W154–W159 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh352

Riboswitch finder—a tool for identification of riboswitch RNAs

Peter Bengert[1,2] and Thomas Dandekar[1,2,3,4,*]

[1]Biochemiezentrum, Project Group B2 (Dandekar/Schirmer), University of Heidelberg, Im Neuenheimer Feld 504, 69120 Heidelberg, Germany, [2]Department of Bioinformatics, Biocenter, Am Hubland, University of Würzburg, Germany, [3]EMBL, Postfach 102209, D-69012 Heidelberg, Germany and [4]Max Delbrueck Centre for Molecular Medicine, Robert-Roessle-Strabe 10, 13092 Berlin-Buch, Germany

Received January 27, 2004; Revised and Accepted February 20, 2004

ABSTRACT

We describe a dedicated RNA motif search program and web server to identify RNA riboswitches. The Riboswitch finder analyses a given sequence using the web interface, checks specific sequence elements and secondary structure, calculates and displays the energy folding of the RNA structure and runs a number of tests including this information to determine whether high-sensitivity riboswitch motifs (or variants) according to the *Bacillus subtilis* type are present in the given RNA sequence. Batch-mode determination (all sequences input at once and separated by FASTA format) is also possible. The program has been implemented and is available both as local software for in-house installation and as a web server at http://www.biozentrum.uni-wuerzburg.de/bioinformatik/Riboswitch/.

INTRODUCTION

Riboswitches are metabolic binding domains within certain mRNAs to sense concentrations of their corresponding ligands (metabolites) (1). Upon ligand binding, allosteric rearrangement of mRNA structure modulates gene expression. Recently, a number of papers reported the identification of different RNA riboswitches, including the high-sensitivity *Bacillus subtilis* riboswitch type to sense guanine (which senses 5 nM metabolite concentrations).

Based on these results, it would be desirable to identify new riboswitches by specific RNA motif searches for this element. Here, we describe a strategy to identify such regulatory elements. The present paper demonstrates

(i) that such a strategy is feasible using the riboswitch motif and consensus from Mandal *et al*. (2);

(ii) the resulting predicted new riboswitches matching the reported consensus and includes a number of experimentally confirmed positive controls.

We make the full strategy available as both a server and program package, 'riboswitch finder'.

The identification of such elements is becoming topical as several recent experimental papers show (2–7). We provide, in addition to the server, the program as a flexible source code for database screening, which can be adapted and modified to related regulatory RNA structures if desired by the researcher. For our study, we concentrated on the high-sensitivity *B.subtilis*-like riboswitch, since (i) it is well characterized, (ii) it has a well-confirmed test set of RNA structures bearing the element, and (iii) exploiting this, we can provide a list of strong new candidates for experimental testing. Using a strategy considering primary sequence, secondary structure as well as a fast and accurate folding routine, we derive here a specific program package to identify such elements.

MATERIALS AND METHODS

An RNA identification program was implemented (P.B.) to search for riboswitches. The program flow is given in Figure 1. As a test set a selection of bona fide riboswitches as described by Mandal *et al*. (2) was used (Table 1). In the final version, the program used sequence, secondary structure and folding routines to define and identify riboswitches. For the program package a web interface was written and a server implemented. It runs on a Pentium processor machine. In addition, source code and a simple installation protocol are available for Linux on request from the authors.

We used (Table 1) a consensus-set of 13 known *B.subtilis*-like riboswitches to establish and optimize the program. After optimization, the program was able to detect all these as well as the 19 riboswitches of the larger second test set (both in the web-server-based version and in the database search version,

*To whom correspondence should be addressed at Department of Bioinformatics, Biocenter, Am Hubland, University of Würzburg, Germany. Tel: +49 931 888 4551; Fax: +49 931 888 4552; Email: dandekar@biozentrum.uni-wuerzburg.de

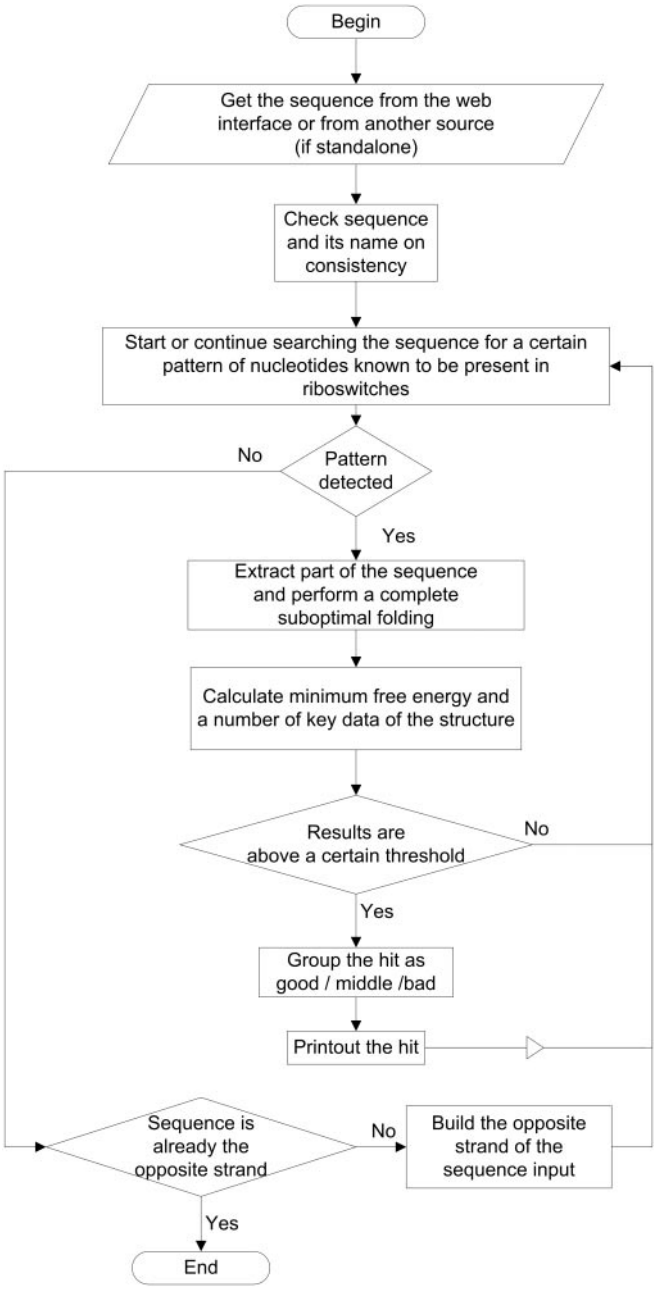


Figure 1. Program flow. After getting the sequence from the web interface, a pattern search including secondary structure and RNA folding including energy values are done. All results are collected and it is scored whether a good, middle or low/tentative riboswitch is present.

Table 1. Bona fide riboswitches used in this study

Consensus-set	BH1, BH2, BH3, BH4, BH5,BS1, BS2, BS3, BS4, BS5, CA1, CA2, CA3
Test-set	CP1, CP2, CP3, CP4, FN1, LL1, LM1, LM2, OI1, OI2, OI3, OI4, SA1, SE1, STA1, STPY1, STPN, TE1, VV1

where it detected all these against a large background of unrelated sequences). Another output example is shown in Figure 4, where riboswitch elements are detected and examined in the test file STPY1 (this sequence is Genbank entry AE006556 of *Streptococcus pyogenes*; the entry is section 85 from 167 covering the complete genome sequence).

We included RNA folding routines as used and implemented by Stiegler and Zuker (8,9) and made available through the Vienna RNA package (10). A detailed scoring function analysed base pairings in the three consecutive stem–loops of the consensus structure (P1, P2 and P3, see Figure 2) of the putative riboswitch. For scoring, paired nucleotides in the three stems were evaluated. Values of P1 $\geqslant$ 5 and P2 $\geqslant$ 5 and P3 $\geqslant$ 5 were classified as 'good', values of P1 $\geqslant$ 3 and P2 $\geqslant$ 3 and P3 $\geqslant$ 3 as 'middle' and values below that as 'bad'.

RESULTS

After validation the program was used to scan the prokaryotic EMBL database (release 75) for new riboswitches. We identified several new riboswitch elements (Table 2) matching the strict consensus structure (elements in Table 1).

The program conducts a specific search including detailed folding and is in this respect highly specific. The new riboswitches identified are available for experimental testing as is the program (server and executable) for public or private in-house screening of new prokaryotic sequences. The identification of all known high-sensitivity guanine riboswitches indicates a very low background of false negatives. No false positives were detected in screening the complete prokaryotic EMBL database, including rRNA or other regulatory and highly structured RNAs mimicking different parts of the riboswitch structure.

Regarding false negatives, the program detects with the option 'strict consensus' (Figure 2) only riboswitch elements matching the structure of the high-sensitivity guanine riboswitch elements as listed in Table 1. However, current research indicates (2–7) that the number and variation of biochemical identified riboswitch elements is steadily increasing. Relaxing the template structure allows the identification of further riboswitches with our program. Using the option 'general consensus' allows for mismatches in the template structure in the loop regions of stem–loop P2 and P3 that are not part of the possible pseudoknot and one nucleotide in the connection between P2 and P3 (Figure 2). This adds to the list of riboswitches while the false positive hits are kept low. The added good hits are the bona fide riboswitch in *Bacillus anthracis*, which is 5′ of purE open reading frame at position 5 374 (EMBL entry AE017025); a hit (at position 673) before the reading frame of the *Spiroplasma citri* SC76 gene (EMBL entry AY136815); and from *Bacillus cereus* (EMBL entry AE016998) a riboswitch at position 298786, which is 5′ of the phosphoribosylaminoimidazole carboxylase open reading frame. Finally, the option 'loose consensus' allows mismatches in both loops P2 and P3, no matter whether the nucleotides are part of the possible pseudoknot or not, allowing a broad screening for related RNA structures potentially matching the template.

Description of the web server

Query. A query is posted by simply pasting the sequence into the query window (Figure 3; accepted formats: Raw, FASTA,

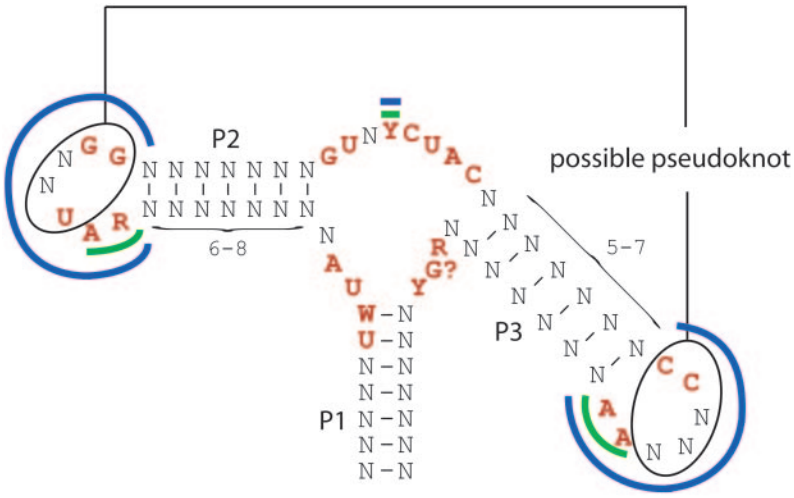


Figure 2. Different consensus options. The general structure of a high-sensitivity guanine riboswitch is shown. The nucleotides marked in red and bold define the strict consensus. No mismatches are allowed here. The option 'general consensus' allows mismatches (marked in green) in the loop regions of P2 and P3 that are not part of the possible pseudoknot and one mismatch in the connection between P2 and P3. Option 'loose consensus' allows mismatches (marked in blue) in both loops P2 and P3, no matter whether the nucleotides are part of the possible pseudoknot or not.

Table 2. Prokaryotic riboswitches identified with riboswitch finder in EMBL database

ID	Pos	Strand	Species	Gram	Function
HI32771	10007(m)	Plus	*Haemophilus influenzae*	Negative	In complement cds
AE017024	260653	Plus	*Bacillus anthracis*	Positive	GMP synthase
AE016998	259613	Plus	*Bacillus cereus*	Positive	GMP synthase
AE017029	46780	Plus	*Bacillus anthracis*	Positive	Xanthine phosphoribosyltransferase
AE017002	301095	Plus	*Bacillus cereus*	Positive	Xanthine phosphoribosyltransferase
AE016954	153881	Minus	*Enterococcus faecalis*	Positive	Xanthine phosphoribosyltransferase
AE017025	52375	Minus	*Bacillus anthracis*	Positive	NupC
AE016999	36552	Minus	*Bacillus cereus*	Positive	NupC
AL935261	262818	Minus	*Lactobacillus plantarum*	Positive	Adenine deaminase
AL935260	2490	Plus	*Lactobacillus plantarum*	Positive	Xanthine/uracil transporter
AE017036	96471	Minus	*Bacillus anthracis*	Positive	Transcriptional regulator GntR family
AE017010	138231	Minus	*Bacillus cereus*	Positive	Transcriptional regulator GntR family
AB008757	103	Minus	*Bacillus stearothermophilus*	Positive	Unclear
AE017024	262613	Plus	*Bacillus anthracis*	Positive	Xanthine/uracil permease
AE016998	261570	Plus	*Bacillus cereus*	Positive	Guanine/hypoxanthin permease
AE015944	141644	Minus	*Clostridium tetani*	Positive	Inosine-5-monophosphate dehydrogenase
AL596165	154144	Minus	*Listeria innocua*	Positive	Xanthine/uracil permease according to pfam
AL596170	223333	Minus	*Listeria innocua*	Positive	Xanthine phosphoribosyltransferase
AP005088	167683	Plus	*Vibrio parahaemolyticus*	Negative	Adenosine deaminase

file of FASTA sequences). Note that for this program, run-time scales only about linearly with sequence length.

Data. Any given RNA sequence can be analysed (no tight length restrictions; up to three million base pairs of nucleotides work well with the server version, beyond that, the database search version is recommended). Also a complete database file or a chunk of genomic DNA can be searched if provided in FASTA format.

RNA analysis routines. Starting (Figure 1) from the input the sequences are searched for the riboswitch pattern. A folding of the RNA structure and a calculation of the mean free energy is included. In the end all resulting elements for the analysed sequence are collected and a coloured sequence output is created.

Analysis results. The output (Figure 4) gives folding results, energy, the exact region where each riboswitch element starts and a graphical display at the bottom of the web page. In most browsers (e.g. Netscape) the obtained web server output file with the complete output can be exported or saved easily in html format, or separately the picture can be saved as jpg and the text as a text file.

Efficiency. Response time of the server is fast. Single sequences are analysed within seconds, depending on the length and web traffic. The server is implemented and ready to use at http://www.biozentrum.uni-wuerzburg.de/bioinformatik/Riboswitch. The search engine can efficiently screen large amounts of data (Figure 1) since it combines a fast pattern-matching routine in PERL with an efficient folding routine and a rapid scoring scheme. Calculation time scales in this way only almost linearly with sequence length since the pattern-matching routine invokes the energy routine only for small subsets of sequence (specific stem–loop regions which could contain a riboswitch). The energy routine itself (10) allows quick calculation of subfoldings.

Figure 3. Web interface for the riboswitch finder.

Identification of new prokaryotic RNA riboswitches

Scanning the EMBL Prokaryotic databases, we were able to detect 19 further putative riboswitches *matching the strict consensus*. Most of them were found in a closely related organism, indicating phylogenetic conservation (Table 2), all identified hits scored 'good' with the exception of hit HI32771, which scored 'middle' (m).

Strong RNA *B.subtilis*-type riboswitches are among others predicted for RNAs encoded in the contigues of AE017024 and AE017025 from *B.anthracis* and AE016998 and AE016999 from *B.cereus*. These riboswitches are phylogenetically conserved in both species and are located in the correct position to be functional (5′-UTR region of the mRNA) of the mRNAs for nucleoside permease nupC (*B.cereus*, AE 16999 and *B.subtilis*, AE017025) and the 5′-UTR of GMP synthase (*B.cereus* AE016998, *B.subtilis* AE014024). Moreover, in all these four instances of a putative riboswitch element, there is good biological context why exact sensing of guanine concentrations would be advantageous, to control either import according to concentration (nucleoside transporter nupC) or the synthesis of GMP (GMP synthase).

Furthermore, and in extension to previous studies, the program readily and with high specificity identifies in other prokaryotic organisms riboswitches of the high-sensitivity guanine sensing type (Table 2). These include further enzymes biochemically involved in purine metabolism, the Xanthine phosphoribosyltransferase (phylogenetically well conserved, examples in the database were *B.anthracis*, *B.cereus*, *Enteroccus faecalis* and *Listeria innocua*), inosin 5-monophosphate dehydrogenase and in the 5′ UTR of the xanthin/uracil permease (*B.anthracis*, *Lactobacilus plantarum*). Further riboswitch-containing enzymes of this type were the guanine/hypoxanthin permease (*B.cereus*) as well as the adenine deaminidase in *L.plantarum* and *Vibrio parahaemolyticus*.

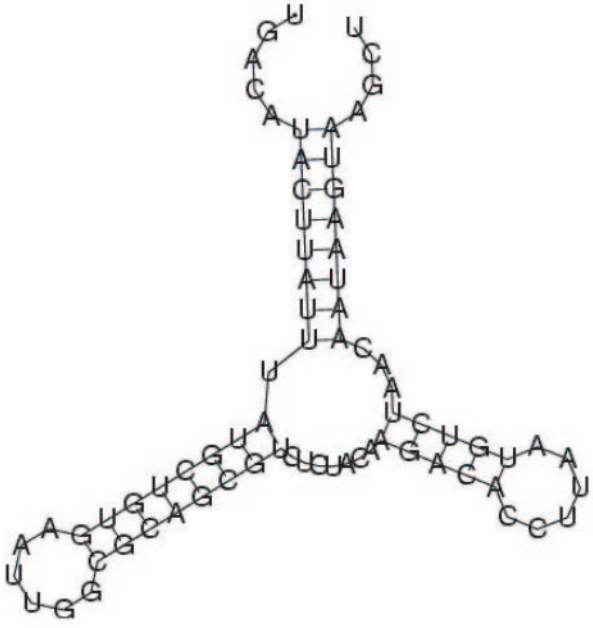
B.anthracis and *B.cereus* include furthermore a riboswitch of the 5′-UTR of a transcriptional regulator of the GntR family. For the HutC/Far-like bacterial transcription factors of the GntR family it is known (11) that they contain a recently described small molecule-binding domain (histidine in HutC, fatty acids in FarR) in the mature protein. Interestingly, the riboswitch element should be able to fulfil a complementary guanine sensor function on the mRNA level in the two RNA

Here are the results for: STPY1

Consensustype used: strict

Hit found:
On **plus** strand at posititon **1288**
Sequence: UGACAUACUUAUUUAUGCUGUGAAUUGGCGCAGCGUCUCUACAAGACACCUUAAUGUCUAACAAUAAGUAAGCU
Structure: .....(((((((.(((((((.....)))))))).....(((((.....)))))...)))))))....
Energy: -19.70 kcal/mol **Paired nucleotides** in P1: 8 P2: 8 P3: 5
Quality: **good**
Structure: .....((((((((((((((.....))))))).....(((((.....)))))....))))))))..
Energy: -19.50 kcal/mol **Paired nucleotides** in P1: 9 P2: 6 P3: 5
Quality: **good**
Structure: .....(((((((.(((((((.....)))))))).....((((.....))))...)))))))....
Energy: -19.10 kcal/mol **Paired nucleotides** in P1: 8 P2: 8 P3: 4
Quality: **middle**

The following picture shows your putative riboswitch:



The blue brackets show Stem P1, the orange brackets stem P2 and the green brackets stem P3.

Figure 4. Result of a search with the riboswitch finder (example file STPY1). Folding results, energy, and base pairings are shown, giving the exact region where the riboswitch starts and a graphical display at the bottom of the web page. In the example the program identifies from the sequence file a known riboswitch in the 5′-UTR of xanthine phosphoribosyltransferase in *S.pyogenes*. Three possible overlapping patterns describing this riboswitch element, each compatible with the consensus structure, are printed out.

molecules identified by the riboswitch finder for *B.anthracis* and *B.cereus* (we confirmed by protein sequence analysis that the described new protein domain should be present in the two encoded proteins).

Finally, the high-scoring riboswitch structure in the cbaAB gene for bo3-type cytochrome *c* oxidase in *Bacillus stearothermophilus* suggests that here also guanine may exert a regulatory effect.

This shows that such riboswitches are widespread, e.g. the structure is also found in Listeria and such structures are predicted in some Gram-negative bacteria. Further identification of such riboswitches in additional databases or new prokaryotic genomes is now easily possible with our server (or alternatively the database search version installed in-house), as demonstrated for another known example by screening Genbank data (Figure 4).

DISCUSSION

A web-based tool for riboswitches has not yet been available that allows the user to identify riboswitches in any new RNA or DNA sequence. Using a strategy considering primary sequence and secondary structure, as well as a fast and accurate folding routine, we derive a useful and specific program to identify riboswitch elements. Their detailed structure shows

some variation, thus our concrete application focuses on the best characterized, high-sensitivity and well-characterized *B.subtilis* riboswitches.

However, the server software and design are established and implemented so that they can further be extended as additional riboswitch motifs are found and characterized, such as the modified motifs reported by Epshtein *et al*. (4) and Winkler *et al*. (7).

The specificity of the program is enhanced by not only looking at sequence and secondary structure but including also energy considerations and RNA folding. This is also shown by the fact that best structures identified by our strategy could often be confirmed by independent evidence such as the correct position of the riboswitch element in the 5′-UTR of mRNA (all examples in Table 2) and further biological context information to support its functionality. Furthermore, the false positive rate is very low; in fact, no rRNA, tRNA or snoRNA was mistakenly assigned as a riboswitch element in all our searches. Interestingly, no such guanine-sensing elements were found in any of the eukaryotic sequences (EMBL database) that we searched with the strict consensus. A list of candidate riboswitches using the general or loose consensus is available on request from the authors.

The success of the detection program shows as well as the biological evidence for the known and the new elements identified in this paper, that high-sensitivity guanine sensing by riboswitches seems to be a widespread mechanism in prokaryotes to regulate metabolism, including both transport and synthesis pathways around specific metabolites (in this case guanine). The modified riboswitches recently identified for regulation of other metabolites suggest that these elements are even more widespread. Modification of the riboswitch program regarding the consensus features is convenient by replacing the regular expressions in the fast pattern-matching routine as desired; three different versions are already provided by the package. Additional stem–loops as a further variation of riboswitch motifs are easily accommodated in the second part of the program, utilizing an efficient folding routine to examine potential riboswitches. This requires, however, sufficient available data on stem–loop arrangement, which is one of the reasons why the well-characterized high-sensitivity guanine riboswitch was taken as the default template structure. The availability of the identification program presented here both as a server and as an easily modifiable in-house search version for new genomes opens up further opportunities for the continued identification of riboswitches.

ACKNOWLEDGEMENT

We thank DFG for support (SFB 544/B2; BO-1099/5-2).

REFERENCES

1. Stormo,G.D. (2003) New tricks for an old dogma: riboswitches as *cis*-only regulatory systems. *Mol. Cell*, **11**, 1419–1420.
2. Mandal,M., Boese,B., Barrick,J.E., Winkler,W.C. and Breaker,R.R. (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, **113**, 577–586.
3. Sudarsan,N., Barrick,J.E. and Breaker,R.R. (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, **9**, 644–647.
4. Epshtein,V., Mironov,A.S. and Nudler,E. (2003) The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl Acad. Sci., USA*, **100**, 5052–5056.
5. Winkler,W.C., Cohen-Chalamish,S. and Breaker,R.R. (2002) An mRNA structure that controls gene expression by binding FMN. *Proc. Natl Acad. Sci., USA*, **99**, 15908–15913.
6. Winkler,W., Nahvi,A. and Breaker,R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952–956.
7. Winkler,W.C., Nahvi,A., Sudarsan,N., Barrick,J.E. and Breaker,R.R. (2003) An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nat. Struct. Biol.*, **10**, 701–707.
8. Zuker,M. (1994) Prediction of RNA secondary structure by energy minimization. *Meth. Mol. Biol.*, **25**, 267–294.
9. Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.
10. Hofacker,I., Fontana,W., Stadler,P., Bonhoeffer,L., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures (The Vienna RNA package). *Monatshefte fuer Chemie (Chemical Monthly)*, **125**, 167–188.
11. Aravind,L. and Anantharaman,V. (2003) HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. *FEMS Microbiol. Lett.*, **222**, 17–23.