Towards Accurate Transcription Start Site Prediction: a modelling approach

Marko Djordjevic
Institute of Physiology and Biochemistry
Faculty of Biology, University of Belgrade
11000 Belgrade, Serbia
dmarko@bio.bg.ac.rs

Abstract: Promoter prediction in bacteria is a classical bioinformatics problem, where available methods for regulatory element detection exhibit a very high number of false positives. We here argue that accurate transcription start site (TSS) prediction is a complex problem, where available methods for sequence motif discovery are not in itself well adopted for solving the problem. We here instead propose that the problem requires integration of quantitative understanding of transcription initiation with careful description of promoter sequence specificity. We review evidence for this viewpoint based on our recent work, and discuss a current progress on accurate TSS detection on the example of sigma70 transcription start sites in *E. coli*.

1 Introduction

Bacterial RNA polymerase is a central enzyme in cell, and initiation of transcription by bacterial RNA polymerase is a major point in gene expression regulation. Core RNA polymerase cannot by itself initiate transcription, so a complex between RNA polymerase core and a σ factor, which is called RNA polymerase holoenzyme (RNAP) is formed. A major σ factor, which is responsible for transcription of housekeeping genes, is called σ^{70} in *E. coli* and σ^{A} in a number of other bacteria. The discussion here will concentrate on this major class of promoter elements [BN01].

Accurate recognition of transcription start sites (TSS) is a necessary first step in understanding transcription regulation. Accurate recognition of bacterial promoters is consequently considered a major problem in bioinformatics, particularly since TSS detection is an important ingredient for number of other bioinformatic applications (e.g. gene and operon prediction). Available methods for TSS search include both standard information-theory based weight matrix searches, and those based on more advanced computational approaches such as neural networks and support vector machines. These methods however show poor accuracy for TSS prediction, i.e. lead to a very high number of false positives [St02]. We here argue that, instead of developing different methods for processing the existing data within the motif search framework, solving the problem requires an integrative approach, which includes: i) quantitatively modelling transcription initiation, which allows calculating kinetic parameters of transcription initiation ii) accurately describing sequence specificity of the promoter elements, so that

the bioinformatics description is consistent with available biophysical measurements iii) characterizing sequence elements outside of the canonical -10 and -35 box. In the text below we concentrate on promoter detection for sigma 70 class of promoters, which is a major promoter class that is responsible for transcription of housekeeping genes.

Our discussion will emphasize the following: *i*) accurately aligning promoter elements is highly non-trivial, so that the promoter specificity may not be accurately reflected by the available alignments *ii*) the promoter specificity is likely determined by additional sequence elements, which are located outside of the canonical -35 and -10 boxes iii) TSS predictions require accurately calculating kinetic parameters of transcription initiation. In addressing these issues we will highlight our recently published work [Dj03,Dj04], and also discuss some of our most recent results on this problem.

2 High degeneracy of promoter elements

Transcription initiation begins with RNAP binding to dsDNA, which is referred to as the closed complex formation [DZR05]. Subsequent to RNAP binding, the two strands of DNA are separated through thermal fluctuations that are facilitated by interactions of RNAP with ssDNA [DR06]. The opening of two DNA strands results in a formation of ~15bps long transcription bubble, which typically extends from -11 to +3 (where +1 corresponds to the transcription start site) [BS07]. After the open complex is formed, RNAP clears the promoter and enters the elongation, which leads to synthesis of RNA from DNA template [BN01].

The main elements that determine functional promoter are -35 element (${}^{-35}$TTGACA^{-30}, where the coordinates in the superscript are relative to the transcription start site), -10 element (${}^{-12}$TATAAT^{-7}), the spacer between these two elements, and the extended -10 element (${}^{-15}$TG^{-14}) [HH08]. Interactions of σ70 with dsDNA of -35 element, extended -10 element, and -12 base of -10 element result in the closed complex formation [MD09]. On the other hand, the downstream bases of -10 element (-11 to -7) interact with σ70 in ssDNA form [MD09], and are directly involved in the open complex formation.

Consequently to better relate involvement of different promoter elements with the kinetic steps of transcription initiation (the closed and the open complex formation), it was recently proposed that the region from -15 to -7 is reorganized in the following way [HH08]: Region from -15 to -12 is connected in a new element that is defined as -15 element; this element includes extended -10 element, the most upstream base in -10 element (base -12), and base -13 that is in-between. Consequently, -10 element is shortened for one base-pair (to the region -11 to -7), which we here refer to as the *short* -10 element. In this way -35 and -15 elements are directly related with σ^{70}-dsDNA interactions, while short -10 element is directly related with σ^{70}-ssDNA interactions.

The basic problem with accurate promoter detection is high degeneracy of the promoter elements (-35, -15 and -10 elements); in addition, variable distance between -35 and -10 element also contributes to the problem. This high degeneracy is illustrated in Table 1, where we show the aligned elements for several randomly selected promoters. For example, if we concentrate on -35 element, we see that the consensus sequence

'TTGACA' does not match any of the promoter instances in the table. Furthermore, only one instance has one mismatch from the consensus, most of the instances have two mismatches, while two of the instances have as much as four mismatches. In order to accommodate such high degeneracy, i.e. to correctly classify majority of the detected promoters, a low value of the detection threshold has to be imposed; this low threshold value than leads to a high number of false positives. One can artificially increase the detection threshold, which would decrease the false positives; however, another problem than emerges, i.e. a number of experimentally detected promoters are than wrongly classified. Consequently, the high degeneracy of the promoter elements, together with the relatively complex core promoter structure (several sequence elements with variable relative distances), is the main reason behind the low prediction accuracy of the available approaches.

Promoter	**-35**	**spacer**	**-15**	**short -10**
accAp	TTGCTA	17	AGGC	AAATT
accBp	TTGATT	17	GACC	AGTAT
accDp	TATCCA	19	TGTT	TTAAT
aceBp	TTGATT	16	GAGT	AGTCT
acnAp1	CTAACA	15	GCCT	TTATA
acnAp2	TCAAAT	19	TGTT	ATCTT

Table 1: Examples of promoter sequence elements

3 Importance of the accurate promoter alignment

A necessary step in accurate TSS prediction is achieving a quantitative understanding of promoter specificity, i.e. accurately defining sequence elements that constitute bacterial promoter. However, aligning the promoter elements presents in-itself a highly non-trivial bioinformatic task due to both complex structure of bacterial promoter and degeneracy of the promoter elements (see above). A major problem with the existing collections of the promoter elements is due to the following: *i*) they are based on initial alignments of a small collection of promoter elements which were performed `by eye' [WB10,HC11,RMC12,MZBM13] *ii*) accurate aligning of -35 element is complicated by both variable distance from -35 element and by a lower conservation of this element [HC11] *iii*) it is non-trivial to produce an alignment with sufficient accuracy for analyzing -15 element, given a weaker conservation of this element compared to both -10 and -35 elements [MZBM13].

Having these problems in mind, we recently performed a systematic `de-novo' alignment of the promoter elements on a large collection of more than 300 experimentally confirmed σ^{70} TSS in *E. coli* [Dj04]. This alignment comes directly from experimentally determined TSS assembled in RegulonDB database [GSPSMSJWGL14]. For this we used Gibbs search algorithm for unsupervised alignment of the promoter elements, which we

consequently improved through supervised search by weight matrices defined through the Gibbs algorithm. The approach was to first align -10 element, and to consequently use this element as an anchor to align -35 element. Alignment of other relevant elements (spacer and -15 element) is directly determined once -10 element and -35 element are aligned. One should note that in addition to the canonical -10 and -35 elements our approach also allowed quantitating specificity of -15 element.

The unbiased alignment that we inferred shows notable differences with previously published alignments, as is discussed in more detail in [Dj04]. Furthermore, while our alignment is in accordance with biophysical data on σ^{70}-DNA interactions, the previously published alignments show notable discrepancies with the interaction data. We therefore next investigated to what extent the improved sequence alignment can in itself improve the prediction accuracy. To that end, we incorporated our improved alignment in the standard (weight matrix based) procedure for TSS detection, and compared the prediction accuracy with those resulting from the previous alignments. The comparison is shown in Figure 2, where we see that our alignment leads to as much as 50% reduction in the number of false positives. However, despite this significant reduction, one can see that the number of false positives is still very high; this then leads us to the next question that we consider, which is to what extent are kinetic effects important in transcription initiation.

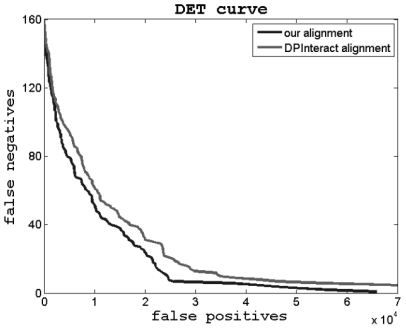

Figure 1: DET (Detection Error Tradeoff) curve, which presents comparison of the method based on our alignment, and the method in [RMC12] that uses the alignment from DPInteract database. The vertical axis presents the number of false positives, which is based on the number of correctly classified sequences in the testing set. The horizontal axis presents the number of false negatives, which is estimated based on the number of hits in the randomized intergenic regions. The blue line and the red line correspond, respectively, to our procedure and the procedure from [RMC12]. The figure adopted from (Djordjevic M, Djordjevic M, to be submitted).

4 Importance of the kinetic effects

We next discuss another factor which may have a major impact on the accuracy of TSS predictions, which are kinetic effects in transcription initiation. As the first step of transcription initiation, RNAP reversibly binds to dsDNA of promoter elements, which is called the closed complex formation, and is described by the binding affinity K_B. This binding of RNAP leads to opening of the two DNA strands (promoter melting), so that a

transcription bubble is formed. This transcription bubble extends from the upstream edge of -10 element to about two bases downstream of the transcription start site, which roughly corresponds to positions -12 to +2 (+1 is transcription start site) [BS07]. The (inverse) time needed to form the transcription bubble (i.e. to open the two DNA strands) is described by the transition rate from closed to open complex (k_f).

An extreme example of the kinetic effects in transcription initiation are poised promoters: These are locations in genome where RNAP binds with high binding affinity (high K_B), but has a low rate of transcription initiation due to a slow transition from closed to open complex (low k_f). It has been proposed that poised promoters may present a major problem for accurate TSS prediction [HC11,SF15]. This is particularly important, given the high number of false positives [St02,HC11,RMC12] that typically originate from computational TSS searches.

We consequently used the kinetic model of transcription initiation that we previously developed [DR06] in order to estimate the importance of the kinetic effects, which is shown in Figure 2 (for more details see [Dj03]).

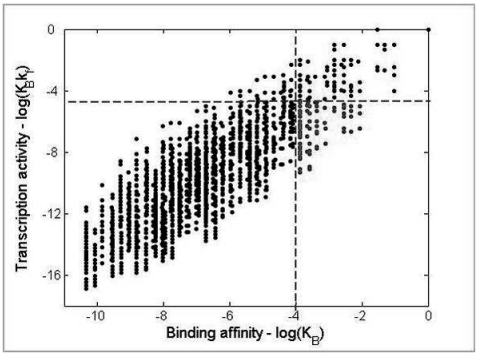

Figure 2: Log transcription rate ($\log(\varphi)$) vs. log binding affinity ($\log(K_B)$) for the intergenic segments. -10 element of lacUV5 promoter is substituted by all 6bp long segments from *E. coli* intergenic regions. $\log(K_B)$ and $\log(\varphi)$ are calculated for each of these substitutions and shown, respectively, on the horizontal and the vertical axes on each of the panels. The horizontal and the vertical dashed lines correspond, respectively, to the transcription rate threshold and the binding affinity threshold. Green and red dots in the figure correspond to the strongly bound DNA sequences that are, respectively, functional promoters and poised promoters. Figure adapted from [Dj03].

From Figure 2 we see that a significant fraction of the strongly bound sequences corresponds to poised promoters: In Figure 2, the blue dots mark strongly bound DNA segments that correspond to the functional promoters (i.e. to sequences that are above both the binding and the transcription activity threshold), while the red dots mark the sequences that correspond to the poised promoters (i.e. to sequences that are above the binding, but below the transcription activity threshold). One can see that a significant fraction of the strongly bound sequences (~30%) correspond to poised promoters. Such poised promoters can be falsely identified as targets by computational searches of core

promoters. Furthermore, our results from [Dj04] strongly suggest that the relevant kinetic parameter that characterizes functional promoters is the overall transcription activity; this is in contrast to some previous models which suggested that the relevant parameter is binding affinity of RNAP to dsDNA.

5 Conclusion

Accurate promoter prediction in bacteria is crucial not only as the first step in understanding transcription regulation, but also as an important ingredient in other bioinformatics applications such as gene and operon prediction. Despite being a classical bioinformatics problem, current methods for transcription start site prediction lead to a very high number of false positives. We here argue that transcription start site detection is a complex problem whose solution requires integrating several levels of knowledge. In particular, the discussion here strongly indicates that the following elements are necessary: *i*) accurately aligning promoter elements *ii*) characterizing sequences outside of canonical -35 and -10 boxes *iii*) estimating kinetic parameters of transcription initiation for a given sequence of interest, in particularly its transcription activity. Our current work is aimed at addressing these issues.

Acknowledgements

MD acknowledges support by Marie Curie International Reintegration Grant within the 7th European community Framework Programme (PIRG08-GA-2010-276996), by the Ministry of Education and Science of the Republic of Serbia under project number ON173052, and by the Swiss National Science foundation under SCOPES project number IZ73Z0_152297.

References

[BN01] Borukhov, S., Nudler, E.: RNA polymerase holoenzyme: structure, function and biological implications. Curr Opin Microbiol 6, 93-100, 2003.

[St02] Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics 16, 16-23, 2000.

[Dj03] Djordjevic, M.: Efficient transcription initiation in bacteria: an interplay of protein-DNA interaction parameters. Integr Biol (Camb) 5, 796-806, 2013.

[Dj04] Djordjevic, M.: Redefining Escherichia coli sigma(70) promoter elements: -15 motif as a complement of the -10 motif. J Bacteriol 193, 6305-6314, 2011.

[DZR05] DeHaseth, P.L., Zupancic, M.L., Record Jr, M.T.: RNA polymerase-promoter interactions: The comings and goings of RNA polymerase. J Bacteriol 180, 3019-3025, 1998.

[DR06] Djordjevic, M., Bundschuh, R.: Formation of the Open Complex by Bacterial RNA Polymerase—A Quantitative Model. Biophysical Journal 94, 4233-4248, 2008.

[BS07] Borukhov, S., Severinov, K.: Role of the RNA polymerase sigma subunit in transcription initiation. Res Microbiol 153, 557-562, 2002.

[HH08] Hook-Barnard, I.G., Hinton, D.M.: Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. Gene Regulation and Systems Biology 1, 275, 2007.

[MD09] Murakami, K.S., Darst, S.A.: Bacterial RNA polymerases: the wholo story. Curr Opin Struct Biol 13, 31-39, 2003.

[WB10] Wang, H., Benham, C.J.: Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. BMC bioinformatics 7, 248, 2006.

[HC11] Huerta, A.M., Collado-Vides, J.: Sigma 70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. J Mol Biol 333, 261-278, 2003.

[RMC12] Robison, K., McGuire, A., Church, G.: A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. Journal of molecular biology 284, 241-254, 1998.

[MZBM13] Mitchell, J.E., Zheng, D., Busby, S.J.W., Minchin, S.D.: Identification and analysis of 'extended–10'promoters in Escherichia coli. Nucleic acids research 31, 4689, 2003.

[GSPSMSJWGL14] Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñiz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J.S., López-Fuentes, A.: RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic acids research 39, D98, 2011.

[SF15] Stormo, G.D., Fields, D.S.: Specificity, free energy and information content in protein-DNA interactions. Trends Biochem. Sci 23, 109–113, 1998.