Microhet Separation

Knut Reinert

1. Overview

The microhet separation module's task is to use the microheterogenity of different repeat copies in order to decide whether an overlap between two fragments is normal or repetitive. If the later is the case the over lap should be removed.

2. High Level Design

3. Methods

3.1. Separating an alignment (method one)

For the described procedure microhet we assume that we are given a multiple alignment **a** of **k** fragments together with its consensus sequence. The length of the alignment is **L**, i.e. **a=(a_{i,j})** with **1<= i <= k** and **1 <=j <= L**. In addition, we know the mate of each fragment and its distance estimation (if it has one). The fragments form **group**s from different regions of the genome that share almost the same sequence. The underlying sequences of the fragments differ at some positions due to point mutations (inserts/deletions/substitutions). This means that each **a_{i,j}** differs from the original (unknown) sequence with probability **m_{i,j}**. This in turn means that for a particular position we expect all fragments stemming from the same group to have the same base. Unfortunately this is not true, because fragments stemming from the same region can differ due to sequencing errors. That means that each **a_{i,j}** differs from its underlying sequence with probability $**s_{i,j}**$, regardless of which group it belongs to.

In addition we are given mate information, that means a fragment has a mate located at a approximately known distance of that fragment. This information is considered to be reliable meaning that the probability of that information being wrong is quite low.

In general we face two problems given the above information:

- Devise a test that determines whether the fragments in the alignment come from different locations or not.
- If they are likely to come from different locations, the more interesting question would be, from how many locations they come and which fragment belongs to which location.

We address the two problems in separate Sections below.

Microhet Testing

If an alignment **a** contains fragments from different locations in the genome, we call it in the following **repetitive** alignment as opposed to the case where all fragments come only from one location. In this case we speak of a **simple** or **normal** alignment. Since fragments from different locations in the genome are mutated, all fragments in a **group**, i.e., coming from the same location, should have the same mutated character (except for sequencing errors). This feature distinguishes the different groups from each other.

It contains part of a repetitive alignment generated using a simulator where dots represent the underlying character displayed in the first row. In the displayed alignment two consecutive rows

are in one group.

```
ATACAT-C-AG-T--CGCAAAGA-CTTGT-CCG

1 T-.......T......C.............
2 T.............-.............
3 T............T.....A....G......
4 T............T.....A....G......
5 ......G...C......G.........C...-.
6 ..............G........A....-.
7 ..........T..............T...
8 ..........T....A.............
```

The only hope to distinguish a repetitive alignment from a simple alignment is that there are at least two groups that contain more than one fragment. Therefore **4** is a lower bound on the number of fragments that have to be in the multiple alignment **a**.

In the following we describe a one-sided statistical test of the null hypothesis **H_0** stating that the observed alignment is simple.

We assign to each column in the alignment a predicate **P_{i,j}** that depends on rows **i** and **j** and which is explained below. This predicate occurs with probability **p(s)** in a simple alignment where **p(s**) depends on the average sequencing error rate. Hence the number of columns that have this predicate is a random variable **Q** that is binomially distributed with a mean of **l*p(s).** The predicate **P_{i,j}** such that a column in a repetitive alignment ought to have a much higher probability to have this predicate for at least one pair of rows. Assume the observed alignment contains **x** columns with the predicate **P**. Then we test whether we accept or reject the hypothesis **H_0: x<= l*p(s)** for a given level of significance **alpha**.

This is done by computing the critical value **z** for which the probability **1-P(Q<= x)** is smaller than **alpha**. Since **p(s)** is quite small in our test we cannot approximate the binomial distribution by a normal distribution but rather have to compute the critical value on the fly.

Now we describe how we choose the predicate **P_{i,j}.** We say that a column has the predicate **P_{i,j}** if it contains the same character **c** in rows **i** and **j** different from the majority character of the column. Given the sequencing error rate **s**, which can be estimated from the given alignment, the probability of a column having this predicate can be computed as follows.

- In the rows different from row **i** and row **j** anything can happen as long as the character **c** occuring in row **i** and **j** does not become the majority character. That means we can compute the probability of all possible events in the remaining rows by correctly summing up the probabilities of a fournomial distribution yielding a probability **P(c).**
- For the two rows $**i**$ and $**j**$ there are two possibilities. Either the character **c** in the two fixed rows has changed from the underlying sequence. This event happens with probability **(s/4)^2** for each of the four characters the underlying character can change into. Or, the character **c** in the two fixed rows has not changed from the underlying sequence. This event happens with probability **(1-s)^2**.
- Hence, assuming that **c_1,c_2,c_3,c_3** are characters different from the underlying character **u**, the overall probability of the predicate **P_{i,j}** is **(1-s)^2* P(u)+ sum_{i=1}^4 P(C_i)* (s/4)^2**.

Now we know the probability that a column has the predicate **P_{i,j}** for two fixed rows **i** and **j**. Unfortunately it is not necessarily true that a repetitive alignment has more columns with predicate

P_{i,j} than a simple alignment for a particular pair **i**, **j**. The key observation is that a repetitive alignment is very likely two have **some** pair of rows, say **i'** and **j'** that have the predicate **P_{i',j'}.**

Therefore we conduct our test for each pair **i,j** of rows. If we reject the null hypothesis for one pair, then we assume that the alignment is repetitive. Since we conduct now **t=l\choose 2** tests, the chance of incorrectly rejecting the null hypothesis rises from **alpha** to **1-(1-alpha)^t**. Or stated differently, if we still want to have a confidence level of **alpha** we have to conduct each of the test with a level of significance of **1-(1-alpha)^{1/t**}. (NOTE: all distributions an probabilities were cross-checked with the microhet-simulator).

Microhet Grouping