Sequence logo

A *sequence logo* is a graphical representation for a multiple alignment of DNA, RNA or protein sequences, from which regions with conserved residues can immediately be read. Sequence logos are, for example, used to detect conserved regions in the DNA because there, transcription factors can bind.

Sequence logo that represents the most conserved bases around the start codon of all human mRNAs. Note that the start codon itself is not shown to scale, because otherwise the letter AUG would each have a height of 2 bits.

In making sequence logos one starts from related DNA, RNA or protein sequences, or from DNA sequences with conserved binding regions. In a first step, these sequences are aligned relative to each other, wherein the most conserved residues are put under each other. Thereafter, the frequency of the residues is calculated per position in this multiple alignment. The sequence logo shows for each position how good the residues are conserved: the more residues of a particular type, the higher the letter, because the preservation of the residue at that position is larger. The letters of the residues at the same position are scaled according to their frequency. The height of all the letters in the same position corresponds to the information expressed in bits (entropy).

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
```

Below is an example of a FASTA file with multiple sequences. Note that in this case, there are

multiple description lines - lines that start with a ">" - which indicate that a new sequence begins thereafter.

```
>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

- Write a function readFasta with which the sequences from a FASTA file can be read. The location of the FASTA file must be passed to the function as string-argument. The function must return a list of strings, in which the successive strings correspond to the consecutive sequences as they are listed in the file. A sequence that was split over multiple lines in the file must be displayed in the list as a single string with no whitespace.
- Write a function sequenceLogo to which a list of aligned sequences must be passed. This includes that all the strings in the list have the same length. See the example below to see how the function should react if this condition is not met. The second optional argument alphabet (default ACGT) can pass another string to the function, in which the residue letters are listed from which the sequences exist. See the example below to see how the function should respond, if there is residue in given sequences that does not belong to the given alphabet. The function cannot distinguish between uppercase and lowercase letters in the letter representation of the residues, nor in the given sequences, nor the given alphabet.
  The function should return a list of which each element at position i is itself a list containing the frequencies of residues at position i in the aligned sequences. The order of the residues in each frequency table is the same as the order of the residues in the given alphabet.

Example

In the following example we assume that the file seq.fasta is in the current directory.

```
>>> sequences = readFasta('seq.fasta')
>>> sequences
['ATG', 'gtg', 'CTA', 'TTa', 'ATG']

>>> sequenceLogo(sequences)
[[0.4, 0.2, 0.2, 0.2], [0.0, 0.0, 0.0, 1.0], [0.4, 0.0, 0.6, 0.0]]

>>> sequenceLogo(sequences, alphabet='gcta')
[[0.2, 0.2, 0.2, 0.4], [0.0, 0.0, 1.0, 0.0], [0.6, 0.0, 0.0, 0.4]]
```

```
>>> sequenceLogo(sequences, alphabet='GCUA')
Traceback (most recent call last):
AssertionError: invalid residu

>>> sequenceLogo(['AGCTGC', 'TCGT', 'CGTATGATAG'])
Traceback (most recent call last):
AssertionError: not all sequences have the same length
```

The construction of the first sequence logo from the above example is shown in the table below. Such a table - for DNA sequences sometimes shortened to the last four columns - is referred to as *position-specific scoring matrix* in bioinformatics.

seq1	seq2	seq3	seq4	seq5	A	C	G	T
A	g	C	T	A	0.4	0.2	0.2	0.2
T	t	T	T	T	0.0	0.0	0.0	1.0
G	g	A	a	G	0.4	0.0	0.6	0.0

Een *sequentielogo* is een grafische voorstelling voor een meervoudige aligning van DNA-, RNA- of eiwitsequenties, waaruit onmiddellijk de regio's met geconserveerde residu's kunnen afgelezen worden. Sequentielogo's worden bijvoorbeeld gebruikt om geconserveerde regio's in het DNA op te sporen, omdat daar transcriptiefactoren kunnen binden.

Sequentielogo dat de meest geconserveerde basen rond het startcodon van alle menselijke mRNAs weergeeft. Merk op dat het startcodon zelf niet op schaal wordt weergegeven, want anders zouden de letter AUG elk een hoogte van 2 bits hebben.

Bij het maken van sequentielogo's vertrekt men van gerelateerde DNA-, RNA- of eiwitsequenties, of van DNA-sequenties met geconserveerde bindingsregio's. In een eerste stap worden deze sequenties ten opzichte van elkaar gealigneerd, waarbij de meest geconserveerde residu's onder elkaar gezet worden. Daarna wordt per positie in deze meervoudige aligning de frequentie van de residu's berekend. Het sequentielogo geeft voor elke positie weer hoe goed de residu's geconserveerd zijn: hoe meer residu's van een bepaald type, hoe groter de letter, omdat de conservering van het residu op die positie groter is. De letters van de residu's op eenzelfde positie worden geschaald volgens hun frequentie. De hoogte van alle letters op eenzelfde positie komt overeen met de informatie uitgedrukt in bits (entropie).

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 gesplitst 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
```

Hieronder staat een voorbeeld van een FASTA bestand met meerdere sequenties. Let op het feit dat er in dit geval meerdere beschrijvingsregels — regels die starten met een >-karakter — zijn, die aangeven dat er daarna een nieuwe sequentie begint.

```
>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-, RNA- of eiwitsequenties voorstellen, en ze kunnen gaten bevatten die worden voorgesteld door een minteken (-).

Opgave

- Schrijf een functie `leesFasta` waarmee de sequenties uit een FASTA bestand kunnen uitgelezen worden. De locatie van het FASTA bestand moet als string-argument aan de functie doorgegeven worden. De functie moet als resultaat een lijst van strings teruggeven, waarbij de opeenvolgende strings corresponderen met de opeenvolgende sequenties zoals ze in het bestand staan opgelijst. Een sequentie die over verschillende regels gesplitst werd in het bestand moet in de lijst als één enkele string zonder witruimte weergegeven worden.
- Schrijf een functie `sequentieLogo` waaraan een lijst van gealigneerde sequenties moet doorgegeven worden. Hierbij moeten alle strings uit de lijst dus dezelfde lengte hebben. Bekijk onderstaand voorbeeld om te zien hoe de functie moet reageren als deze voorwaarde niet voldaan is. Als tweede optionele argument `alfabet` (standaardwaarde `ACGT`) kan aan de functie nog een string meegegeven worden, waarin de residuletters staan opgelijst waaruit de sequenties bestaan. Bekijk onderstaand voorbeeld om te zien hoe de functie moet reageren als er in de gegeven sequenties residu's voorkomen die niet tot het gegeven alfabet behoren. De functie mag geen onderscheid maken tussen hoofdletters en kleine letters bij de lettervoorstelling van de residu's, noch in de gegeven sequenties, noch

in het gegeven alfabet.
De functie moet als resultaat een lijst teruggeven, waarvan elk element op positie i zelf ook een lijst is die de frequenties bevat van de residu's op positie i in de gealigneerde sequenties. De volgorde van de residu's in elke frequentielijst is dezelfde als de volgorde van de residu's in het gegeven alfabet.

Voorbeeld

Bij onderstaand voorbeeld gaan we ervan uit dat het bestand seq.fasta zich in de huidige directory bevindt.

```
>>> sequenties = leesFasta('seq.fasta')
>>> sequenties
['ATG', 'gtg', 'CTA', 'TTa', 'ATG']

>>> sequentieLogo(sequenties)
[[0.4, 0.2, 0.2, 0.2], [0.0, 0.0, 0.0, 1.0], [0.4, 0.0, 0.6, 0.0]]

>>> sequentieLogo(sequenties, alfabet='gcta')
[[0.2, 0.2, 0.2, 0.4], [0.0, 0.0, 1.0, 0.0], [0.6, 0.0, 0.0, 0.4]]

>>> sequentieLogo(sequenties, alfabet='GCUA')
Traceback (most recent call last):
AssertionError: ongeldig residu

>>> sequentieLogo(['AGCTGC', 'TCGT', 'CGTATGATAG'])
Traceback (most recent call last):
AssertionError: niet alle sequenties hebben dezelfde lengte
```

De constructie van het eerste sequentielogo uit bovenstaand voorbeeld wordt weergegeven in onderstaande tabel. Een dergelijke tabel — voor DNA sequenties soms ook ingekort tot de laatste vier kolommen — wordt in de bioinformatica een *position-specific scoring matrix* genoemd.

seq1	seq2	seq3	seq4	seq5	A	C	G	T
A	g	C	T	A	0.4	0.2	0.2	0.2
T	t	T	T	T	0.0	0.0	0.0	1.0
G	g	A	a	G	0.4	0.0	0.6	0.0