G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(3):65-70
https://doi.org/10.5808/GI.2018.16.3.65

ORIGINAL ARTICLE

A Short Report on the Markov Property of DNA Sequences on 200-bp Genomic Units of ENCODE/Broad ChromHMM Annotations: A Computational Perspective

Hyun-Seok Park[1,2]*

[1]Bioinformatics Laboratory, ELTEC College of Engineering, Ewha Womans University, Seoul 03760, Korea,
[2]Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul 03760, Korea

The non-coding DNA in eukaryotic genomes encodes a language which programs chromatin accessibility, transcription factor binding, and various other activities. The objective of this short report was to determine the impact of primary DNA sequence on the epigenomic landscape across 200-base pair genomic units by integrating nine publicly available ChromHMM Browser Extensible Data files of the Encyclopedia of DNA Elements (ENCODE) project. The nucleotide frequency profiles of nine chromatin annotations with the units of 200 bp were analyzed and integrative Markov chains were built to detect the Markov properties of the DNA sequences in some of the active chromatin states of different ChromHMM regions. Our aim was to identify the possible relationship between DNA sequences and the newly built chromatin states based on the integrated ChromHMM datasets of different cells and tissue types.

Keywords: chromatin maps, computational epigenetics, Markov chain, noncoding DNA, nucleotide frequency patterns

Availability: We are using OSF (Open Science Framework) pages to host the study design, analysis plan, and data for our article. The datasets generated during the current study are available in the Open Science Framework repository, (https://osf.io/9anpd/).

Introduction

In 2011, the Encyclopedia of DNA Elements (ENCODE) consortium released the ChromHMM chromatin state annotations for 9 consolidated epigenomes, where ChromHMM is software developed by ENCODE labs, to integrate multiple chromatin datasets of various histone modifications to discover *de novo* the major combinatorial and spatial patterns of marks [1, 2]. The 15-chromatin-state model of the ENCODE Project consists of 15 states that are publicly available through 9 Browser Extensible Data (BED) files [3]. Since, large-scale epigenetic datasets such as ENCODE have become publicly available, a growing interest has been shown in predicting the function of non-coding DNA regions directly from sequence by utilizing these large-scale ChromHMM annotations [4-7].

On the other hand, many researchers have shown that formal language theory is an appropriate tool in analyzing various biological sequences [1, 2]. The hidden Markov model (HMM) is most closely related to regular grammars, because an n-gram is a subsequence of n items from a given sequence, and language models that are built from n-grams are actually (n-1)-order Markov models. We therefore proposed n-gram probabilistic language models for predicting the functions of ChromHMM regions of ENCODE [8]. In our previous study, we performed preliminary experiments to test whether the DNA sequences contained in each different chromatin unit of the ENCODE project possess the Markov property by applying Markov chains built from the two BED files of ENCODE tier 1 cell lines (GM12878, a B-lymphocyte lymphoblastoid cell line; and K562, a leukemia cell line) [8]. Our rationale for using the n-gram

Received August 13, 2018; Revised September 6, 2018; Accepted September 18, 2018
***Corresponding author:** Tel: +82-2-3277-3513, Fax: +82-2-3277-2306, E-mail: neo@ewha.ac.kr

model was that each of the sequences contained in the ChromHMM chromatin states can follow a linguistic grammar, not merely as a form of short fragments of motifs or DNA signatures, but as a continuous and longer fragment of sequences.

Our simulation studies showed that some of these chromatin states possessed strong Markov properties of DNA sequences, and could even be predicted by the naïve Bayesian classifier. However, our model could have been biased, as our n-gram analyses were conducted only on two of the cell lines.

Thus, as a follow-up to our preliminary study on ENCODE datasets [8], we extend our previous study and continue our ongoing efforts to build comparative nucleotide frequency profiles to detect Markov properties by analyzing the datasets of the full range of 9 cells and tissue types provided by ENCODE. It was therefore critical to propose a new functional annotation framework that can be generalized to different cell types.

A generalizable framework can be achieved through statistically-justifiable models. We downloaded BED files from ENCODE and combined all the annotations spread out through 9 different BED files, into a single integrated BED file. Based on the newly integrated BED file, we assigned a *dominant* chromatin state for each 200-bp unit. We then rebuilt newer Markov chains by iteratively analyzing the *variability count* of the chromatin states of each 200-bp unit. By eliminating the highly variable 200-bp units, in our simulation studies we finally analyzed the active chromatin states that showed a strong Markov property.

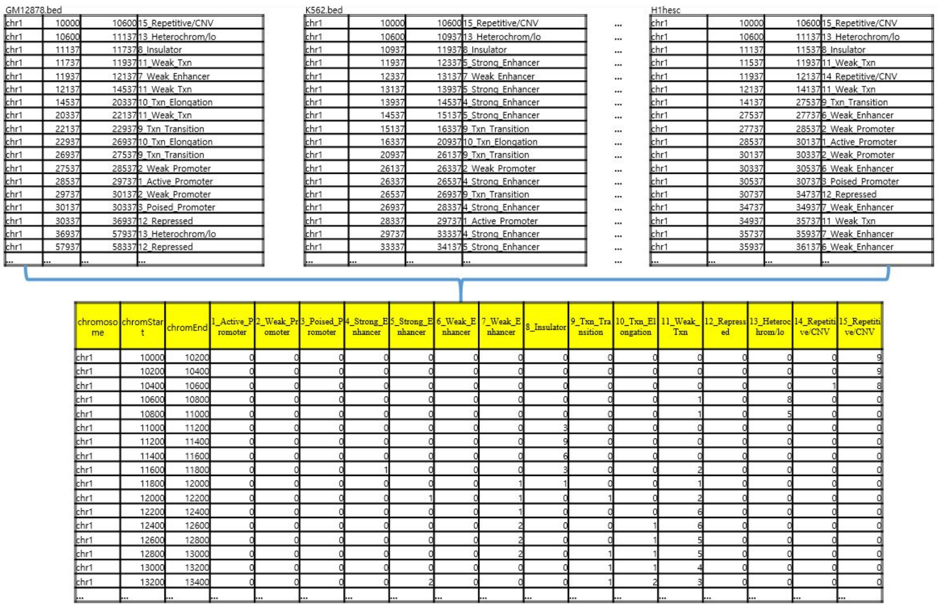
Methods

When making 15-state ChromHMM BED files, the ENCODE consortium uses a core set of 9 chromatin markers [1]. We investigated whether some subsets of the annotated ENCODE 15-state model can be predicted by simply creating n-gram models of DNA sequences, in reverse [9]. To achieve this, ChromHMM blocks of human genome were initially dissected into a nucleosome resolution of 200-bp units and, by analyzing the 9 BED files of ChromHMM, each individual unit was assigned one dominant chromatin state. The process is explained in detail in the following sections: combining 9 BED files into a single file, filtering out highly variable 200-bp units, and finally building 5th order Markov Models.

Combining 9 BED files into a single file

The ENCODE consortium released a 15-state model BED file from an analysis of consolidated epigenomes, resulting in a total of 9 epigenomes for public download in ChromHMM BED files [3].

Fig. 1 shows the chromatin states of the 9 cell lines from chr1: 10,000 to chr1: 30,000 displayed in the University of California Santa Cruz (UCSC) genome browser (with human genome GRCh35/hg19). The BED format shown at the bottom of Fig. 1 provides a flexible way to define the data lines that are shown in an annotation track. The four BED fields shown in each BED file represents *chrom* (name of the chromosome), *chromStart* (starting position of the feature in the chromosome), *chromEnd* (ending position of the feature in the chromosome), and *state* (15 chromatin states, numbered



GM12878.bed

chr1	10000	10600	15_Repetitive/CNV
chr1	10600	11137	13_Heterochrom/lo
chr1	11137	11737	8_Insulator
chr1	11737	11937	11_Weak_Txn
chr1	11937	12137	7_Weak_Enhancer
chr1	12137	14537	11_Weak_Txn
chr1	14537	20337	10_Txn_Elongation
chr1	20337	22137	11_Weak_Txn
chr1	22137	22937	9_Txn_Transition
chr1	22937	26937	10_Txn_Elongation
chr1	26937	27537	9_Txn_Transition
chr1	27537	28537	2_Weak_Promoter
chr1	28537	29737	1_Active_Promoter
chr1	29737	30137	2_Weak_Promoter
chr1	30137	30337	3_Poised_Promoter
chr1	30337	36937	12_Repressed
chr1	36937	57937	13_Heterochrom/lo
chr1	57937	58337	12_Repressed
...	...	...	...

K562.bed

chr1	10000	10600	15_Repetitive/CNV
chr1	10600	10937	13_Heterochrom/lo
chr1	10937	11937	8_Insulator
chr1	11937	12337	5_Strong_Enhancer
chr1	12337	13137	7_Weak_Enhancer
chr1	13137	13937	5_Strong_Enhancer
chr1	13937	14537	4_Strong_Enhancer
chr1	14537	15137	5_Strong_Enhancer
chr1	15137	16337	9_Txn_Transition
chr1	16337	20937	10_Txn_Elongation
chr1	20937	26137	9_Txn_Transition
chr1	26137	26337	2_Weak_Promoter
chr1	26337	26537	4_Strong_Enhancer
chr1	26537	26937	9_Txn_Transition
chr1	26937	28337	4_Strong_Enhancer
chr1	28337	29737	1_Active_Promoter
chr1	29737	33337	4_Strong_Enhancer
chr1	33337	34137	5_Strong_Enhancer
...	...	...	...

...

H1hesc

chr1	10000	10600	15_Repetitive/CNV
chr1	10600	11137	13_Heterochrom/lo
chr1	11137	11537	8_Insulator
chr1	11537	11937	11_Weak_Txn
chr1	11937	12137	14_Repetitive/CNV
chr1	12137	14137	11_Weak_Txn
chr1	14137	27537	9_Txn_Transition
chr1	27537	27737	6_Weak_Enhancer
chr1	27737	28537	2_Weak_Promoter
chr1	28537	30137	1_Active_Promoter
chr1	30137	30337	2_Weak_Promoter
chr1	30337	30537	6_Weak_Enhancer
chr1	30537	30737	3_Poised_Promoter
chr1	30737	34737	12_Repressed
chr1	34737	34937	7_Weak_Enhancer
chr1	34937	35737	11_Weak_Txn
chr1	35737	35937	7_Weak_Enhancer
chr1	35937	36137	6_Weak_Enhancer
...	...	...	...

chromosome	chromStart	chromEnd	1_Active_Promoter	2_Weak_Promoter	3_Poised_Promoter	4_Strong_Enhancer	5_Strong_Enhancer	6_Weak_Enhancer	7_Weak_Enhancer	8_Insulator	9_Txn_Transition	10_Txn_Elongation	11_Weak_Txn	12_Repressed	13_Heterochrom/lo	14_Repetitive/CNV	15_Repetitive/CNV
chr1	10000	10200	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
chr1	10200	10400	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
chr1	10400	10600	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8
chr1	10600	10800	0	0	0	0	0	0	0	0	0	0	1	0	8	0	0
chr1	10800	11000	0	0	0	0	0	0	0	0	0	0	1	0	5	0	0
chr1	11000	11200	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
chr1	11200	11400	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
chr1	11400	11600	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0
chr1	11600	11800	0	0	0	1	0	0	0	3	0	0	2	0	0	0	0
chr1	11800	12000	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0
chr1	12000	12200	0	0	0	0	1	0	1	0	1	0	2	0	0	0	0
chr1	12200	12400	0	0	0	0	0	0	1	0	0	0	6	0	0	0	0
chr1	12400	12600	0	0	0	0	0	0	2	0	0	1	6	0	0	0	0
chr1	12600	12800	0	0	0	0	0	0	2	0	0	1	5	0	0	0	0
chr1	12800	13000	0	0	0	0	0	0	2	0	1	1	5	0	0	0	0
chr1	13000	13200	0	0	0	0	0	0	0	0	1	1	4	0	0	0	0
chr1	13200	13400	0	0	0	0	2	0	0	0	1	2	3	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Fig. 1. Chromatin states of 9 cell lines from chr21: 33,031,600 to chr21: 33,041,600, shown in University of California Santa Cruz (UCSC) genome browser (GRCh37/hg19): the 15 chromatin states shown in the 4th field are numbered and abbreviated as: 1_Active_Promoter, 2_Weak_Promoter, 3_Poised_Promoter, 4_Strong_Enhancer, 5_Strong_Enhancer, 6_Weak_Enhancer, 7_Weak_Enhancer, 8_Insulator, 9_Txn_Transition, 10_Txn_Elongation, 11_Weak_Txn, 12_Repressed, 13_Heterochrom/lo, 14_Repetitive/CNV, and 15_Repetitive/CNV. These are the probabilistic categories based solely on the nine chromatin marks [1, 2].

from 1 to 15).

For example, the chromatin state of Gm12878 shown at the bottom of Fig. 1, for the block from chr1: 10,600 to chr1: 11,137, is 13_Heterochrom/lo, whereas the chromatin state of K562, for the block from chr1: 10,937 to chr1: 11,937, is 8_Insulator.

We attempted to build comparative nucleotide frequency profiles to detect their Markov property. Thus, it became critical to devise a functional annotation framework that can be generalized to different cell types. To design good predictive models in building the Markov chain atlas of the human genome, we modified the original BED files by dissecting the ChromHMM blocks in each BED file into 200-bp units. When the size of a dissected unit near the ChromHMM boundary is less than 150-bp, we discarded the unit, whereas when the size of dissected unit was greater than 150-bp, we rounded it up to a 200-bp unit.

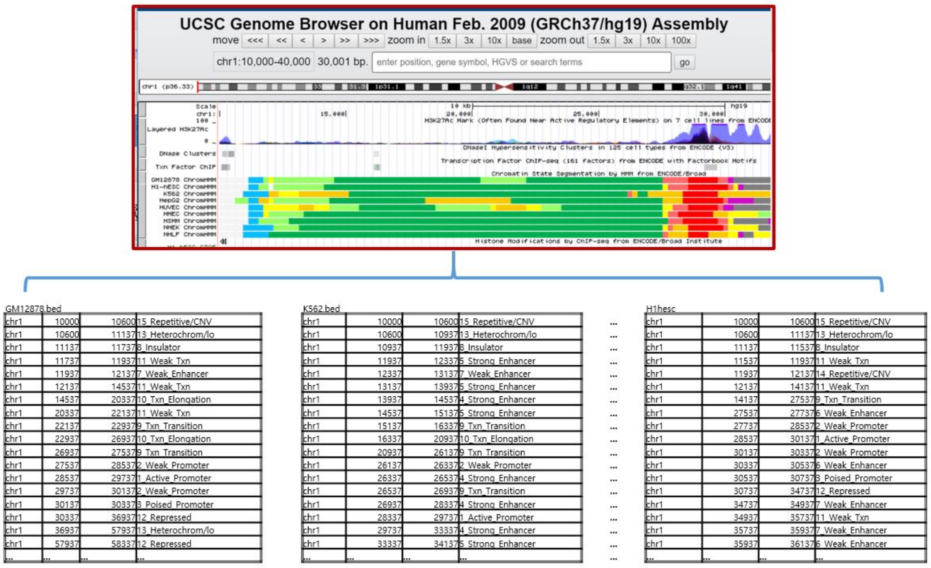
For example, the original Gm12878 block in Fig. 1, from chr1: 10,600 to chr1: 11,137 (a block size of 537 bp), was dissected into two units of 200-bp blocks (from chr1: 10,600 to chr1: 10,800; from chr1: 10,800 to chr1: 11,000), in a new BED file, by discarding the last unit. Likewise, the original K562 block in Fig. 1, from chr1: 10,937 to chr1: 11,937 (a block size of 1,000 bp), was dissected into five units of 200-bp blocks (from chr1: 11,000 to chr1: 11,200; from chr1: 11,200 to chr1: 11,400; from chr1: 11,400 to chr1: 11,600; from chr1: 11,600 to chr1: 11,800; and from chr1: 11,800 to chr1: 12,000), by rounding up the last unit. Profiling nucleotide frequency tables into units of 200-bp is a convenient way to build a general framework and test various Markov properties simply by combining these 200-bp frequency tables in various ways for specific purposes.

Dissecting the blocks uniformly made it possible to combine all the annotations spread out through 9 different BED files into a single integrated BED file [10], as shown in Fig. 2. Each row of the integrated BED file shown at the bottom of Fig. 2 is composed of eighteen entries: chromosome number, starting block number, ending block number, and the remaining fifteen entries that show the number of annotation frequencies of each of the chromatin states, in the original BED files. For example, the chr1: 12,800–13,000 unit at the bottom of Fig. 2 shows that this specific 200-bp unit is annotated two times as state 7 (Weak_Enhancer), one time as state 9 (Txn_Transition), one time as state 10 (Txn_Elongation), and 5 times as state 11 (Weak_Txn) throughout the original 9 BED files, whereas all of the occurrence count numbers of the remaining chromatin states for this unit are zero (in 1, 2, 3, 4, 5, 6, 8, 12, 13, 14, and 15 states).

Filtering out highly variable 200-bp units

After integrating the BED files, we defined the *variability count* of the chromatin states of a given 200-bp unit as the number of states where counts of occurrences were non-zeroes, to define and compare the observed consistency of each chromatin state at any given genomic position across all 9 epigenomes. For example, the chromatin state variability count of the chr1: 12,800–13,000 block in Fig. 2 would be four, as there are four non-zero states (i.e., 7, 9, 10, and 11), whereas the chromatin state variability count of the chr1: 10,200–10,400 block in Fig. 2 would be one, as there is only one non-zero state.

We could then use variability count statistics and


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
100x
chr1:10,000-40,000
30,001 bp.
enter position, gene symbol, HGVS or search terms
go
chr1 (p36.33)
33
31.3
1p31.1
1q12
q32.1
1q41
Scale
chr1:
100
Layered H3K27Ac
0
10 kb
hg19
15,000
20,000
25,000
30,000
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn Factor ChIP
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Chromatin State Segmentation by HMM from ENCODE/Broad
GM12878 ChromHMM
H1-hESC ChromHMM
K562 ChromHMM
HepG2 ChromHMM
HUVEC ChromHMM
HMEC ChromHMM
HSMM ChromHMM
NHEK ChromHMM
NHLF ChromHMM
Histone Modifications by ChIP-seq from ENCODE/Broad Institute

GM12878.bed

chr1	10000	10600	15_Repetitive/CNV
chr1	10600	11137	13_Heterochrom/lo
chr1	11137	11737	8_Insulator
chr1	11737	11937	11_Weak_Txn
chr1	11937	12137	7_Weak_Enhancer
chr1	12137	14537	11_Weak_Txn
chr1	14537	20337	10_Txn_Elongation
chr1	20337	22137	11_Weak_Txn
chr1	22137	22937	9_Txn_Transition
chr1	22937	26937	10_Txn_Elongation
chr1	26937	27537	9_Txn_Transition
chr1	27537	28537	2_Weak_Promoter
chr1	28537	29737	1_Active_Promoter
chr1	29737	30137	2_Weak_Promoter
chr1	30137	30337	3_Poised_Promoter
chr1	30337	36937	12_Repressed
chr1	36937	57937	13_Heterochrom/lo
chr1	57937	58337	12_Repressed
...	...	...	...

K562.bed

chr1	10000	10600	15_Repetitive/CNV
chr1	10600	10937	13_Heterochrom/lo
chr1	10937	11937	8_Insulator
chr1	11937	12337	5_Strong_Enhancer
chr1	12337	13137	7_Weak_Enhancer
chr1	13137	13937	5_Strong_Enhancer
chr1	13937	14537	4_Strong_Enhancer
chr1	14537	15137	5_Strong_Enhancer
chr1	15137	16337	9_Txn_Transition
chr1	16337	20937	10_Txn_Elongation
chr1	20937	26137	9_Txn_Transition
chr1	26137	26337	2_Weak_Promoter
chr1	26337	26537	4_Strong_Enhancer
chr1	26537	26937	9_Txn_Transition
chr1	26937	28337	4_Strong_Enhancer
chr1	28337	29737	1_Active_Promoter
chr1	29737	33337	4_Strong_Enhancer
chr1	33337	34137	5_Strong_Enhancer
...	...	...	...

...

H1hesc

chr1	10000	10600	15_Repetitive/CNV
chr1	10600	11137	13_Heterochrom/lo
chr1	11137	11537	8_Insulator
chr1	11537	11937	11_Weak_Txn
chr1	11937	12137	14_Repetitive/CNV
chr1	12137	14137	11_Weak_Txn
chr1	14137	27537	9_Txn_Transition
chr1	27537	27737	6_Weak_Enhancer
chr1	27737	28537	2_Weak_Promoter
chr1	28537	30137	1_Active_Promoter
chr1	30137	30337	2_Weak_Promoter
chr1	30337	30537	6_Weak_Enhancer
chr1	30537	30737	3_Poised_Promoter
chr1	30737	34737	12_Repressed
chr1	34737	34937	7_Weak_Enhancer
chr1	34937	35737	11_Weak_Txn
chr1	35737	35937	7_Weak_Enhancer
chr1	35937	36137	6_Weak_Enhancer
...	...	...	...

Fig. 2. Combining the 9 Browser Extensible Data (BED) files into an integrated single file: the annotations are contained in nine separate BED files: embryonic stem cells (H1hesc.bed), erythrocytic leukaemia cells (K562.bed), B-lymphoblastoid cells (Gm12878.bed), hepatocellular carcinoma cells (Hepg2.bed), umbilical vein endothelial cells (Huvec.bed), skeletal muscle myoblasts (Hsmm.bed), normal lung fibroblasts (Nhlf.bed), normal epidermal keratinocytes (Nhlf.bed), and mammary epithelial cells (Hmec.bed) [1, 2].

maximum likelihood decision rule to create an optimal classification Markov model, as uniform priors can be assumed if 200-bp units with stable chromatin state are used. In this way, the highly variable 200-bp units, with which different chromatin states were frequently switched to other states across different tissues and cell types, could be eliminated in the training Markov transition tables. Our rationale behind this was that, compared to the highly variable 200-bp units, the 200-bp units that are less frequently changed would show a strong Markov property.

When the human genome was dissected into a 200-bp unit, there were originally 14,148,124 units (see Table 1). Among these units, all of the variability counts of the chromatin states of 5,721,116 units were one, indicating that all 9 cell lines were annotated with the same state in these 200-bp units, and the variability counts of the chromatin states of 4,557,108 units were two. It also indicates that all of the 9 cell lines were annotated as either of the two states in these 200-bp units.

This means that most of these 200-bp units have strong preferences for certain dominant chromatin state, where a dominant state of a 200-bp unit is the most frequently annotated chromatin state among the 15 chromatin states. This provided good heuristic insight for designing new Markov models for our study. Thus, it was possible to assign only one or two dominant chromatin states for most of the 200-bp units of the entire human genome.

Building fifth order Markov models

After we assigned a dominant chromatin state for each 200-bp unit, frequency counts were used to build fifteen initial transition tables for the fifth order Markov models [10]. For example, a uniform fifth order Markov chain is specified by a vector with initial probabilities $P(X_{n-5}, X_{n-4}, X_{n-3}, X_{n-2}, X_{n-1})$ for 4,096 components as well as a matrix of transitional probabilities $P(X_n \mid X_{n-5}, X_{n-4}, X_{n-3}, X_{n-2}, X_{n-1})$ with a size of 4,096 × 4. These tables were used to build a global Markov chain classifier to explore and rank sub-optimal predictions of the chromatin states. Based on the nucleotide frequency profiles, given a random sequence $x_1, x_2, \cdots, x_{200}$ in the state of a cell line, we compared sequences $\pi_1, \pi_2, \cdots, \pi_{200}$ of chromatin states that maximized the following probability of the initial 15 Markov chain models, where $a_{\pi i \pi i+1}$ is a transition probability:

$$P(x, \pi) = \alpha_{0\pi_1} \prod_{i=1}^{200} \alpha_{\pi_i \pi_{i+1}}$$

By trial and error, we rebuilt newer Markov chains by iteratively analyzing the *variability count* of the chromatin states of a given 200-bp unit, and by eliminating the highly variable 200-bp units in training.

Fig. 3 summarizes our process of building Markov chains. When the human genome was dissected into 200-bp units, there were originally 14,075,448 units. By trial and error, we rebuilt newer Markov chains by eliminating the highly variable 200-bp units in training. We finally excluded 200-bp units that showed more than two different chromatin state signatures when training our transition tables. Thus, our result is based on 7,038,863 units, which accounted for approximately 49.75% of the entire human genome. However, determining whether the remaining 50.25% of highly variable 200-bp units of the genome would show a Markov property is beyond the scope of this paper.

By this process, we found that some inactive chromatin states were highly constitutive and marked in most of the 9 epigenomes. For example, state 13 (Hetero_Chromatin state), which covered on average 70.48% of each reference epigenome, was excluded when considering the variability count of the chromatin states. We also excluded units in which a transcribed state showed both promoter and enhancer signatures. Mostly, we profiled each 200-bp with chromatin states and built new transition tables by training

Table 1. Statistical distribution of variability counts of each chromatin state of 200-bp units

Variability counts	Annotation frequencies
1	5,721,116
2	4,557,108
3	2,534,396
4	934,644
5	311,234
6	77,050
7	11,658
8	890
9	28

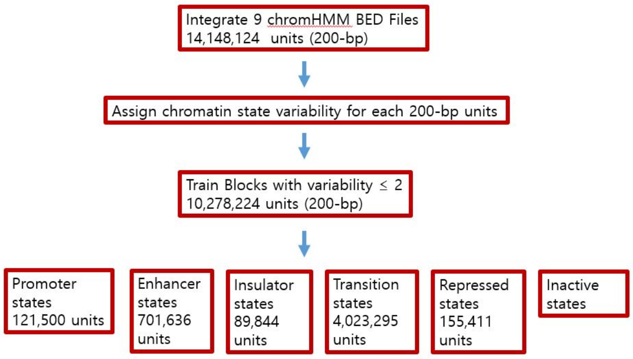

Integrate 9 chromHMM BED Files
14,148,124 units (200-bp)
Assign chromatin state variability for each 200-bp units
Train Blocks with variability ≤ 2
10,278,224 units (200-bp)
Promoter states 121,500 units
Enhancer states 701,636 units
Insulator states 89,844 units
Transition states 4,023,295 units
Repressed states 155,411 units
Inactive states

Fig. 3. Flowchart of building Markov chains by iteratively eliminating highly variable 200-bp units.

the 200-bp blocks with a chromatin variability of less than 2 (and containing at least one active state).

These fifteen chromatin states were then merged into six broad states: *Promoter*, *Enhancer*, *Insulator*, *Transition*, *Repressed*, and *Inactive*. Our final transition tables for the *Promoter*, *Enhancer, Insulator, Transition* and *Repressed* state (excluding inactive states) were built from 121,500, 701,636, 89,844, 4,023,295, and 155,411 200-bp units, respectively. As these Markov chains could be used as a Naive Bayes classifier, we calculated the sequence of each 200-bp unit that maximized our Markov models. We defined a correctly predicted unit as one in which the predicted result matched one of the dominant chromatin states in the same broad state.

Results

As a means to proving Markov property, we directly investigated whether our sequence-based Markov chain models for each chromatin state have the discriminating power necessary to identify different chromatin states.

The samples were stratified according to chromosomes into strictly non-overlapping training and testing sets. A total of 6,334,977 200-bp units were trained, and 703,886 200-bp units were tested for prediction accuracy. At this time, reverse complements of sequences were not considered when building the Markov models, since the backward Markov chain could show similar properties for solutions.

Table 2 shows the result: 52.86% precision for *Promoter* states, 37.95% precision for *Transcribed* states, and 59.82% for *Enhancer* states. These percentages were obtained by adding all units that were predicted correctly as a dominant state in each of the 200-bp units divided by the number of all testing units in the same broad group.

By estimating the prediction accuracy of chromatin states, we infer that the *Promoter* states showed reasonable Markov property, the *Repressed* state did not seem to display Markov property, and those units related to the *Enhancer* states (4, 5, 6, and 7 states) were the most tissue specific, whereas those related to the *Transcription* states (9, 10, and 11 states) were highly constitutive.

Discussion

In this short report, we did not provide any interpretable biological meanings for our statistically defined *dominant* state, yet. Therefore, our study should only be considered from a computational perspective, and, is thus a preliminary work. Still, it is important to note that we only used DNA sequences contained in the epigenetic datasets in modeling the Markov chains. Our study showed that once a dominant state for each 200-bp unit is assigned, a generalizable Markov framework can be achieved. Based on the framework, we showed that some subsets of the active chromatin states possessed a strong Markov property. We are currently investigating the overall co-occurrence of the 200-bp chromatin states for ENCODE ChromHMM datasets together with Roadmap Genomics datasets [11].

Table 2. Prediction accuracy of newly built transition tables of six broad states by analyzing the variability of the chromatin states of 9 BED files

Broad chromatin states	Chromatin states	No. of training units	No. of testing units	Prediction accuracy (%) for unit variability ≤ 2
Promoter state	1_Active_Promoter	66,513	7,390	59.42
	2_Weak_Promoter	41,279	4,587	37.74
	3_Poised_Promoter	13,708	1,523	66.57
Enhancer state	4_Strong_Enhancer	53,192	5,910	60.49
	5_Strong_Enhancer	144,691	16,077	62.90
	6_Weak_Enhancer	140,044	15,560	61.38
	7_Weak_Enhancer	363,710	40,412	57.90
Insulator state	8_Insulator	898,44	9,983	27.37
Transition state	9_Txn_Transition	40,417	4,491	26.84
	10_Txn_Elongation	552,758	61,418	35.31
	11_Weak_Txn	3,430,120	381,124	38.51
Repressed state	12_Repressed	1,398,701	155,411	2.30
Inactive state	13_Heterochrom/lo	NA	NA	NA
	14_Repetitive/CNV	NA	NA	NA
	15_Repetitive/CNV	NA	NA	NA

BED, Browser Extensible Data; NA, not available.

ORCID: Hyun-Seok Park: http://orcid.org/0000-0002-1237-8831

References

1. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;473:43-49.
2. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9:215-216.
3. ENCODE. Encode chromatin state segmentation by HMM from broad institute, MIT and MGH. Santa Cruz: UCSC Genome Bioinformatics, Accessed 2018 Aug 30. Available from: http://moma.ki.au.dk/genome-mirror/cgi-bin/hgTrackUi?db=hg18&g=.
4. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012;40:D930-934.
5. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nat Methods* 2014;11:294-296.
6. Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 2015;5:10576.
7. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12:931-934.
8. Lee KE, Park HS. Preliminary testing for the Markov property of the fifteen chromatin states of the Broad Histone Track. *Biomed Mater Eng* 2015;26 Suppl 1:S1917-S1927.
9. Park HS, Galbadrakh B, Kim YM. Recent progresses in the linguistic modeling of biological sequences based on formal language theory. *Genomics Inform* 2011;9:5-11.
10. Park HS. Epigenetic HMM models. Open Science Framework, 2018. Accessed 2018 Aug 30. Available from: https://osf.io/9anpd/.
11. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317-330.