From: Lars Holden
Date: 1.1.09

1 Single track tests

1.1 Unmarked points

1. Track 1: unmarked points

1.1.1 Questions

Are the points independently, uniformly distributed in each bin?

Comment:

- It is possible to test whether the points in each bin are independently, uniformly distributed. However, it may be difficult to separate between dependent and non-uniform if the hypothesis is rejected.
- When comparing different tracks, the test becomes even more complicated. We may fix track 1 (of any kind) and then test whether the points in track 2 are independent of track 1, independent of each other and uniformly distributed in each bin. Alternatively, we test whether the points in track 2 are independent of track 1 and uniformly distributed in each bin without assuming independence between the points in track 2. The difference between these to tests are that we simulate the points in track 2 from different distributions. Assuming one of these distributions when it is not correct, leads to wrong p-value. But it is not easy to identify which assumption that is best.
- The p-values may be found by simulation or by an approximate calculation.

1.1.2 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.

1.1.3 Hypothesis tested

Question, independently, uniformly distributed
For each bin i, we have the following hypothesis:
$\mathbf{H}_0$: Points in track 1 are independently, uniformly distributed

If the points are independently, uniformly distributed, then the distances D_j between the points are exponentially distributed, i.e. with density $\lambda \exp(-\lambda x)$. Estimate $\hat{\lambda} = (1/n) \sum_j D_j$ where n is number of points in bin. Use the

Kolmogorov-Smirnov goodness-of-fit test whether all D_i are from the exponential distribution. Let $F(x) = (1 - \exp(-\lambda x))$ be the cumulative distribution of D_j, and let $D_{(j)}$ be the ordered distances. Define $F_n(x) = j/n$ if $D_{(j)} \leq x < D_{(j+1)}$ and

$$K_n = \sqrt(n) \sup_x |F_n(x) - F(x)|$$

K_n converges to the Kolmogorov distribution independently of the distribution F. Hence, we may find critical values for $K_n > c_\alpha$ that leads to rejection of the hypothesis. We assume that n is sufficient large such that the asymptotic properties are valid. If the hypothesis is rejected, then a plot with $F_n(x)$ and $F(x)$, may indicate why the hypothesis is rejected. If there are too many small distances, the points may cluster and if there are too many long distances, then there may be repulsion. If there is a non-uniform density, for example due to dependence with another track, this may also lead to many small distances even if each point is independent of the other points.

A scatter plot with D_j and D_{j+1} may also give valuable information. This may indicate that neighboring distances are dependent. If the points cluster and there are more than two points in the cluster, then this is easily seen from this plot.

There are also other goodness-of-fit tests.

1.2 Unmarked segments

1. Track 1: unmarked segments

1.2.1 Questions

Question 1

Does the length of the segments and the intersection follow a stationary Markov process in each bin?

Question 2

Does the length of the segments and the intersection follow a normal distribution in each bin?

Comment:

- If the length of the segment and the intersection are a stationary Markov processes, then there is a stationary probability that if a base pair is inside a segment, then the next base pair is inside the segment independently of the length of the segment. There is a corresponding probability, such that if a base pair is between segments, then the next base pair also is between segments. If the length are Markov processes, then the length of the segments and the intersections are exponentially distributed. Hence, we may use exactly the same goodness-of-fit test as the distance between points above.

1.2.2 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.

1.2.3 Hypothesis tested

Question 1

Does the length of the segments and the intersection follow a stationary Markov process in each bin?

Question 2

Does the length of the segments and the intersection follow a normal distribution in each bin?

In both cases may we perform the same goodness-of-fit test as for points and reject the hypothesis if $K_n > c_\alpha$. If we test for the normal distribution, then $F(x)$ should be the cumulative normal distribution.

If the hypothesis is rejected, then a plot with $F_n(x)$ and $F(x)$, may indicate why the hypothesis is rejected.

Let L_j be the length of the segments and S_j be the distances between the segments. Scatter plot of L_j by L_{j+1} and S_j by S_{j+1} and L_j by S_j may also give valuable information. This may indicate that neighboring segments/distances are dependent.

1.3 Marked points and segments

1. Track 1: marked points or segments

We may perform exactly the same tests as with unmarked points and segments described above. In addition, we may test whether the marks M_j are locally dependent by making a scatterplot of M_j by M_{j+1}.

1.4 Function

1. Track 1: Function

1.4.1 Questions

Does the function follow a stationary Markov process in each bin?

Comment:

- This is a natural question if the function takes discrete values or there are sudden large changes followed by intervals with small changes. If the function is a stationary Markov process, then the length between the changes are exponentially distributed. Hence, we may use exactly the same goodness-of-fit test as the distance between points above. It is also of interest to find the transition matrix, e.g. the probability of changing from state (value) i to state (value) j.