Published online 11 June 2010 Nucleic Acids Research, 2010, Vol. 38, Web Server issue W299–W307
doi:10.1093/nar/gkq531

RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach

Pavel S. Novichkov[1,*], Dmitry A. Rodionov[2,3,*], Elena D. Stavrovskaya[3,4], Elena S. Novichkova[1], Alexey E. Kazakov[1,3], Mikhail S. Gelfand[3,4], Adam P. Arkin[1,5], Andrey A. Mironov[3,4] and Inna Dubchak[1,6]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [2]Burnham Institute for Medical Research, La Jolla, CA 92037, USA, [3]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, [4]Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow 119992, Russia, [5]Department of Bioengineering, University of California, Berkeley, CA 94704-3224 and [6]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

Received March 9, 2010; Revised and Accepted May 26, 2010

ABSTRACT

RegPredict web server is designed to provide comparative genomics tools for reconstruction and analysis of microbial regulons using comparative genomics approach. The server allows the user to rapidly generate reference sets of regulons and regulatory motif profiles in a group of prokaryotic genomes. The new concept of a cluster of co-regulated orthologous operons allows the user to distribute the analysis of large regulons and to perform the comparative analysis of multiple clusters independently. Two major workflows currently implemented in RegPredict are: (i) regulon reconstruction for a known regulatory motif and (ii) *ab initio* inference of a novel regulon using several scenarios for the generation of starting gene sets. RegPredict provides a comprehensive collection of manually curated positional weight matrices of regulatory motifs. It is based on genomic sequences, ortholog and operon predictions from the MicrobesOnline. An interactive web interface of RegPredict integrates and presents diverse genomic and functional information about the candidate regulon members from several web resources. RegPredict is freely accessible at http://regpredict.lbl.gov.

INTRODUCTION

Accurate reconstruction of transcriptional regulatory networks and annotation of regulatory elements is one of critical aspects of microbial genomics in the era of large-scale genome sequencing. Despite the growing number of genome-wide gene expression studies, our abilities to convert the results of these studies into accurate regulatory annotations (such as 'gene A is transcriptionally regulated by regulator B') and to project such annotations from model organisms to related genomes are extremely limited. Likewise, despite the abundance of gene annotation resources and tools, very few of them support regulatory annotations beyond a handful of model species. In our vision, the genomics-driven reconstruction of regulatory interactions in a well-populated taxonomic group should include two main steps: (i) accurate reconstruction of a reference set of regulons in a representative set of genomes from the taxonomic group and (ii) automated propagation of the inferred regulons to all closely related genomes in this group.

Genes and operons directly co-regulated by the same transcription factor (TF) or RNA structure (e.g. a metabolite-sensing riboswitch) are considered to be a part of a regulon. Activation or repression of gene expression is achieved via sequence-specific binding of TFs to regulatory sites (or TF-binding sites, TFBSs) located in promoter regions of genes. The TFBSs often possess an intrinsic symmetry, e.g. palindromic, due to mostly

*To whom correspondence should be addressed. Tel: +1 510 495 2913; Fax: +1 510 486 5614; Email: psnovichkov@lbl.org
Correspondence may also be addressed to Dmitry A. Rodionov. Tel: +1 858 646 3100; Fax: +1 858 795 5249; Email: rodionov@burnham.org

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

dimeric conformation of bacterial TFs in the DNA-bound state. The size of a single TFBS usually varies between ~12 and 30 nt, with the most common length being 16–20 nt. Significant variations in binding sites of particular TFs co-regulating multiple genes pose a challenge to the identification of all TFBSs that conform to a consensus motif in the same genome. Nucleotide frequency positional weight matrices (PWMs) are used by many computational algorithms as a more sensitive and more precise way for TFBS recognition in microbial genomes (1).

Comparative genomic approaches are becoming widely used for genome-scale inference of *cis*-acting regulatory elements (e.g. TFBSs, promoters and RNA regulatory sites) and reconstruction of transcriptional regulatory networks in numerous groups of bacteria (2). The consistency check and phylogenetic footprinting techniques are based on the assumption that regulons have a tendency to be conserved between the genomes that contain orthologous TFs (3–7). The presence of similar candidate TFBSs upstream of orthologous genes is an indication that it is a true regulatory site, whereas candidate TFBSs scattered at random in the genome are considered false positives. Simultaneous analysis of multiple genomes from the same taxonomic group allows one to make reliable predictions of TFBSs even with weak recognition rules. In the general approach, after validation of TFBSs by comparative genomics, the improved PWM is constructed and used to scan the genomes for the presence of additional instances of the TFBS motif, yielding regulon expansion beyond the initial training sets. This leads to another cycle (or several cycles) of iterative refinement aimed at maximizing coverage and consistency of the reconstructed regulons.

In our previous work, we provided multiple examples of the power of the comparative genomic approach applied to a large number of different TF- and riboswitch-driven regulons reconstructed for a broad spectrum of diverse bacteria (2). In particular, this approach was applied to: (i) propagate previously known regulons from model organisms to many others; (ii) *ab initio* predict and reconstruct novel regulons either for a metabolic pathway of interest or for a particular family of TFs; and (iii) perform wide-ranging reconstruction of multiple regulons in a group of closely related species (8–13). These, and other previous studies, helped us to determine major workflows for genomic reconstruction of regulons and develop the RegPrecise database for collection and visualization of the accumulated regulatory reconstructions (14).

In this paper, we present the RegPredict web server designed to bring our extensive experience to the scientific community and facilitate the procedure of regulon inference in prokaryotic genomes from diverse taxonomic groups by providing a highly integrated set of computational modules for semi-automated and accurate analysis of regulons. RegPredict is designed to approach the goal of fast accumulation of reference sets of regulons in diverse groups of microorganisms.

CLUSTERS OF CO-REGULATED ORTHOLOGOUS OPERONS

In a typical scenario of comparative genomic analysis of regulation, candidate TFBSs are identified in the upstream regions of orthologous genes and subsequent consistency check procedure is carried out (3–7). Two potentially challenging issues that could hinder automatic implementation of this procedure are (i) high conservation rate in closely related genomes (source of false positives), and (ii) frequent operon rearrangements and appearance of additional TFBSs in distantly related genomes (false negatives). To address these issues and facilitate comparative assessment of putative regulon members in a group of closely related genomes, we developed an automatic procedure to split the entire regulon into a set of candidate subregulons that are defined as Clusters of co-Regulated Orthologous operoNs (CRONs). For each analyzed regulon, the set of constructed CRONs is prioritized based on the level of conservation of regulatory interactions, allowing focusing on the most prominent regulon members. At the next step, the functional and genomic context analysis of each CRON is conducted using the advanced web interface (see below), facilitating the decision on CRON inclusion in the final regulon model.

The procedure of CRON construction is outlined in Figure 1. First, the upstream regions of all putative operons are scanned for candidate TFBSs. Second, all orthologous operons with candidate TFBSs are linked together in clusters, allowing one to evaluate the level of conservation of the candidate regulatory interactions. Two operons are considered orthologous if they share at least one orthologous gene. Finally, the initial clusters are extended by adding orthologous operons lacking candidate TFBSs to build a candidate CRON. Each constructed CRON can then be analyzed separately as an independent unit. Combining all accepted CRONs for a given TFBS motif yields the reconstructed TF regulon for a group of target genomes. The CRON-based splitting of a regulon into subregulons is especially important for the analysis of large regulons for global TFs such as CRP in Proteobacteria and CcpA in Firmicutes.

The procedure of CRON construction in RegPredict is based on the genomic data available in the MicrobesOnline database including: (i) high-quality orthologs based on the analysis of phylogenetic trees of protein domains and (ii) predicted operons (15,16).

WORKFLOWS FOR REGULON INFERENCE

The RegPredict web server was designed to provide a set of integrated tools for the complete pipeline of regulon inference in a set of taxonomically related prokaryotic genomes. The server implements two well-established workflows that are widely used for the comparative genomic analysis of regulation: (i) regulon reconstruction for known regulatory motifs with available PWMs and (ii) *de novo* inference of regulons for previously unknown regulatory motifs (Figure 2).

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

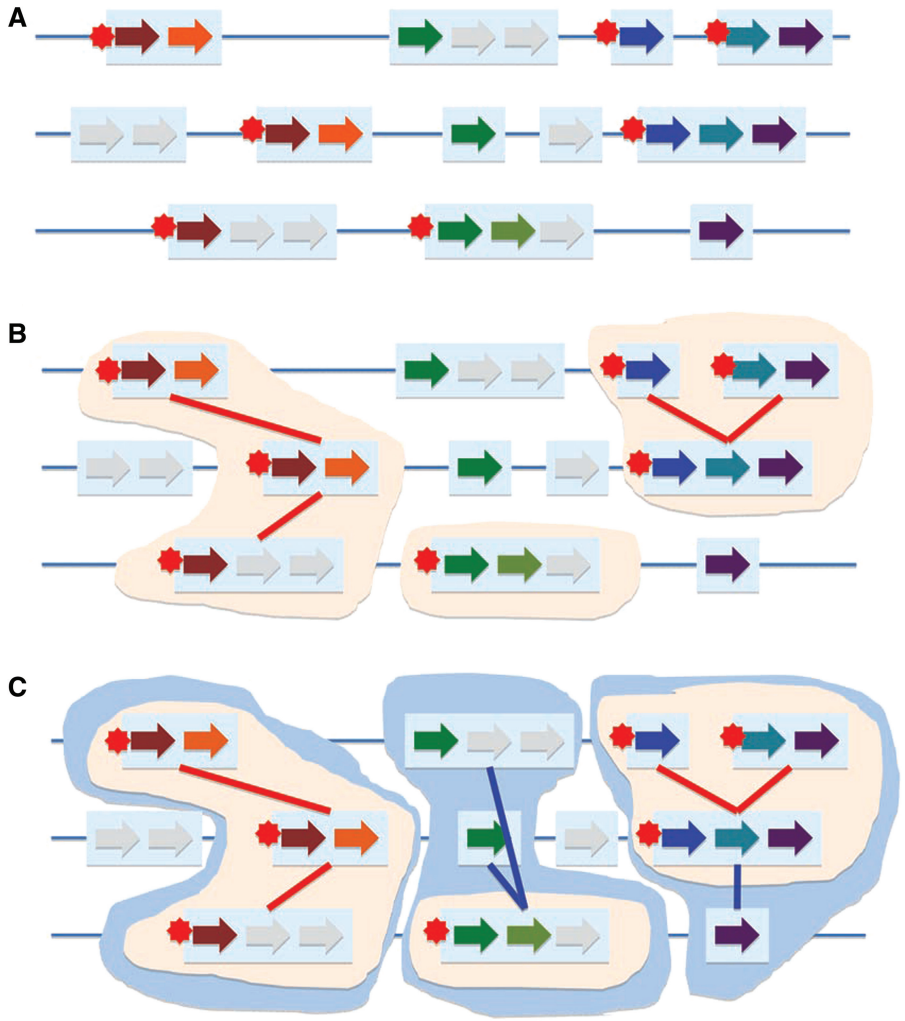

Figure 1. Major steps in building Clusters of co-Regulated Orthologous operoNs (CRONs): (**A**) search for putative TF binding sites, (**B**) building a core of a CRON, (**C**) extension of the core to form a final CRON. Three genomes are represented by straight lines. Putative operons are shown by rectangles. Genes are shown by colored arrows. The orthologous genes are depicted by the same color. Genes marked by light gray color do not have orthologs in other two genomes. Red circles depict the putative TFBSs. A core of a CRON is constructed as a set of operons containing orthologous genes preceded by putative TFBSs (highlighted by pink background). A final CRON is expanded by inclusion of orthologous genes without putative TFBSs (highlighted by light blue background).

'Regulon inference based on known PWM workflow' utilizes our collection of PWMs for known TFBS motifs. All genomes in the analyzed group of species are scanned by a particular PWM and candidate TFBSs are identified. Then, CRONs are automatically computed, ranked by the number of genomes with candidate TFBSs and by site scores, and finally provided as an output for manual curation in the interactive RegPredict web interface (see below). The PWM-based regulon inference is straightforward when the PWM training set belongs to genomes from the same taxonomic group as target genomes. In a more general case, the applicability of a particular PWM, initially built based on a set of known TFBS from one group of genomes, to another group of genomes can be determined by preliminary checking for the presence of orthologous TFs in the analyzed genomes.

The current version of RegPredict provides a comprehensive collection of PWMs from three publicly available web resources. The RegPrecise database (http://regprecise.lbl.gov), recently developed by our group (14), captures and visualizes predicted TF regulons that were reconstructed by the comparative genomic approach in a wide variety of prokaryotic genomes (~11 500 TFBSs for ~400 orthologous groups of TFs from over 350 prokaryotic genomes). The majority of these data are represented by genome-wide reconstructions in several diverse taxonomic groups of bacteria (including *Shewanellaceae*, *Streptococcaceae*, *Staphylococcaceae*, *Desulfovibrionales* and *Thermotogales*) and a large-scale prediction of regulons for the LacI family of TFs. The literature-based database RegTransBase (http://regtransbase.lbl.gov) (17) captures experimental knowledge on regulatory sequences and interactions published for a variety of microorganisms using a controlled vocabulary. The RegTransBase database provided data for ~140 PWMs based on published TFBSs. The RegulonDB database

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

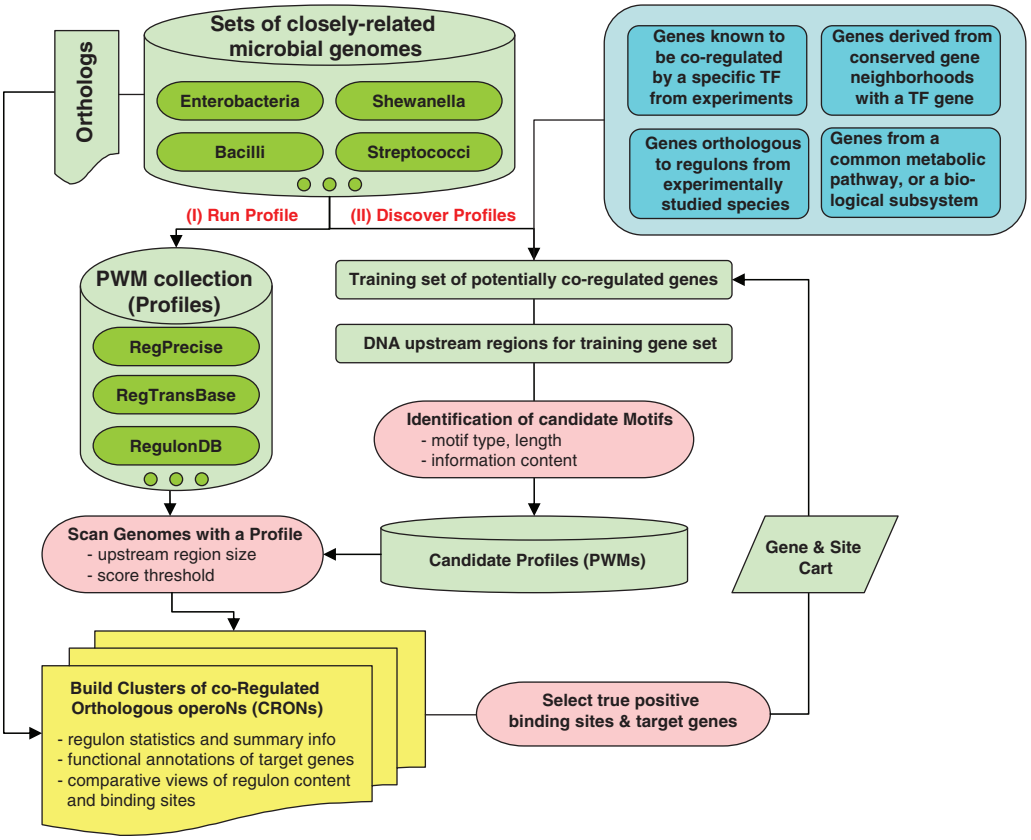

Figure 2. Diagram of regulon inference workflows implemented in the RegPredict web server. (**I**) regulon reconstruction for known regulatory motif profile, (**II**) *ab initio* regulon inference when a regulatory motif is not known.

(http://regulondb.ccg.unam.mx) (18) contains models of the TF regulons that were experimentally characterized in *Escherichia coli* (~70 TFs with PWMs).

The current list of microbial genomes available for regulon reconstruction in RegPredict includes representative set of species from major taxonomic groups of Bacteria available in the MicrobesOnline database. Currently, RegPredict allows to analyze up to 15 genomes simultaneously.

'*De novo* regulon inference workflow' is intended for prediction of regulons when there is no available PWM. The procedure starts with a set of potentially co-regulated genes from one or multiple related genomes from a particular taxonomic group. The input training set might be composed based on many sources including (i) genes that make up a functional pathway; (ii) genes homologous to regulons from a well-studied species (e.g. *E. coli* and *Bacillus subtilis*); (iii) genes derived from conserved chromosomal loci or operons containing orthologous TF genes; and (iv) genes with similar expression profile as determined by microarray experiments (Figure 2). Subsequent important steps are identification of candidate TFBS motifs within a set of DNA upstream regions for genes from the training set and construction of corresponding PWM profiles. The common approach is to find profiles of different types, such as palindromes of different length, or direct repeats, using the MEME-like iterative algorithm previously implemented by us in the SignalX software (4,19) and massively applied for microbial regulon inference (2,20).

The list of candidate motif profiles ranked by information content is provided as an output of the 'Discover Profiles' procedure in RegPredict. The selection of a correct motif is a critical point in the *de novo* regulon inference workflow, and it can be usually obtained by application of the selected candidate motifs to the target genomes and subsequent analysis of the resulting regulon content. RegPredict provides an interface to expedite the evaluation of candidate motifs, as described for the PWM-based regulon reconstruction workflow. Once the optimal motif is identified, its application to the target genomes may result in identification of new candidate regulon members that can be iteratively added to the initial training set. The iterative procedure of PWM construction is simplified by the 'Gene cart' for the collection of candidate target genes identified during the *de novo* regulon inference analysis.

WEB INTERFACE

The RegPredict web server is publicly accessible through a web interface at http://regpredict.lbl.gov. To start the session of regulon inference, the user selects a set of genomes from one or several taxonomic groups of interest. Organisms in the 'Select Genome' dialog are now grouped by taxonomy. The selected genome set remains constant

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

for the entire working session and is used in all regulon reconstruction procedures. Then the user is provided with a wizard to start from one of the two main strategies of regulon inference.

For regulon inference by known PWM

'Run Profile' procedure allows one to scan the genomes with a selected PWM profile using optional parameters and to automatically generate a set of candidate CRONs (use 'Profiles' tab, Figure 3A). The collection of PWMs available for the analysis is classified by the source database and taxonomy, where each motif profile is described by its name, length, training set size, information content per nucleotide position and consensus. Optional parameters for the genome scanning with a PWM are the upstream-region intervals to be searched and the threshold for the TFBS score, which by default is set to the minimal score among all sites in the training set for a given PWM. 'Exclude overlap' parameter allows for excluding any coding region in the selected upstream region interval, thus narrowing the upstream region search area. In addition to the collection of PWM profiles available in RegPredict, the user may submit any PWM profile uploaded from a file, as well as any set of pre-aligned TFBSs in the FASTA format as a training set for a new PWM (use 'Sequences' tab).

Positional nucleotide weights w in PWM profiles are calculated as $W(b,k) = log[N(b,k)+0.5] - 0.25\Sigma_{i=A,T,G,C}\log[N(i,k)+0.5]$, where $N(b,k)$ is the count of nucleotide b in position k (4). The candidate site score Z is defined as the sum of the respective positional nucleotide weights $Z(b_1 \ldots b_L) = S_{k=1.L} W(b,k)$, where k is the length of the site. The score threshold in 'Run Profile' is automatically updated each time when a user selects a particular PWM. The default score threshold is set to a minimal score among all binding sites in the training set which was used to build the PWM.

De novo regulon inference

'Discover Profiles' allows for inferring candidate motifs in a set of DNA fragments provided in the FASTA format and for generating the input set of upstream gene regions automatically from various gene sets (Figure 3B). Currently, the two sources of gene sets are available for the extraction of upstream regions: (i) genes collected in the 'Gene cart' (see below), and (ii) genes from a particular metabolic pathway or a subsystem [uploaded via the web service from the SEED database, http://theseed.uchicago.edu/FIG/ (21)]. To facilitate the procedure of profile discovery, the user may select one or several DNA motif types to be searched simultaneously. For each motif type, the user specifies a set of parameters to be used for the motif discovery procedure including the motif length, the minimal number of palindromic site positions, the size of the training set to be included in the profile and the minimum number of GC pairs among palindromic site positions. The 'Discover Profiles' procedure results in a set of candidate profiles that are prioritized based on their information content per position. All constructed candidate profiles are immediately available for the comparative genomic analysis by performing the same 'Run Profile' procedure, described in the previous section.

Prioritization of automatically generated CRONs

All CRONs constructed after the profile run are shown in the table 'Clusters of co-regulated orthologous operons' (Figure 4, left). For each CRON, the table provides a brief summary including: the number of genomes with candidate TFBSs (the 'Genomes' column), the number of operons, genes and sites in the CRON as well as the maximum TFBS score in the CRON (the 'Max score' column). The 'Genomes' and 'Max score' columns are used to automatically prioritize all CRONs based on: (i) the conservation of candidate TFBSs in the group of target genomes; and (ii) the similarity of candidate TFBSs to the PWM profile.

An alternative way of selecting a subset of CRONs for the detailed analysis is using filters. Two filters currently available in the 'Filter Operon Clusters' menu allow one to select CRONs containing either a particular subset of genes (using the 'By Genes' filter dialog, where multiple gene ids or locus tags can be submitted in a line-delimited format), or the candidate TFBSs in a specified subset of genomes (via the 'By Sites' filter dialog to specify the genomes of interest). In addition, 'Filter by Locus Tag' fast search option is available at the top of the CRONs panel.

Detailed analysis of a selected CRON

All information related to a particular CRON is shown in the 'Genomic context', 'Summary info', and 'Gene/Site properties' panels (Figure 4).

The interactive 'Genomic context' panel visualizes the genes (colored bars), sites (circles) and operons (gray background bars) for the selected CRON in the set of analyzed genomes. Orthologous genes are shown in the same color. The operons may be either from the same locus, or from multiple, distant from each other, genomic loci. All genes in this representation are shown in the same orientation from 5′ to 3′, while the actual strand of each gene in the genome is indicated in the 'Gene properties' panel. Operons located next to each other on the chromosome are separated by a short space, whereas remotely located operons are shown with long spaces in between. Three additional filters for the analysis of the CRON content are available in the main 'Filter' menu. 'Show Operons with Sites Only' restricts the representation of CRON content by co-regulated operons only. 'Show Orthologs of Selected Operon Only' restricts the gene coloring to the genes from a currently selected operon and their orthologous genes in other genomes, simplifying the analysis of the CRON composition. These two filters are especially valuable for the analysis of large and complicated CRONs. Finally, 'Show Minor Sites' visualizes additional weak TFBSs with scores within 10% of the threshold (minor sites). Although minor sites are not accounted for in the CRON construction procedure, their analysis may help to adjust the score threshold and to repeat the 'Run Profile' procedure.

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

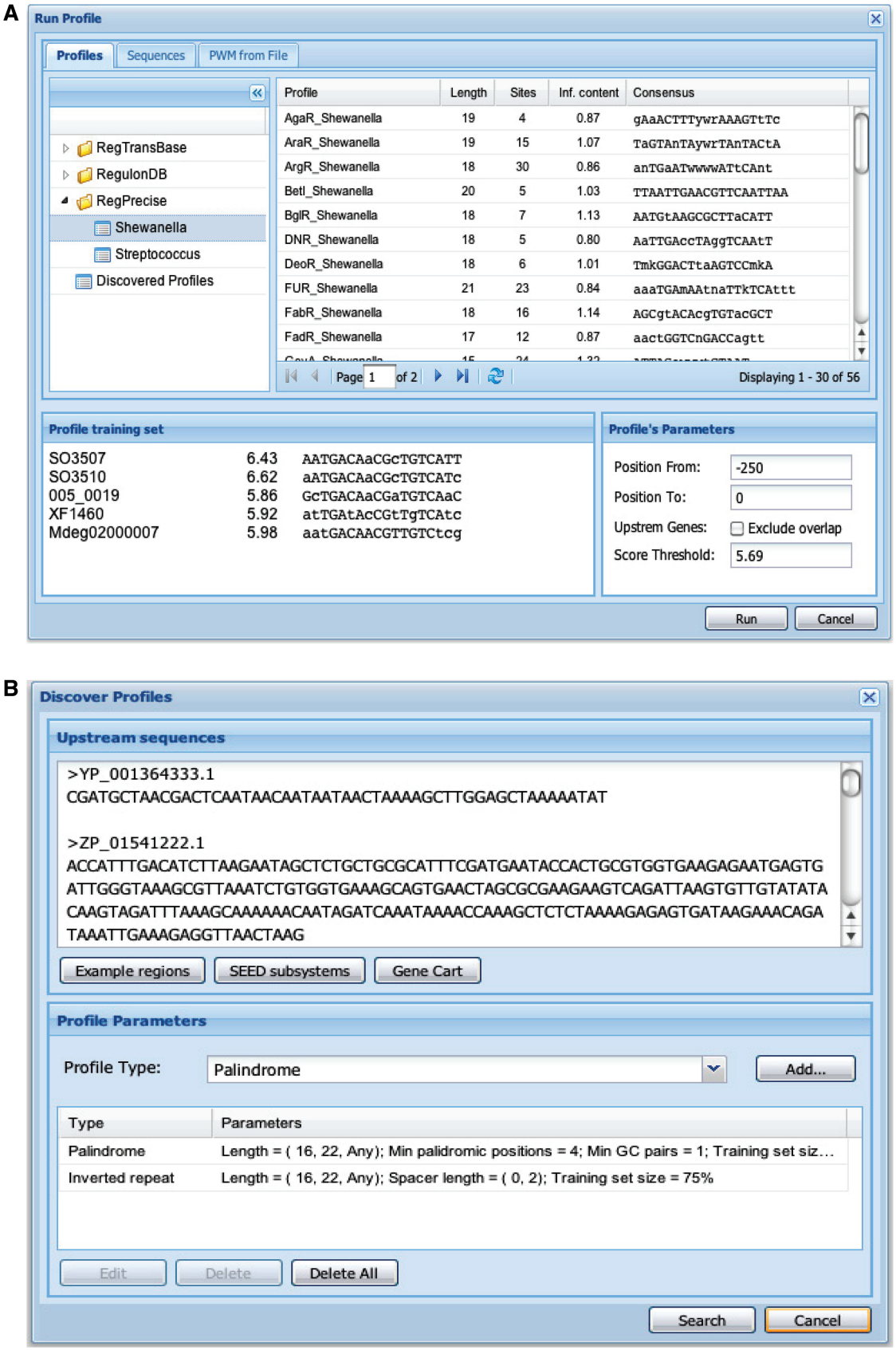

Figure 3. The parts of interface devoted to user input in two main strategies of regulon inference. (**A**) 'Run Profile' allows the user to start regulon reconstruction for a known PWM profile, (**B**) 'Discover Profiles' initiates the procedure for inferring candidate motifs in a training set of sequences.

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

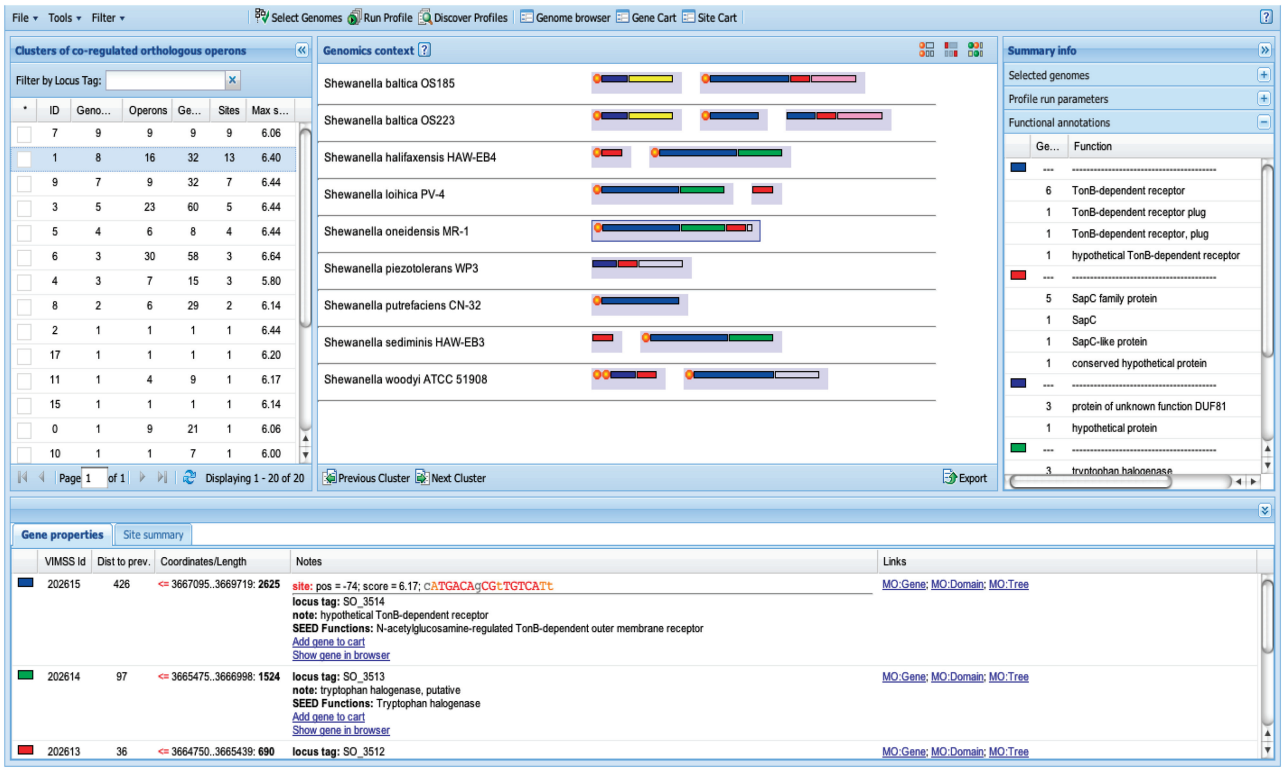

Figure 4. Interactive workspace of the RegPredict web server. Example shows the main output obtained by scanning nine *Shewanella* genomes with PWM for the NagR transcription factor. Prioritized clusters of co-regulated orthologous operons (CRONs) are listed in the left panel table. Genomic context and summary gene function information for the first selected CRON are shown in the central and right panels, respectively. Detailed genomic information for the selected NagR-regulated operon in *S. oneidensis* MR-1 is shown in the bottom panel.

The 'Gene properties' panel provides details on genomic and functional properties of each gene in a selected operon. Gene functional annotations and affiliations with SEED subsystems are automatically uploaded from the SEED database using web services (21). Finally, each gene in this panel has multiple links to web pages in the MicrobesOnline database (15), including the central gene page ('MO:Gene'), the domain page ('MO:Domain') and the gene phylogentic tree page ('MO:Tree').

The 'Summary info' panel shows additional information about the analyzed group of genomes, current parameters of the 'Run Profile' procedure and an overview of functional annotations of orthologs populating a particular CRON.

The results of one round of CRON analysis can be exported to a text file using the 'Export' button in the 'Genomic context' panel. The exportable CRON content may be narrowed by marking genes/operons intended for exclusion from the CRON (using 'Alt' button). The export file contains detailed description of genes (gene ID in MicrobesOnline, locus tag, gene name, ortholog ID in MicrobesOnline and function note) and putative TFBSs (site sequence, score and position relative to the following gene start) for each analyzed genome (with listed genome name and its NCBI taxonomy ID). Genes are grouped in putative operons.

Additional components

To provide more flexibility during the regulon reconstruction, the RegPredict provides three additional components, 'Genome browser', 'Gene cart' and 'Site cart'.

'Gene cart' is used to work with a collection of genes selected by the user. Genes can be added to the Gene cart either from the 'Gene properties' panel or from additional 'Gene search' or 'SEED subsystem' dialogs. In the 'Gene cart' dialog one can delete/add any gene, and export/import sets of genes to/from a text file. 'Gene search' implements a simple text search in the analyzed group of genomes. In the 'SEED subsystem', the user may open the hierarchical classification of current subsystems available from the SEED database via a web service, select the functional roles of interest from a particular subsystem and extract the list of genes that correspond to the selected roles in the selected subsystem in the analyzed set of genomes from the SEED database. Once a collection of genes is compiled, it can be used as a training set in the *de novo* regulon inference 'Discover Profile' workflow.

'Genome browser' allows for navigating among all genes in any of the analyzed genomes. By opening the genome browser by clicking 'Show gene in browser' link available in the 'Gene properties' panel, one can analyze the selected gene context and location of TFBSs in the genome or to search by gene id, name or locus tag.

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

Any gene in the genome browser can be added to the 'Gene cart'.

The 'Sites summary' panel collects all putative TFBSs that were identified in the selected CRON. 'Site cart' provides a tool for the comparative analysis of any subset of TFBSs selected during the regulon analysis session. Representation of TFBSs together with their flanking regions allows for estimating the overall conservation of gene upstream regions. 'Site cart' provides export of all TFBS sequences to a file, and can be used in 'Run Profile' procedure.

EXAMPLE USAGE ON RECONSTRUCTION OF HRCA REGULON IN STAPHYLOCOCCI

Typical scenario of the RegPredict usage is the identification of TFBSs and regulon reconstruction in a group of genomes for a TF with a known motif in a model organism. This scenario can be demonstrated by the inference of the heat shock regulon HrcA, previously described in *Streptococcus* spp., in a novel taxonomic group, for instance, the *Staphylococcaceae* (currently includes SEVEN nonredundant genomes).

First, the potential HrcA regulon in the *Staphylococcaceae* can be estimated by preliminary 'Run Profile' scan using the *Streptococcus*-specific PWM available in the RegPrecise-based collection of profiles and the default profile scan parameters (threshold = 6.3). Three CRONs highly conserved in all *Staphylococcaceae* will be obtained including two heat shock response operons, *groES-groEL* and *hrcA-grpE-dnaK-dnaJ-SA1407-SA1406-SA1405*, and the hypothetical gene *SA1838*, which shares the same TFBS with the divergently transcribed *groES* gene. Then additional putative HrcA regulon members can be checked by repeating 'Run Profile' with the reduced score threshold.

Second, the 'Discover profiles' procedure can be applied to build the *Staphylococcaceae*-specific HrcA motif. The training set of genes initially found with candidate sites during the previous step in all Staphylococcaceae genomes can be easily collected using 'Add gene to cart' function in 'Gene properties' panel (three genes in seven genomes). Upstream regions of 21 genes from 'Gene cart' are extracted in the 'Discover profiles' dialog and the resulting training set is used to build the profile with the parameters specified for the HrcA motif (palindrome; length = 27, 27, any; min palindromic positions = 4; min GC pairs = 1; training set size = 100%). The best resulting palindromic profile (information content = 1.04) can be immediately applied to automatically calculate set of CRONs for HrcA regulon in *Staphylococcaceae* (profile parameters: score threshold = 7). Sites with scores slightly below the threshold could be easily checked using 'Show minor sites' option. For instance, the third CRON including *SA1838* has a minor site with score = 6.98 upstream of the orthologous gene in *Macrococcus caseolyticus*.

It should be noted that the final regulon reconstruction in a particular case could be achieved by manual assessment of each CRON taking into consideration individual scores and overall conservation of TFBSs across the genomes, as well as the genomic context and functional annotations of regulated genes.

SUMMARY AND PERSPECTIVES

RegPredict is a highly interactive web server specifically designed for fast and accurate comparative analysis of microbial regulons and identification of TFBSs in multiple taxonomically related genomes in semi-automatic way. The server has user-friendly interface and multiple functionality including: (i) scanning of microbial genomes with annotated DNA motifs and collections of aligned binding sites, (ii) collecting upstream DNA regions for any gene set and identification of novel TFBS motifs and (iii) comparative analysis of the genomic context of candidate TFBSs in the group of genomes. Among few other web tools available for genomic analysis of TFBS, RSAT suite of tools (22) allows for sequence retrieval, pattern discovery and pattern matching. Key features of RegPredict include: (i) simultaneous analysis of TFBSs in multiple genomes and (ii) distribution of the analysis of large regulons using CRONs. RegPredict is an extremely powerful tool for fast accumulation of reference sets of regulons in diverse groups of microorganisms.

We are planning further development and extension of the RegPredict web server in three main directions. First, we will achieve tight integration of the RegPredict server with the RegPrecise database (14), including direct deposition of the inferred regulons and PWMs to the database using the individual sign-on capability. Such two-way integration will facilitate accumulation of the reference sets of regulons that will further be used for large-scale automated propagation of regulatory annotations to closely related genomes. Second, we will add new types of input gene sets for the *de novo* regulon inference, including: (i) genes with similar expression profiles using microarray experiments deposited in the MicrobesOnline database (15); (ii) genes derived from conserved gene clusters and operons containing a putative TF gene; (iii) genes that are homologous to previously inferred regulons in other lineages described in the RegPrecise database; and (iv) experimentally known sets of co-regulated genes from the RegTransBase (17), DBTBS (23), and CoryneRegNet (24) databases. Finally, we are planning to add a module for the analysis of RNA regulatory elements described in the RFAM database (25).

ACKNOWLEDGEMENTS

We are grateful to the MicrobesOnline and SEED teams for facilitating access to data, to Andrei Osterman for useful discussions and encouragement, to Ekaterina Ermakova, Dmitry Ravcheev, Anna Gerasimova, Olga Tsoy for thorough testing of the server and useful suggestions, Alex Poliakov for technical assistance, and to Tatiana Paley for assistance in web interface development, and to Tatiana Smirnova for web design of the home page.

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010

FUNDING

The US Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program: GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy; National Science Foundation (award DBI-0850546 to D.A.R.); Howard Hughes Medical Institute (55005610 to M.S.G.); Russian Foundation for Basic Research (08-04-01000 to A.E.K., 09-04-92745 to M.S.G., 10-04-01768 to D.A.R., 09-04-92742 to A.A.M.); Russian Academy of Sciences (program 'Molecular and Cellular Biology' to D.A.R and M.S.G.); Federal Agency on Education (P2581 to E.D.S.); Russian Science Agency (contract 2.740.11.0101 to M.S.G.); and Russian President's grant for young scientists (MK-422.2009.4 to D.A.R.). Funding for open access charge: The US Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program.

Conflict of interest statement. None declared.

REFERENCES

1. D'Haeseleer,P. (2006) How does DNA sequence motif discovery work? *Nat. Biotechnol.*, **24**, 959–961.
2. Rodionov,D.A. (2007) Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.*, **107**, 3467–3497.
3. Alkema,W.B., Lenhard,B. and Wasserman,W.W (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res.*, **14**, 1362–1373.
4. Mironov,A.A., Koonin,E.V., Roytberg,M.A. and Gelfand,M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **27**, 2981–2989.
5. Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G.D. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, **11**, 566–584.
6. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.
7. Manson McGuire,A. and Church,G.M. (2000) Predicting regulons and their cis-regulatory motifs by comparative genomics. *Nucleic Acids Res.*, **28**, 4523–4530.
8. Fredrickson,J.K., Romine,M.F., Beliaev,A.S., Auchtung,J.M., Driscoll,M.E., Gardner,T.S., Nealson,K.H., Osterman,A.L., Pinchuk,G., Reed,J.L. *et al.* (2008) Towards environmental systems biology of Shewanella. *Nat. Rev. Microbiol.*, **6**, 592–603.
9. Kazakov,A.E., Rodionov,D.A., Alm,E., Arkin,A.P., Dubchak,I. and Gelfand,M.S. (2009) Comparative genomics of regulation of fatty acid and branched-chain amino acid utilization in proteobacteria. *J. Bacteriol.*, **191**, 52–64.
10. Rodionov,D.A., Dubchak,I., Arkin,A., Alm,E. and Gelfand,M.S. (2004) Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. *Genome Biol.*, **5**, R90.
11. Rodionov,D.A., Dubchak,I.L., Arkin,A.P., Alm,E.J. and Gelfand,M.S. (2005) Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. *PLoS Comput. Biol.*, **1**, e55.
12. Rodionov,D.A. and Gelfand,M.S. (2005) Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling. *Trends Genet.*, **21**, 385–389.
13. Rodionov,D.A., Gelfand,M.S., Todd,J.D., Curson,A.R. and Johnston,A.W. (2006) Computational reconstruction of iron- and manganese-responsive transcriptional networks in alpha-proteobacteria. *PLoS Comput. Biol.*, **2**, e163.
14. Novichkov,P.S., Laikova,O.N., Novichkova,E.S., Gelfand,M.S., Arkin,A.P., Dubchak,I. and Rodionov,D.A. (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.*, **38**, D111–D118.
15. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
16. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
17. Kazakov,A.E., Cipriano,M.J., Novichkov,P.S., Minovitsky,S., Vinogradov,D.V., Arkin,A., Mironov,A.A., Gelfand,M.S. and Dubchak,I. (2007) RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.
18. Gama-Castro,S., Jimenez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muniz-Rascado,L., Martinez-Flores,I., Salgado,H. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.
19. Gelfand,M.S., Koonin,E.V. and Mironov,A.A (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res.*, **28**, 695–705.
20. Gelfand,M.S. (2006) Evolution of transcriptional regulatory networks in microbial genomes. *Curr. Opin. Struct. Biol.*, **16**, 420–429.
21. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
22. Thomas-Chollier,M., Sand,O., Turatsinze,J.V., Janky,R., Defrance,M., Vervisch,E., Brohee,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.
23. Sierro,N., Makita,Y., de Hoon,M. and Nakai,K. (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, D93–D96.
24. Baumbach,J. (2007) CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics*, **8**, 429.
25. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.

Downloaded from http://nar.oxfordjournals.org at Lawrence Berkeley Lab on June 29, 2010