DOI 10.18699/SBB-2022-034

Promoter motif inference and annotation of promoter sequences in bacterial genomes based on the analysis of structures of alternative sigma factor-promoter complexes

Gromyko D.*, Vychyk P., Nikolaichik Y.
Department of Molecular Biology, Belarusian State University, Minsk, Belarus
** grom.dima.grom@gmail.com*

Key words: sigma factor, promoter, DNA motif, DNA binding domain

Motivation and Aim: Bacterial genome annotation pipelines can handle protein-coding sequences, rRNA and tRNA genes, but no pipeline deals with regulatory sequences. To close this void, we developed SigmoID [1, 2] software for regulatory sequence analysis and annotation in bacterial genomes. SigmoID can annotate transcription factor (TF) binding sites and terminators. It also contains a pipeline for *de novo* inference of TF binding site motifs, which could be used for annotating binding sites for uncharacterised TFs. The idea behind this pipeline relies on the conservation of positions of DNA bases-contacting amino acid residues of the TF. These conserved residues constitute the CR tag – the unique identifier of TF pair with its binding site. Due to the autoregulatory nature of most TFs, their DNA recognition motif could be inferred within a set of regulatory sequences of genes coding TFs with the same CR tag.
The same approach faces difficulties if applied to promoter identification due to a different mode of promoter recognition. Rather than making few specific contacts within a single short region of double-stranded DNA, sigma factors make many more base contacts within at least two promoter regions. In addition, many specific contacts are made within the single-stranded region of open promoter complexes. Therefore, CR tags derived from 3D structures of sigma factors bound to their promoters are excessively long for our pipeline resulting in failure due to insufficient diversity of the regulatory regions it collects. This work aims at modifying our pipeline to reliably infer promoter sequence motifs. We approach this problem by carefully analysing known sigma factor-promoter contacts and calculating covariation between CR tag and promoter sequence changes. We show that restricting the CR tag to the most significant positions allows efficient inference of promoter motifs for four families of sigma factors.
Methods and Algorithms: As changes in the promoter DNA sequences are followed by amino-acid replacements in the DNA-binding domain and *vice versa*, a correlation between residues in DNA-binding domains and positions in promoter sequences could reveal residues most important for specific promoter recognition. We looked for correlations with Prot-DNA-Korr [3], which counts Z-score for co-correlating amino-acid residue – DNA nucleotide pairs. Data sets for Prot-DNA-Korr were acquired via a custom Python script that collects sigma factor sequences and their promoter regions from the GenBank database and splits them into matching R2 domain/-10 region and R4 domain/-35 sets. PFAM and ECFhub were used as the source for the sigma factor family and promoter hidden Markov models. Collected sequences were aligned with Clustal Omega and further processed with Prot-DNA-Korr.

Results: To test our approach, four enterobacterial sigma factor families were selected. The main selection criterion was the availability of 3D structures of initiator complexes and characterized promoters. The four families had significant differences in DNA recognition mechanisms: ECF factors are characterized by the typical contacts described above between the promoter and the SR2 and SR4 domains, FliA has an insertion of the SR3 domain, also capable of contacting the promoter, while both DNA-binding domains of RpoN interact with double-stranded DNA.

With the help of Prot-DNA-Korr, we determined CR tag positions that are most important for specific promoter recognition (Fig. 1). This analysis allowed us to define CR tags for four sigma factor families: RpoE, HrpL, FliA and RpoN. The resulting CR tags have shortened from over 40 original contacts to 19 (or less) most important ones. The shortened CR-tags allowed the SigmoID pipeline to successfully infer promoter motifs (Fig. 2) with high similarity to those described in the literature [4–8].

Fig. 1. Heatmap of correlations between SR2 domain amino acids and -10 promoter region nucleotides for RpoE sigma factor family

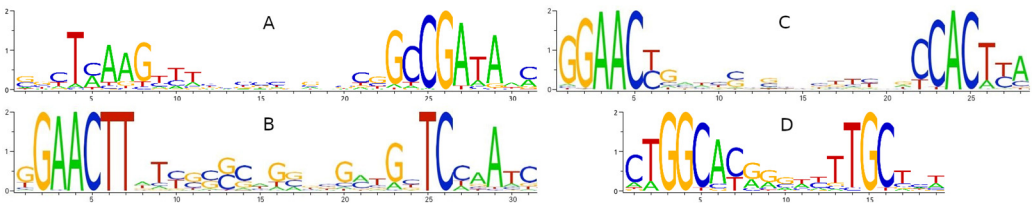

Fig. 2. Inferred promoter motifs for FliA (A), RpoE (B), HrpL (C) and RpoN (D)

Conclusion: We successfully expanded SigmoID annotation capabilities, adding support for alternative sigma factor promoter inference. We are currently working to extend this approach to other families of bacterial sigma factors. Modified SigmoID program code and calibrated profiles for identifying promoters of five alternative sigma factors are available in the repository github.com/nikolaichik/sigmoid. Calibrated profiles are suitable for automated annotation of promoters in bacterial genome sequences.

References

1. Nikolaichik Y., Damienikan A.U. *PeerJ.* 2016;4:E2056.
2. Nikolaichik Y., Vychyk P. Abstr. BGRS/SB-2020:5-76. doi: 10.18699/BGRS/SB-2020-046.
3. Korostelev Y.D. et al. *PLOS One*. 2016;11(9) E0162681.
4. Lam H.N. et al. *PLoS One*. 2014;9(8):E106115.
5. Fitzgerald D.M. et al. *Mol Microbiol.* 2018;108(4):361-378.
6. Missiakas D., Raina S. *Mol Microbiol*. 1998;28(6):1059-1066.
7. Markel E. et al. *J Bacteriol*. 2016;198(17):2330-2344.
8. Wilson M.J., McMorran B.J., Lamont I.L. *J Bacteriol*. 2001;183(6):2151-2155.