Calculating mean quality scores for FASTQ records

Quality scores in FASTQ records are encoded as ASCII characters. As shown in the fourth line of the FASTQ record below.

```
@chr1_245318753_245318470_0_0_0_0_0:0:0_0:0:0_0/1
CTGAAATTTCATGAGGCCAGAGTAGGTCAACAGGGGCTAATGAGACAAGATCTAACTAAACAGAGTAAAGAGTTTCTCGGTCAC
+
I?>DC:>@?IDC9??G?>EH9E@66=9<?@E?DC:@<@BBFG>=FIC@F9>7CG?IC?I;CD9>>>A@C7>>8>>D9GCB<;?D
```

These ASCII characters directly correspond to a number from 0-93. These are the Phred scores. Each phred score stands for a probability

To calculate the probability P for any phred score x:

$$P(x) = 10^{\frac{x}{-10}}$$

Phred scores can not be averaged naively. For instance a score of 10 and 30 do not average 20. 10 stands for $\frac{1}{10^{1.0}} = 0.1$ and 30 for $\frac{1}{10^{3.0}} = 0.001$. Averaging these probabilities gives 0.0505. $-10 \cdot^{10} log(0.505) = 12.97$.

To calculate the average quality score for an entire record we can build a formula. Starting with the formula for a single ASCII character.

$$P(x) = 10^{\frac{x - offset}{-10}}$$

Where offset is the Phred offset. (+33 is added to the phred scores to push them into the ASCII printable range).

To calculate the average probability P for all quality scores a in a vector we take the sum of all the probabilities and divide it by the number of characters:

$$P_{average} = \frac{1}{n} \sum_{i=1}^{n} 10^{\frac{a_i - offset}{-10}}$$

To calculate the phred score for the average P.

$$Phred_{average} = -10 *^{10} log\left(P_{average}\right)$$

The entire formula for calculating the average phred score from the base qualities.

$$Phred_{average} = -10 *^{10} log\left(\frac{1}{n} \sum_{i=1}^{n} 10^{\frac{a_i - offset}{-10}}\right)$$

It can be implemented in python with numpy as follows:

```python
import math
import numpy as np

def qualmean(qualities: bytes, phred_offset: int = 33) -> float:
    phred_scores = np.frombuffer(qualities, dtype=np.int8)
    probabilities = np.power(10, ((phred_scores - phred_offset) / -10))
    average = np.average(probabilities)
    return -10 * math.log10(average)
```

This requires three operations on the array containing the quality scores.

- Subtracting the phred_offset.

- Dividing by -10
- 10 to the power of each value in the array.

We can simplify the formula to reduce the number of calculations. We can use the following math rule $a^{pq} = (a^p)^q$ to remove the division by -10. This is the same as multiplying with -0.1. We can write $10^{\frac{a_i - offset}{-10}}$ as $\left(10^{-\frac{1}{10}}\right)^{a_i - offset}$. Where we can calculate $10^{-\frac{1}{10}}$ first as a constant C. Reducing the number of total calculations.

$$P_{average} = \frac{1}{n}\sum_{i=1}^{n}\left(10^{-\frac{1}{10}}\right)^{a_i - offset}$$

$$P_{average} = \frac{1}{n}\sum_{i=1}^{n} C^{a_i - offset}$$

where $C = 10^{-\frac{1}{10}}$.

We can use $a^{p-q} = \frac{a^p}{a^q}$ to get rid of the `-offset` calculation for each base.

$$P_{average} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{C^{a_i}}{C^{offset}}\right)$$

Since $\frac{a}{c} + \frac{b}{c} = \frac{a+b}{c}$ we can move the offset outside of the sum

$$P_{average} = \frac{1}{C^{offset}} \cdot \frac{1}{n}\sum_{i=1}^{n} C^{a_i}$$

We now have eliminated the need to subtract the offset and divide by -10 for each element in the array of scores. The total formula for the average phred score now looks as follows:

$$Phred_{average} = -10 *^{10} log\left(\frac{1}{C^{offset}} \cdot \frac{1}{n}\sum_{i=1}^{n} C^{a_i}\right)$$

since $log(a/b) = log(a) - log(b)$ we can bring the offset outside of the log.

$$Phred_{average} = -10 * \left({}^{10}log\left(\frac{1}{n}\sum_{i=1}^{n} C^{a_i}\right) -^{10} log(C^{offset})\right)$$

We can simplify further: ${}^{10}log(C^{offset})$ equals ${}^{10}log\left(\left(10^{-\frac{1}{10}}\right)^{offset}\right)$ equals ${}^{10}log\left(10^{-offset/10}\right)$ and the log is cancelled out so $-\frac{offset}{10}$

$$Phred_{average} = -10 * \left({}^{10}log\left(\frac{1}{n}\sum_{i=1}^{n} C^{a_i}\right) - -\frac{offset}{10}\right)$$

`--` becomes `+`. Also we can remove the braces by multiplying both terms in the braces with `-10`

$$Phred_{average} = -10 *^{10} log\left(\frac{1}{n}\sum_{i=1}^{n} C^{a_i}\right) - offset$$

$$Phred_{average} = -10 *^{10} log\left(\frac{1}{n}\sum_{i=1}^{n}\left(10^{-\frac{1}{10}}\right)^{a_i}\right) - offset$$

It can be implemented as follows in python:

```python
import math
import numpy as np

def qualmean(qualities: bytes, phred_offset: int = 33) -> float:
    phred_scores = np.frombuffer(qualities, dtype=np.int8)
    probabilities = np.power((10 ** -0.1), phred_scores)
    average = np.average(probabilities)
    return -10 * math.log10(average) - phred_offset
```

This implementation is about 20% faster as the implementation at the beginning of this document.