UP - US, uniform positioning of points within segments

Tracks

1. Track 1: unmarked points
2. Track 2: unmarked segments

Questions

Q2-1 Where in the genome are the points in track 1 positioned more towards the borders of the segments in track 2 than expected by chance?

Q2-2 Where in the genome are the points in track 1 positioned more towards the middle of the segments in track 2 than expected by chance?

Q2-3 Where in the genome are the points in track 1 positioned more towards the left part of the segments in track 2 than expected by chance?

Q2-4 Where in the genome are the points in track 1 positioned more towards the right part of the segments in track 2 than expected by chance?

Q2-5 Where in the genome are the points in track 1 positioned more non-uniformly inside the segments in track 2 than expected by chance?

Comments:

- We assume that segments are fixed and regard points as random and independent.
- Significance is determined by means of p-values. For Q2-1, small p-values identify regions where the points in track 1 are closer to the borders of the segments in track 2 than expected. Similar for Q2-2, Q2-3, Q2-4 and Q2-5.

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

H_0: Points have a uniform distribution within segments.

Remarks

- If points are uniformly and independently distributed over segments, this will also be the case even if we rescale all segments to the same length.
- The tests described here are restricted to situations where such rescaling appears reasonable. This may not always be the case; biologists may for instance be interested in the distribution of the length in absolute terms from the start of the segments. This will, however, often be an estimation problem rather than a testing problem.

- If points are uniformly distributed, they are symmetrically distributed around the mean/median value. This may be used to construct tests.

Alternative hypotheses

$\mathbf{H}_1$: Points tend to be positioned towards the borders of the segments.

$\mathbf{H}_2$: Points tend to be positioned towards the middle of the segments.

$\mathbf{H}_3$: Points tend to be positioned towards the left part of the segments.

$\mathbf{H}_4$: Points tend to be positioned towards the right part of the segments.

$\mathbf{H}_5$: Points are unequally distributed within segments.

Testing against $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$ and $\mathbf{H}_4$

When testing against H_1 and H_2, let d_i, $1 = 1, \ldots n$, be the relative position, but now scaled such that the value is -1 at both borders and 1 in the middle of the segment (and thus 0 halfway between the middle and the border).

When testing against H_3 and H_4, let d_i, $1 = 1 \ldots n$, be the relative position of points within segments scaled such that the value is -1 at the left end and 1 at the right end.

To test the first four hypotheses above, we may use the Wilcoxon sign-rank test. For n larger than 20-30, we may also use the t-test, which is markedly less time-consuming.

The Wilcoxon test is done in the following way: Rank the d_i without regard to sign; with 1 assigned to the observation closest to 0 (any zeros are neglected). Then compute $W+$ and $W-$ as the sums of the value of the ranks of the originally positive and negative observations, respectively. Significance levels are based on the fact that if H_0 is true, then there are $2n$ equally likely ways for the n ranks to receive signs. As test statistic, we use $W = MIN(W-, W+)$. For small samples ($N \leq 30$), the critical regions must be found from some table. For $N > 30$, the test statistic W approaches a normal distribution with a mean of $n(n+1)/4$ and a variance of $n(n+1)(2n+1)/24$. However, to increase speed, we should consider using the t-test when $n > 20$. The t-test to use is the standard one-sample test.

Testing $\mathbf{H}_5$

To test against the alternative H_5, one may use the Kolmogorov test.

Remark

The alternatives are formulated such that a one-sided test may appear most appropriate, except for H_5. This is hardly an important point, however.