UP - F, value in points

Tracks

1. Track 1: unmarked points
2. Track 2: function

Question

In the unmarked points of track 1, is the average value of the function in track 2 smaller/different/larger than expected by chance?

Comment:

- The test is analytic and assumes that the function is white noise. The assumption is also satisfied if the distance between points are so large that there is the correlation between neighbouring points is small.
- We assume the function in track 2 is fixed and that the points in track 1 are independent of the function values in track 2.
- Significance is determined by means of p-values. Small p-values identify bins where the function values are smaller/different/larger than expected in the points of track 1.
- If the points in track 1 depend on other tracks, it is possible to condition the test on an intensity track using this information.

Bins

The genome (or the areas of the genome under study) are divided into small regions, called bins. The tests are performed in each bin.

Hypothesis tested

For each bin i we have one null hypothesis
$\mathbf{H}_0$: *In the unmarked points of track 1, the average value of the function in track 2 is the same as the average function value.*

There are three alternative hypotheses:
$\mathbf{H}_1$: *In the unmarked points of track 1, the average value of the function in track 2 is smaller than the average function value.*
or
$\mathbf{H}_2$: *In the unmarked points of track 1, the average value of the function in track 2 is different than the average function value.*
or
$\mathbf{H}_3$: *In the unmarked points of track 1, the average value of the function in track 2 is larger than the average function value.*

Statistics and rejection of the null hypothesis

Let n be the number of base pairs in the bin, and let Y_i, $i = 1, 2, \cdots, n$, be the function values in the bin. We assume $Y_i \sim N(\mu, \sigma^2)$. Define the average $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Also, define $G_i = 1$ if there is a point in track 1 and $G_i = 0$ otherwise. Define $X = \frac{1}{m}\sum_{i=0}^{n} G_i Y_i$ where m is the number of unmarked points in track 1. Notice that under the null hypothesis $EX = E\bar{Y}$ and

$$X - \bar{Y} = \sum_{i=1}^{n}(\frac{G_i}{m} - \frac{1}{n})Y_i.$$

Define the constant K from the following expression

$$Var(X - \bar{Y}) = \sum_{i=1}^{n}(\frac{G_i}{m} - \frac{1}{n})^2\sigma^2 = (K\sigma)^2$$

and the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Under the null hypothesis $S^2(n-1)/\sigma^2$ is χ^2-distributed with $n-1$ degrees of freedom. Then the variable

$$T = (X - \bar{Y})/SK$$

is t-distributed with $n - 1$ degrees of freedom. We find the p-value from this distribution depending on the alternative hypothesis.

The test described above is very similar to a standard two sample t-test and the tests will probably give almost identical result. The test described above may be better in reusing previous calculated data and hence reduce CPU time.

Alternative assumption using an intensity track

Assume both the position of the points in track 1 and the function values in track 2 depend on a third track, denoted track 3. We then want to find out if the average value of the function in the points of track 1 is different from the average function values when we also take track 3 into consideration. Track 3 is used for making an intensity track W that gives a weight to each base pair.

For each bin i we have one null hypotheses
$\mathbf{H}_0$: *In the unmarked points of track 1, the average value of the function in track 2 is the same as the weighted average function value.*

There are three alternative hypotheses:
$\mathbf{H}_1$: *In the unmarked points of track 1, the average value of the function in track 2 is smaller than the weighted average function value.*
or
$\mathbf{H}_2$: *In the unmarked points of track 1, the average value of the function in track 2 is different than the weighted average function value.*
or
$\mathbf{H}_3$: *In the unmarked points of track 1, the average value of the function in track 2 is larger*

than the weighted average function value.

We define an intensity track W_i, $i = 1, 2, \cdots, n$, for the points in track 1 conditioned on track 3. W_i is the probability for a point in track 1, conditioned on the value in track 3. We assume that W_i takes k discrete values. Let $q(i)$ be a function that from base pair number i finds the index $1, 2, \cdots, k$ to the discrete value of W_i. Then $W_i = W_j$ if and only if $q(i) = q(j)$. Then we assume $Y_i \sim N(\mu_{q(i)}, \sigma^2)$ where the expectation depends on W_i. Define the variable

$$Z = \frac{1}{\sum_{i=1}^{n} W_i} \sum_{i=1}^{n} W_i Y_i.$$

Notice that $EX = EZ$ and

$$X - Z = \sum_{i=1}^{n} (\frac{G_i}{n} - \frac{W_i}{\sum_{j=0}^{n} W_j}) Y_i.$$

Define the constant K from the following expression

$$Var(X - Z) = \sum_{i=1}^{n} (\frac{G_i}{n} - \frac{W_i}{\sum_{j=0}^{n} W_j})^2 \sigma^2 = (K\sigma)^2.$$

Furthermore, define $\bar{Y}_j$ as the average of Y_i in the base pair where $q(i) = j$ and

$$S^2 = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - \bar{Y}_{q(i)})^2.$$

Under the the null hypothesis $S^2(n-k)/\sigma^2$ is χ^2-distributed with $n-k$ degrees of freedom. Then the variable

$$T = (X - Z)/SK$$

is t-distributed with $n-k$ degrees of freedom. We find the p-value from this distribution depending on the alternative hypothesis.

There is a similar standard two sample t-test and the tests will probably give almost identical result. The test described above may be better in reusing previous calculated data and hence reduce CPU time.

A slightly better approach is to use a control track. Instead of assuming W_i takes k discrete values and the definition of $q(i)$ above, we may use a control track Q_i taking categorical values. We then assume $Y_i \sim N(\mu_{Q_i}, \sigma^2)$ where the expectation depends on Q_i which is more general than when we use intensity as described above. The only change using a control track instead of an intensity track is the definition of $\bar{Y}_j$ and S^2. Define $\bar{Y}_j$ as the average of Y_i in the base pair where $Q_i = j$ and

$$S^2 = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - \bar{Y}_{Q_i})^2.$$