From Lars Holden
Date: 20.10.2009

F, finding local min and max, and intervals of interest

Track

1. Track : function

Questions

Q-1 Where in the genome are the local minimum/maximum of the function F?

Q-2 Where in the genome are there large changes in the value of the function F?

Q-3 Where in the genome is the function F low/high?

Q-4 Where in the genome has the function F small/large variations?

Comments:

- We study different properties of the function F and represent these as marked point and marked segments. These new properties may be of interest by itself. In addition, these properties are stored in new tracks and may then be used in tests with the other tracks in the Hyperbrowser. This note describes how to find the new tracks. Statistical tests with the new tracks are described in the tests with marked points and marked segments.
- We do not need to consider bins when generating the new tracks.

Identify points and segments

Let F_i be the function value in base pair i.

- Let G_i be the average of F in the interval $(i-n, i+n)$.
- Find the quantile q_c where the fraction c of the G_i values are below q_c.

- Define a local minimum point as a marked point at base pair i if
  - $G_k < q_{0.1}$ for $|k - i| < n + 1$
  - $G_{k+2n+1} > G_k$ and $G_{k-2n-1} > G_k$
  - F_i is the smallest value of F in the interval $(k - n, k + n)$

  Hence, the algorithm is to first find k and then find the index i such that the last criteria is satisfied. The point gets the three marks F_i, G_k and $G_k - (G_{k+2n+1} + G_{k-2n-1})/2$.

- We define a local maximum point correspondingly
  - $G_k > q_{0.9}$ for $|k - i| < n + 1$
  - $G_{k+2n+1} < G_k$ and $G_{k-2n-1} < G_k$
  - F_i is the largest value in of F the interval $(k - n, k + n)$

  The point gets the three marks F_i, G_k and $G_k-(G_{k+2n+1}+G_{k-2n-1})/2$.

- Define a contrast point as a point i if $|G_{i-n} - G_{i+n}| > q_{0.7} - q_{0.3}$ and where this difference is larger than the corresponding difference forthe n nearest base pairs. Define a marked point track with all the contrast points and let the mark be $G_{i-n} - G_{i+n}$.

- Define an interval with low values as a marked segment S if the following is satisfied
  - $G_i < q_{0.1}$ for all $|i - j| < n + 1$. Set $S = (i - n, i + n)$
  - Extend the segment S by adding one and one base pair while the average value of F_j inside the segment is lower than G_i. Try first to extend in the end with low index values, then for large index values, then for small index values and continue changing between the two ends.

  Let the marked segment track consists of all the segments generated by the above algorithm. If some of the segments overlap, remove segments with highest marks. The segments has average value equal or lower than $G_i < q_{0.1}$ and the interval is at least $2n + 1$ long with the mark equal to the average value of F_j in the segment.

- Define an interval with high values as a marked segment S if the following is satisfied

- $G_i > q_{0.9}$. Set $S = (i - n, i + n)$
- Extend the segment S by adding one and one base pair while the average value of F_j inside the segment is higher than G_i. Try first to extend the segment in the end with low index values, then for large index values, then for small index values and continue changing between the two ends.

Let the marked segment track consists of all the segments generated by the above algorithm. If some of the segments overlap, remove segments with lowest marks. The segments has average value equal or higher than $G_i > q_{0.9}$ and the interval is at least $2n + 1$ long with the mark equal to the average value of F_j in the segment.

- We want to define a marked segment track with segment where the variation of F_i is small or large. We use the following algorithm to identify these.
  - Define first the two function tracks L_i and H_i with the lowest and highest value respectively F_i in the interval $(i - n, i + n)$.
  - Identify segments $S = (i - n, i + n)$ where $H_i - L_i > q_{0.7} - q_{0.3}$ or $H_i - L_i < q_{0.5} - q_{0.4}$
  - Extend the segment S by one and one base pair while the difference between the highest and lowest F_j value in the segment is higher, respectively lower than $H_i - L_i$.

  Let the marked segment track consists of all the segments generated by the above algorithm with the marks equal to the difference between the highest and lowest of the function values F_j in the segment. If some of the segments overlap, remove the less significant segments first.