From: Lars Holden
Date: 21.10.2009

US - n x UP, all points have same distribution between the segments in US

Tracks

1. Track 1: unmarked segments
2. Track 2 - n+1: unmarked points

Question

Does track 2 - n+1 have the same distribution of points between the segments in track 1?

This question may also be rephrased as:

Is the probability for a points to be inside a particular segment in track 1 the same for all the track 2 - n+1?

We also test the slightly extended question:

Is the probability for a points to be inside a particular segment or outside the segments in track 1 the same for all the track 2 - n+1?

Comment:

- We assume track 1 as fixed and track 2 - n+1 have the same distribution.
- Significance is determined by means of p-values. Small p-values identify bins where the track 2 - n+1 have different distribution between the segments of track 1.
- The p-values are found by an analytic calculation.

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
$\mathbf{H}_0$: *Track 2 - n+1 have the same distribution of points between the segments in track 1.*

The alternative hypothesis is:
$\mathbf{H}_1$: *Track 2 - n+1 have different distribution of points between the segments in US?*

We may also consider the slightly extended hypothesis where we also consider the probability to be outside all the segments.

Statistics and rejection of the null hypothesis

If we also consider the probability for points outside the segments, then we denote all the area outside the segments as an additional segment. Then it is possible to use the same description for the two alternative versions. Let r be the number of segments and $O_{i,j}$ the number of points from track 2 - n+1 inside segment i from track $j+1$. The table with the $O_{i,j}$ values is denoted a contingency table with r rows and n columns.

Let N be the total number of points from track 2 - n+1 i.e. $N = \sum_{i=1}^{r}\sum_{j=1}^{n} O_{i,j}$. If the points have the same distribution between the segments 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 $(r-1)(n-1)$ 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 gives the largest contribution to X is where we have the most significant contribution to reject the hypothesis. The combinations of i and j where $\frac{|O_{i,j}-E_{i,j}|}{E_{i,j}}$ is largest is where the deviation from the same distribution is largest.