BMC Genomics

PROCEEDINGS Open Access

A linear time algorithm for detecting long genomic regions enriched with a specific combination of epigenetic states

Kazuki Ichikawa*, Shinichi Morishita

From The Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015)
HsinChu, Taiwan. 21-23 January 2015

Abstract

Background: Epigenetic modifications are essential for controlling gene expression. Recent studies have shown that not only single epigenetic modifications but also combinations of multiple epigenetic modifications play vital roles in gene regulation. A striking example is the long hypomethylated regions enriched with modified H3K27me3 (called, "K27HMD" regions), which are exposed to suppress the expression of key developmental genes relevant to cellular development and differentiation during embryonic stages in vertebrates. It is thus a biologically important issue to develop an effective optimization algorithm for detecting long DNA regions (*e.g.*, >4 kbp in size) that harbor a specific combination of epigenetic modifications (*e.g.*, K27HMD regions). However, to date, optimization algorithms for these purposes have received little attention, and available methods are still heuristic and *ad hoc*.

Results: In this paper, we propose a linear time algorithm for calculating a set of non-overlapping regions that maximizes the sum of similarities between the vector of focal epigenetic states and the vectors of raw epigenetic states at DNA positions in the set of regions. The average elapsed time to process the epigenetic data of any of human chromosomes was less than 2 seconds on an Intel Xeon CPU. To demonstrate the effectiveness of the algorithm, we estimated large K27HMD regions in the medaka and human genomes using our method, ChromHMM, and a heuristic method.

Conclusions: We confirmed that the advantages of our method over those of the two other methods. Our method is flexible enough to handle other types of epigenetic combinations. The program that implements the method is called "CSMinfinder" and is made available at: http://mlab.cb.k.u-tokyo.ac.jp/~ichikawa/Segmentation/

Background

Epigenetic modifications have been shown to play a vital role in regulating gene expression. Recent genome-wide studies have revealed that in vertebrates, although most CpG sites in DNA sequences are highly methylated, hypomethylated CpG islands proximal to genes are involved in regulating gene expression [1]. Specifically, hypermethylated CpG islands in promoter regions are relevant to gene silencing, while hypomethylated CpG islands are in an active or permissive state for transcription [2]. In addition to cytosine methylation of CpG sites, some histone modifications around promoter regions also are known to affect the regulation of gene expression [3,4].

It was found recently that long hypomethylated regions enriched with H3K27me3 were likely to overlap with regions encoding key genes essential for cell development and differentiation in human embryonic stem cells [5], mouse hematopoietic stem cells [6], early *Xenopus tropicalis* embryos demonstrates [7], and medaka fish blastula (half-day) embryos [8]. Although many hypomethylated domains (HMD) are subjected to modification of the active histone mark H3K4me2 that promotes gene expression [9-12], it is remarkable that ~300 HMDs of length >4 kb rarely have H3K4me2 histone marks but

* Correspondence: ichikawa@cb.k.u-tokyo.ac.jp
Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-0882, Japan

have repressive H3K27me3 histone marks, and are found in association mostly with developmental genes [8]. Promoters in HMD with H3K27me3 marks (called, "K27HMD") are in a 'poised' state, in which the genes are not simply silenced but are ready for activation immediately during cell differentiation, which is important for sustaining the pluripotency of pluripotent cells [13,14]. Figure 1 shows four examples of long K27HMD regions that include developmental genes such as *cbx4*, *cbx8*, *hoxa* genes, *six2*, *hnf6*, and *zic1/4*.

Thus, there has been considerable interest in long K27HMD regions with biologically important characteristics. However, computational methods for detecting long K27HMD regions are still heuristic and *ad hoc*, emphasizing the need to develop an effective algorithm from a profound background in computation theory. For example, to identify K27HMD, Nakamura et al. proposed a heuristic method that used certain *ad hoc* parameter settings to define hypomethylated regions and H3K27me3 peak detection [8]. The method is not guaranteed to output K27HMD regions longer than a given threshold, and it often generates regions of differing lengths. ChromHMM [15] is a statistical method that classifies epigenetic modifications into classes of combinations and divides a DNA sequence into sub-regions such that each sub-region has a uniform combination of epigenetic states while neighboring sub-regions have distinct characteristics. ChromHMM has been used successfully to partition regions surrounding genes into active/inactive promoters, exons, and introns by analyzing epigenetic codes. Although ChromHMM can be used for K27HMD detection by setting its parameters to find regions that are hypomethylated and marked by H3K27me3, ChromHMM often generates many short regions and thus is not suitable for detecting large K27HMD regions. Overall, these previous methods have

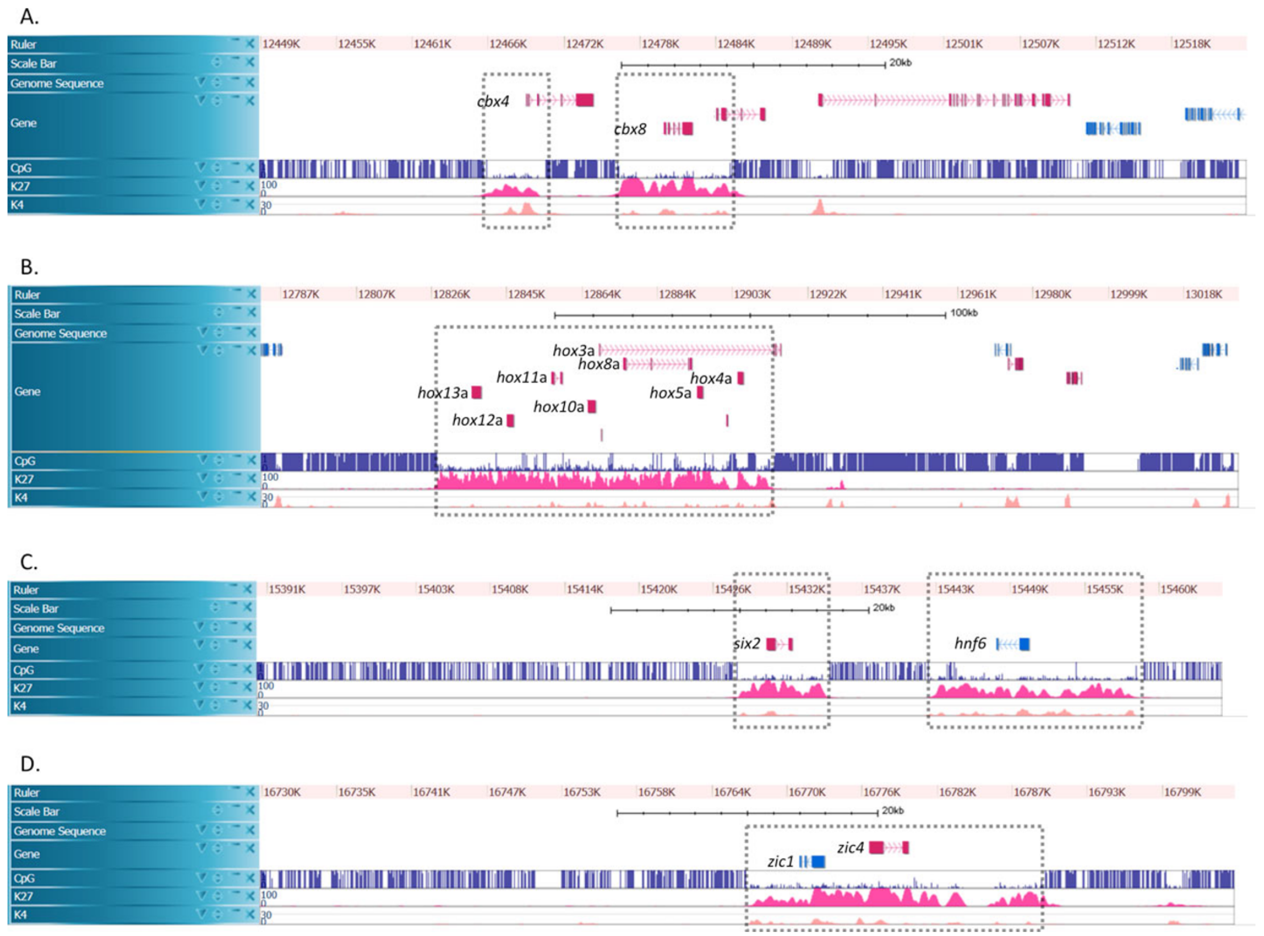

Figure 1 Examples of long K27HMD regions in the medaka genome. Examples of K27HMD regions enclosed in dashed boxes. Each screen capture shows an image in a medaka genome browser that displays tracks of gene structures, CpG methylation levels observed by bisulfite sequencing, and levels of H3K27me3 and H3K4me2 in blastula cells (half-day embryos). **A**. A K27HMD region of length ~4 kbp with *cbx4*, and a ~8 kbp region with *cbx8*. **B**. A large region of length ~90 kbp with *hoxa* genes. **C**. A ~6 kbp region with *six2*, and a ~14 kbp region with *hnf6*. **D**. A ~20 kbp region with *zic1* and *zic4*.

simply not been designed to output regions of lengths greater than or equal to a given minimum threshold.

To address this problem, we propose a linear time algorithm for calculating a set of non-overlapping regions such that the set maximizes the score of focal combinations of epigenetic modifications (*e.g.,* K27HMD) and the length of each region is greater than or equal to a given minimum threshold (*e.g.,* 4 kb). We define the score of a focal combination of epigenetic modifications at each DNA position as the similarity between the vector of focal epigenetic states and the vector of raw epigenetic states at the position. We then define the similarity score of a set of regions as the sum of similarity scores of all positions in the set. This method solves several issues in previous heuristic methods because it allows us to set a minimum region length for detecting 'long' regions of biological importance and guarantees the output of an optimal set of long regions that maximizes an objective function.

We implemented the algorithm. We call the program CSMinfinder (Chromatin State with minimum length finder). With CSMinfinder, we identified large K27HMD regions in the medaka and human genomes [8,16,17] that overlapped many developmental genes. CSMinfinder can be applied to epigenetic data from other vertebrates for understanding cell development and differentiation.

CSMinfinder runs in time proportional to the size of the genome, and it can process vertebrate genomes in feasible amounts of time. Although we applied CSMinfinder specifically to K27HMD, it can be used for the detection of regions with other types of epigenetic combinations by defining the vector of focal epigenetic states appropriately.

Methods

To detect long regions of focal epigenetic states, we formulated this as a problem of finding an optimal set of disjoint (non-overlapping) regions in a sequence that maximizes the sum of similarity scores in all regions. Our method calculates a similarity score between a vector of epigenetic modifications at each position and the feature vector of a focal epigenetic state, such as K27HMD, and outputs the set of regions with the highest sum of similarity scores.

Calculating a similarity vector

We need to generate a modification vector at each position from epigenomic signal data. For example, to create benchmark datasets in this study, we binarized the modification signal level at each position using BinarizeBed in ChromHMM [15], which classified the signal at each position into 0 or 1 according to a Poisson background model. Subsequently, we defined a modification vector as the vector with binary scores of modifications at each position.

Definition 1. Let $w_1,w_2...,w_n$ be non-overlapping windows of the same length (e.g., 200 bp in this study) in a DNA sequence. Let $s_i^1, \ldots, s_i^k$ be binary or real-valued signals of k modifications in window i. The modification vector of w_i is defined as $M_i = (s_i^1, \ldots, s_i^k)$. Let F denote the feature vector of a focal modification pattern with k elements. The similarity score of M_i and F is defined as their inner product minus a given threshold τ.

Example. Suppose that $k = 3$, $\tau = 1.3$, $F = (1,1,0)$, $M_1 = (1,1,0)$, $M_2 = (1,0,1)$ and $M_3 = (0,0,1)$. Similarity scores of F and M_i are 0.7, -0.3, and -1.3 for $i = 1,2,3$

When the inner product of M_i and F is positive for all $i = 1,...,n$, the optimal set of regions that maximizes the sum of similarity scores in the regions becomes the entire region, $[1,n]$, which may not be informative. If we want to select a set of regions whose modification vectors are closer to the feature vector F, we can set the threshold τ to an appropriate positive value to yield a negative similarity score for the inner product that is lower than τ. Positions with negative similarity scores are less likely to be included in the optimal set of regions. A higher threshold is likely to divide the entire genome into smaller regions with a higher precision, while a lower threshold yields an opposite trend. In this manner, for a series of windows $w_1,w_2...,w_n$ in a DNA sequence, we generate a series of similarity scores.

Detecting an optimal set of disjoint regions

To detect regions of focal epigenetic states such as K27HMD, we present an algorithm for calculating an optimal set of disjoint regions in a sequence that maximizes the sum of similarity scores for all regions. In addition, to identify sufficiently long regions, we define a minimum length threshold of regions such that each region is longer than or equal to the minimum length. The problem can be defined as follows.

Definition 2. Let $L = \{L_i | i = 1,2,...,n\}$ be a series of real valued weights L_i (*e.g.,* similarity scores). Let C be a series of non-overlapping regions I_j ($j = 1,...k$) of L such that the length of each I_j is greater than or equal to a given minimum threshold m_1, and the length of the interval between I_{j-1} and I_j is greater than or equal to another given minimum threshold m_0. That is, C is a series of regions of the form $\{[a_1, b_1], \cdots [a_k, b_k]\}$ $(1 \leq a_1 < b_1 < a_2 < b_2 \cdots < a_k < b_k \leq \text{n})$ such that

1. $a_t + m_1 - 1 \leq b_t$ for $t = 1,...,k$ (the minimum length constraint on regions),
2. $b_{t-1}+m_0 < a_t$ for $t = 2,...,k$ (the minimum length constraint on intervals between regions), and
3. $a_1 = 1$ or $a_1 > m_0$ (the first region start at position 1 or at position larger than m_0).

Readers may find the last condition strange because it appears to disallow the situation that the first region starts at position $a_1 > m_0$. We used the condition to simplify the presentation of our linear-time algorithm, which is described later. To obtain such an optimal series of regions that the first region starts at $a_1 > m_0$, for example, you can temporarily add m_0 negative weights in front of L, calculate the optimal series, and restore the coordinate.

To calculate a C that maximizes the sum of weights in C, $\sum_{i \in I \in C} L_i$, we used a dynamic programming algorithm developed by Csuros [18]. Here, we outline the algorithm.

Definition 3. We assume that all series meet the conditions given in Definition 2. Let $w(C)$ denote the sum of weights in C, $\sum_{i \in I \in C} L_i$. We consider two cases: that in which the last region of C ends at I and that in which it does not. When the last region does not end at i, let $C^0_{i,m}$ denote a series of regions that maximizes w (C) among all series, such that the last region ends at position $b_k \le i$ - m, where $m \ge 1$. When the last region ends at i, let $C^1_{i,m}$ denote a series of regions that maximizes $w(C)$ among all series, such that the last region is of length $\ge m$ ($\ge$1); specifically, $a_k + m - 1 \le i$ ($= b_k$).

Example. When i = 12, and L = (1,1,-3,1,1,-3,1,1,1,1,-2,1), we have

$$C^0_{12,1} = \{[1,2],[4.5],[7,10]\}, \quad C^0_{12,4} = \{[1,2],[4,5],[7,8]\},$$
$$C^1_{12,7} = \{[1,2],[4,12]\}, \quad C^1_{12,12} = \{[1,12]\}.$$

According to this definition, C maximizing $w(C)$ is either $C^0_{n,1}$ or C^1_{n,m_1}. For calculating these two series, we define here $w(C^0_{i,m})$ and $w(C^1_{i,m})$ recursively for $i = 1,\ldots,n$ and $m \ge 1$.

Definition 4. We define the following four types of weight sums, $W^0_{short}(i)$, $W^0_{long}(i)$, $W^1_{short}(i)$, and $W^1_{long}(i)$, depending on whether the last region ends at i or not (denoted as 1 or 0, respectively) and whether the minimum length constraint is satisfied or not (denoted as *long* or *short*, respectively):

$$W^0_{short}(1) = 0, \quad W^1_{short}(1) = L_1,$$
$$W^0_{short}(i) = w\left(C^0_{i,1}\right), \quad W^0_{long}(i) = w\left(C^0_{i,m_0}\right),$$
$$W^1_{short}(i) = w\left(C^1_{i,1}\right), \quad W^1_{long}(i) = w\left(C^1_{i,m_1}\right)$$

Csuros showed that these four types of weight sums can be calculated recursively as follows [18]:

$$W^0_{short}(i) = max\{W^0_{short}(i-1), W^1_{long}(i-1)\} \text{ for } i \in [2,n]$$
$$W^1_{short}(i) = L_i + max\{W^0_{long}(i-1), W^1_{short}(i-1)\} \text{ for } i \in [2,n]$$
$$W^0_{long}(i) = W^0_{short}(i - m_0 + 1) \text{ for } i \in [m_0, n]$$
$$W^1_{long}(i) = W^1_{short}(i - m_1 - 1) + \sum_{j=i-m_1+2}^{i} L_j \text{ for } i \in [m_1, n]$$

Recall that C maximizing $w(C)$ is either $C^0_{n,1}$ or C^1_{n,m_1}. From $W^1_{long}(n)$, we can build the series of regions, C^1_{n,m_1}, by tracing back the process of calculating $W^1_{long}(n)$. Similarly, from $W^0_{short}(n)$, we can obtain $C^0_{n,1}$.

We implemented the above idea. We call the program CSMinfinder.

Results

Data sets

To compare CSMinfinder with other available methods for detecting large K27HMD, we used real biological datasets from the medaka-fish and human genomes, each of which was a set of vectors of DNA methylation levels at CpG sites, determined by bisulfite sequencing, and H3K4me3 and H3K27me3 histone modification Chip-seq data [8]. We set the window size to 200 bp, normalized the data using a Poisson distribution model, and set the binarized score of a window to 1 if its probability was < 0.0001 and to 0 otherwise.

Detecting large K27HMD in medaka epigenomic data

We compared CSMinfinder with ChromHMM [15] and Nakamura's method [8].

- Using ChromHMM, we estimated six chromatin states and divided the given DNA sequence into these six states. Specifically, ChromHMM asks users to input the number of epigenetic states beforehand. Thus, we started with inputting a small number into ChromHMM, increased the number gradually one by one until we found a state similar to K27HMD, hypo-methylated DNA modification and H3K27me3 histone modification, and called the number *sufficient*. Inputting a value larger than the sufficient number into ChromHMM did not make much sense because it just output a state similar to K27HMD. The sufficient number was six. Among the six states, one represented hypomethylated DNA modifications and the H3K27me3 histone modification. We therefore treated the state as K27HMD.
- Nakamura's method detects a hypomethylated domain on a DNA sequence that has more than nine contiguous CpG sites with low methylation (methylation level <0.4) and no more than four contiguous highly methylated CpG sites. Parameters are selected heuristically. A hypomethylated domain is treated as a K27HMD if it contains H3K27me3 peaks detected by QuEST [19], such that each peak is more than three times larger than the average.
- In CSMinfinder, we used two types of minimum length thresholds, 4 kbp and 8 kbp, to evaluate the effect of this constraint. We set the minimum length of any interval between regions to 600 bp.

Comparing the performance in detection of large K27HMD around genes in the medaka genome

Large K27HMD regions of length >4 kbp suppress the expression of many developmental genes [8]. Thus, we verified the effectiveness of CSMinfinder for detecting large K27HMD regions surrounding genes in the medaka genome. Nakamura's method could detect 246 large K27HMD regions containing the promoter regions of developmental genes (e.g., *hox* clusters) that were relevant to transcriptional regulation and the developmental process. CSMinfinder detected 911 K27HMD regions, and of these, 386 regions contained promoter regions of >4 kbp in size and contained 242 of the 246 regions identified using Nakamura's method. Indeed, our regions covered 91% of bases in the entire regions detected by Nakamura's method. Specifically, although the exact boundaries of individual regions estimated by the two methods were unlikely to be consistent, these regions largely overlapped each other. These results demonstrate the high concordance between CSMinfinder and Nakamura's methods as well as the ability of CSMinfinder to identify more K27HMD regions than did Nakamura's method.

We assessed the quality of each K27HMD region in terms of their low average DNA methylation level because this property is considered to be essential in maintaining the suppression of developmental gene expression in embryonic cells [8]. Indeed, Figure 2 shows the tendency of the average methylation level in the vertical axis to become lower for a longer K27HDM region, the length of which is displayed in the horizontal axis. This trend was also observed with all three methods.

We then compared the performance of the three methods by examining the length distributions of K27HMD regions in the medaka genome. Figure 3A shows the length distributions of large K27HMD regions (>4 kb in size) estimated by each of the three methods. Setting the minimum length threshold to 4 kbp in CSMinfinder detected more regions of length > 6 kbp but fewer regions of length >7 kbp compared with Nakamura's method. CSMinfinder allows us to output longer regions by setting the minimum length threshold to a higher value. For example, setting the minimum length to 8 kbp, CSMinfinder found more regions than did Nakamura's method (Figure 3C).

Analysis of large K27HMD regions in human epigenomic data

We also compared CSMinfinder with the other two for processing human epigenomic data. For ChromHMM, we calculated the sufficient number for the human data according to the procedure described before, and we classified epigenetic modification data into seven states rather than six so as to identify a state similar to K27HMD. The sufficient numbers of epigenetic states in the human and medaka data differed due to the difference in data quality. The sufficient number in the medaka data was smaller than that in the human data presumably because epigenetic state signals in the medaka data were clearer.

In CSMinfinder, we set the minimum length threshold to 8 kbp and the interval between regions to 600 bp. We also searched an ideal value of threshold τ by repeated trials to detect large continual regions, and we set τ to 1.3 and 1.6 in the respective medaka and human data.

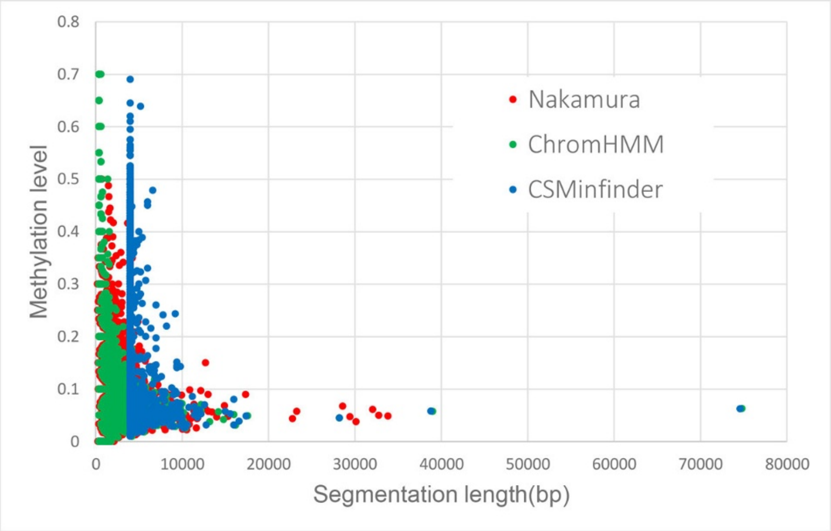

Figure 2 Lengths and average methylation levels of K27HMD regions in the medaka genome. Each dot represents a region that is identified by CSMinfinder, ChromHMM, and Nakamura's method in the medaka genome. The *x*-axis shows the length of a K27HMD region and the *y*-axis presents the average methylation level of the region.

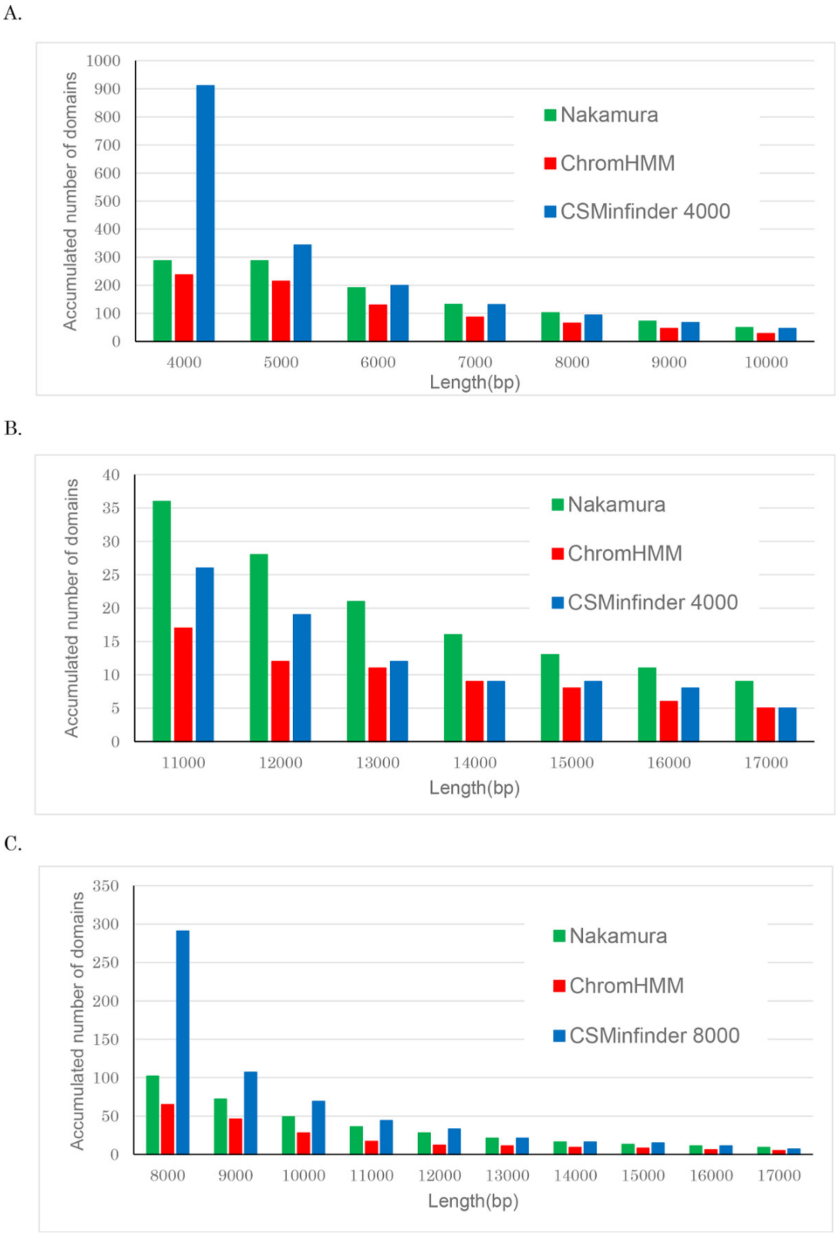

Figure 3 Length distribution of large K27HMD regions in the medaka genome. **A-B**. Comparison between CSMinfinder (minimum length threshold of 4 kbp), ChromHMM, and Nakamura's method. The *x*-axis shows the minimum length of K27HMD regions, and the *y*-axis shows the accumulated number of K27HMD regions longer than or equal to the threshold in the *x*-axis. Because of the space limitations, the histogram is divided into two sub-histograms **A** (threshold is ≤ 10 kbp) and **B** (threshold ≥ 11 kbp). **C**. In this case, we set the minimum threshold to 8 kbp using CSMinfinder.

Because the human genome is longer than the medaka genome, we focused on large K27HMD regions of length > 8 kbp. Nakamura's method detected 314 regions, and CSMinfinder identified 542 regions, including 291 of those found using Nakamura's method. Again, there was high concordance between the results obtained by the two methods. Figure 4 shows examples of large K27HMD regions detected around developmental genes. Although CSMinfinder and Nakamura's method yielded slightly different regions with distinct boundaries in the output, each created regions of similar sizes. In contrast, ChromHMM yielded shorter regions than the other two did. Specifically we compared the length distribution of large K27HMDregions estimated by each of the three methods (Figure 5). We found that CSMinfinder and Nakamura's method were comparable. Precisely, although the number of

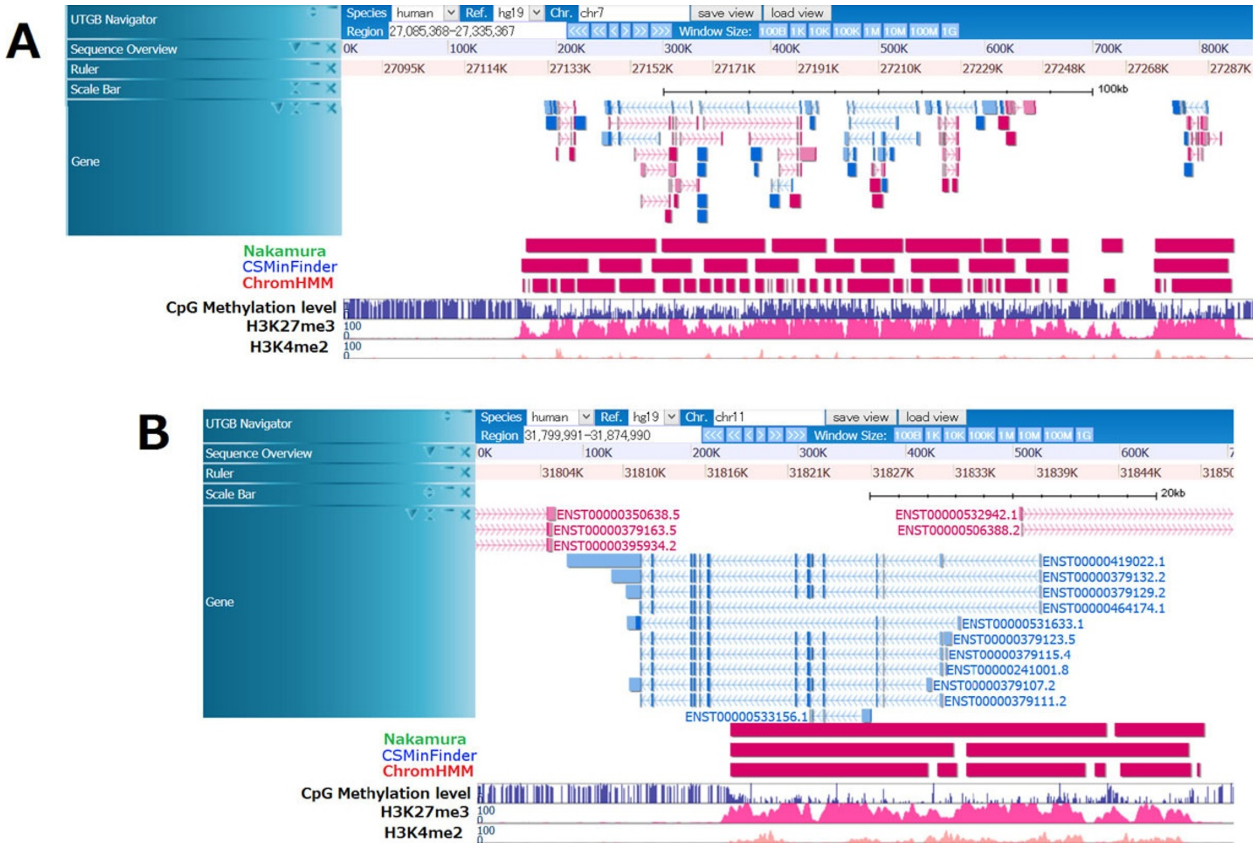

Figure 4 Examples of large K27HMD regions around developmental genes in the human genome. **A**. The figure displays several K27HMD regions in the human chromosome 11 around *pax6*, a gene that regulates eye and brain development. CSMinfinder and Nakamura's method detected large K27HMD regions of >4 kbp in size and output large regions that largely overlapped; however, ChromHMM divided these regions into smaller ones. **B**. These large K27HMD on human chromosome 7 were located around a cluster of *hox* genes that regulate the body plan of the head-tail axis. ChromHMM yielded much smaller K27HMD regions as output than did the other two methods.

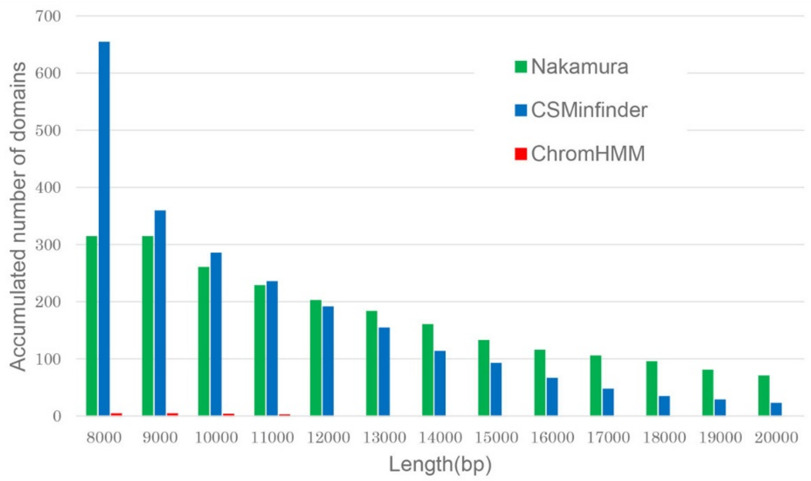

Figure 5 Length distribution of large K27HMD regions in the human genome. Comparison between CSMinfinder (minimum length threshold of 8 kbp), ChromHMM, and Nakamura's method. The *x*-axis shows the minimum K27HMD region length threshold, and the *y*-axis shows the accumulated number of K27HMD regions longer than or equal to the threshold on the x-axis.

extremely large regions longer than 12 kbp is slightly smaller than the number found by Nakamura's method, CSMinfinder could detect similar numbers of large regions between 8 kbp to 12 kbp. Later we will discuss the reason why ChromHMM were inferior to the other two methods.

Computational performance and software availability

We observed the computational performance of CSMinfinder using Intel Xeon CPU E5-2670 processor with a clock rate of 2.60 GHz and 66GB of main memory. The computation time to calculate the optimal series of regions was negligible. Figure 6 shows that the average elapsed time was less than 2 seconds when we processed the epigenetic data of any of human and medaka chromosomes. Furthermore, Figure 6 also illustrates that the elapsed time is almost proportional to the size of each chromosome, thereby confirming experimentally that the worst-case time complexity of the algorithm is linear in the input size. CSMinfinder does not consume a large amount of main memory. CSMinfinder is made available at the following site: URL: http://mlab.cb.k.u-tokyo.ac.jp/~ichikawa/Segmentation/

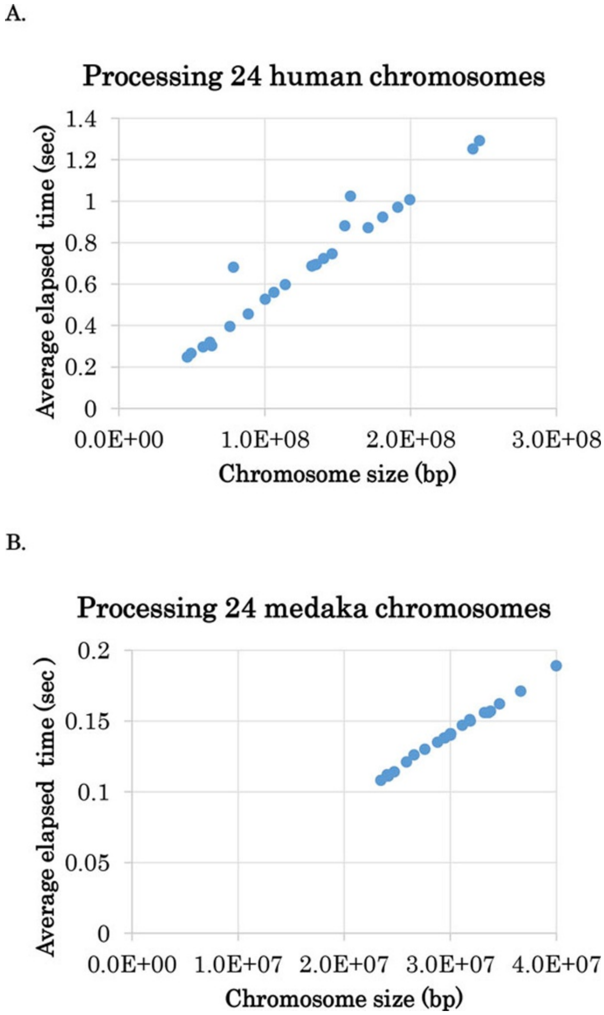

Figure 6 Average elapsed time of processing human (A) and medaka (B) chromosomes ten times by using CSMinfinder. The minimum threshold is set to 8 kbp for handing the human genome, and 4 kbp for the medaka genome. Each dot represents a chromosome, the x-axis value shows the size of the chromosome, and the y-axis value is the average elapsed time.

Conclusions and discussion

In this work, we proposed a method that estimates large K27HMD regions [5-8,12] by calculating the similarity between the vector of focal epigenetic states and that of raw epigenetic states at each DNA position. The advantage of this algorithm (CSMinfinder) is the output of an optimal series of regions while allowing us to set the minimum length threshold on individual regions. We estimated large K27HMD in the medaka and human genomes and verified that CSMinfinder was comparable to Nakamura's heuristic method [8] designed to detect K27HMD and was advantageous over ChromHMM in terms of the lengths of K27HMD regions.

For the medaka genomic data, ChromHMM performed well and could detect as many long regions as CSMinfinder did; however, for the human genomic data, ChoromHMM found a smaller number of large K27HMD regions of length > 8 kbp than the other two methods did. This was likely due to the differences in characteristics between the medaka and human genomic data. In the medaka genome, the data were collected from an inbred stain in which the genomic differences between the two alleles were quite small. Thus, methylation levels were bimodal and were clearly divided into two states, hypomethylated and hypermethylated, making it relatively easy to identify blocks of hypomethylated domains. In the human genome, however, the majority of methylation levels were poised because the human genome is diploid intrinsically and allele-specific methylation is prevalent, making it more difficult to detect clear boundaries between hypermethylated and hypomethylated domains. Although many DNA methylation levels are ambiguous in the human genome, ChromHMM attempts to assign one state to each position. Positions with vague DNA methylation levels are assigned only a single state by ChromHMM. Thus, ChromHMM is likely to output too many short regions.

One advantage of CSMinfinder is that we can set the minimum region length for specific purposes. For example, in the medaka genome, using an 8-kbp minimum length threshold merged some of the shorter regions that were generated using a 4-kbp minimum threshold into a longer continuous region. Thus, we could obtain longer regions using a higher minimum length threshold. Similarly, we can also adjust the minimum threshold for defining similarity scores between modification vectors and the feature vector for a variety of purposes. Setting the minimum threshold to a lower value generates more regions that are less similar to the feature vector of interest. Having more than one series of regions that may overlap can be informative. We can therefore tune CSMinfinder easily to meet various demands.

In this paper, we demonstrated the advantages of our algorithm by detecting large K27HMD regions that have attracted much interest because of their importance in characterizing the behavior of developmental genes and confirmed the performance of our algorithm. CSMinfinder is not limited to the identification of large K27HMD regions but can be used for the detection of other large DNA regions that have different types of epigenetic state combinations associated with regulating gene functions.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Professor Hiroyuki Takeda, Dr. Tatsuya Tsukahara, and Dr. Ryohei Nakamura for stimulating discussions. This work was supported in part by CREST, JST (Japan Science and Technology Agency).

This article has been published as part of *BMC Genomics* Volume 16 Supplement 2, 2015: Selected articles from the Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S2

Published: 21 January 2015

References

1. Hendrich B, Tweedie S: **The methyl-CpG binding domain and the evolving role of DNA methylation in animals.** *Trends in Genetics* 2003, **19(5)**:269-277.
2. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes Dev* 2002, **16(1)**:6-21.
3. Vastenhouw NL, Schier AF: **Bivalent histone modifications in early embryogenesis.** *Curr Opin Cell Biol* 2012, **24(3)**:374-386.
4. Zhou VW, Goren A, Bernstein BE: **Charting histone modifications and the functional organization of mammalian genomes.** *Nat Rev Genet* 2011, **12(1)**:7-18.
5. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, *et al*: **Epigenomic analysis of multilineage differentiation of human embryonic stem cells.** *Cell* 2013, **153(5)**:1134-1148.
6. Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, Zhang X, Chavez L, Wang H, Hannah R, *et al*: **Large conserved domains of low DNA methylation maintained by Dnmt3a.** *Nature genetics* 2014, **46(1)**:17-23.
7. Bogdanovic O, Long SW, van Heeringen SJ, Brinkman AB, Gomez-Skarmeta JL, Stunnenberg HG, Jones PL, Veenstra GJ: **Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis.** *Genome research* 2011, **21(8)**:1313-1327.
8. Nakamura R, Tsukahara T, Qu W, Ichikawa K, Otsuka T, Ogoshi K, Saito TL, Matsushima K, Sugano S, Hashimoto S, *et al*: **Large hypomethylated domains serve as strong repressive machinery for key developmental genes in vertebrates.** *Development* 2014, 141(13):2568-2580.
9. Hu JL, Zhou BO, Zhang RR, Zhang KL, Zhou JQ, Xu GL: **The N-terminus of histone H3 is required for de novo DNA methylation in chromatin.** *Proc Natl Acad Sci USA* 2009, **106(52)**:22187-22192.
10. Cedar H, Bergman Y: **Linking DNA methylation and histone modification: patterns and paradigms.** *Nat Rev Genet* 2009, 10(5):295-304.
11. Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD, *et al*: **DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA.** *Nature* 2007, **448(7154)**:714-717.
12. Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, Grutzner F, Odom DT, Patient R, Ponting CP, *et al*: **Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates.** *eLife* 2013, **2**:e00348.
13. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, Choo A, Orlov YL, Sung WK, Shahab A, Kuznetsov VA, *et al*: **Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells.** *Cell Stem Cell* 2007, **1(3)**:286-298.
14. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, *et al*: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125(2)**:315-326.
15. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization.** *Nat Methods* 2012, **9(3)**:215-216.
16. Qu W, Hashimoto S, Shimada A, Nakatani Y, Ichikawa K, Saito TL, Ogoshi K, Matsushima K, Suzuki Y, Sugano S, *et al*: **Genome-wide genetic variations are highly correlated with proximal DNA methylation patterns.** *Genome research* 2012, **22(8)**:1419-1425.
17. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, *et al*: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462(7271)**:315-322.
18. Csuros M: **Maximum-Scoring Segment Sets.** *Transactions on Computational Biology and Bioinformatics* 2004, **1(4)**:139-150.
19. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.** *Nat Methods* 2008, **5(9)**:829-834.

doi:10.1186/1471-2164-16-S2-S8
Cite this article as: Ichikawa and Morishita: **A linear time algorithm for detecting long genomic regions enriched with a specific combination of epigenetic states.** *BMC Genomics* 2015 **16**(Suppl 2):S8.