



# Progress in the development of the RegTransBase database and the comparative analysis system

Michael Cipriano, Pavel Novichkov, Alexey Kazakov, Dmitry Ravcheev, Adam Arkin, Mikhail Gelfand\*, Inna Dubchak\* (gelfand@litt.ru, ildubchak@lbi.gov)  
Genomics Division, Lawrence Berkeley National Laboratory, The Virtual Institute of Microbial Stress and Survival, The Research and Training Center on Bioinformatics (Moscow)



## Overview



<http://regtransbase.lbl.gov>

RegTransBase, a database describing regulatory interactions in prokaryotes, has been developed as a component of the MicrobesOnline/RegTransBase framework successfully used for interpretation of microbial stress response and metal reduction pathways. It is manually curated and based on published scientific literature. RegTransBase describes a large number of regulatory interactions and contains experimental data which investigates regulation with known elements. It is available at <http://regtransbase.lbl.gov>.

Over 1000 additional articles were annotated last year resulting in the total number of 5118 articles. We specifically focused on annotating the facts of regulation in energy-related bacteria such as: Clostridia, Thermoanaerobacter, Geobacillus stearothermophilus, Zymomonas, Fibrobacter, Ruminococcus, Prevotella, Acetobacter, Anaeromyxobacterium, Streptomyces, Ralstonia.

Currently, the database describes close to 12000 experiments (30% growth in the last year) in relation to 531 genomes. It contains data on the regulation of ~39000 genes and evidence for ~10000 interactions with ~1130 regulators. We removed redundancy in the list of Effectors (currently the database contains about 630 of them) and turned them into controlled vocabulary.

RegTransBase additionally provides an expertly curated library of 150 alignments of known transcription factor binding sites covering a wide range of bacterial species. Each alignment contains information as to the transcription factor which binds the DNA sequence, the exact location of the binding site on a published genome, and links to published articles. RegTransBase builds upon these alignments by providing a set of computational modules for the comparative analysis of regulons among related organisms.

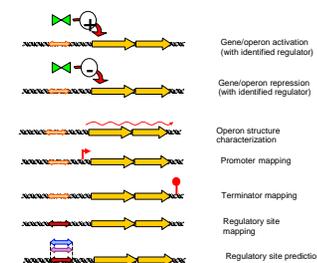
The new tool - "advanced browsing" was developed to allow a user to search the data contained in RegTransBase in a step-by-step manner. Different types of classifications, such as taxonomy, effectors, the type of experimental result, a phenotype, and genome relevance allow for creating and applying complex search criteria. We are planning to include additional classifications such as metabolic pathways and types of experimental techniques in this scheme.

There is an increasingly tight coupling of RegTransBase with MicrobesOnline in reporting cis-regulatory sites and regulatory interactions, and integrating RegTransBase searches into MicrobesOnLine cart functions.

Main page of RegTransBase.

## Data

A decision of whether to include each putative site in a particular regulon is made after consultation with scientific literature by a human expert. RTB contains the following types of experimental data:



Classifications: multi-step search, Experiments in Proteobacteria

## Complex queries

## Information

RegTransBase contains structured information obtained directly from experiments explained in published literature. Articles contain multiple experiments. Each experiment contains multiple elements that make up that experiment. Elements themselves can have a hierarchical relationship (operons-genes). Elements may be linked to other elements (sites are linked to regulators). We provide the tools to view this experiment, and then obtain a global view of the genomic region. view features/elements in that region. list effectors that act on these elements. Provide tools for comparisons between species.

The correlation between an article/experiment and how it appears in RTB. a) An actual article, b) Experiment view, c) Element view, d) Site view, e) Genome view using Gbrowse, f) Graphviz diagram based around the relationship of elements described in literature, g) View of the VISTA Genome Browser comparing the genomes of multiple species.

## Comparison

RTB has a manually curated collection of close to 200 position weight matrices and alignments. We provide the ability to search sequenced genomes using these matrices or user-submitted alignment. Using a collection of interfaces we aim to provide a tool for the following situations:

- One matrix + one genome of interest
  - Show predicted binding sites which match this matrix, while providing additional information.
- One gene + multiple genomes
  - Predict binding sites for orthologous genes using certified matrices.
- One matrix + multiple genomes
  - Compare the predicted binding sites across genomes for a particular matrix, highlighting orthogonal similarities.
- Multiple matrices + multiple genomes
  - Compare the predicted binding sites across genomes for a set of matrices.

These tools guide a user through a semi-automated process which will highlight conserved transcription factor binding sites.

## Alignments of Binding Sites

In addition to publication data, RTB provides its users with a growing collection of curated binding site alignments. Each alignment was created by an expert curator who provided descriptions explaining all alignments, specific sequence locations referenced to NCBI RefSeq genomes, available publications, and recommended options for using this alignment to search new genomes. This data is available for download

## Prediction

The process for comparing hits of a particular motif against multiple genome. a) A predefined alignment is chosen to create a position weight matrix from (custom alignment option is also available), b) Genomes to compare are selected, c) Results will be filtered by the options given, d) The result is a table with rows being orthologous genes, and hits specified within each row. For each orthologous row, additional analysis tools are available, such as sequence logs, sequence alignments in graphical and text formats, phylogenetic trees and the ability to view the alignment in the feature rich application JalView.

## ACKNOWLEDGEMENT

ESPP2 (MDCASE) is part of the Virtual Institute for Microbial Stress and Survival (VIMSS) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.