Integrating Sequence Analysis with Biophysical Modelling for Accurate Transcription Start Site Prediction

Marko Djordjevic[*]

Institute of Physiology and Biochemistry, Faculty of Biology, University of Belgrade, Serbia

Summary

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, and discuss a current progress on these issues 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 [1].

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 [2]. 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.

[*] To whom correspondence should be addressed. Email: dmarko@bio.bg.ac.rs

doi:10.2390/biecoll-jib-2014-240

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. These points will be further discussed in the text below.

2 High degeneracy of promoter elements

Transcription initiation begins with RNAP binding to dsDNA, which is referred to as the closed complex formation [3]. Subsequent to RNAP binding, the two strands of DNA are separated through thermal fluctuations that are facilitated by interactions of RNAP with ssDNA [4]. 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) [5]. After the open complex is formed, RNAP clears the promoter and enters the elongation, which leads to synthesis of RNA from DNA template [1].

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}) [6]. Interactions of σ70 with dsDNA of -35 element, extended -10 element, and -12 base of -10 element result in the closed complex formation [7]. On the other hand, the downstream bases of -10 element (-11 to -7) interact with σ70 in ssDNA form [7], 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 [6]: 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.

Table 1: Examples of promoter sequence elements

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
acnB	TTAACA	17	TGCT	ATTCT
adhEp1	CTAATG	17	TACT	ACAAT

3 Importance of the promoter element 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' [8-11] *ii*) accurate aligning of -10 element is complicated by both variable distance from -35 element and by a lower conservation of this element [9] *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 [11].

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* [12]. This alignment comes directly from experimentally determined TSS assembled in RegulonDB database [13]. 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.

Specificities of the aligned promoter elements are shown in Figure 1, which is generated by EnoLogos [14]. The overrepresentation of -35 element bases obtained from our alignment (Table 1 and Figure 1) is consistent with the available data on interactions between σ^{70} and -35 element [15]: The largest overrepresentation is obtained for bases -35, -34, -33 and -31, which are bound to σ subunit residues with hydrogen bonds; the overrepresentation is notably smaller for bases -32 and -30 which interact with σ^{70} with weaker van der Waals interactions. Finally, there is a statistically significant overrepresentation of G at position -36; this might seem unexpected, since position -36 is not part of -35 element; however, this conservation is consistent with the interaction data that indicate van der Waals interactions between -36 and σ^{70} residues [15].

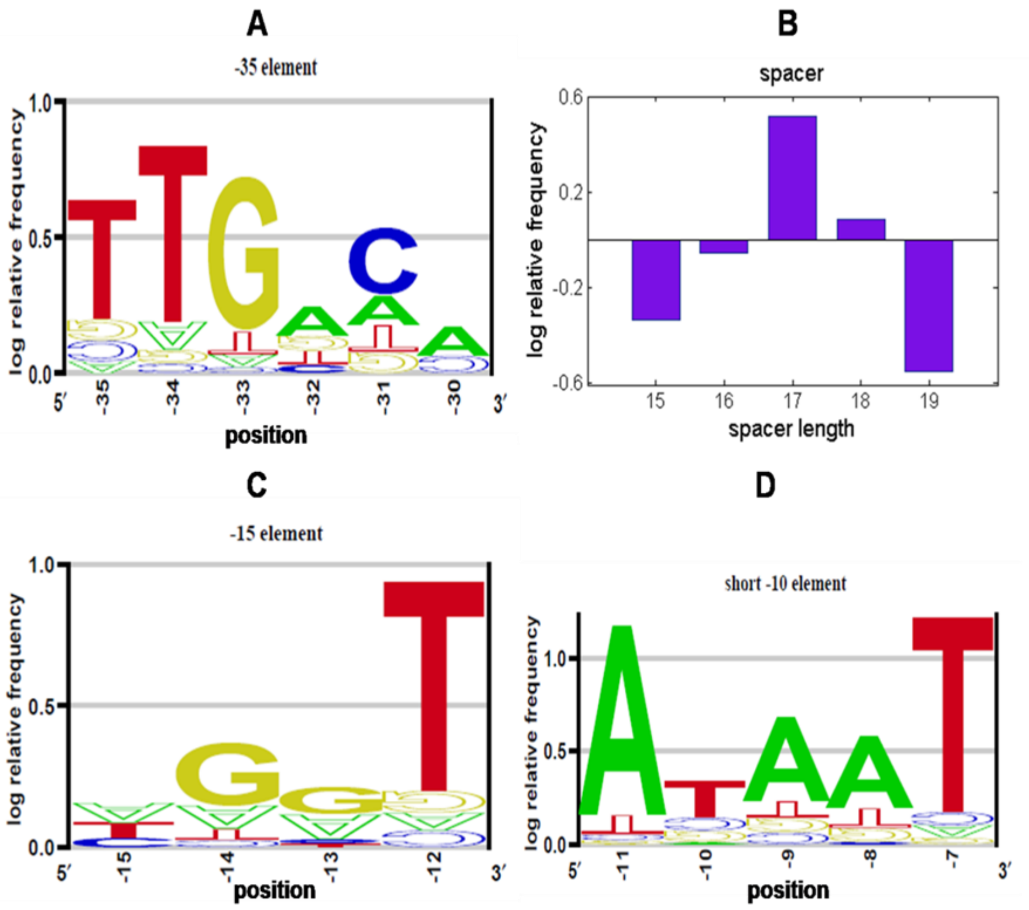

Figure 1: Logarithm of the ratio of the base frequencies in the alignment, relative to the background base frequencies is shown in the figure. For spacers (Figure 1B), log ratios are also presented, where the background distribution is equiprobable. Sequence logos correspond to specificities of A) -35 element B) Spacer between -35 and -10 element C) -15 element D) short -10 element. Figure adapted from [12].

We note that a recent alignment of -35 elements presented in [16] shows notable discrepancies with -35 element alignment presented here. Specifically, in [16] base 'C' at position -31 is significantly less conserved compared to 'A' at -32; this is inconsistent with the available data on interactions between σ^{70} and -35 element which indicate that base -31 interacts with σ^{70} through hydrogen bonds, while interactions with position -32 involve weaker van der Waals interactions. Furthermore, in [16] bases 'A' and 'T' show a larger conservation compared to 'C' and 'A' at positions -31 and -30, which is inconsistent with both the interaction data [15] and with -35 element consensus ('$^{-35}$TTGACA^{-30}') established through previous studies [6]; this is in contrast to our alignment where consensus '$^{-31}$C' and '$^{-30}$A' are clearly distinguished from the other bases at positions -31 and -30.

Our inferred specificity of -10 element is also consistent with available biophysical data: We see that the largest conservation corresponds to positions -11 and -7, which were shown in a number of studies to be of special importance for σ^{70}-ssDNA interactions (see e.g. [17, 18]). On the other hand, mutations at position -10 showed no notable effect on σ^{70}-ssDNA binding [18], consistent with the smallest base overrepresentation at this position.

Finally, we also briefly address specificity of the binding positions within -15 motif. We first note a high degeneracy at position -15 where bases T and C are similarly overrepresented (1.18 and 1.14 relative to the background frequencies). Therefore, it is more appropriate to represent the extended -10 motif with a weight matrix, or qualitatively with a degenerate consensus, than with a consensus sequence. Next, we note a conservation of base 'G' at position -13, which appears at the frequency that is 1.4 times larger than the background frequency, which is statistically highly significant (P~10^{-3}). We also note that conservation of the base at position -13 is larger than conservation at -15, which is a canonical base within the extended -10 motif (the 'T' in 'TG'). Conservation of base -13 at this position had not been reported before. Actually, the consensus sequence for the extended -10 motif is presented in the literature as 'TGn', where 'n' at position -13 indicates no conservation [6]. Consequently, we conclude that -15 element presents a conserved stretch of sequence, which has to be included in promoter search for a complete description of promoter specificity.

In summary, we have seen above that important parts of the previous alignments (in particular -35 element) are not in accordance with biophysical data of σ^{70} interactions. On the other hand, through unbiased alignment of the promoter elements one can obtain a significant improvement in the alignment, which is consistent with biophysical data. Moreover, such careful alignment also allows describing specificity of the sequences outside of canonical -10 and -35 box; such description can be included in TSS search in order to increase its specificity. Consequently, one can expect that a more accurate alignment of the promoter elements can significantly improve the prediction accuracy.

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)[5]. 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 $\mathrm{K_B}$), but has a low rate of transcription initiation due to a slow transition from closed to open complex (low $\mathrm{k_f}$). It has been proposed that poised promoters may present a major problem for accurate TSS prediction [9, 19]. This is particularly important, given the high number of false positives [2, 9, 10] that typically originate from computational TSS searches.

The number of poised promoters in the genome depends on mutual relation between the binding affinity K_B and the transition rate from closed to open complex k_f. In particular, RNAP binding affinity (K_B) depends on interactions of -10 box dsDNA with σ^{70} subdomain 2.4 [7], where the stronger interaction leads to larger binding affinity. On the other hand, stronger interaction of σ2.4 with dsDNA of -10 element leads to slower transition from closed to open complex[4]. The transition rate (k_f) also depends on interactions of -10 box ssDNA with σ^{70} subdomain 2.3 and on -10 element melting energy, both of which are physically independent from σ2.4[4, 20]. Due to this, K_B and k_f should *a priori* be negatively correlated, and there may be a large number of sequences in the genome that correspond to high K_B but low k_f. This naturally raises a possibility that there are many poised promoters in genome.

In order to quantitatively estimate the extent of RNAP poising in genome, one must be able to investigate kinetics of transcription initiation on a genome wide scale. This analysis cannot be done through experiments, since K_B and k_f have to be measured through work-intensive τ plot measurements [21], individually for each sequence of interest. In [22] we approached the problem computationally, where we used a recently developed biophysical model of the open complex formation[4], which allows the calculation of the kinetic parameters (K_B and k_f) for each sequence of interest. This model showed a very good agreement with both biochemical and genomics data, with no free parameters used in comparing the model with the experimental data[4].

To estimate the number of poised promoters in genome, we started from the sequence of lacUV5 promoter, and substituted its consensus -10 element with all 6bp long segments from *E. coli* intergenic regions. For all these substitutions we calculated the binding affinity (K_B) and the transcription initiation rate (φ), by using the method discussed above. In the widely used unsaturated approximation, transcription activity is proportional to the product of K_B and k_f , which is commonly defined as a measure of promoter strength. The relationship between logarithms of K_B and φ (approximated as the product of K_B and k_f) is shown in Figure 2, so that the quantities on the two axis correspond to the appropriate interaction energies that determine the relevant kinetic parameters. Specifically, the horizontal axis (log(K_B)) corresponds to σ70-dsDNA binding energy, while the vertical axis corresponds to a combination of the energy terms that we refer to as the effective energy and which directly determines the transcription initiation rate.

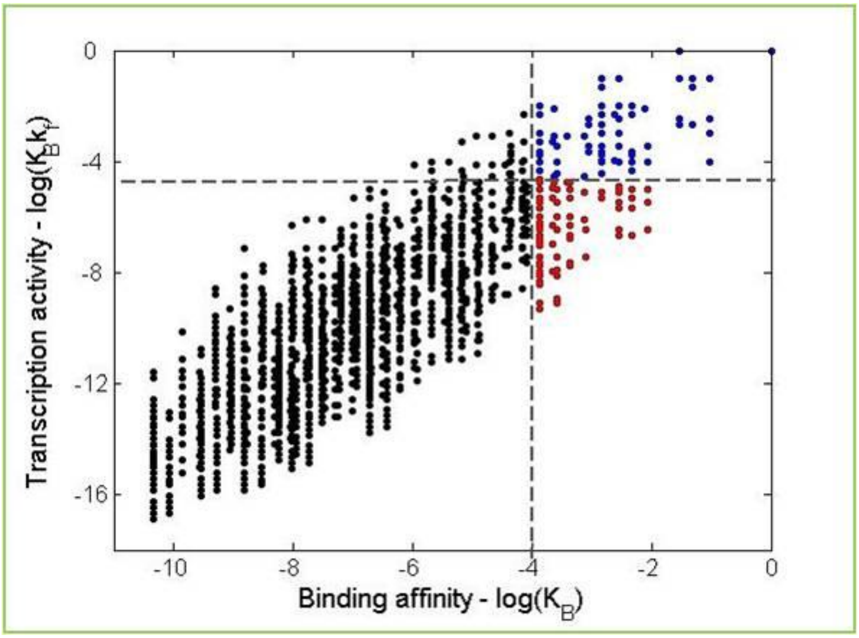

Figure 2: Log transcription rate ($\log(\varphi)$) vs. log binding affinity ($\log(K_B)$) for the intergenic segments. -10 element of lacUV5 promoter is substituted by all 6bp long segments from *E. coli* intergenic regions. $\log(K_B)$ and $\log(\varphi)$ are calculated for each of these substitutions 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 [22].

Both K_B and φ, which are shown in Figure 2, are calculated relative to the binding affinity and the transcription initiation rate of lacUV5 promoter. Note that we substitute (vary) only -10 element of lacUV5 promoter, and that -10 element of this promoter corresponds to the consensus sequence ('$^{-12}$TATAAT^{-7}'). Consequently, zeros on the horizontal and the vertical axis correspond to the consensus -10 element, and stronger interaction energies correspond to larger (less negative) values on the two axes. The horizontal line in Fig. 2 (transcription rate threshold) indicates the transcription rate below which transcript levels cannot be detected, while the vertical line (binding threshold) indicates the binding affinity above which a sequence is considered to be strongly bound by RNAP. The binding threshold is set so that it corresponds to the binding affinity of a weak Plac promoter, which is in accordance with an intuitive notion that functional promoters - a large majority of which bind more strongly than Plac - should be above the binding threshold.

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.

5 Relevant kinetic parameter

In the previous section we reviewed evidence showing that kinetic effects must be taken into account for accurate TSS detection. From this, however, follows a question of which exactly kinetic parameters characterize functional promoters. For example, the recent `mix-and-match' model proposes that the promoter elements, which interact with RNAP in dsDNA form, complement each other strengths so as to achieve a sufficient value of K_B; this would imply that the relevant kinetic parameter is the binding affinity K_B, though this issue is clearly untested and remains open. Testing this issue is further discussed below.

In [12] we formed weight matrices for each of the promoter elements, by using the alignment shown in Fig. 1. Each of the weight matrices are then used to calculate strengths of the promoter elements obtained in the alignment. Consequently, the weight matrices are used to calculate strengths of -35 elements, -15 elements, short -10 elements and the overall promoter strength. The estimated strengths of -15 promoter elements are plotted against the corresponding strengths of -35 element and short -10 element, which is shown in Figure 3A and 3B.

Figure 3A and 3B show that strengths of short -10 element and -35 element are negatively correlated with the strength of -15 element. There is consequently a tendency to have a stronger -15 element when weaker -10 element or weaker -35 element are present. We furthermore see that the negative correlation is stronger for -10 elements than for -35 elements: In the case of -35 element we have a correlation constant of -0.10, which is marginally significant (P value of 0.06); the correlation in the case of short -10 element is -0.17, which is highly statistically significant (P value of $2*10^{-3}$). The stronger correlation in the case of -10 element seems surprising, having in mind that both -35 and -15 element are involved in RNAP-dsDNA interactions, while short -10 element is involved in the open complex formation through RNAP-ssDNA interactions. This issue will be further discussed below.

Figure 3C shows correlation of -15 element strength with overall promoter strength in the absence of -15 element. The overall promoter strength in the absence of -15 element is estimated as a sum of strengths that correspond to -35 element, strength of short -10 element and the spacer weight. We see a highly significant negative correlation between -15 element strength and the overall promoter strength (correlation constant of -0.20, with P value of $3*10^{-4}$); this correlation is stronger than for the individual promoter elements. One should also note that the strength of σ^{70}-dsDNA interactions in the absence of -15 element is simply given by the strength of -35 element. Therefore, by comparing Figure 3A and 3C, we see that a much stronger negative correlation is associated with the overall promoter strength than with σ^{70}-dsDNA interactions. Consequently, the results imply that it is the total promoter strength (which approximately corresponds to the transcription activity), rather than the binding affinity to dsDNA, which defines a functional promoter.

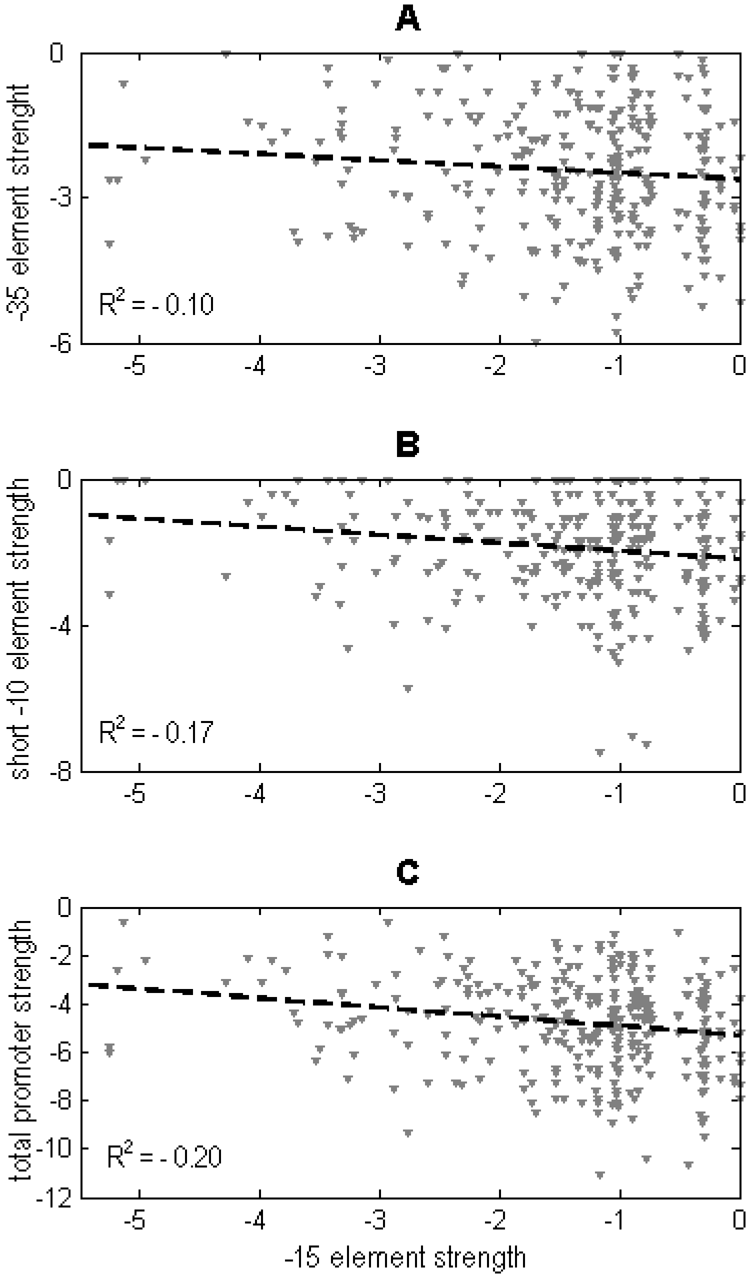

Figure 3 - Correlation of -15 element strength with other promoter strengths. Weight matrices inferred from the alignment are used to calculate strengths of -35 elements, -15 elements, short -10 elements, and the overall promoter strength. Correlations between the following strengths are then shown: A) -15 element and -35 element B) -15 element and short -10 element C) -15 element and the overall promoter strength. Correlation constants are indicated in the figures. Figure adopted from [12].

6 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. The last goal remains the most challenging; with respect to that, note that while interactions of RNAP with -10 element sequence are extensively characterized by experiments, such information is absent for interactions with -35 and -15 elements. Consequently, one has to use mixed (experimental and computational) parametrization of the biophysical model, so that interactions with -10 element come from the experimental data, while interactions with -35 and -15 elements are inferred from sequence data by statistical means. Developing such method for accurate TSS detection, which is based on explicit calculation of promoter kinetic parameters, is our current goal.

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

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

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

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

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

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

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

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

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

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

[10] K. Robison, A. McGuire, and G. Church. 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.

[11] J. E. Mitchell, D. Zheng, S. J. W. Busby, and S. D. Minchin. Identification and analysis of ‘extended–10’promoters in Escherichia coli. *Nucleic Acids Research,* 31:4689, 2003.

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

[13] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, and A. López-Fuentes. 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.

[14] C. T. Workman, Y. Yin, D. L. Corcoran, T. Ideker, G. D. Stormo, and P. V. Benos. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Research,* 33:W389-392, 2005.

[15] E. A. Campbell, O. Muzzin, M. Chlenov, J. L. Sun, C. A. Olson, O. Weinman, M. L. Trester-Zedlitz, and S. A. Darst. Structure of the bacterial RNA polymerase promoter specificity [sigma] subunit. *Molecular Cell,* 9:527-539, 2002.

[16] R. K. Shultzaberger, Z. Chen, K. A. Lewis, and T. D. Schneider. Anatomy of Escherichia coli s 70 promoters. *Nucleic Acids Research,* 35:771-788, 2007.

[17] D. L. Matlock and T. Heyduk. Sequence determinants for the recognition of the fork junction DNA containing the-10 region of promoter DNA by E. coli RNA polymerase. *Biochemistry,* 39:12274-12283, 2000.

[18] M. S. Fenton and J. D. Gralla. Function of the bacterial TATAAT-10 element as single-stranded DNA during RNA polymerase isomerization. *Proc Natl Acad Sci USA,* 98:9020-9025, 2001.

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

[20] K. S. Murakami, S. Masuda, E. A. Campbell, O. Muzzin, and S. A. Darst. Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA Complex. *Science,* 296:1285-1290, 2002.

[21] W. R. McClure. Rate-Limiting Steps in RNA Chain Initiation. *Proc Natl Acad Sci USA,* 77:5634-5638, 1980.

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