A Novel Pipeline for Motif Discovery, Pruning and Validation in Promoter Sequences of Human Tissue Specific Genes

Xiu-Jun GONG, Hua YU
School of Computer Science and Technology
Tianjin University
Tianjin, China
gongxj@tju.edu.cn, yuhua@tju.edu.cn

Fei-Fei ZHAO
State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Science
Beijing, China
zhaofeifei09@gmail.com

***Abstract*—Identification and analysis of tissue-specific (TS) genes and their regulatory activities play an important role in the understanding of mechanisms of organisms, disease diagnosis and drug design. In this paper, we designed a pipeline for the discovery of promoter motifs for tissue-specific genes. The pipeline consists of three phases: motif searching, motif merging and motif validation. The motif searching phase integrated three algorithms: MEME, AlignACE and Gibbs Sampling. In the second phase, we proposed a motif merging method, which is based on Bayesian probabilistic principles, to reduce redundancies of motifs from the first phase. Lastly, the motif validation phase verified the statistical significance of discovered motifs using a Bayesian Hypothesis Test approach. We performed the analysis on the sequences of promoter regions (-449bp-1000bp) of 4,552 human tissue-specific genes across 82 tissues and 924 housekeeping genes. The distributions of motifs in different promoter regions show that most motifs prefer to be in the proximal region (+500~50bp, -50bp~-500bp) of promoters.**

Keywords- tissue specificity; motif discovery; Bayesian; hypothesis test; promtoer sequences.

I. INTRODUCTION

Tissue specificity is the foundation for cells to form specific tissues and functional organs [1]. The identification and analysis of tissue-specific (TS) genes and their regulatory activities play an important role in understanding the mechanisms of organism, disease diagnosis and drug design [2]. However, it remains a challenging question to understand the mechanisms underlying regulation of tissue-specific gene expression. Despite it is not clear entirely, alternative splice, epigenetic characteristics (DNA methylation [3] and histone binding sites [4]) and the characterization of gene promoter regions [5] are becoming important clues for inferring the inner mechanisms of tissue specificity.

TATA-box and CPG-island are the most relevant promoter sequence patterns. By applying statistical length sequence distribution method on human genes' promoter sequence and P-value tests to verify the significance, a batch of 8-base long promoter significant patterns is obtain in [6]. Lawson et al. [7] studied the statistics of the distribution of single sequence repeats (SSRs) frequency and uses the chi-square test method to verify the significance of discovered patterns. It discovered some meaningful SSRs that have certain effect on gene expression and tissue specificity.

Discovering motifs in DNA sequences is a key problem in computational biology that has been addressed by multiple algorithms, including MEME [8], AlignACE [9, 10], MDScan [11] and Gibbs Sampling [12]. Each algorithm has its unique advantage on individual species or datasets. Tompa et al [13] conducted a study that compares the performance of 13 different motif finders by using a variety of real and synthetic sequence sets covering a range of genomes. The study showed that no single motif finder consistently outperforms others. Moreover, the results indicated that a pairwise combination of motif finders can result in improvement over the use of a single motif finder, although the choice of motif finders is important. A common practice is to apply several such algorithms simultaneously to improve coverage at the cost of increased redundancy [14].

In this paper, we present a novel motif discovery pipeline that benefits from combining existing motifs discovering algorithms and at the same time overcomes the problems of redundancy. In this pipeline, we (1) integrate three motif discovering algorithms to generate motif candidates, (2) prune abundant motifs output based on a similarity function, and (3) verify the statistical significance of discovered motifs with a Bayesian Hypothesis Test approach.

II. METHODS

The method consisted of four procedures: data preparation, motif searching, motif merging, and motif validation, see figure 1 for more details.

A. Data preparation

The gene expression datasets, such as GNF, SAGE, and EST, are very widely used as data sources for classifying HK and TS genes. For a given gene, the identification of whether it is HK and TS is usually done by applying predefined thresholds of its expression levels. However, because of the noise in expression datasets and human involvement in defining thresholds, the reliability of the identifications is often not high. In this paper, we attempted to rectify this problem by obtaining lists of the two classes (HK and TS) of genes from two sources: scientific publications from PubMed and known tissue-specific

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

databases such as TisGeD [15] and TiGER [16]. We obtained 4,552 human tissue-specific genes across 82 human tissues and 924 human housekeeping genes. The gene's promoter sequences were downloaded from DBTSS [17] and EPD [18]. The promoter region with length 1500bp (-499bp-1000bp around TSS) is used for motif discovery.

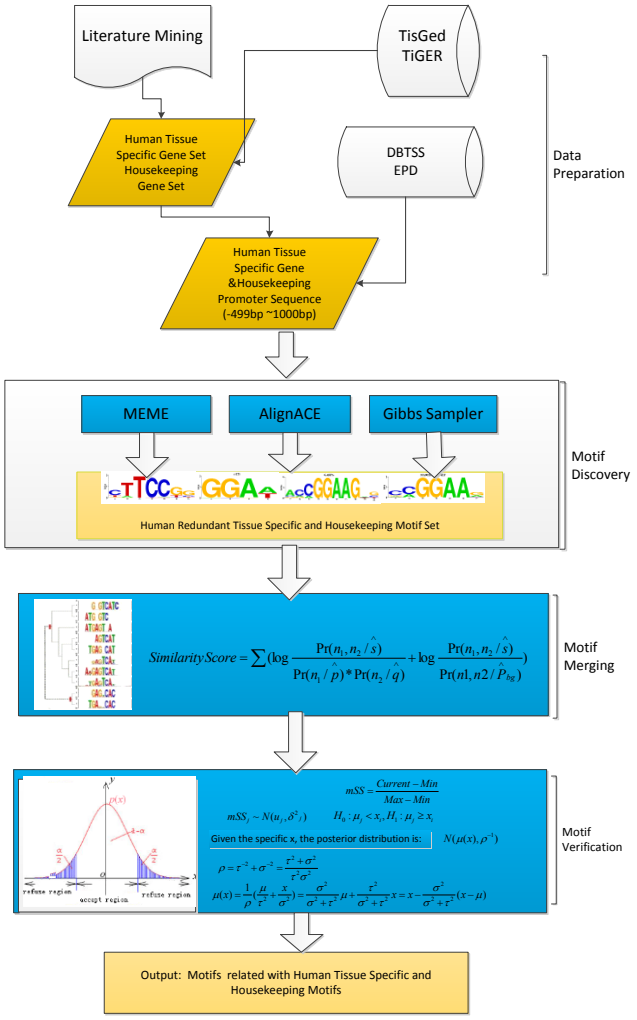


Figure 1. Figure 1. the motif discovery pipeline

B. *Motif searching*

As stated in the previous section, different motif finding algorithms have different characteristics. In this work, we integrated three commonly used motif finding programs: MEME, alignACE and Gibbs Sampling. The length of candidate motifs is fixed to 8-12bp while other parameters are left at their default values. Collecting all results from the three programs, we obtained 6,794 motifs in total. However, there are overlaps among the motifs produced by the three programs. Such redundancy needs to be further pruned.

C. *Motif merging*

1) *Motif Representation*

The outputs of the three programs have different motif representation formals. To compare their similarities in merging motifs, we need to convert all the motif representations into a uniform one. A simple yet flexible representation of motif is the position matrix (equation 1), in which each row represents one residue (A, C, G, T), and each column represents the score (weight) that the aligned residue occurs in the position.

$$PWM_m = \begin{pmatrix} & p_1 & p_2 & p_3 & p_4 \cdots \\ A: & w_{A1} & w_{A2} & w_{A3} & w_{A4} \cdots \\ T: & w_{T1} & w_{T2} & w_{T3} & w_{T4} \cdots \\ G: & w_{G1} & w_{G2} & w_{G3} & w_{G4} \cdots \\ C: & w_{C1} & w_{C2} & w_{C3} & w_{C4} \cdots \end{pmatrix} \quad (1)$$

Where m is the length of a motif, w_{ij} is the weight that a residue i is aligned to position j.

There are several matrix representations, which differ in the weighting schemes. However, all matrix representations assume that the choice of nucleotides in each position of a motif is independent of all other positions. A common weighting scheme is to use frequencies. This matrix is also called Position-specific Weight Matrix (PWM).

2) *Motif Similarity Comparison*

A key problem in the motif merging phase is to compare the similarity between two PWMs. To compare two PWMs, we can utilize the position-independence assumption to decompose the similarity score of two motifs into the sum of similarities score of single aligned positions (a column vector with 4 rows). Assumed that there are two motifs m_1 and m_2 with PWM PWM_{1m} and PWM_{2m}, the similarity score can be calculated according to formula (2).

$$SimilarityScore_{12} = \sum_{\substack{1\le i\le m1 \\ 1\le j\le m2}} similarity(Vector_{1i}, Vector_{2j}) \quad (2)$$

Where $Vector_{ij}$ denotes the jth column of the PWM of motif i.

Two motifs may be of different lengths or reverse complement with each other. Hence all possible alignments should be considered. The similarity score between two motifs is the highest score of all possible alignments of the motifs.

3) *Similarity Score*

Many methods have been proposed to compare the similarity between the two vectors. Pearson correlation coefficient, which is based on statistical measures and Euclidean distance, which measures the distance between geometrical metrics, are two widely used methods. However, according to [19], all of them do not deal with the following two situations: informative columns and non-informative columns. A pair of informative columns means they have equally weighted positions with similar nucleotide distributions that are specific (e.g. a strong preference for an A). A pair of non-informative columns means although they have similar position distributions, they are not specific, i.e., they do not have strong preference for any one of the four nucleotides (e.g. identical to the background distribution). We used the method proposed in [19] to differentiate between the two situations. Such a distinction is desirable because the two positions whose similarity is due to a resemblance to the background distribution are less relevant to motif similarity. This method takes into consideration the DNA se-

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

quence similarity between two motifs, and also considers them to be different from the background distribution.

$$SimilarityScore = \sum \left(\begin{array}{l} \log \frac{\Pr(n_1, n_2 / \hat{s})}{\Pr(n_1 / \hat{p}) * \Pr(n_2 / \hat{q})} + \\ \log \frac{\Pr(n_1, n_2 / \hat{s})}{\Pr(n1, n2 / \hat{P}_{bg})} \end{array} \right) = \sum \left(\begin{array}{l} \log \frac{\Pr(n_1 / \hat{s}) * \Pr(n_2 / \hat{s})}{\Pr(n_1 / \hat{p}) * \Pr(n_2 / \hat{q})} + \\ \log \frac{\Pr(n_1 / \hat{s}) * \Pr(n_2 / \hat{s})}{\Pr(n_1 / \hat{bg}) * \Pr(n_2 / \hat{bg})} \end{array} \right) \quad (3)$$

where n_1 and n_2 are the corresponding alignment positions for two motifs. $\hat{p}$, $\hat{q}$ and $\hat{s}$ are the estimators for the source distribution of n_1, n_2 and the common source distribution respectively. Finally $\hat{P}_{bg}$ is the background nucleotide distribution.

The formula (3) has two components: the first part measures whether the two motifs are generated from a common distribution, and the second part measures the distance of that common distribution from the background.

We assumed that positions of pattern matrix are independent of each other. So the similarity score can be decomposed into a sum of local position scores that examine only the distribution of nucleotides at one position in both motifs. This problem then turns into one that computes the similarity score of two vectors with length of four (A, T, C, and G).

The probability $\Pr(n_1 / \hat{t}) \hat{t} \in \{\hat{s}, \hat{p}, \hat{q}, \hat{bg}\}$ can be calculated using equation (4).

$$\Pr(n \mid \hat{t}) = \Pr(x_A, x_T, x_C, x_G) = \frac{n!}{x_A! x_T! x_C! x_G!} p_A^{x_A} p_T^{x_T} . p_C^{x_C} p_G^{x_G} \quad (4)$$

where $[x_A, x_T, x_C, x_G]$ refers to the occurrence frequency of the nucleotides [A, T, C, G], $n = x_A + x_T + x_C + x_G$ and $[P_A, P_T, P_C, P_G]$ refers to the occurrence probability of the corresponding nucleotides, which can be calculated according to the Dirichlet distributions.

We used a standard Dirichlet prior $a = (a_1, a_2, ..., a_n)$ to estimated distribution for the position *n*, it can be calculated using formula (5):

$$\hat{p}_i = \frac{n_i + a_i}{\sum_{j \in \{A,C,G,T\}} (n_j + a_j)} (i \in \{A, C, G, T\}) \quad (5)$$

We need to estimate the independent distribution and common source distribution for each position. For estimating the common source distribution, we used a five-component mixture of Dirichlet prior, merging five standard Dirichlet priors using uniform weights. The four components, which represent uni-nucleotide distributions, give high probabilities for a single DNA nucleotide: {0.7, 0.1, 0.1, 0.1} for residue A, {0.1, 0.7, 0.1, 0.1} for residue T, {0.1, 0.1, 0.7, 0.1} for residue C and {0.1, 0.1, 0.1, 0.7} for residue G. We also use {0.25, 0.25, 0.25, 0.25} as the background distribution. Therefore, the common .source distribution is given in formula (6).

$$\hat{p}_i = \sum_{k=1}^{5} (\Pr(a^k / n) \frac{n_i + a_i^k}{\sum_{j \in \{A,C,G,T\}} (n_j + a_j^k)}) \quad (6)$$

Where $\Pr(a^k / n)$ can be calculated using formula (7).

$$\Pr(a^k / n) = \frac{\Pr(n / a^k)}{\sum_j \Pr(n / a^j)} \quad (7)$$

D. Motif Validation

We use the Bayesian Hypothesis Test method to verify the statistical significance of discovered motifs. Given the motifs from tissue i, the sample mean x_i of tissue i can be calculated. To verify whether a motif is significant at tissue $j (j \neq i)$ or not, we first defined the matching score [20] between a motif and a sub-sequence, it can be calculated using formula (8) [21].

$$mSS = \frac{Current - Min}{Max - Min} \quad (8)$$

Where $Current: \sum_{i=1}^{L} I(i) f_{i,B}$, $Min: \sum_{i=1}^{L} I(i) f_i^{\min}$, and $Max: \sum_{i=1}^{L} I(i) f_i^{\max}$. $f_{i,B}$ is the frequency of residue B at position i; similarly, $f_i^{\min}$ is the smallest frequency of residue at position i and $f_i^{\max}$ is the largest frequency of residue at position i. $I(i) = \sum_{B \in \{A,T,G,C\}} f_{i,B} \ln(4 f_{i,B})$ describes the information content of residue B at position i.

It has been shown in [22] that the matching score satisfies Gaussian distribution. We assume that different matching scores are independent from each other. Hence, we assumed that the matching score between motif and tissue-specific genes follows a Gaussian distribution, i.e. $S_{ij} \sim N(\mu_i, \sigma_i^2)$.

We obtain the estimator of mean and variance through the *Moment Method*, defined in formula (10).

$$\mu = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \sigma^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad (10)$$

As stated above, the matching score of a motif in tissue j follows a Gaussian distribution $N(\mu_j, \sigma_j^2)$ [22], where σ_j^2 is known, which can be replaced by the sample variance. To verify whether this motif is significant at tissue $j (j \neq i)$, we construct two hypotheses:

$$H_0 : \mu_j < x_i, H_1 : \mu_j \geq x_i \quad (11)$$

such that, if we accept H_0, it shows that this motif is not significant at tissue j; otherwise this motif is significant at tissue j.

Assumed that $X \sim N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known, $\pi(\theta) \sim N(\mu, \tau^2)$, where both μ and τ^2 are known. We obtain that the post distribution of θ follows $N(\mu(x), \rho^{-1})$ [19]

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

by giving sample x , and the corresponding parameters are calculated as below:

$$\rho = \tau^{-2} + \sigma^{-2} = \frac{\tau^2 + \sigma^2}{\tau^2 \sigma^2} \quad (12)$$

$$\mu(x) = \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)$$

$$= \frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2}x \quad (13)$$

$$= x - \frac{\sigma^2}{\sigma^2 + \tau^2}(x - \mu)$$

For discovered motifs, we also calculate their occurrence positions on the promoter region, we first divide the promoter sequence (-499bp ~ 1000bp) of a gene into three regions: core-promoter region (-50~+50bp around TSS), proximal-promoter region (+500~ -500bp around TSS), and distal-promoter region (+1000~ +500bp around TSS). For a discovered motif, we already know which tissue it comes from. We use this motif to scan the promoter sequence of these genes which are specific to this tissue, then compute the similarity score (mSS) which has been introduced in the Motif Tissue Specificity Verification section. We use the maximal score as the final score for the comparison. We ignore the gene if the maximal score is smaller than the similarity threshold. If the maximal score is bigger than or equal to a threshold (such as 0.85), this position is recorded. For an example, if motif1 is found in liver tissue with 100 tissue-specific genes. By scanning these 100 genes' promoter sequences, 70 genes with maximal similarity scores more than threshold are obtained. Then we count the number of these positions in different promoter regions. For instances, there are 50 genes in the core promoter region, 10 genes in the proximal promoter region and 10 genes in distal promoter region.

III. RESULTS AND DISCUSSION

A. *The number of discovered motifs from tissue specific genes*

In this experiment, by setting 1.0 as the similarity score for motif merging threshold, we found a total of 3,233 motifs after the merging phase. We designated a candidate motif to be tissue-specific if it occurs in less than 3 tissues. After filtering with Bayesian Hypothesis Test, 877 tissue specific motifs were discovered. The numbers of candidate tissue specific motifs listed in figure 2.

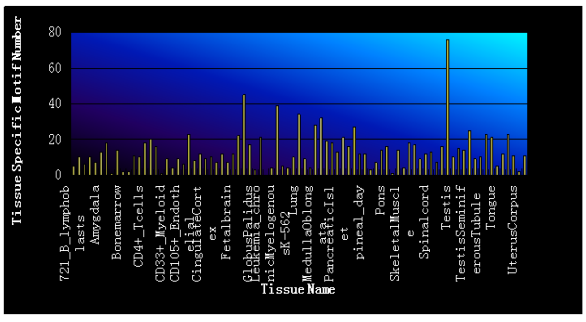


Figure 2. the distribution of the numbers of motifs from 82 human tissue genes

We also count the number of motifs that are present in 1-82 tissues significantly. The result is listed in figure 3. From the figure we can see that the number of motifs is higher on the two extremes and lower in the middle. It shows that most motifs are significantly specific in less than 4 tissues or more than 70 tissues. Put more precisely, we define motifs that are specific significantly in less than 3 tissues as candidate tissue specific motifs, and motifs that are specific significantly in more than 70 tissues as the candidate housekeeping motifs.

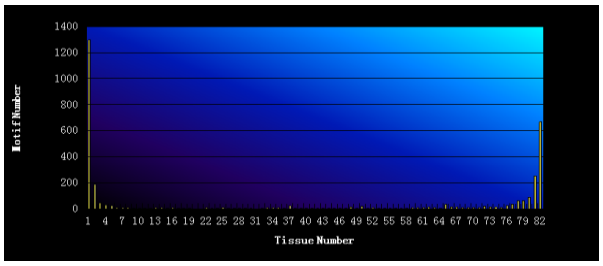


Figure 3. Number of motifs which are specific significantly in tissues (numbered 1-82) respectively

B. *The distributions of motif occurrence positions*

Candidate housekeeping motifs refer to those motifs that are specific significantly in more than 70 tissues. The number percentage distribution of candidate motifs on promoter region is showed in figure 6. We found that 71% of these candidate housekeeping motifs come from proximal promoter regions, and the density (motifs number/promoter region length) is 0.8, while only 29% come from distal promoter regions, and the density is 0.57. Hence, it can be seen that these candidate housekeeping motifs mostly come from proximal promoter regions.

We also calculated the number distribution of candidate tissue-specific motifs on promoter regions. The result is shown in figure 3, which is very similar to the distribution of candidate housekeeping motifs number.

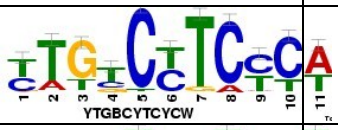
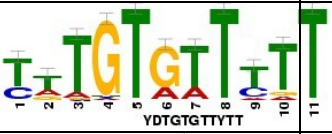
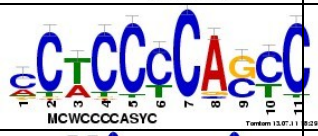
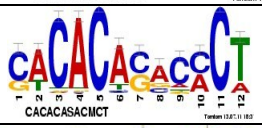
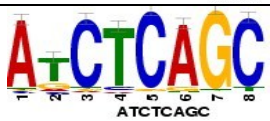
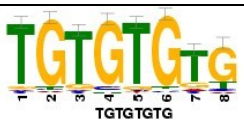
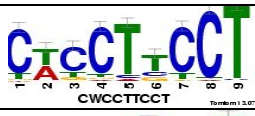
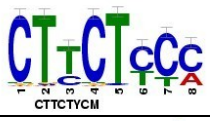
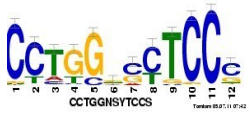
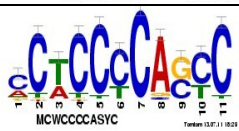
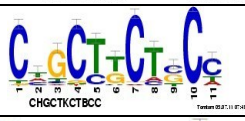
C. *Motifs from brain and colon tissue*

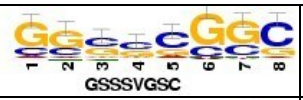
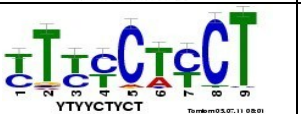
We discovered motifs associating with a total of 82 human tissues. In this subsection, we use motifs discovered in whole brain tissue and colon tissue to demonstrate the results. Other tissues' specific motifs can be found in the appendix at the end of the paper. In those tables, FIT denotes that the motif is found in this tissue, and the posterior probability comprises by two parts, according to the validation method in Motif Discovery section. We used Bayesian Hypothesis Test methods to verify the significance of discovered motifs. In this method, we construct two hypotheses: H0: that a motif is not significant in tissue j; and H1: that a motif is significant in tissue j. The former part of posterior probability means the probability to accept H0 while the latter part of posterior probability means the probability to accept H1.

In Table 1, motifs found in colon tissue are [ATCTCAGC] with the TRANSFAC [23] ID M00361 and [HWTTT] with JASPAR [24]ID MA0398.1, all of which are already known motifs and they occur significantly in colon tissue. We transform the motif matrix to log expression using LocoMotif [14].

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

TABLE I. MOTIFS DISCOVERED FROM HUMAN COLON AND WHOLE BRAIN SPECIFIC GENES

Motif Logo	Other Tissue:{Posterior Probability}
	FIT
	FIT
	FIT
	FIT
	FIT
	TemporalLobe: {0.382436536595506 0.6222119365118208 } Pancreas :{0.39785002108694 0.6073644925584932 }
	Leukemia_promyelocytic-HL-60 : {0.02973306868422576 0.9713917753133016 }
	FIT
	MedullaOblongata : {0.03649960590909652 ,0.9666341358036432}
	FIT
	FIT
	FIT
	FIT

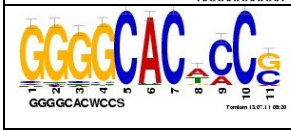
	FIT
	Fe-tallung:{0.397114414572689, 0.6043014481229988} MedullaOblonga-ta:{0.2468135415995074, 0.7551477782544601}

(In this table, the first 8 motifs are from colon tissue and the remained are from the whole brain tissue)

D. Motifs from Housekeeping genes

By applying this pipeline on promoter region of 924 human housekeeping genes, we discovered 3 motifs, see Table 4. Among the three motifs [TTTNTT] is an existing motif which was discovered by previous research, with JASPAR ID MA0049.1 and TRANSFAC ID M00022. Other two motifs are new candidate motifs never reported in literatures. This finding needs to be further verified.

TABLE II. MOTIFS DISCOVERED IN HUMAN HOUSEKEEPING GENES

Motif Logo	Motif Length
	6
	11
	11

IV. CONCLUSIONS AND FURTHER WORKS

Identification and analysis of tissue-specific (TS) genes and their regulatory activities play an important role in the understanding of the mechanisms of organisms, disease diagnosis and drug design. However, Understanding the mechanisms underlying regulation of tissue-specific gene expression remains a challenging question. In this study, we present a novel motif discovery, pruning and verification pipeline that integrates, merge and statistically verify the results of three motif discovering algorithms. Our experiment results demonstrate the advantage of integrating existing motifs discovery algorithms, the usefulness of merging redundant motif sets and verification of the statistical significance of a motif in different tissues using Bayesian Hypothesis Test methods.

In future, we plan to improve this pipeline from the following aspects. First, more motifs finding algorithms can be used to produce candidate motifs for comprehensive results. Second, we will investigate the best combination of motif finding algorithms. Third, we will develop better merging algorithms through clustering of motifs. Candidate motifs are firstly clustered. Motifs in a specific cluster are then merged to form new motifs, which express common characteristics of the motifs in that cluster maximally. Fourth, from the perspective of under-

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

standing the inner mechanism of tissue-specific genes regulation, we plan to combine motifs obtained in our pipeline with known tissue-specific regulatory elements, including enhancer, silencer and TFs, to construct tissue-specific gene regulation networks. Lastly, we will construct a comprehensive and continuously curated database to store all data relevant to tissue specificity. Such a database will serve as a central repository that provides vital tissue specificity information to other researchers.

ACKNOWLEDGEMENT

This research is partly supported by the Natural Science Funding of China under grand number 61170177 and innovation funding of Tianjin University.

REFERENCES

[1] 1. Dezso Z, Nikolsky Y, Sviridov E, et al. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC biology* 2008, 6.

[2] 2. Zhu J, He F, Hu S, Yu J: **On the nature of human housekeeping genes**. *Trends in genetics* 2007.

[3] 3. Schilling E: Analysis of tissue-specific & allele-specific DNA methylation. *Strain* 2010.

[4] 4. Hebert C: Nucleosome rotational setting is associated with transcriptional regulation in promoters of tissue-specific human genes. *Genome Biology* 2010, 11.

[5] 5. Schug J, Schuller W, Kappen C: Promoter features related to tissue specificity as measured by Shannon entropy. *Genome* 2005, 4.

[6] 6. Fitzgerald PC, Shlyakhtenko A, Mir AA, Vinson C: Clustering of DNA Sequences in Human Promoters Clustering of DNA Sequences in Human Promoters. *Genome Research* 2004:1562-1574.

[7] 7. Lawson MJ, Zhang L: Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5 ′ -UTR region. *Gene* 2008, 407:54 - 62.

[8] 8. Bailey TL, Boden M, Buske FA, et al. **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Research* 2009, **37**:W202-W208.

[9] 9. Hughes JD Tavazoie S EPW: Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000, 296:1205-1214.

[10] 10. Roth FP Estep PW, Church GM HJD: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 1998, 16:939-945.

[11] 11. Liu X Liu J BD: An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* 2002, 20:835-839.

[12] 12. Neuwald a F, Liu JS, Lawrence CE: **Gibbs motif sampling Detection of bacterial outer membrane protein repeats**. *Protein science : a publication of the Protein Society* 1995, **4**:1618-1632.

[13] 13. Tompa M, Li N, Bailey TL, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* 2005, 23:137-144.

[14] 14. Clements M: Manual Creating motifs with LocoMotif. *Scanning* 2007.

[15] 15. Shen-Jian Xiao Quan Zou,Zhi-Liang Ji CZ: **TiSGeD:a database for tissue-specific genes**. *Bioinformatics* 2010, **26**:1273-1275.

[16] 16. Xiong Liu Donald J Zack,Heng Zhu and Jiang Qian XY: **TiGER:A database for tissue-specific gene expression and regulaiton**. *BMC Bioinformatics* 2008, **9**:271.

[17] 17. Suzuki Y, Yamashita R, Sugano S, Nakai K: **DBTSS, DataBase of Transcriptional Start Sitesprogress report 2004**. *Nucleic acids research* 2004, **32**:D78-81.

[18] 18. Schmid CD, Praz V, Delorenzi M, Périer R, Bucher P: **The Eukaryotic Promoter Database EPD**. *Nucleic acids research* 2004, **32**:D82-5.

[19] 19. Naomi Habib Hanah Margalit,Nir Friedman TK: **A Novel Bayesian DNA Motif Comparison Method for Clustering and Retrieval**. *PLoS Compu Biol* 2008, **4(2)**:e1000010.

[20] 20. MacIsaac KD, Fraenkel E: **Practical Strategies for Discovering Regulatory DNA Sequence Motifs**. *PLoS computational biology* 2006, **2**:e36.

[21] 21. Kel AE: MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research* 2003, 31:3576-3579.

[22] 22. Bailey TL, Gribskov M: **score distribution for simultaneous matching to multiple motifs**. *Journal of computational biology : a journal of computational molecular cell biology* 1997, **4**:45-59.

[23] 23. E.Wingender R.Hehl,H.Karas,I.Liebich,V.Matys XC: **TRANSFAC:an integrated system for gene expression regulation**. *Nucleic Acids Research* 2000, **28**:316-319.

[24] 24. Sandelin A EngstrÖm P, Wasserman WW, Lenhard B AW: JASPAR:an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* 2004, 32(Database issue):D91-4.

V. ADDITIONAL FILES

1) Additional file 1 – motifs discovered in 82 human tissues which have significant occurrence in less than 3 tissues

File name is "CandidateTissueSpecificMotifs.txt", and each motif is described in the following format: it starts with a row "Motifs:", followed by a matrix. This matrix is named PWM and the ascii alphabet is A,C,G, and T. Then the following rows represent which tissue this motif comes from, and if a row looks like this:"Fetallung [0.395172064836799 0.6073584107985984]", it means that this motif occurs significantly in fetal lung and its posterior probability is [0.395172064836799 0.6073584107985984].

2) Additional file 2 – motifs discovered in 82 human tissues which have significant occurrences in more than 70 tissues

File name is "CandidateHousekeepingMotifs.txt", and each motif is described in the following format: it starts with a row "Motifs:", followed by a matrix. This matrix is named PWM and the ascii alphabet is A, T, C, and G. Then the following rows represent which tissue this motif comes from, and the posterior probability in other human tissues.

3) Additional file 3 – 82 human tissues list

File name is "TissueList.txt", and each row in this file represents a human tissue name.