Bio-Medical Materials and Engineering 26 (2015) S1917–S1927
DOI 10.3233/BME-151494
IOS Press

Preliminary testing for the Markov property of the fifteen chromatin states of the Broad Histone Track

Kyung-Eun Lee[a,b] and Hyun-Seok Park[a,b,c,*]
[a]*Bioinformatics Laboratory, Ewha Womans University, Seoul, Korea*
[b]*Ewha Information and Telecommunication Institute, Ewha Womans University, Seoul, Korea*
[c]*Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul, Korea*

Abstract. Epigenetic computational analyses based on Markov chains can integrate dependencies between regions in the genome that are directly adjacent. In this paper, the BED files of fifteen chromatin states of the Broad Histone Track of the ENCODE project are parsed, and comparative nucleotide frequencies of regional chromatin blocks are thoroughly analyzed to detect the Markov property in them. We perform various tests to examine the Markov property embedded in a frequency domain by checking for the presence of the Markov property in the various chromatin states. We apply these tests to each region of the fifteen chromatin states. The results of our simulation indicate that some of the chromatin states possess a stronger Markov property than others. We discuss the significance of our findings in statistical models of nucleotide sequences that are necessary for the computational analysis of functional units in noncoding DNA.

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

1. Introduction

Recent technical advances in computational epigenetics have allowed for junk DNA to be converted into highly annotated functional areas [1, 2], and the chromatin architecture has been described via genome-wide maps of histone modifications that reveal that a large portion of the genome has different chromatin states. However, since epigenetics is the study of traits not caused by changes in the DNA sequence, little has been published to explore nucleotide sequence patterns associated with genome-wide chromatin maps, except for several studies that analyze a collection of putative human boundary elements [3-5].

To the best of our knowledge, the Markov property of nucleotide sequences associated with genome-wide chromatin maps has never been tested in the literature. Thus, the aim of this paper is to construct comparative nucleotide frequency profiles of the different chromatin states and to investigate whether the nucleotide sequences in the published chromatin maps possess the Markov property. A series of tests are conducted for Markovian dynamics across the chromatin states of the entire human

[*] Address for correspondence: Hyun S. Park, Bioinformatics Laboratory, Ewha Womans University, Seoul, Korea. Tel.: +82)2-3277-3513; Fax: +82)2-3277-2306; E-mail: neo@ewha.ac.kr.

genome to investigate whether these nucleotide frequency profiles can be used as discriminative signatures to classify the fifteen chromatin states while checking for a coincidence with the existing chromatin-assay profiles.

We briefly explain how we extracted the nucleotide sequences of the fifteen chromatin states by parsing the BED files (.bed) of the Broad Histone Track, which is the source of our input data [1, 2, 6]. The BED files of ChromHMM [6] are parsed, and the comparative nucleotide frequencies of the regional chromatin blocks are thoroughly analyzed to detect the Markov property in them. We show that some of the DNA regions that differ in chromatin states also differ in nucleotide frequencies. Then, we check the predictive accuracy of the Markov chain classifiers according to the nucleotide frequency profiles. We also test for homogeneity in the spatial dimension of each of the chromatin states.

2. Parsing the BED files of ChromHMM to extract the nucleotide sequences of the fifteen chromatin states

The human genome contains several genomic maps that contain locations for numerous histone modifications as well as binding sites for various proteins. Of the numerous sources of chromatin annotations from epigenomic studies, we used the ChromHMM dataset, which converts biological assays into discrete annotation maps of fifteen chromatin elements across the human genome. Since our input data source consists of the nucleotide sequences based on the start and end field of the ChromHMM dataset, we explain the dataset in more detail in this section.

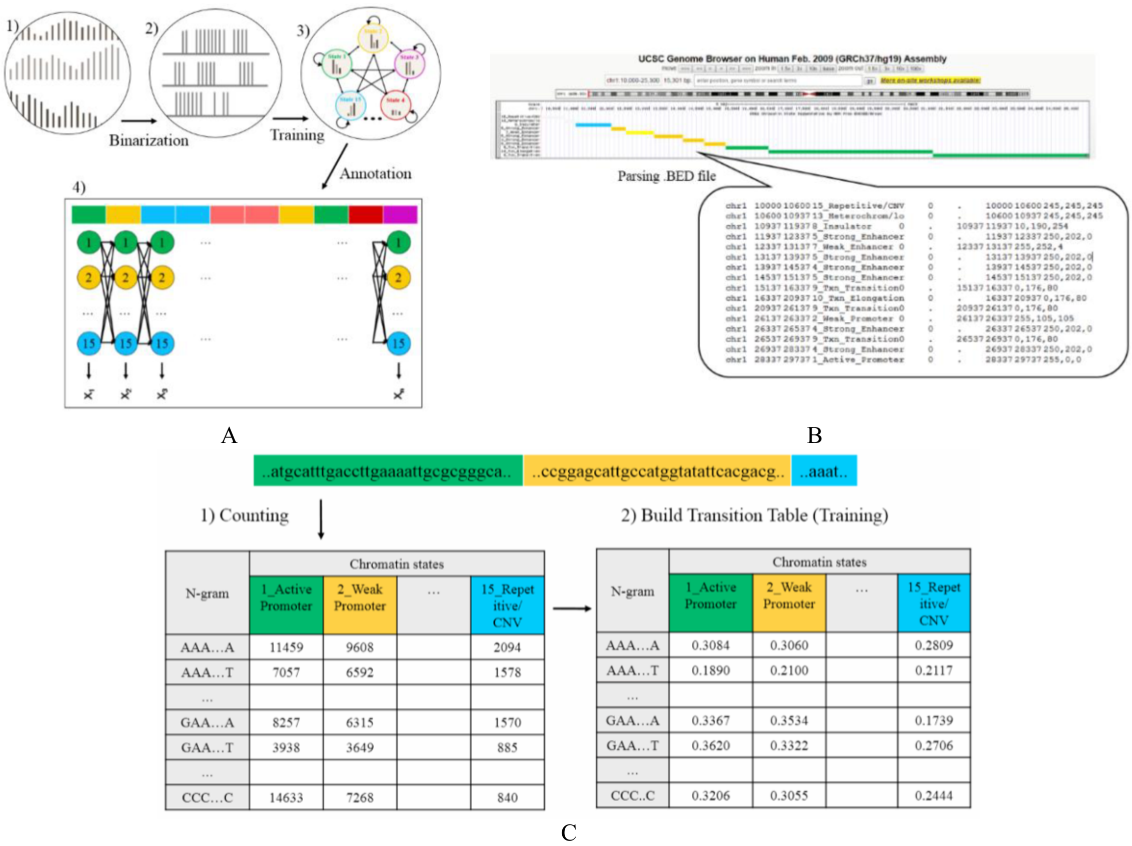
2.1. Parsing the bed files of ChromHMM

Ernst, et al. applied unsupervised learning methodologies to convert the ChIP-seq dataset from the Broad Histone track into discrete annotation maps of fifteen chromatin elements across the human genome [6]. The combinations of histone modifications that are biologically meaningful were discovered by taking an approach that was drastically different from previous approaches in computational epigenetics, particularly in two aspects: hidden states were automatically determined and a massive amount of multivariate data was used for the observed sequences.

Figure 1A summarizes the four steps to build the ChromHMM bed files, including genomic profiling, binarization, model learning, and annotation. The initial input datasets for ChromHMM consist of a list of aligned reads for each chromatin mark (ChIP-seq assays) that were converted into presence calls for each mark across the genome at a 200-base-pair resolution according to a Poisson background model [6]. Finally, each 200-base-pair interval was assigned to its most likely state under the model. The model assumes a fixed number of fifteen distinct hidden states. Figure 1B shows the files resulting from their work hosted on the ENCODE Analysis Data Hub for public download under the ENCODE Data Release Policy (http://genome.ucsc.edu/ENCODE/downloads.html). We downloaded the BED files for the chromatin signal tracks of the ENCODE cell types. Finally, as in Figure 1C, we parsed the BED files and built various nucleotide frequency profiles (with the human genome GRCh35/hg19), which is explained in more detail in the following section.

2.2. Building comparative nucleotide frequency profiles of the 15 chromatin states

The advent of numerous chromatin annotation projects and recent advances in computational


1)
2)
3)
Binarization
Training
Annotation
4)
Parsing .BED file
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
A
B
..atgcatttgaccttgaaaattgcgcgggca..
..ccggagcattgccatggtatattcacgacg..
..aaat..
1) Counting
2) Build Transition Table (Training)
C

```
chr1 10000 10600 15_Repetitive/CNV    0    .    10000 10600 245,245,245
chr1 10600 10937 13_Heterochrom/lo    0    .    10600 10937 245,245,245
chr1 10937 11937 8_Insulator     0    .    10937 11937 10,190,254
chr1 11937 12337 5_Strong_Enhancer    0    .    11937 12337 250,202,0
chr1 12337 13137 7_Weak_Enhancer 0    .    12337 13137 255,252,4
chr1 13137 13937 5_Strong_Enhancer    0    .    13137 13937 250,202,0
chr1 13937 14537 4_Strong_Enhancer    0    .    13937 14537 250,202,0
chr1 14537 15137 5_Strong_Enhancer    0    .    14537 15137 250,202,0
chr1 15137 16337 9_Txn_Transition0    .    15137 16337 0,176,80
chr1 16337 20937 10_Txn_Elongation    0    .    16337 20937 0,176,80
chr1 20937 26137 9_Txn_Transition0    .    20937 26137 0,176,80
chr1 26137 26337 2_Weak_Promoter 0    .    26137 26337 255,105,105
chr1 26337 26537 4_Strong_Enhancer    0    .    26337 26537 250,202,0
chr1 26537 26937 9_Txn_Transition0    .    26537 26937 0,176,80
chr1 26937 28337 4_Strong_Enhancer    0    .    26937 28337 250,202,0
chr1 28337 29737 1_Active_Promoter    0    .    28337 29737 255,0,0
```

N-gram	Chromatin states			
	1_Active Promoter	2_Weak Promoter	…	15_Repetitive/CNV
AAA…A	11459	9608		2094
AAA…T	7057	6592		1578
…				
GAA…A	8257	6315		1570
GAA…T	3938	3649		885
…				
CCC…C	14633	7268		840

N-gram	Chromatin states			
	1_Active Promoter	2_Weak Promoter	…	15_Repetitive/CNV
AAA…A	0.3084	0.3060		0.2809
AAA…T	0.1890	0.2100		0.2117
…				
GAA…A	0.3367	0.3534		0.1739
GAA…T	0.3620	0.3322		0.2706
…				
CCC..C	0.3206	0.3055		0.2444

Fig. 1. An overview of our input data processing and the resulting nucleotide frequency profiles. A. Four Steps of Building ChromHMM BED Files [6]; B. The Format of a Bed File (chr1:10,000-25,300): each BED line has three required fields and nine optional fields; C. Building Nucleotide Frequency Profiles.

epigenetics provide a new opportunity to revisit n-gram statistics. Sequence-level comparisons of the fifteen chromatin states can be valuable resources to investigate whether or not n-gram preferences are conserved across different chromatin states of the human genome [7] since large numbers of integral parameters can be obtained through statistical analyses [8, 9]. We built various nucleotide frequency profiles based on n-gram counts to explore the existence of the Markov property or of genomic signatures in order to enable the classification and grouping of chromatin states purely based on nucleotide sequence information.

It is possible that for the spatially uneven distribution of the nucleotide frequencies in the chromatin states to be inconsistent with the model of a nucleotide sequence that we will be building as a uniform Markov chain. For this reason, it is convenient to profile the nucleotide frequency tables in 200bp units in order to build various combinations of transition tables and to test the various Markov properties of these. For example, if we want to select common regions where chromatin states are not altered in all nine cells, we can simply combine the 200bp regions where all chromatin states of the 9 cell lines are identical. Likewise, we could always combine the pre-built 200bp frequency tables in different combinations to produce the necessary databases or tables to test the population homogeneity of the Markov chains in this way. An indexing technique was used to reduce the preprocessing time and the number of comparisons. To join each of the frequency tables in an efficient manner, hashes of the array subscript values were used.

In summary, we constructed 120,570,489 (13,396,721 units of 200bp frequency tables, multiplied by the 9 cell lines) nucleotide frequency profiles together with information of different chromatin annotations of the 9 cell lines (H1ESC, K562, GM12878, HepG2, HUVEC, HSMM, NHLF, NHEK, and HMEC).

3. Preliminary testing for the Markov property of the 15 chromatin states

In the previous section, we have explained how and why we built the basic 200bp nucleotide frequency profiles. Considering that nucleotide sequences in each of the 15 chromatin states are arranged over chromosomes, the analysis must account for their spatial characteristics. Approaches that are based on Markov chains can integrate dependencies between regions that are directly adjacent in the genome. The key property of a Markov chain is that the probability of each symbol Xi depends only on the value of the preceding symbol X_{i-1} (i.e., $\mathrm{P}(X_i|X_{i-1})$) and not on the entire prior sequence (i.e., $\mathrm{P}(X_i|X_{i-1},\ldots,X_1)$). This can be used to model a random system that changes states according to a transition rule that only depends on the current state.

In this section, we explain how we built the various transition tables of the Markov chains, and this is the basis for how we later determine the chromatin states that possess the Markov property. We also explain how we have tested for homogeneity in the spatial dimension of the fifteen chromatin states of the Broad Histone Track.

3.1. Definition of Markov chain, Markov property, and homogenity

A discrete-time chain X_n: n = 0, 1,… is said to be a Markov chain of order m if it satisfies the following Markov property of order m:

$$P(X_{n+1}=j|X_0=i_0,\ldots,X_{n-1}=i_{n-1},X_n=i_n)=P(X_{n+1}=j|X_{n-m+1}=i_{n-m+1},\ldots,X_n=i_n)$$

A Markov chain is virtually a memoryless chain. If the process is constant over time or space, the Markov chain is completely determined by the following Markov transition matrix that summarizes all N^2 transition probabilities p_{ij}(i, j = 1, …, N) with N as the state space of the chain.

$$\Pi=\begin{bmatrix} P_{11} & \cdots & P_{1N} \\ \vdots & \ddots & \vdots \\ P_{N1} & \cdots & P_{NN} \end{bmatrix}, P_{ij}\geq 0, \sum_{j=1}^{N} P_{ij}=1$$

The Markov property of order 1 (i.e., m = 1) is usually referred to as a strong Markov property. A Markov chain is spatially homogeneous if the transition probabilities do not vary across regions and is population homogeneous if it is not comprised of subpopulations for which the transition probabilities sufficiently differ from one another.

3.2. Building various N-gram transition tables of the 15 chromatin states

Markov chains provide a sound mathematical framework to model and analyze genome-wide nucleotide sequences. Uniform Markov chains of the zeroth, first, and up to the eighth order of the chromatin states of all cell lines were generated by building nucleotide frequency profiles of 1 to 9 grams by combining the 200bp frequency tables that were described in the previous section. We considered the simplest positional statistical model under the assumption that the even the distribution of the nucleotide frequencies in most of the chromatin states follow the model of a nucleotide sequence in the form of a uniform Markov chain.

From the training set, the frequency of the occurrence of all possible substring patterns of length 1 to *8* were tabulated for all 15 chromatin states of all 9 cell lines by joining the necessary tables of the 200bp frequency tables. In general, 4^{k+1} probabilities need to be identified in a kth-order Markov model. The problem that arises in building transition tables is that the number of probabilities that should be estimated from the data increases exponentially when higher order models are present. For example, a uniform sixth order Markov chain is specified by a vector with initial probabilities $\mathrm{P}(X_{n-6,}X_{n-5}, X_{n-4}, X_{n-3}, X_{n-2}, X_{n-1})$ for 42496 components as well as a matrix of transitional probabilities $\mathrm{P}(X_n | X_{n-6}, X_{n-5}, X_{n-4}, X_{n-3}, X_{n-2}, X_{n-1})$ with a size of 42496 x 4.

3.3. Testing for the homogeneity in the spatial dimension

Before making predictions using the Markov chain models, we must ensure that the process does indeed possess the Markov property. Despite innumerable studies of the Markov property, only a small number of tests are available in the literature to verify such an assumption [10-13]. Here, we employed a chi-square test to evaluate the reliability of the estimated transition matrices, focusing particularly on the homogeneity and independence of the chromosomes over the spatial dimensions.

In principle, the homogeneity in the spatial dimension can be checked by dividing all of the samples into several geographically different regions and testing whether the transition matrices that are estimated from each of the sub-samples differs significantly from the matrix that is estimated from all samples. Specifically, it tests the null hypothesis $H_0: \forall t: p_{ij}(t) = p_{ij}(t = 1, \ldots, R)$ against the alternative hypothesis $H_a: \exists r: p_{ij}(t) \neq p_{ij}$ of transition probabilities that differ between different regions using the following chi-square statistics:

$$CS^{(R)} = \sum_{r=1}^{R} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\left(P_{ij}(r) - P_{ij}\right)^2}{P_{ij}}$$

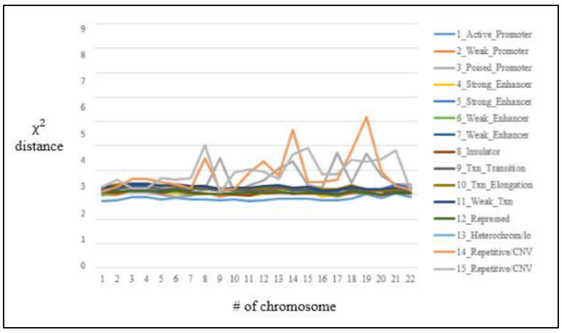


Fig. 2. An exemplary chi-square test against each of the transition table~~s~~ of a single chromosome against the transition table of the entire genome of the k562 cell line.

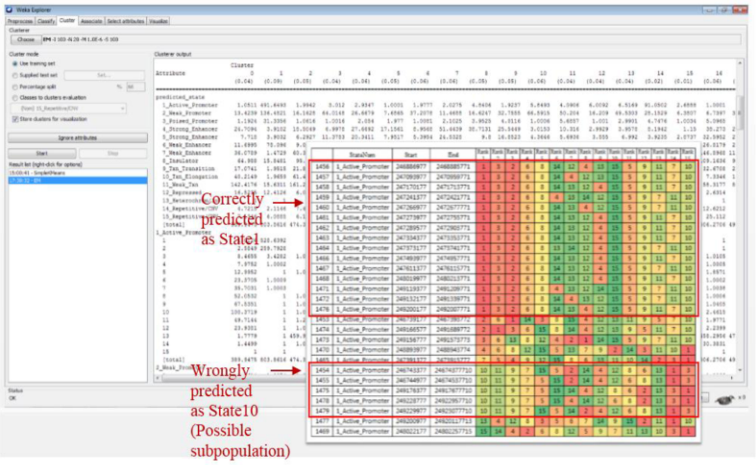
where P_{ij} denotes the probability of a transition from the ith to the jth class that is estimated from the entire sample (pooled across all R regions), and $p_{ij}(r)$ denotes the corresponding transition probability estimated from the rth sub-sample (i.e., training samples in the rth chromosome in our case).

Figure 2 shows an example of a series of chi-square tests of transition tables of the k562 cell line that were built from each chromosome against the 5th order Markov transition table of the entire chromosome. Most of the chromatin states exhibited relatively stable chi-square distances, with the exception of the chromatin states 3, 14, and 15, which showed relatively larger chromosomal variances. The relatively smaller numbers of available training sequences of state 3 (6266 sequences), 14 (7027 sequences), and 15 (7989 sequences) are postulated to have caused larger variances, compared to the average number of sequences of the overall states (approximately, 41,483 sequences).

3.4. Preliminary tests for the existence of subpopulations within the 15 chromatin states

A Markov chain is population homogeneous if it is not comprised of subpopulations in which the transition probabilities sufficiently differ from one another. The transition probabilities are estimated simultaneously from the entire sample, and those that are estimated from sub-samples obtained by dividing the entire sample into mutually independent groups of observations have to be scrutinized.

For this purpose, we used a couple of publicly available classification and clustering tools (such as Weka: http://www.cs.waikato.ac.nz/ml/weka/ or Classias: http://www.chokkan.org/software/classias/), to investigate the possible existence of the subpopulation within the 15 existing chromatin states. Figure 3 shows the exemplary clustering results obtained using Weka for the testing sequences (the sequences of the chromatin state 1 of the K562 cell line) based on the EM (Expectation Maximization) algorithm. Clearly, some of the 15 chromatin states were comprised of subpopulations in which the transition probabilities sufficiently differ from the original states and distinct Markov models have to be built on each subset of the learning data associated with a specific class. Unfortunately, the number of training sequences was not sufficient to construct new transition tables for groups of possible



	StateNam	Start	End	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10	Rank 11	Rank 12	Rank 13	Rank 14	Rank 15
1456	1_Active_Promoter	246886977	2468885771	1	3	2	6	8	14	12	4	13	15	5	9	11	7	10
1457	1_Active_Promoter	247093977	2470959771	1	3	2	6	8	14	4	12	13	15	5	9	11	7	10
1458	1_Active_Promoter	247170177	2471715771	1	3	2	6	8	14	13	12	4	15	5	9	11	7	10
1459	1_Active_Promoter	247241377	2472421771	1	3	2	6	8	4	13	14	12	15	5	9	7	11	10
1460	1_Active_Promoter	247266977	2472677771	1	3	2	6	8	14	13	4	12	15	5	9	7	11	10
1461	1_Active_Promoter	247273977	2472755771	1	3	2	6	8	14	13	12	4	15	5	9	11	7	10
1462	1_Active_Promoter	247289577	2472905771	1	3	2	6	8	14	13	12	4	15	5	9	11	7	10
1463	1_Active_Promoter	247334377	2473353771	1	3	2	6	8	14	13	12	4	15	5	9	11	7	10
1464	1_Active_Promoter	247373177	2473741771	1	3	2	6	8	13	14	12	4	15	5	9	7	11	10
1466	1_Active_Promoter	247493977	2474957771	1	3	2	6	8	14	13	12	4	15	5	9	11	7	10
1467	1_Active_Promoter	247611377	2476115771	1	3	2	6	8	14	13	12	4	15	5	9	7	11	10
1468	1_Active_Promoter	248019977	2480215771	1	3	2	6	8	14	13	12	4	15	5	9	11	7	10
1471	1_Active_Promoter	249119377	2491209771	1	3	2	6	8	4	13	12	14	15	5	9	7	11	10
1472	1_Active_Promoter	249132177	2491339771	1	3	2	6	8	14	4	13	12	15	5	9	7	11	10
1476	1_Active_Promoter	249200177	2492007771	1	3	2	6	8	14	13	4	12	15	5	9	11	7	10
1453	1_Active_Promoter	[illegible]	[illegible]	2	6	1	14	3	8	15	4	12	13	11	9	5	7	10
1474	1_Active_Promoter	249166577	2491689772	2	1	3	6	15	8	14	4	12	13	9	5	11	7	10
1473	1_Active_Promoter	249156777	2491573773	3	6	13	8	12	4	2	1	14	15	5	9	7	11	10
1470	1_Active_Promoter	248893977	2488943774	4	6	8	12	15	5	13	7	9	2	14	3	11	10	1
1465	1_Active_Promoter	[illegible]	[illegible]	7	5	4	9	12	15	8	6	13	11	10	14	2	3	1
1454	1_Active_Promoter	246743377	24674377710	10	11	9	7	15	5	2	14	4	12	8	6	13	1	3
1455	1_Active_Promoter	246744977	24674537710	10	11	9	7	5	15	2	14	4	12	6	8	13	1	3
1475	1_Active_Promoter	249176377	24917677710	10	11	9	7	5	15	14	4	12	8	6	2	13	3	1
1478	1_Active_Promoter	249228777	24922957710	10	11	9	7	5	15	4	14	12	6	8	2	13	3	1
1479	1_Active_Promoter	249229977	24923077710	10	11	9	7	15	5	14	2	4	12	6	8	13	1	3
1477	1_Active_Promoter	249200977	24920117713	13	4	12	8	3	5	6	7	14	9	15	2	11	1	10
1469	1_Active_Promoter	248022177	24802257715	15	14	4	2	6	8	12	5	9	7	11	13	10	3	1

Fig. 3. EM (expectation maximization) clustering results of the chromatin states, using weka 3 (http://www.cs.waikato.ac.nz/ml/weka/). The table in the middle shows the rankings of the fifteen Markov models, based on the preliminary prediction accuracy evaluation of the sequences of the chromatin state 1 of the K562 Cell Line. For example, the 1456th row of the table (start: 246886977, end: 246888577) shows that the Active Promoter state of the K562 Cell Line is correctly predicted as chromatin state 1 while the 1475th row of the table (start: 249176377, end: 249176777) shows that it is wrongly predicted as the chromatin state 10.

subpopulations. The existence of the subpopulations certainly affected the results from the evaluation of the prediction accuracy, as described in the next section.

4. Prediction accuracy evaluation of the fifteen chromatin states built from the tier 1 cell types

In the previous section, we tested whether each of the fifteen chromatin states of the Broad Histone Track have the Markov property. In this section, we are interested in using the Markov chains to sequence the prediction rather than classification. The Markov chains, also known as n-grams, form a generalization of a Naive Bayes classifier when longer histories are considered. We present representative results for applying the third, fourth, fifth, and sixth order Markov chains built from the common regions of the ENCODE Tier 1 cell types, including erythrocytic leukemia cells (K562) and B-lymphoblastoid cells (GM12878). When the Markov chains are used to make a prediction, the quality of the modeling is generally measured by computing the (log-) likelihood. These values measure how well the model fits the underlying unknown process distribution that is represented with independent test samples, and better models can offer higher test likelihoods.

4.1. Prediction accuracy evaluation using fixed-order Markov chains

Though there is some skepticism for the use of chi-square statistics to test for the Markov property under certain conditions [14], we directly investigated whether our sequence-based Markov chain models for each chromatin state have the discriminating power necessary to identify different chromatin states, as described in Section 3.2. We used the frequency counts to build fifteen preliminary transition tables for the third to sixth order Markov models. To this end, we developed a global Markov chain classifier to provide values for the stochastic evaluation in order to explore and rank the sub-optimal predictions.

Thus, we initially built 540 (15 chromatin states x 4 n-gram order x 9 cell lines) Markov chain models for each chromatin state. Given a random sequence $x_1, x_2, \dots, x_k$ in the state S_n, we calculated the sequence $\pi_1, \pi_2, \dots, \pi_k$ of hidden states that maximize P[x, |M], where M is one of the fifteen Markov chain models by using transition probabilities $a_{\pi_i \pi_{i+1}}$, as follows:

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

It is more complex to interpret the nucleotide frequency information contained in each chromatin state than to interpret a large vector with numerical values of different biochemical assays. We randomly selected training blocks and systematically increased the percentage of the testing blocks when training the classifiers. We used the remaining blocks as our testing blocks to naturally exclude the regions of testing blocks when each of the transition tables was built. To prepare for the test set of the sequences, random DNA fragments were chosen from each of the fifteen chromatin states.

We explored whether this type of analysis would enable classification and grouping for the chromatin states according to the similarities in the numerous n-gram counts and to whether preferences were conserved across different chromatin states for a given chromosome. Since the spatial variations of the Markov chains were minimal, as was shown in Section 3.3, training was performed for approximately 59% of the sequences, and we randomly picked the testing sequences

Table 1

Prediction Accuracy Evaluation of 3rd, 4th, 5th, and 6th order Markov Models (Excluding the Chromatin State 13)

15 Chromatin States	# of total blocks	# of training blocks	# of testing blocks	4 gram (Average Ranking)	5 gram (Average Ranking)	6 gram (Average Ranking)	7 gram (Average Ranking)
1_Active_Promoter	8970	4839	1479	63.62%(3.55)	65.86%(3.26)	67.68%(3.18)	34.42%(5.39)
2_Weak_Promoter	17349	10121	3120	4.55%(8.22)	4.10%(7.70)	3.91%(7.73)	8.37%(6.88)
3_Poised_Promoter	3744	2488	1114	26.48%(7.14)	27.47%(6.47)	29.17%(6.18)	34.83%(6.38)
4_Strong_Enhancer	17643	10142	3966	9.78%(5.45)	14.83%(4.87)	18.31%(4.55)	7.59%(6.30)
5_Strong_Enhancer	22735	13455	4829	8.10%(5.40)	11.16%(4.72)	13.11%(4.51)	8.78%(6.20)
6_Weak_Enhancer	34047	20715	5960	1.76%(7.74)	2.10%(7.03)	2.48%(6.73)	9.65%(6.63)
7_Weak_Enhancer	70956	43681	13192	13.63%(5.65)	14.71%(5.08)	15.99%(4.79)	10.37%(6.29)
8_Insulator	23851	14470	3870	6.64%(6.15)	19.92%(5.49)	21.73%(5.18)	6.82%(6.40)
9_Txn_Transition	12547	7562	3372	6.26%(5.03)	6.70%(4.64)	8.63%(4.45)	5.60%(6.21)
10_Txn_Elongation	14614	9080	3120	46.54%(5.44)	46.47%(5.22)	46.09%(5.16)	18.11%(6.23)
11_Weak_Txn	58733	35626	10693	16.38%(5.88)	15.37%(5.63)	13.95%(6.23)	16.62%(7.29)
12_Repressed	23122	14674	4777	5.65%(5.07)	15.53%(4.54)	26.43%(3.99)	7.12%(5.37)
13_Heterochrom/lo	45045	29464	6141	NA	NA	NA	NA
14_Repetitive/CNV	4216	1926	880	15.57%(8.04)	0.00%(15.00)	0.00%(15.00)	4.32%(14.40)
15_Repetitive/CNV	4793	2524	781	3.84%(7.46)	3.20%(7.80)	2.56%(8.69)	11.27%(8.22)

from the remaining 41% in order to evaluate the classification accuracy.

Due to the page limit for this paper, we provide an explanation for only some examples of the prediction accuracy performance of the global Markov classifier built from the common regions of the chromatin signal tracks of the ENCODE Tier 1 cell types (K562 and GM12878), and we thus used this model to assign testing sequences to their most likely state.

Table 1 shows the prediction accuracy performance of the 15 states with the 3rd, 4th, 5th, and 6th order Markov chains. The performance differed greatly, depending on the chromatin states. The test fragments were extracted from state 1 and 10 and were predicted correctly, peaking at the 5th order model (or 6-gram model) with a prediction accuracy of 67.68% and 46.09%, respectively. However, less than 10% of the sequence fragments that were extracted from chromatin states 2, 6, and 9 were predicted correctly for all n-gram tests. Inactive states, such as states 13, 14, and 15, were not population homogeneous. In particular, state 13 needs to be subdivided into many subpopulations, which is beyond the scope of this paper.

For a 6th-order model that requires probabilities for all 7-mers, there is a substantial number of 7-mers that do not occur sufficiently often, and the problem is worse for higher order models. Therefore, a more sophisticated method (such as an interpolated Markov Model [15, 16]) needs to be considered in the future, and the parameters would need to be carefully calibrated to address the data sparseness problem when more training data becomes available.

4.2. Prediction accuracy evaluation of the broader chromatin states

According to Ernst, et al. (2011), the fifteen chromatin states can be represented as six broader classes of chromatin states. These are referred to as the Promoter, Enhancer, Insulator, Transcribed, Repressed, and Inactive states [6]. The Active, Weak and Poised Promoters (states 1-3) differ in their expression levels, the Strong and Weak candidate Enhancers (states 4-7) differ in the expression of their proximal genes, and the Strongly and Weakly Transcribed regions (states 9-11) also differ in their positional enrichment along the transcripts. Similarly, the polycomb-repressed regions (state 12) differ from the heterochromatic and the repetitive/CNV states (states 13-15), which are also enriched for H3K9me3.

Thus, many of the wrongly predicted cases in Table 1 were predicted as other states in the same broad state groups. For example, 33.6% of the sequences extracted from state 11 were wrongly predicted as state 10; 32.5% of the sequences extracted from state 2 were also wrongly predicted as either state 1 (17.66%) or state 3 (14.84%), and 21.3% of the sequences extracted from state 1 were also wrongly predicted as either state 2 (13.18%) or state 3 (8.11%), 33.36% of the sequences extracted from state 11 were also wrongly predicted as a state.

Figure 4 represents the relative chi-square scatter plots of 6-grams of tier 1 cell lines, which confirms that the chromatin states in the same broad group are indeed closer than other broader groups in covariance distance. These general patterns suggest that despite a variation in the activity levels or expression levels, similar chromatin regions tend to preserve their Markov properties as regions of their broader states. Our analysis in Figure 4 suggest that some of the 15 chromatin state models do not have a Markov property that is distinguishable from other chromatin states in the same broad group, and the Markov models might have to be reconstructed according to the 6 broader state models.

Thus, we explored two models with different structures in terms of the means and variances of our states based on the fifteen existing states and the six broader states (when these fifteen states are merged into the six broader Promoter, Enhancer, Insulator, Transition, Repressed, and Inactive states). Table 2 shows the prediction accuracy of these two models. The precision accuracy for both models peaked at the fourth- or fifth-order Markov model.

The prediction accuracy for the individual chromatin states improves for the first model (e.g., Active Promoter state: 85.46%; Strong Enhancer state: 43.12%; and the Txn_Elongation state: 50.25%). That number was obtained by adding all the blocks that were predicted correctly as the chromatin states in the same group, divided by all testing blocks. For example, the Promoter state prediction accuracy was calculated using the following formula:

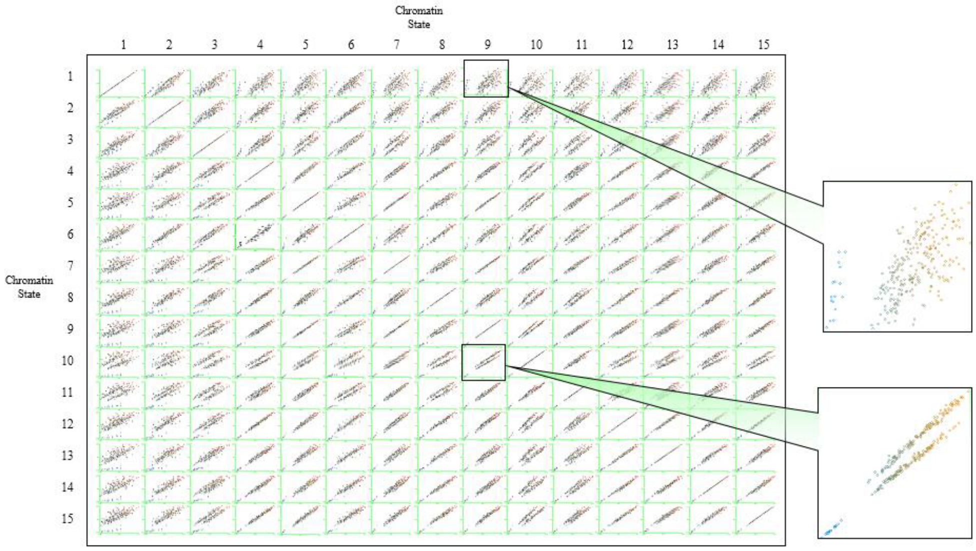


Fig. 4. The 15 x 15 Chi-square scatter plots of the 6-gram transition tables of 15 chromatin states. For example, this shows that the scatter plot between states 1 and 9 belongs to different broader states (Promoter State and Transition State, respectively), is more scattered (wider) than the scatter plot between states 9 and state 10, and belongs to the same broader state (Transition State).

Table 2

The Six Broader Chromatin States and Their Prediction Accuracy (Excluding Chromatin State 13)

Broader Chromatin States	15 Chromatin States	# of total blocks	# of training blocks	# of testing blocks	4gram		5gram		6gram		7gram	
Promoter State	1(Active)	5713	17448	5713	85.80%	50.18%	85.60%	50.48%	85.46%	49.99%	61.66%	49.61%
	2(Weak)				37.05%		37.31%		36.28%		43.88%	
	3(Poised)				39.68%		40.75%		41.29%		49.64%	
Enhancer State	4(Strong)	27947	87993	27947	30.61%	25.07%	37.62%	31.99%	43.12%	34.75%	17.40%	17.64%
	5(Strong)				25.89%		32.78%		35.22%		16.75%	
	6(Weak)				21.04%		26.56%		28.69%		18.44%	
	7(Weak)				24.93%		32.46%		34.80%		18.24%	
Insulator State	8(Insulator)	23851	14470	3870	6.64%	6.64%	19.92%	19.92%	21.73%	21.7%	6.82%	6.82%
Transition State	9(Transition)	85894	52268	17185	48.93%	55.22%	48.16%	53.47%	47.63%	50.25%	32.62%	35.32%
	10(Elongation)				66.86%		65.74%		65.03%		47.47%	
	11(Weak)				53.81%		51.57%		44.84%		26.36%	
Repressed State	12(Repressed)	23122	14674	4777	5.65%	5.65%	15.53%	15.53%	26.44%	26.44%	6.41%	6.41%
Heterochrom State	13(Heterochrom/lo)	45045	29464	6141	NA	NA	NA	NA	NA	NA	NA	NA
Inactive State	14(Repetitive/CNV)	9009	4450	1661	20.11%	16.01%	4.89%	4.09%	5.34%	4.03%	15.11%	13.31%
	15(Repetitive/CNV)				11.40%		3.20%		2.56%		11.27%	

$$\text{The Promoter State Prediction Accuracy} = \frac{\#\ of\ States\ Predicted\ either\ as\ the\ Chromatin\ States\ 1{,}2\ or\ 3}{\#\ of\ all\ the\ States\ 1{,}2\ and\ 3\ testing\ block}$$

We also explored the second model by building new transition tables based on the broader states. As a whole, the broader states also showed reasonable prediction accuracies (Promoter state: 49.99%; Enhancer state: 34.75%; Transition state: 50.25% for the case of the 6-gram). For a 6th-order model (or 7-gram), there is still a substantial number of 7-mers that do not occur sufficiently often, and the prediction accuracy was also influenced by the data sparseness.

5. Conclusion

In this paper, we have provided a conditional characterization of the Markov property of the 15 chromatin states of the Broad Histone Track by performing preliminary tests for the Markov properties of the common regions of the ENCODE Tier 1 cell types. It was not possible to enter into a more detailed discussion of the questions associated with the population homogeneity of our Markov chain models within the framework of this paper, so we note only the following.

We present a pioneering line of research to address the characteristics of the genome-wide Markov properties of nucleotide sequences in noncoding DNA regions. We showed that the patterns of the nucleotides in noncoding DNA are non-uniform for fifteen different chromatin states of the common regions of the ENCODE Tier 1 cell types. We also showed that it is possible to establish Markov chain models to distinguish some of the functionally significant regions. The characteristics of these patterns are postulated to reflect a specific path of evolution, selecting different rules of fixation for mutations in different functional regions. Our study is significant in that it has potential for use to construct special statistical models that are necessary to develop algorithms to build nucleotide-sequence-based classifiers of chromatin states.

Acknowledgments

The work is partially supported by the Small and Medium Business Administration (Project Number: 2-2014-2443-001-1) of Korea and the Interdisciplinary Graduate Research Program in Systems Genomics (Project Number: 2012M3A9D1054744). We also would like to thank Dr. Philippe Meunier of Ewha Womans University, for valuable comments.

References

[1] ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project, Science **22** (2004), 636–640.

[2] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature **489** (2012), 57–74.

[3] J. Wang, V.V. Lunyak and I.K. Jordan, Genome-wide prediction and analysis of human chromatin boundary elements, Nucleic Acids Research **40** (2011), 511–529.

[4] V.V. Lunyak, G.G. Prefontaine, E. Nunez, T. Cramer, B.G. Ju, K.A. Ohgi, K. Hutt, R. Roy, A. Garcia-Diaz, X. Zhu, et al., Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis, Science **317** (2007), 248–251.

[5] S. Cuddapah, R. Jothi, D.E. Schones, T.Y. Roh, K. Cui and K. Zhao, Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains, Genome Research **19** (2009), 24–32.

[6] J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shoresh , L.D. Ward, C.B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis and B.E. Bernstein, Systematic analysis of chromatin state dynamics in nine human cell types, Nature **473** (2011), 43–49.

[7] K.E. Lee and H.S. Park, Identifying genomic signatures of N-gram nucleotide sequences to classify the chromatin states of broad histone track, Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, Indonesia, 2015, 7-14.

[8] T.F. Smith, M.S. Waterman and J.R. Sadler, Statistical characterization of nucleic acid sequence functional domains, Nucleic Acids Research **11** (1983), 2205–2220.

[9] M.Y. Borodovskii, Y.A. Sprizhitskii, E.I. Golovanov and A.A Aleksandrov, Statistical patterns in primary structures of functional regions in the in E. coli genome: 1. Frequency Characteristics, Molecular Biology **20** (1986), 826–833.

[10] T.W. Anderson and L.A. Goodman, Statistical inference about Markov chains, The Annals of Mathematical Statistics **28** (1975), 89–110.

[11] L.A. Goodman, Simplified runs tests and likelihood ratio tests for Markov chains, Biometrica **45** (1958), 181–197.

[12] Y. Ait-Sahalia, Nonparametric tests of the Markov hypothesis in continuous-time models, The Annals of Statistics **38** (2010), 3129–3163.

[13] M. Fernandes and R.G. Flˆores, Nonparametric tests for the Markov property, Getulio Vargas Foundation, 1999.

[14] R.N. Hiscott, Chi-square tests for Markov chain analysis, Mathematical Geology **13** (1981), 69–80.

[15] S.L. Salzberg, A.L. Delcher, S. Kasif and O. White, Microbial gene identification using interpolated Markov models, Nucleic Acids Research **26** (1998), 544–548.

[16] S.L. Salzberg, M. Pertea, A.L. Delcher, M.J. Gardner and H. Tettelin, Interpolated Markov models for eukaryotic gene finding, Genomics **59** (1999), 24–31.