886–893 *Nucleic Acids Research, 2017, Vol. 45, No. 2* *Published online 29 August 2016*
doi: 10.1093/nar/gkw749

Computational prediction of regulatory, premature transcription termination in bacteria

Adi Millman, Daniel Dar, Maya Shamir and Rotem Sorek*

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

Received April 19, 2016; Revised August 08, 2016; Accepted August 18, 2016

ABSTRACT

A common strategy for regulation of gene expression in bacteria is conditional transcription termination. This strategy is frequently employed by 5′UTR *cis*-acting RNA elements (riboregulators), including riboswitches and attenuators. Such riboregulators can assume two mutually exclusive RNA structures, one of which forms a transcriptional terminator and results in premature termination, and the other forms an antiterminator that allows read-through into the coding sequence to produce a full-length mRNA. We developed a machine-learning based approach, which, given a 5′UTR of a gene, predicts whether it can form the two alternative structures typical to riboregulators employing conditional termination. Using a large positive training set of riboregulators derived from 89 human microbiome bacteria, we show high specificity and sensitivity for our classifier. We further show that our approach allows the discovery of previously unidentified riboregulators, as exemplified by the detection of new *LeuA* leaders and T-boxes in *Streptococci*. Finally, we developed PASIFIC (www.weizmann.ac.il/molgen/Sorek/PASIFIC/), an online web-server that, given a user-provided 5′UTR sequence, predicts whether this sequence can adopt two alternative structures conforming with the conditional termination paradigm. This webserver is expected to assist in the identification of new riboswitches and attenuators in the bacterial pan-genome.

INTRODUCTION

Conditional transcription termination is a common mechanism for gene expression regulation in bacteria (1). Conditional transcriptional terminators usually occur in the 5′UTR of genes or operons, such that in some conditions an intrinsic premature transcriptional terminator is formed, preventing the transcription into the downstream gene (Figure 1). It is estimated that a significant fraction of all bacterial genes are regulated by conditional premature termination (2).

Several types of *cis*-acting RNA-based regulation systems (riboregulators) employ conditional premature termination as part of their mechanism of action: (i) riboswitches (3), which are non-coding RNA elements that directly bind small molecule ligands and alter their structure accordingly; (ii) attenuators, which encode short upstream open reading frames (uORFs) that sense the stalling of the ribosome in case of shortage in amino acids (4) or in presence of antibiotics (5,6); (iii) T-boxes (7), which sense amino acid availability by directly binding uncharged tRNAs and (iv) RNA leaders that bind specific antitermination proteins (8).

While the regulatory architypes outlined above differ significantly in their sensory strategies, they all control premature termination by switching between two alternative and mutually exclusive RNA conformations. In the repressive conformation (‘closed-state’), the riboregulator assumes a terminator form, generating a hairpin structure immediately followed by a uridine rich tract (Figure 1A). Alternatively, in the active conformation (‘open-state’), the RNA folds into an antiterminator stem-loop structure that effectively decouples the uridine tract from the terminator hairpin, therefore promoting transcription read-through into the gene (Figure 1B). The choice between the two possible RNA folds is determined by the presence or absence of the regulating metabolite (for riboswitches), the presence of ribosomes stalled on the riboregulator (for attenuators) or binding of a specific antitermination protein to the riboregulator (for protein-binding RNA leaders).

Recent studies show that riboregulators that function via conditional, regulated termination are more abundant than originally thought (5), conforming with previous estimates that such RNA elements are very common in bacteria (9). Several computational tools have been developed to predict the presence of such riboregulators, most of them using comparative genomics, relying on consensus secondary structures and utilizing covariance models to search for new elements (10–12). Such approaches perform well when the riboregulator is highly conserved between distant organisms, but are expected to miss RNA elements that are rare or evolutionarily diverged (9,13). Several tools use thermodynamics-based methods to search for RNA ele-

*To whom correspondence should be addressed. Tel: +972 8 9346342; Fax: +972 8 934 4108; Email: rotem.sorek@weizmann.ac.il

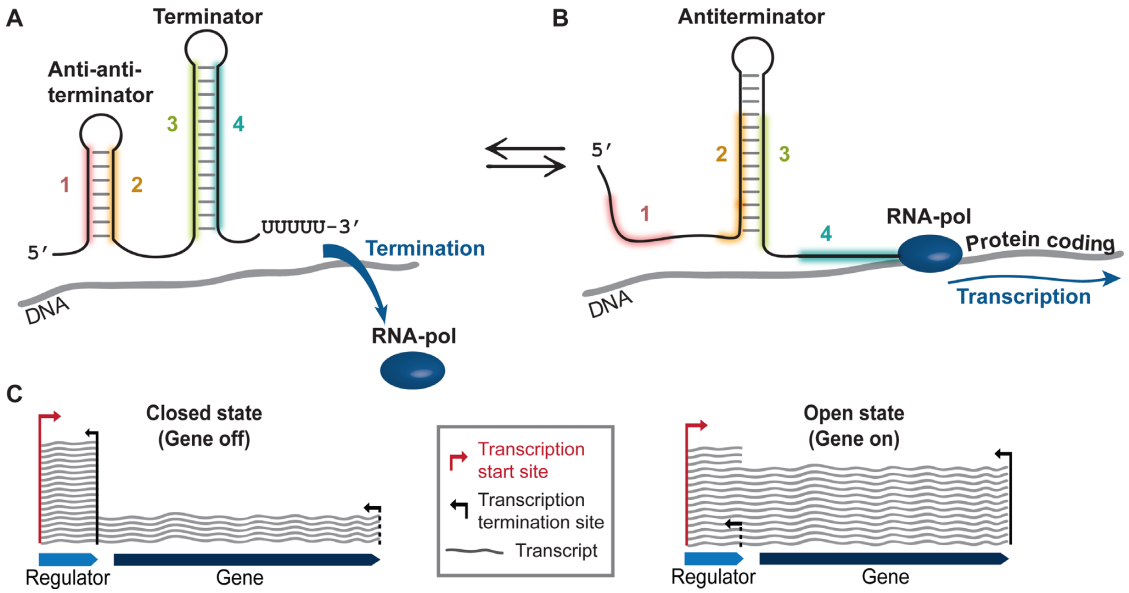

Figure 1. The principle of regulation by conditional termination. A riboregulator that functions through conditional transcription termination can assume two mutually exclusive structural conformations: (**A**) A 'closed' conformation, that entails an intrinsic terminator, which is a stem-loop (strands #3 and #4) followed by a poly-U. This structure causes the RNA-polymerase to terminate transcription prematurely. The terminator structure is usually preceded by another stem-loop structure called the anti-antiterminator or P1 (formed by strands #1 and #2). (**B**) An 'open' conformation, in which an antiterminator stem (not immediately followed by a poly-U) is generated from pairing of strands #2 and #3, allowing the RNA-polymerase to continue transcription into the downstream gene. (**C**) The closed state ('Gene off') typically results in higher amounts of the short, prematurely terminated transcript, which can be measured by RNA-seq (5). In the open state ('gene on') more full-length transcripts are observed.

ments that can adopt alternative conformations (14–17), but these methods are not specifically directed towards finding features of conditional terminators, and may be less effective in detecting riboswitches in which one of the conformations is only stable when bound to the ligand (18). To our knowledge there is currently no tool that utilizes the basic concept of mutually exclusive terminator-antiterminator conformations in order to predict riboregulators that function via regulated termination.

We developed PASIFIC (prediction of alternative structures for identification of *cis*-regulation), an online tool that, given a user-provided bacterial 5′ UTR, searches this sequence for terminator-antiterminator alternative structures enabling RNA structure prediction of known and novel *cis*-acting riboregulators. Combining machine learning with prediction of alternative RNA secondary structures, this tool can detect riboswitches, attenuators and leaders in a manner not dependent on their sequence conservation in other species.

MATERIALS AND METHODS

Collection of the positive set

Known riboregulators were identified in the reference genomes of 169 human microbiome bacterial species studied in (5). For each of these genomes, the Infernal (19) cmscan tool was run with 'trusted' cutoff (–cut_tc) using all Rfam (20) models defined as Cis-reg, Riboswitch, or Leader. Results were screened for riboregulators of classes that are known by the literature to function via conditional termination. Results were further screened for hits that have a premature TTS (transcription termination site) within a 5′UTR. For Rfam models that contained the terminator stem-loop within the model, a TTS was assigned if it was adjacent to the hit (up to 50 bases downstream). For Rfam models that do not contain the terminator stem-loop within the model, a TTS was searched starting 30 bases downstream to the Rfam model and up to 80 bases downstream to the model.

Collection of the negative set

Intergenic regions sized up to 600 bases were extracted from the same genomes as above. Segments of 80–400 bases that have TSSs (transcription start sites) and TTSs (with over five reads), and were expressed in the opposite orientation to the downstream gene were extracted. These segments were then scanned using the Infernal (19) cmscan tool with 'trusted' cutoff for Rfam models and we verified that this set does not contain any elements with positive Rfam hits. The set was supplemented with 27 gene terminators that were obtained by looking for TTS of genes that are regulated by the positive set and taking the last 200 bases, and 30 tRNAs from the discarded sequences.

Alternative folds prediction

Positive and negative sets were filtered for elements that contain features of an intrinsic terminator at their 3′ end. For this, sequences not having a poly-U (defined as a stretch of at least three consecutive uridines and no more than 2 consecutive non-uridine bases) were discarded, as well as sequences not presenting a stem-loop structure upstream to the poly-U. For this, the RNALfold program of the ViennaRNA Package (21) was used to search for stem-loops of

15–45nt adjacent to the poly-U, with the most stable of these selected as a terminator.

To search for the antiterminator structure, the poly-U and the second strand of the terminator (strand #4 in Figure 1) were removed and the sequence was scanned again using RNALfold looking for stem-loops that (i) begin before the terminator stem-loop but overlap it; (ii) the distance between strand #2 and #3 was 50nt or shorter and (iii) the structure was a simple stem-loop (allowing bulges but not allowing multiple internal stem loops). Among these stem-loops the most stable one was selected as an antiterminator. To search for the anti-antiterminator structure, the entire terminator (strand #3 and #4 in Figure 1) was removed from the sequence and the remaining sequence was scanned again by RNALfold for a stem loops that begins before the antiterminator and overlaps it. The longest such stem-loop was then chosen as the anti-antiterminator. Only sequences fulfilling all these structural requirements were included in the training and test sets.

Machine learning

The positive and the negative sets were split randomly to 80% training set and 20% test set. The Random Forest 'cross-validation for feature selection' function (rfcv) of the randomForest R package (22) was used to assess how many features should be used for the classification, using a threshold of 0.2 estimated out of the bag (OOB) error rate. A classifier was built using the randomForest function in the randomForest R package (22) with all the features, and the 15 features with the highest mean decrease in Gini index were chosen for the final classifier.

For the 10-fold cross validation the training set was then divided into 10 groups, and the classifier was built 10 times, each time leaving one group out. Each classifier was then tested on the left-out group and the AUC was calculated. A final classifier was then built using all the training data and the 15 best features selected before. The classifier was then tested on the test set to assess the final AUC, sensitivity and specificity of the tool.

The classifier code is available through github: https://github.com/adimil/PASIFIC. DOI: 10.5281/zenodo.56651.

Terminators prediction within PASIFIC

Sequences were scanned using the RNALfold program of the ViennaRNA Package (21) for stem-loops of 15–45nt which have adjacent poly-U of 11 bases, defined according to (23) as: spacer (0–2 bases), proximal part (five bases of which at least three uridines), distal part (four bases that are not four cytosines or four purines) and overall at least four uridines.

Prediction of new riboregulators

Intergenic regions of 100–600 bases were extracted from the 169 analyzed genomes described above. Segments of 80–400 bases that have TSSs and TTSs (with over 20 reads each), and of the same orientation as the downstream gene were extracted. These segments were then scanned using the Infernal (19) cmscan tool with 'trusted' cutoff for Rfam models typed as Gene, Ribozyme, rRNA, tRNA, sRNA, Cis-reg, Riboswitch and Leader, and positive hits were filtered out. The remaining sequences were scanned using the PASIFIC tool, and sequences showing scores above the threshold of 0.5 were further reported (Supplementary Table S3).

Depiction of predicted RNA structures in the PASIFIC web server

The alternative structures were predicted for each sequence using RNAfold (21) with structure constraints (-C), once with the predicted terminator stem-loop and anti-antiterminator stem-loop as constraints and once with the antiterminator as a constraint. Same coloring of the four stems were then added to the two RNAfold output postscript files.

Estimation of accuracy of predicted PASIFIC structures

For each sequence in the positive set the Infernal (19) cmscan tool was run with 'trusted' cutoff (–cut_tc) using all Rfam (20) models defined as Cis-reg, Riboswitch, or Leader. Indels and truncated areas were removed from the Rfam predicted structure and the coordinates of the second strand of the anti-antitermiterminator were extracted. These coordinates were compared to the coordinates of the first strand of the PASIFIC predicted antiterminator, and the overlap was calculated as the percentage of bases in the second strand of the Rfam anti-antiterminator that are covered by the first strand of the PASIFIC antiterminator.

Previously validated positive control

Sequences of regulators from Dar *et al.* (5) Supplementary Table S2 were extracted and were scanned using the PASIFIC tool with the three preset folding options. Sequences passing the threshold with at least one of the preset folding options were counted as positive hits.

RESULTS

The hallmarks of a riboregulator that utilizes conditional termination can be schematically depicted as a two-stem structure, where the second stem is followed by a poly-U sequence (forming a terminator), and the second strand of the first stem can base pair with the first strand of the second stem (Figure 1). We sought to use a machine learning approach that would learn the properties of known such structures (e.g. the length of the terminator and antiterminator, the free energy of the local structures, the ratios between the stabilities of the local structures, probabilities of the different folds, and more, see Table 1) and will enable prediction of new such elements.

To extract positive and negative training sets for the machine learning we used a recently published, large dataset of transcriptomes sequenced for >100 bacteria belonging to the human oral microbiome, where the transcription start sites (TSSs) and transcription termination sites (TTSs) were determined to the single-base resolution (5). We ran the Rfam Infernal tool (19) on the genomes comprising this set,

Table 1. Selected features for the Random Forest classification

Feature name	Explanation
ΔG_{open}	The free energy of the 'open state' RNA fold
ΔG_{closed}	The free energy of the 'closed state' RNA fold
$\Delta G_{open}/\Delta G_{closed}$	The ratio between the free energy of the two alternative folds
ΔG_{open}/length	The free energy of the 'open state' RNA fold normalized to the length of the sequence
$\Delta G_{antiterminator}$	The strength of antiterminator stem-loop
$\Delta G_{antiterminator}$/length	The strength of the antiterminator stem, normalized to its length
$\Delta G_{antiterminator}/\Delta G_{terminator}$	The ratio between the strength of the antiterminator and terminator stem-loops
$End_{antiterminator}$	The distance, in nt, of the antiterminator stem end from the TSS
$\Delta G_{closed}/\Delta G_{MFE}$	The ratio between the free energy of the 'closed state' RNA fold and the most stable folding of the RNA molecule retrieved from RNAFold (21)
$\Delta G_{open}/\Delta G_{MFE}$	The ratio between the free energy of the 'open state' RNA fold and the most stable folding of the RNA molecule retrieved from RNAFold (21)
ΔG_{open}-ΔG_{MFE}	The difference in kcal/mol between the free energy of the 'open state' RNA fold and the most stable folding of the RNA molecule
ΔG_{closed}-ΔG_{MFE}	The difference in kcal/mol between the free energy of the 'closed state' RNA fold and the most stable folding of the RNA molecule
ΔG_{P1}	The strength of the anti-antiterminator (P1) stem-loop
ΔG_{P1}/length	The strength of the P1 stem, normalized to its length
length $_{P1}$	The length of P1

and extracted known riboregulators including riboswitches, attenuators, and protein-based regulators (Methods). We then took regulators for which a clear terminator structure was present, and for which a potential antiterminator stem loop structure was identifiable using the RNALfold (21) RNA secondary structure predictor (Methods). Altogether, the positive set contained 312 known riboregulators, belonging to 18 regulator families, in 89 bacteria (Supplementary Table S1).

As a negative set, we collected non coding RNAs sized 80–400 bases, for which a TSS was identified as well as a TTS that conformed with a structure of an intrinsic terminator (Methods). For this set we only selected non coding RNAs that did not reside within the 5′UTR of a protein coding gene and were in a 'tail-to-tail' relationship with their nearby genes, to avoid cases of possible conditional terminators. These small RNAs were further filtered to remove known riboregulators. The negative set included 273 non-coding RNAs of unknown function. This set was further supplemented by 30 additional tRNAs, and 27 200bp 3′ ends of protein coding genes that end with a clear intrinsic terminator (Supplementary Table S2).

For each RNA in each of the sets we extracted a large set of features describing the composition of the RNA. Using cross validation for feature selection (Methods) we selected the 15 features with the highest mean decrease in Gini index (24), namely those features that could best differentiate between the positive and the negative sets (Supplementary Figure S1A; Table 1). Among the strongest classifying features was the ratio between the free energy of the two alternative folds, which we found, in the positive set, to average 0.79 (±0.2) suggesting that in a *bona fide* regulatory RNA that employs conditional termination, the 'closed state' conformation is typically stronger, but not much stronger, than the 'open state' one.

We used a Random Forest algorithm (22,25) to generate a classifier that differentiates between the positive and negative sets. Each dataset was randomly divided into 80% training set and 20% test set, and the classifier was trained using the selected 15 features (Table 1). We performed a 10-fold cross-validation on the training set and found an average AUC of 0.9 with standard deviation of 0.03, suggesting little to no over fitting (Supplementary Figure S1B). Applying the resulting classifier on the test set yielded classification with high sensitivity (82.5%) and low false positive rate (specificity of 80.6%, Figure 2). To check whether the method can perform well in the absence of term-seq data that accurately identifies the 3′ end of the riboregulator, we added 50 genomic bases to the 3′ ends of the test set sequences and predicted the position of the premature terminators using previously described guidelines (23). This resulted in slightly lower sensitivity of 73% and specificity of 88.1%, suggesting that the classifier can perform well even when the accurate 3′ end of the regulatory 5′UTR is not known.

To examine whether the classifier can detect new riboregulators, we searched, among the available transcriptomes of the oral microbiome, for genes showing long 5′UTRs (>100 bases) in which a TTS was observed, and for which Rfam did not identify any known regulators (Methods). Among the 5′UTRs conforming with these conditions, 47 5′UTRs passed the classifier threshold predicting a dual terminator/antiterminator conformation (Supplementary Table S3).

Four of these predicted riboregulators were found upstream of the 2-isopropylmalate synthase gene in closely related *Streptococcus* species, a genus common in the human oral cavity (Figure 3A) (26). The 2-isopropylmalate synthase (*leuA*) gene encodes for an enzyme that participates in biosynthesis of L-leucine. In *Escherichia coli*, this gene is known to be regulated by ribosome-mediated attenuation, via a riboregulator (*leuL*) encoding a uORF leader peptide that contains four consecutive leucine codons (27). Under leucine-limiting conditions the ribosome stalls over the leucine-rich uORF. This, in turn, enables the formation of the antiterminator stem-loop and leads to the transcription of the full length leucine biosynthesis gene. While the *leuL* leader was described in *E. coli* (27) as well as in other proteobacteria (28), to date it was not reported in Gram-positive Firmicutes, except for rare predictions in *Lactococcus lactis* (29). Our finding indicates that the leucine biosynthesis operon in *Streptococcus* species is regulated using the exact same mechanism as in *E. coli* (Figure 3A-C), suggesting that this mode of regulation has been established early in the evolution of bacteria. While the *Streptococcus* leader shows no significant sequence similarity to the *E.*

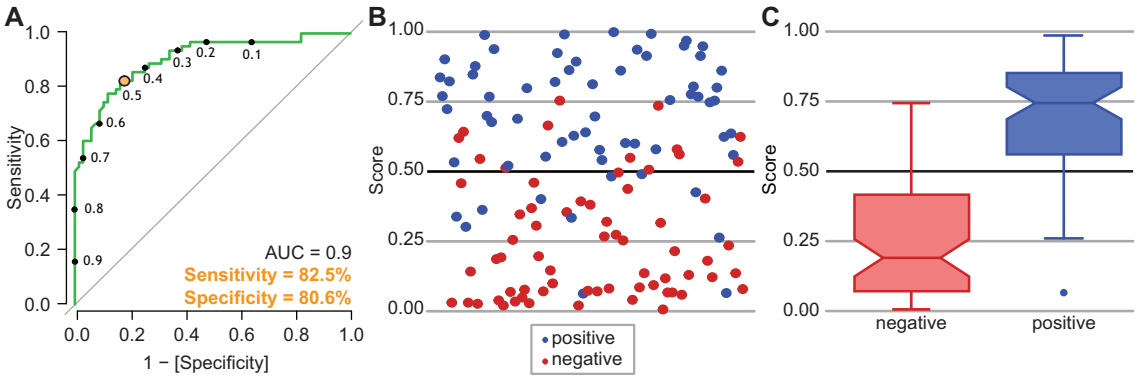

Figure 2. Classification of riboregulators (positive set) and other small ncRNAs (negative set) using a Random Forest classifier. (**A**) Receiver Operating Characteristic (ROC) curve depicting the performance of the Random Forest classifier in differentiating riboregulators from other small RNAs. The area under the curve (AUC) is 0.9. Sensitivity and specificity are specified for the score threshold (0.5) chosen as the classifier threshold. (**B**) Prediction results for the test set. Individual elements belonging to the positive and negative sets are depicted by blue and red points, respectively. Y-axis depicts the classifier score. Thick horizontal line depicts the classifier threshold (**C**) Box plot describing the classification score distribution of the positive and negative test sets.

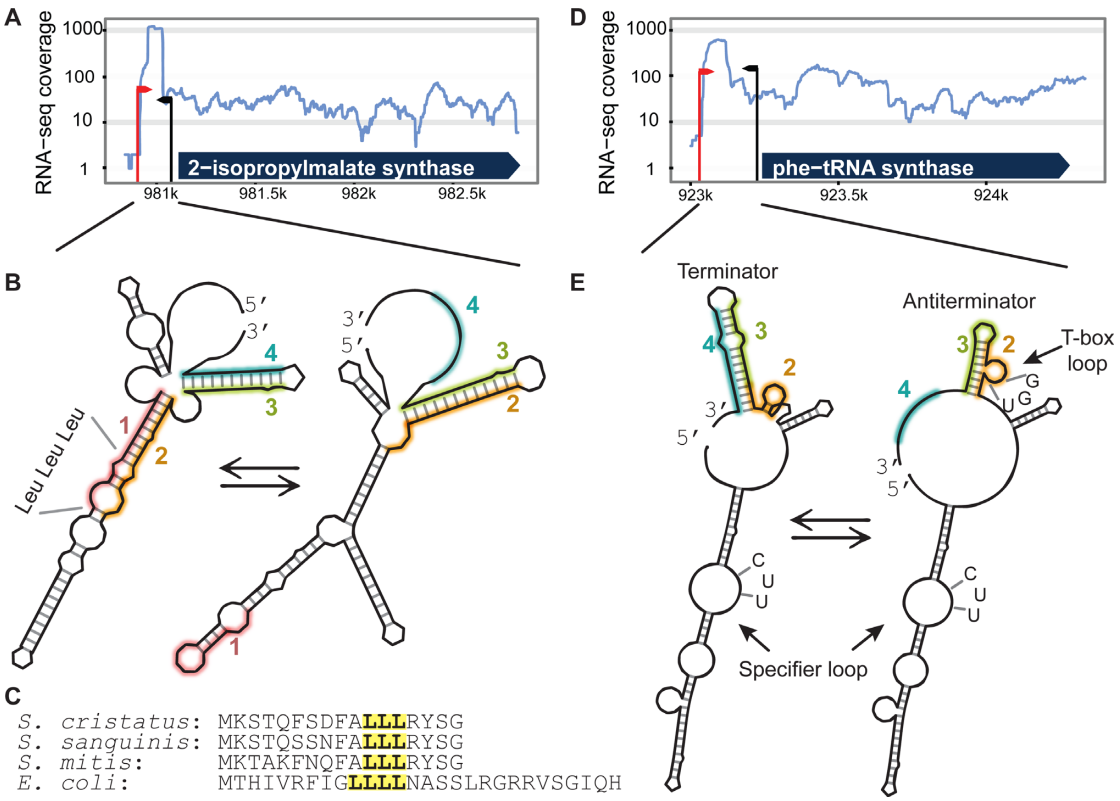

Figure 3. Identification of riboregulators in bacteria belonging to the human oral microbiome. (**A**) The leucine leader in *Streptococcus sanguinis* SK36. Data are shown for the *Streptococcus sanguinis* 2-isopropylmalate synthase (*leuA*) gene. Shown are RNA-seq data (blue curve), TSS inferred from 5′ end sequencing data (red arrows) and TTS inferred from term-seq data (black arrows) (5). X-axis, position on the *Streptococcus sanguinis* chromosome (NC_009009). (**B**) Predicted alternative conformations of the *LeuA* 5′UTR: Left – the 'closed' state where strands #3 and #4 form a terminator. Right – the 'open' state, with the predicted antiterminator. (**C**) Sequence of leucine-rich uORFs upstream of *LeuA* in various *Streptococcus* species and *E. coli*. (**D**) RNA seq data for the *S. sanguinis* phenylalanyl-tRNA synthase gene. (**E**) Predicted alternative conformations of the phenylalanyl-tRNA synthase 5′UTR, characteristic of a T-box leader structure. Left and right, the 'closed' and 'open' states, respectively.

coli one, the regulatory principle remains conserved: both leaders maintain the terminator/antiterminator structure, with a leader peptide containing several consecutive leucine codons overlapping the anti-antiterminator. These results demonstrate the evolutionary flexibility of attenuators, as recently observed also for attenuators regulating antibiotic-resistance genes (5).

In addition to the *leuL* leader, our predictions also identified terminator/antiterminator structures for the phenylalanine tRNA synthetase gene in *Streptococci* (Figure 3D and E). In Gram positive bacteria, this gene is known to be regulated by an upstream T-box leader, which switches between a closed or an open state when bound to charged or uncharged phe-tRNAs, respectively. However, Rfam search using the Infernal tool (19) failed to identify the 5′UTRs of these genes in *Streptococcus* species as containing T-box leaders, even when using the lowest threshold. Nevertheless, the terminator/antiterminator structure predicted by our algorithm showed a typical T-box conformation, with the antiterminator encoding the T-box sequence and the specifier loop properly encoding the phenylalanine codon. These results suggest that the *Streptococcus* phenylalanine tRNA synthetase is regulated by a diverged T-box structure that could not have been detected by commonly used tools.

To facilitate visual representation of terminator/antiterminator structures in RNA, we developed the PASIFIC web server (prediction of alternative structures for identification of *cis*-regulation), available at www.weizmann.ac.il/molgen/Sorek/PASIFIC/. For a user-provided sequence, PASIFIC presents the probability of the sequence being a functional riboregulator based on the classification results, and visually shows RNA secondary structure predictions of the two alternative conformations (Figure 4A). The user can enter sequences up to 400 bases long in FASTA format, browse the alternative conformations, download the figure of the structures as well as download the detailed results in CSV format. Several parameters can be set within the PASIFIC query form to best suit the user-provided sequence. For example, the search for P1, the helix that holds the aptamer, is only relevant for riboswitches and can be disabled for suspected T-box leaders, attenuators or protein binding riboregulators. In addition, while the length of the antiterminator is usually limited in riboswitches, it can be longer in attenuators, and accordingly the user can choose to restrict the antiterminator length parameter within the PASIFIC prediction. Finally, in case the exact 3′ end of the input sequence is unknown, PASIFIC can predict the position of the terminator if the relevant parameter is selected.

To estimate the accuracy of the structures predicted by PASIFIC, we compared the structures derived from PASIFIC predictions for the 312 riboregulators in our positive training set to the experimentally determined model structures in Rfam. Specifically, we asked to what extent our antiterminator prediction (RNA strand #2 in Figure 1B) overlaps with the anti-antiterminator in the Rfam model (RNA strand #2 in Figure 1A). In 49% of the cases there was a complete agreement between the PASIFIC prediction and the Rfam model (Methods). Moreover, in additional 16% of the cases, complete agreement was observed when the default PASIFIC parameters were altered (see above). These results suggest that, to a large extent, the PASIFIC–reported structures overlap with experimentally determined ones. For example, Figure 4B shows a comparison between the experimentally determined structure of the Purine riboswitch (30), and the structure derived from the PASIFIC prediction. While complex aptamer conformations such as pseudoknots cannot be predicted by PASIFIC, the general terminator/antiterminator relationship predicted by the webserver conforms well with the known structure.

In a recent study, we used term-seq to detect known and novel riboregulators in three model organisms: *Bacillus subtilis*, *Listeria monocytogenes* and *Enterococcus faecalis* (5). We examined the 164 elements experimentally detected in these three organisms as having premature termination using the PASIFIC webserver. We found that 104 of the 164 term-seq validated riboregulators (63%) pass the PASIFIC threshold for having terminator/antiterminator structures, including the novel antibiotic-sensitive riboregulator found upstream to the *lmo0919* gene in *L. monocytogenes*, which was experimentally validated as working via a termination/antitermination mechanism (5). Out of the 60 sequences that did not pass the PASIFIC threshold, 20 did not have a detectable intrinsic terminator, and the rest were both known and novel regulators. These results further validate the utilization of the PASIFIC webserver for detection of such riboregulators.

DISCUSSION

We present a machine-learning based approach (and an accompanying online tool) for discovery of riboregulators that encode conditional termination. Our approach is based on a search for mutually exclusive, alternative RNA conformations followed by a machine learning classification. The method does not rely on sequence or structural conservation and therefore can potentially enable detection of rare classes of riboregulators that are challenging to detect using comparative genomics (9). Since the basic mechanism of alternative RNA folding is common among the different types of riboregulators (riboswitches, attenuators, protein-based leaders), this method is not limited to a specific system. Moreover this method can predict the functional segments of the regulator—terminator and antiterminator and thus provides an easy framework for experimental validation of the predicted elements. Indeed, using this method we predicted a set of putative novel regulatory elements in various bacterial genomes (Supplementary Table S3), which would require further experimental verification in the future.

A number of studies have attempted to use structural predictions for the discovery of riboswitches. Some of these studies are conservation based, requiring comparative genomics for structural prediction (10,11,31,32). Our approach requires only a single sequence for prediction of alternative structure conformation, and is hence useful for highly diverged riboregulators, as well as for identifying alternative structures within synthetically designed riboregulators prior to experimental testing. Other studies utilized single sequences, searching for clusters of predicted suboptimal structures (15,16) or local energy minima within the RNA conformations landscape (33,34) to predict alterna-

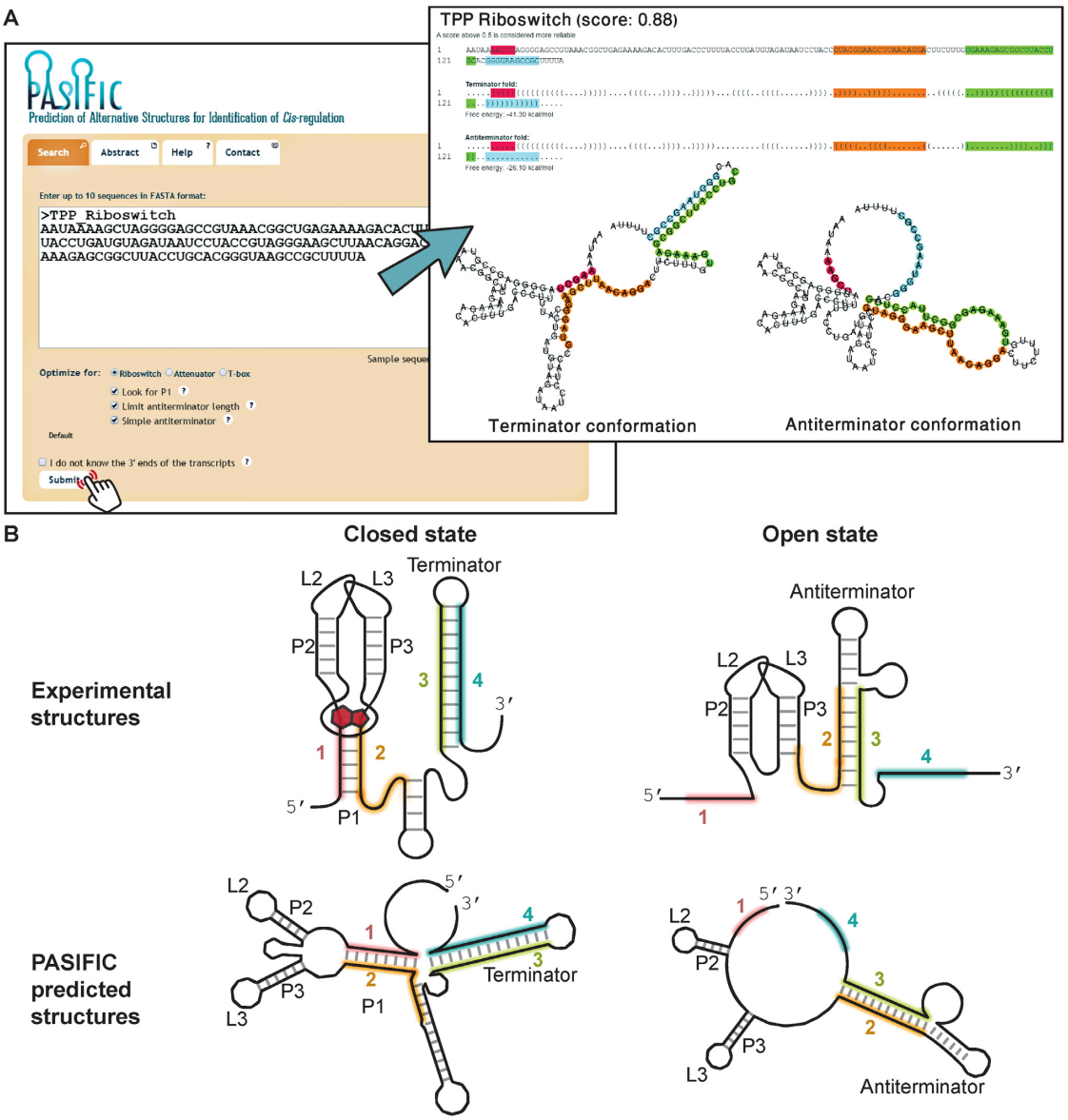

Figure 4. The PASIFIC web server. (**A**) Screenshot of the query page (left) and the results page (right) with predicted alternative conformations of a TPP riboswitch given as an example. (**B**) Comparison of the experimental alternative structures of the purine riboswitch (adapted from (38)) (top) and its PASIFIC-predicted structures (bottom). Left: the 'closed' state where the anti-antiterminator (P1) is stabilized by the bound purine and strands #3 and #4 form a terminator. Right: the 'open' state, where in the absence of purine the P1 stem is not stabilized and strand #2 is free to form an antiterminator with strand #3.

tive structures. While these approaches generally predict alternative possible conformations, our approach is focused on detecting conditional termination, and is less computationally expensive.

For a successful prediction of the two alternative conformations our method requires *a-priori* knowledge on the premises of the studied 5′ UTR regulator, and specifically the transcription start point and premature termination point. While methods that determine the exact TSS of bacterial transcripts are commonly used (35), methods to determine the 3′ ends of bacterial RNAs were only recently introduced. In this study we found the positions of the premature termination points using data derived from term-seq, an RNA sequencing method that determines RNA 3′ end positions in bacteria at a single-base resolution (5). In case term-seq data is absent, PASIFIC also allows prediction of the premature terminator using a specific parameter. In future applications, it would be possible to use one of the computational tools available for prediction of transcriptional terminators, such as TransTermHP (36) or WebGeSTer (37) to predict the 3′ end of the riboregulator. In such cases the sensitivity will depend on the sensitivity of the terminator prediction tool used.

A number of recent studies have predicted that conditional termination is a much more common mode of regulation in bacteria than previously thought (5). However, it

was shown that in many cases, current computational methods are not efficient in the discovery of such elements. The online server we developed, PASIFIC, can be a useful tool in discovery and validation of new such elements in the future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Shany Doron, Omer Zuqert, Gal Ofir, Zohar Erez and Yinon Bar-On for insightful discussion.

FUNDING

Israel Science Foundation [303/12, I-CORE 1796/12]; European Research Council -StG program [260432]; Human Frontier Science Program [RGP0011/2013]; Abisch-Frenkel foundation; Pasteur-Weizmann council grant; Minerva Foundation; Leona M. and Harry B. Helmsley Charitable Trust; Deutsche Forschungsgemeinschaft [DIP grant]. Funding for open access charge: Israel Science Foundation.
Conflict of interest statement. None declared.

REFERENCES

1. Merino,E. and Yanofsky,C. (2005) Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet.*, **21**, 260–264.
2. Naville,M. and Gautheret,D. (2010) Transcription attenuation in bacteria: theme and variations. *Brief. Funct. Genomics*, **9**, 178–189.
3. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic Control by a Metabolite Binding mRNA. *Chem. Biol.*, **9**, 1043–1049.
4. Grundy,F.J. and Henkin,T.M. (2006) From Ribosome to Riboswitch: Control of Gene Expression in Bacteria by RNA Structural Rearrangements. *Crit. Rev. Biochem. Mol. Biol.*, **41**, 329–338.
5. Dar,D., Shamir,M., Mellin,J.R., Koutero,M., Stern-Ginossar,N., Cossart,P. and Sorek,R. (2016) Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science*, **352**, aad9822.
6. Ramu,H., Mankin,A. and Vazquez-Laslop,N. (2009) Programmed drug-dependent ribosome stalling. *Mol. Microbiol.*, **71**, 811–824.
7. Green,N.J., Grundy,F.J. and Henkin,T.M. (2010) The T box mechanism: tRNA as a regulatory molecule. *FEBS Lett.*, **584**, 318–324.
8. Stülke,J. (2002) Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures. *Arch. Microbiol.*, **177**, 433–440.
9. Breaker,R.R. (2011) Prospects for Riboswitch Discovery and Analysis. *Mol. Cell*, **43**, 867–879.
10. Weinberg,Z., Barrick,J.E., Yao,Z., Roth,A., Kim,J.N., Gore,J., Wang,J.X., Lee,E.R., Block,K.F., Sudarsan,N. *et al.* (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.*, **35**, 4809–4819.
11. Weinberg,Z., Wang,J.X., Bogue,J., Yang,J., Corbino,K., Moy,R.H. and Breaker,R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*, **11**, R31.
12. Yao,Z., Barrick,J., Weinberg,Z., Neph,S., Breaker,R., Tompa,M. and Ruzzo,W.L. (2007) A Computational Pipeline for High- Throughput Discovery of cis-Regulatory Noncoding RNA in Prokaryotes. *PLoS Comput. Biol.*, **3**, e126.
13. Naville,M. and Gautheret,D. (2010) Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation. *Genome Biol.*, **11**, R97.
14. Voss,B., Meyer,C. and Giegerich,R. (2004) Evaluating the predictability of conformational switching in RNA. *Bioinformatics*, **20**, 1573–1582.
15. Voss,B., Giegerich,R. and Rehmsmeier,M. (2006) Complete probabilistic analysis of RNA shapes. *BMC Biol.*, **4**, 5.
16. Freyhult,E., Moulton,V. and Clote,P. (2007) Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, **23**, 2054–2062.
17. Manzourolajdad,A. and Arnold,J. (2015) Secondary structural entropy in RNA switch (Riboswitch) identification. *BMC Bioinformatics*, **16**, 133.
18. Badelt,S., Hammer,S., Flamm,C. and Hofacker,I.L. (2015) Thermodynamic and kinetic folding of riboswitches. *Methods Enzymol.*, **553**, 193–213.
19. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
20. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
21. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
22. Liaw,A. and Wiener,M. (2002) Classification and Regression by randomForest. *R news*, **2**, 18–22.
23. Lesnik,E.A., Sampath,R., Levene,H.B., Henderson,T.J., McNeil,J.A. and Ecker,D.J. (2001) Prediction of rho-independent transcriptional terminators in Escherichia coli. *Nucleic Acids Res.*, **29**, 3583–3594.
24. Svetnik,V., Liaw,A., Tong,C. and Wang,T. (2004) Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli,F, Kittler,J and Windeatt,T (eds). *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, Vol. **3077**, pp. 334–343.
25. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
26. Mark Welch,J.L., Rossetti,B.J., Rieken,C.W., Dewhirst,F.E. and Borisy,G.G. (2016) Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E791–E800.
27. Wessler,S.R. and Calvo,J.M. (1981) Control of leu operon expression in Escherichia coli by a transcription attenuation mechanism. *J. Mol. Biol.*, **149**, 579–597.
28. Vitreschak,A.G., Lyubetskaya,E. V, Shirshin,M.A., Gelfand,M.S. and Lyubetsky,V.A. (2004) Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis. *FEMS Microbiol. Lett.*, **234**, 357–370.
29. Godon,J.J., Chopin,M.C. and Ehrlich,S.D. (1992) Branched-chain amino acid biosynthesis genes in Lactococcus lactis subsp. lactis. *J. Bacteriol.*, **174**, 6580–6589.
30. Edwards,A.L. and Batey,R.T. (2009) A structural basis for the recognition of 2′-deoxyguanosine by the purine riboswitch. *J. Mol. Biol.*, **385**, 938–948.
31. Gruber,A.R., Neubock,R., Hofacker,I.L. and Washietl,S. (2007) The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.*, **35**, W335–W338.
32. Cruz,J.A. and Westhof,E. (2011) Identification and annotation of noncoding RNAs in Saccharomycotina. *C. R. Biol.*, **334**, 671–678.
33. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
34. Flamm,C., Hofacker,I.L., Stadler,P.F. and Wolfinger,M.T. (2002) Barrier Trees of Degenerate Landscapes. *Zeitschrift Phys. Chem.*, **216**, 155.
35. Sorek,R. and Cossart,P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.*, **11**, 9–16.
36. Kingsford,C.L., Ayanbule,K. and Salzberg,S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.
37. Mitra,A., Kesarwani,A.K., Pal,D. and Nagaraja,V. (2011) WebGeSTer DB–a transcription terminator database. *Nucleic Acids Res.*, **39**, D129–D135.
38. Gilbert,S.D., Stoddard,C.D., Wise,S.J. and Batey,R.T. (2006) Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *J. Mol. Biol.*, **359**, 754–768.