Genomic Segmentation

Overview

The genomic segmentation algorithm finds a segmentation according to the following criteria:

1. Neighboring regions have statistically significantly different average intensities (as defined by user-specified p-value)
2. Breakpoints (region boundaries) are chosen to give optimal statistical significance (smallest p-value)
3. Detected regions must contain a user-specified minimum number of data points

In addition to specifying a p-value threshold, the user also specifies a the minimum magnitude of change to be detected relative to the noise estimate for each chromosome. The signal to noise parameter allows the one parameter to represent the desired magnitude of change for all samples without regard to the samples' noise. Partek estimates the amount of noise for each sample using the difference between neighboring probes. This provides a good estimate of local variance with very minor influence of true biological changes.

While segmentation does not strictly produce a unique solution, it does produce a locally optimal solution, which we have found to out-perform HMM in sensitivity and specificity while requiring very little time and comparable results when compared to other algorithms such as CBS, etc. The algorithm has been developed to handle large amounts of genomic data efficiently while dealing with many artifacts found in microarray data.

After determining the segmentation result, two one-sided t-tests are performed on the probes in each region—one test above a given threshold, and one below a threshold. The minimum p-value of these two tests will be used to determine if the region is a significant deviation from the expected normal. The specified report p-value threshold will determine if a region is reported in the detected region result. For example, if the goal is to detect regions of copy # gain or loss, one may set the upper threshold to 2.1 and the lower threshold to 1.9 so that the p-value reflects the probability of being > `2.1` or < 1.9.

Parameters

The genomic segmentation procedure is a two step process.

1. Find a segmentation that produces significantly different neighboring regions
2. Filter these regions to only report those that are of interest

Segmentation parameters:

- The minimum number of probe sets specified will search for regions that contain at least a number of probe sets
- The p-value specifies the level of significance that the regions are different
- The signal to noise parameter describes the magnitude of significant region differences relative to the noise level in each sample. Increasing this parameter will report fewer breakpoints caused by small differences between neighboring regions

Report parameters:

- Below specifies the lower test filter. Any regions with means significantly below this value will be reported
- Above specifies the upper test filter. Any regions with means significantly above this value will be reported
- The p-value specifies the level of significance required in the above two tests