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 σ70 and σ54 in *Escherichia coli*. While most bacteria possess multiple members of the σ70 family, they contain a single representative of the σ54 family, which is involved in nitrogen metabolism. Cyanobacteria lack any σ54-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 σ70, 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 σ70 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*
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/](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 $\approx 99\%$, specificity $\approx 98\%$, F1 score $\approx 99\%$ and Matthews correlation coefficient (MCC) 0.97. In the case of TATA-less promoters, we also achieved quite a good performance (with sensitivity $\approx 82\%$, specificity $\approx 97\%$, F1 score $\approx 89\%$ 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](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 $\approx 72\%$, F1 $\approx 47\%$), followed by TSSP (Sn $\approx 63.6\%$, F1 $\approx 41.2\%$), NNPP (Sn $\approx 51\%$, F1 $\approx 31\%$), EP3 (Sn $\approx 20\%$, F1 $\approx 31\%$) and Proscan (Sn $\approx 9\%$, F1 $\approx 15\%$).

**bTSSfinder: the bacterial promoter prediction tool**

The largest collection of experimentally vali-
dated promoters of *E. coli* in RegulonDB was classified into seven different sigma classes (σ^{20}, σ^{34}, σ^{38}, σ^{39}, 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^{20}), σ^{C} (analogous to σ^{38}), σ^{F} (analogous to σ^{28}), σ^{G} (analogous to σ^{24}) and σ^{H} (analogous to σ^{35}).

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

<table>
<thead>
<tr>
<th>Tool</th>
<th>Genes with ≥ 1 TSSpr</th>
<th>Total number of TSSpr</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Sn, %</th>
<th>F1-score, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSSPlant</td>
<td>54</td>
<td>115</td>
<td>40</td>
<td>75</td>
<td>15</td>
<td>72.7</td>
<td>47.1</td>
</tr>
<tr>
<td>TSSP</td>
<td>45</td>
<td>105</td>
<td>35</td>
<td>80</td>
<td>20</td>
<td>63.6</td>
<td>41.2</td>
</tr>
<tr>
<td>NNPP2</td>
<td>47</td>
<td>122</td>
<td>28</td>
<td>97</td>
<td>27</td>
<td>50.9</td>
<td>31.1</td>
</tr>
<tr>
<td>EP3^2</td>
<td>16</td>
<td>16</td>
<td>11</td>
<td>5</td>
<td>44</td>
<td>20.0</td>
<td>31.0</td>
</tr>
<tr>
<td>Proscan^2</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>50</td>
<td>9.1</td>
<td>15.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Promoter prediction tool</th>
<th>Genes with ≥ 1 TSSpr</th>
<th>Genes without any TSSpr</th>
<th>Genes with true TSSpr</th>
<th>Total number of TSSpr</th>
<th>Sn, %</th>
<th>TSSP density</th>
</tr>
</thead>
<tbody>
<tr>
<td>bTSSfinder</td>
<td>197</td>
<td>3</td>
<td>143</td>
<td>355</td>
<td>71.5</td>
<td>620</td>
</tr>
<tr>
<td>BPROM</td>
<td>200</td>
<td>0</td>
<td>130</td>
<td>569</td>
<td>65.0</td>
<td>386</td>
</tr>
<tr>
<td>NNPP2</td>
<td>175</td>
<td>25</td>
<td>109</td>
<td>500</td>
<td>54.5</td>
<td>440</td>
</tr>
<tr>
<td>PromPredict</td>
<td>74</td>
<td>126</td>
<td>0</td>
<td>149</td>
<td>0.0</td>
<td>14772</td>
</tr>
</tbody>
</table>

Notes: TSSan: annotated TSS position 1001). TSSpr: predicted TSS. ^ Prediction is true, if distance between annotated and predicted TSSs is 50 bp or less. ^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 EIF-2010-11-4027-3 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
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.cbr.cau.st.edu/sa/download/](http://www.cbr.cau.st.edu/sa/download/). bTSSfinder predicts promoters for five classes of σ factors in Cyanobacteria (σ7, σ32, σ21, σ28 and σ24) and for five classes of sigma factors in *E. coli* (σ32, σ24, σ28, σ21 and σ21). Comparing to currently available tools, bTSSfinder achieves highest accuracy. bTSSfinder is available standalone and online at [http://www.cbr.cau.st.edu/btssfinder](http://www.cbr.cau.st.edu/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.