BIOINFORMATICS

Vol. 00 no. 00
Pages 1–10

Applying High-Dimensional Hidden Markov Model and Clustering Analysis in *Drosophila melanogaster* Chromatin Classification

Fengchong Wang [1,*], Supervisors:Sach Mukherjee [2,3] and Thomas Nichols [3,4*]

[1]Complexity Center of the University of Warwick, Zeeman Building, University of Warwick, Coventry CV4 7AL, the United Kingdom
[2]Netherlands Cancer Institute (NKI), Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[3]Department of Statistics, Zeeman Building, University of Warwick, Coventry CV4 7AL, the United Kingdom
[4]Warwick Manufacturing Group, International Manufacturing Centre, University of Warwick, Coventry, CV4 7AL, the United Kingdom

ABSTRACT

The traditional way of classifying chromatin is questioned by several recent epigenetic evidences. Some researchers propose one way of reclassify the chromatin which is the application of high dimensional Hidden Markov Model (HMM). The results of two models based on HMM are examined of their biological meaning in this report. To figure out the optimal cluster number of classifying chromatin, cluster number from 1 to 53 is investigated in the report. Finally, optimal cluster number is suggested by integrating biological meaning of two models based on HMM with 53 models based on k medoid clustering.

Contact: fengchongwang@gmail.com

1 INTRODUCTION

Traditionally, chromatin is considered to have two types —heterochromatin and euchromatin (Bolsover et al., 2011). The heterochromatin "tends to remain condensed in the metabolic or interphase nucleus and in prophase"(Rothwell [1988]) and is transcriptionally inactive(Swanson et al. [1967],Miglani [2007]) as opposed to euchromatin. Cytogenetically, one can distinguish them by keeping in mind that the heterochromatin is more intensely stained with DNA-specific stains(Miglani [2007]). Figure 1 shows the heterochromatin and euchromatin in the fourth chromosome of *Drosophila melanogaster* observed by microscope(Locke [1999]).

However, recent epigenetic evidences indicate that a finer classification may be more plausible. For instance, the heterochramtin in Rye (*Secale cereale*) B chromosomes is found to be transcriptionally active(Carchilan et al. [2007]). And evidence in *Drosophila melanogaster* shows that the heterochromatin can be divided into at least two nonoverlapping types which are marked by different proteins(Hediger and Gasser [2006],Sparmann and Van Lohuizen [2006],Coop et al. [2008]).

Guillaume *et al* do a purely data driven research in *Drosophila melanogaster* chromatin classification by applying Hidden Markov Model(Filion et al. [2010]). They gain the DNA-protein binding force data of the 53 proteins by applying DNA adenine methyltransferase identification (DamID) technology, a technology used to identify protein-DNA binding loci(Orian et al. [2009]). To give the readers a quick insight in the raw data, the DamID data of chromosome 2L (chr2L) of *Drosophila melanogaster* are shown by Figure 2. Figure 2 is plotted by using R language(R Development Core Team [2012a], Seidel). One can see big difference of protein-DNA binding force of different proteins in the same genomic loci.

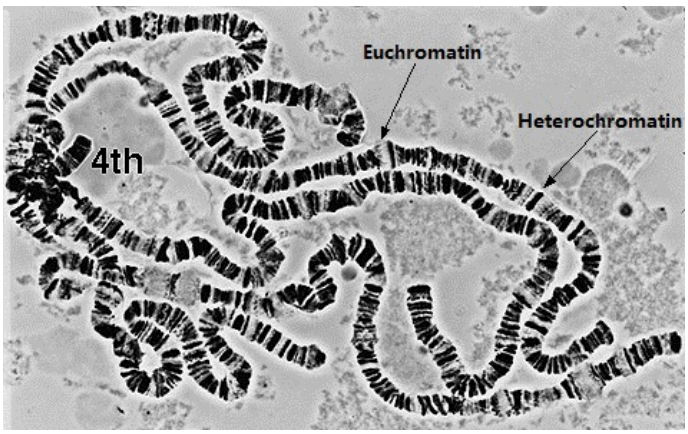


Fig. 1. The fourth chromosome of *Drosophila melanogaster* observed by microscope. The dark regions in chromosome are the heterochromatin while the light regions are the euchromatin.

Guillaum *et al* assume that there is a Markov Chain which is related to the observed data and that the emission distribution of the HMM is Student's distribution. With the initial condition of a two-state HMM and the application of Baum-Welch Algorithm (Baum et al. [1970]), optimal state number is estimated to be 5. These five principal chromatin types revealed by them are called black, blue, red, green and yellow states. But is this new classification of chromatin biologically meaningful? Can one get different classifications based on HMM or some other methods?

I investigate the relation between their classification and the

*to whom correspondence should be addressed

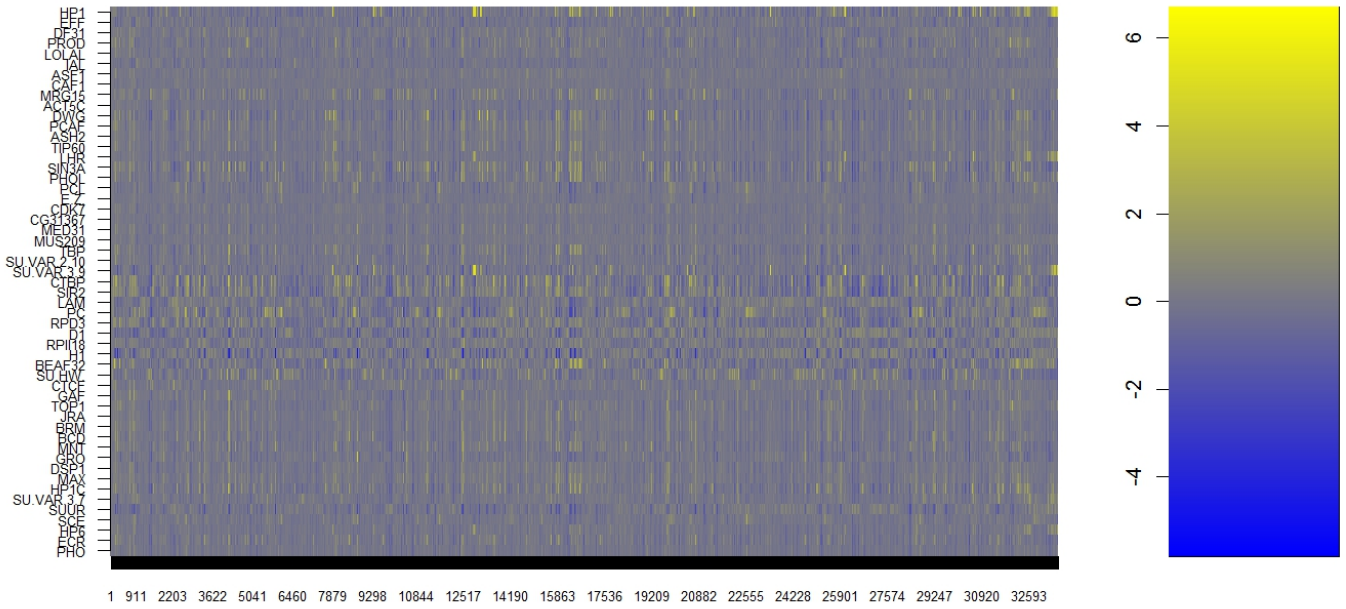


Fig. 2. Location maps of 53 proteins in chr2L of *Drosophila melanogaster*. The x axis indicates the genomic loci. The y axis indicates the different protein names. The binding forced is shown in the graph by colours. Yellow colour means high binding force. Blue colour means low binding force.

known genes in the chromosome. Similarly, a 20-state Hidden Markov Model is studied which is proposed by Nicolas Städler, the postdoc of my supervisor Sach Mukherjee. Finally, 1-cluster to 53-cluster models based on k medoid clustering are studied to figure out the optimal number of cluster.

2 METHODS

2.1 Definition

2.1.1 Biologically meaningful In this study, we think a model is biologically meaningful if all or most of the regions that cover a gene belong to the same state. In the paper, sometimes, we call a model is "good" when it is biologically meaningful.

This definition is illustrated by an example in Figure 3. In a good classification model, most of the genes are like in situation the A and B in Figure 3. Yet in reality, one cannot expect there exists a classification model in which all genes are like in the situation A in Figure 3, unless the model is a 1-state model.

2.1.2 Coverage Proportion c_{ij} The Coverage Proportion c_{ij} of state j of a certain gene, say the i^{th} gene, is given by the following equation:

$$c_{ij} = \frac{p_i}{r_i} \tag{1}$$

in which n is the total number of genes we investigate, r_i the total number of regions in the i^{th} $(1 \leqslant i \leqslant n)$ gene and p_i the number of regions in the i^{th} gene belonging to state j.

2.1.3 Coverage Proportion Matrix $\boldsymbol{C}$ Coverage Proportion Matrix $\mathbf{C}$ is a $n \times k$ matrix in which cell of i^{th} row and j^{th} column in the Coverage Proportion Matrix $\mathbf{C}$ equals Coverage Proportion c_{ij} defined by Equation (1). k is the total number of states of the model.

2.1.4 maximum Coverage Proportion m_i The maximum Coverage Proportion m_i is a n dimensional vector in which the i^{th} element of m_i equals the largest element of row i in the Coverage Proportion Matrix $\mathbf{C}$:

$$m_i = max(c_{i1}, c_{i2}, ..., c_{ir_i}) \tag{2}$$

Recall: r_i is the total number of regions in the i^{th} $(1 \leqslant i \leqslant n)$ gene.

2.1.5 average of maximum Coverage Proprotion s_k The average of maximum Coverage Proportion s_k of a k-state model is given by the following equation:

$$s_k = \frac{1}{n}\sum_{i=1}^{n} m_i \tag{3}$$

in which m_i is given in Equation (2).

2.2 Hidden Morkov Model (HMM)

The HMM of a discrete form can be understood in the following way(Rabiner and Juang [1986]):

A system have several states $\{ s_1, s_2, ..., s_k \}$. Each time t the system can only be in a state. The state u_t at time t is only dependent on the state at time (t-1), namely

$P(u_t = s_{i^*} | u_{t-1} = s_{j_{t-1}}) =$

$$P(u_t = s_{i^*} | (u_{t-1} = s_{j_{t-1}}, u_{t-2} = s_{j_{t-2}}, ..., u_0 = s_{j_0})) \tag{4}$$

Though the state cannot be observed directly, an observer can gain some data by measuring the system at each time t. And there is a certain probablity distribution controlling the emission from the state to the data. So one can estimate which state the system is most likely to be in at time t.

Now let us have some denotations in a formal way. Recall that k is the number of states in the model. The state space is S = $\{s_1, s_2, ..., s_k\}$ Let u_t be the state of t^{th} observation o_t where o_t is a 53-dimensional vector in our case. Let A be the number of different o_t's. Let the state transition

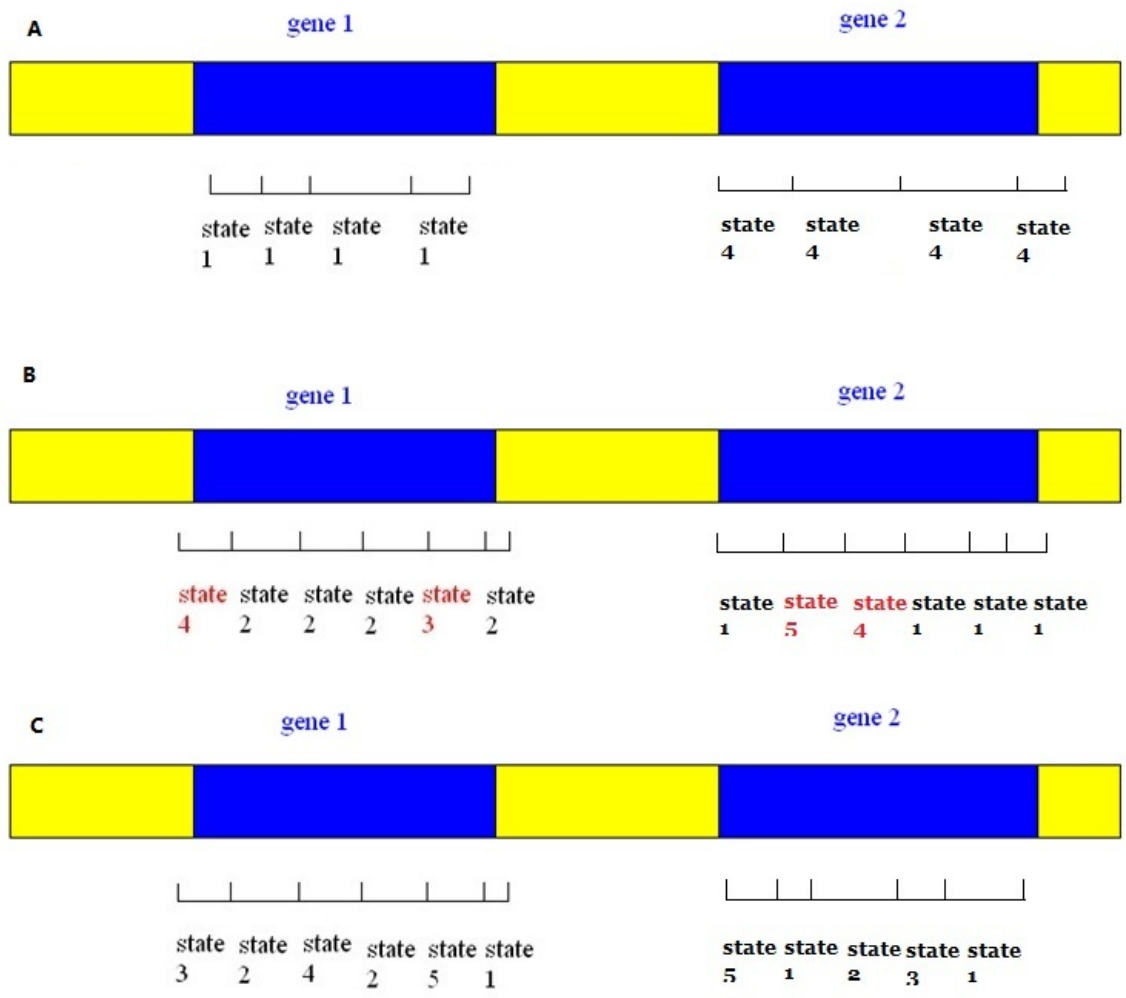


Fig. 3. Three different classification models of two genes. The blue areas represent genes. The yellow areas represent introns in chromosomes. **(A)**Both genes are covered with regions that belong to the same state (gene 1 covered by state 1, gene 2 covered by state 4). If the coverage profile of all genes is like of these two genes, this model is extremely biologically meaningful. **(B)**Each gene is covered by regions belong to a dominant state (state 2 dominates gene 1 while state 1 dominates gene 2) though there are some regions belong to other states in the gene. If the coverage profile of all genes is like of these two genes, this model is biologically meaningful. **(C)**There is no single dominant state in the gene. The state profile seems to be totally random. If the coverage profile of all the genes is like of these two genes, the model is not a "good" model.

probability distribution **Tr**$= tr_{ij}$ where

$$tr_{ij} = P(u_{t+1} = s_j | u_t = s_i), 1 \leqslant i, j \leqslant k \tag{5}$$

Let D be a set, that all o_t 's can be found in t and D = $\{d_1, d_2, ..., d_A\}$.

Let emission distribution matrix **E**=$\{e_j(l)\}$ where

$$e_j(l) = P(o_t = d_l | u_t = s_j), 1 \leqslant j \leqslant k, 1 \leqslant l \leqslant A \tag{6}$$

2.3 Baum-Welch Algorithm

Baum-Welch Algorithm was clearly explained in Rabiner's tutorial(Rabiner [1989]). The basic ideas of the algorithm are:

According to Rabiner(Rabiner [1989]),

$$P(u_t = s_i, u_{t+1} = s_j) = \frac{tr_{ij} e_j(o_{t+1}) f_{t,i} b_{t+1,j}}{\sum_{i=1}^{k} \sum_{j=1}^{k} tr_{ij} e_j(o_{t+1}) f_{t,i} b_{t+1,j}} \tag{7}$$

where

$$f_{t,i} = \sum_{j=1}^{k} f_{t-1,j} tr_{ji} e_i(o_t), f_{1,j} = P(u_1 = s_j) e_j(o_1) \tag{8}$$

known as the forward variable(Baum et al. [1970]) and

$$b_{t+1,j} = \sum_{i=1}^{k} b_{t+2,i} tr_{ji} e_i(o_{t+2}), b_{t,j} = 1 \tag{9}$$

known as the backward variable(Baum et al. [1970]). One can get the estimated $\hat{P}(u_1 = s_i)$, $\hat{tr}_{ij}$ and $\hat{e}_j(l)$ as:

$$\hat{P}(u_1 = s_i) = \sum_{j=1}^{k} P(u_t = s_i, u_{t+1} = s_j) \tag{10}$$

$$\hat{tr} = \frac{\sum_{t=1}^{t-1} P(u_t = s_i, u_{t+1} = s_j)}{\sum_{t=1}^{t-1} \sum_{j=1}^{k} P(u_t = s_i, u_{t+1} = s_j)} \tag{11}$$

$$\hat{e}_j(l) = \frac{\sum_{t=1}^{t}[\delta(o_t, d_l)\sum_{i=1}^{k} P(u_t = s_j, u_{t+1} = s_i)]}{\sum_{t=1}^{t}\sum_{i=1}^{k} P(u_t = s_j, u_{t+1} = s_i)} \quad (12)$$

where $\delta(o_t, d_l)$ is the delta function:

$$\delta(o_t, d_l) = \begin{cases} 1 & \text{if } o_t = d_l \\ 0 & \text{otherwise} \end{cases}$$

One can begin with initial guess of $P(u_1 = s_i)$, tr_{ij} and $e_j(l)$ and substitute it into the Equations (10) to (12) iteratively. It can be proven that the results converge to a model that fits the observed data better than the initial guess. In the 20-state model done by the postdoc Nicolas Städler, the emission distribution is assumed to be normal distribution.

2.4 Viterbi Algorithm

When one gets the HMM of some observed data, one can use Viterbi Algorithm(Viterbi [1967]) to generate the most likely sequence of states which fit the parameters of the HMM best. So each observation will finally correspond to a state.

How does Viterbi Algorithm work?

First, let

$$\alpha_{1,j} = e_j(o_1)P(u_1 = s_j) \quad (13)$$

$$\gamma_{1,j} = 0 \quad (14)$$

Then calculate the Equation (15) and Equation (16) recursively.

$$\alpha_{t,i} = e_i(o_t))max_{j\in[1,k]}(\alpha_{t-1,j}tr_{ji} \quad (15)$$

$$\gamma_{t,i} = argmax_{j\in[1,k]}(tr_{ji}\alpha_{ji}) \quad (16)$$

This process ends when t reaches the T we want. So the state at T is

$$s_T = argmax_{i\in[1,k]}(\alpha_{T,i}) \quad (17)$$

Then we can gain optimal state of time t (s_t) by recalling the results of every recursive step.

$$s_t = \gamma_{t,s_{t+1}} \quad (18)$$

2.5 k Medoids Algorithm

Here are the steps of k medoids algorithm (ROUSSEEUW [1987], Friedman et al. [2001], Theodoridis et al. [2010]):

Let there be T observations in total.

(1) k medoids (observations) are chosen to be the initial medoids. We call the medoids which are chosen $\{o_{j_1}, o_{j_2}, ..., o_{j_k}\}$, the observations that are not chosen $\{o_{i_1}, o_{i_2}, ..., o_{i_{(T-k)}}\}$

(2) Assign each observation o_{i_p} to a medoid o_{j_q} (q = 1, 2, ..., k) that minimizes the distance function $d(o_{i_p}, o_{j_q})$. If there are more than one o_{j_q}'s that can minimize the distance function, assign the observation o_{i_p} randomly to one of them.

(3)For x in 1 to k

{For each $o \notin o_{i_1}, o_{i_2}, ..., o_{i_{(T-k)}}$

{swap o and o_{j_x} and compute the cost function}}

(4) Choose k medoids o_j's that minimize the cost function c.

(5) Do (2) to (4) iteratively until the k medoids o_j's do not change any more.

Notice: (1)The distance function in our case is defined to be Euclidean distance. (2)cost function is defined to be

$$c = \sum_{q=1}^{k}\sum_{p=1}^{T-k} d(o_{j_q}, o_{i_p}) \quad (19)$$

The clustering function clara (short for Clustering LARge Applications) in the R package I use is based on the k medoid algorithm(Kaufman et al. [1990], R Development Core Team [2012a], Maechler et al. [2012], R Development Core Team [2012b]).

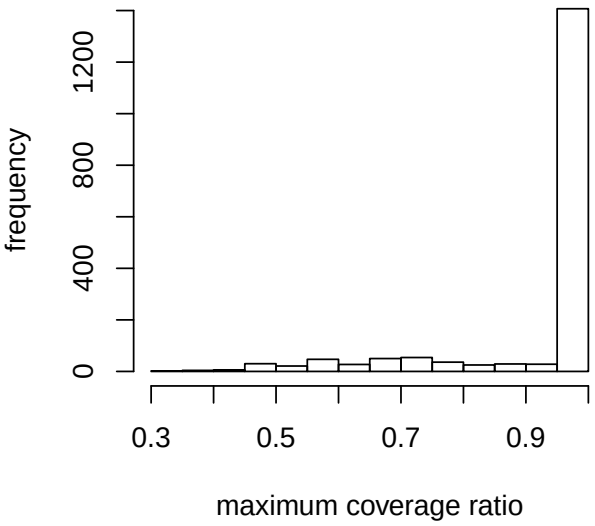


Fig. 5. Distribution of maximum coverage proportion in each gene over the whole genome in 5-state model.

3 RESULTS

3.1 How good the 5-state model is

3.1.1 Heat map of coverage proportion matrix ***C*** By observing heat map of **C**, one will see immediately some properties of the 5-state model (Figure 4). If Figure 4 is nearly homochromatic, the model is not biologically meaningful. Contrarily, if most of the colours in Figure 4 are bright yellow or dark blue, the model is biologically meaningful. Figure 4 is in accordance with the second situation, so it is biologically meaningful.

In is shown by Figure 4 that yellow state is the most "popular" in genes while green state is the most "unpopular" in genes. This will be explored in detail later.

3.1.2 Distribution of maximum coverage proportion of the 5-state model The maximum coverage proportion can reflect how small the gene is fragmented by different states. If the peak of the distribution of maximum coverage proportion is very high, say almost 1, this means that the model is very biological meaningful. Otherwise, it is not a very good model.

A remarkably high peak is observed around 1 in Figure 5. This means that the 5-state model is very biological meaningful.

3.1.3 Investigate by states The classification results of the 5-state model are studied.

The first interesting thing to investigate is how many cells of a certain column in the Coverage Proportion Matrix C have a certain range of value. If the classification is totally random and has no biological meaning, the value distribution should have a peak at 0.2. If the model is completely biologically meaningful, the values should only equal 0 and 1.

Figure 6 to Figure 10 are the histograms of the distribution of coverage proportion in each state. It is happy to see that all the histograms are very similar to our guess of ideal histogram – the peaks only occur at 1 and 0 while the number of genes of other coverage proportion are very small.

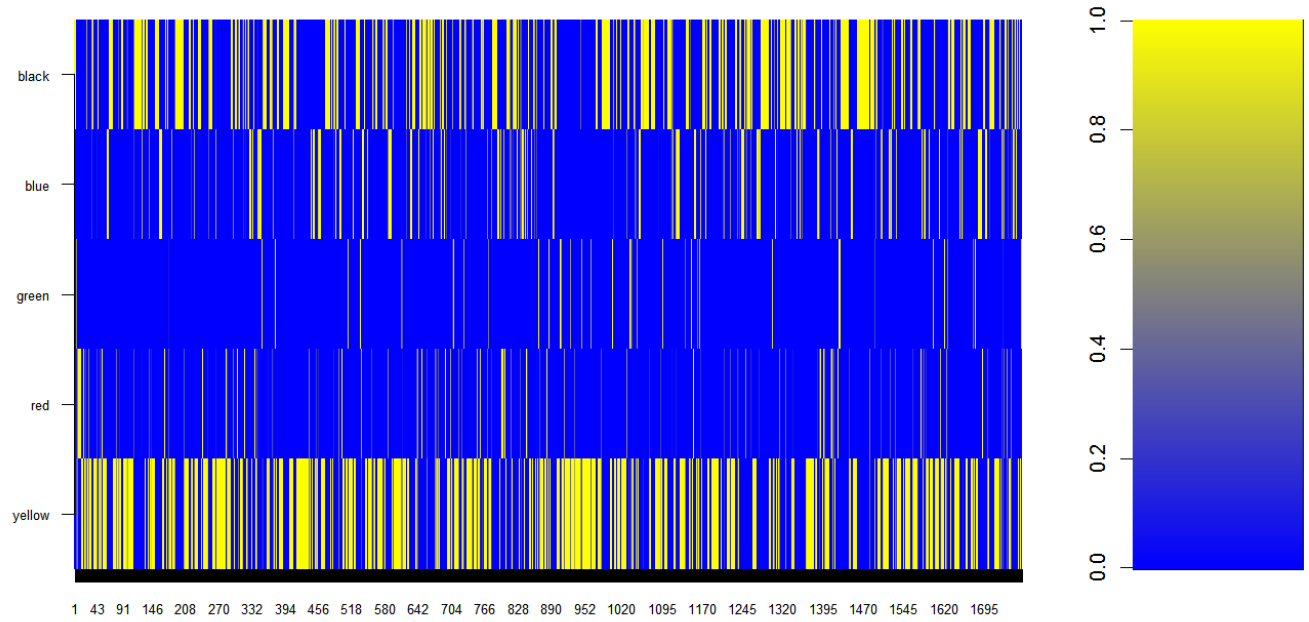


Fig. 4. Coverage Proportion of each state in each gene in chr2L of the 5-state model. x axis indicates the number of gene. y axis indicates the state name. The colours in the heat map indicate how much is the coverage proportion of each state in each gene.

This indicate that the 5-state model proposed by Guillaum *et al* is very biologically meaningful.

Shown by Figure 9, green state is the rarest in genes – there is almost no green state in genes. This is in accordance with expectation because green state is thought to correspond to the classic heterochromatin (Filion et al. [2010]).

Comparing to other states, yellow state is the most "popular" in genes. This is in accordance with the fact that yellow state corresponds to classic euchromatin (Filion et al. [2010]).

An interesting finding is that though red state is thought to correspond to classic euchromatin (Filion et al. [2010]), it is not abundant in genes.

3.1.4 States on the boundaries of genes Some regions belonging to a state are just on the boundaries of genes. Figure 11 is the bar plotting of the boundary-regions.

Figure 11 shows that yellow state is more than twice as high as any other states. This compelling property of yellow states indicates that the regions belonging to yellow state might be related to transcription initiation or termination.

3.2 The 20-state model

3.2.1 Summary of the 20-state model The classification results of the 20-state model are provided by Nicolas Städler. Unlike the emission distribution of the 5-state model, Städler assumes Gaussian distribution to be the emission distribution.

The average of the DamID data in each state in the 20-state model is shown in Figure 12. One might notice that proteins which belong to the same family tend to behave similar in a state. For example, E(Z),PC,PCL,SCE, which are PcG proteins, unlike proteins of other families, are of high average binding value in state 5 and state 7. This indicates that the 20-state model, in a way, is good.

One can see how much is the coverage proportion of each state in each

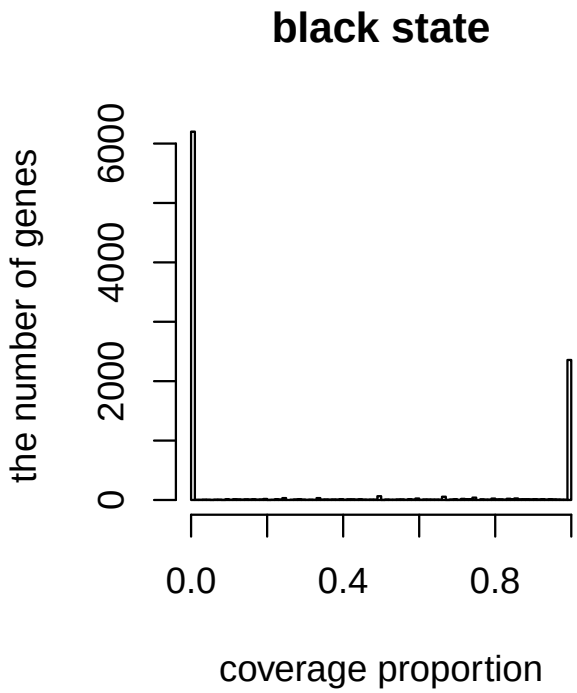


Fig. 6. Number of genes of different ranges of coverage proportion of black state in the 5-state model.

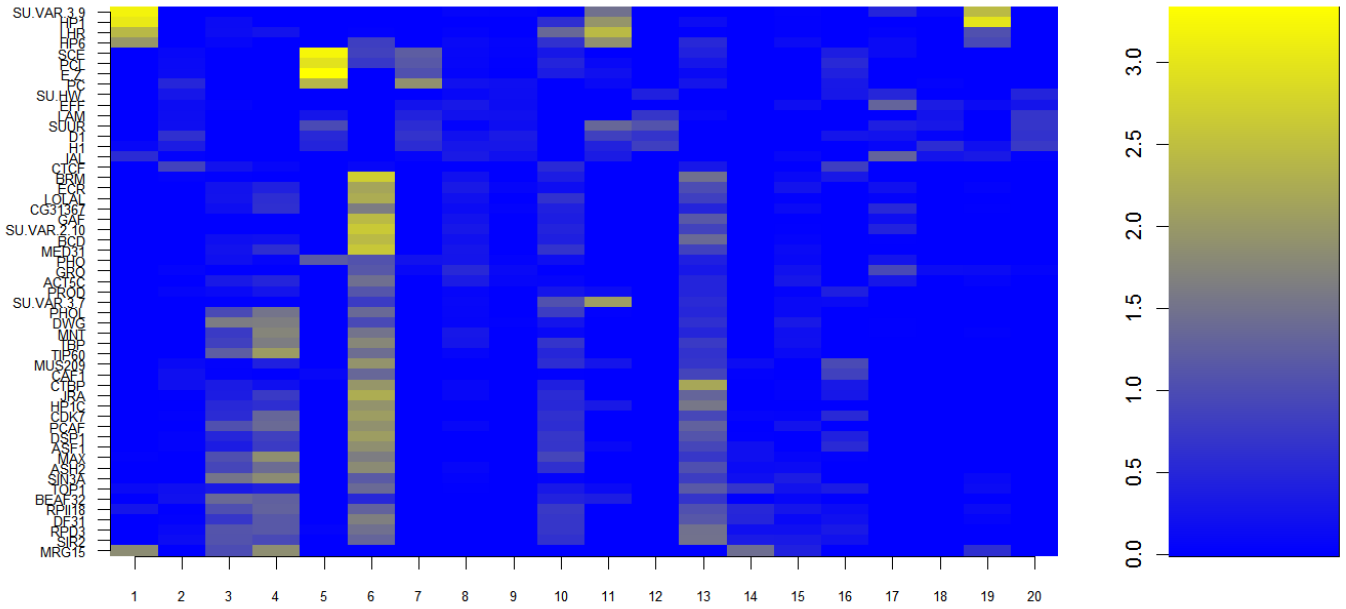


Fig. 12. Average of DamID data in each state in the 20-state model. x axis indicates the states. y axis indicate the protein names. Colour of blue indicates low binding force while yellow indicates high binding force.

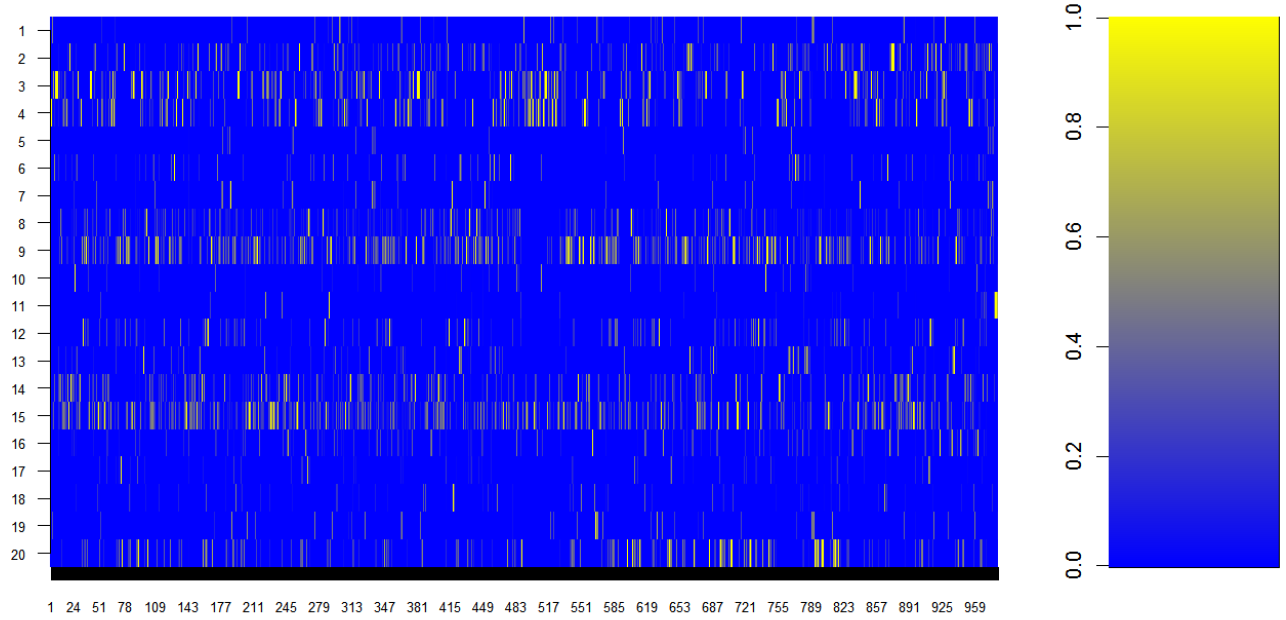


Fig. 13. Coverage Proportion of each state in each gene in chr2L in the 20-state model. x axis indicates the which gene it is. y axis indicates the states. The colours in the heat map indicate how much is the coverage proportion of each state in each gene.

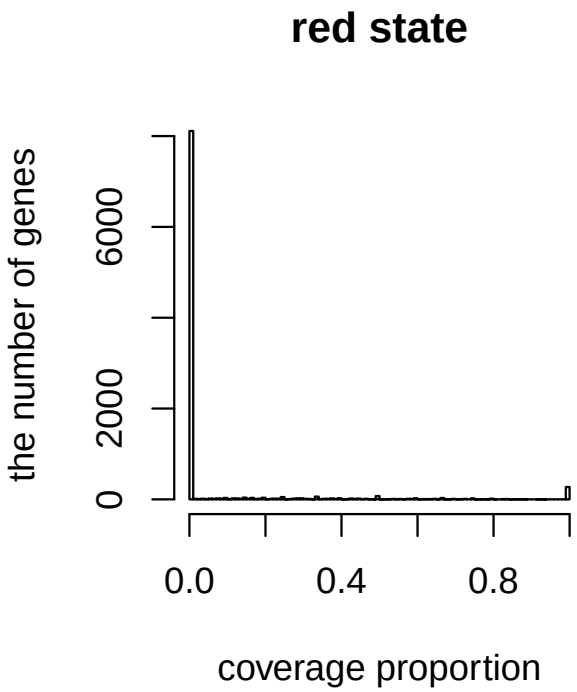


Fig. 7. Number of genes of different ranges of coverage proportion of red state in the 5-state model.

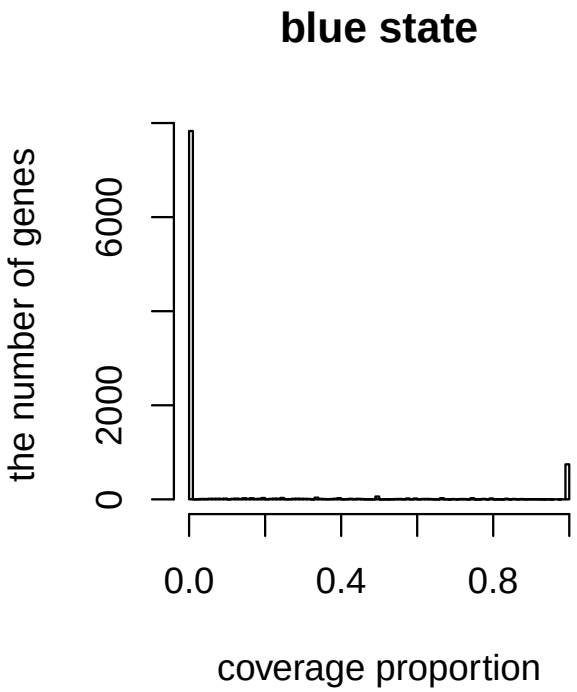


Fig. 8. Number of genes of different ranges of coverage proportion of blue state in the 5-state model.

gene by observing Figure 13. By comparing Figure 13 and Figure 4, one could see that though 20-state model is biologically meaningful, it is worse than the 5-state model. This is natural because the less states there are, the more likely that the coverage proportion will be big. An extreme case is that if there is just 1 state, the coverage proportion of that state will be 100% everywhere.

3.2.2 Distribution of maximum coverage proportion of the 20-state model Figure 14 shows the distribution of maximum coverage proportion in each gene over the whole genome in the 20-state model. Comparing to the similar plotting (Figure 5) of the 5-state model, Figure 14 seems worse. And the pie plot in Figure 14 shows that the maximum coverage proportions that are larger than 0.5 are less than 50% in all the maximum coverage proportion. These results shows that this model is less biologically meaningful than the 5-state model.

3.2.3 States on the boundaries of genes Figure 15 shows the state distribution on the boundaries of genes. We can see that some states have much higher probability of being on the boundary than other states. State 1, 2, 3, 19 together occupied more than half of the gene boundaries that have regions on them. This might indicate some unusual properties of these states.

3.3 Clustering analysis result

I did a clustering analysis of the DamID data as an alternative way to classify the chromatin.

I investigate the situations from 1 state to 53 states (k=1,2, ... 53) and plot the average of maximum coverage proportion s_k against k (Figure 16). Theoretically, the curve is likely to be monotonically decreasing if the classification does not bare much biological meaning. Surprisingly, there are some rises in the curve when the s_k is still high. Figure 16 indicates that 8 or 9 state number might be the optimal classification strategy.

4 DISCUSSION

We have already seen that the 5-state HMM works better than the 20-state model. And the cluster analysis reveals that the state number of 8 or 9 might be optimal. By taking all these results into consideration, the optimal state number might not be a very large number, say less than 10.

Also, we reveal that some states are more prone to be on the boundaries of genes than other states. These states might play critical roles in regulation. Also, the proteins (if there exist such proteins) that uniquely mark these states might have some regulatory functions concerning transcription initiation or termination.

Many further interesting researches can be done concerning the HMM application in classifying the chromatins. One can investigate other assumptions of emission distribution of the HMM or other initial conditions of the HMM. Moreover, because the Baum-Welch Algorithm can only locally maximize likelihood (Rabiner [1989]), more investigations into globally maximized likelihood might be essential.

Other clustering methods like fuzzy clustering(Bezdek [1981]) might also be reasonable. The fuzzy clustering might be ideal to reflect the phenomenon that some genes involve only in some particular stages of the body development while some keep being active

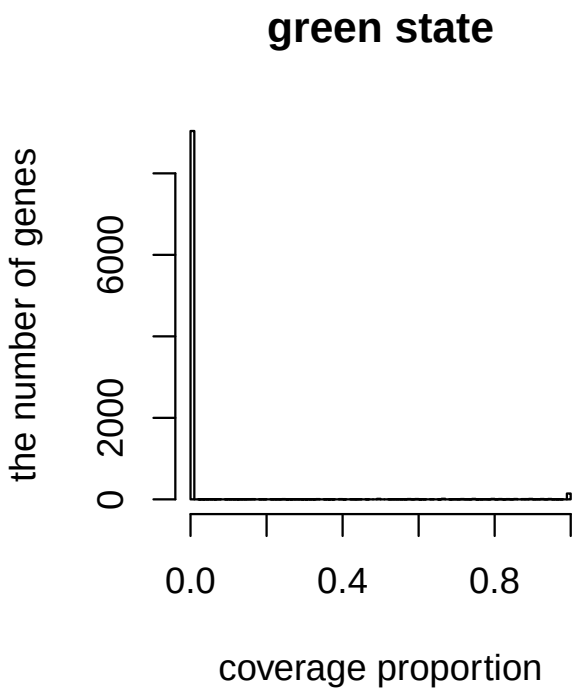


Fig. 9. Number of genes of different ranges of coverage proportion of green state in the 5-state model.

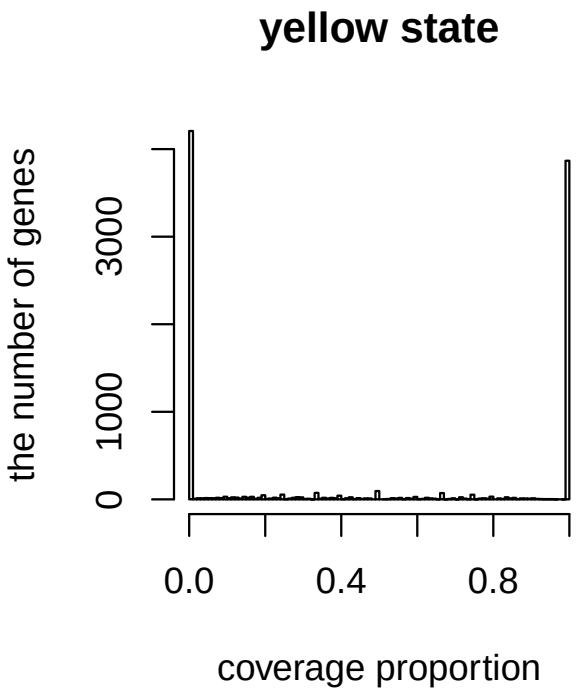


Fig. 10. Number of genes of different ranges of coverage proportion of yellow state in the 5-state model.

all the time. Each clustering might correspond to the stage of life. If a region simultaneously belong to more than one clustering, it might indicate that this region involves in more than one stages of the body development.

Furthermore, one can investigate the state distribution in introns, known regulatory domains of DNA, some unique 3 dimensional domains of DNA, the DNA regions that code microRNA ...

Also, the similar work might be done in the chromatin of other organisms, say human(*Homo sapiens*).

Finally, once one gets satisfied enough clustering results, he or she can do gene ontology analysis (GO) to check whether each clustering correspond to some specific gene functions.

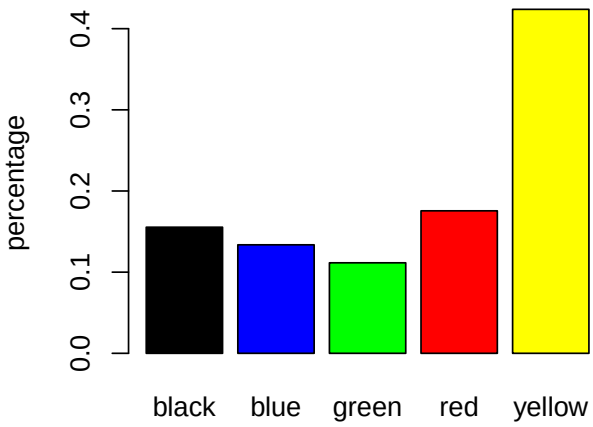
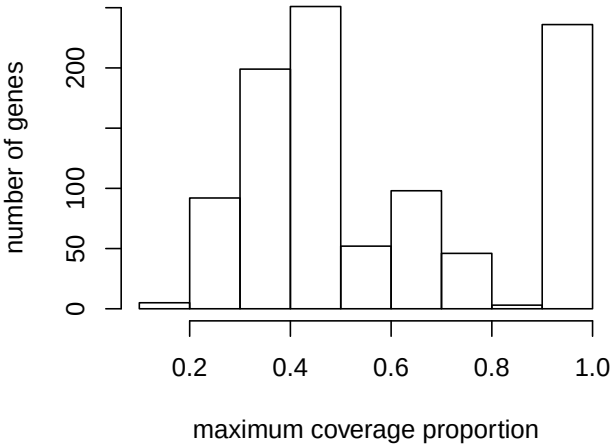


Fig. 11. Profile of states on the boundaries of genes in chr2L in the 5-state model.

ACKNOWLEDGEMENT

I thank Nicolas Städler for providing me with the 20-state HMM clustering data, the matrix of average of the DamID data in each state, raw data of the 5-state model and raising thought-provoking questions, my supervisor Professor Sach Mukherjee for his patient great instructions, my supervisor Dr Thomas Nichols for his instruction on R programming. Also I should thank Christopher Tjoeng for he giving me a lot of learning stuff. I am also grateful to Guillaume *et al* because our work is mainly based on their work and data.



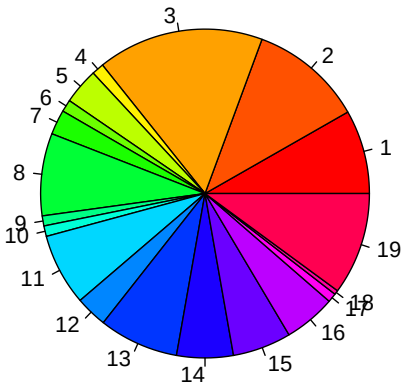


Fig. 15. State distribution on the boundaries of genes in chr2L in the 20-state model. The number indicate the state name. the area of a sector represents the percentage of a certain on-boundary state in all the on-boundary states.

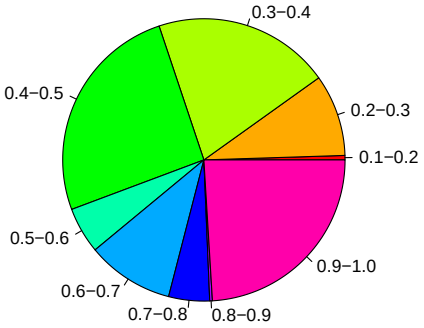


Fig. 14. Distribution of maximum coverage proportion in each gene over the whole genome in the 20-state model.

Funding: Erasmus Mundus Master Program in Complex Systems.

REFERENCES

L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

S.R. Bolsover, E.A. Shephard, H.A. White, and J.S. Hyams. *Cell biology: a short course*. Wiley-Blackwell, 2011.

M. Carchilan, M. Delgado, T. Ribeiro, P. Costa-Nunes, A. Caperta, L. Morais-Cecílio, R.N. Jones, W. Viegas, and A. Houben. Transcriptionally active heterochromatin in rye b chromosomes. *The Plant Cell Online*, 19(6):1738–1749, 2007.

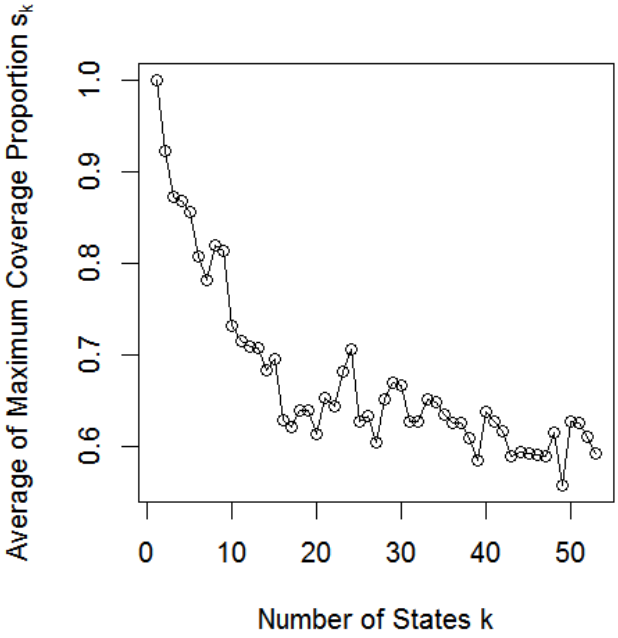


Fig. 16. Plotting of average of maximum coverage proportion against the number of states.

G. Coop, X. Wen, C. Ober, J.K. Pritchard, and M. Przeworski. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868):1395–1398, 2008.

G.J. Filion, J.G. Van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L.D. Ward, W. Brugman, I.J. De Castro, R.M. Kerkhoven, H.J. Bussemaker, et al. Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, 143(2):212–224, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.

F. Hediger and S.M. Gasser. Heterochromatin protein 1: dont judge the book by its cover! *Current opinion in genetics & development*, 16(2):143–150, 2006.

L. Kaufman, P.J. Rousseeuw, et al. *Finding groups in data: an introduction to cluster analysis*, volume 39. Wiley Online Library, 1990.

John Locke, 1999. URL `http://www.biology.ualberta.ca/locke.hp/research.htm`.

Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2012. R package version 1.14.2 — For new features, see the 'Changelog' file (in the package source).

G.S. Miglani. *Advanced genetics, second edition*. Narosa, 2007.

A. Orian, M. Abed, D. Kenyagin-Karsenti, O. Boico, et al. Damid: a methylation-based chromatin profiling approach. *Methods in molecular biology (Clifton, NJ)*, 567:155, 2009.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012a. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012b. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

L. Rabiner and B. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

N.V. Rothwell. *Understanding genetics, fourth edition*. Williams & Wilkins, 1988.

L.K.P.J. ROUSSEEUW. Clustering by means of medoids. *Statistical data analysis based on the L1-norm and related methods*, page 405, 1987.

Chris Seidel.

A. Sparmann and M. Van Lohuizen. Polycomb silencers control cell fate, development and cancer. *Nature Reviews Cancer*, 6(11):846–856, 2006.

C.P. Swanson, T. Merz, W.J. Young, et al. Cytogenetics. *Englewood Cliffs, NJ: Prentice-Hall, Inc.*, 1967.

S. Theodoridis, K. Koutroumbas, A. Pikrakis, and D. Cavouras. *Introduction to pattern recognition: a matlab approach*. Academic Pr, 2010.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.