Markov model on cell reprogramming of cells with iPSC gene data

qin.he

Dec 2019

1 Introduction

Cell reprogramming is usually a time lapse studied through gene expression data and DNA sequence data.Through computing the reprogramming rate, it is shown that more reprogramming happen under the condition of inhibition of DNA methylation or the knowckdown of somatic transcription factors. In the first model, one fixed-variable-order Bayesian tree is constructed for the identification of transcription factor binding sites(TFBSs), while in the second model, the focus is on the expression data of the cell and transcription factors. This is mainly based on the process that, the promoters of ESCs can not only bounded by their own products but also activate other pluripotent genes and inhibit lineage specific genes[1]. Thus, a markov model is applied to induce the reconstruction of transcription regulation in embryonic stem cell states, by the ectopic expression of factors and reprogramm the differentiated cells.

2 Homogeneous VOB

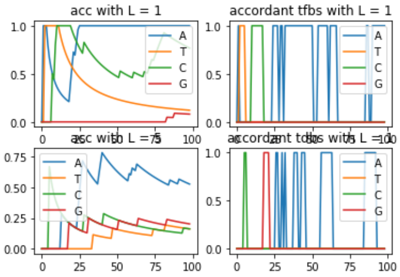

Figure 1: L = 1

To realize the detection of binding site of DNA transcrition factor, wIth homogeneous order 1 and 5 VOM(0.65,10), I first get the binding sites with the highest accuracy prediction peaks. X = A,T,C,G, and thus the d = 4. And using the mix model with pruning on KL scaled by threshold c on log scaled odds , the bayesian tree is for specifically to predict those at binding site(pruned level less than 3). For the foreground dataset, L = 1, and each DNA sequence from X is predicted directly through the one before it while using the most easiest critieria, chossing the one with largest probability. For the background dataset, L = 5, and each DNA sequence from X is predicted through the previous five sequences before it while using the frequency of the respective subsequencies combining Bayes Theory. And again, the highest peak frequency is chosen to be the biding site with its probability approximated: As the results, generally, the KL converges faster with L = 1 and the mix model of order 1 and 5. In addition, the model is accumulated instead of step-wise, leading to the convergence not rising with the iteration. It is obvious that, on some specific TFBSs, the prediction is even higher than others.

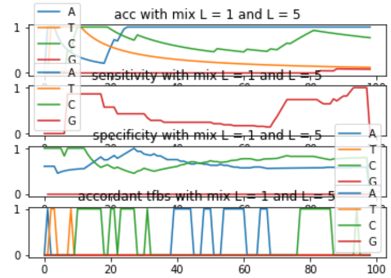

Figure 2: L = 5

$$\hat{P}\left(X_j = x_j \middle| X_{j-L}^{j-1} = x_{j-L}^{j-1}, \Omega_k\right) = \frac{n_k\left(x_{j-L}^{j}\right)}{n_k\left(x_{j-L}^{j-1}\right)},$$

Figure 3: frequency

$$P(x_1^N) = \prod_{j=1}^{N} P\left(X_j = x_j \middle| X_{j-L_j}^{j-1} = x_{j-L_j}^{j-1}\right),$$

Figure 4: approximation

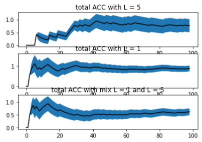

Figure 5: convergence

3 Stepwise Markov Ising model with lineage tree

The first step is to configure the epigenetic states of sequencies through their expression, whehter "close" of "open" on any temporary cell state. (In our data, there are 64 cell states in all.)

The second step is to check the final state of each lineage tree, as the expression of modules maybe conflicted with each other, influenced by other cell's transcriptional regulatory network which might lead to cell death(we denoted as 0) while ϵ denoted as the last state to the last states, which making 0 and n(66th) as the absorbing states.

The transition matrix is built according to reference [2]: Note that in the first step, The transition states predicted through the highest probability. And in the second step, either probability(0 or ϵ and further n) over 0.5 will lead to the end of the markov process. Basically, it is based on the Ising model, choosing direction among P1,P2,P3,P5 instead of the original random spin choise from -1,1 and the final fate of the cell is predicted through either 0 or n after ϵ, i.e. P4. Among the 200 simulations, first level cell expression takes fewest amounts while the highest convergence.(Although the variance to energy ratio are quite similar for the three levels.).

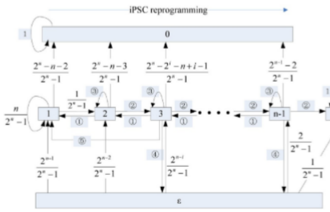

Figure 2 Reprogramming Markov Chain

(1) $P_{i,i-1}=\frac{2}{2^n-1}\times\frac{1}{2^i-3}$ $1<i<n$

(2) $P_{i,i+1}=\frac{1}{2^n-1}\times P_{i1}, P_{i1}=\begin{cases}\frac{1}{2^i-3} & 1<i<n\\ 1 & i=1\end{cases}$

(3) $P_{i,i}=\begin{cases}(1+n-i-1+\frac{1}{2^i-3})/(2^n-1) & 1<i<n\\ 1 & i=n\\ \frac{n}{2^n-1} & i=1\end{cases}$

(4) $P_{i,\varepsilon}=\frac{4}{2^n-1}\times\left(1-\frac{1}{2^i-3}\right)$ $2<i<n$

(5) $P_{ij}=\frac{2^{i-j}}{2^n-1}$ $0<j<i-1, 2<i<n$

Figure 6: Transition matrix

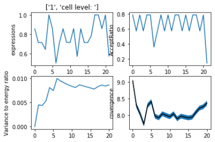

Figure 7: cell level 1

Generally, the Accept ratio is higher after the expression level goes to the lowest which is either when inhibition of DNA methylation exhists or the knowckdown of somatic transcription factors occurs.

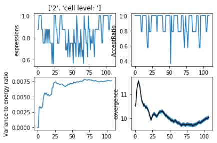

Figure 8: cell level 2

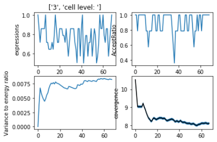

Figure 9: cell level 3

4 Disccussion

The transcriptional factor are quite useful in either identifying the transcription factor binding sites or the reprogramming of the cell. As for our first model, the approximation can be modified so that the model can be based on dynamic data, instead of using the last state only. On the other hand, for the second model, there can be improvements applied with the accept ratio checked on expression and DNA methylation both (as refered in the original article, it first check whether there is the methylation of histone and then the expression of the protein.)

References

[1]Novel Markov model of induced pluripotency predicts gene expression changes in reprogramming Zhirui Hu, Minping Qian, Michael Q Zhang*

[2]Identication of transcription factor binding sites with variable-order Bayesian networks, I. Ben-Gal,, A. Shani1, A. Gohr, J. Grau, S. Arviv1, A. Shmilovici, S. Posch, and I. Grosse

[3]Jaenisch Rudolf, Young Richard: Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 2008, 132:567-582. 2. Koche Richard P, Smith Zachary D, Adli Mazhar, Gu Hongcang