Exercise 13
Hi-C data.

Ewa Szczurek
MIM UW

January 12, 2016

Exercise 1. Today we will look at the topologically associated domains identified by Dixon et al. (2012) for the human IMR90 Fibroblasts (the total.combined.domain file). Full list of datasets available for this paper can be found at `http://chromosome.sdsc.edu/mouse/hi-c/download.html`.

It has been reported that not only CTCF, but a whole large group of architectual proteins can be found at the borders of topological domains (Bortle et al. 2014; `http://www.genomebiology.com/2014/15/6/R82`). In fact, the "border signal" is stronger, when more proteins are found bound to the border area (see Figure 4 of that paper, `http://www.genomebiology.com/2014/15/6/R82/figure/F4`)

We will use the matrix of binding domains of different proteins from that paper (the ArchBinding.csv file).

Write a program (R or python) that

1. Reads in the (unique) set of border points for chromosome 1 from the total.combined.domain file. These correspond to both the start and end points listed in each row.

2. Reads in, for chromosome 1, from the ArchBinding.csv file:

   - the ends of binding domains(start, end)
   - the binding event information (columns named after the proteins) for CTCF and Rad21.

3. computes the mid points of binding domains: (start+end)/2

4. collects such binding events where CTCF binds but Rad21 does not (lets call them "CTCF only")

5. collects such binding events where CTCF binds and Rad21 binds (lets call them "CTCF Rad21")

6. computes the distances from the borders to the nearest a) CTCF only mid points and b) CTCF Rad21 mid points.

7. for bins between 1 and 40000 and width 10000 compute the frequencies: number of distances that fall within each bin in the points a) and b).

8. plot the frequencies for a)- titled CTCF only and b) CTCF Rad21. On the x axes there should be the distances (1, 10001, 20001, ..., 39001), and on the y axes the computed frequencies.

9. Draw the same number of random positions (from 1 to the maximum observed border position on chromosome 1).

10. Repeat the same analysis as above for the random borders.

11. Add the curves for the distance frequencies for random borders to the respective plots.

Homework 1. Is the difference between the expected and the observed distances in the two cases (1. CTCF only and 2. CTCF Rad21) significant? To assess that for both cases, report the p-value from the Wilcoxon test comparing the sets of distances.