MP - MP, similar marks in nearby points

Tracks

1. Track 1: marked points
2. Track 2: marked points

We assume that either the marks in both tracks are categorical or the marks in both tracks are continuous, discrete, ordered categorical and not ordered categorical.

Question

Is the mark of a points in track 1 and the mark of its nearest neighbour point in track 2 independent?

Comment:

- We assume the position of the points in track 1 and the track 2 are fixed. We permutate only the marks of the points in one or both tracks.
- We identify the point in tracks 2 that is the nearest to each point in track 1. There are several different options. Nearest in the direction of lower base pair number, in both directions and in the direction of higher base pair number. It is necessary with a rule of preference if there are points at same distance in both directions. We may neglect neighbours that are further apart than a maximum distance. Some point in track 2 may be the nearest neighbour to several points in track 1 and some points may not be the nearest neighbour to any points in track 1. We assume that this occurs so seldom that it does not dominate the statistics.
- Significance is determined by means of p-values. Small p-values identify bins where the marks of the points in track 1 are not independent of the marks of the points of track 2.
- The p-values are found by an analytic calculation or MC simulation.

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 the null hypothesis
$\mathbf{H}_0$: *The mark of a points in track 1 and the mark of its nearest point in track 2 are independent.*

The alternative hypothesis is:
$\mathbf{H}_1$: *The marks of a points in track 1 depends the mark of its nearest point in track 2.*

Statistics and rejection of the null hypothesis, categorical variables

In this section we assume that the marks of both tracks are categorical variables. Let r be the number of categories for marks of points in track 1 and let c be the number of categories for marks of points in track 2. Furthermore, let $O_{i,j}$ be the number of observations of points from track 1 with mark equal i where its nearest neighbour in track 2 has mark j. In this test we consider these pairs of marks and neglect that some points in track 2 may be part of several pairs and that some points in both tracks may be part of no pairs. The table with the $O_{i,j}$ values is a contingency table with r rows and c columns.

Let N be the total number of pairs, i.e. $N = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{i,j}$. If the marks of the pairs are independent, we expect $O_{i,j} \approx E_{i,j}$ where

$$E_{i,j} = \frac{1}{N} \sum_{k=1}^{r} O_{k,j} \sum_{k=1}^{c} O_{i,k}.$$

Let

$$X = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

Under the null hypothesis X is χ^2-distributed with $(r-1)(c-1)$ degrees of freedom. This is an approximation that is considered accurate if all $O_{i,j} > 10$. (ref. Wikipedia/Pearson's chi-square test). We find the p-value from this distribution. The combinations of i and j that give the largest contribution to the double sum in X, are the cells where the deviation from the independence assumption is largest.

Statistics and rejection of the null hypothesis, continuous or discrete variables

In this section we assume that the marks of both points and segments are continuous or discrete variables. Let X_i be the mark of a point in track 1 and Y_i the mark of its nearest neighbour in track 2, $i = 1, 2, \cdots, n$. We use the following test statistic:

The sample correlation

$$r_{x,y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)(Y_i - \bar{Y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_x s_y},$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Spearman's rank correlation is defined as the sample correlation except that it uses the ranks x_i and y_i instead of the original data X_i and Y_i.

Kendall τ rank correlation is then defined as

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}.$$

where n_c is the number of concordant pairs i.e. the number of pairs where $(X_i - X_j)(Y_i - Y_j) > 0$ and n_d is the number of of disconcordant pairs i.e. the number of pairs where $(X_i - X_j)(Y_i - Y_j) < 0$. The pairs where both $X_i = X_j$ and $Y_i = Y_j$ are both condordant and disconcordant, but are in fact not critical for the definition of Kendall τ.

The distribution for the sample correlation, Spearman's rank correlation and Kendall τ are known and we may find the p-value from these distributions.

In addition, we may use the test statistics

$$Z_1 = \sum_{i=1}^{n} (X_i - Y_i)^2,$$

and

$$Z_2 = \sum_{i=1}^{n} |X_i - Y_i|$$

The distribution for these test statistics are not known and it is necessary with MC simulations in order to decide whether to reject the hypothesis.