UNIVERSITÀ DEGLI STUDI ROMA TRE

Faculty of Engineering
Master Degree Course in Computer Engineering

Single Nucleotide Polymorphism

Thesis of the course of Biomedical Informatics

Authors
D. Tosoni , D. Sicignani, U. Buonadonna

Academic Year
2013/2014

Capitolo 1

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 *al-*

ternating 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.

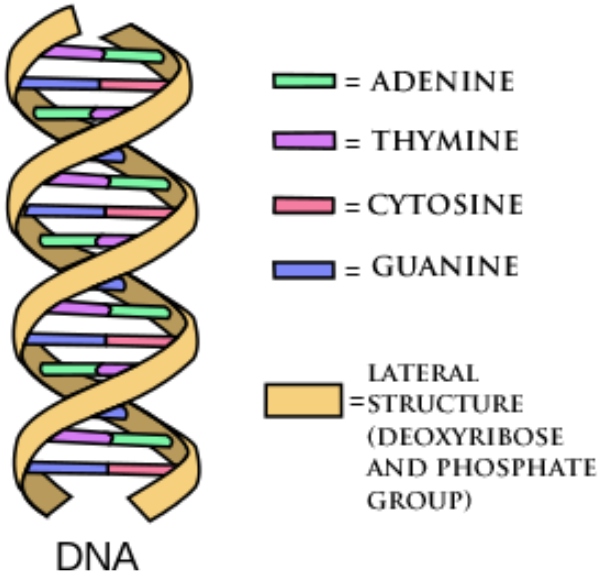

Figura 1.1: DNA structure

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 sepa-rated. A significant portion of DNA (more than 98 per cent 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.

Capitolo 2

Single Nucleotide Polymorphism: what is it?

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

For example, if we have two sequenced DNA fragments from different individuals (see *Picture 2.1*):

- AAGCCTA
- AAGCTTA

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

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

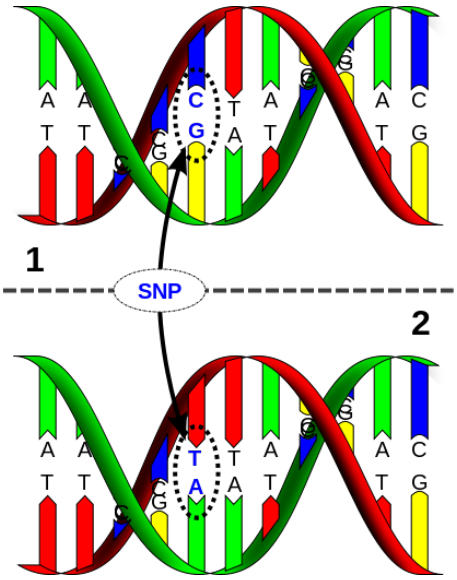

Figura 2.1: SNP example

2.1 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).

2.1.1 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

2.1.2 SNPs in the coding sequences

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

1. Synonymous
2. Nonsynonymous
    - Missense
    - 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	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 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 mutation 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** (beta-globin)
- **Hurler syndrome**

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

2.1.3 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*).

Capitolo 3

How to found SNPs: DNA Sequencing

CAPITOLO 2

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

For example, if we have two sequenced DNA fragments from different individuals (see *Picture 2.1*):

- AAGCCTA
- AAGCTTA

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

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

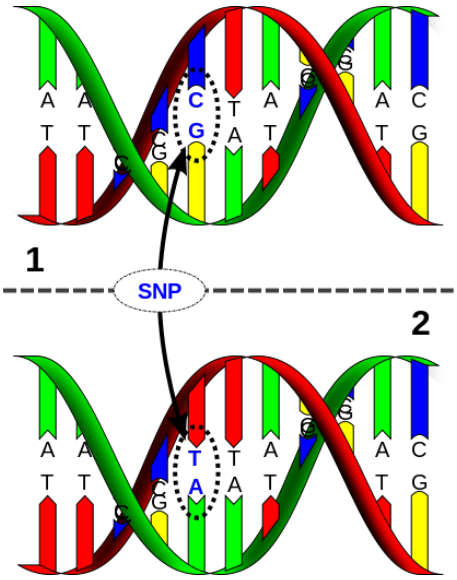

Figura 3.1: SNP example

3.1 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).

3.1.1 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

3.1.2 SNPs in the coding sequences

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

1. Synonymous
2. Nonsynonymous
    - Missense
    - 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	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 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 mutation 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** (beta-globin)
- **Hurler syndrome**

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

3.1.3 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*).