GC content

For this task we ask to write a number of Python functions that can be used to generate an overview with the *% GC* of any *DNA sequence* from a given file in *FASTA format*. Terms used in the preceding sentence that were displayed in italic font are explained below in detail.

DNA sequence

A DNA sequence can be represented by a string consisting only of the letters A, C, G and T (but in practice also other letters are used, that represent uncertain base pairs or gaps in the sequence). The% GC is calculated as the ratio of the number of letters G and C in the string, with respect to the total number of A's, C's, G's, and T's (other letters are ignored). For example, for GCCTGCAG the %GC = 75%.

%GC

In genetics, the guanine-cytosine content (%GC) is a characteristic property of the genome of a given organism, or any other piece of DNA or RNA. Normally, this property is expressed as a percentage, and gives the ratio of the GC base pairs in the DNA molecule or genome sequence that is being examined. G stands for guanine and C for cytosine. The remaining base-pairs of a DNA molecule will then consist of the bases A (adenine) and T (thymine), so that the calculation of the%GC in an indirect manner provides the calculation of the%AT (% GC = 58% means% AT = 42%) as well. GC base pairs in the DNA are bound to three hydrogen bonds, rather than two in the case of the AT base pairs. This ensures that GC pairs are stronger and more resistant to denaturation at high temperatures, so that the %GC usually is greater in hyperthermophiles.

FASTA format

In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences.

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). There should be no space between the ">" and the first letter of the identifier. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence. A simple example of one sequence in FASTA format:

```
>118480563|DQ207729|Bacillus cereus|16S ribosomal RNA gene
AGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAATGGATTA
AGAGCTTGCTCTTATGAAGTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTGCCCATAAGACTGGG
ATAACTCCGGGAAACCGGGGCTAATACCGGATAACATTTTGAACCGCATGGTTCGAAATTGAAAGGCGGC
TTCGGCTGTCACTTATGGATGGACCCGCGTCGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCAA
CGATGCGTA
```

The description line - the line that begins with a ">" - contains the following fields, which are separated by a vertical bar (|): *i)* unique identifier (database-specific), *ii) accession number* (universally unique identifier) , *iii)* generic name *iv)* description. An example of a FASTA file

with multiple sequences, generated for the GenBank database of the NCBI, looks like this:

```
>571435|U16165|Clostridium acetobutylicum|16S ribosomal RNA gene
TGGCGGCGTGCTTAACACATGCAAGTCGAGCGATGAAGCTCCTTCGGGAGTGGATTAGCGGCGGACGGGT
GAGTAACACGTGGGTAACCTGCCTCATAGAGGGGAATAGCCTTTCGAAAGGAAGATTAATACCGCATAAG
ATTGTAGTGCCGCATGGCATAGCAATTAAAGGAGTAATCCGCTATGAGATGGACCCGCGTCGCATTAGCT
AGTTGGTGAGGTAACGGCTCACCAAGGCGACGATGCGTAGCCGACCTGAGAGGGTGATCGGCCACATTGG
GACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTG
>996091|L07834|Geobacter metallireducens|16S ribosomal RNA gene
AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGAGTGCCTAACACATGCAAGTCGAACGTGAAGGGGG
CTTCGGTCCCCGGAAAGTGGCGCACGGGTGAGTAACGCGTGGATAATCTGCCCAGTGATCTGGGATAACA
TCTCGAAAGGGGTGCTAATACCGGATAAGCCCACGGAGTCCTTGGATTCTGCGGGAAAAGGGGGGGACCT
TCGGGCCTTTTGTCACTGGATGAGTCCGCGTACCATTAGCTAGTTGGTGGGGTAATGGCCCACCAAGGCT
ACGATGGTTAG
```

After each description line one or more lines that describe the sequence follow. Sequences can represent both DNA sequences and protein sequences, and they can contain holes that are represented by a minus sign (-).

Assignment

1. Write function readFasta to which a file location must be passed as an argument. At this location a text file must be retrievable, that contains one or more DNA sequences in FASTA format. As a result, the function should return a list, that for each record from the FASTA file contains a tuple containing the *accession number*, the generic name and the DNA sequence.
2. Write a percentGC function that for a given DNA sequence - that is to be passed to the function as an argument - calculates the% GC and returns it as a real value.
3. Write a function showOverview, to which a list of tuples as generated by the function readFasta must be passed as an argument. This function should display a list that first of all writes a line for each sequence from the given FASTA file (read as a list of tuples) bearing the generic name (for which 30 characters reserved, left aligned), followed by the %GC of the sequence (rounded off to two decimal places), a space and the *accession number* in parentheses. This is followed by a blank line, and successively also the minimum, maximum and mean %GC of all sequences are printed on separate lines. See the examples below for an illustration of the format in which the overview is to be displayed.

Example

This interactive Python session uses the file seq1.fasta.

```
>>> fasta = readFasta('seq1.fasta')
>>> fasta
[('ABCDE', 'elephant', 'AGAGTTTGATAGAGCTTGCT'), ('FGHIJ', 'donkey', 'GAACGCTGGCGGCATGCCTT')]
>>> percentGC('AGAGTTTGATAGAGCTTGCT')
40.0
>>> showOverview(fasta)
elephant                    40.00% (ABCDE)
donkey                      65.00% (FGHIJ)

minimum                     40.00%
maximum                     65.00%
```

```
mean                52.50%
```

This interactive Python session uses the file seq2.fasta.

```
>>> fasta = readFasta('seq2.fasta')
>>> showOverview(fasta)
Bacillus cereus              53.49% (DQ207729)
Burkholderia xenovorans      56.41% (U86373)
Clostridium acetobutylicum   52.68% (U16165)
Geobacter metallireducens    56.40% (L07834)
Listeria welshimeri          53.61% (X98532)
Methanosarcina acetivorans   56.63% (M59137)
Oceanobacillus iheyensis     52.85% (AB010863)
Thermus thermophilus         63.96% (X07998)
Xanthomonas campestris       55.13% (X95917)
Bacillus sporothermodurans   54.38% (U49078)

minimum                 52.68%
maximum                 63.96%
mean                55.55%
```

Voor deze opgave wordt gevraagd om een aantal Python functies te schrijven die kunnen gebruikt worden om een overzicht te genereren met het *%GC* van elke *DNA sequentie* uit een gegeven bestand in *FASTA formaat*. Termen die in voorgaande zin met een cursief lettertype werden weergegeven worden hierna in detail uitgelegd.

DNA sequentie

Een DNA sequentie kan voorgesteld worden door een string die enkel bestaat uit de letters A, C, G en T (maar in de praktijk worden ook nog andere letters gebruikt, die onzekere baseparen of gaten aangeven in de sequentie). Het %GC wordt berekend als de verhouding van het aantal letters G en C in de string, ten opzichte van het totaal aantal A's, C's, G's en T's (andere letters worden genegeerd). Voor GCCTGCAG is bijvoorbeeld het %GC = 75%.

%GC

In de genetica is de guanine-cytosine inhoud (%GC) een karakteristieke eigenschap van het genoom van een gegeven organisme of van elk ander stuk DNA of RNA. Normaalgezien wordt deze eigenschap uitgedrukt als een percentage, en geeft ze de verhouding weer van de GC baseparen in de DNA molecule of genoomsequentie die onderzocht wordt. G staat hierbij voor guanine en C voor cytosine. De overblijvende baseparen van een DNA molecule zullen dan bestaan uit de basen A (adenine) en T (thymine), zodat de berekening van het %GC op een indirecte manier ook de berekening van het %AT oplevert (%GC = 58% betekent %AT = 42%). GC-baseparen zijn in het DNA verbonden met drie waterstofbindingen in plaats van twee in het geval van de AT-baseparen. Dit zorgt ervoor dat GC-paren sterker en beter resistent zijn tegen denaturatie bij hoge temperaturen, waardoor het %GC dus meestal groter is bij hyperthermofielen.

FASTA formaat

FASTA is een tekstgebaseerd bestandsformaat dat gebruikt wordt in de bioinformatica om DNA of eiwitsequenties op te slaan. Individuele baseparen of eiwitresidu's worden daarbij voorgesteld door één-letter codes. Het formaat laat ook toe om de verschillende sequenties te laten voorafgaan door sequentienamen en andere informatievelden.

Een sequentie in FASTA formaat begint met een één-regel beschrijving, gevolgd door de

eigenlijke sequentiegegevens die eventueel kunnen gesplist worden over verschillende regels. De regel met de beschrijving wordt onderscheiden van de sequentiegegevens door een "'groter dan"' symbool (>) in de eerste kolom. Elke sequentie eindigt waar een nieuwe regel begint met een >-karakter, wat de start van een nieuwe sequentie aangeeft, of op het einde van het bestand. Een eenvoudig voorbeeld van één enkele sequentie in FASTA formaat:

```
>118480563|DQ207729|Bacillus cereus|16S ribosomal RNA gene
AGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAATGGATTA
AGAGCTTGCTCTTATGAAGTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTGCCCATAAGACTGGG
ATAACTCCGGGAAACCGGGGCTAATACCGGATAACATTTTGAACCGCATGGTTCGAAATTGAAAGGCGGC
TTCGGCTGTCACTTATGGATGGACCCGCGTCGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCAA
CGATGCGTA
```

De beschrijvingsregel — de regel die begint met een >-karakter — bevat de volgende informatievelden, die van elkaar gescheiden worden door een verticale streep (|): *i*) unieke identifier (databank-specifiek), *ii*) *accession number* (universeel unieke identifier), *iii*) soortnaam en *iv*) omschrijving. Een voorbeeld van een FASTA bestand met meerdere sequenties, gegenereerd voor de GenBank databank van het NCBI, ziet er als volgt uit:

```
>571435|U16165|Clostridium acetobutylicum|16S ribosomal RNA gene
TGGCGGCGTGCTTAACACATGCAAGTCGAGCGATGAAGCTCCTTCGGGAGTGGATTAGCGGCGGACGGGT
GAGTAACACGTGGGTAACCTGCCTCATAGAGGGGAATAGCCTTTCGAAAGGAAGATTAATACCGCATAAG
ATTGTAGTGCCGCATGGCATAGCAATTAAAGGAGTAATCCGCTATGAGATGGACCCGCGTCGCATTAGCT
AGTTGGTGAGGTAACGGCTCACCAAGGCGACGATGCGTAGCCGACCTGAGAGGGTGATCGGCCACATTGG
GACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTG
>996091|L07834|Geobacter metallireducens|16S ribosomal RNA gene
AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGAGTGCCTAACACATGCAAGTCGAACGTGAAGGGGG
CTTCGGTCCCCGGAAAGTGGCGCACGGGTGAGTAACGCGTGGATAATCTGCCCAGTGATCTGGGATAACA
TCTCGAAAGGGGTGCTAATACCGGATAAGCCCACGGAGTCCTTGGATTCTGCGGGAAAAGGGGGGGACCT
TCGGGCCTTTTGTCACTGGATGAGTCCGCGTACCATTAGCTAGTTGGTGGGGTAATGGCCCACCAAGGCT
ACGATGGTTAG
```

Na elke beschrijvingsregel volgen één of meerdere regels die de sequentie beschrijven. Sequenties kunnen zowel DNA-sequenties als eiwitsequenties voorstellen, en ze kunnen gaten bevatten die worden voorgesteld door een minteken (-).

Opgave

1. Schrijf een functie leesFasta waaraan een bestandslocatie als argument moet doorgegeven worden. Op deze locatie moet een tekstbestand terug te vinden zijn, dat één of meerdere DNA sequenties in FASTA formaat bevat. Als resultaat moet de functie een lijst teruggeven, die voor elke record uit het FASTA bestand een tuple bevat met daarin het *accession number*, de soortnaam en de DNA sequentie.
2. Schrijf een functie procentGC die voor een gegeven DNA sequentie — die als argument aan de functie moet doorgegeven worden — het %GC berekent en als een reële waarde teruggeeft.
3. Schrijf een functie toonOverzicht waaraan een lijst van tuples zoals die gegenereerd wordt door de functie leesFasta als argument moet doorgegeven worden. Deze functie moet een

overzicht weergeven dat in eerste instantie voor elke sequentie uit het gegeven FASTA bestand (ingelezen als een lijst van tuples) een regel uitschrijft met daarop de soortnaam (waarvoor 30 lettertekens gereserveerd worden, links uitgelijnd), gevolgd door het %GC van de sequentie (afgerond tot op twee cijfers na de komma), een spatie en het *accession number* tussen ronde haakjes. Daarna volgt een lege regel, en worden achtereenvolgens ook het minimum, maximum en gemiddelde %GC van alle sequenties op afzonderlijke regels uitgeschreven. Zie onderstaande voorbeelden voor een illustratie van het formaat waarin het overzicht moet weergegeven worden.

Voorbeeld

Onderstaande interactieve Pythonsessie maakt gebruik van het bestand seq1.fasta.

```
>>> fasta = leesFasta('seq1.fasta')
>>> fasta
[('ABCDE', 'olifant', 'AGAGTTTGATAGAGCTTGCT'), ('FGHIJ', 'ezel', 'GAACGCTGGCGGCATGCCTT')]
>>> procentGC('AGAGTTTGATAGAGCTTGCT')
40.0
>>> toonOverzicht(fasta)
olifant                 40.00% (ABCDE)
ezel                    65.00% (FGHIJ)

minimum                   40.00%
maximum                   65.00%
gemiddelde                52.50%
```

Onderstaande interactieve Pythonsessie maakt gebruik van het bestand seq2.fasta.

```
>>> fasta = leesFasta('seq2.fasta')
>>> toonOverzicht(fasta)
Bacillus cereus              53.49% (DQ207729)
Burkholderia xenovorans      56.41% (U86373)
Clostridium acetobutylicum   52.68% (U16165)
Geobacter metallireducens    56.40% (L07834)
Listeria welshimeri          53.61% (X98532)
Methanosarcina acetivorans   56.63% (M59137)
Oceanobacillus iheyensis     52.85% (AB010863)
Thermus thermophilus         63.96% (X07998)
Xanthomonas campestris       55.13% (X95917)
Bacillus sporothermodurans   54.38% (U49078)

minimum                   52.68%
maximum                   63.96%
gemiddelde                55.55%
```