UDC 579:252

SHAHMURADOV I.A.
Institute of Molecular Biology and Biotechnologies, ANAS,
Azerbaijan, AZ1073, Baku, Matbuat Ave., 2a, e-mail: ilhambaku@gmil.com, (+994-50)3842496

NOVEL TOOLS FOR THE PREDICTION OF PROMOTERS IN PLANTS AND BACTERIA

The promoter is a chromosome region that determines where the transcription of a particular DNA region is initiated. Promoter recognition is important in defining the transcription units responsible for specific pathways and gene regulation. Initiation of transcription is a dynamic partnership between RNA polymerase (RNAP) and promoter.

In nuclear genomes of eukaryote organisms, transcription process is conducted by multiple types of RNA polymerases. In particular, all protein genes and most noncoding RNA genes, as well as DNA regions of unknown functions are transcribed by RNA polymerase II. 30–50 % of all known promoters contain a TATA-box located from 40 to 18 bp upstream of the TSS. However, promoters of many large groups of genes (e.g. housekeeping genes) lack the TATA-box; the corresponding promoters are referred to as TATA-less promoters [1-3].

In contrast to eukaryotes, bacteria have a single form of the RNAP core enzyme [4]. However, this RNAP alone is not able to recognize and bind to promoters to initialize transcription. Different σ-factors are required that temporarily binds the RNAP core enzyme, determine the RNAP-promoter binding specificity and transcription start site (TSS), depending on nutritional or environmental conditions or developmental stage [5, 6].

Bacterial σ factors are classified into two families with distinct structure and function, termed as σ^{70} and σ^{54} in *Escherichia coli*. While most bacteria possess multiple members of the σ^{70} family, they contain a single representative of the σ^{54} family, which is involved in nitrogen metabolism. Cyanobacteria lack any σ^{54}-like factors [5, 7, 8].

Due to the development of advanced experimental technologies a great progress was made in analysis of gene regulatory sequences [9–11]. However, a detailed experimental exploration of transcripts is still a quite expensive and difficult procedure. Therefore, in addition to experimental efforts, accurate computational identification of putative promoter regions remains an important task of genomics and post-genomics studies.

Over the last decade various promoter prediction programs have been developed. Recent studies indicate that there is often no single TSS, but rather a whole transcription start region (TSR) with multiple TSSs [12, 13]. However, for genome annotation projects predicting TSRs spanning several hundred (from 250 up to 1000) nucleotides is less useful to identify a gene start point. For such tasks, finding the TSSs seems to be more informative.

To date, various computer programs aimed to predict plant promoters have been developed [14–18]. In particular, previously we developed the TSSP-TCM program that showed a quite high accuracy of TSS prediction in the test sequences with experimentally validated TSSs: 87.5 % and 84 % for TATA and TATA-less promoters, respectively.

The first attempt to predict bacterial promoters was by position weight matrices (PWM), which relied on the conservation of the -35 and the -10 elements for σ^{70}, combined with the distribution of the distance between them [19, 20]. Later, more accurate bacterial promoter prediction tools have been developed [21–28]. Despite these efforts, all these tools tend to produce many false positives or show poor sensitivity, particularly when they are applied to long sequences or whole genomes. Another restriction of these tools is that they are limited to the prediction of σ^{70} promoters in the model organism *E. coli*, and very rarely can extend to other bacterial species. Therefore, novel, more accurate and efficient tools are required for the computational recognition of different classes of promoters in a broader taxonomical scope.

In this paper, two new computer tools, TSSPlant for prediction of plant promoters for RNA polymerase II, and and bTSSfinder for predicting TSSs in *E. coli* and three cyanobacterial species are briefly described.

Materials and methods

12,467 TSSs assigned to the annotated protein coding genes of *Arabidopsis thaliana* and *Oryza sativa japonica* were obtained from the Plant Promoter Database (ppdb), version 2.0 [29, http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi]. Using the genome annotations of Arabidopsis (https://www.arabidopsis.org/) and rice (http://rapdb.dna.affrc.go.jp/) genomes, the promoter set of 251 bp sequences (200 bp upstream and 51 bp downstream of a TSS) was created. Besides, 567 plant promoter sequences of 251 bp lengths with experimentally validated TSS, including 50 rice and 106 Arabidopsis promoters, were selected from the current release of PlantProm DB (http://www.softberry.com/plantprom2016/). With merging ppdb and PlantProm DB sets, the final set of 12,948 plant promoters was created. Out of them, 11893 promoters (including 426 sequences from the PlantProm DB) were used as the learning set. For testing, we selected 2 sets: Set 1 of 1000 sequences from ppdb only, and Set 2 of 55 promoters only from PlantProm DB. Length of sequences in the Learning set and Test set 1 was 251 bp. Promoter sequences of the Set 2 were extended up to 1101 bp: 1,000 bp upstream of TSS and 101 bp downstream. As a negative dataset, Arabidopsis and rice genomic sequences composed of CDSs and introns were used.

Data on plant transcription factor binding sites (TFBSs) were obtained from the Regsite DB (www.softberry.com; Plant division) that contained 1,976 TFBSs.

Bacterial TSS/promoter sets were created for *E. coli* K12 MG1655 (RegulonDB, version v8.0), the non-marine cyanobacterium *Nostoc* sp. PCC 7120 [30], the freshwater cyanobacterium *S. elongatus* PCC 6301 [31] and the freshwater cyanobacterium *Synechocystis sp*. PCC 6803 [32]. The final TSS count was as follows: 1) *E. coli*: 1,544 for σ^{70}, 140 for σ^{38}, 237 for σ^{32}, 135 for σ^{28}, 412 for σ^{24}; 2) *Nostoc*: 11,386, 3) *S. elongatus*: 1,471 and 4) *Synechocystis*: 343.

Data on the bacterial TFBSs were obtained from three sources: 2,953 sites for *E. coli* from RegulonDB, 30 cyanobacterial sites from CollecTF (33), and 63 sites from the literature.

To compute Nucleotide Frequency Matrices (NFM) for the core-promoter elements (TATA-box, -35 box, -10 box, etc.), the Expectation Maximization (EM) algorithm [34] was utilized. It was applied also for the promoter classification.

To distinguish promoter and non-promoter sequences, we explored about 30 different sequence characteristics of positive and negative learning sets. Based on the values of Mahalanobis distances [35] of individual characteristics, we selected up to 21 features for different classes of plant [36] and bacterial [37] promoters.

To get recognition function (classifier) parameters giving the best separation of promoter and non-promoter sequences, separately for two promoter classes, we applied Neural Network (NN) technique [36, 37].

Results and discussion

TSSPlant: the plant promoter prediction tool

The program analyzes each sliding window of 251 bp over the query sequence, one nucleotide at a time, where position 201 is assigned to be classified as TSS or non-TSS. The classification is based on a threshold that was computed during the training. Predictions with the score higher than the threshold are marked as putative TSSs. The program performs additional filtering by discarding all but the highest-scoring TSS in intervals of a user-specified length (default is 300 bp). TSSPlant is available to download as a standalone program at http://www.cbrc.kaust.edu.sa/download/.

We tested our TSS finder on positive and negative sets of TATA and TATA-less promoter classes. For TATA class, we observed very high prediction accuracy, with sensitivity ≃99 %, specificity ≃98 %, F1 score ≃99 % and Matthews correlation coefficient (MCC) 0.97. In the case of TATA-less promoters, we also achieved quite a good performance (with sensitivity ≃82 %, specificity ≃97 %, F1 score ≃89 % and MCC 0.83).

We compared our TSSPlant predictor with tools available for on-line execution or downloading and run locally: NNPP [38], Proscan [39], EP3 [40] and TSSP (http://www.softberry.com/berry.phtml). First, we compared NNPP, Proscan, TSSP and TSSPlant tools on short (251 bp) sequences, randomly chosen from TATA and TATA-less test sets. For both TATA and TATA-less cases, the accuracy data clearly indicate that TSSPlant has significantly higher prediction accuracy (results not shown). Another tests were performed on long sequences of 1,100 bp regions of 55 plant protein coding genes with experimentally validated TSS collected in PlantProm DB (Table 1). TSSPlant produced the best accuracy (Sn ≃ 72 %, F1 ≃ 47 %), followed by TSSP (Sn = 63.6 %, F1 ≃ 41.2 %), NNPP (Sn = 51 %, F1 ≃ 31 %), EP3 (Sn ≃ 20 %, F1 ≃ 31 %) and Proscan (Sn ≃ 9 %, F1 ≃ 15 %).

bTSSfinder: the bacterial promoter prediction tool

The largest collection of experimentally validated promoters of *E. coli* in RegulonDB was classified into seven different sigma classes (σ^{70}, σ^{54}, σ^{38}, σ^{32}, σ^{28}, σ^{24} and σ^{19}). Unfortunately, no such classification exists for cyanobacterial promoters. Our preliminary comparison of *E. coli* and cyanobacterial promoters indicate that there is a level of conservation, based on which we used *E. coli* PWMs for the classification of cyanobacterial promoters. Based on the inter-phyla orthology (data not shown), we propose the classification for cyanobacterial promoters into five classes: σ^{A} (analogous to σ^{70}), σ^{C} (analogous to σ^{38}), σ^{F} (analogous to σ^{28}), σ^{G} (analogous to σ^{24}) and σ^{H} (analogous to σ^{32}).

Using a combination of features for each promoter class, we built 10 NN classifiers, one for each promoter class in *E. coli* and in cyanobacteria. Then, we implemented these models into the bTSSfinder program. The program slides a window of 251 bp over the query sequence, one nucleotide at a time (analogous to TSSPlant tool). bTSSfinder is available standalone and online at http://www.cbrc.kaust.edu.sa/btssfinder.

We tested bTSSfinder on sets for every promoter class in *E. coli* and cyanobacteria (Table 2). We observed good performance for all promoter classes in *E. coli* (251 bp, a single search window size). In the case of cyanobacteria, we observed the highest accuracy in σ^{A} promoters (F_1-score: 0.94).

Table 1. Comparison of four promoter prediction tools assessed on 1,100 bp region of 55 plant protein coding genes with experimentally validated TSS

Tool	Genes with ≥ 1 TSSpr	Total number of TSSpr	TP[1]	FP	FN	Sn,%	F1-score,%
TSSPlant	54	115	40	75	15	72.7	47.1
TSSP	45	105	35	80	20	63.6	41.2
NNPP	47	122	28	97	27	50.9	31.1
EP3[2]	16	16	11	5	44	20.0	31.0
Proscan[2]	10	10	5	5	50	9.1	15.4

Notes: [1]Prediction is considered true, if the distance between annotated TSS and predicted TSS (TSSpr) is 50 bp or less. [2]EP3 and Proscan programs perform a search for transcription start region (of 250 nt and 400 nt, respectively).

Table 2. Comparison of available promoter prediction programs assessed on the 1,100 bp upstream region of 200 *E. coli* σ^{70} promoters with experimentally validated TSSs

Promoter prediction tool	Genes with ≥1 TSSpr	Genes without any TSSpr	Genes with true TSSpr[1]	Total number of TSSpr	Sn, %	TSSpr density
bTTSSfinder	197	3	143	355	71.5	620
BPROM	200	0	130	569	65.0	386
NNPP2	175	25	109	500	54.5	440
PromPredict	74	126	0	149	0.0	1477[2]

Notes: TSSan: annotated TSS position 1001), TSSpr: predicted TSS. [1]Prediction is true, if distance between annotated and predicted TSSs is 50 bp or less. [2]But, no any true prediction.

We could only evaluate bTSSfinder against previously published promoter prediction tools for σ^{70} promoter class in *E. coli*. For fairness, we assessed all tools on a single testing dataset. The following promoter prediction tools were available for comparison: BPROM (27), NNPP2 (24), and PromPredict (26). All other promoter prediction tools that we checked were no longer available. Results of the comparison for short (251 bp) sequences clearly indicate that bTSSfinder has significantly higher prediction accuracy (data not shown). However, using short sequences to predict TSSs is not sufficient in evaluating the accuracy and efficiency (especially the real false positive rate) of a prediction tool. It should also be tested on longer sequences. We run the four programs on longer DNA sequences to search for putative σ^{70} TSSs in 200 test sequences of 1,101 bp from *E. coli*. As presented in Table 2, bTSSfinder produced the best performance (Sn ≃ 72 %, F1 ≃ 52 %), followed by BPROM (Sn = 65 %, F1 ≃ 34 %) and NNPP2 (Sn ≃ 54 %, F1 ≃ 33 %). Surprisingly, PromPredict failed to produce a single true positive prediction (Se = 0, F1 = 0).

This publication is based upon work supported by (1) the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Awards No URF/1/1976-02 and FCS/1/2448-01; (2) Grant EİF-2010-1(1)-40/27-3 from the Science Development Foundation under the President of the Republic of Azerbaijan; (3) Science and Technology Center of Ukraine, (STCU) Project #6154; (4) Azerbaijan National Academy of Sciences.

References

1. Solovyev V.V., Shahmuradov I.A., Salamov A.A. Identification of promoter regions and regulatory sites. In: Computational Biology of Transcription Factor Binding (Methods in Molecular Biology). Editor: Istvan Ladunga. Springer Science+Business Media, Humana Press, 2010, 674, Chapter 5. doi: 10.1007/978-1-60761-854-6_5.
2. Hernandez-Garcia C.M., Finer J.J. Identification and validation of promoters and cis-acting regulatory elements // Plant Science. – 2014. – V. 217–218. – P. 109–119.
3. Roy A.L., Singer D.S. Core promoters in transcription: old problem, new insights // Trends Biochem Sci. – 2015. – V. 40. – P. 165–171.
4. Schneider G.J., Hasekorn R. RNA polymerase subunit homology among cyanobacteria, other eubacteria and archaebacteria // J Bacteriol. – 1988. – V. 170. – P. 4136–4140.
5. Imamura S., Asayama M. Sigma factors for cyanobacterial transcription // Gene Regul Syst Bio. – 2009. – V. 3. – P. 65–87.
6. Ruff E.F., Record M.T., Jr., Artsimovitch I. Initial events in bacterial transcription initiation // Biomolecules. – 2015. – V. 5. – P. 1035–1062.
7. Wosten M.M. Eubacterial sigma-factors // FEMS Microbiol Rev. – 1998. – V. 22. – P. 127–150.
8. Gruber T.M., Gross C.A. Multiple sigma subunits and the partitioning of bacterial transcription space // Annu Rev Microbiol. – 2003. – V. 57. – P. 441–466.
9. Mundade R., Ozer H.G., Wei H., Prabhu L., Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond // Cell Cycle. – 2014. – V. 13. – P. 2847–2852.
10. Suryamohan K., Halfon M.S. Identifying transcriptional cis-regulatory modules in animal genomes // Wiley Interdiscip Rev Dev Biol. – 2015. – V. 4. – P. 59–84.
11. Levati E., Sartini S., Ottonello S., Montanini B. Dry and wet approaches for genome-wide functional annotation of conventional and unconventional transcriptional activators // Comput Struct Biotechnol J. – 2016. – V. 14. – P. 262–270.
12. Frith M.C., Valen E., Krogh A., Hayashizaki Y., Carninci P., Sandelin A.A code for transcription initiation in mammalian genomes // Genome Res. – 2008. – V. 18. – P. 1–12.
13. Abeel T., Van de Peer Y., Saeys Y. Toward a gold standard for promoter prediction evaluation // Bioinformatics. – 2009. – V. 25. – P. i313–i320.
14. Scherf M., Klingenhoff A., Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach // J. Mol. Biol. – 2000. – V. 297. – P. 599–606.
15. Ohler U., Niemann H. Identification and analysis of eukaryotic promoters: recent computational approaches // Trends Genet. – 2001. – V. 17. – P. 56–60.
16. Shahmuradov I.A., Solovyev V.V., Gammerman A.J. Plant promoter prediction with confidence estimation // Nucleic Acids Res. – 2005. – V. 33. – P. 1069–1076.
17. Azad A.K.M., Shahid S., Noman N., Lee H. Prediction of plant promoters based on hexamers and random triplet pair analysis // Algorithms for Molecular Biology. – 2011. – V. 6. – P. 19.
18. Zuo Y-C., Li Q-Z. Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility // Genomics. – 2011. – V. 97. – P. 112–120.
19. Hertz G.Z., Stormo G.D. *Escherichia coli* promoter sequences: analysis and prediction. Methods in enzymology. – 1996. – V. 273. – P. 30–42.
20. Huerta A.M., Collado-Vides J. Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals // J Mol Biol. – 2003. – V. 333. – P. 261–278.
21. Gordon L., Chervonenkis A.Y., Gammerman A.J., Shahmuradov I.A., Solovyev V.V. Sequence alignment kernel for recognition of promoter regions // Bioinformatics. – 2003. – V. 19. – P. 1964–1971.
22. Gordon J.J., Towsey M.W., Hogan J.M., Mathews S.A., Timms P. Improved prediction of bacterial transcription start sites // Bioinformatics. – 2006. – V. 22. – P. 142–148.
23. Knudsen S. Promoter2.0: for the recognition of PolII promoter sequences // Bioinformatics. – 1999. – V. 15. – P. 356–361.
24. Reese M.G. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome // Comput Chem. – 2001. – V. 26. – P. 51–56.
25. Li Q.Z., Lin H. The recognition and prediction of sigma70 promoters in Escherichia coli K-12 // J Theor Biol. – 2006. – V. 242. – P. 135–141.
26. Rangannan V., Bansal M. Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition // Mol Biosyst. – 2009. – V. 5. – P. 1758–1769.
27. Solovyev V., Salamov A. Automatic annotation of microbial genomes and metagenomic sequences // In: Metagenomics and its applications in agriculture, biomedicine and environmental studies. (Ed. R.W. Li), Nova Science Publishers. – 2011. – P. 61–78.

28. Song K. Recognition of prokaryotic promoters based on a novel variable-window Z-curve method // Nucleic Acids Res. – 2012. – V. 40. – P. 963–971.
29. Hieno A., Naznin H.A., Hyakumachi M., Sakurai T., Tokizawa M. et al. ppdb: plant promoter database version 3.0 // Nucleic Acids Res. – 2014. – V. 42. – P. D1188–D1192.
30. Mitschke J., Vioque A., Haas F., Hess W.R., Muro-Pastor A.M. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in Anabaena sp. PCC7120 // Proc Natl Acad Sci USA. – 2011. – V. 108. – P. 20130–20135.
31. Vijayan V., Jain I.H., O'Shea E.K. A high resolution map of a cyanobacterial transcriptome // Genome Biol. – 2011. – V. 12. – P. R47.
32. Mitschke J., Georg J., Scholz I., Sharma C.M., Dienst D. et al. An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803 // Proc Natl Acad Sci USA. – 2011. – V. 108. – P. 2124–2129.
33. Kilic S., White E.R., Sagitova D.M., Cornish J.P., Erill I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria // Nucleic Acids Res. – 2014. – V. 42. – P. D156–160.
34. Cardon L., Stormo G. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments // J. Mol. Biol. – 1992. – V. 5. – P. 159–170.
35. Afifi A.A., Azen S.P. Statistical analysis: a computer oriented approach. Academic Press. – 2014. – 384 p.
36. Shahmuradov I.A., Umarov R.Kh., Solovyev V.V. TSSPlant: a new tool for prediction of plant Pol II promoters // Nucleic Acids Research. – 2017. – Published online 13 Jan 2017. doi: https://doi.org/10.1093/nar/gkw1353.
37. Shahmuradov I.A., Mohamad Razali R., Bougouffa S., Radovanovic A., Bajic V.B. bTSSfinder: a novel tool for the prediction of promoters in *Cyanobacteria* and *Escherichia coli*. Bioinformatics. – 2016. – Published online 30 Sep. doi: 10.1093/bioinformatics/btw629.
38. Reese M.G., Harris N.L., Eeckman F.H. Large Scale Sequencing Specific Neural Networks for Promoter and Splice Site Recognition. *Biocomputing: Proceedings of the 1996 Pacific Symposium* (edited by Lawrence Hunter and Terri E. Klein), World Scientific Publishing Co, Singapore, 1996, January 2–7.
39. Prestridge D.S. Predicting Pol II promoter sequences using transcription factor binding sites // J Mol Biol. – 1995. – V. 249. – P. 923–932.
40. Abeel T., Saeys Y., Bonnet E., Rouze P., Van de Peer Y. Generic eukaryotic core promoter prediction using structural features of DNA // Genome Research. – 2008. – V. 18. – P. 310–323.

SHAHMURADOV I.A.
Institute of Molecular Biology and Biotechnologies, ANAS,
Azerbaijan, AZ1073, Baku, Matbuat Ave, 2a, e-mail: ilhambaku@gmail.com

NOVEL TOOLS FOR THE PREDICTION OF PROMOTERS IN PLANTS AND BACTERIA

Aim. The computational search for promoters remains an attractive problem in bioinformatics. Despite the attention it has received for many years, the problem has not been addressed satisfactorily. These studies were aimed to develop novel computer tools for prediction of promoters (transcription start sites, TSSs) in plants and bacteria. ***Results***. Two novel tools for prediction of RNA polymerase II promoters in plants (TSSPlant) and bacteria (bTSSfinder) have been developed. TSSPlant achieves significantly higher accuracy compared to the next best promoter prediction program for both TATA and TATA-less promoters; it is available to download as a standalone program at http://www.cbrc.kaust.edu.sa/download/. bTSSfinder predicts promoters for five classes of σ factors in Cyanobacteria (σ^A,σ^C,σ^H,σ^G and σ^F) and for five classes of sigma factors in *E. coli* (σ^{70},σ^{38},σ^{32},σ^{28} and σ^{24}). Comparing to currently available tools, bTSSfinder achieves highest accuracy. bTSSfinder is available standalone and online at http://www.cbrc.kaust.edu.sa/btssfinder. ***Conclusions***. To date, TSSPlant and bTSSfinder are most accurate promoter predictors in plants and bacteria, respectively.

Keywords: transcription, RNA polymerase, promoter, TSS, promoter prediction.