Resource for the exploration of regulons accurately predicted by the methods of comparative genomics.

Pavel S. Novichkov[1*], Mikhail S. Gelfand[2], Inna Dubchak[1,4], Dmitry A. Rodionov[2,3,*]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia; [3]Burnham Institute for Medical Research, La Jolla, CA 92037,USA; [4]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA.

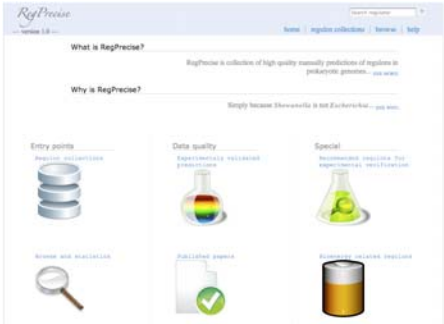
* PSNovichkov@lbl.gov, rodionov@burnham.org

Overview

Identification and reconstruction of various transcriptional regulons in bacteria using a computational comparative genomics approach is coming of age. During the past decade a large number of manually-curated high quality inferences of transcriptional regulatory interactions were accumulated for diverse taxonomic groups of bacteria. These data provide a good foundation for understanding molecular mechanisms of transcriptional regulation, identification of regulatory circuits, and interconnections among circuits within the cell. Traditional experimental methods for regulon analysis have certain limitations both in terms of productivity and feasibility. While the development of high-throughput transcriptome approaches allow to obtain genome-scale gene expression patterns, in many cases the complexity of the interactions between regulons makes it difficult to distinguish between direct and indirect effects on transcription. The availability of a large number of closely related genomes allows one to apply comparative genomics to accurately expand already known regulons to yet uncharacterized organisms, and to predict and describe new regulons. Due to fast accumulation of such valuable data, there is a need for a specialized database and associated analysis tools that will compile and present the growing collection of high quality predicted bacterial regulons.

The RegPrecise database was developed for capturing, visualization and analysis of transcription factor regulons that were reconstructed by the comparative genomic approach. The primary object of the database is a single regulon in a particular genome, which is described by the identified transcription factor, its DNA binding site model (a profile), as well as the set of regulated genes, operons and associated operator sites. Regulons for orthologous transcription factors from closely related genomes are combined into the collections that provide an overview of the conserved and variable components of the regulon. A higher level representation of the regulatory interactions is also provided for orthologous regulons described in several bacterial taxonomic groups enabling comparison and evolutionary analysis of the transcription factor binding motifs. Another view of complex data in the database is a general overview of multiple regulons inferred in a set of closely related group of genomes.

http://regprecise.lbl.gov

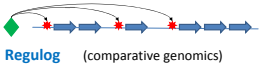


Data structure & statistics

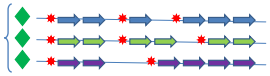
Transcription factors (TFs) are proteins that recognize specific binding sites to either activate or repress transcription of genes.

Binding sites (TFBSs) (or operators) are *cis*-regulatory DNA sequences located in 5'-untranslated regions (UTRs) of genes.

Regulon (single genome) is a set of genes in a particular genome directly regulated by the same transcription factor



Regulog (comparative genomics) is a group of regulons operated by orthologous TFs in closely-related genomes.
- involves set of genes with conserved TFBSs.
- is associated with a distinct TFBS-recognition profile.



Collections of regulogs

Regulogs are organized into collections of three major types:

By Taxonomic group
Complete set of regulogs analyzed within a particular taxonomic group of closely-related organisms.

By Transcription factor
Set of similar regulogs operated by orthologous TFs across various taxonomic groups of organisms.

By Functional classification
Set of various regulogs that control genes from a particular metabolic pathway or biological process.

Database content	
Transcription factors	87
Regulogs /profiles	169
TF families	27
Genomes	278
Binding sites	6223

Basic elements

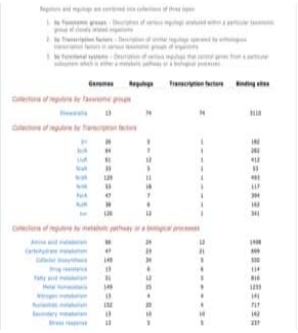
Regulog page: a comparative genomic view on
- a current reconstruction of orthologous TF regulons in a set of closely related genomes;
- a common DNA binding site profile (logo).

Regulon page: a detailed genomic view on
- a regulon gene content in a selected genome;
- individual DNA binding sites located in 5'-UTRs;
- a binding site profile (logo) used in the analysis.

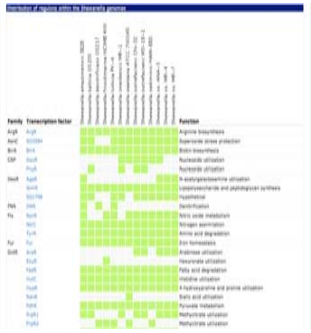
Binding sites page: provides
- an alignment of all DNA binding sites described for a particular regulog in a set of genomes;
- a binding site profile (logo) used in the analysis.

Regulog Collections

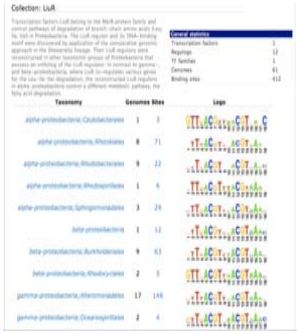
Main page with a list of collections:
- shows the essential content of the database classified by regulon collections of three major types.



Collections of regulons by Taxonomic groups:
- Currently involves a single collection of *Shewanella* regulons;
- Shows distribution and properties of regulons across taxonomic group;
- Allows to overview the regulatory network described for a group of genomes and analyze a statistics by genomes, TF families, etc..



Collections of regulons by Transcription factor:
-shows the taxonomic distribution of a TF in the genomes;
-combines multiple regulogs operated by orthologous TFs;
- allow to analyze evolution of binding site profiles (logos) for a TF across various taxonomic groups of genomes.



Collections of regulons by Functional classification:
- provides list of regulogs for a particular metabolic pathway or a biological process;
- shows an effector attributed to a particular TF and statistics.

Quality of data

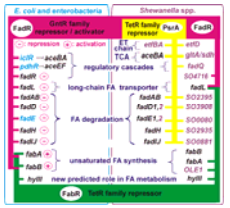
Manual curation of data represented in RegPrecise is essential to challenge the flexibility of regulation across multiple genomes

***Shewanella* is not *Escherichia* ...**

- Thorough comparison of the reconstructed regulons in *Shewanella* species with known regulons in model bacterium *Escherichia coli* demonstrated striking differences in the overall regulatory strategy in these two lineages of gamma-proteobacteria. Multiple interesting trends in diversification and adaptive evolution of TRNs between lineages were detected including regulon "shrinking", "expansion", "mergers", and "split-ups", as well as multiple cases of using nonorthologous regulators to control equivalent pathways or orthologous regulators to control distinct pathways.

- Bacterial transcriptional regulation is very flexible and mostly not conserved between various taxonomic groups. Therefore investigation of regulation by computational genomic approaches need to be done with careful curation of data. RegPrecise database is a depository of manually curated genomic-based reconstructions of bacterial regulons.

Example of variability of regulons for Fatty acid metabolism



The fatty acid degradation *fad* genes are controlled by two nonorthologous regulators, FadR in *E. coli* and PsrA in *Shewanella*. Novel regulon PsrA controls not only *fad* genes in *Shewanella* but also other genes from the central metabolism. Though a FadR ortholog is present in *Shewanella*, its regulon is significantly reduced (2 targets). The fatty acid biosynthesis regulon FabR has both conserved and variable parts between *E. coli* and *Shewanella*.

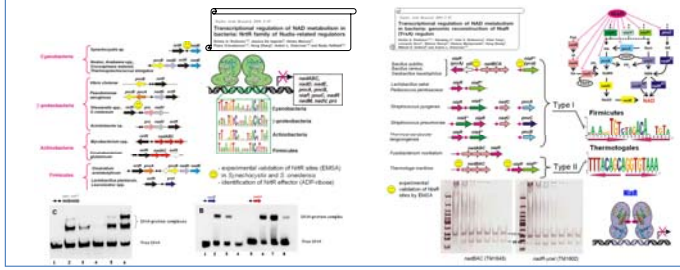
Experimentally verified regulon predictions

NrtR regulon for NAD metabolism

A novel regulator for the NAD biosynthesis (termed NrtR) displays a mosaic distribution across taxonomic groups of bacteria, and shows a substantial variability in its DNA binding profiles and regulon composition. NrtR was shown to operate as a ADP-ribose-responsive transcriptional repressor of the NAD biosynthesis and salvage operons. Using EMSA assays, the predicted NrtR binding sites in two bacterial genomes (*Synechocystis* and *Shewanella oneidensis*) were validated.

NiaR regulon for NAD metabolism

A novel regulator for the NAD biosynthesis (termed NiaR) was identified in Firmicutes and Thermotogales. Using EMSA assay, NiaR was shown to operate as a niacin-responsive transcriptional repressor of the NAD biosynthesis operon *nadABC*. Detailed analysis of the NiaR regulon by a combination of comparative genomic and experimental techniques allowed us to refine its DNA-binding motifs and identify additional regulon members in bacteria.



The manually curated regulatory data were successfully applied for reconstruction of multiple biological networks and metabolic pathways.

List of selected publications (from more than 25)

Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. Rodionov DA, Mironov AA, Gelfand MS. (2002) Genome Res, 12: 1507:16.

Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. Rodionov DA, Dubchak I, Arkin A, Alm E, Gelfand MS. (2004) Genome Biol, 5: R90.

Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling. Rodionov DA, and Gelfand MS. (2005) Trends Genet, 21: 385:9.

Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. Rodionov DA, Dubchak IL, Arkin AP, Alm EJ, Gelfand MS. (2005) PLoS Comput Biol, 1: e55.

Computational identification of BioR, a transcriptional regulator of biotin metabolism in Alphaproteobacteria, and of its binding signal. Rodionov DA, Gelfand MS. (2006) FEMS Microbiol Lett, 255: 102:7.

Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of Shewanella oneidensis. Yang C, Rodionov DA, Li X, Laikova ON, Gelfand MS, Zagnitko OP, Romine MF, Obraztsova AY, Nealson KH, Osterman AL. (2006) J Biol Chem, 281: 29872:85.

Comparative Genomics of Regulation of Fatty Acid and Branched-chain Amino Acid Utilization in Proteobacteria. Kazakov AE, Rodionov DA, Alm E, Arkin AP, Dubchak I, Gelfand MS. (2008) J Bacteriol, 190: Sep 26.

Acknowledgments

We are grateful to Andrey Osterman for useful discussions, Andrei Mironov for software for genome analysis, and Katya Panina, Alexey Kazakov, Elizaveta Permina, and Anna Gerasimova for contributions to regulon reconstructions.

This project was supported by US Department of Energy (DOE).

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.