Community-based approach for genome-wide regulon annotation in Bacteria

Dmitry A. Ravcheev[1,2]* (dravcheev@sanfordburnham.org), Aaron A. Best,[3] Mikhail S. Gelfand,[2,4] **Pavel S. Novichkov**,[5] **Dmitry A. Rodionov**[1,2]

[1]Sanford-Burnham Medical Research Institute, La Jolla, CA; [2]Institute for Information Transmission Problems RAS, Moscow, Russia; [3]Hope College, Holland, MI; [4]Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; [5]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA

http://regprecise.lbl.gov

Project Goals:

1. **Develop integrated platform for genome-scale regulon reconstruction utilizing the comparative genomics approach and analysis of RNA and DNA regulatory sites.**
2. **Infer the regulatory interactions and reconstruct transcription regulatory networks (TRNs) in several groups of microbial species important for DOE mission.**
3. **Develop RegKnowledgeBase on microbial transcriptional regulation.**

In the light of the constantly growing number of complete genomic sequences accurate genome-scale annotation of regulatory features is a one of the critical task of modern genomics and system biology. Inference of regulatory interactions in newly sequenced genomes allows reconstruction of the transcriptional regulatory networks (TRNs) and effective modeling of cellular metabolic and signal pathways. TRN is a fine-tuned system including controllers, such as transcription factors (TFs), their binding sites (TFBSs), and RNA regulatory elements (e.g., riboswitches), and sets of target genes whose expression is regulated by these controllers. Genome-scale reconstruction of TRNs in bacteria requires accurate prediction of a large number of regulatory interactions between controllers and their targets. A regulon, defined as a set of genes under the direct control of a certain controller, constitutes a 'building block' of each TRN. The comparative genomics approach was successfully applied for reconstruction of multiple regulons in diverse bacterial groups. A regulog concept is used to represent a regulon inferred and projected in a group of closely-related bacterial genomes. High-quality reconstruction of each regulog constituting the combined TRN in a group of taxonomically-related genomes is a labor intensive work that can be carried out by the scientific community.

We developed a workflow for coordinated reconstruction of multiple bacterial regulons by the community annotators (Fig. 1). The workflow utilizes the comparative genomics-based approach implemented in the RegPredict Web-server (http://regpredict.lbl.gov). This reconstruction pursues two main tasks: (i) propagation of the known regulon to new genomes; (2) expansion of the known regulon by prediction of new regulon members. Initial datasets of regulatory interactions including information about TFs, the previously mapped TFBSs and sets of regulated genes (mostly experimentally determined) are extracted from literature and public databases, such as DBD, EcoCyc, CoryneRegNet, DBTBS, RegTransBase, and RegPrecise. These data are mapped to complete genomes forming initial sets of tasks for community annotators. Each regulon is reconstructed by a community annotator in the reference set of taxonomically-related genomes using the computational platform RegPredict. After the completion of regulon annotation task, each output regulog undergoes quality control by curators. Finally approved regulogs include information on regulators and their targets in a group of taxonomically-related genomes.

The community-based workflow was efficiently used for initial TRN inference in four taxonomic groups of bacteria, including Enterobacteriaceae, Lactobacillaceae, Streptococcaceae, and Corynebacteriaceae. The reference set of inferred transcription factor regulogs is available in the RegPrecise database (http://regprecise.lbl.gov) and includes 173 regulogs described in 50 genomes.

In the Enterobacteriaceae lineage (12 species including *Escherichia coli*), 64 regulons were reconstructed by a community of 27 undergraduate students from Moscow State University (MSU) in Russia. These include the global regulons ArgR, Crp, Fur, FruR, LexA, and PurR that control from tens to a hundred of targets per genome. The obtained collection of the Enterobacteriales regulogs contains more than 600 target genes, including ~100 new regulon members. Another community of 36 MSU students performed annotation of 43 regulons in the Corynebacteriaceae lineage (8 species including *Corynebacterium glutamicum*). The final set of reconstructed regulons comprises more than 150 target operons, including multiple novel members of regulons. Streptococcaceae and Lactobacillaceae are two closely related lineages of the Firmicutes phylum. Regulon reconstruction in 30 selected genomes from these two taxonomic groups was carried by a community of 18 undergraduate students from Hope College (Holland, MI). As a result of this community effort, 33 regulons were inferred in each of these two lineages. Among them, 22 regulons are shared between the two lineages, whereas the remaining 18 regulons are lineage-specific. Final collections of regulons in Streptococcaceae and Lactobacillaceae contain 200-300 regulated operons per genome with multiple novel predicted regulon members.

The community-based approach for *in silico* reconstruction of multiple regulons applied in this work to four taxonomically diverse groups of bacterial genomes is a promising approach for large-scale annotation of regulatory features. Detailed TRNs obtained by this approach for DOE-mission genomes will constitute an important dataset for the forthcoming DOE Systems Biology KnowledgeBase.

This research is supported by the Genomic Science Program (GSP), Office of Biological and Environmental Research (OBER), U.S. Department of Energy (DOE), under contract DE-SC0004999 with Sanford-Burnham Medical Research Institute and Lawrence Berkeley National Lab.
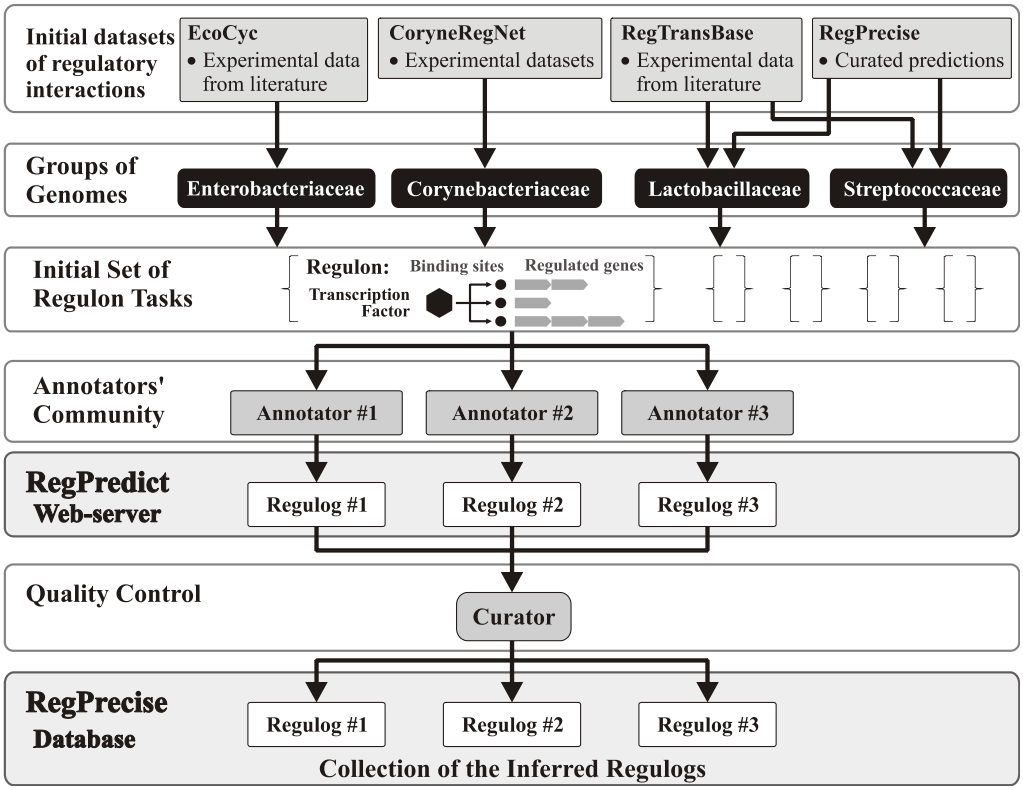
Fig.1. Workflow for genome-wide regulon annotation by scientific community