# Identification of Histone Modification Function using Hidden Markov Models

BMI 576 Class Project Final Report

December 9, 2014

**Matthew Long, Minsub Shin, and Quansheng Yang**

**Contributions:**

Matthew Long: write in your contribution

Minsub Shin:write in your contribution

Quansheng Yang: 1) Convert aggregate tag counts to binary data; 2) Calculate the Possion threshold levels for each modification. 3) Implement the forward, backward and EM algorithm. 4) Compare the effect of Possion confidence levels on the hidden states of the 5 states Hidden Markov model. 5) Learn the hidden states for 2, 5, 10 states HMM models. 6) Write and revise the final report.

### Introduction

Recently, much of microbiology has been focused on genetics and the interaction between DNA, RNA and proteins and their impact on the phenotype of a cell or organ. While cellular regulation of RNA and proteins by factors such as promoters, siRNAs, and ribosome binding sites have been studied extensively, focus has turned to epigenetics and the cellular regulation of the primary DNA sequence by histone proteins. The DNA sequence inside of human cells is packaged into chromatin by proteins called histones. In chromatin, DNA sequences of 200 basepairs are wound around 4 histone proteins, thereby limiting the availability of this section of DNA for cellular functions such as transcription. Histone proteins are post-translationally modified, with more than 100 histone modifications having been discovered (Ernst & Kellis, 2010). Specific combinations of those modifications have been shown to be linked with biological functions of the underlining DNA sequence. In addition, these histones have been shown to correlate with RNA expression between differentiated cell types (Heintzman et al., 2009). The goal of this project is to learn biological functions from the known histone modifications and their combinations.

**Approach:**

**Aggregate numbers of tags of each modification into 200 bp interval bins**

Histone modification tags are mapped to 200 bp intervals on chromosomes. The assignment of a tag to an interval is based on whether the sequence of the tag, 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. The tag counts for each modification will be tallied.

**Possion distribution model is used to convert aggregated counts of tags into a binary presence or absence call for the modification in an 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.

When the tag number of a modification in an interval bin is larger or equal to the threshold, a presence call is made. Otherwise, the modification is considered to the absence in the interval. While data for all acetylation are identical as described, a different dataset of methylation, polII and CTCF are used in our analysis. Thus the thresholds for those modifications are slightly different from what are shown in the paper. The thresholds are therefore recalculated for those modifications.

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.

**Hidden states of multivariate Hidden Markov Model and model learning.**

After we obtain the binary data for all 41 modifications, we randomly initiate following hidden states, with a random number generator.

1. Hidden State 1, 2, ……., K (K can be set from 2 to 80 states)
2. Emission probability from a state k to emit a modification represent with a binary data 0 or 1: A random number between 0 and 1.
3. Transition probability from state k to l: a random number is generated and the total sum from one state k to all other states have to sum up to 1.
4. Transition probability to the start of the chromosome: a random number between 0 and 1 and the total must sum up to 1.
5. Transition probability from the end of the chromosome: a random number between 0 and 1 and the total must sum up to 1.

The training of models is based on the forward-backward multivariate Hidden Markov model. The emission probability for an interval is simplified by aggregating individual binomial distribution of 41 modifications by multiplication into a single value. We learn the HMM with 2, 5, 10 states. The hidden states quickly converge after training the models with several chromosomes. It is not a problem for converging for a low number state HMM model.

**Calculation of very small float numbers**

The HMM calculates the probability to observe the chromosome modification state for 41 marks. The observed probability quickly becomes zero if no adjustment to our calculation is made. This is because, for every 200 bp interval, each modification is modeled on binomial distribution, and joined probability of these 41 binomial distributions is calculated by multiplication of individual binomial probabilities, which yields a small value close to ~ 1e-20. Chromosome contains at least millions of such 200 bp intervals, and multiplication millions of times of such small value cannot represent by any float type value of a programming language, with its float numbers in the range of ~1e+/-308. Thus after calculating ~15 interval bins, the number becomes 0.

There are three ways to prevent this problem. 1). First, once the average of emission probability from K states falls below certain threshold such as 1e-50 , the average is to be brought to an acceptable range by multiplying every number in K states by the same scale factor such as 1e+100. Every interval from K states thus associates with a scale factor, and the scale factor has to be factored in the latter numeric calculation. 2). Second, we can store a real number in a class with two variables, a as coefficient, b as the power in a form of a \times 10^b. We then overload numeric operators (+-\*/) and assignment operators(+=, \*=, =) with the this number specific calculation. Thus we can calculate extremely small number up to power(10,-2^64). Third, converting the probability to logarithmic expression. We implement the HMM model with the first and second approaches.

The first approach can be modeled up to 10 states without any problem. But for a 51 state model, the float calculation error will occur. For example, when multiplication of several big number with scale factors, it can exceed the ~1e+308, thus produces a “not a number” error, “nan”. Alternatively multiplication of several very small numbers will lead to a result, zero. Thus the every following calculation with scale factors will have to be rescaled very carefully, which proves to be very challenging. Without rescaling the data either exceed the float limits or become zero, which will cause a “nan” error when calculating power(0,0). So we implement a Number class that can calculate extremely small values. For 2, 5, 10 states models, those two approaches yields identical models, but the latter approach is much more robust for models with larger number of states.

**Results and Disucssion**

We aggregate tag counts for all modifications and the empirical means for modifications as shown in the Methods 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 thresholds are 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, shown below.

> 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

Similarly, the threshold levels for confidence levels 0.99,0.999,0.99999 are also calculated and used in comparison study of the effect of this cutoff threshold on the model’s hidden states as shown below.

**Effect of threshold on the HMM model’s hidden states**

The threshold of setting the histone modification is modeled on Possion. 99.99% confidence level is chosen to learn the model. We assess the effect of the other confidence levels from 99% to 99.999% on the status of model’s hidden states. The hidden states for other confidence levels are learned and correlation is calculated among those models. Correlation is calculated for emission states and transition states.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 1 Correlation of emission states of 5 state HMM models | | | | |
| Conf. indicates the confidence level for Possion distribution of Tag counts | | | | |
| Conf. | 0.99 | 0.999 | 0.9999 | 0.99999 |
| 0.99 | 1 | 0.96 | 0.92 | 0.86 |
| 0.999 | 0.96 | 1 | 0.98 | 0.86 |
| 0.9999 | 0.92 | 0.98 | 1 | 0.85 |
| 0.99999 | 0.86 | 0.86 | 0.85 | 1 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 2 Correlation of transition states of 5 state HMM models | | | | | |
| Conf. indicates the confidence level for Possion distribution of Tag counts | | | | | |
| Conf. | 0.99 | 0.999 | 0.9999 | 0.99999 |  |
| 0.99 | 1 | 0.995 | 0.993 | 0.957 |  |
| 0.999 | 0.995 | 1 | 1 | 0.961 |  |
| 0.9999 | 0.993 | 1 | 1 | 0.959 |  |
| 0.99999 | 0.957 | 0.961 | 0.959 | 1 |  |

It can be seen that models based on 99.9% and 99.99% confidence are the most highly correlated models. Decreasing the confidence to 99% decreases the correlation slightly on emission from 0.98 to 0.92 and almost unchanged on transition from 1 to 0.993. But increasing confidence levels to 99.999% has major adverse effect on emission states from 0.98 to 0.86 and transition from 1 to 0.959. Thus the best confidence levels is 99.9% and 99.99%. Too high or too low is not desirable.

**Effect of number of states on the precision of the HMM model to predict chromosomes’ function state**

We first generate models with 2, 5, 10 states. Then we use Vitebi alogorithm to assign our model’s hidden states to chromosomes.