Computational and Structural Biotechnology Journal 20 (2022) 3347–3350

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Fasim-LongTarget enables fast and accurate genome-wide lncRNA/DNA binding prediction

Yujian Wen [a,1], Yijin Wu [a,1], Baoyan Xu [a,1], Jie Lin [a,*], Hao Zhu [a,b,*]

[a] *Bioinformatics Section, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China*
[b] *Guangdong-Hong Kong-Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence, Southern Medical University, Guangzhou 510515, China*

ARTICLE INFO

Article history:
Received 7 March 2022
Received in revised form 9 June 2022
Accepted 9 June 2022
Available online 18 June 2022

Keywords:
TDF
LongTarget
RNA/DNA binding
Triplex
lncRNA
Epigenetic regulation

ABSTRACT

Many long noncoding RNAs (lncRNAs) can bind to DNA sequences proximal and distal to abundant genes, thereby regulating gene expression by recruiting epigenomic modification enzymes to binding sites. Because a lncRNA's target genes scattering in a genome have correlated functions, epigenetic analyses should often be genome-wide on both genome and transcriptome levels. Multiple tools have been developed for predicting lncRNA/DNA binding, but fast and accurate genome-wide prediction remains a challenge. Here we report Fasim-LongTarget (a revised version of LongTarget), compare its performance with TDF and LongTarget using the experimental data of the lncRNA MEG3, NEAT1, and MALAT1, and describe a case of genome-wide prediction. Fasim-LongTarget is as accurate as LongTarget and more accurate than TDF and is 200 times faster than LongTarget, making accurate genome-wide prediction feasible. The code is available on the Github website (https://github.com/LongTarget/Fasim-LongTarget), and the online service is available on the LongTarget website (https://lncRNA.smu.edu.cn).

1. Introduction

Many long noncoding RNAs (lncRNAs) epigenetically regulate gene expression by binding to DNA sequences and recruiting epigenomic modification enzymes to binding sites. LncRNAs and these enzymes show enriched distributions at binding sites [1], indicating that a DNA binding site (DBS) may host multiple lncRNAs. Following specific base-pairing rules, a lncRNA binds to a duplex DNA sequence by forming a triplex that comprises triplex-forming oligonucleotides (TFO) in the lncRNA and a triplex-targeting site (TTS) in the DNA sequence. Thus, overlapping TTSs indicate a DBS, and overlapping TFOs indicate a DNA binding domain (DBD). At or near DBSs are lncRNAs' target genes. Abundant studies have revealed that lncRNA's target genes have correlated functions (e.g., multiple imprinted genes regulated by the lncRNA H19 control embryonic growth) [2] and that lncRNAs regulate target genes genome-wide (e.g., the lncRNA XIST regulates the inactivation of nearly all genes on the X chromosome in female mammals) [3].

The specific base-pairing rules (i.e., Hoogsteen and reverse Hoogsteen rules) [4] make triplexes, TFOs, TTSs, DBDs, and DBSs computationally predictable. Largely two kinds of methods have been developed to predict triplexes. Upon canonical Hoogsteen/reverse Hoogsteen base-pairing rules, *Triplexator* uses substring search to find triplexes (which are consecutive paired nucleotides with a small error rate such as <=2 consecutive mismatches) [5]. The nature of the substring search makes the triplexes very short (16–20 bp) and prone to occur, making it hard to judge whether triplexes form a DBD/DBS. To help predict DBS/DBD, *TDF* (Triplex Domain Finder), which runs upon *Triplexator* or *TRIPLEXES* (*TRIPLEXES* performs substring search faster than *Triplexator* does), was developed. *TDF* statistically tests if several triplexes form a DBD/DBS [6].

LongTarget took another approach. It first translates the DNA sequence into RNA sequences upon 24 Hoogsteen/reverse Hoogsteen rulesets, then uses a variant of the Smith-Waterman algorithm to identify all local alignments in each lncRNA/translated RNA pair [7]. This local alignment can flexibly identify very long triplexes but is more time-consuming. In addition, because a lncRNA can be parallel or anti-parallel to a DNA sequence, a pair of lncRNA/DNA sequences generate 48 lncRNA/translated RNA pairs and demands 48 local alignments.

* Corresponding authors at: Bioinformatics Section, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China (J. Lin).
E-mail addresses: JL.linjie@outlook.com (J. Lin), zhuhao@smu.edu.cn (H. Zhu).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.csbj.2022.06.017

The two kinds of methods have pros and cons. By using substring search to identify triplexes, *Triplexator*/*TRIPLEXES* + *TDF* is fast but the triplexes are short (14–20 bp) and less overlapped; thus, *TDF* spends extra time on statistically testing whether triplexes likely form DBD/DBS. By using local alignment to identify triplexes, *LongTarget* is slow but the triplexes are long (>60 bp); thus, no statistical test is needed because long triplexes often overlap at a DBS. Although *LongTarget* is integrated into a platform supported by multiple genomes and a lncRNA database [8], true genome-wide prediction is infeasible due to time consumption. So far, no methods has be satisfactorily used for genome-wide lncRNA/DNA binding.

Experimental studies have gone from single genes to gene sets and further to the whole genome, and abundant lncRNAs have been identified in mammalian genomes. The two factors drive genome-wide analysis of lncRNA-mediated epigenetic regulation on the genome and transcriptome levels (e.g., the regulatory relationship between differentially expressed protein-coding genes and lncRNA genes in cancer cells). We report *Fasim-LongTarget* (*abbr*. *Fasim*), which is about 200 times faster than yet almost equally powerful as *LongTarget*. First, we introduce the revised alignment algorithm; then, we use three experimentally generated lncRNA/DNA binding datasets to evaluate the performance of *TDF*, *Fasim*, and *LongTarget*; finally, we describe a case of genome-wide prediction.

2. The *Fasim* algorithm

The Smith-Waterman algorithm has been revised in two ways. On the one hand, Waterman and Eggert extended the algorithm by outputting multiple non-intersecting local alignments [9], and Huang and Miller greatly reduced the space and time consumption of the Waterman-Eggert algorithm (the *SIM* program) [10]. On the other hand, Farrar used the Single-Instruction Multiple-Data (SIMD) instruction to parallelize the Smith-Waterman algorithm (Striped Smith-Waterman) [11], and Zhao et al. developed a C/C++ library for the SIMD Smith-Waterman algorithm [12]. *LongTarget* calls *SIM* to identify multiple local alignments (triplexes) between a translated DNA sequence and a lncRNA sequence [7–8]. Upon the five works, *Fasim* is developed. In the command line of *Fasim* (i.e., *fasim*(*Ms, Ns, Gs, Qs, triplex length, DNA sequence list, lncRNA sequence*)), *Ms*, *Ns*, *Gs*, *Qs* are the scores of match, mismatch, gap open, and gap extension, and *triplex length* is the minimal length of triplexes. *DNA sequence list* and *lncRNA sequence* specify the files of DNA and lncRNA sequences. In accordance, variables *Mn*, *Nn*, *Gn*, *Qn* are the numbers of match, mismatch, gap open, and gap extension. *Fasim* outputs two files; one contains the distribution of DBSs and can be uploaded to the UCSC Genome Browser as a custom track, and the other contains detailed information of TTSs, TFOs, DBSs, and DBDs.

First, given a translated DNA sequence and a lncRNA sequence, *Fasim* uses SIMD to compute the scoring matrix and identifies and outputs the best alignment using the standard Smith-Waterman algorithm. Second, *Fasim* uses the scoring matrix to identify and output the remaining local alignments whose number is determined by the parameter *Threshold*. For example, if the lncRNA sequence = CGATTGTTGT, the translated DNA sequence = ACGCGATGAATTGGACTT, and the *Threshold* = 0.8 (i.e., 80% of the best alignment' score), then *Threshold* would be 20 which determines the number of the remaining local alignments. Our revised scoring matrix adds three rows: *max*, *tmpscore*, and *finalscore*, which store the maximal score of each column, the maximal scores that are >=*Threshold*, and the maximal scores that are local maximum (Fig. 1A). Third, upon the ordered values of *finalscore*, *Fasim* first identifies the ending position of a local alignment, then uses the following computation to determine the starting position of the local alignment. For an ending position *Pi*, the starting position is at *Pi*-(*Mn* + *Nn* + *Gn* + *Qn*). Because the $identity = \frac{M_n}{M_n+N_n+G_n+Q_n}$ and $score = M_n * M_s - N_n * N_s - G_n * (G_s + Q_s) - Q_n * Q_s$, in this example if Gn<=3, Ms = 5, Ns = 4, Gs = 8, Qs = 4, then $M_n + N_n + G_n + Q_n = \frac{finalscore-(N_s-G_s-Q_s)*G_n-(N_s-Q_s)*Q_n}{(N_s+M_s)*identity-N_s} = \frac{finalscore+8*G_n}{9*identity-4} \leq \frac{finalscore+24}{9*identity-4}$. *Fasim* assumes $identity \in (0.6, 1)$ and *step* = 0.02 and uses *identity* = *identity* + *step* to try *identity* values. For each *identity* value (which determines a *Pi*-(*Mn* + *Nn* + *Gn* + *Qn*)), an alignment is examined using the sequences between *Pi*-(*Mn* + *Nn* + *Gn* + *Qn*) and *Pi*, and a local alignment is found if the alignment score equals the *finalscore* value. Finally, by examining all values in the *finalscore* row, *Fasim* identifies and reports all non-intersecting local alignments without revising the scoring matrix. Two key revisions make *Fasim* faster than *SIM*. (a) *Fasim* computes the large scoring matrix for a pair of sequences using the SIMD instructions only once. (b) *SIM* finds multiple local alignments by revising the scoring matrix; by using extra rows to store critical information, *Fasim* finds multiple local alignments without revising the large scoring matrix. These revisions make *Fasim* report fewer alignments than *SIM*. For example, if a short alignment with a higher score lies within a large alignment with a lower score, *SIM* reports both, but *Fasim* reports only the short one. The reduction of time consumption depends on sequence length. If the lncRNA and translated RNA are long, identifying the starting positions of local alignments by testing multiple short alignments (i.e., potential triplexes) is much faster than identifying the starting and ending positions of local alignments by revising and checking the whole scoring matrix.

3. Performance evaluation

We used the experimentally detected DNA binding regions (called peaks) of three lncRNAs to evaluate *TDF*, *LongTarget*, and *Fasim*. MEG3 (ENST00000451743), NEAT1 (ENST00000501122), and MALAT1 (ENST00000534336) have 532, 3692, and 670 peaks in three cell lines, ranging from 500 to 1500 bp, 500–1500 bp, and about 10 Kb, respectively [13–14]. We used these peaks as the target DNA sequences and used their 532*2, 3692*2, and 670*2 neighboring sequences (1000 bp for MEG3 and NEAT1, 15000 bp for MALAT1) as the negative controls.

We let *TDF* call *TRIPLEXES* to predict triplexes, with the parameters *triplex length*>=14 for MEG3 but >=16 for NEAT1 and MALAT1 (as the original authors did), *maximum of mismatch*<=3, *consecutive errors*<=2, and *repeat time* = 100. We ran *LongTarget* and *Fasim* with the default parameters (*triplex length*>=60 and *identity*>=0.6). *Fasim* and *TDF* predicted similar numbers of TTSs per DBS. *Fasim* reported fewer TTSs per DBS (and per peak) than *LongTarget,* due to not revising the scoring matrix; however, the length and number of DBD/DBS predicted by *Fasim* are only slightly shorter and fewer than the length and number of DBD/DBS predicted by *LongTarget*. These indicate that *Fasim* and *LongTarget* can equally well predict long triplexes, DBDs, and DBSs. Although *TDF* predicts DBSs/DBDs upon statistically testing TTSs/TFOs, the length of DBDs/DBSs predicted by *TDF* is significantly shorter than the length of DBDs/DBSs predicted by *LongTarget* and by *Fasim* (Supplementary Fig. 1).

Next, we examined the time consumption of the three methods. *Fasim* is about 200–300 times faster than *LongTarget* and even much faster than *TDF* upon these datasets (Fig. 1B). Finally, we used the Precision-Recall curves (PRC) and Receiver Operating Characteristic (ROC) curves to evaluate the power of the three methods. To this end, we defined several quantitative measures. *TTSscore* is the scores of triplexes reported by *TRIPLEXES* and scores of local alignments reported by *LongTarget* and *Fasim*. *DBSscore* is computed as $\sum(TTS_1score,\ TTS_2score,\ldots,TTS_kscore)$, where these

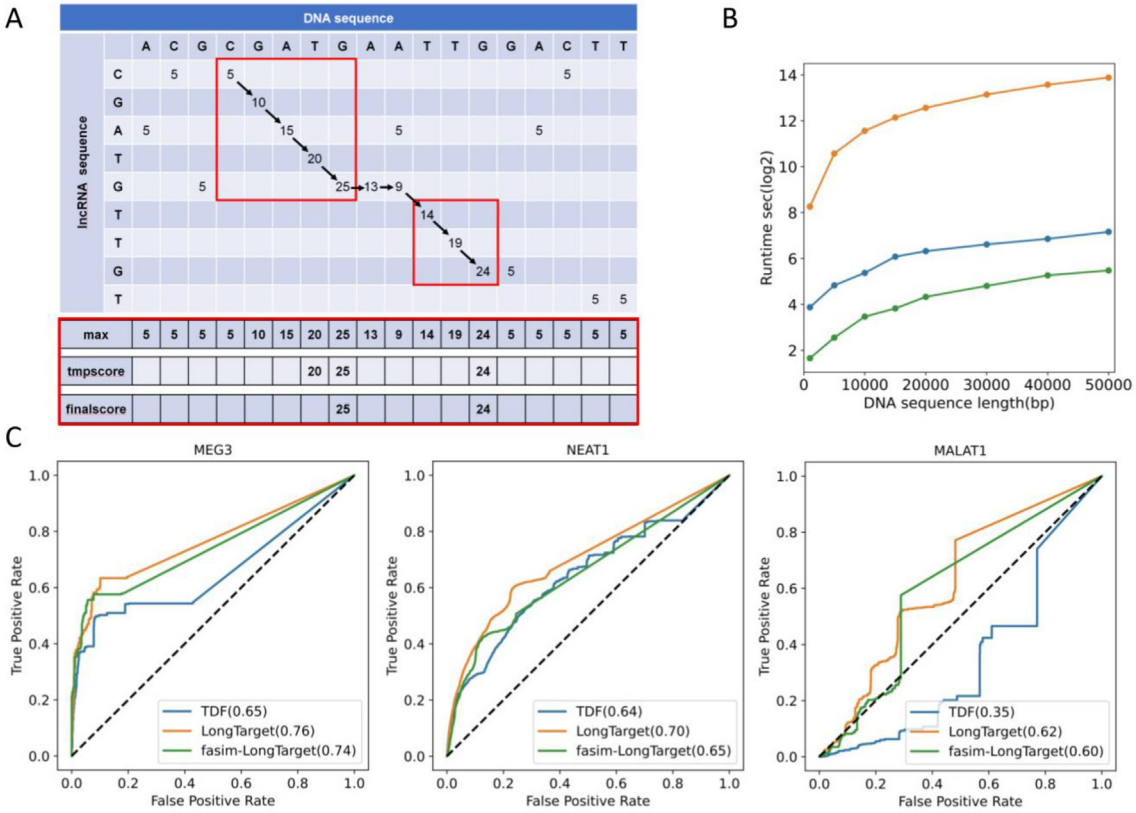

Fig. 1. (A) The scoring matrix and extra rows *Fasim* uses when identifying and outputting multiple local alignments. The scores 25 and 24 in the 8th and 13th columns are local maximum thus are the two local alignments' ending positions. (B) The time consumption (seconds, the log2 form) of *LongTarget*, *TDF*, and *Fasim* (from top to bottom indicated by orange, blue, and green lines). (C) The ROC curves of *TDF*, *LongTarget*, and *Fasim* (generated upon the ranking of *NPeakscores* that indicates the triplex signal in the experimentally detected regions). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

TTSs overlap within the DBS. *NPeakscore* is the normalized scores of peaks, computed as $\sum(DBS_1score,\ DBS_2score,\ldots,DBS_kscore)/(\text{peak length})$. *TTSscore*, *DBSscore*, and *NPeakscore* quantify the strength of each TTS, DBS, and peak (a peak may contain multiple DBSs). Upon these quantitative measures and PRC and ROC curves, *Fasim* slightly underperforms *LongTarget* and clearly outperforms *TDF* (Fig. 1C; Supplementary Fig. 1) (because the peaks' neighboring regions were used as the negative controls, the PRC and ROC curves of MEG3 are somewhat different from those where random regions were used as the negative controls) [6].

4. Application

Early studies revealed that H19 and Airn regulate the imprinted expression of *IGF2* and *IGF2R* to control the embryonic growth of mammals. Later studies revealed that H19 is the master regulator of genomic imprinting by regulating many genes. Thus, predicting H19's DBSs genome-wide is an interesting application. *Fasim* took 817 h (times of all cores, using a Xeon(R) E7-4830 v3, 2.10 GHz) to predict H19's DBSs in the human genome hg38 (22 autosomal and 2 sex chromosomes). Thus, by using modern multi-core CPUs, analyzing lncRNA-mediated epigenetic regulation genome-wide is feasible (e.g., predicting target genes of H19 and Airn in human and mouse genomes and analyzing species-specificity of genomic imprinting).

Using chromosome 21 and chromosomal 22 (which has more 'N' than chromosome 21), we further compared *Fasim*, *TDF*, and *LongTarget* (with parameters mentioned above). *Fasim* took 11.38/11.16 h, and *TDF* took 17.90/20.46 h to predict DBSs on chromosome 21/22, respectively (*LongTarget* would take 34 days to finish chromosome 21). Upon H19 and chromosome 21/22, *Fasim* identified DBD1 (the top-ranked DBD) at 2387–2451 bp (the mean length of DBSs is 82/86 bp), *LongTarget* identified DBD1 at 2386–2451 bp (upon the finished part), but *TDF* had no DBD passed the statistical test. By manual checking, *TDF* identified the best DBD at 2368–2450 bp, but the mean length of DBSs is 20 bp (TTSs were not integrated into DBSs). Of note, we previously used *LongTarget* to analyze H19 and genomic imprinting in mammals. The predicted DBSs in annotated imprinted genes agree with experimental reports, and the predicted DBD1 was at 2366–2465 bp [15]. That *LongTarget* and *Fasim* predicted the same DBD1 upon imprinted genes and chromosome 21/22, respectively, suggests the reliability of the prediction.

5. Brief remarks

Triplexes identified by the two kinds of methods are quite different. Those identified by substring search are shorter and have a higher identity than those identified by local alignment. Which kind of triplexes is biologically more reasonable may await more experimental investigations. Although a higher identity may indicate higher stability, identity alone may not critically determine lncRNA/DNA binding [7]. Triplexes with high identity are inevitably short, and it can be time-consuming to statistically test whether a set of short triplexes, which are less prone to overlap, form a DBS. On the other hand, triplexes identified by local alignment are often long because mismatches are more tolerable, and these triplexes are prone to overlap at DBSs. Long TTSs and DBSs are unlikely to be obtained by chance and are strong signs of true DNA binding sites. Although *Fasim* reports fewer triplexes than

LongTarget, the number and length of DBSs predicted by *LongTarget* and *Fasim* are similar, ensuring that *Fasim* has comparable power with *LongTarget*.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Department of Science and Technology of Guangdong Province (2020A1515010803) and the National Natural Science Foundation of China (31771456).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.017.

References

[1] Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell 2010;38:662–74.
[2] Alipoor B, Parvar SN, Sabati Z, Ghaedi H, Ghasemi H. An updated review of the H19 lncRNA in human cancer: molecular mechanism and diagnostic and therapeutic importance. Mol Biol Rep 2020;47:6357–74.
[3] Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B. Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 2002;36:233–78.
[4] Abu Almakarem AS, Petrov AI, Stombaugh J, Zirbel CL, Leontis NB. Comprehensive survey and geometric classification of base triples in RNA structures. Nucl Acids Res 2012;40:1407–23.
[5] Buske FA, Bauer DC, Mattick JS, Bailey TL. Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome Res 2012;22:1372–81.
[6] Kuo CC, Hanzelmann S, Senturk Cetin N, Frank S, Zajzon B, Derks JP, et al. Detection of RNA-DNA binding sites in long noncoding RNAs. Nucl Acids Res 2019;47:e32.
[7] He S, Zhang H, Liu H, Zhu H. LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. Bioinformatics 2015;31:178–86.
[8] Lin J, Wen Y, He S, Yang X, Zhang H, Zhu H. Pipelines for cross-species and genome-wide prediction of long noncoding RNA binding. Nat Protoc 2019;14:795–818.
[9] Waterman MS, Eggert M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. J Mol Biol 1987;197:723–8.
[10] Huang X-Q, Miller W. A time-efficient linear-space local similarity algorithm. Adv Appl Math 1991;12:337–57.
[11] Farrar M. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 2007;23:156–61.
[12] Zhao M, Lee WP, Garrison EP, Marth GT. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. PLoS ONE 2013;8:e82138.
[13] Mondal T, Subhash S, Vaid R, Enroth S, Uday S, Reinius B, et al. MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures. Nat Commun 2015;6:7743.
[14] West JA, Davis CP, Sunwoo H, Simon MD, Sadreyev RI, Wang PI, et al. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol Cell 2014;55:791–802.
[15] Liu H, Shang X, Zhu H. LncRNA/DNA binding analysis reveals losses and gains and lineage specificity of genomic imprinting in mammals. Bioinformatics 2017;33:1431–6.