UP - US, located inside

Tracks

- Track 1: Unmarked points
- Track 2: Unmarked segments

Questions

Where in the genome is there more points of track 1 inside the segments of track 2 than expected by chance?

Comments:

- This question is used to identify regions of the genome (or the part of it under analysis) where points are over-represented inside the segments and where this over-representation is so strong that it would seldom happen by chance. Such over-representation would be a strong indication that points and segments do not occur independent of each other.
- In the analysis, the genome is divided into bins, and the tests are carried out for each bin
- "more" may be changed into "less" or "either more or less".

Hypothesis tested

The model valid under the null hypothesis is given by:

1. a preservation rule for each track,
2. a probability law on how the non-preserved elements are randomized. This rule implicitly imply independence of the positioning of the points and segments in the two tracks.

Consider the points in track 1 in one bin. A challenge in formulating models is that the structure of the interpoint distances may be crucial. If points (under the null hypothesis) are randomly distributed, or more precisely, conform to a Poisson process, then there is a simple solution to the testing problem (see Null hypothesis 1 below). However, there will probably often be more structure in the sequence of points: the points may occur more regularly than in a Poisson process, but probably more importantly, they may form clusters of points. Such clustering is very difficult to model. Thus, in the solutions presented below we either make the strong assumption of random positioning of the points (Null hypothesis 1) or we preserve the point positions in track 1 and base the tests on specific assumptions regarding the random segmental structure of track 2.

Let N be the total number of points in track 1 in the bin under consideration, and let T be the number falling within the segments defined by track 2.

Null Hypothesis 1, points in track 1 are randomly distributed

The model valid under the null hypothesis is:

1. The number of points in track 1 is preserved and the points are assumed uniformly distributed (typically arising from a simple Poisson process).

2. A fraction of the bp in track 2 equal to the observed one is included in segments.

Note that assumption 1 above is a very strong assumption on lack of structure for the points. The gain from this strong assumption is that we get a simple test and only need to make a very weak assumption for track 2.

The p-value is $P(T \geq k)$ where k is the observed value of T in the data. (If the question would be "less", we would use $\leq$; if "different" we would multiply times two.)

Let θ be the fraction of bp in track 2 that belongs to segments. Then

$$P(T \geq k) = \sum_{h=k}^{N} \left(\begin{array}{c} N \\ h \end{array} \right) \theta^h (1-\theta)^{N-h}$$

gives the exact p-value.

It is possible to make an asymptotic approximation, to avoid computing these sums. Here we use that the binomial is approximated by a normal. More precisely, a Binomial(n,θ) random variable has approximately a normal distribution

$$N(n\theta, n\theta(1-\theta)).$$

Hence we may approximate by

$$P(T \geq k) \sim 1 - \Phi(\frac{k - 0.5 - N\,\theta}{\sqrt{N\,\theta(1-\theta)}})$$

asymptotically. The 0.5 is a continuity correction, making the approximation better.

Null Hypothesis 2

The model under the null hypothesis is given by:

1. Track 1 is preserved as observed

2. In track 2, we preserve the segment lengths, but not the segment ordering or positions.

The test statistics remain the same as above; the number T among the N points that falls within a segment. However, now T has a distribution under the null hypothesis which we are not able to find exactly.

We thus compute p-values using Monte Carlo simulations. We do this by generating many new configurations of track 2 (in the bin we are working on). Each repetition has the same segments as the data (same collection of segment lengths), but now with random ordering of the segments within the bin and with random distances between the segments.

Preserving the lengths of the segments, means that we know the total length of the inter-segments too. If there are K segments, the algorithm starts by splitting the total intersegment lengths L into K parts (intersegments) by drawing K-1 points on [0, L]. A realization of track

2 is then obtained in a two step process. The reason for the two steps is that the borders of the bins represent a challenge in the implementation: using the trivial solutions, points in track 1 close to the segment border will either have larger or smaller probability of being included in a segment. Therefor, in the first step we make a sequence: the first intersegment, followed by a randomly drawn segment, then the next intersegment, the next randomly drawn segment and so on. Then we connect the borders of track 2 (if both borders are covered by segments, we still regard them as two segments). Finally, we randomly draw a starting position on the circle and use this as the starting point for the bin.