Tissue Significances Tests on DNA Binding Sequence Motifs for Human Genes

Hua Yu*† and Xiu-Jun GONG*† ‡
*School of Computer Science and Technology, Tianjin University, Nankai, Tianjin, China, 300072
†Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China, 300072
‡Corresponding author:gongxj@tju.edu.cn

***Abstract*—DNA binding sequence motifs are becoming increasingly important in the analysis of gene regulation, disease diagnosis and drug design. Although so far there are amount of tools available to discover these kinds of motifs, little was done to identify the biological functions, especially in tissue or cell type specific contributions, of those motifs. In this paper we used an integrated pipeline to discover sequences motifs for the promoter regions of human genes. Then we distinguished two types of motifs: tissue rich motifs (TRM) and tissue even motifs (TEM), using hypotheses test approaches including Bayesian hypothesis, Binomial distribution and traditional z-test. We finally got 233 overlapped TRMs and 56 TEMs. Most of those motifs are validated against JASPAR databases.**

Keywords*—tissue specificity, tissue rich motifs, tissue even motifs and hypothesis test.

I. INTRODUCTION

DNA binding sequence motifs are becoming increasingly important in the analysis of gene regulation, disease diagnosis and drug design [1]. In last years, many research projects were performed to study expressions and regulatory mechanisms of TS genes including transcription factor and their binding sites, sequence features of promoter regions [2] alternative splicing [3] and Epigenetics features [4] of those genes.

Although so far we are not completely clear about the mechanisms of the gene tissue specificity, the sequence features of TS genes are becoming an important clue[2]. P. FitzGerald et.at calculated the statistics of Simple Sequence Repeats (SSR) and identified that the SSR could be an important factor to the tissue specificity [5]. F. Song et.al pointed that methylation changes during development are dynamic, involve demethylation and methylation, and may occur at late stages of embryonic development or even postnatal using mouse genome data [6]. C. Heber et.al showed that Nucleosome rotational setting is associated with transcriptional regulation in promoters of tissue-specific human genes [4].

With the completion of the whole human genome project, various algorithms have been developed for discovering patterns or motifs of huge volume genome sequences. Those algorithms typical include three phases: motif searching, redundant motif pruning and motif significance testing. The methods for motif discovery may be grouped into two categories [7]: enumerative methods and alignment-based methods. Enumerative methods typically involve exhaustive enumeration of words up to some maximum size in a dataset, and are thus best suited to consensus sequence motif models, like Consensus, PROJECTION and PDEM. Alignment methods take on a wide variety of forms, but often involve development of a probabilistic model of the observed sequence data and optimization to finding motifs common to all input sequences, such as MEME program [8], the expectation-maximization (EM) algorithm and Gibbs sampling [9]. Each algorithm has its unique advantage on individual species or datasets. Tompa et al 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[7]. A common practice is to apply several such algorithms simultaneously to improve coverage at the cost of increased redundancy.

In this paper, we first applied an integrated motif searching approach to find motifs for human genes. As we known, it is first time to search sequence motifs for tissue specific genes. Then we merged the similar motifs using the method in literature[7]. To test the significances of those motifs in each tissue, we used three hypothesis test methods: Bayesian hypothesis, Binomial distribution and traditional z-test. We also distinguish two kinds of significant motifs: tissue rich motifs (TRM) and tissue even motifs (TEM). The former refer to motifs only showing significance in few tissues, and the later refer to motifs in most of tissues. We finally got 3/233 overlapped TRMs and 56/35 TEMs from 3954 TS genes across 83 human tissues respectively. Most of those motifs are validated against JASPAR databases.

II. METHOD

A. Data preparing

Tissue specific genes were obtained mainly by querying the tissue specific gene expression database TiGER [10] against the tissue names. Some of them came from TisGED database [11]. All of the TS genes with PubMed IDs were used in the experiment. We finally got 3954 human tissue specific genes across 83 human tissues. The gene's promoter sequences are downloaded from DBTSS [12] and EPD [13]. The promoter region with 1500bp (-499bp-1000bp around TSS) length is used for motif searching.

B. Motif searching

In this phase, we integrated three motif searching programs: MEME, AlignACE and Gibbs Sampler. The length of candidate motifs is fixed to 6-12 bp, other parameters as the default setting. In this phase, we get 6794 motifs.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

C. Motif tissue significance testing

To identify whether a motif is really related with tissue specificity or not, we statistically distinguish two kinds of motifs: tissue rich motifs (TRM) and tissue even motifs (TEM). The former refer to motifs only showing statics significance in less than 3 tissues, and the later refer to motifs in more than 70 tissues. We used hypothesis approaches to test the significance of motifs in each tissue. To do the hypothesis test, the distributions of motifs in a given sequence must be estimated. Therefore, a key step is to calculate the statistic of a motif in a given sequence.

For a given motif m with length w from tissue T_0 , in which the motif is discovered, our purpose is to judge whether its occurrence in in tissue T_1 is significant or not. Therefore we have to take a measure on the motif occurrences.

We define two kinds of match scores for a given motif to a promoter sequence.

Definition 1. *for a given motif m, its matching score with a Promoter sequence segMent x of the gene from tissue T_1 (PMS1) is defined in equation 1.*

$$PMS1(m,x) = \sum_{i=1}^{w} s(m, x[i]) \tag{1}$$

Definition 2. *for a given motif m , its matching score with a Promoter Sequence s of the gene from tissue T_1 (PSS1) is defined in equation 2.*

$$PSS1(m,s) = \frac{1}{n}\sum_{i=1}^{n} PMS1(m, s[i]) \tag{2}$$

Where $s[i]$ whose PMS1 score is more than a predefined threshold is a segment of S by sliding a widow with length w, n is the number of $s[i]$ in sequence s.

Definition 3. *for a given motif m, its matching score with a Promoter sequence segMent x of the gene from tissue T_1 (PMS2) is defined in equation 3.*

$$PMS2(m,x) = \frac{Current - Min}{Max - Min} \tag{3}$$

where

$$\begin{aligned} Current &= \sum_{i=1}^{w} I(i) f_{i,B} \\ Min &= \sum_{i=1}^{w} I(i) f_i^{\min} \\ Max &= \sum_{i=1}^{w} I(i) f_i^{\max} \\ I(i) &= \sum_{B\in\{A,T,G,C\}} f_{i,B}\ln(4 f_{i,B}) \end{aligned} \tag{4}$$

In the equations 3 and 4, $f_{i,B}$ is the frequency of residue B at position i, which is from PWM; $f_i^{\min}$ /$f_i^{\max}$ is the smallest/largest frequency of the residue at position i; and $I(i)$ describes the information content of residue B at position i.

Definition 4. *for a given motif m , its matching score with a Promoter Sequence s of the gene from tissue T_1 (PSS2) is defined in equation 5.*

$$PSS2(m,s) = \sum_{i=1}^{n} PMS2(m, s[i]) \tag{5}$$

Where $s[i]$ whose PMS2 score is more than a predefined threshold is a segment of S by sliding a widow with length w, n is the number of $s[i]$ in sequence s.

1) Classical Z-test

In the classical z-test, we estimated the mean and variance of the match score in tissue T_1 , and then calculated the z-value using 6

$$z = \frac{\overline{PSS1} - \mu_0}{\sigma/\sqrt{n}} \tag{6}$$

where μ_0 and σ are the mean and variance of the match score in tissue T_0. In the experiment, we set the confidence degree 0.05.

2) Bayesian Hypothesis Test

Assumed that matches of a motif at tissue T_0 follows a Gaussian distribution $N(\mu_0, \sigma_0^2)$. To test that whether the motif is significant at tissue T_1 , we constructed two hypothesizes as the followings:

$$H_0 : \mu_0 < x_1, H_1 : \mu_0 \geq x_1 \tag{7}$$

Where x_1 is the mean of the match score in tissue T_1 .

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. The post distribution of θ is fol-lowed $N(\mu(x), \rho^{-1})$ according [16], where

$$\rho = \tau^{-2} + \sigma^{-2} = \frac{\tau^2 + \sigma^2}{\tau^2\sigma^2} \tag{8}$$

$$\begin{aligned} \mu(x) =& \frac{1}{\rho}(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}) \\ =& \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}x \\ =& x - \frac{\sigma^2}{\sigma^2+\tau^2}(x-\mu) \end{aligned} \tag{9}$$

3) Binomial distribution test

In Binomial distribution test, we need the number of matches between the motif and the promoter sequence of a gene. A match between a motif and a sequence is defined if the match score of the motif with a segment of the sequence is larger than a predefined value. We counted all the matches in tissue T_0 and T_1 , represented the numbers of matches by K_0 and K_1 respectively. The Binomial distribution test is to seek a value k_value holding:

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

$$\sum_{m=0}^{k\_value} \binom{n_1}{m} p^m (1-p)^{n_1-m} = \sum_{m=0}^{k_0} \binom{n_0}{m} p^m (1-p)^{n_0-m} \quad (10)$$

Where n_0 and n_1 are the numbers of promoter sequences in tissue T_0 and T_1 re-respectively and p is fixed to 0.5 in the experiment.

III. RESULTS

A. Data sources

The gene expression datasets, such as GNF, SAGE, and EST, are very widely used as data sources for the identifications of TS genes. 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, tissue specific genes were obtained mainly by querying the tissue specific gene expression database TiGER against the tissue names. Some of them came from TisGED database. All of the TS genes with PubMed IDs were used in the experiment. We obtained 3954 TS genes across 83 human tissues. Because of the limitation of page size, the gene lists for all the tissues are available on request to authors.

The genes promoter sequences were downloaded from DBTSS and EPD. The promoter region with length 1500bp (-499bp-1000bp around TSS) is used for motif discovery.

B. Motifs discovered by three test methods

After merging phase, we get total 3244 motifs. The number of motifs in each tissue is shown in table I. After hypothesis phase, the number of TRMs and TEMs is show in table II, the more details of the numbers in each tissue are shown in figure 1 and 2. Looking at the tables and figures, we conclude that Bayesian hypothesis test and binomial distribution test using PSS1 scoring can get more TRMs and Bayesian hypothesis test using PSS2 scoring can get more TEMs.

TABLE II. NUMBER OF TRMS AND TEMS

	PSS1 scoring		PSS2 scoring	
	TRM	TEM	TRM	TEM
Classic Hypothesis Test	430	167	539	164
Binomial Distribution Test	1629	290	279	925
Bayesian Hypothesis Test	1534	1270	412	2390

C. Overlapped motifs in three testing methods

To find the overlaps of motifs among different testing methods, we draw the Venn diagrams for the numbers of the overlapped motifs, see figure 3 and 4. From the diagrams, there are 233 overlapped TRMs using PSS2 scoring, only 3 using PSS1 scoring. So we presumed that PSS2 scoring is better than PSS1 in identifying TRMs. There is no big difference for the numbers of overlapped TEMs using PSS1 and PSS2 scoring. We validate the overlapped motifs against JASPAR database[14]. All 3 TRMs in the left of Figure 3 have corresponding JASPAR IDs. 25 matches out of 56 overlapped motifs in right of Figure 3 are found. The numbers of matches with JASPAR in Figure 4 are 109 and 22

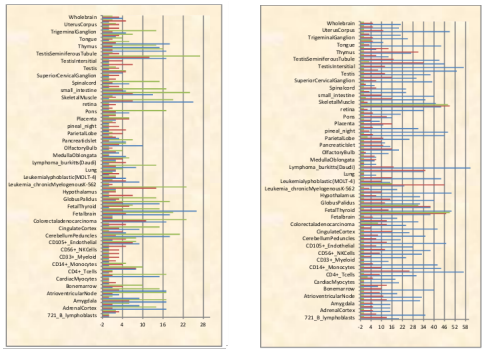

Fig. 1. The numbers of TRM (left) and TEM(right) using PSS1 scoring

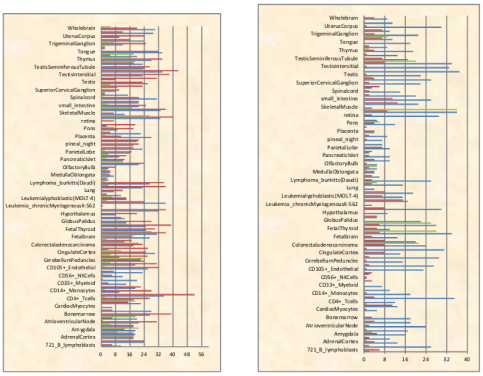

Fig. 2. The numbers of TRM (left) and TEM(right) using PSS2 scoring

respectively. For an example, $[CCCCNCCCCC]$ is a motif which was discovered by previous researches with JASPAR ID $MA0079.2\_SP1$, and $[GGGGAATCCCC]$ with JASPAR ID $MA0105.1\_NFKB1$.

IV. CONCLUSION

Tissue specificity is the foundation for cells form specific tissues and functional organs. Identification and analysis of tissue-specific genes and their regulatory activities play an important role in understanding mechanisms of the organism, disease diagnosis and drug design. And finding accurate and meaningful motif with tissue specificity still remains a big challenge.

In this paper we used an integrated pipeline to discover sequence motifs for the promoter regions of TS genes. To test the significances of those motifs in a specific tissue, we used hypotheses test approaches including Bayesian hypothesis, Binomial distribution and traditional z-test by two scoring schemas. We finally got 3/233 overlapped TRMs and 56/35 TEMs respectively. Most of those motifs are validated against JASPAR databases.

ACKNOWLEDGMENT

This research is partly supported by the Natural Science Funding of China under grand number 61170177, National Basic Research Program of China under grand number 2013CB32930X and innovation funding of Tianjin University.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

TABLE I. NUMBER OF MOTIFS IN EACH TISSUE AFTER MOTIF MERGING

tissue	#	tissue	#	tissue	#	tissue	#
721_B_lymphoblasts	22	Cerebellum	53	FetalThyroid	50	pineal_day	50
Adipocyte	53	Cerebellum_Peduncles	42	Fetallung	53	pineal_night	48
AdrenalCortex	32	CiliaryGanglion	41	GlobusPalidus	37	Pituitary	31
Adrenalgland	31	CingulateCortex	44	Heart	40	Placenta	33
Amygdala	35	Colon	44	Hypothalamus	32	PrefrontalCortex	29
Appendix	43	Colorectalade_nocarcinoma	43	Kidney	62	Pons	22
Atrioventricular_Node	47	DorsalRootGanglion	34	Leukemia_chronicMyelogenousK-562	19	Prostate	26
BDCA4+_Dentritic Cells	38	SkeletalMuscle	41	Leukemia_promye _locytic-HL-60	63	Tongue	53
Bonemarrow	50	Skin	54	Leukemialyphoblastic(MOLT-4)	30	Tonsil	13
BronchialEpithelialCells	28	small_intestine	48	Liver	44	TrigeminalGanglion	33
CardiacMyocytes	20	SmoothMuscle	42	Lung	12	Uterus	34
Caudatenucleus	35	Spinalcord	46	Lymphnode	45	UterusCorpus	49
CD4+_Tcells	41	SubthalamicNucleus	18	Lymphoma_burkitts(Daudi)	64	WholeBlood	40
CD8+_Tcells	77	SuperiorCervicalGanglion	39	Lymphoma_burkitts(Raji)	10	Wholebrain	44
CD14+_Monocytes	44	TemporalLobe	36	MedullaOblongata	13	Fetalbrain	42
CD19+_BCells (neg._sel.)	43	Testis	54	OccipitalLobe	10	Fetalliver	65
CD33+_Myeloid	21	TestisGermCell	53	OlfactoryBulb	14	retina	34
CD34+	24	TestisIntersitial	63	Ovary	33	Salivarygland	69
CD56+_NKCells	33	TestisLeydigCell	68	PancreaticIslet	22	Thymus	49
CD71+_Early Erythroid	39	TestisSemini ferousTubule	44	Pancreas	19	Thyroid	29
CD105+_Endothelial	51	Thalamus	36	ParietalLobe	29		

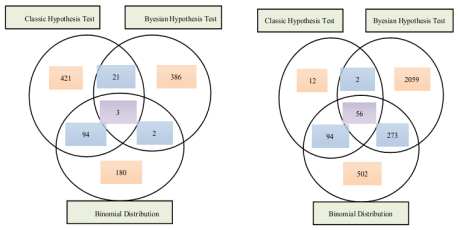

Fig. 3. Venn diagrams of numbers of TRM (left) and TEM(right) using PSS1 scoring

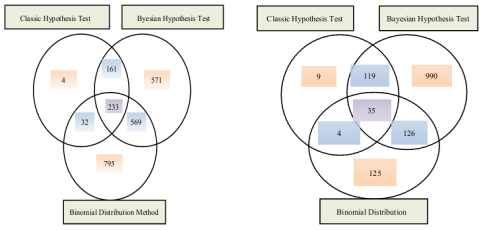

Fig. 4. Venn diagrams of numbers of TRM (left) and TEM(right) using PSS2 scoring

REFERENCES

[1] Z. Dezso, Y. Nikolsky, E. Sviridov, W. Shi, T. Serebriyskaya, D. Dosymbekov, A. Bugrim, E. Rakhmatulin, R. J. Brennan, A. Guryanov, K. Li, J. Blake, R. R. Samaha, and T. Nikolskaya, "A comprehensive functional analysis of tissue specificity of human gene expression," *BMC biology*, vol. 6, no. 49, Jan. 2008.

[2] D. Kuzmin, E. Gogvadze, R. Kholodenko, D. P. Grzela, M. Mityaev, T. Vinogradova, E. Kopantzev, G. Malakhova, M. Suntsova, D. Sokov, Z. Ivics, and A. Buzdin, "Novel strong tissue specific promoter for gene expression in human germ cells," *BMC biotechnology*, vol. 10, no. 1, p. 58, Jan. 2010.

[3] A. Grosso, A. Gomes, and N. Barbosa, "Tissue-specific splicing factor gene expression signatures," *Nucleic Acids*, vol. 36, no. 15, pp. 4823–4832, 2008.

[4] C. Hebert, "Nucleosome rotational setting is associated with transcriptional regulation in promoters of tissue-specific human genes," *Genome Biology*, vol. 11, no. 5, Jan. 2010.

[5] M. J. Lawson and L. Zhang, "Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5 -UTR region," *Gene*, vol. 407, pp. 54 – 62, 2008.

[6] F. Song, S. Mahmood, S. Ghosh, P. Liang, D. J. Smiraglia, H. Nagase, and W. A. Held, "Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development," *Genomics*, vol. 93, no. 2, pp. 130–9, Feb. 2009.

[7] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnology*, vol. 23, no. 1, pp. 137–144, 2005.

[8] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. Web Server, pp. W202–W208, 2009.

[9] A. F. Neuwald, J. S. Liu, and C. E. Lawrence, "Gibbs motif sampling Detection of bacterial outer membrane protein repeats," *Protein science : a publication of the Protein Society*, vol. 4, no. 8, pp. 1618–1632, 1995.

[10] X. Liu, X. Yu, D. J. Zack, H. Zhu, and J. Qian, "TiGER : A database for tissue-specific gene expression and regulation," *BMC Bioinformatics*, vol. 7, pp. 1–7, 2008.

[11] S.-J. Xiao, C. Zhang, and Z.-L. Ji, "TiSGeD: a Database for Tissue-Specific Genes," *Bioinformatics*, vol. 26, no. 9, pp. 1273–1275, Mar. 2010.

[12] Y. Suzuki, R. Yamashita, K. Nakai, and S. Sugano, "DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs." *Nucleic acids research*, vol. 30, no. 1, pp. 328–31, Jan. 2002.

[13] R. C. Périer, V. Praz, T. Junier, C. Bonnard, and P. Bucher, "The Eukaryotic Promoter Database (EPD)," *October*, vol. 28, no. 1, pp. 302–303, 2000.

[14] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard, "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." *Nucleic acids research*, vol. 32, no. Database issue, pp. D91–4, Jan. 2004.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE