DNA representation

DNA representation

The complementarity between base pairs (**A = T** and **G ≡ C**) implies that if you know one sequence you can deduce the complementary sequence.

It is common to represent DNA sequences by 4-letter strings:

TGCTAATGCCGCTACTCTATCTGC

By convention, we write sequences **from 5' to 3'** end.

5'- TGCTAATGCCGCTACTCTATCTGC - 3'

Source: Jacques van Helden

DNA representation

Don't forget the second strand!

When we analyze a DNA sequence represented by

ATGCGCGGATG

we should keep in mind that the corresponding molecule is a double strand helix with the following base pairs:

```
5' - ATGCGCGGATG - 3'        (upper strand)
     |||||||||||
3' - TACGCGCCTAC - 5'        (lower strand)
```

Note that the concept of *upper strand* and *lower strand* are purely artificial. A DNA molecules is a 3D structure and there is no reason to consider preferentially one or the other strand.

This does not mean however that the two strands are functionally equivalent: in coding regions for example, only one strand will serve as a template for the synthesis of RNA.

Source: Jacques van Helden

DNA representation

Reverse complementarity

Reverse complementary sequences represent the two strands of the same DNA molecule.

The reverse complement is obtained by transposing each nucleotide into its complementary nucleotide (**A → T, T → A, C → G, G → C**), and then reversing the string.

For example the sequences **`ATGCGCGGATG`** and **`CATCCGCGCAT`** are mutually reverse complementary. These strings describe the two strands of the same DNA molecule. Consequently, the two following double strand schemes represent the same molecule:

```
5'- ATGCGCGGATG - 3'        5'- CATCCGCGCAT - 3'
    |||||||||||                 |||||||||||
3'- TACGCGCCTAC - 5'        3'- GTAGGCGCGTA - 5'
```

Source: Jacques van Helden

Symmetries in DNA sequences

Symmetries in DNA sequences

Tandem repeat

```
GATAAGATAAGATAAGATAA = 2 x GATAAGATAA = 4 x GATAA

GATAAGATAAatgtagGATAAGATAA = 2 x GATAAGATAA separated by
                                 a non repeated sequence.
```

Tandem repeats are presumed to occur frequently in genomic sequences, comprising perhaps **10%** or more of the human genome (Benson, *NAR* 27:573,1999).

Tandem repeat are sometimes associated to a **repeated structure of a protein** (Ex: some ABC transporters have been shown to contain a tandem repeat of six transmembrane helices, Tusnady et al, *FEBS Lett* 402:1-3,1997).

In recents years, the discovery of **short tandem repeat polymorphisms** are involved in various diseases (e.g. Cancer, Huntington, Parkinson,..., Zhang & Yu, *Eur J Surg Oncol*,33:529-34,2007).

ATP-binding Cassette (ABC) transporters

Source: Jacques van Helden

Symmetries in DNA sequences

Symmetries in DNA sequences

Textual palindromes

ATGGCCGGTA = ATGGC|CGGTA

Note that the corresponding DNA molecule does not contain any axis of symmetry since in 3D space a nucleotide cannot be superimposed on its own image. Therefore, searching for palindromes is not relevant for detecting biological features.

5'- ATGGC - 3' ≠ 5'- CGGTA - 3'

Source: Jacques van Helden

Symmetries in DNA sequences

Symmetries in DNA sequences

Reverse complementary palindromes

A reverse complementary palindrome is a sequence identical to its reverse complement.

Example: **ATGGGGCCCCAT**

Reverse complementary palindromes correspond to 3D symmetries in DNA molecules. In the following 2-strand representation, a 180° rotation around the center would swap the two strands, and each letter would take place of an identical letter on the complementary strand.

```
5'- ATGGGG CCCCAT - 3'
    ||||||.||||||
3'- TACCCC GGGGTA - 5'
```

Note that this sequence is not a textual palindrome.

Note that reverse complementary palindromes can be separated by a stretch of non-symmetrical nucleotides.

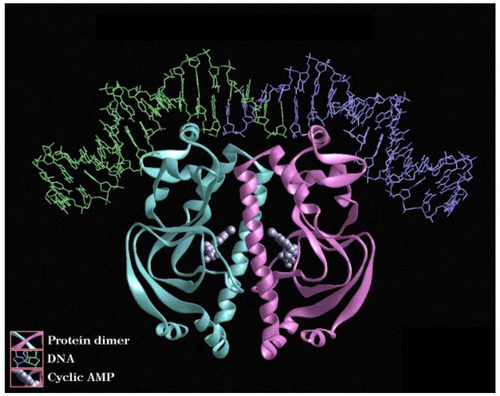
Source: Jacques van Helden

Symmetries in DNA sequences

Symmetries in DNA sequences

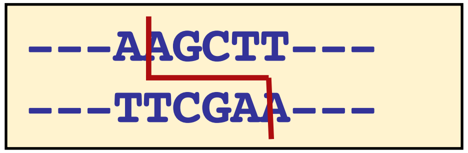
Reverse complementary motifs play important roles in biological mechanisms.

Example 1: some classes of transcription factors (e.g. *helix-turn-helix*) typically form homodimers whose tridimensional structure is symmetrical. These protein complexes specifically recognize reverse complementary motifs in gene promoters.



cAMP Receptor Protein (CRP)
TGTGA-N_6-TCACA

Example 2: In bacteria, hexamers with reverse complementary palindromic structure also play an essential role as recognition sites for *restriction enzymes*.



The restriction enzyme HindIII specifically cuts DNA at instance of AAGCTT

Symmetries in DNA sequences

Symmetries in DNA sequences

Reverse complementary motifs play important roles in biological mechanisms.

Example 3: Reverse complementary motifs separated by a stretch are frequent in RNA, where they mediate the pairing between distant segments of the molecules.

```
5'- UCGGGcucauaaCCCGA - 3'    --folding-->       a
                                               c   u
                                              u     a
                                               c   a
                                                GC
                                                GC
                                                GC
                                                CG     stem loop structure
                                                UA
                                                ||
                                                5'3'
```

Source: Jacques van Helden