is a coding?

To understand the difference between SNPs' types, we have to see what a coding is.

The main concept to analyse is the **Genetic Code**: it is the *set of rules* by which information encoded within genetic material (DNA or even mRNA sequences) is *translated* into proteins by living cells.

During the translation, the sequence of nitrogenous bases is treated in groups of three at a time; a group of three nitrogenous bases is called a **codon**. The code defines how codons specify which amino acid will be added next during protein synthesis. Generally, three-nucleotide codon in a nucleic acid sequence specifies a single amino acid. On the other hand, **a single amino acid can be specified by more than one codon**: this is the key concept that we will need in the following.

To understand better, here is the table that shows, for each amino acid (20 in total + START and STOP), the sequences that can generate it:

Amino acid	Codons
Ala/A	GCT, GCC, GCA, GCG
Arg/R	CGT, CGC, CGA, CGG, AGA, AGG

Asn/N	AAT, AAC
Asp/D	GAT, GAC
Cys/C	TGT, TGC
Gln/Q	CAA, CAG
Glu/E	GAA, GAG
Gly/G	GGT, GGC, GGA, GGG
His/H	CAT, CAC
Ile/I	ATT, ATC, ATA
Leu/L	TTA, TTG, CTT, CTC, CTA, CTG
Lys/K	AAA, AAG
Met/M	ATG
Phe/F	TTT, TTC
Pro/P	CCT, CCC, CCA, CCG
Ser/S	TCT, TCC, TCA, TCG, AGT, AGC
Thr/T	ACT, ACC, ACA, ACG
Trp/W	TGG
Tyr/Y	TAT, TAC
Val/V	GTT, GTC, GTA, GTG
START	ATG
STOP	TAA, TGA, TAG

Table 1 – Inverse genetic code

SNPs in the coding sequences

SNPs that fall in this category can be divided into two subcategories:

1. **Synonymous**
2. **Nonsynonymous**
 - a. **Missense**
 - b. **Nonsense**

First ones does not result in a change in the protein sequence, because the “original” sequence and the real sequence of bases both code the same amino acid.

Second ones, instead, change the amino acid sequence of protein. In their turn, they can be of two types: *Missense*, in which a single nucleotide change results in a codon that codes for a different amino acid (that can render the resulting protein non-functional), and *Nonsense*, that results in a premature stop codon, or a nonsense codon and then in a truncated, incomplete, and usually non-functional protein product.

Let us see an example of *Missense mutation*:

Original DNA code for the amino acid sequence:

C	A	T	C	A	T	C	A	T	C	A	T	C	A	T	C	A	T
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Resulting amino acids:

His	His	His	His	His	His	His
-----	-----	-----	-----	-----	-----	-----

If we had, for example, a replacement of the eleventh nucleotide:

C	A	T	C	A	T	C	C	T	C	A	T	C	A	T	C	A	T
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Resulting amino acids will be:

His	His	His	Pro	His	His	His
-----	-----	-----	-----	-----	-----	-----

This is, instead, an example of *Nonsense mutation*:

Original DNA code for the amino acid sequence:

A	T	G	A	C	T	C	A	C	C	G	A	G	C	G	C	G	A	A	G	C
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Resulting amino acids:

Met	Thr	His	Arg	Ala	Arg	Ser
-----	-----	-----	-----	-----	-----	-----

If we had, for example, a replacement of the tenth nucleotide:

A	T	G	A	C	T	C	A	C	T	G	A	G	C	G	C	G	A	A	G	C
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Resulting amino acids will be:

Met	Thr	His	Stop			
-----	-----	-----	------	--	--	--

Nonsense mutations are jointly responsible for many diseases; they can cause a genetic disease by damaging a gene responsible for a specific protein (for example *dystrophin* in *Duchenne muscular dystrophy*).

Examples of diseases in which nonsense mutations are known to be among the causes include:

- **Cystic fibrosis**
- **Duchenne muscular dystrophy** (dystrophin)
- **Beta thalassaemia** (β -globin)
- **Hurler syndrome**

On the other hand, cancer associated Missense mutations can lead to drastic destabilisation of the resulting protein.

SNPs not in coding regions

SNPs that are not in protein-coding regions may still affect:

1. gene splicing
2. transcription factor binding
3. messenger RNA degradation
4. ...

Gene expression affected by this type of SNP is referred to as an **eSNP** (*expression SNP*).