Bcheck: a wrapper tool for detecting RNase P RNA genes

Dilmurat Yusuf[1] , Manja Marz[2,3], Peter F. Stadler[3,4,5,1,6], Ivo L. Hofacker[*,1]

[1] Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[2] Institut für Pharmazeutische Chemie, Philipps Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany
[3] Bioinformatics Group, Department of Computer Science University of Leipzig, Härtelstrasse 16-18, D-01407, Leipzig, Germany
[4] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22 D-04103 Leipzig, Germany
[5] Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany
[6] Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA

Email: Dilmurat Yusuf- dilmurat@tbi.univie.ac.at; Manja Marz - manja@staff.uni-marburg.de; Peter F. Stadler - studla@bioinf.uni-leipzig.de; Ivo L. Hofacker - ivo@tbi.univie.ac.at;

[*]Corresponding author

Abstract

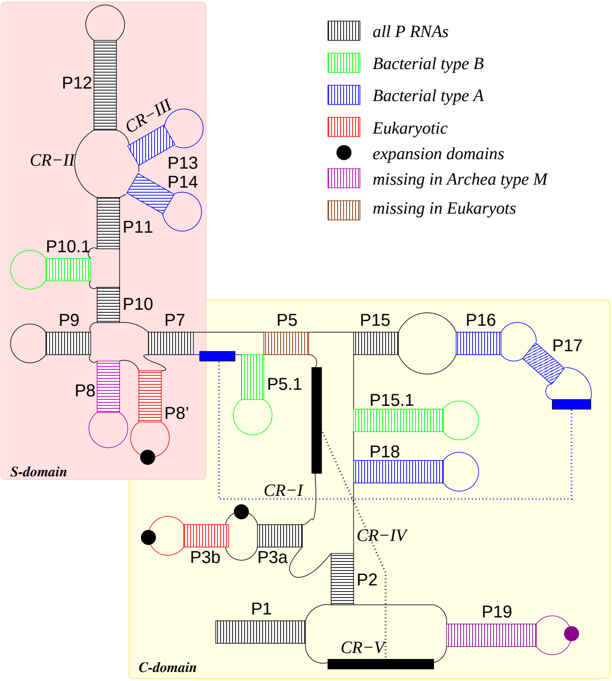
Background: Effective bioinformatics solutions are needed to tackle challenges posed by industrial-scale genome annotation. We present Bcheck, a wrapper tool for predicting RNase P RNA genes by combining the speed of pattern matching and sensitivity of covariance models. The core of Bcheck is a library of subfamily specific descriptor models and covariance models.

Results: Scanning all microbial genomes in GenBank identifies the RNase P RNA in 98% of 1024 microbial chromosomal sequences within just 4 hours on single CPU. Comparing to existing annotations found in 387 of the GenBank files, Bcheck predictions have more intact structure and are automatically classified by subfamily membership. For eukaryotic chromosomes Bcheck could identify the known RNase P RNA gene in 84 out of 85 metazoan genomes, 19 out of 21 fungi genomes. Bcheck predicted 37 novel eukaryotic RNase P RNAs, 32 of which are from fungi organisms. Gene duplication events are observed in at least 20 metazoan organisms. Scanning of meta-genomic data from the Global Ocean Sampling Expedition comprising over 10 million sample sequences (18 Gigabases), predicted 2909 unique genes, 98% of which falls into ancestral bacteria A type of RNase P RNA and 66% of which have no close homolog to known prokaryotic RNase P RNAs.

Conclusions: The combination of efficient filtering by means of a descriptor-based search and subsequent construction of a high-quality gene model by means of a covariance model provides an efficient method for the detection of RNase P RNAs in large-scale sequencing data.
Bcheck is implemented as webserver and can also be downloaded for local use from http://rna.tbi.univie.ac.at/bcheck/index.html

1 Introduction

In recent years, biological sequence databases have grown exponentially. These data include a rapidly increasing number of completely sequenced genomes as well as large-scale metagenomic data set that await annotation. For instance, the Global Ocean



Element	*Bact.* *A*	*B*	*Arch.* *A*	*M*	*Euk.* *P*
P1	●	●	●	●	●
P2	●	●	●	●	●
P3a	●	●	●	●	●
P3b					●
P5	●	●	●	●	
P5.1		●			
P7	●	●	●	●	●
P8	●	●	●		●
P8'					●
P9	●	●	●	●	●
P10	⬮	●	⬮	⬮	⬮
P11		●			
P10.1		●			
P12	●	●	●	●	●
P13	●				
P14	●				
P15	●	●	●	●	●
P16	●		⬮		
P17	●				
P15.1		●			
P18	●	●			
P19	●	●	●		●
P4	●	●	●	●	●
P6	●		●		
CR−I	●	●	●	●	●
CR−II	●	●	●	●	●
CR−III	●	●	●	●	●
CR−IV	●	●	●	●	●
CR−V	●	●	●	●	●

Figure 1: Schematic drawing of the consensus structures of RNase P RNA. Adapted from [1–4]. The table indicates the distribution of structural elements. A black circle in the table represents the occurrence of a particular element. An ellipse shows two elements merged and cannot be separated unambiguously.

Sampling Expedition (GOS) deposited more than 18 G metagenomic sequences already [5]. The analysis of these data calls for new and more efficient methods of data analysis [6].

Non-protein-coding RNA (ncRNA) genes are abundant in genomic sequences, playing diverse important biological roles [7]. The genomic annotation of ncRNA genes is attracting strong research focus, in particular in the context of genome annotation [8, 9] and metagenomics [10, 11]. Methods for homology-based annotation have dramatically improved over the last years. In particular, `Infernal` 1.0 [12] outperform the previous methods by orders of magnitude in speed. Nevertheless, such general purpose approaches do not reach the performance levels of customized class-specific tools, in particular `tRNAscan-SE` [13] in terms of both speed and quality. Manual strategies in some cases [14] reach superior results, but are too time-consuming for larger projects and in most cases are hard to generalize.

`tRNAscan-SE` is not a single algorithm but rather a wrapper tool that combines a series of increasingly complex and expensive filters. Similarly, the major searching strategy of `Rfam` [15] is a combination of a `blast`-based filter and followed by `Infernal`. The pre-filtering at sequence level with `blast`, however, is not ideal in particular in applications to distance homologs [16]. Another common approach is to apply a descriptor of sequence and structural motif to predict ncRNA homologs. The descriptor construction is a manual process, requiring expert knowledge. Several descriptor languages have been developed, e.g., `RNAmot` [17], `PatScan` [18] `HyPaL` [19], `RNAMotif` [20] and Sean Eddy's `rnabob` [21], which is also used here.

RNase P RNA, possibly a remnant of the RNA world [22], is an important ribozyme involved in the processing of pre-tRNAs [23]. Its gene is usually designated as rnpB in eubacteria. A variable number of protein components [24] facilitates substrate binding [25]. RNase P RNA exists in almost all organisms. So far, there is compelling evidence for the loss of RNase P RNA only in a single organism, the archaeon *Nanoarchaeum equitans* [22]. It is not unlikely, however, that plants, red algae, and heterokonts [26], some eubacteria (e.g. *Aquifex aeolicus*) [27, 28]) and some additional archaea (e.g. *Pyrobaculum aerophilum* [27]) have lost their RNAse P RNA. The archaeon *Methanothermobacter thermoautotrophicus* may be a transition towards the loss of RNAse P RNA, which is catalytically inactive in this organism but can be "repaired" by a few substitutions [29].

The length of RNase P RNA ranges from 250 nt to 550 nt. It is divided into two structural domain: S-domain for binding and C-domain for catalysing

[30], see Fig. 1. The secondary structure of RNase P consists of up to 19 conserved stems, denoted P1 to P19, of which P7 to P14 form the S-domain, which is flanked by the C-domain [31]. There are five regions with strong sequence conservation, designated CR-I to CR-V, including the P4 pseudoknot composed by CR-I and CR-V [32].

The RNase P RNA structures can be broadly assigned to five subfamilies: the eubacterial classes A and B (bacA and bacB), the archaeal types A and M (arcA and arcM), and a single eukaryote group (nucA) [33]. In addition, two eukaryotic subtypes in fungi (fugA & fugB) can be identified [34]. The types arcA and bacA, which have been identified as ancestral states [35], cover the majority of microbial rnpB genes, forming diverse sets in terms of both sequence and structure variation. In contrast, the derived types arcM and bacB, in contrast have more uniform members. The diversity is largest among the eukaryotic RNase P RNAs.

In eukaryotes, RNase P RNA is transcribed by polymerase III [36]. The human promoter elements were described recently [3] to contain TATA-box, PSE, Oct and SP1/SPH element within 100 nt upstream of transcription initiation site. A comparison of all eukaryotic promoter elements showed weak similarities only in TATA box.

In the contribution we are concerned with the detection of RNase P RNAs in genomic data from all domains of life. In [32], a pattern matching based pipeline for efficient rnpB gene prediction has been proposed. It is not applicable to large-scale database searches in practise, however. Here, we present `Bcheck`, a wrapper, to perform efficient rnpB gene prediction by combining the fast filtering with `rnabob` [21] and sensitive validation of `Infernal`. The construction of such a method entails two tasks: the design of an efficient yet sensitive descriptor model (DM) that acts as a filter, and the derivation of a sensitive statistics covariance model (CM). Both components are based on a careful analysis of published RNase P RNA sequences and structures. The success of `Bcheck` depends on the efficiency and predictive power of both models, as well as a sensible wrapping algorithm that optimizes the interplay of DM and CM.

2 Algorithm and models

The construction of effective models of RNase P RNAs is a non-trivial task because of the lack of strong family-specific conservation. Our strategy is to first classify the training sequences into the seven sub-families identified in the literature: arcA, arcM, bacA, bacB, nucA, fugA and fugB. The training set consists of sequences from the RNase P RNA Database [37] with intact and complete secondary structures and additional RNase P RNA sequences from the `Rfam` and from two recent publications [26, 38]. A set of randomized decoys as well as randomized genomic sequences were constructed using `ushuffle` [39] in order to determine the false positive rates.

The training of both DM and CM requires structural alignments, whose quality is crucial both for the automatic learning procedures of the `Infernal` CMs and the manual construction of the DMs. We adopted a multi step strategy: The RNase P sequences were first divided into structural elements, then folding regions were structure-aligned manually, and loop regions were sequence-aligned by means of `MUSCLE` [40]. Local alignments were then recombined into a "raw" global alignment for each subfamily. These alignments contain errors, mainly caused by local foldings which are not fitting to the conservation patterns shared by the majority of members. We adopted two correction methods. First, we applied `RNAfold` [41, 42] to check the thermodynamic plausibility of local structure elements. Construction of DMs starts with these alignments. In the course of DM construction, outliers are temporarily removed from the alignments, searched with the preliminary DMs, which provided additional information to guide the re-insertion of the outlier into the alignment.

Since efficiency is the major focus of DM training, we focus on selected features of local regions. To gain consensus sequence information, each alignment column was summarized and assigned with standard, "ambiguous" IUPAC nucleotides (taking into account every nucleotide appearing in the column) or the gap character (whenever the column contained at least one gap). The sequence was edited to take established structural knowledge into account. The resulting consensus sequence was then annotated in the alignment. The RALEE mode [43] in the emacs editor was used for visually inspecting alignments, consensus structures and conservation patterns. Regions with rich conservation in se-

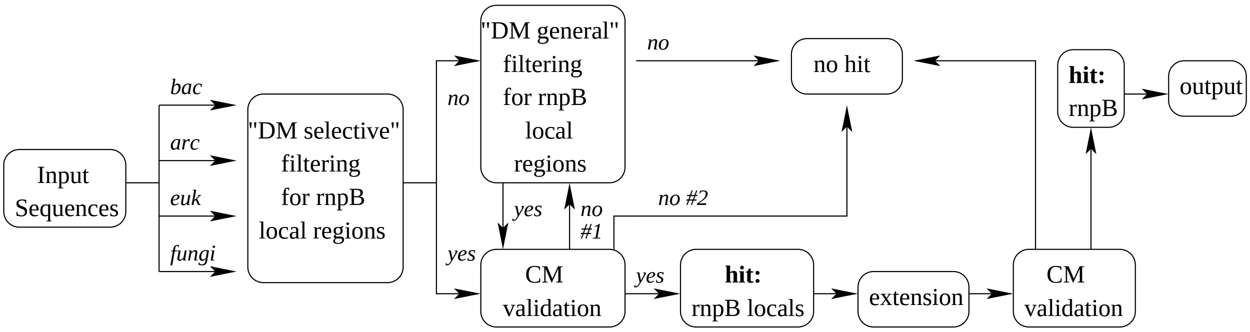


Figure 2: `Bcheck` wrapping algorithm follows *local to global* and *selective to general* strategy. See text for details. arc – archaea descriptor, bac – bacterial descriptor, fungi – fungal descriptor, euk – eukaryotic descriptor, DM – Descriptor Model, CM – Covariance Model.

quence and/or structure were selected for inclusion in the descriptor. A simple example of constructing DM from alignment is shown in Fig. 3. The DM for RNase P RNAs mainly consists of the S-domain and its flanking conserved sequences. Once the feature selection was completed, we carried out a interactive process between DM building and DM testing to adjust the parameters of feature variables balancing between false positive rate and efficiency. Among several descriptor languages we chose `rnabob` as search engine for our DMs because of its convenient syntax and its speed.

For the subfamilies arcA, bacA, and nuc with strong variation, we constructed two variants, "DM selective" and "DM general", with different parameter settings. The selective DMs miss a few aberrant RNase R RNAs, while the "DM general" models have a larger false positive rate. For each subfamily, only one CM is needed and automatically generated based on global structural alignment using the tools of the `Infernal` package.

The `Bcheck` wrapping algorithm takes the strategy of *local to global* and *selective to general*, Fig. 2. At first, subfamily-specific DMs locate candidate genes. If no valid hit was produced by the "selective" model, the corresponding "DM general" is applied. Then the CM is applied in local alignment mode to validate the candidate. Valid hits, i.e., those recognized by the CM, are extended by 150 nt and 300 nt at 5' and 3' ends, respectively, and fed to the CM in global alignment to produce better estimates of the ends. At both phases, an E-value threshold of $E \leq 10^{-10}$ must be reached.

To distinguish functional copy and pseudogene of eukaryotes, we analyzed their promoter regions. For this purpose we aligned 100nt upstream of Polymerase III transcripts of the same organism and compared the RNase P predictions. Pseudogenes are marked as such within the `Bcheck` webserver, see sect. 3.4.

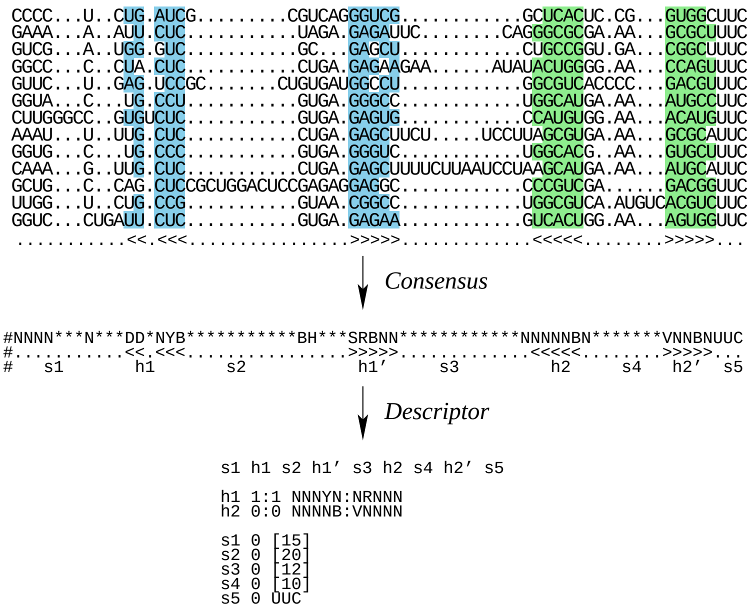
3 Applications

3.1 Procaryote rnpB genes in GenBank

We used `Bcheck` to scan the genomic sequences of 956 bacteria and 68 archaea organisms from `GenBank`. The entire computation, which surveyed 3.1 G of input sequence, took approximately 4 hours to complete with single core of 2.4 GHz Intel(R) Core(TM)2 CPU. `Bcheck` produced one hit per organism for 98% (1005) of organisms, see Tab. 1. The default algorithm yielded no prediction in 29 organisms, for 10 of which a direct CM search was successful. `Bcheck` predictions for three members of the phylum *Chloroflex* (*Roseiflexus castenholzii*, *Roseiflexus RS-1*, and *Chloroflexus aggregans*) are only partial rnpB regions including partial-P11, P12 and junctions between two stems.

After removing duplicate sequences from closely related strains, we obtained 777 unique rnpB genes of which 45 belong to arcA, 10 to arcM, 621 to bacA, and 101 to bacB, see Tab. 3 below.

The `GenBank` files contained annotated rnpB genes for 365 eubacteria and 22 archaea, all of which were among the `Bcheck` predictions. We then compared start-end positions of `Bcheck` predictions and `GenBank` annotations. Only 25% of the annotations agree within a discrepancy of 5 nt or less at



```
CCCC...U..CUG.AUCG.........CGUCAGGGUCG............GCUCACUC.CG...GUGGCUUC
GAAA...A..AUU.CUC...........UAGA.GAGAUUC........CAGGGCGCGA.AA...GCGCUUUC
GUCG...A..UGG.GUC...........GC...GAGCU............CUGCCGGU.GA...CGGCUUUC
GGCC...C..CUA.CUC...........CUGA.GAGAAGAA......AUAUACUGGGG.AA...CCAGUUUC
GUUC...U..GAG.UCCGC.......CUGUGAUGGCCU............GGCGUCACCCC...GACGUUUC
GGUA...C...UG.CCU...........GUGA.GGGCC............UGGCAUGA.AA...AUGCCUUC
CUUGGGCC..GUGUCUC...........GUGA.GAGUG............CCAUGUGG.AA...ACAUGUUC
AAAU...U..UUG.CUC...........CUGA.GAGCUUCU.....UCCUUAGCGUGA.AA...GCGCAUUC
GGUG...C...UG.CCC...........GUGA.GGGUC............UGGCACG..AA...GUGCUUUC
CAAA...G..UUG.CUC...........CUGA.GAGCUUUUCUUAAUCCUAAGCAUGA.AA...AUGCAUUC
GCUG...C..CAG.CUCCGCUGGACUCCGAGAGGAGGC............CCCGUCGA......GACGGUUC
UUGG...U..CUG.CCG...........GUAA.CGGCC............UGGCGUCA.AUGUCACGUCUUC
GGUC...CUGAUU.CUC...........GUGA.GAGAA............GUCACUGG.AA...AGUGGUUC
 ...........<<.<<<.................>>>>>.............<<<<<........>>>>>...
```

```
#NNNN***N***DD*NYB***********BH***SRBNN************NNNNNBN*******VNNBNUUC
#...........<<.<<<.................>>>>>............<<<<<........>>>>>...
#   s1       h1       s2           h1'     s3         h2     s4   h2'  s5
```

```
s1 h1 s2 h1' s3 h2 s4 h2' s5

h1 1:1 NNNYN:NRNNN
h2 0:0 NNNNB:VNNNN

s1 0 [15]
s2 0 [20]
s3 0 [12]
s4 0 [10]
s5 0 UUC
```

Figure 3: Construction of a descriptor model (DM). A simple example based on a partial RNase P sequence is shown here. The refined alignment columns are annotated with consensus structural and sequence information. The DM is then constructed by manual inspection of the best-conserved regions, taking into account both sequence and structure variation observed in the alignment.

Table 1: Summary of predicted microbial rnpB for GenBank genome data set. "Known" refers to organisms with annotated rnpB genes, whereas "unknown" refers to organisms with rnpB genes unannotated.

Domain	known	unknown	total	CM only
Eubacteria	365/365	581/591	946/956	7/946
Archaea	22/22	37/46	59/68	3/59
Total	387/387	618/637	1005/1024	—

Table 2: Evaluation of the five major discrepancies between `GenBank` annotation and `Bcheck` results. `Rfam` scores are bit-scores for `Infernal` using `Rfam`'s CM models. The discrepancies column lists features missing in the region annotated in `GenBank`.

Organism	`Rfam` scores		Discrepancy
	`GenBank`	`Bcheck`	
M. acetivorans	-63.35	167.88	P1, P7
A. cellulolyticus	-132.16	221.11	all
E. coli(CFT073)	-72.51	282.72	most
R. typhi wilmington	-97.66	264.53	P1, P3, P9, P10
B. halodurans	-110.59	300.87	P1, P9

both ends. Inspection of sequences and predicted secondary structures shows that the published sequences are in general inaccurate: At the 5' end, 66% known annotations miss flanking regions of P4, ranging from 30 to 90 nucleotides. At the 3' end, 56% known annotations miss flanking regions of P4', ranging from 10 to 20 nucleotides. A few of the `GenBank` annotations, furthermore, have promoter or terminator sequences included. `Bcheck` thus provides a substantial improvement also of the existing annotations in most cases.

The published annotation is more accurate than the `Bcheck` prediction only in a single case: *Roseiflexus RS 1*. In five cases, the published annotation and the Bcheck results differ dramatically. In order to evaluate these cases further, we employed the CM model of the Rfam, which supported the authenticity of the Bcheck predictions, Tab 2.

3.2 RNase P RNAs in metagenomic sequences

The GOS metagenomic sequences were obtained from the CAMERA project [44]. Due to the taxonomic uncertainty of the GOS data set, all models of archaea, eubacteria, and eukaryotes were applied to search over 10 million sequences comprising about 18 G. No hit was produced by any of the eukaryotic models.

In total, Bcheck predicted 4675 rnpB genes with median E-valuse of 10^{-78}. In 211 cases two models overlapped. In these cases there was a clear difference in E-values, so that the assignment to domains was unambiguous in all positive cases. After duplication removal, 2909 rnpB sequences are unique, 2857 of which belong to bacA, 49 to arcA, 3 to bacB, but none for arcM, see Tab. 3.

The ancestral types arcA and bacA are clearly predominant in both GenBank and GOS data set. In the marine samples, the number of bacA rnpBs exceeds 95%. We compared rnpB sequences from two datasets w.r.t. their GC content, Fig. 4. Differences are particularly obvious in eubacteria, where the majority of GOS bacA sequences have low GC-content, while the median GC content of GenBank rnpB is high, with ≈ 0.6.

We use the detected rnpB genes as a marker to infer the taxonomic distribution of GOS samples. We used blast to find the closest orthologs of 2909 unique GOS rnpB genes the among the 777 GenBank sequences using an E-value cutoff of $E < 10^{-50}$. High scoring orthologs are found for 1003 GOS rnpB genes, 914 of which have only one ortholog making species assignment possible, and 39 of which have multiple orthologs from a single genus. These species assignments and genus assignments are shown in Fig. 5. The identified organisms are mostly eubacteria belonging to the three phyla proteobacteria, cyanobacteria and bacteroidetes. Only a single archaeon, *Nitrosopumilus maritimus*, was recognized. Among eubacteria, most sequences belong to *Pelagibacter ubique* (75%) and *Prochlorococcus marinus* (13%). For 1906 GOS rnpBs (66%), no close homologs are known, suggesting that they derive from unknown species. Of these 1859 (97.5%) belonging to bacA subfamily, 44 (2.3%) to arcA subfamily, and 3 (0.1%) to bacB subfamily.

3.3 RNase P RNAs in eukaryotic genomes

We investigated 237 eukaryotic genomes, Tab. 4. Of the previously annotated genes, we recovered 84 of 85 metazoan and 19 of 21 fungal RNase P RNAs. We miss the *Otolemur garnetti* sequence because of a 3 nt insertion within the highly conserved P4, which is used as a block in all descriptors. For the two related fungi *Coprinus cinereus* and *Laccaria bicolor* hypothetical RNase P RNAs have been reported [26]. Both sequences, however, differ substantially from the CM and are not recognized by Bcheck. On the other hand, Bcheck made novel predictions for 32 fungi and 4 metazoans (*Meloidogyne hapla*, *Aedes aegypti*, *Canis familiaris* and *Taeniopygia guttata*) and the choanoflagellate *Monosiga brevicollis.*

Strong promoter signals were identified for *Meloidogyne hapla* and *Monosiga brevicollis*, supporting that these candidates are functional copies. For 36 metazoan genomes, Bcheck made multiple predictions. In at least 16 cases, the additional predictions seem to be due to assembly errors rather than constituting true paralogs. In the other 20 cases differences in the flanking regions and the RNase P RNA itself indicate that we see the results of gene duplications. In each of these cases, a presumably functional RNase P RNA like promoter structure was found for only one of the copies. Similar duplication patterns are observed in closely related primate, fish and rodent. For instance, both *Homo sapiens* and *Pan troglodytes* have functional copies on chromosome 14 and a pseudogene on chromosome 4. Among teleosts, both *Danio rerio* and *Gasterosteus aculeatus* have their functional copies and pseudogenes on the chromosome 2. In rodent family, *Rattus norvegicus* and *Mus musculus* have the pseudogenes spreading on at least 4 different chromosomes.

Novel RNase P RNA genes were detected by Bcheck in many fungi and several newly sequenced genomes, which had not been analyzed in much detail so far. Only eight sequences which were reported before were not recognized by Bcheck. In 119 sequences Bcheck found 37 novel RNase P RNA genes. The remaining 82 genomes are either unfinished drafts, so that the RNase P RNA is not contained in the data, or they belong to clades where RNase P RNA may be absent. In plants, red al-

Table 3: The subfamily distribution of microbial rnpB. No hit was obtained with any of the eukaryotic DMs.

Data set	arcM	arcA	bacB	bacA	total
GenBank	10	45	101	621	777
GOS	0	49	3	2857	2909

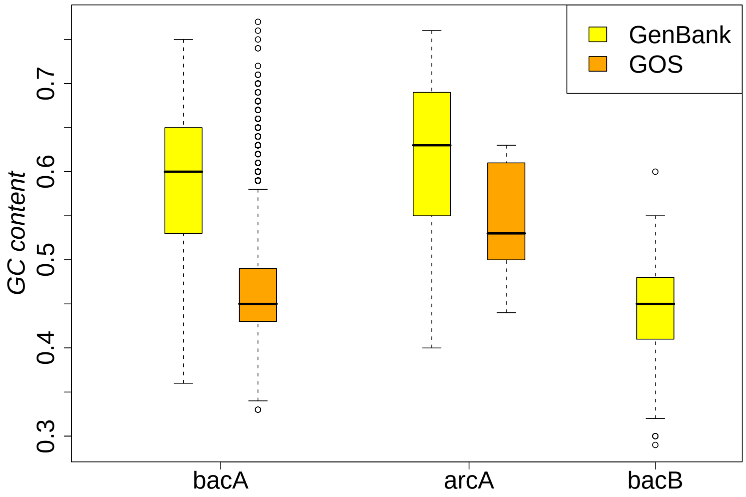


Figure 4: Comparison of GC contents in the GOS and `GenBank` data sets. The statistics were calculated based on unique genes with intact secondary structures. The difference of GC-content is particularly obvious in the eubacteria domain.

gae, and heterokonts RNase MRP RNA, an ancient paralog of RNase P RNA, is well described [3, 26]. One may speculate that it substitutes for RNase P RNA in these clades, in particular given that multiple copies of RNase MRP RNA are present in plant genomes. Incomplete genome assemblies explain e.g. the deviant RNase P in the genome of the elephant (*Loxodonta africanus*), which shows a canonical sequence interrupted by a run of Ns in the latest assembly (Loxafr3.0). We suspect that we missed the RNase P RNA in some fungi and some of the basal eukaryotes due to highly divergent sequence and secondary structure.

3.4 Software, webserver, and database

`Bcheck` was written in Python (version 2.5.2). Input consists of DNA or RNA sequences in `fasta` format, rnpB predictions are output with `fasta` format or with secondary structure annotated. Besides the default searching algorithm, `Bcheck` also gives the option for searching with CM only. However, CM-only search is at least 100 times slower.

We set up a `Bcheck` webserver to facilitate online RNase P RNA gene prediction. A searchable rnpB database was developed, including genes for 1005 microbial organisms, 147 eukaryote organisms and 4756 GOS sample sequences. The predicted pseudogenes for eukaryote organisms are also included.

The "rnpB database" uses a hierarchical tree structure, consisting of 5 tables, implementing preorder tree traversal algorithm to process query efficiently. `blast` is also offered in the server for homology search against the database compromising 777 unique rnpB genes. The sever can be accessed at http://rna.tbi.univie.ac.at/bcheck/. The `Bcheck`-pipeline can also be download from the same location for the local usage in a Linux environment.

4 Discussion and conclusions

The rapidly increasing size of sequence databases requires efficient tools for data analysis. In particular, homology annotation of small ncRNAs, with their short and often poorly conserved sequences poses a severe problem for large-scale annotation. Here,

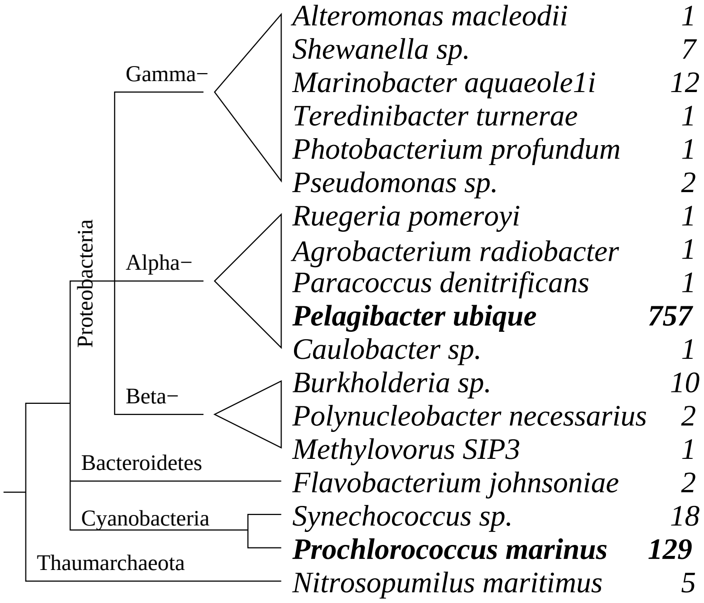


Figure 5: Phylogenetic distribution of rnpB genes detected in the GOS data set. 99.5% of the sequences are of eubacterial origin, with three quarters deriving from *Pelagibacter ubique* and another 13% coming from *Prochlorococcus marinus*. Only 5 hits are of Archaeal origins.

Table 4: Summary of predicted RNase P transcripts in eukaryotes. "Known" refers to organisms with annotated rnpB genes, whereas "unknown" refers to organisms with unannotated rnpB genes.

	known	unknown	Sum
Metazoans	84/85	4*/13	88/98
Fungi	19/21	32/49	51/70
Heterokonts	0/0	0/6	0/6
Plants	0/0	0/30	0/30
Other Eukaryotes	7/12	1/21	8/33
Sum	110/118	37/119	147/237

* No common promoter signals observed: 3 out of 4.

we describe Bcheck, an efficient pipeline to determine RNase P RNAs across all three domains of life. In order to deal with the high variability of the RNase P RNA sequences and structures, we employ descriptor-based models specific for sub-families instead of a single pattern to construct more efficient filters. In the second step, improved covariance models are used to validate the candidates from the DM step and to determine nearly exact gene boundaries.

With Bcheck, we were able to determine the RNase P RNA sequences of 59 out of 68 archaea, 946 out of 956 eubacteria, and 147 out of 237 eukaryotes. 61% of the prokaryotic sequences and 25% of eukaryotic result were not annotated previously. The quality of the predicted rnpB gene is much better than a large fraction of the – usually blast-based – annotation available through GenBank. The size and diversity of eukaryote genomes brings with it a particular challenge in finding RNase P RNAs, because this diversity is reflected in many aberrant features of the RNase P RNA itself. Using the fungi-specific DMs, we uncovered 32 previously unannotated sequences. As in previous studies, we did not find RNase P RNA candidates in plants and Heterokonta.

Since Bcheck is more than 100 times faster than the direct application of Infernal (version 1.0), it is suitable in particular as tool to screen large high-throughput sequencing data. With only a handful of false negatives (10 out of 1005 prokaryotes), Bcheck provides a highly efficient way to annotate newly sequences genomes. A particular strength of Bcheck is its applicability to metagenomics data.

Among the 19 prokaryotic genomes for which Bcheck failed to detect a candidate, 15 have a size below 2.0 Mbp. One of them, *Nanoarchaeum equitans*, is among three organisms having extremely

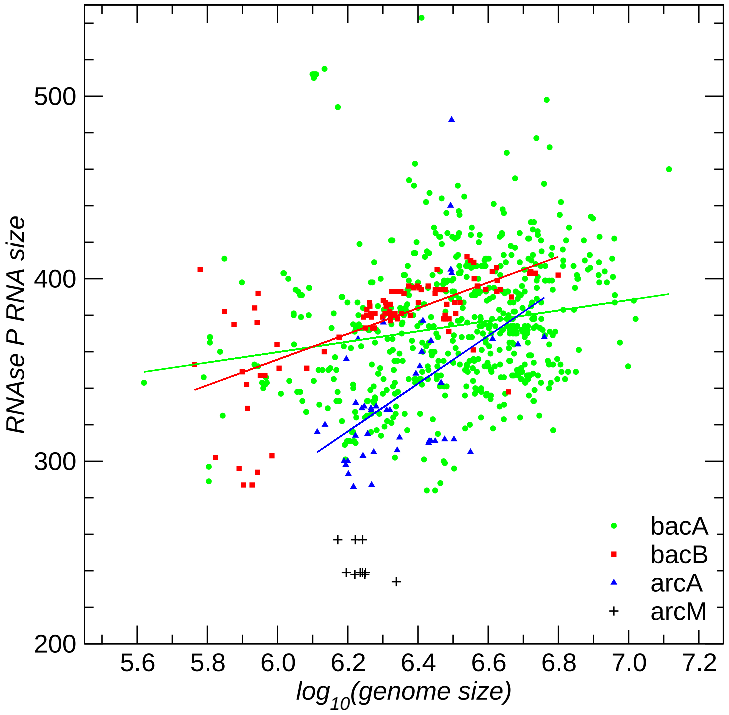


Figure 6: Correlation of the size of RNase P RNA genes with genome size. For bacA ($\rho = 0.16$) bacB ($\rho = 0.66$), and arcA ($\rho = 0.47$) there is a weak but significant positive correlation. The few sequences of type arcM are significantly shorter and are restricted to genomes in a very narrow size range.

condensed genomes with length even below 0.5 Mbp. *Nanoarchaeum equitans* appears to have lost its RNase P RNA, whose function has been taken over by the protein components [22].

In Fig. 6 we summarize the correlations between genome size and the size of prokaryotic RNase P RNA. Even though there is no strong correlation indicated in arcA and bacA subfamilies, the evolutionarily younger bacB and arcM seem to be more strongly affected by changes in genome size.

At present, Bcheck models were built on the conserved sequence and secondary structure features of a large sample of RNase P RNAs. Conceivably, the predictive power of the pipeline could be improved further by include additional information. For instance, promoter and terminator regions might be utilized. A recent survey for 7SK RNAs capitalized largely on the conserved features of the characteristic pol-III promoter signals of this ncRNA class [45]. A similar strategy might allow a further relaxation of the DM pattern in favour of a second filter utilizing the promoter and terminator motifs.

Authors contributions

I.L.H. designed the study. D.Y. constructed descriptors, designed and implemented the pipeline. M.M. prepared the training sets for eukaryote models and performed eukaryote hit validation with promoter analysis. All authors collaborated on analysing the data and drafting the manuscript.

Acknowledgement

We are grateful to Thomas Rattei (Department of Genome Oriented Bioinformatics, Technische Universität München) for making the GOS dataset available to us in convenient form, and Dipl.-Inf. Henry Wirth (Interdisciplinary Centre for Bioinformatics, University of Leipzig) for producing the boxplot figure. This work has been funded by the Austrian GEN-AU projects “bioinformatics integration network III”, and in part supported by *Landesstipendium Sachsen*.

References

1. Marquez SM, Harris JK, Kelley ST, Brown JW, Dawson SC, Roberts EC, Pace NR: **Structural implications of novel diversity in eucaryal RNase P RNA**. *RNA* 2005, **11**:739–751.

2. Walker SC, Engelke DR: **Ribonuclease P: the evolution of an ancient RNA enzyme**. *Crit Rev Biochem Mol Biol.* 2006, **41**:77–102.

3. Woodhams MD, Stadler PF, Penny D, Collins LJ: **RNAse MRP and the RNA Processing Cascade in the Eukaryotic Ancestor**. *BMC Evol. Biol.* 2007, **7**:S13.

4. Zhu Y, Pulukkunat DK, Li Y: **Deciphering RNA structural diversity and systematic phylogeny from microbial metagenomes**. *Nucleic Acids Res.* 2007, **35**:2283–2294.

5. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcn LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific**. *PLoS Biol.* 2007, **5**:e77.

6. Rust AG, Mongin E, Birney E: **Genome annotation techniques: new approaches and challenges**. *Drug Discov. Today* 2002, **7**:S70–76.

7. Eddy SR: **Non-coding RNA genes and the modern RNA world**. *Nat. Rev. Genet.* 2001, **2**:919–929.

8. The Athanasius F Bompfünewerer RNA Consortium:, Backofen R, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, Mosig SJ Axel Prohaska, Rose D, Stadler PF, Tanzer A, Washietl S, Sebastian W: **RNAs Everywhere: Genome-Wide Annotation of Structured RNAs**. *J. Exp. Zool. B: Mol. Dev. Evol.* 2007, **308B**:1–25.

9. Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF: **Non-Coding RNA Annotation of the Genome of *Trichoplax adhaerens***. *Nucleic Acids Res.* 2009, **37**:1602–1615.

10. Weinberg Z, Perreault J, Meyer MM, Breaker RR: **Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis**. *Nature* 2009, **462**:656–659.

11. Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ, Breaker RR: **Identification of candidate structured RNAs in the marine organism *'Candidatus Pelagibacter ubique'***. *BMC Genomics* 2009, **10**:268.

12. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments**. *Bioinformatics* 2009, **25**:1335–1337.

13. Lowe TM, Eddy S: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucl. Acids Res.* 1997, **25**:955–964.

14. Mosig A, Zhu L, Stadler PF: **Customized strategies for discovering distant ncRNA homologs**. *Brief. Funct. Genomics Proteomics* 2009, **8**:451–460.

15. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database**. *Nucleic Acids Res* 2009, **37**(Database issue):136–140.

16. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Res.* 2005, **33**:D121–124.

17. Gautheret D, Major F, Cedergren R: **Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA**. *Comput. Appl. Biosci.* 1990, **6**(4):325–331.

18. Dsouza M, Larsen N, Overbeek R: **Searching for patterns in genomic data**. *Trends Genet.* 1997, **13**:497–498.

19. Gräf S, Strothmann D, Kurtz S, Steger G: **HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns**. *Nucl. Acids. Res.* 2001, **29**:196–198.

20. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R: **RNAMotif, an RNA secondary structure definition and search algorithm**. *Nucl. Acids Res.* 2001, **29**(22):4724–4735.

21. Eddy SR: **RNABOB: a program to search for RNA secondary structure motifs in sequence databases** 1992-1996. [http://selab.janelia.org/software.html].

22. Randau L, Schröder I, Söll D: **Life without RNase P**. *Nature* 2008, **453**:120–123.

23. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S: **The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme**. *Cell* 1983, **35**:849–857.

24. Evans D, Marquez SM, Pace NR: **RNase P: interface of the RNA and protein worlds**. *Trends Biochem. Sci.* 2006, **31**:333–341.

25. Reich C, Olsen GJ, Pace B, Pace NR: **Role of the protein moiety of ribonuclease P, a ribonucleoprotein enzyme**. *Science* 1988, **239**:178–181.

26. Piccinelli P, Rosenblad MA, Samuelsson T: **Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes**. *Nucleic Acids Res.* 2005, **33**:4485–4495.

27. Li Y, Altman S: **In search of RNase P RNA from microbial genomes**. *RNA* 2004, **10**:1533–1540.

28. Marszalkowski M, Willkomm DK, Hartmann RK: **5'-end maturation of tRNA in *Aquifex aeolicus***. *Biol. Chem.* 2008, **389**:395–403.

29. Li D, Willkomm DK, Hartmann RK: **Minor changes largely restore catalytic activity of archaeal RNase P RNA from *Methanothermobacter thermoautotrophicus***. *Nucleic Acids Res.* 2009, **37**:231–242.

30. Loria A, Pan T: **Domain structure of the ribozyme from eubacterial ribonuclease P**. *RNA* 1996, **2**:551–563.

31. Krasilnikov AS, Yang X, Pan T, Mondragón A: **Crystal structure of the specificity domain of ribonuclease P**. *Nature* 2003, **421**:760–764.

32. Li Y, Altman S: **In search of RNase P RNA from microbial genomes**. *RNA* 2004, **10**:1533–1540.

33. Brown JW: **The Ribonuclease P Database**. *Nucleic Acids Res.* 1999, **27**:314.

34. Frank DN, Adamidi C, Ehringer MA, Pitulle C, Pace NR: **Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA**. *RNA* 2000, **6**:1895–1904.

35. Haas ES, Williams D, Frank DN, Brown JW: **New insight into RNase P RNA structure from comparative analysis of the archaeal RNA**. *RNA* 2001, **7**:220–232.

36. Schramm L, Hernandez N: **Recruitment of RNA polymerase III to its target promoters**. *Genes Dev.* 2002, **16**:2593–2620.

37. Brown JW: **The Ribonuclease P Database**. *Nucleic Acids Res.* 1999, **27**:314–314.

38. Marz M, Schoen A, Stadler P: **RNase MRP and RNase P**. *in prep* 2010.

39. Jiang M, Anderson J, Gillespie J, Mayne M: **uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts**. *BMC Bioinformatics* 2008, **9**:192–192.

40. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113–113.

41. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures**. *Monatsh. Chem.* 1994, **125**:167–188.

42. Hofacker IL: **Vienna RNA secondary structure server**. *Nucleic Acids Res.* 2003, **31**:3429–3431.

43. Griffiths-Jones S: `RALEE`—**RNA ALignment editor in** `Emacs`. *Bioinformatics* 2005, **21**:257–259.

44. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: a community resource for metagenomics**. *PLoS Biol.* 2007, **5**:e75.

45. Gruber A, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF: **Arthropod 7SK RNA**. *Mol. Biol. Evol.* 2008, **1923-1930**:25.