CHAPTER 7

TUNING THE DRAGON PROMOTER FINDER SYSTEM FOR HUMAN PROMOTER RECOGNITION

Vladimir B. Bajić

Institute for Infocomm Research
bajicv@i2r.a-star.edu.sg

Allen Chong

Institute for Infocomm Research
achong@i2r.a-star.edu.sg

Discovery of new genes through the identification of their promoters in anonymous DNA and the study of transcriptional control make promoter and transcription start site recognition an important issue for Bioinformatics. The biological process of transcription activation is very complex and is not completely understood. Hence computer systems for promoter recognition may not perform well if we do not pay attention to their tuning.

This chapter explains the tuning of a computer system for the recognition of functional transcription start sites in promoter regions of human sequences. The system is called Dragon Promoter Finder, and it can be accessed at `http://sdmc.i2r.a-star.edu.sg/promoter`. The tuning of this complex system is set up as a multi-criteria optimization problem with constraints. The process is semi-automatic, as it requires an expert assessment of the results.

ORGANIZATION.

Section 1. We briefly discuss the importance and challenges of promoter and transcription start site recognition. Then we briefly describe the performance of several existing systems for this recognition problem.

Section 2. Good performance in promoter recognition requires careful tuning of the recognition system. We use Dragon Promoter Finder, which is one of the best performing system on this problem, to illustrate this tuning process. So, we zoom into a more extensive exposition of the architecture of Dragon Promoter Finder in this section.

Section 3. After that, we identify the parameters of Dragon Promoter Finder which should be tuned. These parameters include the parameters of a nonlinear signal processing block and a threshold on the output node of an artificial neural network.

Section 4. In order to determine the optimal parameter values, carefully chosen tuning data should be used. The selection of our tuning data set, comprising samples of both promoters and non-promoters, is described in this section.

Section 5. We are then ready to dive into the details of the tuning process of Dragon Promoter Finder. This process is a multi-criteria optimization that is cast as a Gembicki goal attainment optimization.

Section 6. Finally, we present the fruit of this tuning process. Specifically, we demonstrate the significant superiority of Dragon Promoter Finder's performance compared to several systems for finding promoters and transcription start sites. On a test set of 1.15Mbp of diverse sequences containing 159 transcription start sites, Dragon Promoter Finder attains several folds less false positives than other systems at the same level of sensitivity.

1. Promoter Recognition

Promoters are functional regions of DNA that control gene expression. The biochemical activity in this region—involving the interaction of DNA, chromatin and transcription factors—determines the initiation and the rate of gene transcription. In eukaryotes, the promoter region is usually located upstream of, or overlaps, the transcription start site.[249, 656, 878] A gene has at least one promoter.[656, 878] One can thus find a gene by first locating its promoter.[879] Discovering new genes through the identification of their promoters in anonymous DNA and the study of transcriptional control make promoter recognition an extremely important issue for Bioinformatics.

Even though promising solutions have been proposed—by Bajic *et al.*,[48, 49, 51] Davuluri *et al.*,[190] Down and Hubbard,[211] Hannenhalli and Levy,[321] Ioshikhes and Zhang,[383] and Scherf *et al.*[753]—computational recognition of promoters still has not yet achieved a satisfactory level of confidence.[249, 802] The reason is that the biological process of transcription activation is very complex and hierarchical, and is not completely understood.[878] There are numerous and functionally diverse transcription factors that individually bind to specific DNA consensus sequences—called transcription factor binding sites—in the promoter to activate the transcriptional machinery in concert with RNA polymerases.

Many simplistic approaches have been taken in computational promoter recognition.[681] Unfortunately, at significant levels of true positive recognition, these have produced a significant number of false positive recognition.[249, 681, 708] A false positive prediction is a prediction that indicates the presence of a promoter at a location where the promoter does not exist. A true positive prediction is one that correctly identifies the location of a promoter. Promoter recognition systems for large-scale DNA screening require an acceptable ratio of true positive and false positive predictions. That is, these systems should maximize true positive recognition while minimizing false positive recognition.

While the boundaries of promoter region are loosely defined, each promoter has at least one strong reference site: the transcription start site (TSS). Promoter search can thus focus either on locating the promoter region,[190,321, 383,753] or on pinpointing the TSS.[48,49,51,211,439,628,707] The system in this chapter is a TSS finder. Existing TSS-finders—like NNPP2.1,[707] Promoter2.0,[439] and McPromoter[628]—produce a lot of false positive predictions,[628,656,681,708] making them unsuitable for locating promoters in large genomic sequences.

A recently reported system, Eponine,[211] has demonstrated very good performance. However, its predictions are very much related to CpG-island associated promoters[81,180,271,474] and to G+C rich promoters. The G+C-content of a DNA segment is the proportion of the total number of G and C nucleotides relative to the length of that segment. CpG islands are unmethylated segments of DNA longer than 200bp, with the G+C content of at least 50dinucleotides—that is, a C followed by a G—being at least 60the G+C content of the segment.[81,180,271,474] CpG islands are found around gene starts in approximately half of mammalian promoters[180,474] and are estimated to be associated with about 60% of human promoters.[181] For this reason it is suggested by Pedersen *et al.*[656] that CpG islands could represent a good global signal to locate promoters across genomes. At least in mammalian genomes, CpG islands are good indicator of gene presence. The G+C content is not uniformly distributed over the chromosomes and the CpG island density varies according to the isochores' G+C content.[673]

Several promising systems have been developed in the last two years. PromoterInspector[753] has been reported to produce a considerably reduced level of false positive recognition compared to other publicly available promoter recognition programs. After its introduction, three other systems with similar performance have also been reported.[190,321,383] These four systems predict regions that either overlap promoter regions, or are in close proximity to the promoter. The localization of TSS is not considered in Hannenhalli and Levy[321] and Ioshikhes and Zhang.[383] However, if we wish to develop a promoter model that can efficiently search for genes of a specific genetic class through the recognition of relevant promoter features for the targeted class of genes, it is necessary to pinpoint the TSS and thus, localize the promoter region.

2. Dragon Promoter Finder

It is well known that TSS-finding systems produce a high number of false positives.[249,628,681,708] The TSS-finding system that we present in this chapter considerably reduces the number of false positives. In contrast to solutions which are aimed at the recognition of specialized classes of promoters—such as CpG-island related promoters,[321,383] or G+C-rich promoters with a TATA-box[211]—our system is aimed at analyzing and identifying general human polymerase II promoters. While the design details of our system, Dragon Promoter Finder, have been published elsewhere,[48,49,51] the tuning process of this complex system has not described previously. We thus present here the details of the tuning process for Dragon Promoter Finder system. The tuning process is performed using Gembicki's goal attainment optimization process.[99,284,754]

Our system is at `http://sdmc.i2r.a-star.edu.sg/promoter`. It is based on a hierarchical multi-model structure with models specialized for

(a) different promoter groups, and
(b) different sensitivity levels.

To the best of our knowledge, this is the first reported composite-model structure used in promoter recognition systems based on (a) and (b) above. First, the short DNA segments around TSS representing the promoter data are separated into the G+C-rich and G+C-poor groups. This separation of data and subsequent development of models for both of these promoter groups, as well as the sophisticated tuning of the models, has resulted in considerably enhanced system performance. The resulting system combines:

- multiple hierarchically organized models optimally tuned for different sensitivity requirements,
- specialization of models to G+C-rich or G+C-poor promoter groups,
- sensor-integration,
- nonlinear signal processing, and
- artificial neural networks (ANN).

This makes it conceptually different from the approaches used in other promoter-finding systems,[249,681] including those that use several region sensors.[488,628,753]. The system is shown to be capable of successfully recognizing promoters that are CpG-island related and those that are not, as well as promoters in G+C-rich and in G+C-poor regions. This makes it quite universal as opposed to solutions which are specialized in recognizing CpG-islands related promoters,[321,383] or to the one in Down and Hubbard.[211]

The practical significance of Dragon Promoter Finder is in its use for identification and annotation of promoters in anonymous DNA, as well as in the enhancement of gene hunting by more accurate determination of the 5' end of the gene and parts of the gene's regulatory regions.

3. Model

The description of the system is presented by Bajic and colleagues.[48, 49, 51] The system possesses three sensors for promoters, coding exons, and introns. Let the produced signals of the promoter, coding exon, and intron sensors be denoted respectively by σ_p, σ_e, and σ_i. These signals enter a nonlinear signal processing block where they are transformed according to

$$\begin{aligned} s_E &= f(\sigma_p - \sigma_e, a_e, b_e, c_e, d_e) \\ s_I &= f(\sigma_p - \sigma_i, a_i, b_i, c_i, d_i) \\ s_{EI} &= f(\sigma_p - \sigma_{ei}, a_{ei}, b_{ei}, c_{ei}, d_{ei}) \end{aligned}$$

In Ver. 1.2 of Dragon Promoter Finder, the function f is defined by

$$f = blin = \begin{cases} c \times x, \text{ if } x > a \\ x, \text{ if } b \leq x \leq a \\ d \times x, \text{ if } b > x \end{cases}$$

In Ver. 1.3 of Dragon Promoter Finder, the function f is defined by

$$f = sat = \begin{cases} a, \text{ if } x > a \\ x, \text{ if } b \leq x \leq a \\ b, \text{ if } b > x \end{cases}$$

The parameters a_k, b_k, c_k, d_k, for $k = e, i, ei$, are part of the tunable system parameters. Also, the signals s_E, s_I, and s_{EI} are subject to whitening in Ver.1.2, and to principal component transform in Ver.1.3. The transformed signals—z_E, z_I, and z_{EI}—are inputs to the feed-forward ANN. The ANN is trained by the Bayesian regularization method[82] for the best separation between the classes of input signals, with initial parameters $a_k = +\infty$ and $b_k = -\infty$ for $k = e, i, ei$. The trained ANN is then used as a part of the system in the final tuning.

4. Tuning Data

In order to determine the optimal parameter values, a set of tuning data composed of promoter and non-promoter sequences must be created. This section briefly describes our tuning data set.

4.1. *Promoter Data*

For the promoter data of the tuning data set, we used 793 different vertebrate promoter sequences from the Eukaryotic Promoter Database[662]. These sequences are extracted from the window $[-250, +50]$ relative to the TSS position. Note that, by convention, there is no nucleotide position "0". The nucleotide at the TSS is assigned the position "1" and the nucleotide immediately preceding the TSS is assigned the position "–1". Additionally, we used $[-250, +50]$ sequence segments of 20 full-length gene sequences with known TSS, whose promoters are not included by the Eukaryotic Promoter Database.

4.2. *Non-Promoter Data*

For the non-promoter data of the tuning data set, we randomly collect from Genbank[76] Rel. 121 a set of non-overlapping human coding exons and intron sequences, each 250bp in length. We also selected non-overlapping human sequences from the 3'UTR regions taken from the UTRdb.[665] All sequences in these three groups are checked for similarity using the BLAST2Sequences program[820] to ensure that any two sequences within the group have less than 50% identity relative to each other. In total, 1300 coding exon sequences, 4500 intron sequences, and 1600 3'UTR sequences are selected. Additionally, from the 20 gene sequences mentioned earlier, we include as non-promoter data all 250bp segments that do not overlap the $[-250, +50]$ regions.

5. Tuning Process

The Dragon Promoter Finder requires careful tuning to achieve the best performance of the system in recognition of TSS in a blind promoter search. The general goal of tuning is to maximize the level of true positives versus false positives over the entire range of sensitivity settings. Different models are trained and each is tuned for the best performance at a predefined sensitivity level. This means that we aim at making the highest positive predictive value (ppv)—sometimes denoted as specificity in bioinformatics[50]—for the predefined sensitivity level. The sensitivity and ppv are given by

$$S_e = \frac{TP}{TP + FN}$$
$$ppv = \frac{TP}{TP + FP}$$

where FN stands for false negatives and equals the number of true promoters not predicted by the promoter prediction programs; TP stands for true positives and equals the number of true promoters predicted by the promoter prediction programs; and FP stands for false positives and equals to the number of non-promoters incorrectly claimed by the promoter prediction programs as promoters.

The tuning process can thus be considered as an optimization process with two goals expressed by

$$\max S_e$$
$$\max ppv$$

However, TP, FP, and FN can be only positive integers. Therefore, the formulation of the tuning of the above-mentioned optimization problem cannot take full advantage of the sophisticated optimization algorithms with continuous criteria functions. So, we need to reformulate the optimization problem for the tuning purpose.

For Ver. 1.3, the set of tunable parameters p of the system consists of the ANN threshold τ and a_k, b_k, c_k, d_k, for $k = e, i, ei$. For Ver. 1.2, additional sensor signal thresholds are also used. These parameters have to be adjusted so that the tuned system achieves the desired performance. The tuning process is conducted 10 times for each selected level of sensitivity, and different models are produced in the process. Then, from all of the models, the selection of the representative model for each sensitivity level is made.

In the tuning process, a sequence from the tuning set is presented to the system and this produces an output signal s. This signal is compared to the desired target value t, which is 1 for promoters and –1 for non-promoters, and the error $e = s - t$ is calculated. The error serves as the signal that determines the change in the system's tunable parameters by means of a feedback process. This tuning process can be considered as a multi-criteria optimization with constraints, more specifically as a Gembicki's goal attainment optimization.[99, 284, 754] The details of which is described in the remainder of this section.

We define several objectives for which we want to achieve the predefined goals in the tuning process. These objectives are captured in the vector F of objectives of the system produced on the tuning set as

$$F = \langle E_p, E_e, E_i, E_{utr}, E_g, S_e, ppv \rangle$$

where

$$\begin{aligned} E_k &= \frac{1}{N_k} \times \sum_{j=1}^{N_k} |e_j^k|, \text{ for } k = p, e, i, utr, g \\ e_j^k &= s_j^k - t_j^k, \\ t_j^k &= \begin{cases} 1, & \text{if } k = p \\ -1, & \text{otherwise} \end{cases} \end{aligned}$$

Here, N_k is the number of presented sequences of a specific class k; s_j^k and t_j^k are the system output signal and target values respectively, for group k when the

j-th sequence is presented to the system; p, e, i, utr, g stand for promoter, coding exon, intron, 3'UTR, and the sequences corresponding to non-promoter positions in the selected 20 genes, respectively.

The goal attainment process is defined as

$$\min_{\gamma \in R,\ p \in \Omega} \gamma$$

subject to the following constraints:

$$\begin{aligned} E_j - w_j \times \gamma &\leq 1, \text{ for } j = p, e, i, utr, g \\ \frac{1}{ppv} - w_f \times \gamma &\leq 1 \\ \frac{1}{S_e} - w_t \times \gamma &\leq \frac{1}{L_s} \end{aligned}$$

Here, L_s is the predefined sensitivity level for which the model is tuned, Ω is the overall parameter space, and w_j, w_t, and w_f are the slack weights in the optimization. The tuning process is repeated 10 times with the tuning parameters randomly initialized.

After the collection of models are produced for the all selected sensitivity levels, then the selection of the appropriate models for each level is made. This second phase of the process is not automated and requires manual selection. The goals in the model selection process are

- the change of the parameter values for the selected models for different successive sensitivity levels has to be gradual, and
- the best possible models, expressed in terms of E_{TP} and E_{FP}, should be selected.

Sometimes it is not possible to choose models that satisfy the gradual change in the parameters. Then the tuning for the critical sensitivity levels is repeated sufficient number of times until this can be done.

6. Discussions and Conclusions

This tuning has resulted in a superior recognition ability of Dragon Promoter Finder. For the purpose of illustration, we present in Figure 1 some of the performance comparison results.

The Dragon Promoter Finder is compared with the Promoter2.0 program,[439] the NNPP2.1 program,[707] and the PromoterInspector program.[753] The first two programs are TSS-finding programs, while the third one is a promoter region

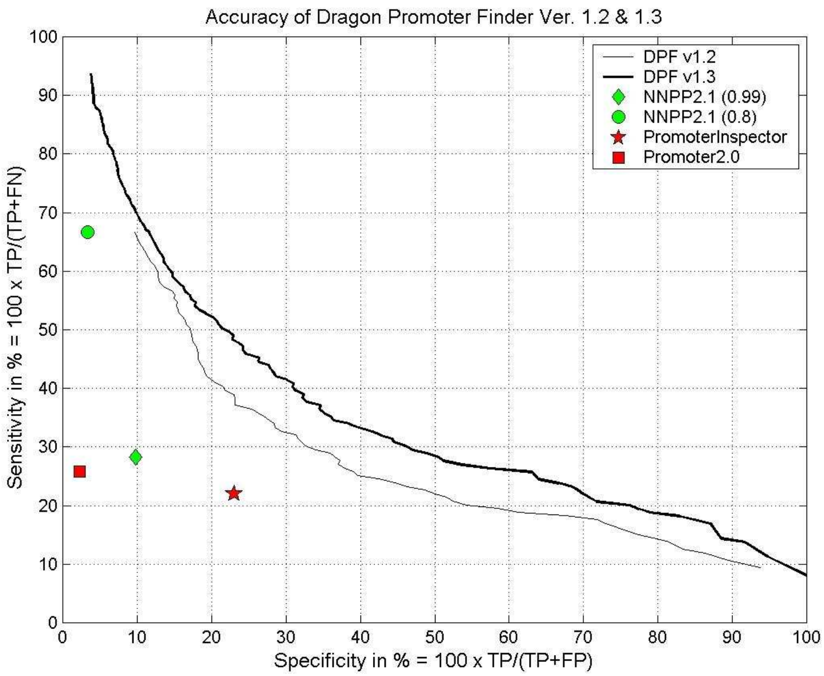

Fig. 1. Comparison of performances of three TSS prediction programs—Dragon Promoter Finder Ver.1.2 and Ver.1.3, Promoter2.0 and NNPP2.1—on the test set from Bajic *et al.*[51] which has a total length of 1.15Mbp and comprises 159 TSS.

finder. The details of the comparisons, as well as of the test set, are given in Bajic *et al.*[48, 49, 51], and at `http://sdmc.i2r.a-star.edu.sg/promoter`. In these tests, Dragon Promoter Finder has produced several folds smaller number of false positive predictions than the other promoter prediction systems.

Due to the very complex structure of the system and its many tunable parameters, it is not possible to make sequential tuning of parameters in the model. A reasonable solution to this problem is to use a general optimization approach. So, as described in the previous section, we have opted for the Gembicki's goal attainment optimization. Choosing random initialization values for parameters 10 times we obtained different models at each of the predefined sensitivity levels. This is necessary as, in general, the model parameters converge to different values. In total, we have generated models for 85 levels of sensitivity, spanning from $S_e = 0.1$ up to $S_e = 0.94$.

One of the obstacles in the final selection of the models is the non-gradual change in the tunable parameter values for the successive sensitivity levels. The reason for requesting a gradual change in parameter values for successive sensitivity levels is that this property indirectly implies that the models for each of the sensitivity levels is not overfitted to the training data, but rather to the general properties of the data classes. These characteristics are necessary for good generalization in the classification and recognition process. In the cases when the best models for the successive sensitivity levels showed abrupt change in tunable parameter values, the tuning process is repeated sufficient number of times until this criterion is satisfied.

For the lower sensitivity range from 0.1 to 0.27, we have used a data-window of 250bp since the models using this sequence length allowed for very high specificity. For the higher sensitivity levels, we have used a data-window of 200bp because—from our previous experience with this window length—we are able to achieve reasonable sensitivity/specificity ratios not possible with the window length of 250.

Let us summarize. We have presented details of the tuning procedure for one of the most efficient transcription start site recognition system in human DNA. The tuning problem is treated as iterative multi-criteria optimization of the goal attainment type. It has resulted in good generalization of the system and superior prediction ability. This tuning approach has opened a way for a more general method of tuning complex prediction systems for computational biology.