Building the Frequency Profile of the Core Promoter Element Patterns in the Three ChromHMM Promoter States at 200bp Intervals: A Statistical Perspective

Heather Lent1,2, Kyung-Eun Lee1, Hyun-Seok Park1*

1Bioinformatics Laboratory, School of Engineering, Ewha Womans University, Seoul 03760, Korea,
2Human Language Technology Program, Department of Linguistics, University of Arizona, Tucson, AZ 85721, USA

Recently, the Encyclopedia of DNA Elements (ENCODE) Analysis Working Group converted data from ChIP-seq analyses from the Broad Histone track into 15 corresponding chromatic maps that label sequences with different kinds of histone modifications in promoter regions. Here, we publish a frequency profile of the three ChromHMM promoter states, at 200-bp intervals, with particular reference to the existence of sequence patterns of promoter elements, GC-richness, and transcription starting sites. Through detailed and diligent analysis of promoter regions, researchers will be able to uncover new and significant information about transcription initiation and gene function.

Keywords: chromatin state, epigenetics, epigenomics, genetic promoter regions
Availability: https://github.com/KyungEunLee/motif

Introduction

Both eukaryotic and prokaryotic cells contain important gene-regulating elements, located within promoter regions [1,2]. It is known that promoter sequences possess special signatures to distinguish them from the rest of the genome sequences. A few core promoter elements have been detected, of which the most common elements are the CpG island, TATA box, initiator (Inr), downstream promoter element (DPE), and TFIIB recognition element (BRE) [2], but these promoter elements may not be universal.

Recently, a group from the Encyclopedia of DNA Elements (ENCODE) project published chromatic maps to mark histone modifications within nine cells of different types from the human genome, drawn from the results from chromatin immunoprecipitation sequencing (ChIP-seq) analyses of these same cells [1,2]. From each of these nine cells, Ernst et al. [3,4] published 15 chromatin states of the human genome. Among them, the first three states represent promoter regions: state 1 (Active_Promoter), state 2 (Weak_Promoter), and state 3 (Poised_Promoter). The information gained from the exploration of promoter elements and transcription starting sites will enable researchers to produce more accurate promoter prediction algorithms, which in turn will yield deeper insight into the process of transcription initiation and gene functions.

Here, we publish frequency profiles of the ChromHMM promoter states of the human genome, at 200bp intervals, paying particular attention to the existence of promoter elements, such as CpG islands, TATA boxes, Inr, DPE, and recognition elements from the transcription factors that bind with polymerase II (BRE) (signal features), together with transcription starting sites, in terms of size, overlapping patterns, and locations.

Results

Ernst et al. [4] applied unsupervised learning methods to convert the Broad Histone track’s results from ChIP-seq analyses into specific points that indicate promoter regions on chromatin maps for 15 chromatin elements within the genome. These 15 chromatin states were chosen because of their rich biological content, as well their well-established
Table 1. Relative coverage of the ChromHMM promoter states in the GM12878 and K562 cell lines

| State             | gm12878 | k562       |          |          |
|-------------------|---------|------------|----------|----------|
|                   | No. of blocks | Size (bp)  | Percentage | No. of blocks | Size (bp)  | Percentage |
| 1_Active_Promoter | 14,674  | 21,367,850 | 0.80     | 14,951  | 18,277,036 | 0.68       |
| 2_Weak_Promoter   | 34,013  | 19,381,338 | 0.72     | 28,916  | 13,448,418 | 0.50       |
| 3_Poised_Promoter | 5,193   | 4,643,842  | 0.17     | 6,240   | 3,592,784  | 0.13       |

Fig. 1. Java code to recognize regular expressions of different promoter elements: TATA-box, Initiator, TFIIB recognition element, and downstream promoter element.

Fig. 2. An output profile format consisting of the 18 required fields. INI, initiator; BRE, TFIIB recognition element; DPE, downstream promoter element; TSS, transcription start site.
Table 2. Summary of the promoter frequency profile

| Motif     | TSS presence | TSS absence | Motif frequency (TSS absence) | Motif frequency (TSS presence) |
|-----------|--------------|-------------|-------------------------------|-------------------------------|
|           | TATA | Inr | BRE | DPE | Blocks, n (%) | Blocks, n (%) | TATA | Inr | Bre | Dpe | GC ratio (%) | TATA | Inr | Bre | Dpe | GC ratio (%) |
| o o o o o | 6,054 (3.6) | 402 (4.3) | 1.59 | 3.30 | 1.13 | 3.39 | 55.7 | 1.60 | 3.33 | 1.14 | 3.31 | 55.1 |
| o o o o x | 190 (0.1) | 8 (0.1) | 1.82 | 3.17 | 0.00 | 55.5 | 1.50 | 3.25 | 1.25 | 0.00 | 59.3 |
| o o x o o | 76,660 (45.9) | 4,000 (42.4) | 2.49 | 3.90 | 0.00 | 35.7 | 42.4 | 2.31 | 3.82 | 0.00 | 3.56 | 43.6 |
| o o x x x | 2383 (1.4) | 118 (1.3) | 3.12 | 4.33 | 0.00 | 38.9 | 3.08 | 4.21 | 0.00 | 0.00 | 40.3 |
| o x o o o | 331 (0.2) | 30 (0.3) | 1.24 | 0.00 | 1.27 | 0.00 | 38.9 | 3.08 | 4.21 | 0.00 | 0.00 | 66.7 |
| o x o x x | 22 (0.01) | 1 (0.01) | 1.41 | 0.00 | 1.27 | 0.00 | 48.7 | 6.50 | 0.00 | 0.00 | 0.00 | 79.4 |
| o x x o o | 1,806 (1.1) | 87 (0.9) | 1.87 | 0.00 | 3.92 | 0.00 | 50.0 | 1.60 | 3.25 | 0.00 | 0.00 | 59.3 |
| o x x x x | 49 (0.03) | 2 (0.02) | 2.82 | 0.00 | 48.7 | 0.00 | 65.0 | 0.00 | 0.00 | 0.00 | 0.00 | 28.9 |
| x o o o o | 16,680 (10.0) | 1,043 (11.1) | 0.00 | 2.16 | 3.67 | 0.00 | 68.7 | 0.00 | 2.10 | 3.56 | 0.00 | 69.2 |
| x o o x x | 452 (0.3) | 41 (0.4) | 0.00 | 1.93 | 1.53 | 0.00 | 71.2 | 0.00 | 1.98 | 1.56 | 0.00 | 71.7 |
| x o x o o | 50,393 (30.1) | 2,864 (30.4) | 0.00 | 2.84 | 4.05 | 0.00 | 58.4 | 0.00 | 2.76 | 3.95 | 0.00 | 59.7 |
| x o x x x | 1,100 (0.7) | 55 (0.6) | 0.00 | 2.84 | 0.00 | 58.1 | 0.00 | 2.62 | 0.00 | 0.00 | 61.7 |
| x x o o o | 4,642 (2.8) | 387 (4.1) | 0.00 | 0.00 | 32 | 0.00 | 74.3 | 0.00 | 0.00 | 3.28 | 75.0 |
| x x o x x | 250 (0.2) | 20 (0.2) | 0.00 | 0.00 | 1.81 | 0.00 | 77.1 | 0.00 | 0.00 | 1.90 | 76.7 |
| x x x o o | 5,703 (3.4) | 359 (3.8) | 0.00 | 0.00 | 3.88 | 0.00 | 66.9 | 0.00 | 0.00 | 0.00 | 68.1 |
| x x x x x | 482 (0.3) | 19 (0.2) | 0.00 | 0.00 | 32.3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 68.1 |

TSS, transcription start site; Inr, initiator; BRE, TFIIB recognition element; DPE, downstream promoter element.

fact that we divided the regions into 200 bp lengths whereas the core promoter elements are known to exist typically within 50 bp (base pairs) upstream to 50 bp downstream of the TSS [2], and that we are including the Poised_Promoter state of the ChromHMM.

From the ChromHMM human promoter blocks, the results of our statistical survey on 176,579 units show that DPE sequences have the highest frequency among the four chosen motifs in promoter elements, appearing in about 97.1% of the ChromHMM promoter regions, meanwhile Inr is seen in 92% of promoter regions, and BRE is seen in only 17.3% of the promoters (It is known that Inr usually occurs only about half of the human promoters [3]). TATA boxes are found in 52.2% of the ChromHMM promoter blocks. Some contextual patterns have been recognized for certain promoter elements. For example, we can expect to find BREs within GC-rich areas, meanwhile areas with TATA boxes are typically GC-poor.

In total, there are 9,435 ChromHMM promoter blocks in the dbTSS locations (approximately 5.34%). The number of 200-bp blocks that contain all four promoter elements was 176,579. Among them, only 9,436 of the blocks contain dbTSS data [7,8]. Some of the blocks showed distinct associations with dbTSS data. For example, the promoter blocks that do not contain any of the promoter elements tend to have a higher GC ratio (0.68) where a dbTSS location exists.

Conclusion

One of the strengths of our work versus previous studies is that we profiled the core elements separately, based on the three classes of promoter states (Active_Promoter, Weak_Promoter, and Poised_Promoter of the ChromHMM). A more detailed analysis of frequency profiles in regard to these three states is beyond the scope of this paper. However, the complete frequency profiles and Java code are published through GitHub (https://github.com/KyungEunLee/motif). We believe that the frequency profiles presented of the three promoter states can be helpful for bioinformaticians to gain a deeper understanding of the characteristics of the various promoter regions and to improve upon old methods of core promoter analysis algorithms for their specific research needs.

Acknowledgments

This work was supported by the Small & Medium Business Corporation (project number : 2-2014-2443-001-1) of Korea.

References

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306:636-640.
2. ENCODE Project Consortium. An integrated encyclopedia of
DNA elements in the human genome. *Nature* 2012;489:57-74.

3. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010;28:817-825.

4. 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.

5. 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.

6. Datta S, Mukhopadhyay S. A composite method based on formal grammar and DNA structural features in detecting human polymerase II promoter region. *PLoS One* 2013;8:e54843.

7. Yamashita R, Sugano S, Suzuki Y, Nakai K. DBTSS: DataBase of transcriptional start sites progress report in 2012. *Nucleic Acids Res* 2012;40:D150-D154.

8. Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, *et al.* Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res* 2011;21:775-789.