Homework assignment #6-1 - Solutions

Problem 1

Sequences 3, 6 (and to a lesser extent, 7) were judged by the sequencer to be a bit noisier than others. This is just a warm-up question to get you to think about quality scores, namely that in the PHRED encoding, "E" is very good, and "A" is also quite good, but e.g. "6" signals that the sequencer judged that base call to be less certain.

Problem 2

BLAST finds many exact matches in its database, almost all of which are from the genus *Lactobacillus*. The most frequent match in the list is *Lactobacillus salivarius*, but this sequence is in fact too short to confidently claim a species-level taxonomy. (Not something you would know, just pointing this out.)

Problem 3 - See attached.

Problem 4

The two major sources of substitution errors in 16S data are: the Illumina sequencer itself, and a molecular machine called **polymerase**, used in a reaction called **Polymerase Chain Reaction (PCR)** to selectively amplify the 16S-encoding DNA prior to its sequencing.

Problem 5

For the sequencer, any physical sequences in the sample are equally real – it does not know what artifacts PCR may have generated. The quality scores can only "catch" potential base-calling errors of the sequencer itself, i.e. the sequencing artifacts. (E.g. if the strongest light flash was green, but the red channel also had a blip, then we're not as confident about this particular base call.)

Problem 6

- Sequences 3,7,9 are identical to Seq_1; if we assumed Seq_1 was real, then so are these.
- Seq_6 differs by just one base, which also happens to have a low quality score → most likely a sequencing error.
- Sequences 2,4,10 are identical and differ by one substitution. Since the same substitution occurs repeatedly and has good quality score, this is not a sequencing artifact. Since it's just one substitution away from Seq_1, Occam's principle points to this probably being a PCR error. But the possibility of this being a true biological sequence cannot be excluded.
- Sequence 5 is the same as 2,4,10, plus an extra sequencing error on top, reinforcing our belief that 2,4,10 was physically present in the sequenced sample.
- Finally, Seq_8 is three substitutions away (from Seq 1 or any other) and so is likely real.

How do we know if Seq_(1,3,7,9) is real, and not Seq(2,4,10)? Our demo file only had 10 sequences, but in a real-life scenario the real sequence will typically be much more abundant than any specific error it generates: even a common PCR error will be an order of magnitude less frequent.

Problem 7

If sequence 2,4,10 is real, it should appear in the replicate. A PCR error is *unlikely* to appear in the same place. It's not unheard of – PCR errors are not, in fact, evenly distributed along a given sequence… But to a first (and reasonably good) approximation, if the same sequence appears again in an independently processed sample, it's probably real.

```
@Seq01  Original (BLAST match: Lactobacillus salivarius)
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCAGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEA<EEEEAEEEEEEEEEEEEEEEEEE

@Seq02  1nt away, PCR error
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCGGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<EEEEEEEEAAAEEEEEEEAEEA

@Seq03  Original
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCAGAAGAGGAGAGTGGAACTC
+
EEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/</AEE/EEEEEEAEEEEAEE6/A6EEEEEEEEAAE/6AE<EEE/EEA/6EE<EEE<E//EE6<A/EEA//E<EE/E/E/E<E/

@Seq04  1nt away, PCR error
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCGGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEAEEAAAAEEEEEEEEE/

@Seq05  1nt away, PCR error + Sequencing error on top
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAAACGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCGGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE6EEEEEE/EEEEEEEEEEEEEEEEEEEEEEEEAEEEEEAAEEEEEEEEEEEAEEAEA

@Seq06  1nt, Seq error
GCAAGCGTTGTCCGGATTTATTGGGAGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCAGAAGAGGAGAGTGGAACTC
+
EEEEEEEEE/EEEEEEEEEEEEEEE2EEEEEEEEEEEEEEEEEEE/E//AAEEEE/AEEEEEEEEEA/E/EEEEEE/EAEEAE/E/EAEEEEEEE/EEEEEEE6//EEEEEEEEE<EAEEEEEEE/6E/EEEE<A<

@Seq07  Original
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCAGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEA/EAE/EEEE/EEAEEEEEEEEEAAEEEEAE/EEEE/A/AEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEAEEEEEEAEEEEEAAEEEEEEEEEEE

@Seq08  3nt away, so presumably real (BLAST matches: Lactobacillus murinus, animalis)
GCAAGCGTTATCCGGATTTATTGGGCATAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGGAGACTTGAGTGCAGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE6EEEEEEEEEEEEAAEEEEEEEEEEEEEAEEEEAEEEEEEEEEEEEEAEEEEAAAEEEEA

@Seq09  Original
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCAGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEAEEEEEEEEEA<AEEEEAEEEEAAEEEAEE

@Seq10  1nt, PCR
GCAAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCAGGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTGCATTGGAAACTGGAAGACTTGAGTGCGGAAGAGGAGAGTGGAACTC
+
EEEEEEEEEEEEEEEEEEE<EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAA<EAEE<AAAEEEEAEEEEE
```