Using Markov chains of nucleotide sequences as a possible precursor to predict functional roles of human genome: a case study on inactive chromatin regions

K.-E. Lee[1], E.-J. Lee[1], and H.-S. Park[1,2]

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

Corresponding author: H.-S. Park
E-mail: neo@ewha.ac.kr

Genet. Mol. Res. 15 (3): gmr.15039004
Received July 21, 2016
Accepted August 1, 2016
Published August 30, 2016
DOI http://dx.doi.org/10.4238/gmr.15039004

Copyright © 2016 The Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution ShareAlike (CC BY-SA) 4.0 License.

ABSTRACT. Recent advances in computational epigenetics have provided new opportunities to evaluate n-gram probabilistic language models. In this paper, we describe a systematic genome-wide approach for predicting functional roles in inactive chromatin regions by using a sequence-based Markovian chromatin map of the human genome. We demonstrate that Markov chains of sequences can be used as a precursor to predict functional roles in heterochromatin regions and provide an example comparing two publicly available chromatin annotations of large-scale epigenomics projects: ENCODE project consortium and Roadmap Epigenomics consortium.

Key words: Chromatin maps; Nucleotide frequency patterns; Markov chain; Noncoding DNA; Computational epigenetics

INTRODUCTION

Our understanding of computational epigenetics and their impact on epigenetic studies has rapidly expanded in recent years as a result of large-scale epigenetic projects. In 2011, the Encyclopedia of DNA Elements (ENCODE) project consortium released discrete annotation maps of chromatin elements from an analysis of the 9 epigenomes by applying unsupervised learning methodologies (Ernst et al., 2011). Later, in 2015, the Roadmap Epigenomics consortium also released datasets from a joint analysis of 111 consolidated epigenomes and an additional 16 epigenomes from the ENCODE project (Roadmap Epigenomics Consortium, 2015), resulting in 127 epigenomes.

The annotations in hg19 resulting from these two consortiums are hosted on the ENCODE Analysis Data Hub (UCSC Genome Bioinformatics, 2013) and on the Roadmap Genomics consortium's supplementary website (Wang Lab at Washington University in St. Louis, 2015), respectively, for public download.

Considering that these datasets provide a vast number of segmented regions on the whole genome scale, a series of recent publications of chromatin maps has provided the opportunity to revisit n-gram language models for analyzing non-coding DNA regions. N-gram analysis of chromatin states of these projects may be valuable resources for investigating whether nucleotide sequences are conserved across the different chromatin states of the human genome, as well as the degree to which they are conserved (Smith et al., 1983; Borodovskii et al., 1986).

N-gram models, most notably Hidden Markov models, have been extensively studied in the bioinformatics field (Yoon, 2009). However, the Markov properties of nucleotide sequences associated with the chromatin states of the whole human genome scale has never been investigated. This is likely because epigenetics is regarded as the study of heritable changes in gene activity that are not caused by changes in the DNA sequence, and thus few studies have explored the nucleotide sequence patterns associated with whole genome-wide chromatin maps. Thus, we recently performed preliminary experiments to test whether each of the different chromatin states of the ENCODE project possesses the Markov properties by applying Markov chains built from the annotation files of ENCODE Tier 1 cell types (Lee and Park, 2015). Our simulation studies showed that some of the chromatin states had stronger Markov properties than other states.

Thus, the aim of this study, which is a follow-up to our previous study (Lee and Park, 2015), is to present the preliminary findings that our model can be used to predict the Markovian chromatin states of functionally unknown regions of the human genome by analyzing the differences in annotations between the ENCODE project consortium (Ernst et al., 2011) and the Roadmap Genomics consortium (Roadmap Epigenomics Consortium, 2015).

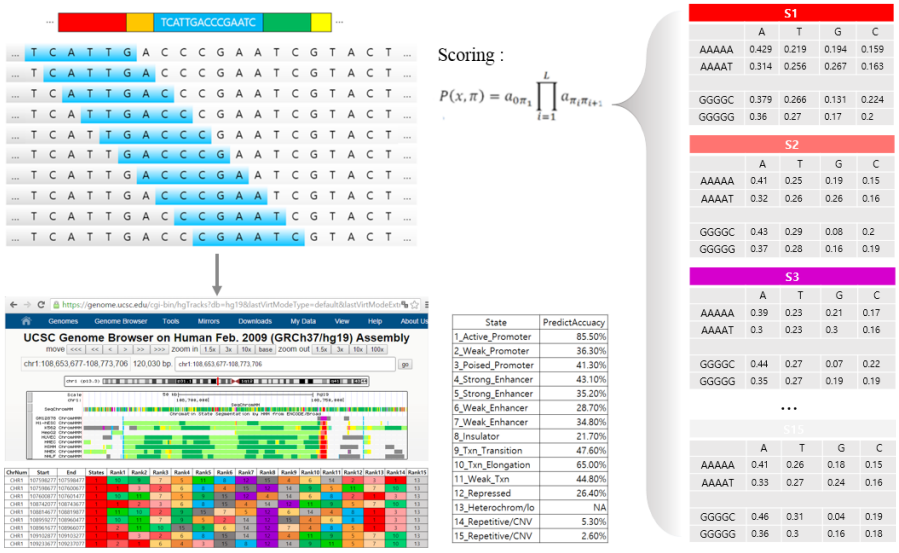
MATERIAL AND METHODS

Building preliminary Markov chains based on a single ChrommHMM BED file of ENCODE

Chromatin is as an instructive information carrier of DNA that can respond to external cues to regulate the many uses of DNA. The organization of chromatin into functionally distinct segments suggests the existence of conserved areas of nucleotide sequence patterns of each chromatin state on the whole genome scale.

In 2011, the ENCODE project consortium published 15 chromatin elements across the human genome (Ernst et al., 2011) by analyzing the data for modified histones. We used the Browser Extensible Data (BED) files generated by ChromHMM software (Ernst and Kellis, 2012; Lee and Park, 2014) to analyze sequence-based profiles to determine the nucleotide sequences in the 15 chromatin states that possess the Markov property (Lee and Park, 2015).

Figure 1 provides an overview of our previous study (Lee and Park, 2015), showing a case of building 15 transition table Markov chains, parsing the BED files of common regions of the 1st tier cell lines of the ENCODE. Initially, nucleotide frequency profiles (with the human genome GRCh35/hg19) were used to build 15 preliminary transition tables for the Markov models, where the 15 chromatin states were active, repressed, and poised promoters (states 1, 2, 3), strong and weak enhancers (states 4, 5, 6, 7), putative insulators (8), transcribed regions (states 9, 10, 11), and large-scale repressed and inactive domains (state 12,13,14,15) (Ernst et al., 2011).



State	PredictAccuacy
1_Active_Promoter	85.50%
2_Weak_Promoter	36.30%
3_Poised_Promoter	41.30%
4_Strong_Enhancer	43.10%
5_Strong_Enhancer	35.20%
6_Weak_Enhancer	28.70%
7_Weak_Enhancer	34.80%
8_Insulator	21.70%
9_Txn_Transition	47.60%
10_Txn_Elongation	65.00%
11_Weak_Txn	44.80%
12_Repressed	26.40%
13_Heterochrom/lo	NA
14_Repetitive/CNV	5.30%
15_Repetitive/CNV	2.60%

Figure 1. Overview of building a global classifier of 15 chromatin states.

We then developed a Markov chain classifier, as the Markov chains can form a Naive Bayes classifier. Figure 1 shows how a chromatin state was predicted based on the nucleotide frequency profiles. Given a random sequence in the state of a cell line, we calculated the sequence of chromatin states that maximize the) among the Markov chain models.

We explored whether this type of analysis would enable the classification of the chromatin states according to similarities in the n-gram counts. The prediction accuracy of each state differed significantly, and the results clearly showed that certain regions benefitted from a stronger Markov property than other regions (Lee and Park, 2015). For example, the prediction accuracy of active promoter blocks (state 1) reached 85%, as shown in Figure 1.

Based on these results, we then created the initial version of a sequence-based Markovian chromatin map, referred to as SeqChromMM and published on the github repository (https://github.com/KyungEunLee/SeqChromMM.git) (Lee and Park, 2016).

Characterizing the overall variability of each chromatin state across the 9 cell lines

In the previous section, we described the overview of our preliminary study (Lee and Park, 2015). However, if we build transition tables based only on cell-line specific models, noise will arise. Thus, the prediction accuracy of our model was trained and measured by only using the information contained in the BED files of erythrocytic leukemia cells (K562) or B-lymphoblastoid cells (GM12878).

However, a total of 9 BED files of ENCODE are publicly available. The other 7 cell lines of ENCODE are embryonic stem cells (H1ES), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblasts (NHLF), normal epidermal keratinocytes (NHEK), and mammary epithelial cells (HMEC) (Ernst et al., 2011) The epigenomic landscape of each cell can vary considerably.

We compared our prediction results in relation to the annotations of the other 7 cell lines (compared together), although the transition tables were built solely from the K562 and GM12878 cell lines (Lee and Park, 2015). Associating our prediction results with nucleotide frequency information contained in each chromatin state of different cell lines is more complex and challenging.

To characterize the overall variability of each chromatin state across the 9 cell lines, the original ChromHMM blocks were uniformly dissected at a nucleosome resolution of 200 base pairs, and each of the 200-bp units was analyzed and assigned a new predicted chromatin state by our Markov classifier.

Next, we measured the consistency of each chromatin state at any given genomic position (200-bp resolution) across all 9 epigenomes together with our prediction results. Figure 2 illustrates some of our prediction results of K562-based Markov chains compared to the other 8 cell lines.

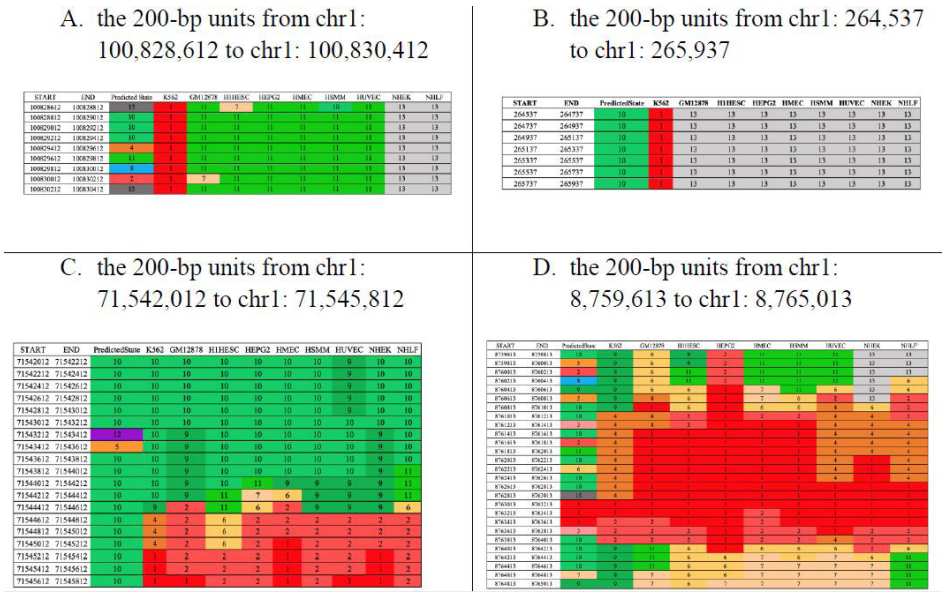


Figure 2. Analysis of the relationship between the predicted states and the 9 cell line chromatin states.

Figure 2A shows the example 200-bp units (from chr1: 100,828,612 to chr1: 100,830,412) where K562 is annotated as state 1 of ENCODE annotation, but our global classifier predicted it as state 10 of ENCODE annotation. However, the other cells were all annotated as 10 or 11 (which are in the same broad group), suggesting that the region has a higher probability of displaying the Markov property of the transition state.

Additionally, the regions similar to that shown in Figure 2B (from chr1: 264,537 to chr1: 265,937) frequently occurred, where K562 is annotated as state 1, but most other cell lines were annotated as inactive states such as the 13 hetero-chromatin states. This is because the coverage of the annotation of K562 was relatively greater than for the 8 other cells.

Overall, coverage was relatively stable across the different cell types. However, there were exceptions. For example, H3K4me1-associated states were the most tissue specific, whereas the active promoter and transcribed states were highly constitutive. We observe that many positions of frequently variable chromatin states were the main reasons for our prediction errors. Figure 2C shows the units (from chr1: 71,542,012 to chr1: 71,545,812) of the genomic position of the most variable chromatin states.

Figure 2D shows example units (from chr1: 8,759,613 to chr1: 8,765,013), in which the weak states were generally adjacent to the strong states. Thus, the potential boundary problem of 200-bp segmentations arises, and accordingly the prediction for weak states is more prone to prediction error.

Reducing the 15-state Markov chain models into 12-state models

Several studies show that the DNA sequence is highly predictive of nucleosome positioning and chromosome functions (Ioshikhes et al., 1996, Lee et al., 2004, Segal et al., 2006, Whitehouse and Tsukiyama, 2006; Grewal and Jia, 2007; Lee et al., 2007; Peckham et al., 2007; Field et al., 2008; Schones et al., 2008; Tolstorukov et al., 2008; Yuan and Liu, 2008; Kaplan et al., 2009). Thus, the SeqChromMM will become an important resource because it has the potential to construct the statistical models necessary to develop algorithms for predicting chromatin states or genes in relation to the vast number of biological assays of large-scale epigenetic projects.

To demonstrate the usefulness of our prediction results, we scrutinized the annotated regions of heterochromatin states of ENCODE and further investigated how these states were later annotated in the corresponding regions of Roadmap Genomics annotations.

To achieve this goal, we initially reduced our 15 state Markov chain model into 12 state Markov chain models, excluding inactive states, such as states 13, 14, or 15. These states were excluded from the training set because the population homogeneity among these 3 chromatin Markov chains observed in our previous study (Yoon, 2009) could not be assumed and low entropy according to the Kullback-Leibler distance test (Kullback, 1987). Next, based on the study of overall variability of each 200-bp unit described in the previous section, we iteratively excluded the most variable 200-bp units from the training set and rebuilt the Markov chains.

Figure 3 illustrates the distance matrix corresponding to the newly built transition tables of the 12 states of SeqChromMM. According to the Chi-square distance plot, state 3 (poised promoter) was somewhat distant from states 1 and 2 (active promoter and weak promoter, respectively), although states 1, 2, and 3 belonged to the same broad group of promoters according to the ChromHMM document (Ernst et al., 2011). The behavior of state 6 (weak enhancer) also differed from the other enhancer states of 4, 5, and 7, while state 8 (insulator) showed a relatively different Markov property.

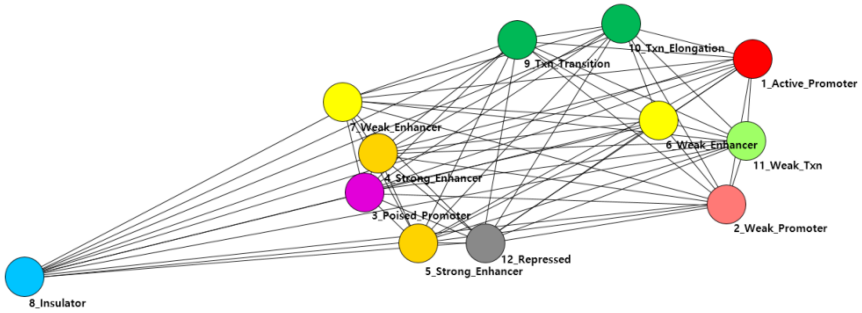


Figure 3. Distance matrix of the newly built transition tables of 12 Markov chains (excluding states 13, 14, and 15 of ENCODE ChromHMM).

RESULTS

Mapping ENCODE annotations to Roadmap Genomics annotations

Not only biological assays but also DNA sequences can play important roles in determining the chromatin states of the human genome. To demonstrate the usefulness of our study, we investigated the annotations of the heterochromatin regions of the Roadmap Genomics project, which were published in 2015 (Roadmap Epigenomics Consortium, 2015), whereas our Markov models were trained from the BED files of the ENCODE consortium, which were published in 2011 (Ernst et al., 2011).

In Table 1A, the 15-state model of ENCODE: ENCODE consortium distinguished 15 different chromatin states. The 11 active states consist of active, weak, and poised promoters (states 1-3), strong and weak candidate enhancers (states 4-7), and strongly and weakly transcribed regions (states 9-11). The 4 inactive states consisted of polycomb repressed regions (state 12), heterochromatic (state 13), and repetitive states (states 14-15) (Ernst et al., 2011).

In Table 1B, the core 15-state model of Roadmap Genomics: The Roadmap Genomics consortium distinguished 15 chromatin states. The 8 active states consist of active transcription start site (TSS) proximal promoter states (TssA, TssAFlnk), a transcribed state showing both promoter and enhancer signatures (TxFlnk), actively transcribed states (Tx, TxWk), enhancer states (Enh, EnhG), and a state associated with zinc finger protein genes (ZNF/Rpts). The 7 inactive states consisted of constitutive heterochromatin (Het), bivalent regulatory states (TssBiv, BivFlnk, EnhBiv), repressed Polycomb states (ReprPC, ReprPCWk), and a quiescent state (Quies) (Roadmap Epigenomics Consortium, 2015).

Table 1 summarizes the differences in annotations between ENCODE and the Roadmap Genomics consortium. Although both used the same number of chromatin states, the individual characteristics of each chromatin state differed. However, these two projects shared similar color codes according to functional roles of the chromatin states. For example, transcribed segments of both of the ENCODE annotations (state 9, 10, and 11) and Roadmap Genomics annotations (state 3, 4, and 5) were all green; inactive segments of both of the ENCODE annotations (state 12, 13, 14, and 15) and Roadmap Genomics annotations (state 13, 14, and 15) were pale colors such as gray, silver, or white. The original 15 states and their associated segment colors can be found on the ENCODE project page (Encode, 2011) and Roadmap Genomics project page (Wang Lab at Washington University in St. Louis, 2015).

Table 1. Fifteen chromatin states of the ENCODE project consortium and Roadmap Genomics consortium.

Chromatin states	Abbreviation
A.	
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_Heterochromatin	
14_Repetitive/CNV	
15_Repetitive/CNV	
B.	
1_Active TSS	TssA
2_Flanking Active TSS	TssAFlnk
3_Transcr. at gene 5' and 3'	TxFlnk
4_Strong transcription	Tx
5_Weak transcription	TxWk
6_Genic enhancers	EnhG
7_Enhancers	Enh
8_ZNF genes & repeats	ZNF/Rpts
9_Heterochromatin	Het
10_Bivalent/Poised TSS	TssBiv
11_Flanking Bivalent TSS/Enh	BivFlnk
12_Bivalent Enhancer	EnhBiv
13_Repressed PolyComb	ReprPC
14_Weak Repressed PolyComb	ReprPCWk
15_Quiescent/Low	Quies

The original 15 states and their associated segment colors can be found on the ENCODE project page (Encode, 2011) and Roadmap Genomics project page (Wang Lab at Washington University in St. Louis, 2015).

Additionally, there are fundamental differences in determining ChromHMM parameters between these two annotations. While the ENCODE consortium used 9 epigenomes and 9 histone markers (H3K4me3, H3K4me2, H3K4me1, H3K9ac, H3K27ac, H3K36me3, H4K20me1, H3K27me3, and CTCF) for setting ChromHMM parameters, the Roadmap Epigenomics consortium used a core set of five chromatin markers (H3K4me3, H3K4me1, H3K36me3, H3K27me3, and H3K9me3) and 60 epigenomes (Ernst et al., 2011, Roadmap Epigenomics Consortium, 2015). Four inactive or repressed states were clearly observed in ENCODE and 7 were observed in Roadmap Genomics.

Next, we examined how the same regions were annotated differently by parsing the ENCODE and Roadmap Genomics BED files.

Table 2 shows how the blocks of the heterochromatin states (state 13) of the ENCODE project were mapped to the corresponding blocks of the Roadmap Genomics consortium across the 9 common cell lines. Using the E123 cell line (K562) as an example, the heterochromatin regions of ENCODE mostly mapped to the similar inactive states such as the Quies (63.92%) or ReprPCWk (10.95%) states in Roadmap Genomics. However, a noticeable portion of the originally heterochromatin regions in ENCODE annotations were annotated differently in Roadmap Genomics; particularly, 12.70% in the E123 cell line was annotated as TxWk (state 5: weak transition) in Roadmap Genomics (Table 2).

Table 2. Nine cell line distribution of the chromatin states of Roadmap Epigenomics corresponding to the heterochromatin states (13 state) of ENCODE: among the 127 cell types and tissues of the Roadmap Genomics consortium, the E3, E118, E119, E121, E122, E123, E127, and E128 datasets correspond to H1ES, K562, GM12878, HepG2, HUVEC, HSMM, NHLF, NHEK, and HMEC of the ENCODE project consortium, respectively.

Annotation		Cell line														
		1_TssA	2_TssAFlnk	3_TxFlnk	4_Tx	5_TxWk	6_EnhG	7_Enh	8_ZNF/Rpts	9_Het	10_TssBiv	11_BivFlnk	12_EnhBiv	13_ReprPC	14_ReprPCWk	15_Quies
H1HESC	E3	0.60%	0.10%	0.00%	3.40%	11.60%	0.10%	2.40%	0.20%	3.10%	0.20%	0.10%	0.20%	0.70%	3.00%	74.40%
GM12878	E116	0.80%	0.90%	0.20%	3.50%	9.00%	0.40%	2.20%	0.40%	2.30%	0.00%	0.00%	0.00%	0.40%	11.90%	68.00%
HEPG2	E118	0.30%	0.70%	0.10%	4.10%	11.30%	0.70%	3.80%	0.10%	8.40%	0.30%	0.10%	0.30%	3.30%	15.20%	51.50%
HMEC	E119	0.50%	0.40%	0.00%	2.80%	12.60%	0.30%	3.70%	0.10%	3.40%	0.00%	0.00%	0.00%	0.30%	9.10%	66.60%
HSMM	E121	0.70%	0.30%	0.00%	2.60%	1.30%	0.20%	2.80%	0.00%	0.80%	0.00%	0.00%	0.00%	2.20%	20.00%	55.90%
HUVEC	E122	0.37%	0.58%	0.04%	2.65%	10.51%	0.34%	2.68%	0.03%	9.33%	0.01%	0.03%	0.04%	2.06%	13.15%	58.16%
K562	E123	0.49%	0.59%	0.12%	3.10%	12.70%	0.51%	3.17%	0.06%	2.94%	0.05%	0.04%	0.04%	1.32%	10.95%	63.92%
NHEK	E127	0.73%	0.23%	0.06%	2.50%	12.39%	0.35%	2.42%	0.04%	1.00%	0.04%	0.01%	0.02%	1.31%	11.43%	67.47%
NHLF	E128	0.37%	0.58%	0.04%	2.65%	10.51%	0.34%	2.68%	0.03%	9.33%	0.01%	0.03%	0.04%	2.06%	13.15%	58.16%

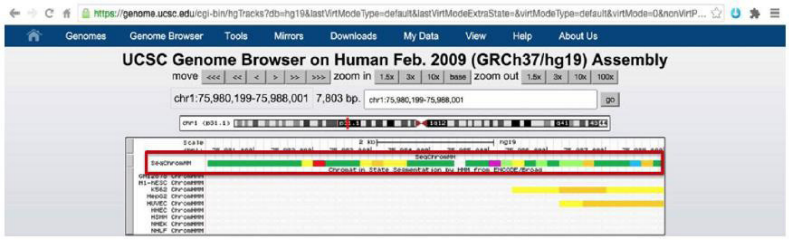
The seventh row (horizontal red box) represents the E123 cell line. The fifth column (vertical red box) represents the 5_TxWk state. The heterochromatin regions of the E123 cell line of ENCODE are mostly mapped to the similar inactive states in Roadmap Genomics.

Case study: analyzing heterochromatin regions of ENCODE annotations

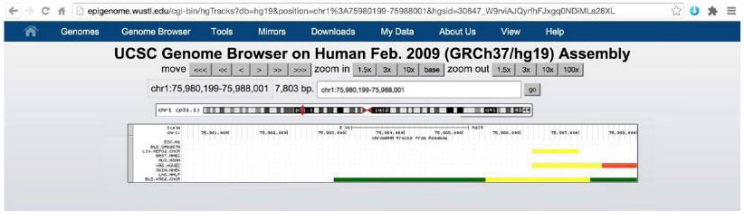
We then examined the genomic regions annotated as inactive states (e.g., heterochromatin states) in ENCODE (Grewal and Jia, 2007), but were annotated as other functional states (e.g., as TxWk) in the Roadmap Genomics project. We also compared the chromatin states of these regions with the SeqChromMM states.

Figure 4 shows an example block (from chr1:75,980,199 to chr1:75,988,001) where the annotations of ENCODE, Roadmap Genomics, and SeqChromMM were substantially different.

Figure 4A shows the annotation of ENCODE from chr1:75,980,199 to chr1:75,988,001 compared to our predicted result of SeqChromMM. There are 10 tracks, where the first track indicates the predicted annotations of SeqChromMM and the other tracks represent the ENCODE annotations for 9 different cell types. While the predicted states of SeqChromMM indicate that the regions were mostly transcribed regions (green color), the ENCODE annotations of 9 cell lines indicate that the states were heterochromatin or inactive states (grey or white color). Thus, the results of the SeqChromMM and ENCODE annotations did not agree.



A. Annotations of ENCODE project together with a SeqChromMM annotation (from chr1:75,980,199 to chr1:75,988,001)



B. Annotations of Roadmap Genomics (from chr1:75,980,199 to chr1:75,988,001)

Figure 4. Snapshot of UCSC browser (Rosenbloom et al, 2015) for an example block from chr1:75,980,199 to chr1:75,988,001, where the annotations of ENCODE, Roadmap Genomics, and SeqChromMM were substantially different.

Figure 4B shows the annotations of Roadmap Genomics of the corresponding region. However, a substantial portion of the same regions of these cell lines was annotated as weak transcription (green color) as shown in Figure 4B. The functional roles approximately agree with the predicted functional roles of SeqChromMM.

Based on this finding, we investigated the entire regions of the ENCODE heterochromatin states and compared the statistical distribution of chromatin annotations between the Roadmap and ENCODE projects across all 9 common cell lines. We also compared the predicted annotations of SeqChromMM with the annotations of Roadmap Genomics across the full range of the cell lines.

Figure 5 shows the distribution of annotations of the 9 cell lines in regard to the regions annotated as TxWk in Roadmap Genomics (corresponding to the vertical red box in Table 2). The figure shows that the regions predicted as Txn_Elongation (state 10 of ENCODE) by SeqChromMM and TxWk (state 5 of Roadmap Genomics) coincided in more than 50% of cases across all 9 cell lines.

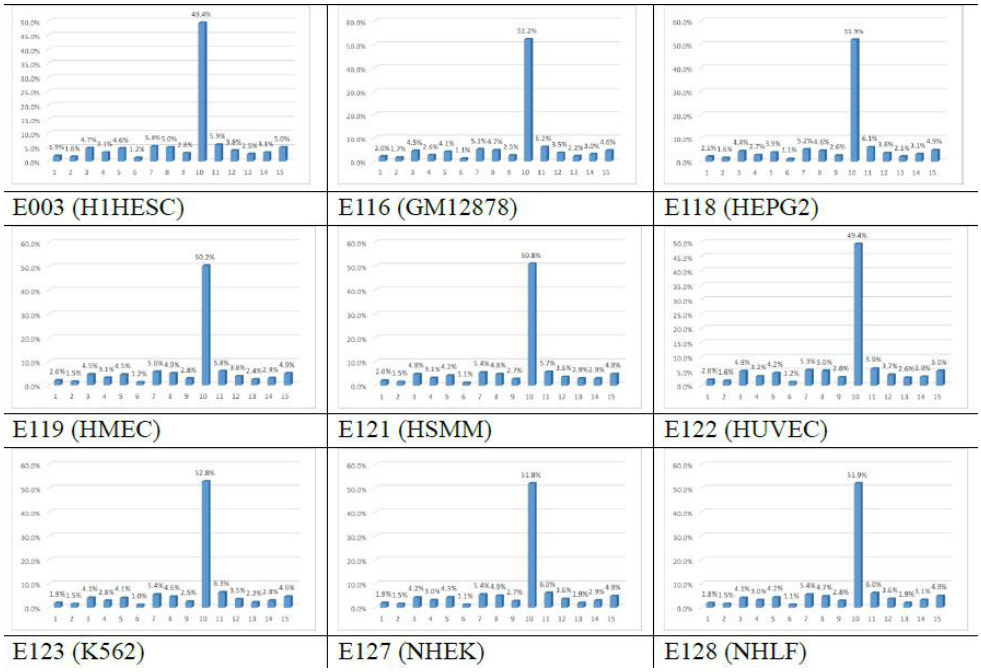


Figure 5. Distribution of the SeqChromMM predictions of the TxWk regions of the Roadmap Genomics across the full range of 9 cell lines.

The result shown in Figure 5 shows that our proposed system can be used as a precursor to predict possible future chromatin states, even when they are inactive.

DISCUSSION

The degree to which genetics plays a role and what proportion of this is epigenetically determined are widely debated topics. Nevertheless, n-gram language analysis of nucleotide sequences can be used to predict epigenetic information.

In this study, we showed that Markov chains of nucleotide sequences can be used as a possible precursor for predicting the functional roles of inactive chromatin regions of the human genome, by providing a representative case in which our prediction results of unknown functional areas of human genome were compared with publicly available large-scale annotations.

Although our study is preliminary and we showed only one example, our results are significant because they prelude the potential use of Markov models of nucleotide sequences necessary for developing algorithms for the prediction of chromatin states in relation to the

vast number of biological assays of large-scale epigenetic projects. Further studies are needed to improve this technique.

ACKNOWLEDGMENTS

Research partially supported by the University-Industry Collaboration Program of Ewha Womans University and Digitaloptics corp. (#2-2016-0872-001-1) and the Interdisciplinary Graduate Research Program in Systems Genomics (#2012M3A9D1054744) of Korea.

REFERENCES

Borodovskii MY, Sprizhitskii YA, Golovanov EI and Aleksandrov AA (1986). Statistical patterns in primary structures of functional regions in the in E. coli genome: 1. Frequency Characteristics. *Mol. Biol.* 20: 826-833.

Encode (2011). Encode Chromatin State Segmentation by HMM from Broad Institute, MIT and MGH. Available at [http://moma.ki.au.dk/genome-mirror/cgi-bin/hgTrackUi?db=hg18&g=wgEncodeBroadHmm].

Ernst J and Kellis M (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9: 215-216. http://dx.doi.org/10.1038/nmeth.1906

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43-49. http://dx.doi.org/10.1038/nature09906

Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, et al. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLOS Comput. Biol.* 4: e1000216. http://dx.doi.org/10.1371/journal.pcbi.1000216

Genome Bioinformatics UCSC (2013). Sequence and Annotation Downloads. Available at [http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm].

Grewal SI and Jia S (2007). Heterochromatin revisited. *Nat. Rev. Genet.* 8: 35-46. http://dx.doi.org/10.1038/nrg2008

Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M, et al. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* 262: 129-139. http://dx.doi.org/10.1006/jmbi.1996.0503

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, et al. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458: 362-366. http://dx.doi.org/10.1038/nature07667

Kullback S (1987). Letter to the Editor: The Kullback-Leibler distance, The American Statistician. 41,4: 340-341.

Lee CK, Shibata Y, Rao B, Strahl BD, et al. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* 36: 900-905. http://dx.doi.org/10.1038/ng1400

Lee KE and Park HS (2014). A review of three different studies on hidden markov models for epigenetic problems: a computational perspective. *Genomics Inform.* 12: 145-150. http://dx.doi.org/10.5808/GI.2014.12.4.145

Lee KE and Park HS (2015). Preliminary testing for the Markov property of the fifteen chromatin states of the Broad Histone Track. *Biomed. Mater. Eng.* 26 (Suppl 1): S1917-S1927. http://dx.doi.org/10.3233/BME-151494

Lee KE and Park HS (2016). Building the SeqChromMM Markov property atlas of the human genome by analyzing the 200bp units of the 15 different chromatin regions. The 5th International Conference on Biomedical Engineering and Biotechnology.

Lee W, Tillo D, Bray N, Morse RH, et al. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39: 1235-1244. http://dx.doi.org/10.1038/ng2117

Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, et al. (2007). Nucleosome positioning signals in genomic DNA. *Genome Res.* 17: 1170-1177. http://dx.doi.org/10.1101/gr.6101007

Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. 518: 317-330.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* 43: D670-D681. http://dx.doi.org/10.1093/nar/gku1177

Schones DE, Cui K, Cuddapah S, Roh TY, et al. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132: 887-898. http://dx.doi.org/10.1016/j.cell.2008.02.022

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, et al. (2006). A genomic code for nucleosome positioning. *Nature* 442: 772-778. http://dx.doi.org/10.1038/nature04979

Smith TF, Waterman MS and Sadler JR (1983). Statistical characterization of nucleic acid sequence functional domains. *Nucleic Acids Res.* 11: 2205-2220. http://dx.doi.org/10.1093/nar/11.7.2205

Tolstorukov MY, Choudhary V, Olson WK, Zhurkin VB, et al. (2008). nuScore: a web-interface for nucleosome positioning predictions. *Bioinformatics* 24: 1456-1458. http://dx.doi.org/10.1093/bioinformatics/btn212

Wang Lab at Washington University in St. Louis (2015). ROADMAP epigenomics project Chromatin state learning. Available at [http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html].

Whitehouse I and Tsukiyama T (2006). Antagonistic forces that position nucleosomes *in vivo*. *Nat. Struct. Mol. Biol.* 13: 633-640. http://dx.doi.org/10.1038/nsmb1111

Yoon BJ (2009). Hidden Markov models and their applications in biological sequence analysis. *Curr. Genomics* 10: 402-415. http://dx.doi.org/10.2174/138920209789177575

Yuan GC and Liu JS (2008). Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* 4: e13. http://dx.doi.org/10.1371/journal.pcbi.0040013