F - F, similarity (correlation)

Tracks

- Track 1: *Function*
- Track 2: *Function*

Question

Where are the two functions similar/associated/correlated ?

Comment:

- Correlation is measured in different ways.
- The question is answered in the setting of statistical hypothesis testing. We perform the test inside a series of bins of the genome.
- Significance is determined by means of a p-value calculated for each subinterval. Small p-values identify regions with significant results, where the tracks differ.
- The p-values are computed as explained below, where the null hypotheses are explained in detail.

Refined questions

Alternative A1

- Where are the two functions associated?

Alternative A2

- Where are the two functions positively associated?

Alternative A3

- Where are the two functions negatively associated?

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.

Hypotheses tested

- A1 H_0: No association against H_a: Association
- A2 H_0: No association against H_a: Positive association
- A3 H_0: No association against H_a: Negative association

Tests and test statistics

Similarity of the two functions can be studied in various ways. We focus here on simple tests for correlation like associations between the two series.

Inside a bin, assume we have n observation pairs (x_i, y_i), where x_i is a data point of track 1 in position i and y_i is a data point of track 2 in position i. We wish to test if certain values of x and y have a tendency to occur together, for instance that both track 1 and track 2 tend to have high values or both low values in the same (intervals of) base pairs, which would be a positive association.

The n observation pairs could be the function values in all base pairs inside the bin. But it is likely that each function exhibits (strong, positive) autocorrelation, that is, dependency between function values in neighbouring sites inside each track, f.ex. between x_i and x_j. This will result in too small p-values if ignored, because the following calculations are based on assumptions of n independent observation pairs. To reduce autocorrelation, we divide each bin into n sub bins and use a representative from each sub bin as the n data points for each track in each bin. Such a representative could be the mean, the median or the function value in the midpoint of the sub bin.

Based on these n pairs of observations, we test for linear or non linear but monotone relationships between the two tracks inside each bin. The number of sub bins n should typically be around 20-30. Non smooth functions require more sub bins.

- **Option 1: Pearson correlation** (Assuming binormality and a linear relationship between x and y.)

  Test statistic

  $$T_n = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

  where r_{xy} is the empirical correlation coefficient

  $$\frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}.$$

  Under the null hypothesis of no correlation, T_n has a $t(n-2)$ distribution.

- **Option 2: Spearman correlation** (No assumption on normality, no linear assumption, measures any monotone relationship between x and y.)

  Substitute $x_1, x_2, ..., x_n$ with their ranks, and the same with $y_1, y_2, ..., y_n$. In the case of ties (equal values for two or more measurements), give the same rank to all of the involved values, which should be the mean of the ranks that they otherwise would have had. Calculate r_{xy} above with the observations substituted by their ranks.

  If $n \geq 20$, we use the test statistic T_n and the $t(n-2)$ distribution above to find a p-value. If $n < 20$, precalculated tables for p-values are available.

  If no ties are present, the Spearman r_{xy} can be very easily calculated as

  $$r_{xy} = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where d_i = rank of x_i - rank of y_i.

- **Option 3: Kendall's tau** (Same assumptions as for Spearman, but different measure of association.)

  Among all $n(n-1)/2$ possible pairwise comparisons $\{i, j\}$, let

  $$C = \#\text{ pairs where } x_j - x_i \text{ and } y_j - y_i \text{ have the same sign (Concordant)}$$
  $$D = \#\text{ pairs where } x_j - x_i \text{ and } y_j - y_i \text{ have the opposite sign (Discordant)}$$

  and

  $$\tau = \frac{C - D}{n(n-1)/2},$$

  which is Kendall's tau. In the case of ties ($x_i = x_j$, $y_i = y_j$, or both), τ is instead defined as

  $$\tau = \frac{C - D}{\sqrt{[\binom{n}{2} - n_x][\binom{n}{2} - n_y]}}.$$

  Here n_x and n_y are the number of ties involving x and y, respectively.

  Test statistic is in both cases

  $$Z_n = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

  which is $N(0, 1)$ when n is large. Should be OK for for $n \geq 20$. Tabulated p-values are available for n up to 50, but only in the situation of no ties.