Bioinformatics, 31(15), 2015, 2426–2433
doi: 10.1093/bioinformatics/btv163
Advance Access Publication Date: 24 March 2015
Original Paper

Genome analysis

Learning chromatin states with factorized information criteria

Michiaki Hamada[1,2,*], Yukiteru Ono[3], Ryohei Fujimaki[4] and Kiyoshi Asai[2,5]

[1]Department of Electrical Engineering and Bioscience, Faculty of Science and Engineering, Waseda University, 55N–06–10, 3–4–1, Okubo Shinjuku-ku, Tokyo 169–8555, Japan, [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2–41–6, Aomi, Koto-ku, Tokyo 135–0064, Japan, [3]Information and Mathematical Science and Bioinformatics Co., Ltd., 4–21–1–601 Higashi-Ikebukuro, Toshima-ku, Tokyo 170–0013, Japan, [4]Department of Media Analytics, NEC Laboratories America, 10080 North Wolfe Road, Suite SW3-350, Cupertino, CA 95014 and [5]Graduate School of Frontier Sciences, University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa 277–8562, Japan

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on October 5, 2014; revised on March 2, 2015; accepted on March 16, 2015

Abstract

Motivation: Recent studies have suggested that both the genome and the genome with epigenetic modifications, the so-called epigenome, play important roles in various biological functions, such as transcription and DNA replication, repair, and recombination. It is well known that specific combinations of histone modifications (e.g. methylations and acetylations) of nucleosomes induce chromatin states that correspond to specific functions of chromatin. Although the advent of next-generation sequencing (NGS) technologies enables measurement of epigenetic information for entire genomes at high-resolution, the variety of chromatin states has not been completely characterized.

Results: In this study, we propose a method to estimate the chromatin states indicated by genome-wide chromatin marks identified by NGS technologies. The proposed method automatically estimates the number of chromatin states and characterize each state on the basis of a hidden Markov model (HMM) in combination with a recently proposed model selection technique, factorized information criteria. The method is expected to provide an unbiased model because it relies on only two adjustable parameters and avoids heuristic procedures as much as possible. Computational experiments with simulated datasets show that our method automatically learns an appropriate model, even in cases where methods that rely on Bayesian information criteria fail to learn the model structures. In addition, we comprehensively compare our method to ChromHMM on three real datasets and show that our method estimates more chromatin states than ChromHMM for those datasets.

Availability and implementation: The details of the characterized chromatin states are available in the Supplementary information. The program is available on request.

Contact: mhamada@waseda.jp

Supplementary information: Supplementary data are available at *Bioinformatics online*.

1 Introduction

In the eukaryotic cell nucleus, the genome forms a complex structure called chromatin. The building blocks of chromatin are nucleosomes, in which a DNA strand is wrapped around an octamer of histone proteins H2A, H2B, H3 and H4. It is known that, in nucleosomes, post-transcription modifications of the histone proteins and substitution of histone proteins with variants both occur frequently (Zhou *et al.*, 2011). Gene regulation and DNA replication, repair and recombination are greatly affected by the chromatin structures determined by not only transcription factors bound to DNA sequences but also post-translation modifications and substitutions (with variants) of the histone proteins. In molecular biology, the study of mechanisms of gene regulation that do not depend on changing the DNA sequence itself is called epigenetics, and it is an important research field in the post-genome-sequencing era.

Changes that affect chromatin structures are called 'histone marks' and include various histone modifications (e.g. methylation, acetylation, citrullination, phosphorylation, SUMOylation, ubiquitination and ADP-ribosylation) and substitution with histone variants (e.g. H2A.Z for H2A histone) (Zhou *et al.*, 2011). Recently, it has become possible to measure these histone marks at high resolution across an entire genome by using specific antibodies in combination with next-generation sequencing (NGS) technologies (Roh and Zhao, 2008; Wang *et al.*, 2008). In 1999, Dr Brian D. Strahl and Dr C. David Allis proposed the 'Histone code hypothesis', which posits that specific combinations of histone modifications correspond to specific functions of chromatins (called 'chromatin states'), in a manner similar to how combinations of nucleotide (codons) correspond to translated amino acids (Rando, 2012; Strahl and Allis, 2000). Recent studies have offered strong evidence in support of this hypothesis (Kouzarides, 2007; Ruthenburg *et al.*, 2007). For example, H3K4me2/3 (H3 lysine 4 di/trimethylation) and H3K9/14/18/23ac (H3 lysine 9/14/18/23 acetylation) characterize transcriptionally active chromatin, while H3K9me3 and H3K27me3 correspond to silent loci; Nucleosomes with both H3K4me3 and H3K27me3 characterize bivalent features of genes, which are frequently found in stem cells (Voigt *et al.*, 2013); and an inactive X-chromosome is marked by the presence of H3K9me2/3, H3K27me3 and H4K20me1.

Although, as described earlier, a large amount of experimental evidence and knowledge has been collected about histone modifications in chromatins, definitive answers to the following questions are still elusive: How many chromatin states exist? What are the possible kinds? (Baker, 2011) The attempt to answer these questions has inspired the development of computational methods to estimate hidden chromatin states from accumulated observational data. Such a method is the focus of this study.

There are several computational methods for characterizing hidden chromatin states from observed experimental chromatin marks (Biesinger *et al.*, 2013; Cielik and Bekiranov, 2014; Ernst and Kellis, 2012; Hoffman *et al.*, 2013; Lai and Buck, 2013; Larson and Yuan, 2010); See also Supplementary Section S4.1 for a review. A pioneering and popular tool for one method of characterization is ChromHMM (Ernst and Kellis, 2010, 2012), which binarizes each chromatin mark with respect to 200-base-pair (bp) intervals in genomes and then models the binarized data by using multivariate hidden Markov models (HMMs) whose hidden states correspond to chromatin states (The size of interval can be changed in ChromHMM). Because the conventional learning algorithm for HMMs (i.e. the Baum–Welch algorithm) does not determine the model structure (e.g. the number of chromatin states), ChromHMM employs Bayesian information criteria (BIC) (Schwarz, 1978) to estimate the variety of chromatin states (i.e. the number of distinct states). Although model selection procedures frequently use BIC, quality guarantees with BIC are available only for probabilistic models that satisfy certain regularity conditions (see, e.g. Konishi and Kitagawa, 2007 for the details of these regularity conditions). Unfortunately, HMMs do not typically satisfy these conditions (Yamazaki and Watanabe, 2005), and so the model selection by ChromHMM is not mathematically well-founded. Ernst and Kellis (2010) has recognized this limitation of BIC and primarily relied on intuition and biological interpretations to determine chromatin states. It is, however, important to develop unbiased methods of characterizing chromatin states to ensure that overly optimistic conclusions are not being reached.

In this study, we propose a new method for automatically estimating the variety of chromatin states and simultaneously characterizing each state. This method is based on a recently proposed model selection method that uses so-called factorized information criteria (FIC) for selection (Fujimaki and Hayashi, 2012). It has been shown that the use of FIC is more appropriate than the use of BIC for selection in non-regular models, including HMMs (which are used in ChromHMM). As an additional benefit, FIC includes only two adjustable parameters and avoids heuristic procedures as much as possible when learning the underlying model structures. Model selection by FIC is therefore expected to obtain a more nearly unbiased chromatin state model than other approaches obtain. To confirm this expectation, we carried out extensive computational experiments on both simulated and real datasets, and the results indicate the usefulness of the proposed method.

The article is organized as follows. In Section 2, we describe the theory and methods used in this study. In Section 3, we describe computational experiments on both simulated and real datasets. Additional results from testing with both simulated and real data are shown in the Supplementary Information (SI).

2 Methods

2.1 Notation

In this article, we use the following notations (cf. Fig. 1). $\mathcal{C}$ denotes a set of chromosomes and C denotes the number of chromosomes in that set (i.e. $C = |\mathcal{C}|$). T_c is the number of disjoint intervals in a chromosome $c \in \mathcal{C}$. A specific interval is indicated by an integer $t \in \{1, 2, \ldots, T_c\}$. In this study, as has been done in previous studies (Ernst and Kellis, 2010; Ernst *et al.*, 2011), each interval indicates a 200-bp (disjoint) interval in the human genome. This width is chosen because about 200 bp is considered to correspond to one nucleosome (DNA sequences wrapped around core histones plus linker DNA sequences). Every binarized chromatin mark is considered to be observed or not with respect to each 200-bp interval across entire genome sequences. $\mathscr{M}$ is a set of binarized chromatin marks (e.g. H3K4me2, H3K9ac and CTCF), and M is the number of such marks (i.e. $M = |\mathscr{M}|$). K is the (unknown) number of chromatin states, which needs to be estimated. The value of x_c^t ($c \in \mathcal{C}$) indicates the

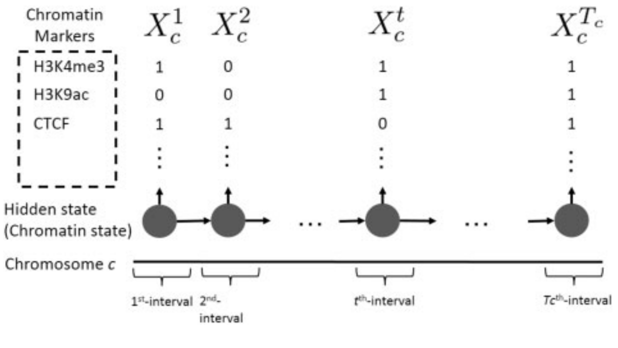

Fig. 1. Illustrated example of chromatin states and chromatin marks. For each interval (disjoint 200-bp segment in genomic sequences), the binarized chromatin marks [e.g. H3K4me3 is present (1) or absent (0)] are considered to be induced by one of the (unknown number of) chromatin states. Model selection strategies are useful for learning the number and types of chromatin states

presence of interval t in chromosome c, and x_c^t is a random variable in $\{0, 1\}^M$, with x_{cm}^t representing the status of the chromatin mark m in interval t of chromosome c. That is,

$$x_{cm}^t = \begin{cases} 1 & \text{if the mark } m \text{ is present in interval } t \text{ of chromosome } c \\ 0 & \text{otherwise.} \end{cases}$$

The value of z_c^t indicates the chromatin state, represented as an integer in $\{1, 2, \ldots, K\}$, of interval t in chromosome c.

2.2 Multivariate HMM

In this study, we use a multivariate HMM for modeling the binarized chromatin marks (Ernst and Kellis, 2010, 2013; Ernst *et al.*, 2011); the HMM is mathematically described by the following.

An observation $x = \{x_c\}_{c\in\mathcal{C}} = \{\{x_c^t\}_{t=1}^{T_c}\}_{c\in\mathcal{C}}$ can be modeled by a multivariate HMM according to Equation (1).

$$p(x|\theta) = \prod_{c\in\mathcal{C}} p(x_c|\theta) \tag{1}$$

Here, θ is a parameter of the model, with $\theta = \{\alpha, \beta, \phi\}$, and $p(x_{\mathrm{c}}|\theta)$ is the marginal distribution with respect to the chromatin states $z_c = \{z_c^t\}$. This can be written in the following form.

$$\begin{aligned} p(x_c|\theta) &= \sum_{z_c} p(x_c, z_c|\theta) \\ &= \sum\nolimits_{z_c} p(z_c^1|\alpha)p(x_c^1|z_c^1, \phi)\prod_{t=2}^{T_{\mathrm{c}}} p(x_c^t|z_c^{t-1}, \beta)p(x_c^t|z_c^t, \phi), \end{aligned}$$

where $p(z_c^1|\alpha) = \alpha_{z_c^1}$ $(\alpha = (\alpha_1, \ldots, \alpha_K) \in [0, 1]^K)$ is an initial probability distribution chosen such that $\sum\nolimits_{k=1}^K \alpha_{\mathrm{k}} = 1$ holds; and $p(z_c^t|z_c^{t-1}, \beta) = p(z_c^t|\beta_{z_c^{t-1}}) = \beta_{z_c^{t-1}z_c^t}$ is a transition probability distribution for which $\sum\nolimits_{k=1}^K \beta_{jk} = 1$ holds for all j. In this study, the emission probability $p(x_c^t|\phi_k)$ is given by the following:

$$p(x_c^t|\phi_k) = \prod_{m=1}^M (\phi_{km})^{x_{cm}^t}(1-\phi_{km})^{1-x_{cm}^t}, \tag{2}$$

where $\phi_k \in [0, 1]^M$. The expression ϕ_{km} for a chromatin state k and a chromatin mark m represents the probability that the mark m is present. Note that, in Equation (2), we assume the independence of emission of chromatin marks in order to reduce the number of parameters.

2.3 Concept of FIC with HMMs

Because $p(x|\theta)$ in Equation (1), which is obtained as a marginal probability over the hidden chromatin states, does not fulfill the regularity conditions (Konishi and Kitagawa, 2007) necessary for theoretical guarantees, neither AIC nor BIC (nor any other set of information criteria with regularity requirements) is applied to this model. Instead, we apply the FIC measure, which was described in a published article (Fujimaki and Hayashi, 2012). In this section, we describe the conceptual basis of FIC as applied to HMMs (FIC-HMMs).

The logarithm of the joint probability of the observed data x (chromatin marks) and the hidden data z (chromatin states) can be computed as follows:

$$\begin{aligned} \log p(x, z|\theta) &= \sum_{c\in\mathcal{C}} \log p(z_c^1|\alpha) \\ &+ \sum_{c\in\mathcal{C}}\sum_{t=1}^{T_{\mathrm{c}}-1} \log\ p(z_c^{t+1}|\beta_{z_c^t}) + \sum_{c\in\mathcal{C}}\sum_{t=1}^{T_{\mathrm{c}}} \log p(x_c^t|\phi_{z_c^t}). \end{aligned} \tag{3}$$

Because $p(\cdot|\alpha)$, $p(\cdot|\beta_k)$ and $p(\cdot|\phi_k)$ satisfy the necessary regularity conditions, we can employ the Laplace method (Laplace, 1986) for each of them (which is similar to the derivation of BIC given in, e.g. Konishi and Kitagawa, 2007), and this leads to

$$\begin{aligned} \log p(x, z|\theta) &\approx \log p(x, z|\hat{\theta}) \\ &- \frac{\mathcal{D}_\alpha}{2}\log C - \sum_{k=1}^K \frac{\mathcal{D}_{\beta_{\mathrm{k}}}}{2}\log n_k(z') - \sum_{k=1}^K \frac{\mathcal{D}_{\phi_k}}{2}\log n_k(z), \end{aligned} \tag{4}$$

where $\hat{\theta}$ denotes the maximum likelihood estimates, $n_k(z)$ is the number of times that chromatin state k appeared in $z = \{z_c\}_{c\in\mathcal{C}}$, the value $n_k(z')$ is the number of times that the chromatin state k appeared in $z' = \{\{z_c^t\}_{t=1}^{T_{\mathrm{c}}-1}\}_{c\in\mathcal{C}}$ (i.e. a sequence of all chromatin states except the last one), and $\mathcal{D}_\alpha$, $\mathcal{D}_{\beta_k}$ and $\mathcal{D}_{\phi_k}$ are the degrees of freedom of the parameters α, β_{k} and ϕ_{k}, respectively (i.e. in our case, $\mathcal{D}_\alpha = \mathcal{D}_{\beta_{\mathrm{k}}} = K - 1$ and $\mathcal{D}_{\phi_k} = M$). In earlier, the term $\log p(x, z|\hat{\theta})$ in the right-hand side is equal to the log likelihood for the maximum likelihood estimate $\hat{\theta}$, and the other terms can be considered as a kind of penalty imposed on the sizes of parameters, which also encompasses the dependency between hidden variables and parameters.

The conventional Baum–Welch algorithm for training parameters in HMMs is derived from the expectation-maximization algorithm (Durbin *et al.*, 1998) with the following lower bounds for the marginal probabilities, with respect to the hidden states, of observed data:

$$\log p(x|\theta) = \log\sum_{\mathrm{z}} p(x, z|\theta) \geq \sum\nolimits_{\mathrm{z}} q(z)\log\left(\frac{p(x, z|\theta)}{q(z)}\right). \tag{5}$$

These hold for any probability distribution $q(\cdot)$ of chromatin states, where the sums in the second and third terms are computed for all possible combinations of chromatin states of z. Although, in the derivation of the Baum–Welch algorithm, $\log p(x, z|\theta)$ in Equation (5) is substituted with Equation (3), in the derivation of the training algorithm for FIC, the value of $\log p(x, z|\theta)$ is substituted with Equation (4).

The detailed methods as well as the algorithm are shown in Supplementary Section S1.

3 Results

3.1 Results for simulated datasets

To evaluate the model selection ability of the proposed method (FIC-HMM), we performed computational experiments on the simulated datasets, where previously proposed chromatin state models were utilized as true model and simulated observed epigenetic data from the models (see Supplementary Section S2.2 for the details of the simulated datasets). Because the true model (which generates the data on observed chromatin marks) is known for the simulated datasets, we can precisely evaluate the correctness of the estimated models. For FIC-HMM, as described in Supplementary Section S1.4, we let $\varepsilon = 1$ (the threshold for reducing the number of chromatin states as described in Supplementary Section S1.2.3) and $K_{\max} = 100$ (the maximum number of chromatin states). Moreover, we tried five randomly initialized parameters of the initialized model to reduce the risk of poor but locally optimal models; the model with the maximum FIC score (Supplementary Section S1.2.5) was selected.

Table 1 summarized the number of chromatin states estimated by FIC-HMM (proposed) and BIC-HMM (used in part by ChromHMM; Ernst and Kellis, 2012) for the three simulated datasets. The results indicate that, for each dataset, our approach (FIC-HMM) successfully learned the number of chromatin states from

Table 1. Numbers of chromatin states in simulated datasets as estimated by FIC-HMM (proposed) and BIC-HMM

Dataset	Id	True	FIC-HMM	BIC-HMM
Ernst2010	1	51	41	102
	2	51	41	102
	3	51	48	102
Ernst2011	1	15	15	29
	2	15	15	30
	3	15	15	28
Hoffman2013	1	25	26	50
	2	25	25	50
	3	25	27	50

Each cell in the columns ‘FIC-HMM’ and ‘BIC-HMM’ shows the number of states estimated by multivariate HMMs (Section 2.2) trained by FIC and BIC, respectively. In BIC-HMM, the number of chromatin states for the model with the maximum BIC value is shown (we computed the BIC from 2 to two times of the number of states of the true model). The column ‘True’ shows the number of chromatin states in the true models. We stochastically generated three simulated datasets (label as No. 1, No. 2 and No. 3) for each true model (Supplementary Section S2.2). See Supplementary Tables S7–S9 in the SI for detailed results from FIC-HMM.

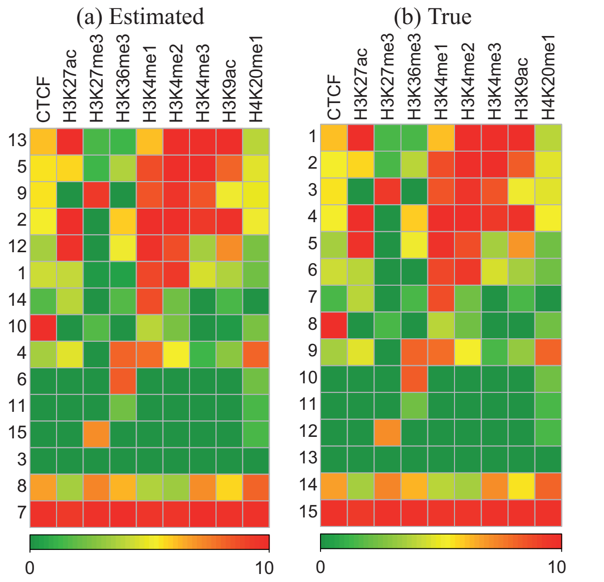

Fig. 2. Comparison of (the logarithms of) emission probabilities ϕ_{km}, where k indicates a chromatin state (vertical axis) and m indicates a chromatin marks (horizontal axis), between **(a)** the probability as estimated by FIC-HMM (with $K_{\max} = 100$) and **(b)** the emission probability of the true model for Ernst2011 dataset (Supplementary Section S2.2.2). Warmer colors indicate that the state is likely to emit a ‘positive’ mark ($\phi_{km} = 1$) for the corresponding histone mark. The states estimated by the FIC-HMM are sorted according to the emission profiles of the true model

the simulated chromatin marks. Moreover, the results obtained from FIC-HMM appear to be robust to stochastic effects of dataset generation and parameter initialization of the model. See Supplementary Tables S7–S9 in the SI for detailed results.

Table 1 also shows that, in contrast to results for the other datasets, FIC-HMM estimated fewer states than the true number of chromatin states for the Ernst2010 dataset; FIC-HMM estimated 41, 41 and 48 states while the number of chromatin states in the true model is 51. We investigated this point in detail. In the dataset, seven states of the true model (37, 40–45) were reduced to one state in FIC-HMM (Supplementary Fig. S3a–c). Those states have similar profiles of emission probabilities (Supplementary Fig. S2d) and transition probabilities (Supplementary Fig. S2d), and we suspect that it is difficult to discriminate states with similar emission profiles by using FIC-HMM.

Moreover, we compared the results from FIC-HMM with those from HMM with BIC (BIC-HMM). The results showed that model selection with BIC failed to estimate the true number of chromatin states, which might be due to applying BIC to an HMM (i.e. to a non-regular model). The results also indicate that BIC-HMM tends to give larger BIC values for models with more chromatin states (see Supplementary Fig. S10 for the details), which indicates that model selection using only BIC does not work in this case. (See the ‘Introduction’ section for an explanation of the mathematical problem with using only BIC.) Because of this tendency toward larger models, post-processing is needed to reduce the number of chromatin states. For example, ChromHMM employs an additional approach to reduce redundant states in the model learned by BIC-HMM (Ernst and Kellis, 2010). Supplementary Figure S10 also shows that the BIC values seem stable around 17, 28 and 60 states, respectively, which are similar with the numbers of chromatin states estimated by FIC-HMM. However, it should be emphasized that there is no justification to choose those models in BIC.

Remarkably, not only the estimated number of chromatin states but also the emission and transition probabilities were accurately estimated by FIC-HMM (Figs 2 and 3; see also Supplementary Figs S1, S2, S4, S5, S7 and S8 for the Ernst2010, Ernst2011 and Hoffman2013 datasets, respectively). Moreover, we confirmed that the correct annotation of the chromatin state for each interval in the human genomes could be successfully recovered by using the trained HMM (Fig. 4; Supplementary Figs S3, S6 and S9 for the Ernst2010, Ernst2011 and Hoffman2013 datasets, respectively).

In summary, FIC-HMM successfully predicted the model structure (the number of chromatin states and the profiles of emission and transition probabilities) in cases for which BIC-HMM failed; we suspect that this is because FIC offers mathematical guarantees when applied to HMM that BIC does not.

3.2 Results for real datasets

3.2.1 Comparison between FIC-HMM and ChromHMM

Table 2 summarizes the number of chromatin states estimated for three real datasets (Supplementary Section S2.1), by ChromHMM (Ernst and Kellis, 2012) (which uses BIC-HMM in part) and FIC-HMM. The results indicate that more chromatin states are estimated by FIC-HMM than by ChromHMM. Specifically, with FIC-HMM (respectively, ChromHMM), 70 (respectively, 51) chromatin states were estimated for Ernst2010_real, 49 (respectively, 15) for Ernst2011_real and 68 (respectively, 25) for Hoffman2013_real. Comparisons of the intervals annotated with the chromatin states determined by FIC-HMM and ChromHMM are shown in Supplementary Figures S15 (Ernst2010_real), S24 (Ernst2011_real) and S29 (Hoffman2013_real). These figures indicate that one chromatin state in the model estimated by ChromHMM corresponds to several chromatin states in the model estimated by FIC-HMM. The FIC-HMM method is more suitable than ChromHMM for chromatin state estimation because ChromHMM is based on BIC-HMM (which is theoretically unsuited to HMMs) with the number of resulting states reduced by a heuristic approach (Ernst and Kellis, 2010). In our experiments in simulated datasets, BIC-HMM consistently estimated many more states than were present in the true model (cf. Table 2).

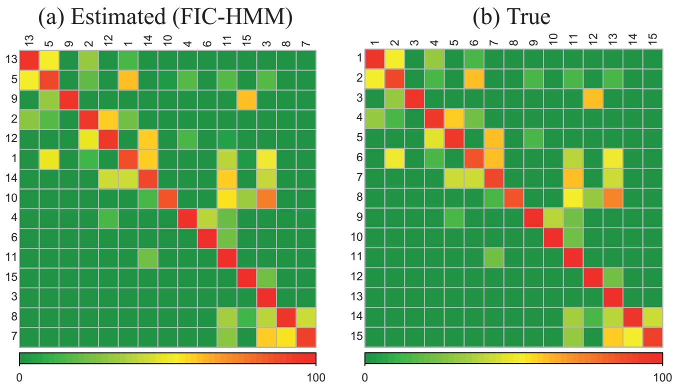

Fig. 3. The (logarithms of) transition probabilities $\{\beta_{kj}\}_{k,j}$ (described in Section 2.2) **(a)** as estimated by FIC-HMM and **(b)** of the true model for the Ernst2011_simulated dataset (Supplementary Section S2.2.2). Warmer colors indicate higher transition probabilities. For FIC-HMM, the initial number of chromatin states was 100, and the number of states was successfully estimated. The chromatin states estimated by FIC-HMM are sorted according to the emission profiles of the true model. See the SI for the results of alternative datasets and the datasets of different models

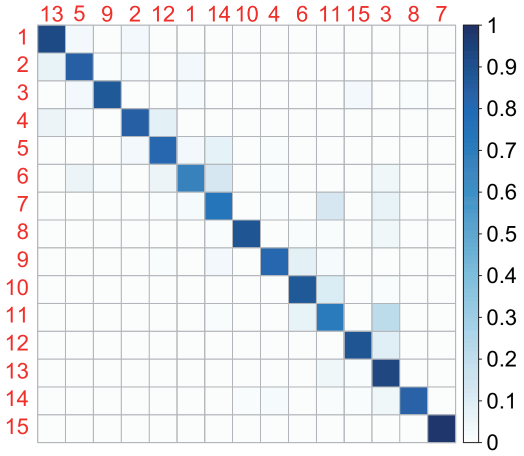

Fig. 4. Comparison of intervals of the Ernst2011_simulated dataset annotated by FIC-HMM and the true model, where we employ Viterbi decoding for our annotation (i.e. each interval is annotated with the chromatin state generating the path with the highest probability). The vertical axis indicates the chromatin states of the true model and the horizontal axis indicates the chromatin states of FIC-HMM. Each cell (*i, j*) shows the probability $p_{ij} = c(i,j)/\sum_i c(i,j)$, where *c*(*i, j*) is the count of the intervals of the chromatin state *i* (in ChromHMM) that corresponds to the chromatin state *j* (in FIC-HMM). The higher probabilities along the main diagonal indicate that most intervals were identically annotated by the two methods

We also remark that although FIC-HMM estimates more chromatin states than ChromHMM for the three real datasets, there are several similarities between the models learned by ChromHMM and FIC-HMM (cf. Supplementary Figs S15, S24 and S29). However, the chromatin states have different emission profiles for the chromatin marks and shows different profiles of enrichment with external genomic functions (cf. Fig. 5; also see Section 3.2.2).

Additionally, in order to show the usefulness of chromatin state models learned by FIC-HMM, we compared FIC-HMM and ChromHMM with respect to predictions of promoters and enhancers. On promoter predictions, both methods achieved similar prediction performance (Supplementary Table S5). On enhancers predictions, FIC-HMM achieved better performance than ChromHMM (Supplementary Table S6), indicating that the usefulness of chromatin state models estimated by FIC-HMM. Specifically, the chromatin state models learned by ChromHMM failed to discriminate between enhancers and promoters in some cases while the ones estimated by FIC-HMM successfully discriminated them, indicating that larger number of chromatin states of FIC-HMM is useful to precise annotation of genomic features from epigenetic data. See Supplementary Section S4.2 for the details.

Table 2. Number of chromatin states estimated for three real datasets

Dataset	FIC-HMM (proposed)	ChromHMM
Ernst2010_real	70	51
Ernst2011_real	49	15
Hoffman2013_real	68	25

FIC-HMM indicates multi-variable HMM in combination with FIC score for estimating the number and types of chromatin states, as proposed in this study, while ChromHMM (Ernst and Kellis, 2012) adopts BIC-HMM with subsequent reduction of the number of states by a heuristic approach (Ernst and Kellis, 2010). Value in cells is the number of estimated chromatin states. The numbers of chromatin states for ChromHMM were taken from previously published results: (Ernst and Kellis, 2010) for Ernst2010_real; (Ernst and Kellis, 2012) for Ernst2012_real; (Hoffman *et al.*, 2013) for Hoffman2013_real. See the main text for details of the estimated model parameters (e.g. emission probabilities for each state and transition probabilities between pairs of chromatin states).

3.2.2 Detailed investigation of chromatin state model learned by FIC-HMM

The detailed emission profiles of the chromatin state models trained by FIC-HMM are shown in Supplementary Figures S11, S20 and S25 for the Ernst2010, Ernst2011 and Hoffman2013 real datasets, respectively. Moreover, according to the method of (Ernst and Kellis, 2010), we computed the enrichment of each chromatin state with respect to various external genomic features, such as generic genomic functions [18, including transcription start site (TSS), transcription end site (TES), Lamina and untranslated region (UTR); Fig. 5b; see Supplementary Section S3.1.1 of the SI], transcription factor-binding sites (TFBSs) (including c-Myc, GABP and RELA; Fig. 5c; see Supplementary Section S3.1.2 of the SI) and repeated elements (37, including SINE, tRNA, rRNA, L1 and L2; Fig. 5d; see Supplementary Section S3.1.3 of the SI), which will be useful in investigating the function and characteristics of each chromatin state. An example, from the Ernst2010 dataset, is shown in Figure 5b–d.

The set of chromatin states specified by (A) in Figure 5a corresponds to the promoter or enhancer region; these have a high

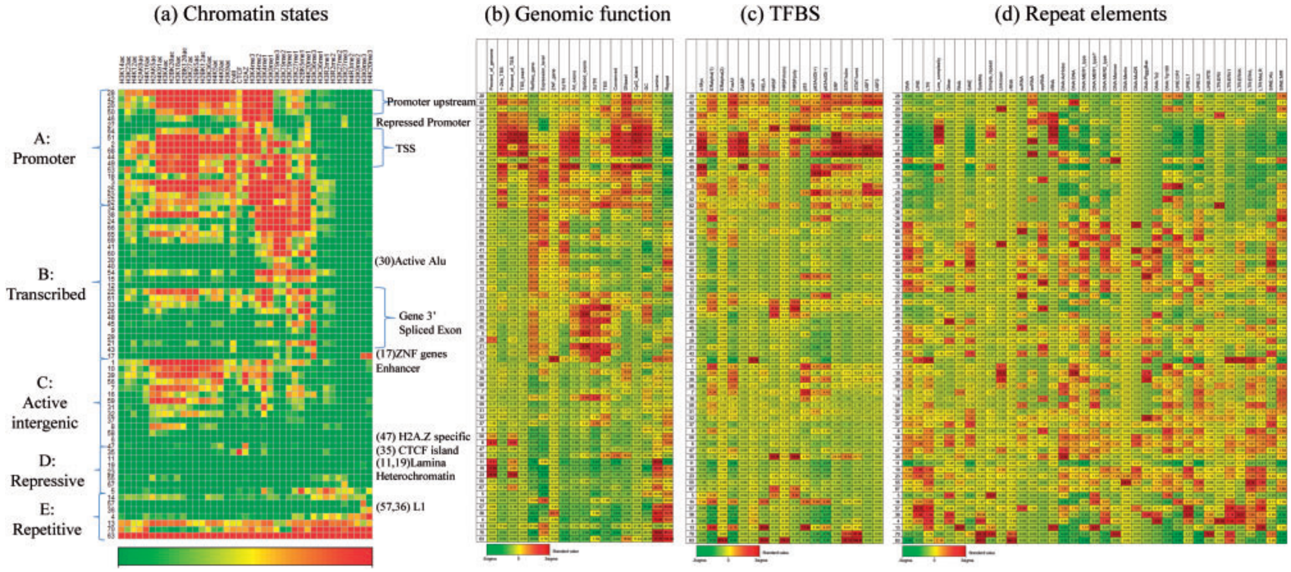

Fig. 5. Chromatin states with emission probability profiles **(a)** as estimated by FIC-HMM and in combination with enrichments by external genomic features for the Ernst_2010_real dataset (Supplementary Section S2.1.1): **(b)** general genomic functions (Supplementary Section S3.1.1 in the SI) including TSSs, TESs and UTRs; **(c)** TFBSs (Supplementary Section S3.1.2 in the SI), such as c-Myc, GABP and RELA; and **(d)** repeated elements (Supplementary Section S3.1.3 in the SI), such as Alu, SINE, L1 and L2. In all images, the numbers along the vertical axis are chromatin states. In each of the subfigures (b)–(d), the enrichment of a chromatin state with respect to a corresponding genomic features is shown, where we estimate the chromatin states of each interval (location) in the genomes by using the Viterbi algorithm (Durbin *et al.,* 1998). Warmer colors show higher degrees of enrichment. As in (Ernst and Kellis, 2010), the estimated chromatin states have been manually partitioned into five categories: (A) Promoter, (B) Transcribed, (C) Active intergenic, (D) Repressive and (E) Repetitive. See the SI, Supplementary Figure S30 for larger figures. The higher resolution figures for other datasets are also available as SI from the journal

probability of containing the chromatin mark H3K4me3 (which is known to indicate active TSSs); In fact, Figure 5b shows enrichment around TSSs for those chromatin states. Those regions also have higher DNAseI sensitivity ('DNaseI" in Fig. 5b), which correspond to open chromatin intervals. Interestingly, in addition to high probability for H3K4me3, chromatin states 46 and 27 possess a higher relative probability for having a H3K27me2/3 mark. Because H3K27me3 is known indicator of repressed regions, this suggests that those states correspond to repressed promoters. In fact, those two states have lower expression levels than other states (Expression_level in Fig. 5b). The set of chromatin states from 34 to 47 [specified by (B)] in Figure 5a corresponds to transcribed chromatin regions, with most chromatin states enriched in RefSeq gene (RefSeq_gene in Fig. 5). State 30, in particular, is enriched in active Alu, which might have some function. Additionally, chromatin state 17 in Figure 5a has a high probability for expressing H3K9me3 and being enriched in the zinc finger gene and the KRAB-associated protein1 (KAP1), which agrees with a previous study (Iyengar and Farnham, 2011).

The set of chromatin states from 1 to 47 [identified by (C) in Fig. 5] correspond to active intergenic regions. Chromatin state 47 has a high probability for only H2A.Z (a variant of H2A histone); it is not, however, observed to have specific enrichment of general genomic functions (Fig. 5b).

The set of chromatin states 35 to 5 [identified by (D)] in Figure 5a correspond to repressive chromatin regions. More than 50% of genomic regions correspond to these (repressive) chromatin states (Percent_of_genome in Fig. 5). Chromatin states 11 and 19 correspond to the genomic region associated with nuclear lamina. Chromatin state 35 corresponds to CTCF islands, in which CTCF has an elevated probability and other chromatin marks have depressed probabilities.

The set of chromatin states from 14 to 63 [identified by (E)] in Figure 5 correspond to regions of high repetition. For example, states 57 and 36 are enriched in retrotransposon L1.

For results on all real datasets, see Supplementary Figures S12–S14 for Ernst2010_real (in comparison with ChromHMM, Supplementary Figs S16–S19); Supplementary Figures S21–S23 for Ernst2011_real; and Supplementary Figures S26–S28 for Hoffman2013_real.

3.2.3 Learning chromatin state models for various cell types

The Ernst2011_real and Hoffman2013_real datasets include chromatin marks of nine and six cell types, respectively, while the Ernst2010_real dataset includes the chromatin marks of only one cell type (Supplementary Table S1). In this study, we performed model learning by FIC-HMM for all cell types separately. The learned models are shown for Ernst2011_real in Supplementary Figures S20–S23; and for Hoffman2013_real in Supplementary Figures S2–S28. In each dataset, it was observed that the number of estimated states, emission probabilities and transition probabilities are quite similar for all cell types (Supplementary Figs S24 and S29). This is consistent with previously reported results (Ernst and Kellis, 2013).

3.3 Summary of computational experiments

In our computational experiments, we first performed experiments on simulated datasets, which showed that FIC-HMM could recover the chromatin state model more accurately than BIC-HMM. We next applied FIC-HMM to three real datasets and investigated the enrichment of estimated chromatin states with a number of genomic features. The full set of results (including high-resolution figures) is available as SI at the journal web site.

4 Discussion

The recent advent of NGS technologies and related technologies ('*-seq' including ChIP-seq, RNA-seq, FAIRE-seq, etc.) has allowed rapid gathering of a huge amount of omics data about topics including genomes, epigenomes, variomes, transcriptomes and interactomes; this has hastened a shift in biology research from hypothesis-based approaches to data-driven approaches. In this study, we aim to contribute to data-driven biology by clarifying the biological model behind data about chromatin states.

In the field of bioinformatics, biological data are often modeled by probabilistic models, in which the probability distribution $p(x;\theta)$ of a characteristic x parameterized with θ. Examples of such models are HMMs, paired HMMs (for sequence alignments), stochastic context free grammars (for RNA secondary structures), neural networks and mixture models (Durbin *et al.,* 1998). To learn not only the parameters but also the model structures of probabilistic models is important in data-driven biology. It is known that many probabilistic models, including HMMs, do not satisfy the regularity conditions necessary for certain performance guarantees; hence, there is no mathematical justification for even some frequently used model selection criteria, such as AIC, BIC and MDL, with these models. Broadly, models that include hidden (latent) structures rarely satisfy regularity conditions. In this study, we therefore applied FIC to learning models of chromatin states by using HMMs. It is noted that, in this study, a multivariate HMMs were taken from a previous study, and we focus mainly on model selection part of the HMM. Developing techniques that use FIC for and learning other types of probabilistic models will be useful and should be pursued in the future.

Some other approaches are suitable for model selection on non-regular models (including HMMs). For example, WAIC (Watanabe, 2010) and WBIC (Watanabe, 2013) are extensions of AIC and BIC, respectively, to non-regular models. Variational Bayes (Attias, 1999; Bishop, 2006) and non-parametric Bayesian approaches (Orbanz and Teh, 2010) based on Dirichlet models also enable us to simultaneously learn model structures and parameters. In this article, we used FIC because FIC is applicable to non-regular models and not computationally burdensome (because the algorithm is similar to the conventional Baum–Welch algorithm for training HMMs), but an application of alternative approaches to the problem addressed in this article would be interesting.

In this article, we compared learning of chromatin states by FIC-HMM with learning by BIC-HMM and ChromHMM (Ernst and Kellis, 2012), and suggested that FIC-HMMs are more appropriate for both theoretical and practical reasons. It should be remarked that our method (FIC-HMM) includes only two parameters, which is expected to lead to unbiased learning of chromatin states. We consider that a method that relies little on heuristics is preferable for model selection. We observed that several differences between models learned by FIC-HMM and those learned by ChromHMM for three real datasets. We suspect that previous studies such as (Ernst and Kellis, 2010) favor a smaller number of chromatin states because a larger number of states complicate understanding. We confirmed the larger number of chromatin states estimated by FIC-HMM are useful for predicting promoters and/or enhancers.

We used binarized emission values (1 or 0) for each chromatin mark in this study, as was done in a previous study (Ernst and Kellis, 2010). It might be useful to directly model the continuous values of observed chromatin marks instead of using binarized values. From a theoretical viewpoint, FIC-HMM could be applied to continuous emission values by using certain probabilistic models (e.g. Gaussian distribution), although the number of parameters would be increased learning them would be more difficult (in particular, poor but locally optimal solutions would be more likely) than with binarized chromatin marks.

In Section 3.2.3, we reported that similar chromatin state models were learned by FIC-HMM for each of the cell types in Ernst_2011_real (nine cell types) and Hoffman_2013_real (six cell types). However, if we investigate those results in detailed, the cell-type specific annotated chromatin states (e.g. a genomic region in some specific cell types may be completely differ from in others; Ernst and Kellis, 2013) can be found, and further study of this would be a worthwhile biological investigation.

5 Conclusion

In this study, we proposed a novel method to learn the latent chromatin states from observed chromatin marks by using FIC-HMM. Our method is more theoretically suitable for this purpose than existing approaches, such as BIC-HMM, which has been adopted by the ChromHMM program. Learning with FIC-HMM is expected to provide unbiased chromatin states model from the observed chromatin marks. Comprehensive computational experiments on simulated and real datasets indicated that FIC-HMM could successfully recover the model of hidden chromatin states. Because model learning methods are important for data-driven biology, with necessitates analyzing a huge amount of accumulated biological data, a similar approach will be useful to other problems in the field of bioinformatics.

Acknowledgements

We are grateful to Dr Manolis Kellis, Dr Michael Hoffman and Dr Jason Ernst for providing the processed datasets and the results of chromatin states as estimated by ChromHMM. Computational experiments was performed on the supercomputer system provided managed by the National Institute of Genetics (NIG), Research Organization of Information and Systems (ROIS), Japan. A part of this work was conducted while MH was at the University of Tokyo; since April 1, 2014, MH has been at Waseda University.

Funding

This work was supported in part by a MEXT KAKENHI Grant-in-Aid for Young Scientists (A) (24680031) to MH and a Grant-in-Aid for Scientific Research on Innovative Areas to MH and KA.

Conflict of Interest: none declared.

References

Attias,H. (1999) Inferring parameters and structure of latent variable models by variational bayes. In: *Proceedings of the 15th Conference on* Uncertainity in Artificial Intelligence. pp. 21–30.

Baker,M. (2011) Making sense of chromatin states. *Nat. Methods,* **8**, 717–722.

Biesinger,J. *et al.* (2013) Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics,* **14**(Suppl. 5), S4.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Cielik,M. and Bekiranov,S. (2014) Combinatorial epigenetic patterns as quantitative predictors of chromatin biology. *BMC Genomics,* **15**, 76.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge: Cambridge University Press.

Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods,* **9**, 215–216.

Ernst,J. and Kellis,M. (2013) Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.*, **23**, 1142–1154.

Ernst,J. *et al*. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* **473**, 43–49.

Fujimaki,R. and Hayashi,K. (2012) Factorized asymptotic bayesian hidden Markov models. In: Proceedings of the 29th International Conference on Machine Learning.

Hoffman,M.M. *et al*. (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.

Iyengar,S. and Farnham,P.J. (2011) KAP1 protein: an enigmatic master regulator of the genome. *J. Biol. Chem.*, **286**, 26267–26276.

Konishi,S. and Kitagawa,G. (2007) *Information Criteria and Statistical Modeling (Springer Series in Statistics)*. Springer.

Kouzarides,T. (2007) Chromatin modifications and their function. *Cell,* **128**, 693–705.

Lai,W.K. and Buck,M.J. (2013) An integrative approach to understanding the combinatorial histone code at functional elements. *Bioinformatics,* **29**, 2231–2237.

Laplace,P.S. (1986) Memoir on the probability of the causes of events. *Stat. Sci.,* **1**, 364–378.

Larson,J.L. and Yuan,G.C. (2010) Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. *BMC Bioinformatics,* **11**, 557.

Orbanz,P. and Teh,Y.W. (2010) Bayesian nonparametric models. In: *Encyclopedia of Machine Learning*. New York, Springer.

Rando,O.J. (2012) Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev.*, **22**, 148–155.

Roh,T.Y. and Zhao,K. (2008) High-resolution, genome-wide mapping of chromatin modifications by GMAT. *Methods Mol. Biol.*, **387**, 95–108.

Ruthenburg,A.J. *et al*. (2007) Multivalent engagement of chromatin modifications by linked binding modules. *Nat. Rev. Mol. Cell Biol.*, **8**, 983–994.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.,* **6**, 461–464.

Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature,* **403**, 41–45.

Voigt,P. *et al*. (2013) A double take on bivalent promoters. *Genes Dev.*, **27**, 1318–1338.

Wang,Z. *et al*. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Watanabe,S. (2010) Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, **11**, 3571–3594.

Watanabe,S. (2013) A widely applicable bayesian information criterion. *J. Mach. Learn. Res.*, **14**, 867–897.

Yamazaki,K. and Watanabe,S. (2005) Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing,* **69**, 62–84.

Zhou,V.W. *et al*. (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, **12**, 7–18.