DOI 10.18699/SBB-2022-054

BacRegDB: a database of bacterial regulatory elements with structural evidence

Vychyk P.[1]*, Duvalov E.[2], Digris A.[2], Skakun V.[2], Nikolaichik Y.[1]

[1] *Department of Molecular Biology, Belarusian State University, Minsk, Belarus*

[2] *Department of Systems Analysis and Computer Modelling, Belarusian State University, Minsk, Belarus*

** p.vychik@gmail.com*

Key words: genome annotation, transcription factor binding sites

Motivation and Aim: None of the current genome annotation pipelines handles regulatory elements like promoters and operators. Although the information about bacterial regulatory elements is collected in several specialized databases, only part of it can be used to annotate regulatory elements, and only in a very limited range of bacteria. Most importantly, no reliable criteria is used for the transfer of regulatory information between genomes. Such transfer is usually based on thorough homology evaluation and heavily depends on human judgement.

Our aim in this work was to develop a new transcription factor binding site database permitting reliable automated transfer of regulatory information between bacterial genome sequences.

Methods and Algorithms: We have previously proposed a 3D structure based strict formal criterion for applying regulatory information to any bacterial genome [1]. This criterion is the CR-tag: the amino acid residues of a transcriptional regulator that specifically contact the nitrogenous bases of the regulatory element in genomic DNA. A motif associated with the CR tag can be correctly applied to annotate such elements in genomes encoding transcriptional regulators with identical CR tags.

The accumulation of motifs associated with CR tags demanded their ordered storage for the convenience of subsequent use in genome annotation. Since none of the known databases use the CR tag concept, a new relational database was developed and implemented in Microsoft SQL Server.

Original regulatory motif information was collected either from the literature of from the regulatory motif databases RegulonDB, CollecTF, Prodoric2, CoryneRegNet [2–5]. For each transcription factor present in these databases, our automated *de novo* transcription factor binding site inference pipeline [1] was run to collect transcription factor gene-linked operators from all genomes encoding TFs with the same CR tag. If the resulting *de novo* inferred motif matched the one present in the databases or described in the literature, automatically collected and experimentally characterised operators were combined in the final profile. Hidden Markov models for each motif were built with threshold cutoff scores manually set to find known operators with minimum or no false positive hits.

A fully automated import procedure was developed within the framework of our SigmoID application [6] for RegPrecise, the largest bacterial regulatory database [7]. Little or no manual curation was performed for the resulting motifs at this time. Cutoff scores for RegPrecise-derived operator profiles were calculated based on manually curated data via regression curves of motif information content versus bit score thresholds.

Results: The database currently covers only transcription factor binding sites and has two divisions: the core and the extended collection. The core includes 237 regulatory elements and has undergone full manual curation including verification of experimental evidence and determination of threshold scores for HMM models of each element. Each core record has manually assigned experimental evidence codes and is linked to the corresponding literature. The extended collection includes information on over 3000 regulatory elements exported from various databases with majority of information coming from RegPrecise. We noted that RegPrecise regulogs quite often combine regulons of TFs with different CR tags. On the other hand, many regulons of TFs with identical CR tags are present in RegPrecise as separate entries. Our RegRrecise import procedure takes special care to correctly split mixed regulogs and combine regulons with the same CR tag, thus improving the quality of the resulting regulatory motifs. The full BacRegDB content is currently accessible via the standalone SigmoID application (https://github.com/nikolaichik/SigmoID) and can be used for annotating regulatory elements in bacterial genomes in semi-automated fashion. We are also completing the development of the web application interfacing BacRegDB and allowing regulatory element annotation in user-submitted genomes.
Conclusion: The main advantage of BacRegBD is the CR-tag concept – a fingerprint uniquely matching transcription factors with their operators. All regulatory motif records in the database are associated with a CR tag and, therefore, can be correctly used to annotate similar elements in any genomes encoding a transcriptional regulator with an identical CR tag.

References

1. Nikolaichik Y., Vychik P. New approach to genome-wide automated inference of bacterial transcription factor binding sites. In: Bioinformatics of Genome Regulation and Structure/Systems Biology. 2020;75-76. DOI: 10.18699/BGRS/SB-2020-046.
2. Santos-Zavaleta A. et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Res.* 2019;47(D1):D212-D220.
3. Kılıç S. et al. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Res.* 2013;42(D1):D156-D160.
4. Dudek C.-A., Jahn D. PRODORIC: state-of-the-art database of prokaryotic gene regulation. *Nucleic Acids Res.* 2022;50(D1):D295-D302.
5. Parise M.T.D. et al. CoryneRegNet 7, the reference database and analysis platform for corynebacterial gene regulatory networks. *Sci Data*. 2020;7(1):142.
6. Nikolaichik Y., Damienikan A.U. SigmoID: a user-friendly tool for improving bacterial genome annotation through analysis of transcription control signals. *PeerJ*. 2016;4:e2056.
7. Novichkov P.S. et al. RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics*. 2013;14:745.