Overlap Quality

Knut Reinert

1. Overview

In several phases of the assembler one faces the situation that overlaps are computed between sequences (either fragment or consensus sequences) that stem from different regions of the genome, i.e. they have a *repetitive* overlap. Among the various strategies for detecting this case as opposed to the case of *normal* overlaps (i.e, the sequences stem really from the same region) we describe here the following one.

Assume you are given the two overlapping sequences S1 and S2 together with a string of their PHRED quality values. First we compute an alignment of the two overlapping regions which we then refine with Using the quality values (NOT DONE YET). This finally yields the to aligned, gapped strings with assigned quality values as shown below.

```
lllllllllllllllllNllllllllllbkkkdgTkkKKKNAKKKKKKKKK
TGATCTATACACTCGTC-TGGGGCTACGACTTACGAGGCATGATCCTGCAC

TGAT-TATACACTCGTCGTGGGGCTACGACTTACGAGGCATGATCCTGCAC
KKKECMKKNKEKK:KKKKKKKIEKKKKOKMKKKKKKLKKPKK5KAKKKKKK
```

The task is now to assign a quality value to such an alignment such that normal alignments get a high quality score whereas repetitive alignments are assigned a low quality score. In order to do so we make use of the fact that repetitive alignments contain, in addition to the mutations caused by sequencing errors, also mutations due to evolutionary events. This should result:

- in a higher overall error rate in the alignment than one would expect given the quality values
- especially into more high quality mismatches, that is mismatches where both positions have a high PHRED value.

2. Method

We use an approach similar to Yandell et al. We use Bayes's formula to compute the probability that the observed data fits one of two models. The two models we consider are:
MN = the alignment is normal and therefore the mutations due to sequencing error.
MR = the alignment is repetitive and therefore the observed mutations are due to sequencing error and polymorphism.
Our observed data point is the number D(qt) of mismatches in the alignment (indels and substitutions) in which both positions have a quality value of at least qt. First we infer the average probability that a function has a qt-mismatch by using the quality values of the matching region where the quality of a dash is arbitrarily set to the quality of a neighboring character (THIS IS NOT THE CASE AT THE MOMENT HAS TO BE TESTED). This gives us an average sequencing error rate of s. Next we stipulate that the difference in between paralogous sequences is at least x percent neglecting SNPs.

If the alignment was normal we would expect the data to support model MN better than if the alignment was repetitive, meaning that P(MN|D) should be higher. Hence this probability can be taken as a quality measure. Assuming model MN is true, we would expect DN=s*l mismatches on average. Assuming model MR is true, we would expect DR=(s+x)*l mismatches on average, where l is the length of the alignment.

Asking for the probability P(MN|D) we can use Bayes’ formula to which yields:

P(MN|D) = (P(D|MN)*P(MN) / (P(D|MN)*P(MN)+P(D|MR)*P(MR))

Since we do not know the prior probabilities of the two models we have to assume them equally weighted. In this case the formula comes down to

P(MN|D) = (P(D|MN)/ (P(D|MN)+P(D|MR))

Hence we have to approximate the probabilities P(D|MN) and P(D|MR). We do that by assuming for MN a Poisson distribution with parameter λ=DN and for MR a Poisson distribution with parameter λ=DR. Doing the algebra this yields

P(MN|D) = 1 / (1 + ((s+x)/s)^D* e^(-l*x)))

This probability becomes more and more discriminating the longer the alignments are. For values see the spreadsheet ..\Analysis\Bayesian.xls.

3. Planned improvements

Since the above probability is depending on the length of the alignment it is not clear where we should set a threshold ofr a given length, such that we would expect a given success rate, where the success rate S is defined as the expected ratio of the number C of correctly classified alignments to the total number T of evaluated alignments.
Or put it differently, the question is whether T/W is the same for alignments of length L1 and alignments of length L2 with L1<L2.

We are planning of empirically evaluating this by conducting the following experiment.

1) We form buckets for PN values, say 1.0-0.98 and 0.98-0.96 and so on.
2) Now given x and s we randomly generate normal and repetitive alignments with a fixed length L.
3) Then we compute the PN value for each of these alignments.
4) This gives a success rate for each of the buckets.
5) Make these computations for different alignment lengths.

If the success rates of the different buckets are different for different lengths we want to adapt for it such We can devise a quality score Q(PN,l) that is not dependent of the length of tha alignment.