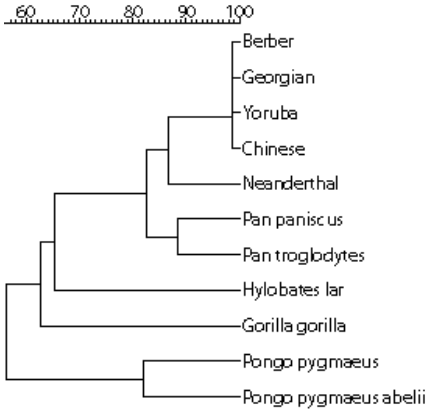
Neanderthals

The discovery of skeletons of the Neanderthal (*Homo sapiens neanderthalensis*) has raised many questions about the human origin in different parts of Europe, including the search for our relationship to this species. Such questions about the origin of humans and primates are normally answered by studying the mitochondrial DNA, and in particular the hypervariable regions within them. These regions show high variability within the human race, making them ideal for the study of relationships between individuals. The mitochondrial DNA exhibits two hypervariable regions, which are designated as HVR-I and HVR-II, respectively.



phylogeny of some anthropoids

In order to determine the similarity (measure of kindship) of DNA sequences, first of all an alignment of all investigated sequences is constructed. In such a sequence alignment the corresponding parts of two or more sequences are placed underneath each other. Below is an example of such a sequence alignment between two sequences.

```
CTG-GGG--GGTGTAC
||  |||   | || |
CTACGGG---GCGTCC
```

Here, the corresponding base pairs (matches) are indicated by vertical lines between the base pairs, and holes are represented by hyphens (-) within the sequences (causing the aligned sequences to always have the same length). Errors (mismatches) can be explained by point mutations and holes (gaps) by insertions or deletions.

On the basis of a sequence alignment, the similarity between two sequences can be calculated in the following way. For corresponding base pairs, a score of +1 is awarded, for errors a score of 0 and for holes a score of -1. Positions where both sequences have a gap will not be taken into account. The scores for each position of the alignment are then added together and divided by the number of positions brought into account. In the above example, there are 9 pairs of corresponding base pairs, 3 errors, 2 holes, and 14 of the 16 positions are taken into account (two positions exhibit a hole in both sequences). The similarity of these two sequences equals $$ \frac{9 \times (+1) + 3 \times 0 + 2 \times (-1)}{14} = 0.5 $$

Assignment

1. Write a function `score`, that returns the corresponding score for two given base pairs (that must be passed to the function as argument) based on the values from the table below:

type	example	score
match	A en A	+1
mismatch	A en G	0
1 hole	A en -	-1
2 holes	- en -	0

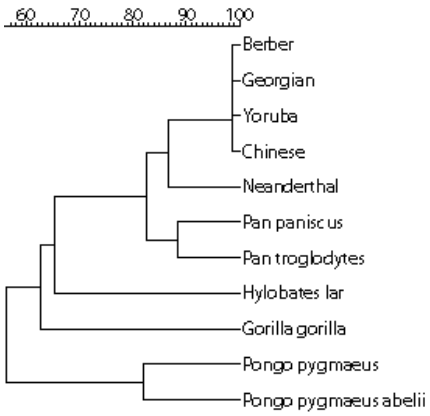
2. Use the function score to write a function similarity that calculates the corresponding similarity for two given DNA sequences of a sequence alignment as was described above. This function should return the value of the similarity. For example, the function should return the value 0.5 for the DNA sequences "CTG-GGG--GGTGTAC" and "CTACGGG---GCGTCC" from the example above.

Note: If you have implemented the similarity function properly it will be used to establish a phylogenetic tree on the basis of the mitochondrial DNA of some primates. Based on this result, can you find out whether modern humans evolved from Neanderthals?

Example

```
>>> similarity('CTG-GGG--GGTGTAC', 'CTACGGG---GCGTCC')
0.5
>>> similarity('CTG-GGG---GTGTAC', 'CTACGGG---GCGTCC')
0.6153846153846154
>>> similarity('CTG-GGG--GGTGTAC', 'CTACGGG---GCGTCA')
0.42857142857142855
```

De ontdekking van skeletten van de Neanderthaler (*Homo sapiens neanderthalensis*) in verschillende delen van Europa heeft veel vragen opgeroepen omtrent de menselijke oorsprong, waaronder de zoektocht naar onze verwantschap met deze soort. Dergelijke vragen over de oorsprong van mensen en primaten worden doorgaans beantwoord door het bestuderen van het mitochondriaal DNA en in het bijzonder de hypervariabele gebieden daarbinnen. Deze gebieden vertonen een hoge variabiliteit binnen het menselijk ras, waardoor ze ideaal zijn voor de studie van verwantschappen tussen individuen. Het mitochondriaal DNA vertoont twee hypervariabele gebieden, die respectievelijk worden aangeduid als HVR-I en HVR-II.



fylogenie van enkele mensapen

Om de similariteit (maat van verwantschap) van DNA sequenties te bepalen, construeert men meestal eerst een alignering van alle onderzochte sequenties. Bij een dergelijke sequentiealignering worden de overeenkomstige onderdelen van twee of meer sequenties onder elkaar geplaatst. Hieronder staat een voorbeeld van een dergelijke sequentiealignering tussen twee sequenties.

```
CTG-GGG--GGTGTAC
||  |||   | || |
CTACGGG---GCGTCC
```

Hierbij worden de overeenkomstige baseparen (*matches*) aangegeven met verticale strepen tussen de baseparen, en worden gaten voorgesteld door koppeltekens (-) binnen de sequenties (waardoor de gealigneerde sequenties altijd even lang zijn). Fouten (*mismatches*) kunnen verklaard worden door puntmutaties en gaten (*gaps*) door inserties of deleties.

Op basis van een sequentiealignering kan de similariteit tussen twee sequenties op de volgende manier berekend worden. Voor overeenkomstige baseparen wordt een score van +1 aangerekend, voor fouten een score van 0 en voor gaten een score van -1. Posities waarop beide sequenties een gat vertonen worden niet in rekening gebracht. De scores voor elke positie van de alignering worden dan bij elkaar opgeteld, en gedeeld door het aantal in rekening gebrachte posities. In het bovenstaande voorbeeld zijn er 9 overeenkomstige baseparen, 3 fouten, 2 gaten en worden in totaal 14 van de 16 posities in rekening gebracht (twee posities vertonen een gat in beide sequenties). Hierdoor is de similariteit van deze twee sequenties gelijk aan $$ \frac{ 9 \times (+1) + 3 \times 0 + 2 \times (-1) }{14} = 0.5 $$

Opgave

1. Schrijf een functie `score`, die voor twee gegeven baseparen (die als argument aan de functie moeten doorgegeven worden) de corresponderende score teruggeeft op basis van de waarden uit de onderstaande tabel:

type	voorbeeld	score
overeenkomstige baseparen	A en A	+1
verschillende baseparen	A en G	0
1 gat	A en -	-1
2 gaten	- en -	0

2. Gebruik de functie `score` om een functie `similarity` te schrijven, die voor twee gegeven DNA sequenties uit een sequentiealignering de corresponderende similariteit berekent zoals hierboven werd beschreven. Deze functie moet de waarde van de similariteit als resultaat teruggeven. Zo moet de functie voor de DNA sequenties "`CTG-GGG--GGTGTAC`" en "`CTACGGG---GCGTCC`" uit bovenstaand voorbeeld de waarde 0.5 teruggeven.

Opmerking: Als je de functie `similarity` correct hebt gemaakt, dan wordt die gebruikt om een fylogenetische stamboom op te stellen op basis van het mitochondriaal DNA van enkele primaten. Kan je op basis van dit resultaat achterhalen of de moderne mens geëvolueerd is uit de Neanderthaler?

Voorbeeld

```
>>> similarity('CTG-GGG--GGTGTAC', 'CTACGGG---GCGTCC')
0.5
>>> similarity('CTG-GGG---GTGTAC', 'CTACGGG---GCGTCC')
0.6153846153846154
>>> similarity('CTG-GGG--GGTGTAC', 'CTACGGG---GCGTCA')
0.42857142857142855
```