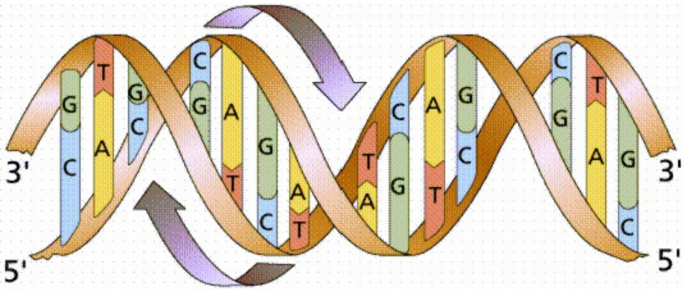
DNA compression

The human body is composed of 10 billion cells, and in the core of each of these cells is a complex set of instructions, which we call the *human genome*. The instructions are stored in a DNA chain which has been divided into 23 pairs of chromosomes. DNA can be represented by a string containing only the letters (bases) A,C,G,T and -. The hyphen represents an unknown base. The human genome contains a total of more than 3 billion bases.



double helix

Since the human genome consists of many repetitions, a DNA chain is usually stored in compressed form. When compressing a series of four or more consecutive identical characters are replaced by a code consisting of three characters: *i)* a hyphen (-), *ii)* a capital letter (A to Z) indicating 1 to 26 repetitions, and *iii)* the repeated character itself. Series of repetitions longer than 26 are replaced by multiple encoding, where all codes except the last one represent a repeated sequence of length 26. Each occurrence of a hyphen in the original DNA chain is encoded as a string of length 1 For example, the DNA chain ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA is encoded as ACCC-A-G-ET-ZA-DA.

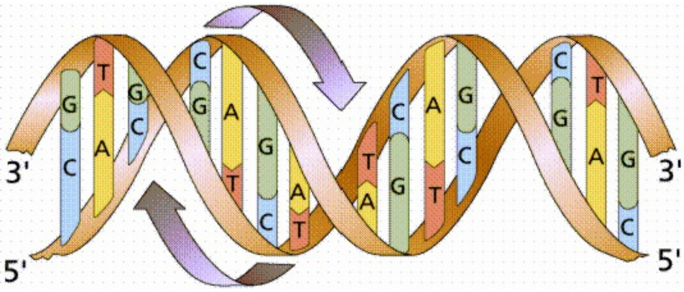
Assignment

1. Write a function DNAcompression that returns the compressed form of a given DNA chain as a result. The given DNA chain is to be passed to the function as parameter. The DNAcompression function should turn the given string ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA , into ACCC-A-G-ET-ZA-DA.
2. Write a function DNAdecompression that returns the original DNA chain for a given compressed DNA chain. The given compressed DNA chain is to be passed to the function a s parameter. For example ACCC-A-G-ET-ZA-DA should be returned as ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA.

Example

```
>>> DNAcompression('ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA')
'ACCC-A-G-ET-ZA-DA'
>>> DNAdecompression('ACCC-A-G-ET-ZA-DA')
'ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA'
```

Het menselijk lichaam bestaat uit 10 miljard cellen en in de kern van elk van deze cellen zit een complexe set van instructies die we het *menselijke genoom* noemen. De instructies zitten opgeslagen in een DNA-keten die is opgedeeld in 23 chromosoomparen. DNA kan voorgesteld worden door een string die enkel de letters (basen) A, C, G, T en - bevat. Hierbij stelt het koppelteken een onbekende base voor. Het menselijke genoom bestaat in totaal uit meer dan 3 miljard basen.



dubbele helix

Omdat het menselijk genoom uit heel wat herhalingen bestaat, wordt een DNA-keten meestal in gecomprimeerde vorm opgeslagen. Bij het comprimeren vervangt men een reeks van vier of meer opeenvolgende gelijke lettertekens door een code die bestaat uit drie lettertekens: *i)* een koppelteken (-), *ii)* een hoofdletter (A tot Z) die 1 tot 26 herhalingen aangeeft, en *iii)* het herhaalde letterteken zelf. Herhalingsreeksen langer dan 26 worden vervangen door meerdere coderingen, waarbij alle coderingen behalve de laatste een herhaalde reeks met lengte 26 voorstellen. Elk voorkomen van een koppelteken in de oorspronkelijke DNA-keten wordt gecodeerd als een reeks van lengte 1. Zo wordt de DNA-keten ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA bijvoorbeeld gecodeerd als ACCC-A-G-ET-ZA-DA.

Opgave

1. Schrijf een functie DNAcompressie die de gecomprimeerde vorm van een gegeven DNA-keten als resultaat teruggeeft. De gegeven DNA-keten moet als parameter aan de functie doorgegeven worden. Zo moet de functie DNAcompressie voor de gegeven string ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA, de string ACCC-A-G-ET-ZA-DA als resultaat teruggeven.
2. Schrijf een functie DNAdecompressie die voor een gegeven gecomprimeerde DNA-keten, de oorspronkelijke DNA-keten als resultaat teruggeeft. De gegeven gecomprimeerde DNA-keten moet als parameter aan de functie doorgegeven worden. Zo moet de functie DNAdecompressie voor de gegeven string ACCC-A-G-ET-ZA-DA, de string ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA als resultaat teruggeven.

Voorbeeld

```
>>> DNAcompressie('ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA')
'ACCC-A-G-ET-ZA-DA'
>>> DNAdecompressie('ACCC-A-G-ET-ZA-DA')
'ACCC-GTTTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA'
```