Data Processing

Chromosomes are divided into 200 bp intervals. For each interval, histone modification data will be assigned, judging from whether the sequence tags, which are pulled down by specific histone modification antibody, are overlapping with the interval sequences. The criterion of a tag assignment to an interval is that the start site of the tag is less than 75 bp upstream of the start site of the interval. Tags of 41 different modifications (H2AK5ac H2AK9ac H2BK120ac H2BK12ac H2BK20ac H2BK5ac H3K14ac H3K18ac H3K23ac H3K27ac H3K36ac H3K4ac H3K9ac H4K12ac H4K16ac H4K5ac H4K8ac H4K91ac CTCF H2AZ H2BK5me1 H3K27me1 H3K27me2 H3K27me3 H3K36me1 H3K36me3 H3K4me1 H3K4me2 H3K4me3 H3K79me1 H3K79me2 H3K79me3 H3K9me1 H3K9me2 H3K9me3 H3R2me1 H3R2me2 H4K20me1 H4K20me3 H4R3me2) will be assigned to each interval.

To assess whether the modification is truly presented in an interval, the tag counts of specific modification in an interval are modeled with a Poison distribution, with the mean set to the empirical mean of the number of specific tags in intervals. A probability of the modification greater than 0.9999 is used to calculate integer threshold of the presence of or absence of the modification. The empirical means for modifications H2AK5ac H2AK9ac H2BK120ac H2BK12ac H2BK20ac H2BK5ac H3K14ac H3K18ac H3K23ac H3K27ac H3K36ac H3K4ac H3K9ac H4K12ac H4K16ac H4K5ac H4K8ac H4K91ac CTCF H2AZ H2BK5me1 H3K27me1 H3K27me2 H3K27me3 H3K36me1 H3K36me3 H3K4me1 H3K4me2 H3K4me3 H3K79me1 H3K79me2 H3K79me3 H3K9me1 H3K9me2 H3K9me3 H3R2me1 H3R2me2 H4K20me1 H4K20me3 H4R3me2 PolII are 0.21, 0.13, 0.21, 0.22, 0.25, 0.21, 0.23, 0.26, 0.16, 0.21, 0.27, 0.22, 0.24, 0.23, 0.44, 0.25, 0.26, 0.20, 0.16, 0.23, 0.49, 0.60, 0.55, 0.53, 0.49, 0.76, 0.61, 0.32, 0.84, 0.32, 0.29, 0.37, 0.55, 0.58, 0.37, 0.57, 0.37, 0.60, 0.27, 0.44, 0.23, respectively. The integer threshold is 3 3 3 3 4 3 3 4 3 3 4 3 4 3 4 4 4 3 3 3 5 5 5 5 5 6 5 4 6 4 4 4 5 5 4 5 4 5 4 4 3, respectively, calculated in the following R script. Because we use a different dataset of methylation, polII and CTCF, the thresholds for those modification are different shown in the paper, while data for all acetylation are identical as described.

> x<-c(0.21, 0.13, 0.21, 0.22, 0.25, 0.21, 0.23, 0.26, 0.16, 0.21, 0.27, 0.22, 0.24, 0.23, 0.44, 0.25, 0.26, 0.20, 0.16, 0.23, 0.49, 0.60, 0.55, 0.53, 0.49, 0.76, 0.61, 0.32, 0.84, 0.32, 0.29, 0.37, 0.55, 0.58, 0.37, 0.57, 0.37, 0.60, 0.27, 0.44, 0.23)

> qpois(0.0001,x,lower.tail=FALSE)

[1] 3 3 3 3 4 3 3 4 3 3 4 3 4 3 4 4 4 3 3 3 5 5 5 5 5 6 5 4 6 4 4 4 5 5 4 5 4 5 4 4 3

It is unclear to me why the regular 95% cutoff is not used in determining the threshold. In the paper, the author claim the models built on thresholds calculated from probabilities of 0.999, 0.9999, 0.99999,0.99999 are all highly correlated, indicating the robustness of the model.

After the binary transformation of the data, for every interval, a 41 member vector V(v1,v2,…vm), each representing one of the 41 known modifications, is generated with a value 1 for a presence call and 0 for an absence call.

Multivariate Hidden Markov Model

K Hidden state from 1 to K and M known histone modifications (1 to 41).

In state k, an emission of each M known modification is modeled with p(k,m), with k from 1 to K and m from 1 to M(41). So for state k, emission of M known modifications is a vector consisting p(k,1), p(k,2), …, p(k,M).

Transition probability from state I to j, b(I,j)