An Algorithm for Motif Discovery in Gene Sequence Study Based on Online AP Clustering

Wei Li[1, a], Chunxiao Sun[1, b, *], Jiayan Deng[1], Weiqin Bao[1] and Qing Zhang[1]

[1]College of Science, Northwest A&F University, Yangling, Shaanxi 712100, China.

[a]lxylw@nwuaf.edu.cn, [b]schxwky@nwuaf.edu.cn

Abstract

The extension of motif discovery to a genome-wide look by Chromatin immunoprecipitation combined with next-generation sequencing (ChIP-Seq) technology brings to the growth of data, which is hard to deal with using traditional algorithms. In this paper, we put forward OAP-Motif algorithm, an algorithm for motif discovery based on online Affinity propagation Clustering Algorithm to deal with the ChIP-Seq data set. Firstly, we divide the data set into a few blocks, and then use traditional and online AP clustering algorithm for analyzing each data block to get candidate motif set. Next we adopt expectation-maximization algorithm for refinement to obtain the motif. Finally, we verify the algorithm OAP-Motif algorithm via the ChIP-Seq data set. The results show that OAP-Motif algorithm is efficient in handling motif discovery in ChIP-Seq data set.

Keywords

CHIP-Set, AP clustering algorithm, OAP-Motif algorithm, clustering center dimensions.

1. INTRODUCTION

Motif discovery is an issue of importance and challenge in studying gene sequence [1]. It is extended to a genome-wide look by Chromatin immunoprecipitation combined with next-generation sequencing (ChIP-Seq) [2]. Compared with traditional algorithm for motif discovery, ChIP-Seq enjoys massive data, which hinders the processing of recognizing motif in the ChIP-Seq data set. Recently, some algorithms have been gradually brought up to cope with motif discovery in the ChIP-Seq data set [3-4], like MEME-ChIP [5], HMS [6] and FMotif [7]. All these algorithms can be used to handle part of discovery issues, but they are not always effective. MEME-ChIP, integrating MEME and DREME, two complementary algorithms for motif discovery, only works in part of sequences (e.g. No.600 input sequence) in the ChIP-Seq data set. HMS adopts random sampling and exhaustive search based on Gibbs sampling algorithm to improve its computational efficiency, but it only using those sequences with good performance when testing. FMotif, facing Big Data, is an exhaustive search algorithm of conservative mode based on suffix tree. It works well in terms of time in handling short motif, but FMotif needs a space of high complexity for mismatching information storage.

In this paper, owing to adopting the strategy of online algorithm, we put forward to OAP-Motif, a new algorithm for motif discovery, aiming to cope with ChIP-Seq data set, which contains thousands of sequences. First, we divide the whole data set into a few blocks, and then we use traditional and online AP clustering algorithm for analyzing each data block to get candidate motif set. Next we adopt expectation-maximization algorithm for refinement to obtain the motif. The results show that OAP-Motif algorithm is able to discover motifs efficiently in ChIP-Seq data set.

2. BACKGROUND

2.1. Definition

Definition 1: Given t DNA sequences with length of n defined on the set {A, C, G, T}, that is S={sl , s2 , ..., st } and two non-negative integers l and d, which meet the condition that 0≤d<l<n, motif discovery is to find a string x with length of l, which lets each sequence si contains a string xi having a position difference of d with x. The string x is called (l, d)-motif, and xi the motif instance of x.

Definition 2: A string s is a sequence defined on the set of {A, C, G, T}. |s| represents the length of the string s.

Definition 3: Given a string s with the length of n and another string x with the length of l, in which l<n. If x is a sub-string of s, x is an l-mer of s, written as $x \in_l s$.

2.2. Online AP Clustering Algorithm

AP clustering algorithm is an unsupervised Clustering algorithm put forward by Frey and others in 2007 [8]. By establishing the similarity matrix for targeted data set, we introduce two types of messages, responsibility and availability, in which, responsibility $r(i,k)$ refers to the transfer of information from data point i to candidate clustering center k, availability $a(i,k)$ refers to the transfer candidate clustering center k to data point i.

$$r(i,k) = s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\} \tag{1}$$

$$a(i,k) = \min\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\} \quad i \neq k \tag{2}$$

$$a(k,k) = \sum_{i' \neq k} \max\{0, r(i',k)\}$$

Each data in the initial targeted set iterates around responsibility and availability as clustering center until the convergence. The clustering result can be obtained via the following equation.

$$c_i = \arg\max_k \{a(i,k) + r(i,k)\} \tag{3}$$

With the growth of data, it takes increasing time to deal with the whole data set using traditional AP clustering algorithm, which is impracticable. Therefore, we put forward online AP clustering algorithm. By establishing the relationship of information transfer between the new data and existing data, we ensure that both of them are in the same level of information transfer to improve the performance of traditional AP clustering algorithm, meeting the requirements of dealing with big data.

Given that there are a similarity matrix $S_{(n-1)\times(n-1)}$, a responsibility matrix $R_{(n-1)\times(n-1)}$ and a availability matrix $A_{(n-1)\times(n-1)}$, we extend $R_{(n-1)\times(n-1)}$ and $A_{(n-1)\times(n-1)}$ respectively to $R_{n\times n}$ and $A_{n\times n}$ using online AP clustering algorithm, which is showed as follows.

$$r_n(i,j)=\begin{cases} r_{n-1}(i,j) & i\le n-1,\ j\le n-1 \\ r_{n-1}(i',j) & i>n-1,\ j\le n-1 \\ r_{n-1}(i,j') & i\le n-1,\ j>n-1 \\ 0 & i>n-1,\ j>n-1 \end{cases} \tag{4}$$

$$a_n(i,j)=\begin{cases} a_{n-1}(i,j) & i\le n-1,\ j\le n-1 \\ a_{n-1}(i',j) & i>n-1,\ j\le n-1 \\ a_{n-1}(i,j') & i\le n-1,\ j>n-1 \\ 0 & i>n-1,\ j>n-1 \end{cases} \tag{5}$$

Where $i'=\arg\max_{i\le n}\{s(i,i')\}$ and $j'=\arg\max_{j\le n}\{s(j',j)\}$.

2.3. Online AP Clustering Algorithm

Input: $S_{(n-1)\times(n-1)}$, $R_{(n-1)\times(n-1)}$, $A_{(n-1)\times(n-1)}$

Output: c_i

(1) Compute that similarity matrix.

(2) Obtain a new responsibility matrix R and a new availability matrix A from the equation (4) and (5).

(3) Update responsibility matrix R and availability matrix A according to equation (1) and (2).

(4) Conduct the updating until the convergence, and obtain the clustering result via equation.

3. ALGORITHM

By using traditional algorithm for motif discovery, we update unknown parameter via iteration with a complete data set. Under this case, all data should be given before the input. But it becomes harder to directly deal with the whole data set along with the growth of high-throughput sequencing data. Also, it will affect the accuracy when only handling one sampling data set. In this paper, we design a OAP-Motif algorithm, an online algorithm for motif discovery based on online AP clustering algorithm. First, we obtain input block B_{input} with the size of t from the input sequences S, and we divide each input block into B_{data} with the size of t' and compute their clustering sub-set. Next we obtain the candidate motif set by using traditional AP clustering algorithm and online AP clustering algorithm respectively, which is called blocked solution. And then we combine the results of each block solution and use expectation-maximization algorithm to deal with them. We keep the computation of clustering sub-set and the following procedure to handle all of the sequences in S. More details are as follows.

3.1. Obtain the Input Block and Data Block

We obtain a few B_{input} with the size of t according to the order from S, a given input sequences set. And then we divide each B_{input} into several B_{data} with the size of t'. An oversized B_{input} will impede the computation of initial motif, and an undersized B_{input} will inevitably depend on new data block. The size of B_{data} should be set according to the processing time of algorithm. In this paper, we set t as 3000 and t' as 600, which means each input block will be divided into 5 data blocks.

3.2. Compute Clustering Sub-Set of Each B_{data}

We build the clustering sub-set consisting of substring with high similarity. The establishment of clustering sub-set is based on the observing fact that the Hamming distance of two instances of the same motif is smaller than or equal to 2d and we need to set a reasonable threshold value k and make sure that $0 \le k \le 2d$. In consideration of general conditions, we choose the first sequence X1 to be the reference sequence, and all $l-mer\ x_k\ (k=1,2,\mathrm{L}\ ,t)$ to be the reference sub-sequence.

Let $B(x_k, X_i) = \{y : y \in_l X_i, d_H(x_k, y) \le k\}$ represents the candidate motif set of x_k in $X_i(i=2,...,t)$.

Let $C(x_k, X) = \{x_k\} \cup \bigcup_{i=2}^{n-l+1} B(x_k, X_i)$ represents the set of all the sequences of which the Hamming distance with x_k is smaller than or equal to k and $C(x_k, B_{data})$ is the clustering sub-set of B_{data} .

3.3. Blocked Solution

We conduct the clustering analysis for each clustering sub-set of B_{data} , taking into consideration historical information of and the effects of new data on estimated parameters. Therefore we put forward the blocked solution. We use traditional AP clustering algorithm and online AP clustering algorithm to deal with clustering sub-set of B_{data} , which are respectively called closed-form solution and online solution. And then we obtain the candidate motif set of B_{data} , $C_{candidate}(x_k, B_{data})$.

3.4. Cluster Refinement Via EM Algorithm

Traditionally, we obtain different local optimal solutions via convergence because of the uncertainty of initial condition. In this paper, we choose $C_{candidate}(x_k, B_{data})$ as the initial condition of each B_{data} , and ensure that we obtain one local optimal solution. More details are as follows. Estimate hidden variables of each l-mer

$$Z_{i,j}^{(T)} = \frac{p(X_i \mid z_{i,j} = 1, \theta^{(T)})}{\sum_{j=1}^{n-l+1} p(X_i \mid z_{i,j} = 1, \theta^{(T)})} \tag{6}$$

Re-estimate the value of pω,m

$$p_{w,m}^{(T)} = \frac{c_{w,m} + \xi_m}{\sum_{w\in\Omega} (c_{w,m} + \xi_m)}, \quad \omega \in \Omega = \{\mathrm{A}, T, C, G\} \tag{7}$$

$$c_{w,m} = \sum_{i=1}^{t} \sum_{j=1,}^{n-l+1} z_{i,j}^{T} I(i, j+k-1) \tag{8}$$

$$c_{w,0} = c_w - \sum_{m=1}^{l} c_{w,m} \tag{9}$$

In these there equations, let $\xi_m = 1$ to avoid the emergence of zero probability.

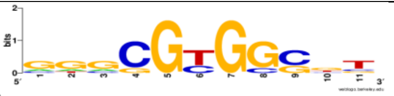
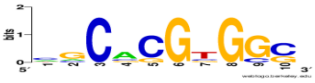
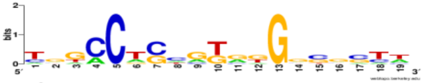
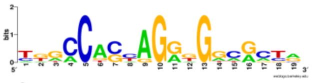
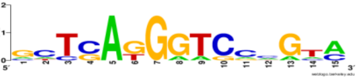
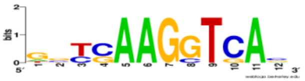
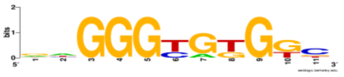
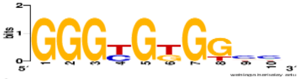
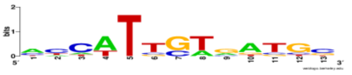
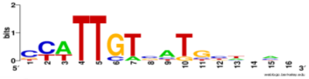
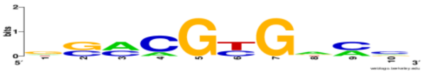
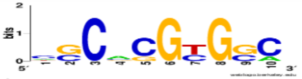
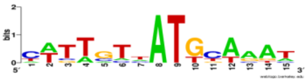
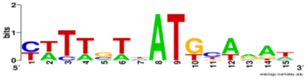
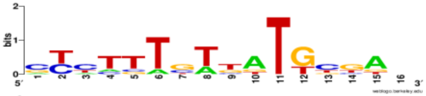
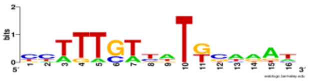
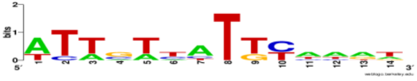
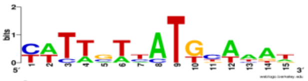
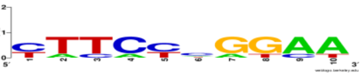
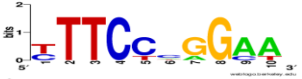
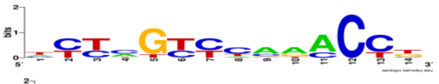
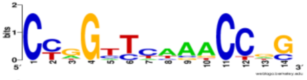
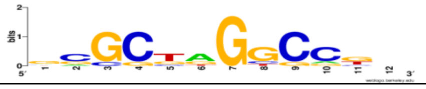
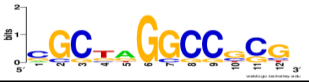
Through Cluster refinement via EM Algorithm, we obtain θ_k, the base distribution of each $C_{candidate}(x_k, B_{data})$ and related target function Q_k. And then we find the maximum value of θ_k, written as θ_{max}. Next we compute the Log Likelihood of all l-mer in each sequence, in which the maximum one is the instance of candidate motif.

$$\log p\left(x_k \mid \theta_{\max}\right) = \max \sum_{m=1}^{l} \log p_{w,m} \tag{10}$$

4. THE RESULT AND ANALYSIS

To verify OAP-Motif in this paper, we choose 12 groups of mES cells of mouse, Nanog, Oct4, Sox2,Esrrb,Zfx,Klf4,c-Myc,n-Myc,Tcfcp21l,Smad1,STAT3and CTCF to conduct an experiment[9]. The Matlab code is used to implement the OAP-Motif algorithm under the Windows system environment. The test environment is 2.67 Hz CPU and 4 GB memory.

Table 1. The testing result of mES data

Data set(seq#)	Predicted motif	Literature
c-Myc(3422)		
CTCF(39609)		
Esrrb(21647)		
Klf4(10875)		
Nanog(10343)		
n-Myc(7182)		
Oct4(3761)		
Smad1(1126)		
Sox2(4525)		
STAT3(2546)		
Tcfcp211(26910)		
Zfx(10338)		

In order to better show the performance of the algorithm, we compare the motif published by the WEEDER algorithm used by Chen et al. with the method in this paper [10]. In order to grasp the similarity between the prediction motif and the published data, a method of expressing statistical significance as a LOGO chart is adopted [11], which is also a widely used comparison method now. The picture shows the comparison between the main motif found by the algorithm and Chen's published motif. It can be seen that the algorithm in this paper can effectively find the main motif in these real data.

5. CONCLUSION

In this paper, we propose an algorithm for motif discovery, OAP-Motif, on large-scale ChIP-Seq data, which can solve high-throughput data sets by combining traditional AP clustering algorithms with online AP clustering strategies. In the experiment using ChIP-Seq data from mouse embryonic stem cells, we prove that the OAP-Motif algorithm can find motifs in large-scale ChIP-Seq sequences. However, these motifs often have different functions in gene sequences, and their functionality needs to be verified in further experiments.

ACKNOWLEDGEMENTS

This paper was supported by The College Students' Innovative Entrepreneurial Training Plan Program of NorthwestA&F University (No. S201910712133).

REFERENCES

[1] Fu, Wenhao; Tang, Linlin; Wei, Gaohu. “Rational Design of pH-Responsive DNA Motifs with General Sequence Compatibility”. ANGEWANDTE CHEMIE INTER- NATIONAL EDITION, 2019, Vol. 46 (58), p16405- 16410.

[2] MARDIS E R. “ChIP-seq: welcome to the new frontier”. Natmethods,2007, Vol.4(8): p613-617.

[3] Wong, K.C. MotifHyades: “Expectation maximization for de novo DNA motif pair discovery on paired sequences”. Bioinformatics, 2017, 33, p3028–3035.

[4] Yu, Q., Wei, D., Huo, H. SamSelect: “A sample sequence selection algorithm for quorum planted motif searchon large DNA datasets”. BMC Bioinform, 2018, 19, 228.

[5] Machanick, P., Bailey, T.L. MEME-ChIP: “Motif analysis of large DNA datasets”. Bioinformatics, 2011,27, p1696–1697.

[6] Hu, M., Yu, J.; Taylor, J.M., Chinnaiyan, A.M., Qin, Z.S. “On the detection and refinement of transcription factor binding sites using ChIP-Seq data”. Nucleic Acids Res, 2010, 38, p2154–2167.

[7] Jia, C., Carson, M.B., Wang, Y., Lin, Y., Lu, H. “A new exhaustive method and strategy for finding motifs inChIP-enriched regions”. PLoS ONE, 2014, 9, p86-94.

[8] Frey, B.J., Dueck, D. “Clustering by passing messages between data points”. Science, 2007, 315, 972–976.

[9] YU Q. “An efficient algorithm for discovering motifs in large DNA data sets”. IEEE transactions on nanobioscience, 2015, Vol.14(5),p535-44.

[10] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. “Integration of external signaling pathways with the core transcriptional network in embryonic stem cells”. Cell, 2008, 133, p1106–1117

[11] Li, Wentian; Thanos, Dimitrios; Provata, Astero. “Quantifying local randomness in human DNA and RNA sequences using Erdos motifs”, Genome research, 2019, 461, p41-50.