Novel Method to Quantify the Distribution of Transcription Start Site

Budrul Ahsan *ahsan@cb.k.u-tokyo.ac.jp*
Department of Neurology, Graduate School of Medicine
The University of Tokyo
Tokyo 113-8655, Japan

Shinichi Morishita *moris@cb.k.u-tokyo.ac.jp*
Department of Computational Biology, Graduate School of Frontier Science
The University of Tokyo
Kashiwa 277-0882, Japan

Abstract

Studies of Transcription Start Site (TSS) show that a gene has several TSSs locally distributed in promoter region. Analysis of this TSS distribution may decipher the gene regulatory mechanism. For that purpose, a numerical representation of TSS distribution is crucial for quantitative analysis of TSS data. To characterize the TSS distribution in quantitatively, we have developed a novel scoring method by considering several significant features that are contributing to shape a TSS distribution. Comparing to other methods, our scoring method describes TSS distribution in a meaningful and effective way. Efficiency of this method to distinguish TSS distribution is evaluated with both synthetic and real dataset.

Keywords: TSS, Transcription Start Site, CAGE, 5'end SAGE, Gene Regulation, Gene Expression.

1. INTRODUCTION

Initiation of transcription is the primary but fundamental step in gene expression process. Regulation of gene expression begins largely from initiation step of transcription. During eukaryotic gene expression process, the assembly of general transcription factors and RNA polymerase enzyme bind around the transcription start site (TSS) to initiate the transcription activity. Generally, these binding sites of transcription factors are defined as promoter region of a gene [1]. Therefore, study of TSSs and their related promoters in genome is essential to unravel transcription regulation riddle. For a global understanding of gene regulation, several novel technologies (CAGE, 5'end SAGE and PEAT) have been developed to capture 5'end of mRNA transcripts [2-4]. Moreover, adaptation of these technologies to the recent high throughput sequencers such as Illumina/Solexa and ABI/SOLiD has given a new momentum in genome-wide TSS studies [5-7]. Depending on the restriction endonuclease, these capturing methods collect about 20~27bp short sequence starting from TSS of mRNA transcript. This short sequence is regarded as 5'end mRNA tag or in short tag in this article. As 5'end of each tag is the starting position of the mRNA transcript, mapping of the tag to the genome provides the TSS position of the original mRNA transcript and the total number of tags that are starting from a TSS gives the expression level of its original mRNA transcript as illustrated in Figure 1. Recent TSS studies demonstrated that most of the genes contain locally concentrated multiple TSSs as depicted in Figure 2. These TSSs and their expression levels create TSS distribution in the promoter region of a gene. TSS distribution in each promoter region implies transcription initiation mechanism of its related gene. Therefore, study of TSS distribution has the potentiality to elucidate the gene regulation mechanisms in cells.

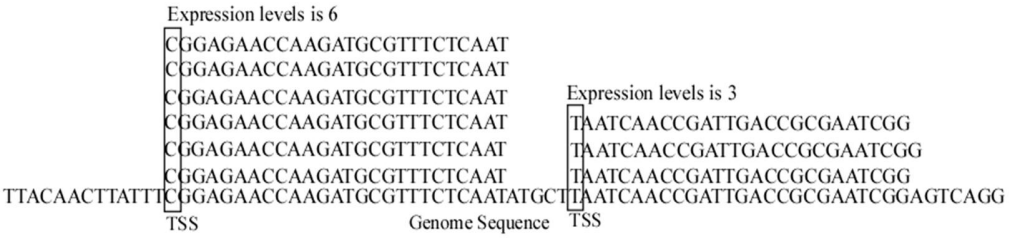

FIGURE 1: 5'end mRNA tags and their expression levels. Aligned 5'end mRNA tags are overlapped in genome. The starting position of each aligned tag is regarded as the Transcription Start Site (TSS). The frequency of each TSS gives the expression level of its original mRNA.

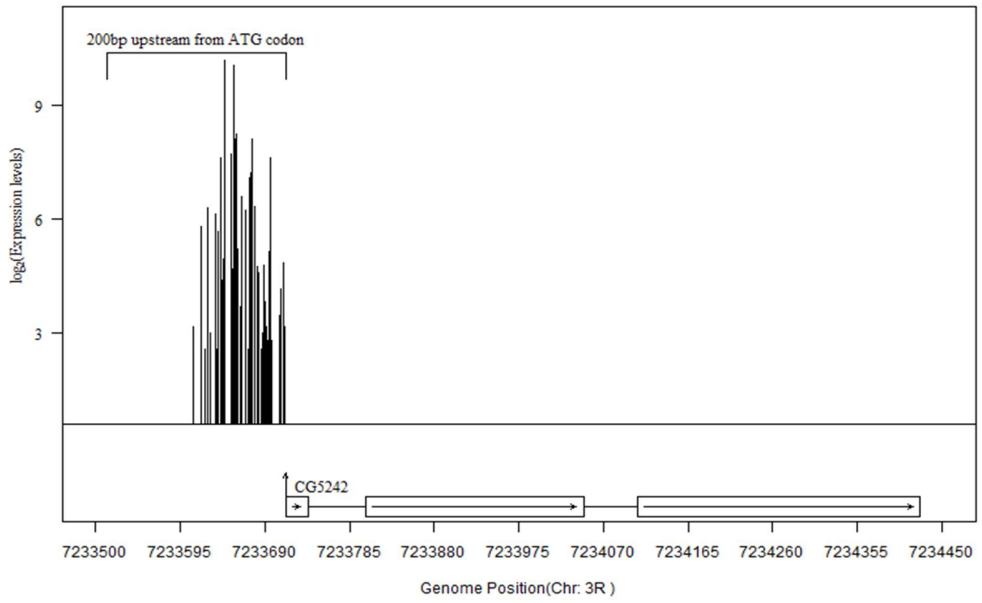

FIGURE 2: Expression distribution at promoter region of *Drosophila melanogaster* mRNA geneCG5242. The vertical arrow at the 5'end of gene CG5242 in bottom row is the initiated site of coding region. In this image, genomic position from 5'end to 3'end is depicted at *x*-axis and the $\log_2$ (Expression levels) is illustrated in *y*-axis.

In TSS studies, it is essential to assign a numerical score to quantitatively classify each promoter region with respect to its TSS distribution. Quantitative characterization of TSS distribution enables gene expression analysis such as clustering genes with respect to their TSS distributions. Quantitative classification of genes also distinguishes differentially expressed genes having disparity in their TSS distributions in case-control studies. Moreover, this quantification method facilitates genome browser to selectively choose and visualize genes having particular type of TSS distributions for further biological studies. To address this problem, *Density Percentile (DP)* within a promoter region has been introduced to categorize TSS distribution [3]. Using *DP* method, promoters having 100 tags or more are categorized into four different classes such as single peak, dominant peak, multimodal peak and broad. As *DP* does not assign score to promoters with respect to their TSS distributions and only classifies them in different groups, it is not efficient for quantitative TSS studies. Recently, *Shape Index (SI)* [8] is introduced to assign a numerical score to the TSS distribution of a promoter. *SI* is defined as follows,

$$SI = 2 + \sum_{i}^{L} p_i \log_2 p_i \quad (1),$$

where p_i is the probability of observing a TSS at base position i within the promoter. L is the number of base positions that have expression levels more than zero. Promoter regions with *SI* score ≥ -1 are classified as peaked and remaining promoters are classified as broad. The principal drawback of *SI* method is that the scoring system considers only expression levels of TSSs, but their spatial orientation is not incorporated in scoring method. From Figure 1 and Figure 2, we can understand that TSS distribution in a promoter region is determined by not only the expression levels of TSSs but also how the expression levels (illustrated as vertical line in Figure1) of TSSs are spatially oriented in the promoter region. As a result, *SI* assigns same score to some TSS distributions, while considerable discrepancy is noticeable among the TSS distributions. In this regard, a numerical representation is essential to precisely quantify the pattern of TSS distribution. The proposed method will benefit if we can consider the significant features such as expression levels of TSSs and spatial orientation of TSSs in a promoter region that are contributing to create the shape of a TSS distribution. By incorporating aforementioned features of a TSS distribution, a scoring method named *Aggregated Index (AI)* is proposed here.

In the following sections, we firstly present the scoring method. Secondly, we experiment the method on both synthetic dataset and real TSS dataset. Finally, we discuss the effectiveness of this scoring method in discussion.

2. METHOD

We define a promoter $C\{y_i, i = 1,2,3,\cdots,K\}$ of K-*mer* length where y_i is the expression level at position i starting from 5'end of the promoter. Total expression in a promoter region C is summed up as $Y = \sum_{i=1}^{K} y_i$. The total expression Y is distributed among K individual bases in that promoter. We discuss how the expression levels and spatial orientation of bases in a promoter are utilized in our scoring method. In the following sub-sections, our proposed method is explained in three steps. Firstly, divergence of TSSs' expression levels is quantified using *Gini Coefficient (GC)*. Secondly, spatial orientation of TSSs is quantified in *Average Neighbourhood Distance (AND)*. Finally, both *GC* and *AND* are used to define the *Aggregated Index (AI)*.

2.1 Divergence of Expression Levels

Observation of Figure 1 and Figure 2 implies that expression level of bases in a promoter region is one of the significant features of a TSS distribution. Therefore, incorporation of expression levels in our scoring method is important to properly quantify a TSS distribution. Our main objective is to consider how disparity of expression levels among the bases in a promoter works to make a TSS distribution highly aggregated or not. To quantify the variability of the expression level in a TSS distribution, we use *Gini Coefficient* [9, 10] in our scoring method. Although, this coefficient is used by economists to illustrate the concentration of wealth distribution in a population, it can be used in all kinds of contexts where size plays a role like gene expression among all bases in a promoter region. The expression levels of a promoter region $C\{y_i, i = 1,2,3,\cdots,k\}$ is ranked in ascending order as, $\tilde{y}_1 \leq \tilde{y}_2 \leq \tilde{y}_3 \leq \cdots \leq \tilde{y}_K$. Kendall and Stuart defined *Gini Coefficient (GC)* as follows [11] :

$$GC = \frac{1}{2K^2\mu} \sum_{i=1}^{K} \sum_{j=1}^{K} |\tilde{y}_i - \tilde{y}_j| \quad (2),$$

μ is the average levels of expression, i=1,2,3,⋯,K and j=1,2,3⋯,K. If there is only one base that has non-zero expression level in a 200bp length promoter region, then *GC* value of that TSS distribution is 1. This implies that the TSS distribution of that promoter is highly concentrated to a single TSS. On the other hand, if the expression levels are equally distributed to all the 200 bases in that promoter, then the *GC* value of that promoter is 0. Therefore, *GC* always takes value between zero and one.

2.2 Spatial Orientation of TSS

Spatial orientation of bases that have non-zero expression level is another important feature of TSS distribution of a promoter. Despite having equal expression levels in the bases of two promoters, orientation of those bases can create different TSS distributions pattern in those promoters. Therefore, the spatial feature of TSSs is incorporated in *AI* scoring method using *Average Neighbourhood Distance (AND). The AND* is defined as below:

$$AND = \frac{1}{L}[1 + (q - p)] \quad (3).$$

In equation 3, p is the first base position that has non-zero expression level, and q is the last base position that has non-zero expression level starting from 5'end of a promoter. Here, L is the total number of bases in the promoter having non-zero expression level. For example, a promoter of length 9 has expression levels of 5,4,0,1,0,2,0,1,3 in the bases position $1,\cdots,9$, starting from 5'end of the promoter. In this example, the number of bases having non-zero expression level is 6. According to the equation 3, $p = 1, q = 9$ and $L = 6$. Therefore, the value of *AND* is 1.5. On the other hand, in an extreme case, if all the bases in a promoter have non-zero expression levels, the value of *AND* will be 1. Except this extreme case, the value of *AND* will be always above one. As a result, the value of *AND* is always one or more than one.

2.3 Aggregated Index

To quantify the TSS distribution, we have targeted at two significant features such as divergence of expression levels and spatial orientation of bases in a promoter of a gene. Firstly, the divergence of expression levels is explained by *GC* of equation 2 that takes score within the range of zero and one. Secondly, spatial orientation of TSSs is quantified in *AND* of equation 3 that takes score one and above. Finally, using *GC* and *AND*, the aggregated index *(AI)* is defined as below:

$$\text{Aggregated Index (AI)} = \text{GC}/\text{AND} \quad (4).$$

AI assigns one single value between zero and one to a TSS distribution in a promoter. For example, if there is only one base having expression level more than zero in a promoter of 200bp length, the total expression level in that promoter is distributed to that single base. In this case, the proposed *AI* assigns value of one that implies the TSS distribution in the promoter is deterministic to a single base position of genome. Moreover, this promoter can be categorized as highly aggregated in its TSS distribution. On the other hand, when all the 200 bases of the promoter have same non-zero expression levels of TSS distribution, *AI* assigns value of zero to the TSS distribution of that promoter. Therefore, the TSS distribution having value of zero or near to zero is categorized as random or nondeterministic TSS distribution.

3. RESULT

To further reinforce the effectiveness of the proposed *AI* scoring method, we tested and verified the *AI* scoring method to distinguish TSS distribution in a promoter region with both synthetic and real TSS dataset.

3.1 Synthetic Dataset

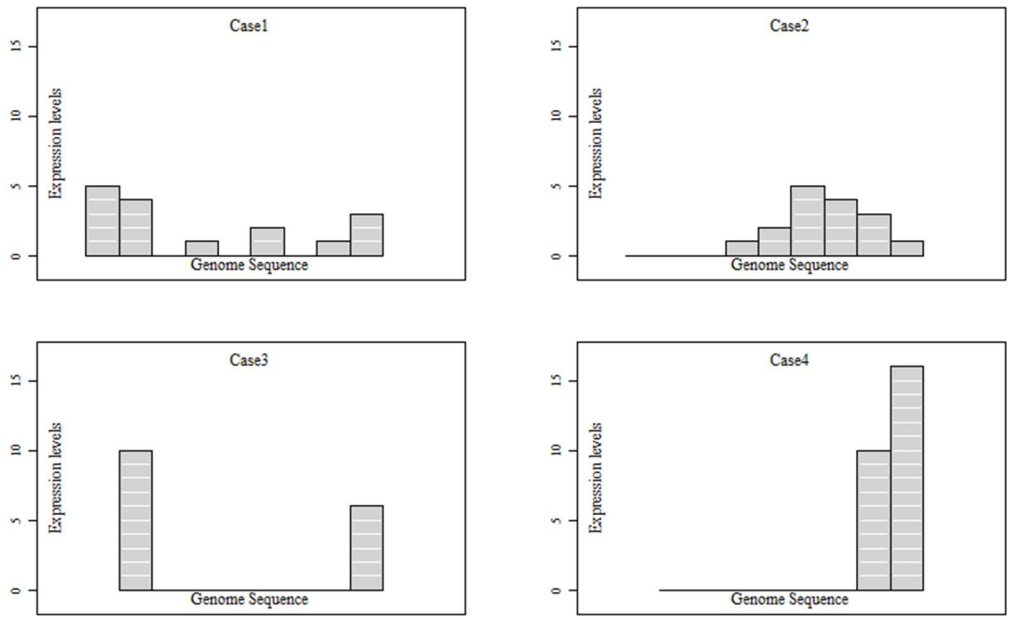

FIGURE 3: Four synthetic examples of TSS distribution with various patterns are showed in this figure, where x-axis is promoter region of a genome and y-axis is expression levels of mRNA transcript.

Example	Promoter	*GC*	*AND*	*AI*	*SI*
Case1	5,4,0,1,0,2,0,1,3	0.54	1.5	0.36	-0.35
Case2	0,0,0,1,2,5,4,3,1	0.54	1	0.54	-0.35
Case3	10,0,0,0,0,0,0,6	0.78	4	0.195	1.04
Case4	0,0,0,0,0,0,6,10	0.78	1	0.78	1.04

TABLE 1: AI values for synthetic promoter examples

Four synthetic examples of promoters that have various TSS distributions are illustrated in Figure 3 & Table1.These examples are presented to examine *AI*'s ability to distinguish TSS distribution by assigning a numerical score. We categorised the four examples in two groups. Firstly, group1 consists of Case1 and Case2. In this group, total expression level of each of the cases is equal; however, the spatial orientation of bases with non-zero expression levels in each promoter is different. Figure 3 shows that TSS distribution in Case1 is random, while in Case2 the distribution is aggregated to make a bell shape pattern. Secondly, group2 is comprised of Case3 and Case4 promoters. All the promoters in group2 also have equally total expression levels; however, two distinct bases with non-zero expression levels are positioned far away from each of the bases in case3 that creates different TSS distribution comparing to Case4 promoter in the same group that have two bases with non-zero expression levels which are located at 3'end of the promoter. In

group1, *GC* scores of Case1 and Case2 promoters are 0.54; on the other hand, *GC* scores in group2 for both Case3 and Case4 are 0.78 (Table1). Figure 3 shows that without the orientation of bases that have non-zero expression levels, the total expression levels in both cases of group1 are same; similarly, both Case3 and Case4 of group2 have equivalent total expression levels. Although the TSS distributions of these promoters are different, their *GC*'s scores are similar in each group. It is because, in the process of *GC* calculation using equation 2, we ranked the expression levels of each base that ignored the spatial information of bases and made the *GC* score similar in both cases of each group. Therefore, in order to have a better scoring method to describe properly the TSS distribution in a promoter, it is necessary to consider the spatial orientation of the bases having expression level more than zero. As a result, spatial orientations are considered through *AND* to properly distinguish each cases of promoters in group1 and group2. In group1, *AND* scores for Case1 and Case2 are 1.5 and 1 respectively; in group2, Case3 and Case4 are 4 and 1 respectively (Figure 3 & Table 1). Finally, *GC* and *AND* are combined at *AI* in equation 4. *AI* scores for all promoter examples are Case1=0.36, Case2=0.54, Case3=0.195 and Case4=0.78 (Table1). By considering significant features of TSS distribution, *AI* successfully assigned scores to each promoter. Especially, *AI* distinguished Case1, Case2, Case3 and Case4 of each properly. In contrast to *AI* score, *Shape Index* (*SI)* assimilated Case1, Case2 and Case3, Case4 by scoring same values in each pair (Table1); because, it does not incorporate information of spatial orientation of bases in a promoter in the scoring method defined in equation 1.

3.2 Real TSS Dataset

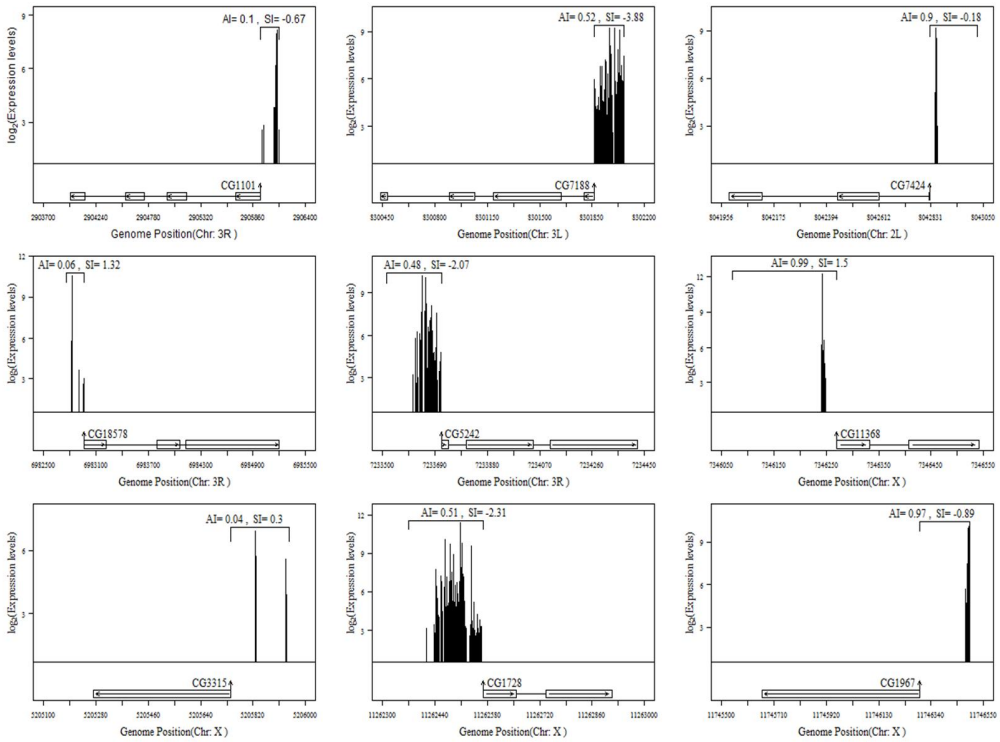

FIGURE 4: AI scores of Drosophila melanogaster genes. TSS distributions of promoter of nine genes are illustrated in Figure 4. Left column is for genes CG1101, CG18578 and CG3315 having AI scores between $\mathbf{0} \leq \boldsymbol{AI} \leq \mathbf{0.1}$. Middle column is for genes CG7188, CG5242 and CG1728 with AI scores range $\mathbf{0.45} \leq \boldsymbol{AI} \leq \mathbf{0.55}$. In right column, genes CG7424, CG11368 and CG1967 are depicted with AI score between $\mathbf{0.9} \leq \boldsymbol{AI} \leq \mathbf{1}$. SI scores for each of the promoter's expression distribution are also presented with AI scores.

Evaluation of *AI* method was performed with TSS data collected from publicly available database called *Machibase* [12]. *Machibase* is a TSS database for *Drosophila melanogaster* that consists of six development stages such as embryo, larva, young male, young female, old male, old female and one culture cell line (S2). All the TSS data from seven libraries were merged and assigned *AI* score to the promoters of *Drosophila melanogaster* genes with respect of their TSS distributions. Promoter information is collected from *Flybase 5.2 [13]* annotated mRNA genes. Promoter region of each mRNA gene is defined as 200bp upstream of coding initiation site (ATG codon). Each bases of promoter region that has more than five expression levels is assigned TSS expression levels from *Machibase* data. Finally, *AI* score of TSS distribution is calculated for all the promoters of genes according to equation 2, 3 and 4. With respect to *AI* scores $(0 \leq AI \leq 0.1, 0.45 \leq AI \leq 0.55, 0.9 \leq AI \leq 1)$ nine genes were illustrated in Figure 4. Among these genes CG1101, CG18578 and CG3315 (illustrated in left column of Figure 4) have *AI* scores between $0 \leq AI \leq 0.1$. The TSS distributions of this group are similar to the example Case3 in synthetic dataset (Figure 3 & Table 1). Genes CG7188, CG5242 and CG1728 (illustrated in mid column of Figure 4) have *AI* scores between $0.45 \leq AI \leq 0.55$. The TSS distributions of this group can be categorized to the example Case2 in the synthetic dataset (Figure 3 & Table1). Finally, TSS distributions of genes CG7424, CG11368 and CG1967 (illustrated in right column of Figure 4) having *AI* score between $0.9 \leq AI \leq 1$ can be categorized to Case4 of the synthetic dataset (Figure 3 & Table1). Examples from real dataset in Figure 4 show how efficiently *AI* score can categorize genes according to their TSS distributions in promoters. On the other hand, *Shaped Index* (*SI)* method categorizes all genes in left and right columns as peaked TSS distribution, where clear disparity exists in their TSS distributions. This result also confirms that *AI* scoring system works well to classify genes by providing numerical score to each gene with respect to its TSS distribution. By assigning well defined scores to TSS distribution of *Drosophila melanogaster* genes, *AI* method obviously outperformed *SI* scoring method in distinguishing TSS distribution pattern of promoter region.

4. DISCUSSION

TSS study has the potentiality to elucidate gene regulation mechanism. In TSS study, it is essential to quantify TSS distribution in a promoter region of a gene. As existing *Density Percentile (DP)* method does not assign any numerical score to TSS distribution, it is not efficient for further quantitative analysis of TSS data. On the other hand, *Shape Index (SI)* method considers only expression levels in its scoring system of equation 1, and resulting score cannot distinguish significant disparity among TSS distributions. After considering all the features that contribute to shape TSS distribution in a promoter region, we proposed *Aggregated Index (AI)* scoring in this study.

AI is a novel scoring method to measure the TSS distribution of a promoter. Evaluation in synthetic data shows the proposed method is able to distinguish distinct patters of TSS distribution in promoter regions. However, the existing *Shape Index* (*SI)* scoring method assigns same scores to some TSS distributions in our synthetic data while significant discrepancy exists among them (Table 1). Furthermore, *AI* also successfully distinguished all the TSS distributions in real TSS dataset as depicted in Figure 4. In contrast, *SI* scores of all the examples in right and left columns of Figure 4 are above -1. As a result, in *SI* scoring system, all of these TSS distributions in right and left columns in Figure 4 are classified as peaked promoters. Thus, *SI* scoring system cannot distinguish obvious disparity among TSS distributions in real dataset. By assigning scores to distinct patterns of TSS distributions, *AI* method allows us to cope with the problem of TSS analysis to a treatable scale. Therefore, using synthetic and real dataset, we verified the advantage of this scoring method in TSS data analysis. In other word, the proposed *AI* method has opened up a new direction for future approaches to genome-wide analysis of gene regulation using TSS data.

The contribution of the proposed *AI* is significant mainly in the following two ways. Firstly, the score can quantify the TSS distribution of promoter region by providing a unique measurement

technique to reduce the ambiguity in TSS analysis. Secondly, the *AI* score can automatically identify the particular pattern of TSS distribution in genome browser, to our knowledge no other scoring method can do like this and that is why *AI* scoring could be an enormous help for biologists working in gene expression and regulation process.

5. REFERENCE

1. Alberts, B., *Molecular biology of the cell*. 5th ed. 2008, New York: Garland Science.
2. Hashimoto, S., et al., *5'-end SAGE for the analysis of transcriptional start sites.* Nat Biotechnol, 2004. **22**(9): p. 1146-9.
3. Carninci, P., et al., *Genome-wide analysis of mammalian promoter architecture and evolution.* Nat Genet, 2006. **38**(6): p. 626-35.
4. Ni, T., et al., *A paired-end sequencing strategy to map the complex landscape of transcription initiation.* Nat Methods. **7**(7): p. 521-7.
5. Fullwood, M.J., et al., *Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses.* Genome Res, 2009. **19**(4): p. 521-32.
6. Hashimoto, S., et al., *High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer.* PLoS ONE, 2009. **4**(1): p. e4108.
7. Valen, E., et al., *Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE.* Genome Res, 2009. **19**(2): p. 255-65.
8. Hoskins, R.A., et al., *Genome-wide analysis of promoter architecture in Drosophila melanogaster.* Genome Res. **21**(2): p. 182-92.
9. Gini, C., ***Measurement of Inequality and Incomes***. The Economic Journal, 1921(31): p. 3.
10. Anand, S., *Inequality and poverty in Malaysia : measurement and decomposition*. A World Bank research publication. 1983, New York: Published for the World Bank [by] Oxford University Press. x, 371 p.
11. Kendall, M.G. and A. Stuart, *The advanced theory of statistics*. [3 vol. ed. 1963, New York,: Hafner Pub. Co.
12. Ahsan, B., et al., *MachiBase: a Drosophila melanogaster 5'-end mRNA transcription database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D49-53.
13. Drysdale, R.A. and M.A. Crosby, *FlyBase: genes and gene models.* Nucleic Acids Res, 2005. **33**(Database issue): p. D390-5.