19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

Identification of Transcription Start Sites via Distribution of A/T-singletons

Phirayu Lanlieng[1*], Chatchawit Aporntewan [1-4], Monnat Pongpanich[1-4†]

[1]Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University Bangkok, Thailand
[2]Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok, Thailand.
[3]Center for Excellence in Molecular Genetics of Cancer and Human Diseases, Chulalongkorn University, Bangkok, Thailand.
[4]Chulalongkorn Academic Advancement into its 2nd Century Project, Chulalongkorn University, Bangkok, Thailand.

Email: *Phirayu.L@gmail.com, †Monnat.P@chula.ac.th

***Abstract*—Transcription start sites (TSSs) are crucial information that determines exact location of genes. However, identifying TSSs in vitro is costly and time consuming. Therefore, there are many attempts to predict TSSs in silico, but they were low in accuracy. Herein, we observed that the distribution of A/T-singletons in the whole genome can be employed to detect the presence of TSS. We found that the distribution pattern is clearly distinct between regions with TSS and without TSS. To assess whether the distribution of A/T-singletons can help detect TSS, we developed a two-step algorithm to detect the specific distribution pattern. Our method was evaluated in terms of sensitivity, specificity and accuracy. The results show that the distribution of A/T-singletons is a useful feature for identifying TSSs. However, using this feature alone is not sufficient to identify all TSSs correctly. Combining this feature with other existing methods should improve their efficacy significantly.**

Keywords—Transcription start site; A/T-singletons

I. INTRODUCTION

Gene expression, the process that DNA directs protein synthesis, requires two major steps—transcription and translation. Transcription produces messenger RNA (mRNA) from DNA, and translation translates mRNA into a protein [1]. Transcription begins when the enzyme called RNA polymerase II (RNA pol II) binds to a region of DNA located upstream of a gene and nearby a transcription start site (TSS), a point where transcription start. These regulatory regions are called promoters. After that, DNA is unwound. This allows RNA pol II to read the template sequence and synthesize RNA by adding nucleotides to the 3' end of the RNA molecule. This process continues until a terminator sequence has been transcribed [2]. The term "upstream/downstream" is used to described a relative position before/after a certain landmark e.g. TSS. Transcriptions take place in a 5' to 3' direction; thus, upstream and downstream is toward the 5' end and 3' end of an RNA molecule respectively.

Based on the knowledge of transcription process, the earlier studies use the location of promoters to predict TSSs as they are nearby [3]. For example, some computer programs locate the TATA box (a DNA sequence found in the promoter regions), which is 25–35 bases before TSS [4]. Some studies used statistical theories or machine learning, e.g. neural network, genetic algorithm, or linear discriminant function [5]. For example, AMOSA used linear discriminant function to find TSSs [6]. Although these programs could identify the TSS, they were low in accuracy. Wang et al. (2008) reported that AMOSA still has accuracy problem [6]. Won et al. (2008) compared many programs and found that there is no best program. Therefore, they combine the results of existing predictors by using three ensemble methods—the majority voting, the weighted voting and the Bayesian approach [5].

Besides employing promoter sequence, we hypothesized that the distribution of adenine (A) could be utilized to identify TSS based on the work of Aporntewan et al. (2013). They found that A-repeats do not randomly distribute around TSS, but they have an interesting pattern [7]. Repeated sequences are sequences that exist in many copies. A-repeat, an adenine base that occurs consecutively, is one example of repeated sequences (non-consecutive A is called A-singletons). Their results show that, in human genome, the distribution of A- and T-repeats is uniform in the upstream and downstream of TSS but sharply drops at TSS. The normalized numbers of A- and T-repeats are the lowest exactly at the bin that contains TSS. In addition, the distribution of A- and T-repeats in the upstream and downstream of TSS is not symmetrical. We extended the findings of Aporntewan et al. (2013) by studying the distribution of A/T singletons and observing that the specific pattern of distribution should help identify TSS. Therefore, it is of our interest to investigate at what level this feature solely can detect TSSs. We developed an algorithm to detect the specific pattern of A/T-singletons distribution and assessed the performance by comparing its sensitivity, specificity and accuracy with Promoter2.0 [8].

II. MATERIALS AND METHODS

A. Human Whole Genome Data

We downloaded human genome sequence build 37 (hg 19) from UCSC Genome Browser database for studying the characteristics of A- and T-singletons around TSS regions and non-TSS regions.

We tested our method on human genome build 38 (hg 38) downloaded from UCSC Genome Browser as well. In

978-1-4673-7825-3/15/$31.00 ©2015 IEEE

addition, we downloaded human TSSs build 38 from UCSC database via Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables?org=human).

B. The Distribution of A- and T-singletons/repeats

We investigated whether the distribution of A-singletons and T-singletons (length = 1) around TSS is similar to the distribution of A- and T-repeats (length = 5 to 30) as reported by Aporntewan et al. (2013). We used the same bin structure as in the previous study, namely, 10,000 bases upstream and downstream of each TSS (20,000 bases for each TSS) were obtained. The 20,000-base-long sequences were divided into 25 bins; each bin was of size 800 bases. The bin structure is shown in Fig. 1. However, for each bin and for each sequence, we simply count the number of A's or T's as shown in Fig. 2. The number of A in each bin is then normalized by the number of genes to obtain the distribution of A-singletons and do the same thing for T. In addition, the distribution of A/T-singletons is also studied. For A/T-singletons, the number of both A and T are counted in each bin and then normalized by the number of genes.

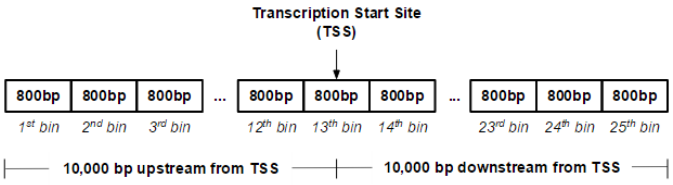

Fig. 1. Bin structure around TSS. There are 25 bins. Each bin contains 800 base pair (bp). The TSS is centered in the 13th bin.

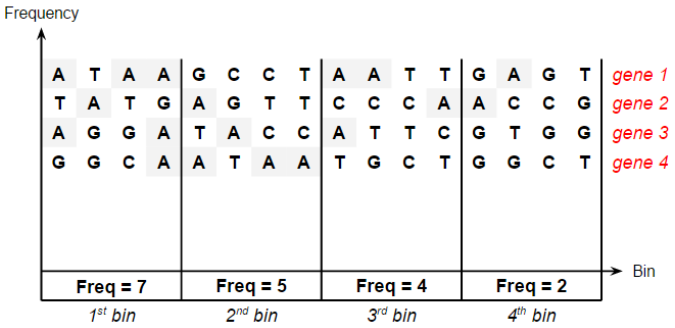

Fig. 2. Illustration of A-singletons counting. Eight bases upstream and downstream of TSSs of 4 genes were made up. The 16 bases were divided into four bins, each with size of four bases. Frequency (Freq) of A-singletons are shown at the bottom for each bin. The mean number of A-repeats (the normalized value) are 7/4 = 1.75, 5/4 = 1.25, 4/4 = 1.00, and 2/4 = 0.50 from the leftmost bin to the rightmost bin.

C. The Proposed Algorithm

The algorithm consists of two major steps: shuffled-difference testing and outliers testing. We applied this algorithm to the whole human genome in a sliding window fashion. We analyzed 20,000 bases at a time by dividing them into 25 bins, each bin of length 800 bases. A/T-singletons were counted in each bin.

1) Shuffled-difference testing

In this step, we will calculate a score for the possibility of containing TSS in the input sequence. Initially, we set score to be zero.

We begin by shuffling the 25 bin numbers to get a new A/T-singletons count. For example, if the first, second, and third bin counts are 100, 120, and 90, respectively, and the shuffle numbers are 3, 1 and 2, after shuffling, the counts become 120, 90, and 100, respectively.

Then, we calculate the difference between the original counts and the new counts in each bin. The rationale is that if the input sequence contains no TSS across all bins. Then, the difference values would be small across all bins. However, if there are, for example, TSS in the 13th bin, the difference would be large in at least two bins. That is, range and standard deviation (SD) of the difference values would be large if TSS is present in the input sequence, otherwise range and SD would be small if TSS is absent from the input sequence.

Next, range and SD of the difference value are calculated. To obtain the appropriate cutoff value for range and SD, we calculate range and SD of the difference values in non-TSS region in the genome. Based on this information, we arbitrarily set the cutoff value for range at 258 and 62 for SD. Therefore, if range is greater than or equal to 258 and SD is greater than or equal to 62 for the input sequence, we increase the score by one.

The process of shuffling, calculating range and SD for the difference values and setting score is repeated for 10 times. If the final score is greater than a threshold (we use 7), then this region is passed to the second step. Otherwise, we mark that TSS is not present in any bin.

2) Outliers Testing

In this step, the region that passed step 2 is checked for outliers. The rationale is that the A/T-singletons count of the bin that contains TSS would be very different from others and can be considered as an outlier. We use the interquartile range (IQR) to check for outlier.

$$Outlier < First\ quartile - 1.5 \times IQR, \qquad (1)$$

where $IQR = Third\ quartile - First\ quartile$

If there are 1–3 outliers occurred in consecutive bin and bin 13 is in one of them, we reported that TSS resided in these bins that contained outliers. However, if there are 1–3 outliers occurred in consecutive bin but bin 13 is not in one of them, we slide the window to reposition the outlier bins at the center and begin step 1 again. Otherwise, we mark that this 20,000-base region contains no TSS.

D. Performance Evaluations

We compared our method to Promoter2.0, which uses neural network and genetic algorithm to identify TSS [9]. We regarded TSSs data obtained from UCSC database as the truth. We applied our algorithm and Promoter2.0 to the whole

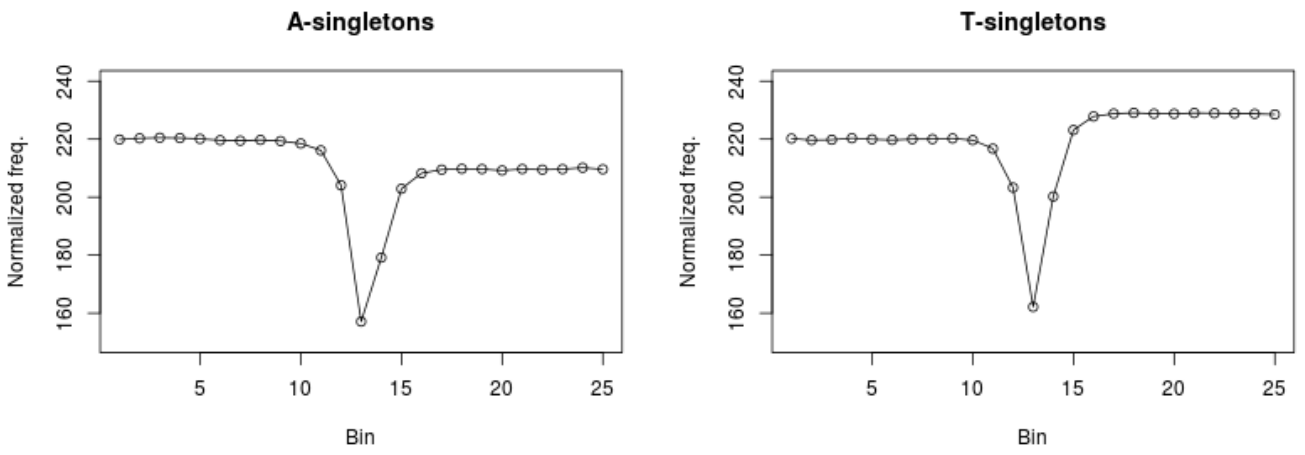

Fig. 3. Distribution of A- and T- singletons in regions with TSS in human. Left: A-singletons distribution. Right: T-singletons distribution. The X-axis represents bin, where TSS resides in the 13th bin. The Y-axis represents the number of singletons normalized by the number of gene.

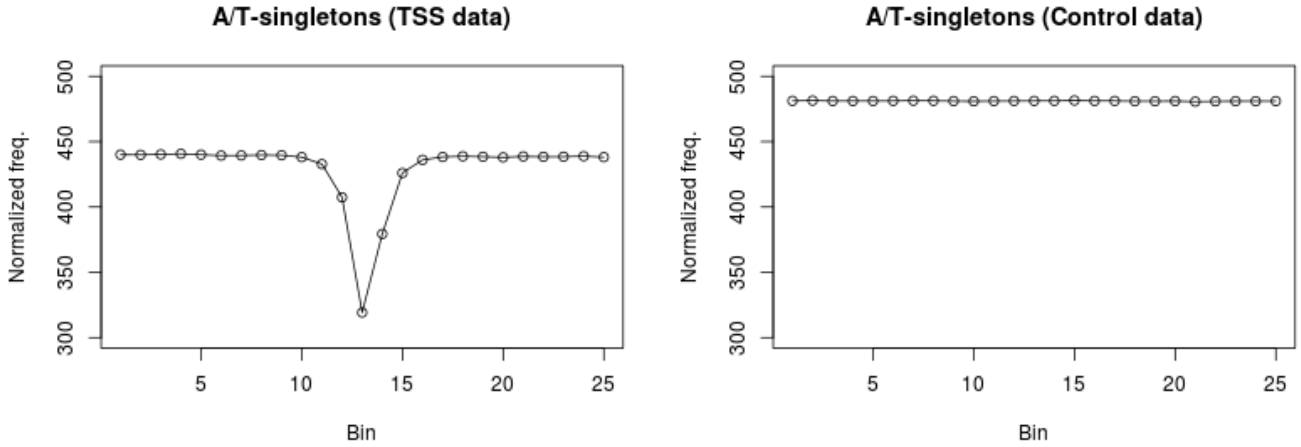

Fig. 4. Distribution of A/T-singletons in human. Left: A/T-singletons distribution in regions with TSS residing in the 13th bin. Right: A/T-singletons distribution in regions without TSS (control). The X-axis represents bin. The Y-axis represents the number of A/T-singletons normalized by the number of gene.

genome in a sliding window fashion as mentioned. The results in each bin are then compared to the truth. This allows us to calculate true positive, true negative, false positive and false negative as shown in the Table I. We evaluate our methods using three criteria—sensitivity, specificity and accuracy.

TABLE I. CONFUSION TABLE

	TSS presence according to UCSC database	**TSS absence according to UCSC database**
TSS presence according to an algorithm	True positive	False positive
TSS absence according to an algorithm	False negative	True negative

III. RESULTS

C. *The Characteristics of A/T-singletons*

We investigated A- and T-singletons distribution and found that they have similar pattern as reported in Aporntewan et al. (2013). That is, V-shaped pattern around TSS still remains and the number of A- or T-singletons is lowest at the 13th bin, which contains TSS. The distributions of A- and T-singletons are also asymmetrical (Fig. 3). A-singletons are more enriched in upstream region more than downstream, while T-singletons are more enriched in downstream region than upstream. Using *t*-test, the difference of singletons between upstream and downstream region is significant with P-value $= 6.32 \times 10^{-19}$ for A-singletons and P-value $= 6.46 \times 10^{-21}$ for T-singletons at the 95% significance level.

Next, we were curious about the distribution of A/T-singletons. We count both A and T in each bin and normalized by the number of gene as before. As expected, the V-shaped

TABLE II. COMPARISONS OF SENSITIVITY, SPECIFICITY AND ACCURACY BETWEEN A/T-SINGLETON AND PROMOTER2.0

Chromosome	Sensitivity (%)		Specificity (%)		Accuracy (%)	
	A/T-singleton	*Promoter2.0*	*A/T-singleton*	*Promoter2.0*	*A/T-singleton*	*Promoter2.0*
1	7.05	18.09	99.65	82.11	98.38	81.24
2	7.96	17.32	99.65	80.01	98.72	79.38
3	8.04	17.21	99.70	80.14	98.57	79.36
4	8.34	19.25	99.59	79.86	98.69	79.27
5	8.45	19.91	99.70	79.67	98.63	78.97
6	7.07	20.32	99.64	80.15	98.33	79.30
7	6.89	18.86	99.67	81.09	98.37	80.21
8	8.25	17.81	99.67	79.92	98.68	79.24
9	7.29	18.58	99.64	83.18	98.66	80.21
10	8.51	18.09	99.68	80.44	98.50	79.24
11	4.89	18.08	99.65	81.54	98.21	80.58
12	6.15	18.61	99.64	79.68	98.20	78.74
13	9.16	20.63	99.67	82.61	98.93	82.10
14	6.38	17.85	99.71	83.38	98.64	82.63
15	6.01	16.41	99.71	83.95	98.13	82.81
16	4.46	15.96	99.74	83.17	98.27	82.13
17	4.44	14.04	99.69	82.62	97.17	80.80
18	9.12	20.27	99.66	77.38	98.99	76.96
19	2.94	14.78	99.70	83.80	96.60	81.59
20	6.38	15.88	99.67	81.25	98.32	80.31
21	5.19	20.28	99.71	83.13	98.68	82.44
22	4.41	14.26	99.82	87.23	98.08	85.90
X	6.88	20.10	99.67	79.06	98.60	78.38
Y	3.62	18.84	99.88	91.43	99.51	91.15

pattern still holds and bin that contains TSS has the lowest frequency as before (Fig. 4, left). In addition, we studied A/T-singletons distribution in regions without TSS (we will refer to these regions as control). We found that the distribution is flat across 25 bins (Fig. 4, right). The results clearly showed that the distribution of A/T-singletons is obviously distinct between region with TSS and without TSS.

D. Performance Evaluations

We compared our algorithm to Promoter2.0 and we refer to our method as "A/T-singleton." The criteria used to evaluate our method are sensitivity, specificity and accuracy. We calculated these values for each chromosome (Table II). Our method has sensitivity, specificity and accuracy across all chromosomes in average 6.5%, 99% and 98% respectively. Promoter2.0 has specificity and accuracy in average 18%, 82% and 81%.

IV. DISCUSSION

We developed a new algorithm, "A/T-singleton," based on a non-random distribution of A/T-singletons around TSS, which contained two steps: shuffled-difference testing and outliers testing. Then, we compared our algorithm with Promoter2.0 using the data from UCSC Genome Browser and regarded this data as the truth. The result showed that our method has very high specificity and accuracy (close to 100%), but low sensitivity, which is expected.

Our algorithm is done in two major steps. The first step, shuffled-difference testing, is designed to filter out regions that are very likely to be absence of TSS. However, the first step alone is not adequate to filter out regions with highly fluctuate patterns, since these regions are more likely to pass the first step. Therefore, we employ the second step, outlier testing. The goal of this step is to detect regions that have a sharp drop at the 13th bin pattern (we allowed 1–3 outliers in consecutive bins). Thus, regions with fluctuate pattern will not be reported as TSS containing region. We are aware that we can lose sensitivity with such a strict constraint in the second step, since there are a number of genes that have V-shaped pattern occurring more than once non-consecutively in 20,000 bases or some TSSs are nearby within 20,000 bases. We traded off sensitivity for specificity as seen in the results; our method has very high specificity (99%).

Although we have lower sensitivity than Promoter2.0, this is not surprising. Promoter2.0 used a much more sophisticated method to detect TSS, while we use only the distribution of A/T-singletons. However, the sensitivity of Promoter2.0 was not tremendously higher than our method. Our findings suggested that existing methods are likely to be improved if

they incorporate the knowledge of A/T-distribution around TSS.

One weak point in this algorithm is that the size of the bin is very big. It requires at least 20,000 bases to detect the presence of TSS and it cannot tell the exact position of TSS. One way to improve this is to lower the bin size. This could be the future works. However, our current algorithm serves to answer the question we had, that is, investigating if the knowledge of A/T distribution can help detect TSS.

Another interesting attribute is C/G-singletons distribution. We expect that the mean C/G-singletons would rise at the 13th bin, opposite of A/T-singletons, as there are 4 bases in the genome. Using both A/T-singletons and C/G-singletons might improve sensitivity.

V. CONCLUSION

We investigated whether the knowledge of A/T-singletons distribution can help detect TSS by developing an algorithm to detect the pattern of A/T-singletons distribution. Our findings demonstrated that A/T-singletons distribution can detect TSS at some level. However, using this knowledge solely is not sufficient to detect all TSSs. Incorporating this knowledge should help improve the efficiency of other TSS detecting methods.

ACKNOWLEDGMENTS

This work was supported by the Honors program, Chulalongkorn University. The authors thank The Center for Biological Sequence Analysis at the Technical University of Denmark for providing Promoter2.0. The authors also would like to thank the anonymous reviewers for their valuable time and comments.

REFERENCES

[1] Scitable by Nature Education. *Transcription / DNA transcription* [Online]. Available: http://www.nature.com/scitable/definition/transcription-87 [Accessed: November 2014]

[2] S. Clancy, "DNA Transcription," *Nature Education*, vol. 1, no. 1, pp. 41, 2008.

[3] Scitable by Nature Education. *Promoter* [Online]. Available: http://www.nature.com/scitable/definition/promoter-259 [Accessed: November 2014]

[4] Scitable by Nature Education. *TATA box* [Online]. Available: http://www.nature.com/scitable/definition/tata-box-313 [Accessed: November 2014]

[5] H.H. Won, M.J. Kim, S. Kim, and J.W. Kim, "EnsemPro: An ensemble approach to predicting transcription start sites in human genomic DNA sequences," *Genomics*, vol. 91, no. 3, pp. 259–266, March 2008.

[6] X. Wang, S. Bandyopadhya, Z. Xuan, X. Zhao, M.Q. Zhang, and X. Zhang, "Prediction of transcription start sites based on feature selection using AMOSA," *Comput. Syst. Bioinformatics Conf.*, vol. 6, pp. 183–193, August 2007.

[7] C. Aporntewan, P. Pin-on, N. Chaiyaratana, M. Pongpanich, V. Boonyaratanakornkit, and A. Mutirangura, "Upstream mononucleotide A-repeats play a cis-regulatory role in mammals through the DICER1 and Ago proteins," *Nucleic Acids Res*, vol. 41, no. 19, pp. 8872–8885, October 2013.

[8] S. Knudsen, "Promoter2.0: for the recognition of PolII promoter sequences," *Bioinformatics*, vol. 15, no. 5, pp. 356–361, February 1999.