Diploma thesis

hIntaRNA - Comparative prediction of sRNA targets in prokaryotes.

By:

Patrick R. Wright

Supervision:

Prof. Dr. Wolfgang R. Hess

Prof. Dr. Rolf Backofen

Dr. Jens Georg

Dipl. Bioinf. Andreas S. Richter

Declaration

The investigations reported in the presented work, were undertaken from July 2011 to January 2012 in the Department of Genetics and Experimental Bioinformatics, Biology III at the Faculty of Biology at the Albert-Ludwigs-University Freiburg im Breisgau. The work was supervised by Prof. Dr. Wolfgang R. Hess, Prof. Dr. Rolf Backofen, Dr. Jens Georg and Dipl. Bioinf. Andreas S. Richter.

I herewith affirm that I completed the work myself, and did not use any other than the mentioned sources and utilities.

Freiburg, March 2012

_____

Patrick R. Wright

Acknowledgements

Firstly I would like to thank my parents Gabriele M. König and Anthony D. Wright as well as other family members for their continued support and guidance not only with respect to my work but also with respect to day to day situations. I value your advice.

I also thank Prof. Dr. Wolfgang R. Hess and Prof. Dr. Rolf Backofen for providing me with the opportunity to work on this project, and for fruitful discussions on the matter. Furthermore I thank them for giving me the chance to attend the RNA meeting in Kassel.

I thank Dr. Jens Georg and Andreas S. Richter for their supervision, patience and the time they invested in me. I am especially thankful to Jens for having the idea behind hIntaRNA in the first place.

Furthermore I acknowledge the help of the IT branch of AG Hess for resolving the technical issues I encountered along the way.

General gratitude is owed to the entire AG Hess for making every day working easy and enjoyable.

I would also like to show my gratitude to Dr. Kai Papenfort who helped me in attaining the only wet lab based result of this thesis.

Finally, I thank Matthias Kopf and Beate Kaufmann both of whom have been great friends and colleagues since week one in October 2006. Work would have been much harder without having you guys doing it alongside.

Contents

1 Abstract

Background: Prediction of targets of bacterial small RNAs (sRNAs) is a very challenging task, addressed by several approaches. Experimental testing and verification of sRNA targets is costly and labour-intensive. Therefore, the reliable algorithmic prediction of putative sRNA targets could vastly reduce the amount of wet lab work. However, due to very short and often imperfect complementarity between the sRNA and its target the prediction is not a trivial task. The **int**e**ra**cting **RNA** (IntaRNA) algorithm is one approach, which frequently, however, does not yield satisfying results yet and therefore demands improvement.

Approach: It has been stated that small RNA targets should be predicted in a comparative manner. Even though this was originally stated for eukaryotic RNAs, the basic idea of this thesis also holds for bacteria. The task of improving the IntaRNA algorithm's prediction quality utilizes exactly this concept, also incorporating the individual phylogenetic distances between the analyzed organisms. For instance, it has been verified experimentally that the MicA and RybB sRNAs in *E. coli* and *Salmonella* each have homologous targets in both organisms, thus indicating conservation on the regulatory level. Here, the implementation of the idea that overlapping target predictions for distinct organisms yield stronger evidence of correct functional prediction is presented. IntaRNA target predictions are combined with phylogenetic information by transforming the IntaRNA energy scores into p-values and combining these in order to attain a combined score for a group of homologous genes.

Results: A Benchmark on a dataset of 74 experimentally verified targets yielded promising results, improving 68.9% of all predictions compared to regular IntaRNA. 37.8% of the true positive predictions were in the top 10, six of these on rank one. Application of hIntaRNA to several sRNAs from bacteria as divergent as alpha-proteobacteria, enterobacteria, or cyanobacteria, suggested a multitude of novel yet unreported sRNA targets. The results hint at the Spot42 sRNA playing a major role in the regulation of citric acid cycle associated genes. Furthermore hIntaRNA analysis suggested the interaction of the RybB sRNA with the yet hardly studied outer membrane protein stm1530 in *Salmonella*, this was subsequently verified experimentally. Additionally, the *Agrobacterium tumefaciens* AbcR1 sRNA seems to control more ABC-transporter related mRNAs than previously conceived. The program is implemented in Perl and R and can be readily used on Linux systems as command line tool. An easy to use webserver is planned for the future.

Zusammenfassung

Hintergrund: Die Vorhersage von Zielen bzw. Targets bakterieller kleiner RNAs (sRNAs) ist eine anspruchsvolle Aufgabe die bereits durch verschiedenste Lösungsansätze angegangen wurde. Experimentelle Überprüfung und Verifikation von sRNA Targets ist zeitlich und finanziell aufwendig. Folglich könnte die verlässliche algorithmische Vorhersage von sRNA Targets zu einer maßgeblichen Reduzierung der experimentellen Arbeit führen. Imperfekte und kurze Paarungen zwischen sRNAs und ihren Zielen erschweren verlässliche Vorhersagen. Der **int**eracting **RNA** (IntaRNA) Algorithmus ist einer der Ansätze zur Lösung des Problems, der aber oft keine verlässlichen Ergebnisse liefert und daher verbessert werden sollte.

Ansatz: Es ist vorgeschlagen worden, die Ziele von kleinen RNAs komparativ vorherzusagen. Dies wurde ursprünglich für Eukaryoten in Erwägung gezogen, gilt aber auch für bakterielle sRNAs. Die hier verfolgte Absicht, IntaRNA Vorhersagen zu verbessern nutzt dieses Konzept und beinhaltet auch eine Betrachtung der individuellen phylogenetischen Abstände der verglichenen Organismen zueinander. Experimentell wurde bestätigt, dass die sRNAs MicA und RybB homologe Targets in *E. coli* und *Salmonella* haben. Dies bestätigt, dass es eine Konservierung auf der regulativen Ebene gibt. In diesem Projekt, wird die Implementierung der Idee, dass überlappende Targetvorheragen für unterschiedliche Organismen erhöhte Evidenz für tatsächlich funktionale Regulation darstellt, vorgestellt. IntaRNA Vorhersagen werden mit phylogenetischer Information kombiniert wobei IntaRNA Energiescores in p-Werte umgewandelt werden. Die p-Werte werden kombiniert um einen finalen Score für eine Gruppe homologer Gene zu erhalten.

Ergebnisse: Ein Benchmark auf einen Datensatz von 74 experimentell bestätigten sRNA Targets, lieferte vielversprechende Resultate. 68.9% aller Vorhersagen verbesserten sich im Vergleich zu IntaRNA. 37.8% der wahr positiven Vorhersagen waren in den oberen 10 ihrer jeweiligen Vorhersage, wobei sich 6 von diesen auf Rang 1 befanden. Die Anwendung von hIntaRNA führte zur Identifikation von vielen putativen noch unbekannten sRNA Targets. Die Ergebnisse deuten auf eine globale Rolle der sRNA Spot42 in der Regulation von Genen des Citratcyklus hin. Des Weiteren wurde die Interaktion des von hIntaRNA vorgeschlagen Targets stm1530 aus *Salmonella* experimentall bestätigt. Das Programm ist in Perl und R implementiert und zurzeit auf Linux Systemen in der Command Line verfügbar. Ein Webserver ist in Planung.

2 Introduction

2.1 Bacterial small RNAs

Ribonucleic acid (RNA) is a key player in every living organism and it is also highly likely that RNA was the carrier of genetic information prior to the advent of deoxyribonucleic acid (DNA) (Gilbert, 1986). RNA can form secondary structures by pairing with other RNA elements in an intra- and/or intermolecular manner. Besides commonly known transfer RNA (tRNA), ribosomal RNA (rRNA), ribozymes and protein coding, messenger RNA (mRNA), recent studies suggest a significant presence of non coding RNAs (ncRNAs) in pro- and eukaryotic organisms, where they can participate in regulative processes, such as controlling translation of mRNA (Eddy, 2001).

Bacterial small RNAs (sRNAs), which are functionally analogous to eukaryotic microRNAs (miRNAs), also belong to this group. They range from 50 – 500 nucleotides (nt) in length, and can in many cases originate from individual genes, controlled by their own promoters and terminators (Waters and Storz, 2009). sRNAs supply an additional module for gene regulation, and are produced faster and "cheaper" than proteins (Shimoni et al., 2007). The *Escherichia coli* MicF RNA was the first reported member of the sRNA group, and was discovered long before the burst in sRNA research (Mizuno et al., 1984). MicF acts on *trans*. This means it is encoded at a locus distinct to its targets. *Cis* acting RNA elements are the counterpart to *trans* acting sRNAs, as they share the same locus with their targets. Examples for this mechanism are Riboswitches, which are often encoded in untranslated regions (UTR) of bacterial mRNAs (Breaker, 2010). Depending on specific signals they can fold into structures that can either enable or disable translation. Antisense RNAs (asRNAs), like Riboswitches, also act on *cis*. They are encoded at the same locus as their targets but on the opposite strand, which clearly leads to perfect complementarity with the target RNA (Georg and Hess, 2011). Conversely, *trans* acting sRNAs often show short and imperfect base pairing with their targets. High complementarity is, in many cases, only given within a seed region of approximately 6-8 base pairs (Gottesmann and Storz, 2010). This seed region is crucial for interaction initiation and is typically located in the 5' regions of the sRNAs (Papenfort et al., 2010). Base pairing with targets usually occurs in the 5' untranslated region of mRNAs, or partially downstream of the start codon.

Several mechanisms for translational control by *trans* acting, base pairing sRNAs are known (Marchfelder and Hess, 2011). Masking of the Shine-Dalgarno (SD) sequence, which is central for the binding of the 30S ribosomal subunit (Shine and Dalgarno 1974), and start codon sequences, inhibits translational initiation. The *Escherichia coli* Spot42 sRNA for instance, targets the SD sequence of the *galK* mRNA and prevents the 30S subunit of the ribosome from associating with its binding site and furthermore leads to a loss of *galK* mRNA stability in experimental over expression of Spot42 (Beisel and Storz, 2011). Vice versa, the DsrA sRNA activates the translation of *rpoS* mRNA by employing an 'anti-antisense' mechanism, which leads to the resolution of an intramolecular base paired structure between the 5'UTR and coding sequence (CDS) of the *rpoS* mRNA, and renders the SD sequence accessible to the 30S ribosomal subunit (Majdalani et al., 1998). sRNAs can also label mRNA targets for degradation by RNases. In this way, the *Salmonella typhimurium* MicC sRNA binds to a region within the CDS of the *ompD* mRNA and flags it for degradation by RNAse-E (Pfeiffer et al., 2009).

Grouping *trans* acting sRNAs into the class of ncRNAs can in some cases be misleading, as they can also serve as templates for translation. This is the case for the SgrS sRNA (Wadler and Vanderpool, 2007) and RNAIII (Benito et al., 2000). In line with this, it has also been suggested that mRNAs may also have, sRNA typical, regulative functions which are yet unknown (Waters and Storz, 2009).

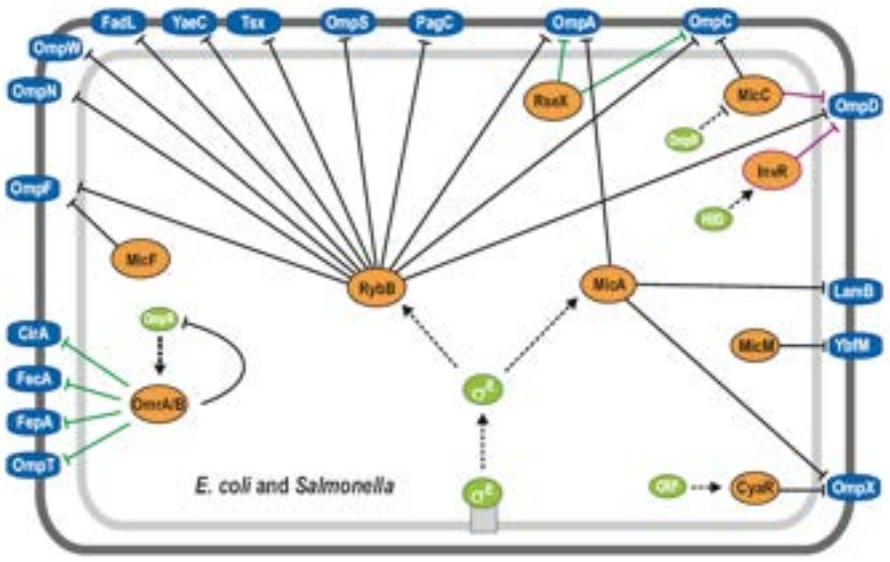


Figure 1: Outer membrane protein (Omp) regulatory circuit in *E. coli* and *Salmonella*. Black lines: homologous sRNAs regulating homologous targets in both organisms, green lines: *E.coli* specific regulation, purple lines: *Salmonella* specific regulation (Corcoran, Papenfort and Vogel, 2011)

Regulation of protein function by sRNAs has also been reported. The 6S sRNA from *E. coli* specifically interacts with the σ_{70} associated RNA polymerase holoenzyme, and thereby down regulates transcription of promoters with weak -35 regions (Wassarmann, 2007).

Despite most examples originating from enteric bacteria, regulation by sRNAs is not limited to this group of microorganisms, and it has been proposed that regulation by sRNAs is spread among all bacteria (Gottesmann, Storz 2010). Organisms such as *Vibrio harveyi* or *Agrobacterium tumefaciens* for example, require sRNAs for quorum sensing and ABC transporter regulation respectively (Tu et al., 2008, Wilms et al., 2011). Especially deep sequencing approaches simplify identification of novel sRNAs in organisms not yet as intensely studied as *E. coli* or *S. typhimurium* (Mitschke et al., 2011a,b).

In many cases, for Gram negative bacteria, the RNA chaperone Hfq is required to allow target recognition and base pairing *in vivo* (Vogel and Luisi, 2011), yet much remains unclear about the exact role of Hfq and about potential other RNA chaperones in the process of mediating sRNA target recognition.

It is apparent that a multitude of sRNAs target more than one mRNA (Fig. 1), simplifying the prediction and identification of new targets for sRNAs by comparing with already known target sites (Modi et al., 2011). This way, the magnitude of potential targets can be narrowed down by incorporation of the preexisting knowledge. Furthermore, sRNAs and their targets are often conserved across certain bacterial species (Fig. 1, 2 and 3), suggesting a comparative approach for identifying new targets. For instance, there is experimentally verified data, which illustrates that the ~80 nt RybB sRNA, which is nearly identical in *E. coli* and *S. typhimurium*, negatively regulates the translation of outer membrane proteins (OMP) in these two organisms (Papenfort et al., 2006, Thompson et al., 2007, Fig. 1).

Direct clinical and biotechnological relevance has been perceived in sRNA research. Reduced pathogenicity in bacterial *hfq* mutants (Romby et al., 2006), controllability of acetate excretion in *E. coli* (Negrete et al., 2011) and accumulation of succinate by RyhB overexpression in *E. coli* (Kang et al., 2012) have been reported.

2.2 Conservation of regulation – RybB, MicA and SgrS

The global outer membrane protein regulator RybB (Johansen et al., 2006, Papenfort et al., 2010) is a suitable example to display how not only the sRNA sequence, but also its target regulation can be conserved across species boundaries.

Figure 2: Alignment of RybB sRNAs from various enteric bacteria. Bold part of the sequence represents the interacting region of the sRNAs. **ST:** *Salmonella typhimurium*, **EC:** *Escherichia coli*, **CR:** *Citrobacter rodentium*, **SD:** *Shigella dysenteriae*, **SB:** *Salmonella bongori,* **KP:** *Klebsiella pneumoniae*, **KO:** *Klebsiella oxytoca*, **CK:** *Citrobacter koseri,* **ES:** *Enterobacter sakazakii,* **YP:** *Yersinia pestis*, **YE:** *Yersinia enterocolitoca*, **SP:** *Serratia proteamaculans*, **SM:** *Serratia marcescens*, **PL:** *Photorhabdus luminescens*, **SG:** *Sodalis glossinidius* , **EW:** *Erwinia carotovora.* (Supplementary figure S3 from Papenfort et al. 2010)

Both in *E. coli* and *Salmonella*, RybB is induced upon heat stress and regulates the same group of targets. Figure 2 shows the extensive sequence similarity of RybB homologs across many enteric bacteria. An alignment of the 5' UTRs of RybB target mRNAs (Fig. 3) clearly outlines that not only the RybB sequences are conserved throughout enteric bacteria but also the sequences of their targets. In this case the binding sites (bold letters) nearly always show perfect conservation, allowing the conclusion that the sRNA as also the target regulation is conserved, which has also been proven experimentally for *E. coli* and *Salmonella*. A study on the MicA sRNA revealed similar results regarding conservation of homologous sRNAs and their targets (Udekwu et al., 2005).

Figure 3: Alignments of RybB target mRNAs. The RybB seed region is printed in bold. **ST:** *Salmonella typhimurium,* **CK:** *Citrobacter koseri*, **EB:** *Enterobacter sp. 638*, **EC:** *Escherichia coli,* **E*S:*** *Enterobacter sakazakii*, **YP:** *Yersinia pestis*, **EW:** *Erwinia carotovora,* **PL:** *Photorhabdus luminescens* (Supplementary Figure S8 from Papenfort et al., 2010)

Yet, not in every case a conservation of target regulation, as in the case of RybB, can be observed. A recent study on SgrS (Rice and Vanderpool, 2011) and its targets in divergent enteric bacteria revealed that target conservation is only given in some cases. *ptsG* regulation by SgrS, for instance, seems to be conserved among all tested, SgrS containing bacteria (Horler and Vanderpool, 2009), while *manX* regulation by SgrS is not.

2.3 Prediction of sRNA targets

Prediction of sRNA targets in bacteria is a choke point in sRNA research, and has proven to be a very challenging task. Imperfect and short base pairing between sRNAs and their targets complicate the matter (Pichon and Felden, 2008). Furthermore, knowledge of sRNA-mRNA interactions is still comparatively limited compared to miRNA-mRNA interactions, and current computational solutions are most likely biased by the limited magnitude of examples available to abstract from. The problem has been addressed by several algorithmic approaches, which tackle the problem from different angles.

Generally the prediction types can be assigned to four different groups, of which all are initially dependent on finding complementary stretches between the sRNAs and their target sequences (Backofen and Hess, 2010).

Firstly, methods such as the basic local alignment search tool (BLAST) (Altschul et al., 1990), TargetRNA (Tjaden et al., 2006) or GUUGle which, unlike BLAST, also allows G-U-base pairs (Gerlach and Giegerich, 2006), are a starting point and solely depend on sequence complementarity.

Secondly, assignment of thermodynamic energy scores, like in RNA secondary structure prediction, to sRNA-target duplexes is an addition to the first approach. This yields a result with increased biological significance and also allows consideration of temperature, which is central for structure and stability of RNA. However, intra-molecular structures are not addressed, which is a major negligence of a biologically important contribution in sRNA-mRNA interactions (Peer and Margalit, 2011, Richter and Backofen 2011). Implementations of this approach are RNAplex (Tafer and Hofacker, 2008) and RNAhybrid (Rehmsmeier et al., 2004).

Concatenating the target RNA and sRNA sequences with an interspacing linker, and predicting the joint secondary structure of both RNAs, is another way of dealing with the problem of sRNA target prediction. In essence, this is similar to general RNA folding algorithms, such as MFOLD (Zuker, 1994) or RNAfold (Hofacker et al., 1994) and has, to name one, been realized in the sRNATarget program (Cao et al., 2009). Yet, the detriments to these concatenation approaches are the same as those intrinsic to the folding algorithms. Certain structures, such as pseudoknots (Fig. 4), which have in vivo relevance, cannot be predicted. This is especially problematic since many interactions are located in hairpin loops of one or both RNAs.

Finally, consideration of accessibility in the sRNA-mRNA interaction has yielded very competitive results (Mückstein et al., 2006, Busch et al., 2008, Eggenhofer et al., 2011, Tafer et al., 2011). Accessibility means that the interaction sites both in the target and the sRNA should not show involvement in intramolecular base pairing if intermolecular base pairing is to occur. If they do show intramolecular base pairing at the putative interaction site, the accessibility of the interaction site is reduced and therefore the interaction is penalized, consequently leading to a worse (i.e. less negative) energy score. RNApredator (Eggenhofer et al., 2011) is a recent target prediction webserver also taking gene ontology

enrichment (Ashburner et al., 2000) into account. Current data indicates, that sRNAs are often part of regulatory networks (Modi et al., 2011), illuminating that ontology enrichment should be a standardized option in every future target prediction software. If significant enrichment can be detected, this allows instantaneous characterization of regulatory networks the sRNA may participate in, and simplifies the evaluation of results.

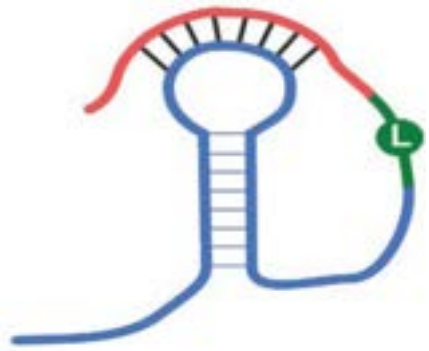


Figure 4: Representation of a pseudoknot between two merged, distinct RNAs (blue and red) with the inter-spacing linker region (L, green) in the concatenation approach. (Figure 1B from Backofen and Hess 2010)

Due to a lack of specificity, leading to high numbers of false positive hits in all available software, target prediction is still not satisfactory and often dismissed by biologists at this time. Therefore, improvement of present algorithms is a pressing concern.

2.4 IntaRNA - **int**era**c**ting **RNA**s

IntaRNA (Busch et al., 2008), which is the underlying sRNA target prediction algorithm in this project, is one of the previously mentioned methods that consider the accessibility of interaction sites. The program calculates the final energy of an interaction by minimizing the sum of the single energy scores. These energies are, (1) the energies required to unfold double stranded stretches in the interacting regions of the sRNA and its target (blue and green in Equation 1), which are positive and therefore penalize the interaction, and (2) the hybridization energy (red in Equation 1) of the interacting region, which is negative and promotes the strength of an interaction. The RNAhybrid energy model is employed in order to calculate the hybridization energy. The accessibilities (i.e. ED-values in Equation 1) can be obtained by using RNAPLFOLD (Bernhart et al., 2006). If the sum of Equation 1 yields a positive result, then this result is dismissed for the final output. IntaRNA also establishes the application of a seed region, which has been shown to be of biological relevance

(Papenfort et al., 2010). Both, length and amount of unpaired bases allowed in the seed region are user definable. Also, the seed is not restricted to a certain location within the sRNA or its target. A disadvantage of IntaRNA is that it cannot predict interactions such as double kissing complexes which form between the OxyS sRNA and the *fhlA* mRNA (Argaman and Altuvia, 2000), thus eliminating these kind of interactions from the scope of predictability.

$$E^{IntaRNA}(i, i', k, k') = E^{hybrid}(i, i', k, k') + ED^{mRNA}(i, i') + ED^{sRNA}(k, k') \quad , \quad (1)$$

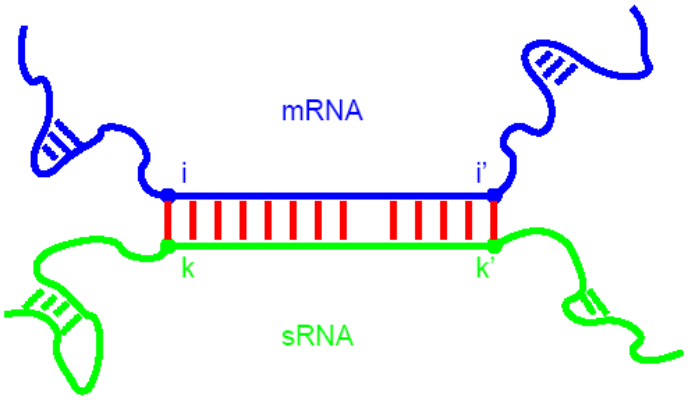


Figure 5: Representation of an mRNA-sRNA interaction between i,i' (bases in the sRNA) and k, k'(bases in the mRNA) (Image supplied by Andreas S. Richter).

A complete approach of IntaRNA requires $O(n^2m^2)$ and $O(nm)$ of time and space respectively. However, a more practical, heuristic approach reduces these complexities to $O(nm')$ for time and $O(nm)$ for space, m' being $\max\{m,L^3\}$, where L is the "size of the sequence window in which both mRNA and sRNA are folded". IntaRNA has proven to be very competitive compared to other sRNA target prediction software in accuracy and complexity, which motivates its use as base line algorithm for the comparative prediction approach presented in this work.

2.5 Comparative approaches in Bioinformatics

In bioinformatics, comparative approaches have generally proven to be very helpful. This is intrinsically obvious, as every group of organisms is thought to have evolved from a common ancestor (Darwin, 1859). In line with this it seems reasonable to assume that organisms, which have diverged from each other, still retain common properties.

Global and local alignments (Needleman and Wunsch, 1970, Smith and Waterman 1981) compare similarity in amino acid and nucleotide sequences and led to approaches as complex as BLAST (Altschul et al., 1990), which is extensively employed in *de novo* characterization of freshly acquired genomic sequences, and is a tool which is indispensable in modern molecular biology. *De novo* genome annotation also follows a comparative methodology (Aziz et al., 2008). Computational prediction of novel sRNAs in bacteria (Voß et al., 2009) and phylogenetic analyses follow this practice too.

A comparative approach also seems viable to solve the eminent problem of lacking reliability in sRNA target predictions, however only if conserved interspecies regulation is present, which is the case for many sRNAs (Corcoran, Papenfort and Vogel, 2011, Fig. 1), but not for all (Richter and Backofen, 2011). This is the focus of the homology IntaRNA (hIntaRNA) program presented in this project.

hIntaRNA is a homology based approach employing the IntaRNA algorithm to increase the reliability of predictions for *trans* acting small RNA targets in prokaryotes. The basic idea of predicting regulatory RNA targets in a comparative manner has been proposed for eukaryotic microRNAs (Rehmsmeier et al., 2004), and prokaryotic sRNAs (Tafer et al., 2011). As already mentioned, in many cases a conservation of targets for conserved sRNAs can be observed across single species boundaries (Figs. 1, 2 and 3). Consequently, under the assumption that regulation is conserved, overlapping target predictions for distinct organisms yield stronger evidence of correct functional prediction.

2.6 The comparative approach in RNAhybrid

Rehmsmeier et al., (2004) introduced a comparative approach for predicting mRNA-miRNA duplexes for orthologous target sequences. They state that the probabilities of two separate predictions occurring by chance (p-values) can be combined to a joint probability by the following equation:

$$P[Z_1 \geq e_1, Z_2 \geq e_2] = (\max\{P[Z \geq e_1], P[Z \geq e_2]\})^2 , \quad (2)$$

with $P[Z \geq e]$ being the probability that the observed energy Z is greater or equal to e. This means that the bigger one of the two p-values is selected and squared.

This equation can be generalized into a form which allows a variable number of organisms as follows:

$$P[Z_1 \geq e_1, \ldots, Z_k \geq e_k] = (\max\{P[Z \geq e_1], \ldots, P[Z \geq e_k]\})^k \text{ , (3)}$$

However, Equation 3 assumes that orthologous targets are statistically independent, which is incorrect when taking the biological background into account. As previously described, the evolutionary idea is based in the concept of species being descended from common ancestors, making it obvious that a certain degree of dependence must still be present. This leads to the need for a measure of dependence between the orthologous target sequences. Hence, Rehmsmeier et al. introduced k-effective (k_{eff}), which lies between 1 and the number of orthologous targets (k):

$$1 \leq k_{eff} \leq k \text{ , (4)}$$

The higher the dependence between the target sequences, the smaller k_{eff} will be. Accordingly the final joint p-value calculated from Equation 3, with the according $k_{eff,}$ will be greater, the higher the dependence between the target sequences is. Assuming an analysis with two identical organisms there is no gain of information, by inclusion of the second organism, consequently leading to a k_{eff} of 1.

k_{eff} is obtained by shuffling the miRNA, predicting interactions with orthologous targets and estimating extreme value distribution parameters to calculate p-values for the predicted energy values. Next, the p-values are joined, employing Equation 3 with k' values instead of k, where k' lies between 1 and k. k_{eff} is the k' which yields the straightest line in the empirical cumulative density function of the joint p-values. Straightness is evaluated using the least squared error measure method.

The approach for combining p-values, which is presented in this work, is strongly based on the methodology presented by Rehmsmeier et al., but has increased sophistication, and shall be presented in the methods part.

2.7 The comparative approach in RNAplex

Compared to the former version of RNAplex (Tafer and Hofacker, 2008), the new version of RNAplex (Tafer et al., 2011) incorporates accessibility calculation into the prediction model. RNAplex can also make alignment based predictions by application of the RNAalifold concept (Bernhart et al., 2008). This model takes evolutionary conservation into consideration as base pairs are only enforced if they can be formed by most sequences in the alignment. Furthermore, a comparative approach is introduced in order to reduce the abundance of false positives. The authors developed their comparative model following the argumentation that target site conservation between organisms sharing homologous sRNAs, and compensatory mutations between target RNA and sRNA – with respect to other organisms' sRNA and target sequences –, are an indicator of increased prediction reliability (Chen et al., 2007). The procedure in the comparative method of RNAplex consists of five steps, starting with aligning the putative target sequences with clustalw (Larkin et al., 2007) and sorting them by similarity. In the following, RNAplex predictions for all sequences are carried out, while the three best predictions for each target sequence are stored. Then these predictions are employed to recursively find the best set of target sites. In order to achieve this, sequence similarity and interaction strengths are used and the final group of target sequences is assessed via backtracking. The arisen cluster of target sites is realigned and the RNAplex alignment version is applied.

Generally RNAplex does not introduce groundbreaking innovations to sRNA target prediction. The authors stress that the strongest gain is on the level of runtime. They also state that genome wide target predictions with other available tools, such as IntaRNA (Busch et al., 2008), are impractical. This observation must be viewed critically especially from an applied angle, as a time difference of a few hours is not central if the quality of predictions demonstrably increases. The comparative approach seems overly restrictive and complicated. While the number of false positives can most certainly be anticipated to decrease, the number of false negatives will most likely increase alongside.

3 Methods

3.1 The concept behind hIntaRNA

The basic concept behind hIntaRNA is that overlapping IntaRNA predictions for homologous targets in distinct organisms yield increased evidence for predictions being of functional relevance. However, IntaRNA predictions yield energy scores that are, due to varying GC-content and dinucleotide frequency, initially not comparable for different organisms (Yakovchuk et al., 2006). A statistical model, based on probability values (p-values), seems promising to combine evidence from various sources. Naively one may think that the p-values of predictions of homologous targets in different organisms could be simply multiplied to assess the joint probability, just as 1/6 can be raised to the power of three to obtain the joint probability of rolling the same number with the dice three times in a row. This would, at least in some cases, certainly give rise to satisfying results. Clearly, a more generally applicable model is desirable. A major disadvantage of the naive method is that it does not take the phylogenetic distance of the organisms into account. If, for instance a comparative prediction were to be made in the naive fashion for three organisms and two of these were of the *Escherichia,* and one of the *Salmonella* genus, it is obvious that without weighting, no statistically and biologically sound result can be attained. To this end, the individual p-values are phylogenetically weighted in hIntaRNA, while the result of this weighting is a phylogenetically weighted mean of all the p-values for one group of homologous targets. However, as this is a mean it does not yet represent the result of a multiplication of probabilities. Consequently this mean must be powered to a certain degree. Here, one also may firstly think that powering the p-value by the amount of participating organisms is acceptable, just as in the example with the dice. This is basically the same as multiplying the individual p-values just that the weighted mean p-value is powered instead. This would be correct in the case of all participating events being entirely independent. Under inclusion of the biological background however, it is clear that there is no complete independence, as it is assumed that all organisms are descended from a common ancestor (Darwin, 1859). Hence, the weighted mean cannot, at least not in every case, be powered by the amount of organisms in the analysis. The degree of dependence can be assessed by

employing a function that describes the products of n uniform distributed random variables between [0,1] which are in this case the p-values (Bailey and Gribskov, 1998).

3.2 P-values

Probability values (p-values) are frequently used in statistical analyses, and describe the probability of an event occurring by chance. Furthermore, p-values are random uniform variables between [0,1] (Murdoch et al., 2008). Hence, p-values can acquire values in the range of 0 and 1 and the probability of an observed event occurring by chance is smaller, the smaller the according p-value is. Here the event is an energy score predicted by IntaRNA and the p-value sheds light on the probability of a particular score being predicted by chance or the probability to get a score better or equal to the score viewed. In order to assess the likelihood of a coincidental event an appropriate background model and a null hypothesis (H_0) need to be established as a basis for the evaluation of chance. In sequence analysis a background model can be generated by shuffling the analyzed sequence or by fitting a model to a dataset of several predictions. If the resulting p-value is smaller or equal to a previously defined significance level, the result is statistically significant with respect to H_0. P-values are also important to make the significance of results from different analyses comparable. The raw data in this project are IntaRNA energy scores, which are useless if evaluating predictions in an interspecies approach as GC-content and dinucleotide abundance, which play a central role in defining the energy of an interaction, can vary between organisms (Yakovchuk et al., 2006). Finally it must be stressed that the p-value is the result of a statistical test, and not the result of a biological experiment. Hence, a good p-value does not necessarily infer biological significance, but rather a biologically interesting statistical result hinting at possible relevance (Mitrophanov and Borodovsky, 2006).

3.3 Generalized and Gumbel extreme value distributions

Extreme value distributions (EVD) are a class of distributions from probability theory. They can be applied to the modeling of the likelihood of extreme events, such as RNA folding energies (Rehmsmeier et al., 2004) or flooding events (Gumbel, 2004). These incidents can be assumed to be extreme value distributed, as low folding energies and floodings are

considered rare and extreme. Consequently, an EVD is a better model for these kinds of occurrences, compared to a normal distribution as the density of extreme events is small. A general extreme value distribution (GEV) is defined by three central parameters location (μ), scale (σ) and shape (ε). μ shifts the location of the maximum, σ changes the width of the function and ε alters the behavior of the tail. The bigger ε is, the stronger the tail converges towards zero.

The cumulative distribution function is defined as:

$$F(x;\mu,\sigma,\varepsilon) = \exp\{-[1+\varepsilon*(\frac{x-\mu}{\sigma})]^{\frac{-1}{\varepsilon}}\} \quad , \quad (5)$$

while the resulting density function is:

$$f(x;\ \mu,\sigma,\varepsilon) = \frac{1}{\sigma}[1+\varepsilon*(\frac{x-\mu}{\sigma})]^{(\frac{-1}{\varepsilon})-1}\exp\{-\left[1+\varepsilon\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{-1}{\varepsilon}}\} \quad , \quad (6)$$

A specialized form of the GEV distribution is the Gumbel extreme value distribution. It differs from the GEV distribution in the detail that the shape parameter (ε) is always zero.

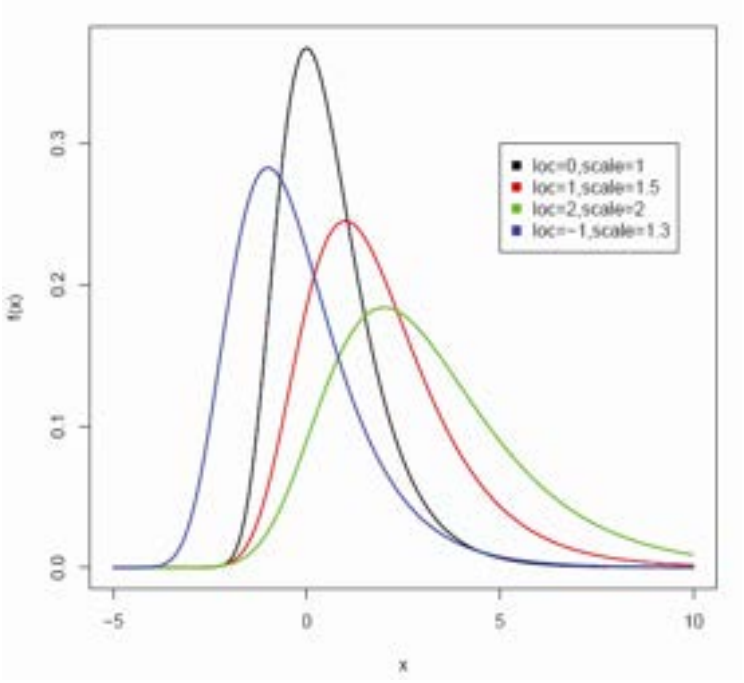


Figure 6: Gumbel distributions with a variety of different location and scale parameters.

Figure 6 shows how extreme value distributions vary under the influence of different parameters. The location parameter clearly shifts the EVD towards the positive region when enlarged, while the width of the EVD increases with bigger scale values. In this project, GEV distributions and Gumbel distributions were used to model density distributions of IntaRNA energy scores and subsequently derive p-values from these.

3.4 Transformation of IntaRNA energies to p-values

Three approaches were tested to transform IntaRNA energies to p-values. These are the empiric approach on shuffled data, the fitting approach on shuffled data and the fitting approach on unshuffled data. The first, empiric, approach created a background model by shuffling the target sequences 10 times, while retaining the dinucleotide frequencies of the shuffled sequences. Shuffling was performed with the shuffle program, which is part of Sean Eddy's squid package (http://selab.janelia.org/, accessed at 01/09/2011). IntaRNA predictions were carried out on the shuffled and unshuffled data for all sRNAs from the benchmarking dataset (see paragraph 3.7) except for InvR. Furthermore, no predictions were made for *Pectobacterium carotovorum*, *Serratia proteamaculans* and *Yersinia pestis* and the MicA *Escherichia fergusonii* prediction was also not considered. Predictions for OmrA were only performed for *E. coli*, *E. fergusonii* and *Salmonella*, yielding a total data set of 92 predictions. The energies predicted for the shuffled data served as basis for calculating p-values for the unshuffled data as described in Equation 7.

The p-value, P(Eu), of a specific IntaRNA energy score Eu is defined by:

$$P(Eu) = \frac{E \leq E_u}{E} \quad , \quad (7)$$

In this equation, E is the total amount of shuffled IntaRNA energy scores. The equation shows the amount of predicted interactions on the shuffled data with an energy score better or equal to each individual, observed energy in the unshuffled data, divided by the total number of energies of the shuffled data.

Previous investigations of the statistical distributions of interaction energies (Rehmsmeier et al., 2004, Schulz 2009) revealed that predicted duplex energies follow extreme value statistics. In line with this, the second approach for p-value generation is fitting of a general

extreme value distribution to the shuffled IntaRNA energy scores. When fitted to the data, p-values can be directly obtained from the extreme value distribution's cumulative distribution function 5.

The third approach is strongly related to the first one, as it also uses extreme value statistics for p-value generation. This method solely differs in the fact, that it does not employ the shuffled data, but rather fits the general extreme value distribution to the unshuffled (i.e. genome wide IntaRNA prediction) data. The whole genome prediction on unshuffled data can also be employed as background model, as most predictions have an arbitrary character and no *in vivo* relevance, as is the case for the shuffled dataset. Also, IntaRNA predicts interactions for most sequences, which supplies a satisfying statistical magnitude (see supplementary Figure 1 or paragraph 3.6 for closer explanation).

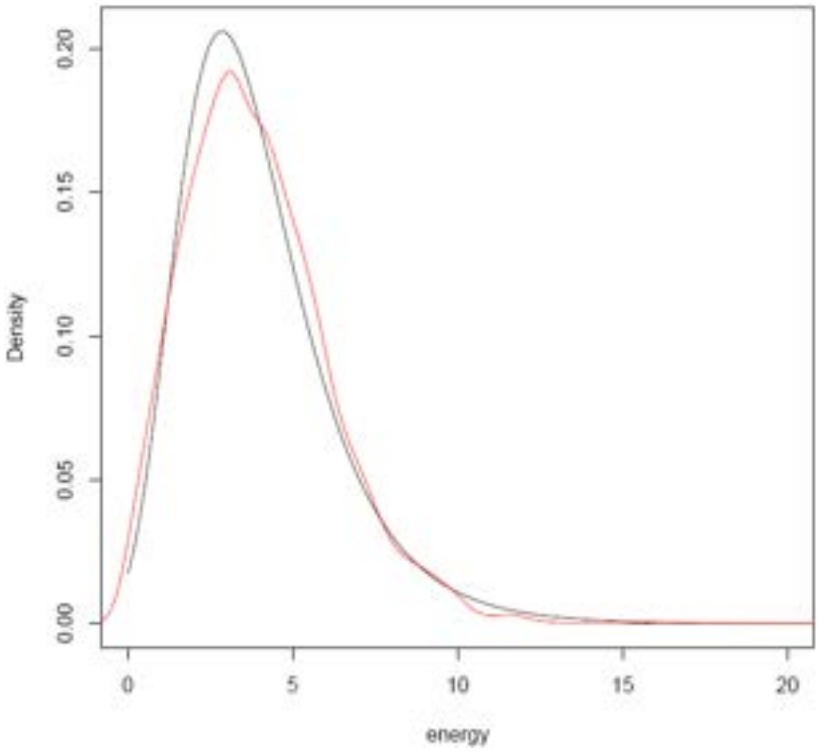


Figure 7: Red: density of IntaRNA energies, black: general extreme value curve fitted to the data. Absolute (i.e. positive) energy values were used for plotting reasons.

Figure 7 graphically illustrates how fitting an extreme value distribution to unshuffled data looks. The red curve shows the densities of individual energies of the unshuffled data, while the black curve shows the fitted density function.

3.5 Method of least squares

In general, the method is used to fit parameters of a model function to obtain the best fit to a dataset of observed values. Here, the method of least squares is a procedure with which an estimation of the deviance between two functions is calculated. Smaller sums indicate higher similarity between compared functions.

Z being the amount of the analyzed data points, F being the calculated error and y_{mi} and y_{ni} being the respective data points of two functions.

$$F = \sum_{i=1}^{z} (y_{mi} - y_{ni})^2 \quad , \quad (8)$$

In this project the method of least squares was employed to assess the similarity of the empiric distribution function and the ideal distribution function 11, in order to retrieve k_{eff}. The evaluation of uniformity of the distribution of the initial p-values was also carried out with this method, using p-values of 92 individual single organism IntaRNA predictions (see paragraph 3.4 for details).

3.6 Weighting and multiplication of p-values according to phylogeny, and retrieval of k_{eff} using the Bailey & Gribskov function for products of independent uniformly distributed random variables

Statistically speaking, the joint probability, Z_n, for n events occurring is the product of the probabilities of the separate events, occurring individually under the assumption that all events are statistically independent (see Equation 10). This can be applied to the p-values generated from the IntaRNA energies, by multiplying the p-values of predictions of homologous targets in distinct organisms. However, as not all the organisms in an analysis are necessarily equidistant from each other and the events are not totally independent, measures of distance and dependence are required. An alignment (emma) (Rice et al., 2000) of genomic 16s-linker-23s regions (Fig. 8) from each participating organism, serves as basis for the subsequent calculation of a distance matrix, using the Jukes-Cantor model (Jukes and Cantor, 1969).

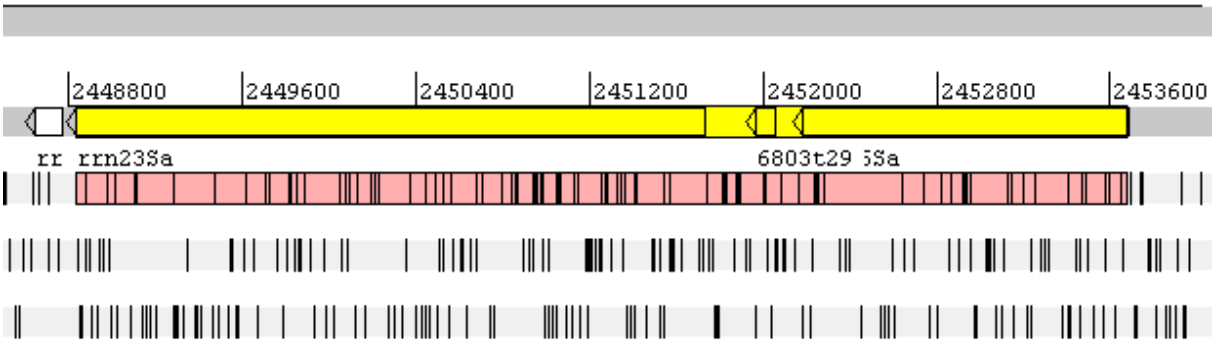


Figure 8: Yellow marked genomic 16s-linker-23s region of *Synechocystis sp.* PCC 6803 in Artemis genome browser (Release 13.2.0) (Rutherford et al., 2000).

The individual p-values, P, are exponentiated with their contribution to the distance matrix, $M[P_n]$, divided by the total size of the distance matrix, M[all], and then multiplied with each other. The individual contribution of an organism to the distance matrix is the sum of distances of one organism to each of the other organisms, while the total size of the distance matrix is the sum of all its entries. This yields a phylogenetically weighted mean of the combined p-values (see paragraph 7.1 for a specific example).

$$P_{all} = P_1^{M[P1]/M[all]} * \ldots * P_n^{M[Pn]/M[all]} \quad , \quad (9)$$

Bailey and Gribskov (Bailey and Gribskov, 1998) introduced a function (11) which describes the cumulative distribution of probabilities of the products of n independent, uniformly distributed random variables between 0 and 1 being smaller or equal to their own value.

Let P_i be a uniform in the interval [0,1] distributed random variable. Then, the product of n of such random variables is defined by:

$$Z_n = \prod_{i=1}^{n} P_i \quad , \quad (10)$$

under the assumption that all P_i are statistically independent.

The probability that Z_n has an observed value ≤ p, $P(Z_n \leq p)$, is given by:

$$F_n(p) = p \sum_{i=0}^{n-1} \frac{(-\ln p)^i}{i!} \quad , \quad (11)$$

In this project, the p-values satisfy the conditions given above and consequently this function is employed in order to find the degree of dependence, k_{eff}, between the separate entities (i.e. organisms) in the hIntaRNA analysis. As function 11 describes how the distributions resulting from products of individual independent p-values should ideally look, it can be used to compare the actual data to. k_{eff} is retrieved in a manner very similar to the previously described method employed in the comparative approach in RNAhybrid (see paragraph 2.6) (Rehmsmeier et al., 2004). The product of the exponentiated p-values is exponentiated with the potential k_{eff} (i.e. k'),

$$P_{k'} = P_{all}{}^{k'} \quad , \quad (12)$$

and the resulting values are used to generate function values for an empiric function, which is then compared to function 11. The empiric probability, $P(P_u)$, of a specific p-value ,P_u , is defined by:

$$P(P_u) = \frac{P \leq P_u}{P} \quad , \quad (13)$$

The empiric function values are compared to the affiliated function values from function 11, and the errors are evaluated with the method of smallest squares.

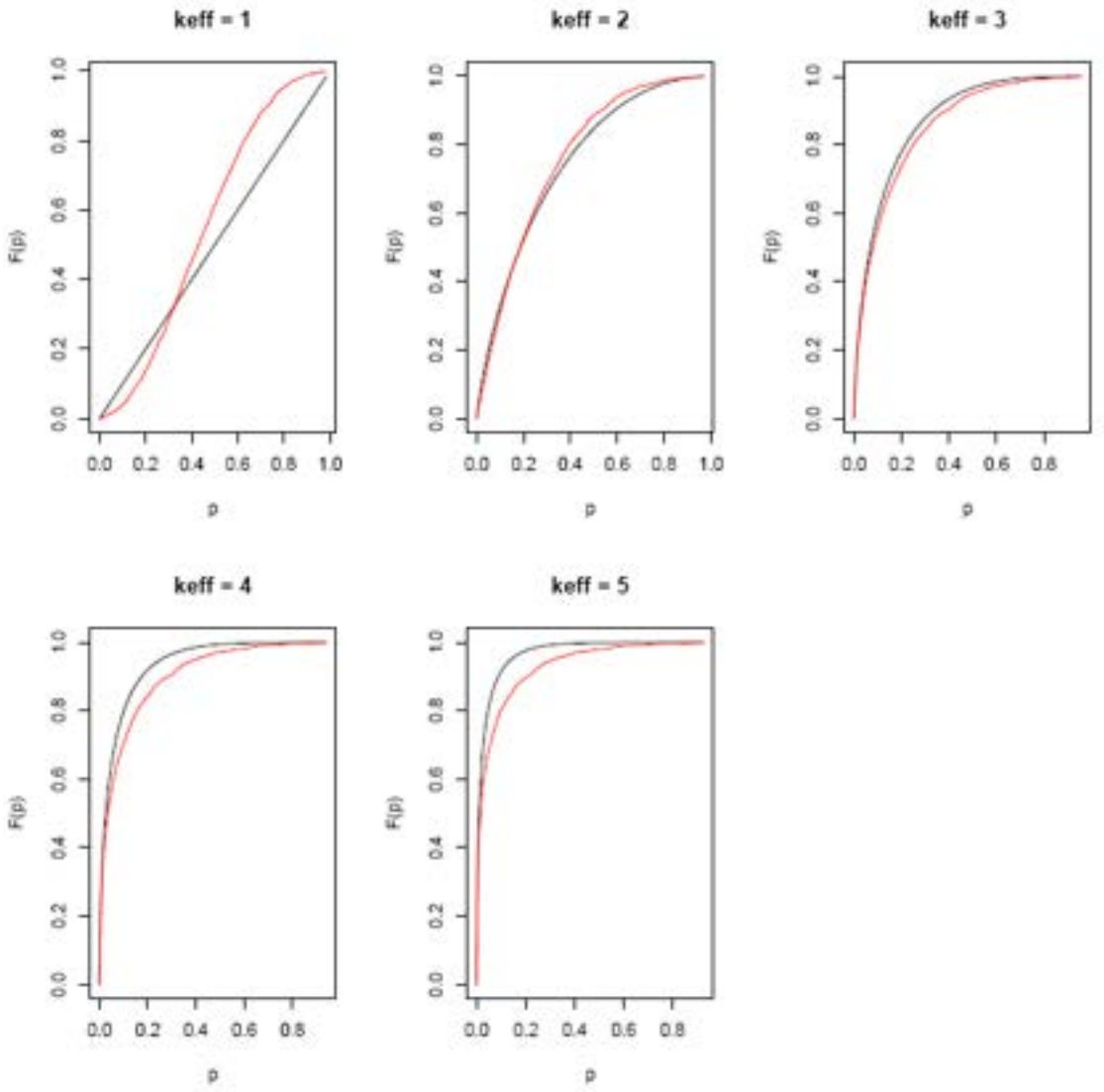


Figure 9: Comparison of empiric (red) and ideal (black) functions for the likelihood of products of uniformly distributed , random variables between [0,1].

The *k'* yielding the smallest error in the comparison of function values is the k_{eff} which is consequently used for the calculation of the final p-values. The data used to assess k_{eff} for each cluster of genes, are the joint p-values from every possible cluster in the range of n down to two, while k_{eff} is assessed for each cluster individually. Figure 9 illustrates the comparison of distribution function (11) (black) to the empiric cumulative distribution function (red) visually. In this case the correct k_{eff} is $k_{eff} = 2$. Similarity of function values can also be visualized by plotting the function values against each other (Fig. 10). The plot yielding the straightest line indicates the two functions with the most similar function values.

Not every gene has homologs in each of the other organisms in the analysis. If, for instance, the analysis comprises of n organisms, many genes will be present in all n organisms, however those only present in n-1 or less organisms complicate the matter, as there are more than n different combinations of these clusters. The previously described analysis must be carried out for each possible cluster < n.

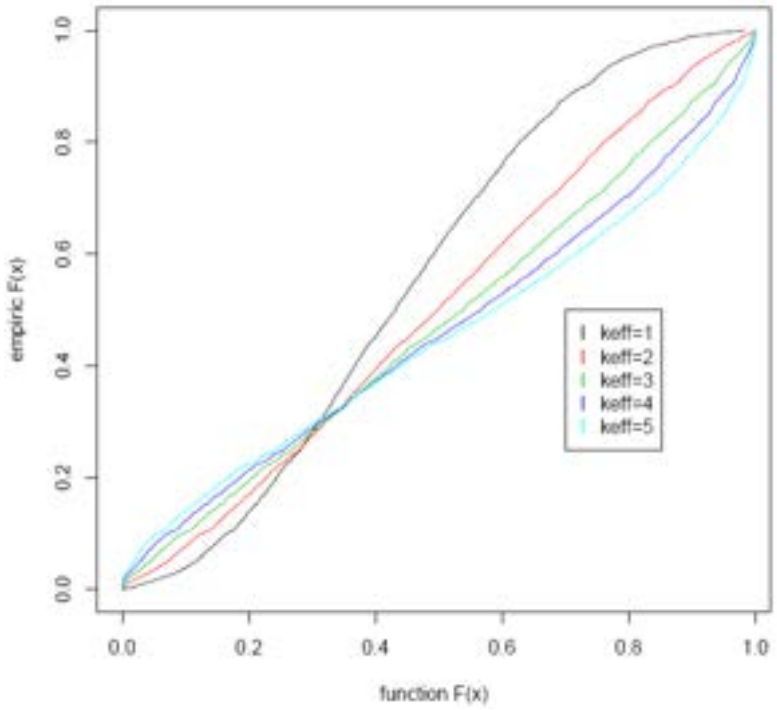


Figure 10: Empiric cumulative density function values plotted against function 11 values.

A problem resulting from these smaller clusters is that prokaryotic genomes are frequently comprised of approximately 4000 genes, making it intrinsically obvious that clusters will lack the statistical magnitude needed, the more organisms participate in an analysis. The problem of insufficient statistical power in smaller clusters can be elegantly solved by extracting subclusters from the bigger clusters. A cluster of five, for example, contains the information of all possible clusters of four and smaller. It only needs to be stripped of the genes not needed in order to create the smaller subclusters or pseudo clusters. These pseudo clusters are not regarded for the final output, but are crucial to achieve a statistically significant population. Supplementary Figure 1, shows that a sample size of 1000 random variables between [0,1] results in an acceptable uniform distribution. This minimum level sample size is always met and mostly significantly overstepped. The procedure of creating pseudo clusters is aided by the Math::Combinatorics Perl module.

3.7 Benchmarking dataset

The Benchmark was executed on 74 experimentally verified 5'UTR targets. However, not only the 5'UTR, in this case assumed to be the stretch 200 nt upstream of the start codon, was used, but also 100 nt downstream of the 5'UTR were parsed, leading to potential target sequences of 300 nt per gene. Of the verified targets, 43 are *E. coli* (NC_000913), 29 are *Salmonella* (NC_003197) and two are *Synechocystis sp.* PCC 6803 targets. A total of 34 different sRNAs were tested (Supplementary table 1). The organism counts varied between different analyses. ArcZ, CyaR, FnrS, GcvB, GlmZ, MicA, OmrB, RprA, RybB and Spot42 were analyzed with eight input organisms (NC_000913, NC_011740, NC_003197, NC_009792, NC_013716, NC_012917, NC_009832, NC_003143). ChiX, DsrA, MicC, MicF, OxyS, RyhB and SgrS were each analyzed with five organisms (NC_000913, NC_011740, NC_003197, NC_009792, NC_013716) as was SyR1 (NC_000911, NC_010296, NC_010546, NC_011729, NC_013161). OmrA was analyzed with three organisms (NC_000913, NC_011740, NC_003197) and InvR with two organisms (NC_003197, NC_003198).

The results of the hIntaRNA analyses were evaluated individually and were also compared to single IntaRNA predictions as a measure of improvement.

Organism name and strain	RefSeq ID
Escherichia coli K-12 MG1655	NC_000913
Salmonella enterica CT18	NC_003198
Salmonella typhimurium LT2	NC_003197
Synechocystis sp. PCC 6803	NC_000911
Escherichia fergusonii ATCC 35469	NC_011740
Citrobacter koseri ATCC BAA-895	NC_009792
Citrobacter rodentium ICC168	NC_013716
Pectobacterium carotovorum PC1	NC_012917
Serratia proteamaculans 568	NC_009832
Yersinia pestis CO92	NC_003143
Microcystis aeruginosa NIES-843	NC_010296
Cyanothece sp. ATCC 51142	NC_010546
Cyanothece sp. PCC 7424	NC_011729
Cyanothece sp. PCC 8802	NC_013161

Table 1: List of organisms employed for the benchmarking dataset with organism name/strain and RefSeq ID.

3.8 Datasets in analyses of novel sRNAs

Several analyses on sRNAs not contained in the Benchmarking dataset were executed. The *Agrobacterium tumefaciens* AbcR1 and AbcR2 sRNAs (organisms: NC_003062, NC_010994, NC_003047), the *Vibrio harveyi* Qrr1 sRNA (organisms: NC_002505, NC_009783, NC_004603, NC_011753, NC_004459) and the *Anabena sp.* PCC 7120 NsiR1 and NsiR3 (organisms: NC_007413, NC_014248, NC_010628, NC_003272) sRNAs were analyzed.

Organism name and strain	RefSeq ID
Agrobacterium tumefaciens C58	NC_003062
Rhizobium etli CIAT 652	NC_010994
Sinorhizobium meliloti 1021	NC_003047
Vibrio cholaerae N16961	NC_002505
Vibrio harveyi ATCC BAA-1116	NC_009783
Vibrio parahaemolyicus RIMD 2210633	NC_004603
Vibrio splendidus LGP32	NC_011753
Vibrio vulnificus CMCP6	NC_004459
Acaryochloris marina MBIC11017	NC_009925
Gleobacter violaceus PCC 7421	NC_005125
Nostoc sp. PCC 7120	NC_003272

Table 2: Organisms employed in further analyses of sRNAs with organism name/strain and RefSeq ID.

3.9 DAVID functional annotation

Functional annotation for gene lists, which are acquired as results of hIntaRNA predictions, can be analyzed using the **d**atabase for **a**nnotation, **v**isualization and **i**ntegrated **d**iscovery (DAVID) (Huang et al., 2009). Due to the fact that sRNAs often act as global regulators in regulatory networks, functional enrichment of top hits (i.e. top 50) is a method to extract information about the regulatory processes an sRNA may participate in. In order to be enriched, a subset of genes must be overrepresented compared to a background. Concretely for a group of genes to be enriched, this means that the percentage of genes of a certain functional group must be significantly (with respect to, for example Fisher's exact test) higher in the subset compared to the percentage of this class of genes represented in the background. The background in this project are all genes derived from a single organism for which IntaRNA interactions can be predicted and homologous genes in other organism are present. The gene identifiers employed in our case are Entrez Gene IDs. Yet, DAVID allows

the use of several other common identifiers. An important value in the enrichment analysis is the enrichment score of an enriched cluster which is defined as "the geometric mean of all the enrichment P-values of each annotation term" in the cluster (citation from Huang et al. 2009), while the p-values are calculated with Fisher's exact test. Huang et al., state that enrichment scores ≥ 1.3 are of increased significance. However, they also stress that gene clusters with lower enrichment scores should not necessarily be dismissed. All enrichments were executed with standard parameters.

3.10 Quality clipping before p-value combination

Preprocessing the data in order to improve predictions seems promising. Preprocessing in this case means that single organism prediction p-values that are bigger or equal to 0.8 are "clipped" from the initial prediction. The idea is to eliminate obviously arbitrary predictions from clusters which may show conservation of sRNA regulation only for some of the genes in the cluster. In order to assess the potential of this preprocessing method, hIntaRNA clipped predictions were taken out for the whole benchmarking dataset and rankings were evaluated in comparison to hIntaRNA predictions without clipping.

3.11 Implementation overview

In the following, the implementation of hIntaRNA shall be explained. Detailed methodology is explained in the preceding paragraphs.

hIntaRNA has a modular character, and is split into 14 individual Perl scripts. All these scripts are successively executed by a master script (homology_intaRNA.pl). The input arguments are firstly a fasta formatted file containing the sRNA homologs, secondly the number of nucleotides upstream and downstream of only the start codon or upstream of the start codon and downstream of the stop codon respectively, depending on the specification of the region for which the prediction shall be made (i.e. 5'UTR or CDS).

Example	Abstraction
perl	[script language]
homology_intaRNA.pl	[program]
0680a.fasta	[sRNA homologs]
200	[upstream start codon]
50	[downstream start codon]
16S-linker-23S.fasta	[molecular chronometer]
5utr	[mRNA region]
NC_008686.gb,NC_008687.gb	[chromosomes org. 1]
NC_007493.gb,NC_007494.gb	[chromosomes org. 2]
NC_009428.gb	[chromosome org.3]
NC_003047.gb	[chromosome org.4]
NC_008209.gb	[chromosome org.5]

Table 3: Exemplary visualization of input arguments for a 5'UTR hIntaRNA analysis.

Furthermore, a fasta formatted file containing a molecular chronometer (i.e. 16s-, linker- and 23s-region, Fig. 8) of the input organisms and RefSeq files of the chromosomes and plasmids of the respective organisms must be supplied (Table 3). An example call of the script is presented in the appendix (see paragraph 7.3 hIntaRNA user manual – Standard Operating Procedure (SOP)). Generally the work flow of the program can be split into two major processes. First, the genome wide IntaRNA predictions for each mRNA sequence are made, and the energies calculated by IntaRNA are transformed into p-values derived from generalized extreme value distributions which are fitted to the IntaRNA energies (Fig. 7). IntaRNA parameters are a seed of at least seven consecutively paired bases, a window size of 140 nt for ED value computation and a maximum distance of two paired bases of 70 nt (IntaRNA webserver standard parameters). If the analysis is run on sequences of different length, a length normalization of the energies is taken out, following the principle presented in RNAhybrid, with m and n being the lengths of the mRNA and the sRNA respectively, and e and e_n being the IntaRNA energy score and the normalized energy score, respectively.

$$e_n = -\frac{e}{\ln(mn)} \quad , \quad (14)$$

For the initial step to be executed, the sequences of interest must be parsed from the RefSeq files, which in this case depends on the Perl Bio::SeqIO module. Then the raw IntaRNA

outputs are processed into a more practical comma separated file format to which the p-values are added.

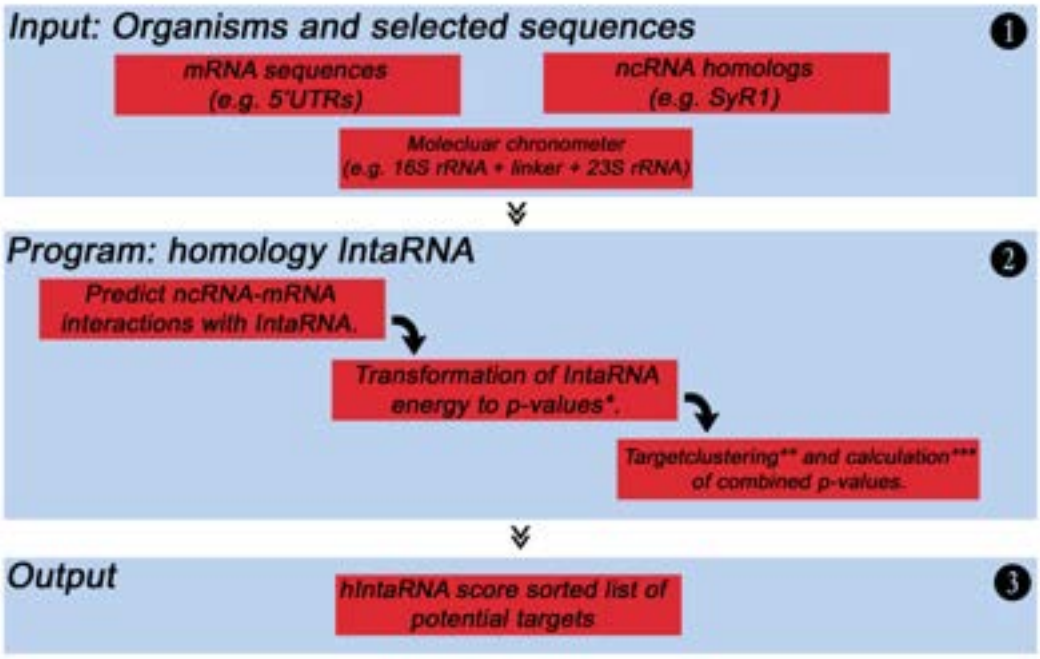


Figure 11: hIntaRNA flow chart, *Explained in 3.4, **using MBGD (Uchiyama et al., 2010) cluster table, ***Explained in 3.6.

Calculation of p-values is achieved by using the R-statistics evir package. Addition of R code into the Perl code was realized utilizing the Statistics::R Perl module. The single organism predictions are the basis for the second part of the work flow. In this part, single predictions of homologous target sequences from distinct organisms are combined. Homology between target genes is assessed, by application of a whole genome homology table calculated on the **M**icro**b**ial **G**enome **D**atabase for Comparative Analysis (MBGD) webserver (Uchiyama et al., 2010), which initially utilizes a BLAST all-against-all approach. If several genes from one organism are homologous only the gene yielding the best interaction energy is used for further processing. For the combination of p-values a distance matrix from the 16s-linker-23s regions of the participating organisms is calculated by firstly aligning the fasta and secondly creating the matrix with the Jukes-Cantor method (Jukes and Cantor, 1969). Alignment (emma) and distance matrix (distmat) software were both taken from the EMBOSS package (Rice et al., 2000). For the combination of p-values, the first step is to calculate the contributions of each organism to the distance matrix and exponentiate the p-value with this contribution. The contribution in this case, is the quotient

of the contribution of the considered organism, and the contributions of all organisms in the analysis. This step yields a somewhat weighted geometric mean of the combined p-values. After the initial combination, the degree of dependence (k_{eff}) in the analysis is assessed and final joint p-values are calculated following a method presented by Bailey and Gribskov (Bailey and Gribskov, 1998). The final output is a hIntaRNA score sorted, comma separated file, which also contains an annotation, Entrez Gene ID, interaction site details and the IntaRNA single prediction energy and p-value of the predicted targets. Furthermore, the sequences of interacting regions for the respective organisms and mRNAs are supplied in FASTA files (*.interacting.fa). For clarification reasons, a detailed, concrete example of a hIntaRNA analysis for a *Synechocystis* sp. PCC6803 SyR1 target is presented in the appendix. A detailed explanation of the output is supplied in paragraph 7.5.

4 Results

4.1 Results of the technical analyses

The first part consists of the results derived from the technical analyses leading to the final implementation of hIntaRNA.

4.1.1 Analysis of uniformity in initial p-value distributions

The theory behind function 11 is based on uniform distributed random variables between [0,1]. In order to test if this holds true for the p-values obtained from the IntaRNA energy scores, the distributions of the initial p-values were tested against a uniform distribution, employing the method of least squares. Figure 12 shows the results of the test of uniformity concerning the distributions of the initial p-values. The box plot illustrates that the p-values generated by fitting general extreme value distributions to the IntaRNA energies have smaller squared errors or deviances, when compared to a uniform distribution, than those generated by the empiric method for p-value generation.

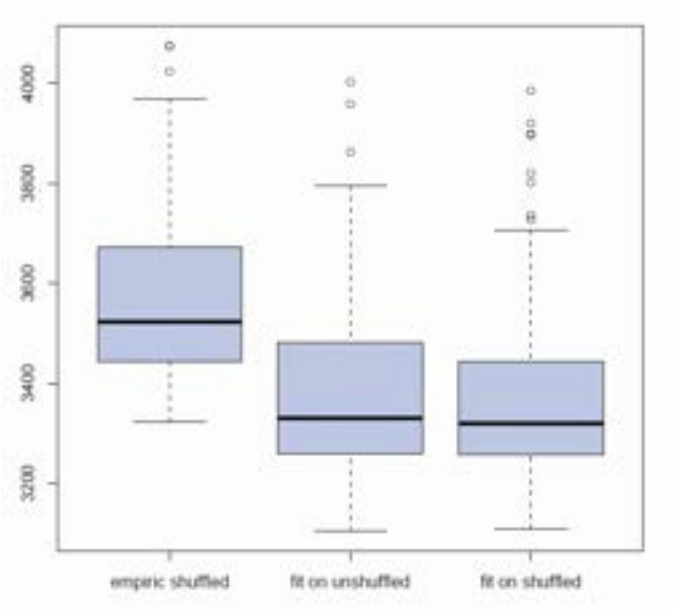


Figure 12: Box plot of the results of the method of smallest squares on initial p-value distributions for 92 data sets.

This means that the distribution of p-values derived from the fitting methods show stronger uniformity. The fitting method on unshuffled data (i.e. genome wide IntaRNA predictions) is used in the final implementation of hIntaRNA due to the major reduction in runtime without the shuffling process being incorporated.

4.1.2 Parameter comparison for EVD fits on shuffled and unshuffled target sequences for different sRNAs

In order to assess possible differences between the EVD parameters for shuffled and unshuffled target sequences, the EVD parameters (μ, σ, ε) of the respective target predictions on shuffled and unshuffled targets, were plotted against each other. Linear dependencies indicate that shuffling does not change any parameters intrinsic to the target sequences. Plotting (Figs. 13-15), reveals a linear dependency between μ and σ, while ε clearly shows two clusters. The first cluster is, similarly to the μ and σ of the first two plots, linearly correlated, while the second group clusters around 0 for the unshuffled data. This indicates that one class of sRNAs seems to follow a Gumbel EVD (i.e. ε=0), while the other group shows general EVD behavior. The shape parameters for the shuffled data do not cluster around 0, indicating that shuffling changes a property of the target mRNAs and shifts Gumbel EVDs to general EVDs. This is an interesting observation, however as the final implementation is realized with unshuffled data, this shift is not of major concern.

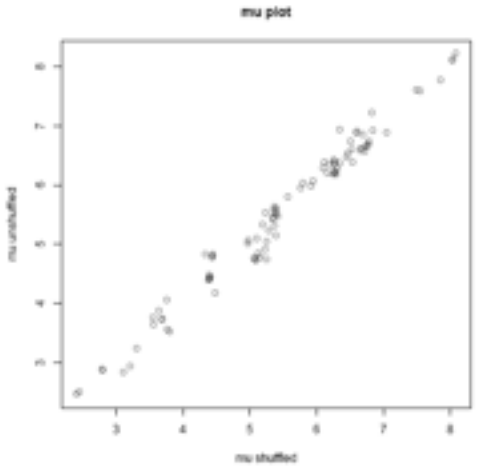


Figure 13: Plot of location (μ) parameters against each other for shuffled and unshuffled data.

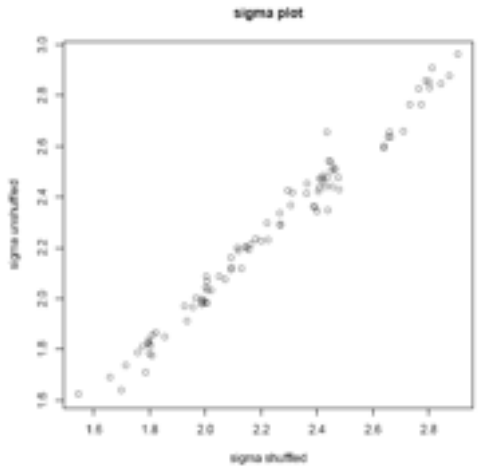


Figure 14: Plot of scale (σ) parameters against each other for shuffled and unshuffled data.

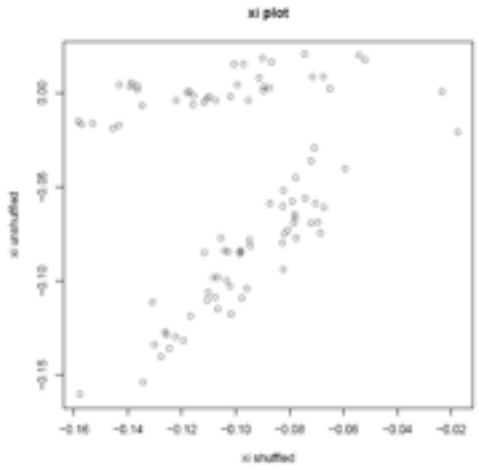


Figure 15: Plot of shape (ε) parameters, here denoted as xi, against each other for shuffled and unshuffled data.

4.1.3 Runtime

The runtime of hIntaRNA is greatly dominated by the runtime of the IntaRNA single organism predictions. The target sequence parsing also takes up a mentionable part of the calculation. The hIntaRNA benchmark prediction of SyR1 targets for instance, takes 182 minutes with NC_010296 having the highest abundance of potential target sequences/genes (6312) and NC_010546 having the longest version of SyR1 with 161 nt. The benchmark prediction for GcvB targets takes significantly more time, lasting 281 minutes. Here NC_009792 has the most potential targets (4980) and the longest sRNA sequence is that of NC_000913 with 202 nt. Table 4 shows the runtimes, sequence lengths and organism counts

while the average amount of target sequences per organism was ~4500 (length 300 nt) for the benchmarking dataset.

sRNA name	longest sRNA version	# Organisms	Runtime (min)
GcvB	202 nt	8	281
FnrS	128 nt	8	159
MicA	73 nt	8	100
CyaR	91 nt	8	114
RybB	80 nt	8	116
Spot42	110 nt	8	148
ArcZ	127 nt	8	168
SyR1	161 nt	5	182
OmrA	88 nt	3	83
InvR	91 nt	2	83
ChiX	84 nt	5	85
DsrA	89 nt	5	95
MicF	94 nt	5	102
OxyS	111 nt	5	111
MicC	110 nt	5	123
SgrS	243 nt	5	305
OmrB	95 nt	8	120
RprA	106 nt	8	137
RyhB	91 nt	5	100
GlmZ	223 nt	8	292

Table 4: Run times for the benchmark dataset. The run times were obtained using the perl "time" command at the start and the end of the master script (homology_intaRNA.pl) and calculating their difference. The processors the calculations were executed on are Quad-Core AMD Opteron™ 8378 processors with 2411 Mhz and 512 KB cache.

A detailed complexity analysis was not performed but the data in Table 4 allows an informed estimation of expected runtimes depending on sRNA length and genome size. Due to the pseudo cluster calculation (described in paragraph 3.6), analyses including more than ten organisms become increasingly large, consequently leading to impracticability with respect to the runtime. The total number of clusters, c, to be calculated in a hIntaRNA analysis of n organisms is defined as follows:

$$c = \sum_{k=2}^{n} \left(\frac{n!}{k!(n-k)!} \right) , \quad (15)$$

Parallel computation on multiple cores is enabled for the IntaRNA single organism predictions. The amount of cores used is the same as the amount of organisms participating in the analysis. This feature greatly reduces the total runtime of hIntaRNA.

4.1.4 Benchmark

When developing refined algorithms for the prediction of biological processes, the central result is always the benchmark, as it clarifies if the refinements to already present methods were successful or not. In this case, the histogram of rank frequencies (Fig. 16) already reveals the advantage of hIntaRNA (blue boxes) over the single organism predictions made with IntaRNA (green boxes) alone.

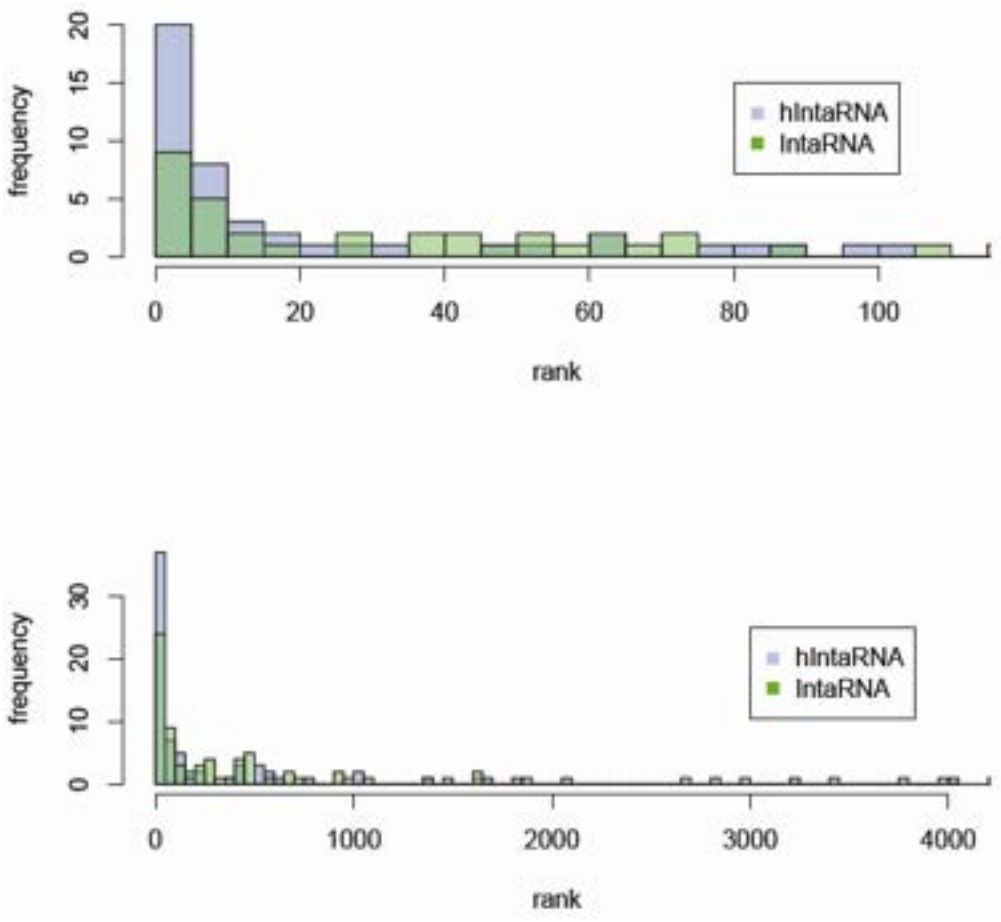


Figure 16: Histograms of the frequencies of predicted ranks in hIntaRNA (blue) and IntaRNA (green) predictions. The first plot shows a magnification of the second plot. The ranking was defined by hIntaRNA score and initial p-value for hIntaRNA and IntaRNA respectively.

The histograms in Figure 16 show the abundance of individual ranks for experimentally verified 5'UTR targets. The abundance of correctly predicted targets in the top ten hits is vastly greater in hIntaRNA. Concretely, IntaRNA predictions on the 74 experimentally verified 5'UTR targets, yield 14 (i.e. 18.9%) true positives in the top ten, while hIntaRNA doubles this number consequently placing 28 (i.e. 37.8%) targets in the top 10, six of these being on rank one. The subsequent ranks, up to rank 25 also show more hIntaRNA predictions, while the latter ranks show higher abundance of IntaRNA predictions. The unmagnified histogram clearly reveals that not all predictions reside within the top 100. 41 (i.e. 55.4%) IntaRNA predictions exceeded the rank of 100, compared to 30 (40.5%) hIntaRNA predictions. Generally, the ranks of 51 predictions (i.e. 68.9%) improve by application of hIntaRNA, while 23 (i.e. 31.1%) target predictions do not improve or worsen. Yet, this number of unimproved predictions is artificially declined due to other correct predictions outranking previous predictions.

The Benchmark dataset did not only yield information pointing at the potential of hIntaRNA. Further analysis of the results suggested a multitude of novel targets for sRNAs for which many targets have already been verified, but were not part of the benchmarking dataset (see paragraph 4.2.1).

4.1.5 Clipped benchmark

The results of the benchmark on the preprocessed single organism p-values are supplied in the supplementary Table 1. The numbers in the Table, as well as the histogram of rank frequencies clearly shows that no major improvement was achieved by pre-clipping the data. In most cases the predictions remained the same or were worse. In this context, pre-processing does not seem to be a viable practice. Furthermore, closer investigation of possible corruption of the statistical model would have been necessary if the benchmark had shown better results. In this case however, pre-clipping can simply be disregarded.

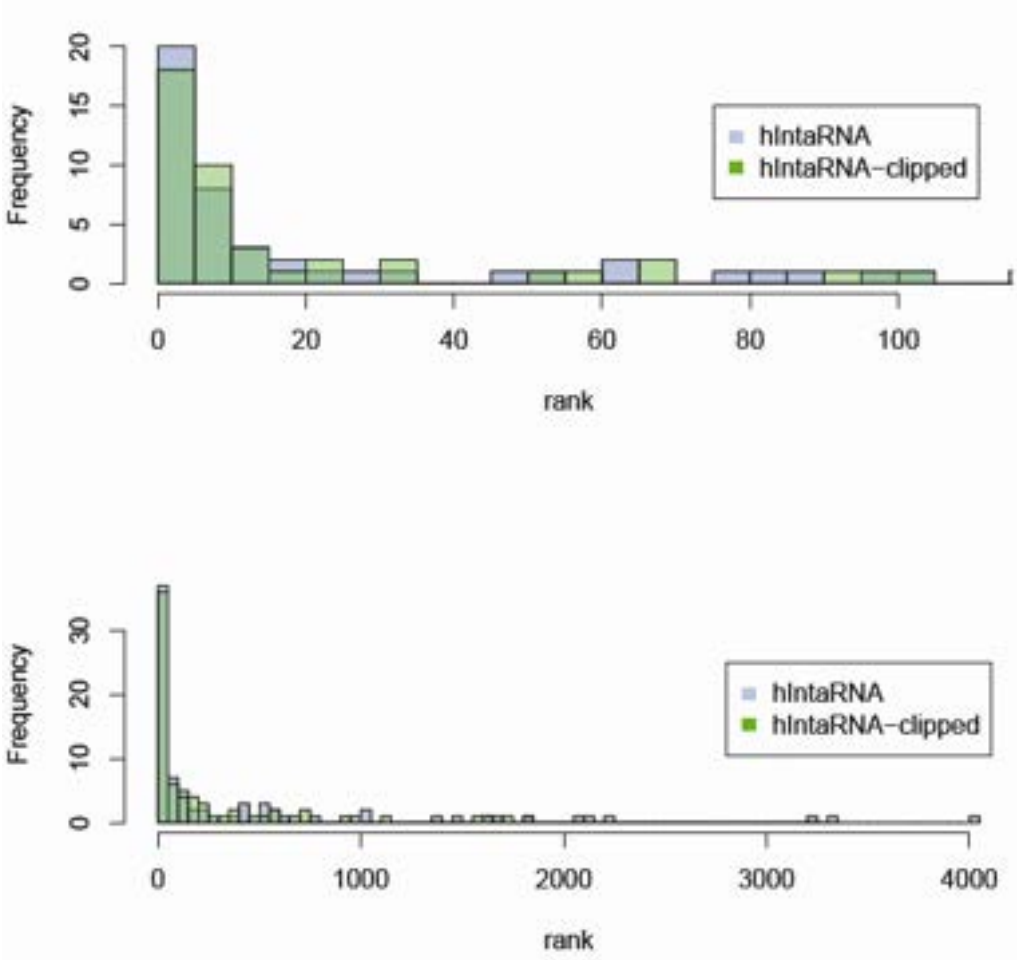


Figure 17: Histograms of the frequencies of predicted ranks in hIntaRNA (blue) and clipped hIntaRNA (green) predictions. The first plot shows a magnification of the second plot. The ranking was defined by hIntaRNA scores.

4.2 Application of hIntaRNA

The second part consists of the results derived from the application of hIntaRNA.

4.2.1 Extended analysis of the benchmarking dataset – GcvB, Spot42, RyhB and RybB

The dataset the Benchmark was executed on, generated additional data suggesting new, yet unconfirmed targets in *E. coli* and *Salmonella*, especially in the context of functional enrichment (see paragraph 3.9).

The hIntaRNA prediction for the *Salmonella* and *E. coli* GcvB sRNA, which plays an important regulative role in amino acid synthesis and amino acid transport (Sharma et al., 2011, Pulvermacher et al., 2008), yielded many new putative targets.

Functional enrichment of the *Salmonella* top 50 showed increased abundance of genes related to amino acid synthesis and transport. New putative targets clustered with genes already reported as GcvB targets. The new putative, synthesis related targets are stm0680 (*asnB*), stm1723 (*trpE*), stm2384 (*aroC*), stm1196 (*acpP*), stm1578 (*narY*), stm1725 (*trpC*), stm3161 (*metC*) and they cluster with the verified targets stm3909 (*ilvC)*, stm1299 (*gdhA)* and stm3903 (*ilvE*). Stm3909 (*yifK*) and stm0150 (*aroP*) are two novel putative transport associated GcvB targets. They clustered with the verified targets stm3564 (*livK*), stm4398 (*cycA*), stm2355 (*argT*), stm1746.s (*oppA*) and stm0399 (*brnQ*).

Clustering of the *E. coli* top 50 yielded a major group of amino acid biosynthetic process related genes, consisting of 17 entries with a high enrichment score of 6.17. Most of these are already reported. However, b1385 (*feaB*) and its homolog in *Salmonella* stm1524 (*yneI*) appear to be yet unreported GcvB targets.

The *E. coli* Spot42 sRNA top 50 predictions, displayed strong overrepresentation of genes that are important in the citric acid cycle. The enrichment score of the cluster containing the verified Spot42 target b0720 (*gltA*) and the putative targets b0728 (*sucC*), b0116 (*lpd*), b0721 (*sdhC*) and b1136 (*icd*), is 2.29. Furthermore the predicted interactions were mainly located in the single stranded regions of Spot42, which are important for base pairing with target RNAs (Beisel and Storz, 2011). Alignment analysis of the respective target

interaction sites, revealed sequence conservation (Fig. 18). Sugar transport related genes were also enriched with an enrichment score of 1.29.

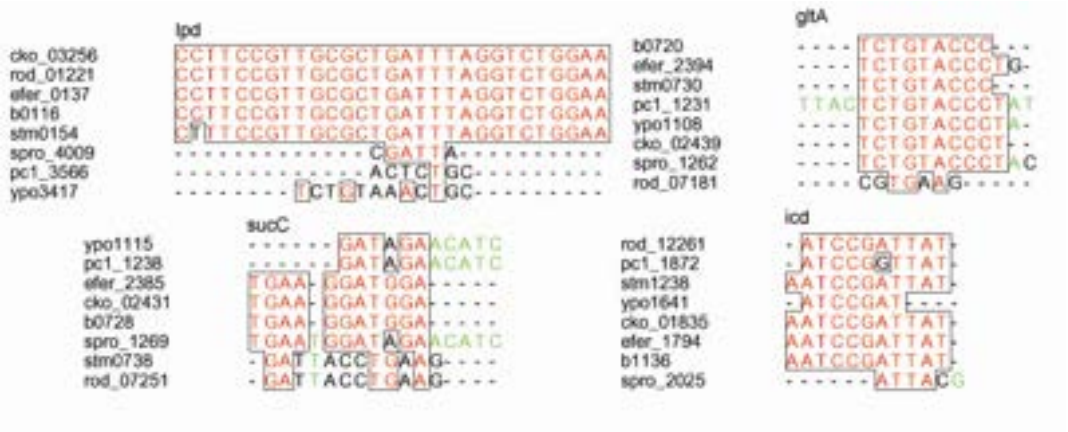


Figure 18: Alignments (created with emma (Rice et al., 2000)) of the interacting regions on the *lpd*, *gltA*, *sucC*, *icd* target mRNAs with locus tags. **cko**: *Citrobacter koseri*, **rod**: *Citrobacter rodentium*, **efer**: *Escherichia fergusonii*, **b**: *Escherichia coli*, **stm**: *Salmonella typhimurium*, **spro**: *Serratia proteamaculans*, **pc1**: *Pectobacterium carotovorum*, **ypo**: *Yersinia pestis*. The letters in red show identical residues, letters in green show similar residues, letters in black show different residues and Ts are Us in RNA context.

Another interesting result is the hIntaRNA prediction for *E. coli* RyhB. Besides the verified targets b4154 (*frdA*), b0722 (*sdhD*) and b1612 (*fumA*) (Richards and Vanderpool, 2011) being located at positions 10, 24 and 32 respectively, functional enrichment showed high abundance of metal-binding affiliated genes, yielding a cluster of 14 genes with an enrichment score of 1.35, which fits directly into the scope of already reported RyhB functions. While all genes in the cluster are certainly interesting, the most striking new potential targets are b3365 (*nirB*), b0156 (*erpA*) and b3867 (*hemN*) as they show predicted interactions with RyhB in which the SD-sequence and the start codon are occluded by base pairing with the sRNA.

Closer investigation of the *Salmonella* RybB hIntaRNA analysis indicated the presence of an outer membrane protein targeted by RybB. The interaction was subsequently verified experimentally (see paragraph 4.2.3).

4.2.2 Novel predictions for sRNAs – AbcR1 & 2, Qrr1, NsiR1 & 3

In order to assess potential functions of the sRNAs AbcR1&2, Qrr1 and NsiR1&3, hIntaRNA predictions were carried out for all three sRNAs and the results were subsequently analyzed employing the DAVID functional enrichment tool.

Table 5 shows the top 15 hIntaRNA predictions for the AbcR1 sRNA in *Agrobacterium tumefaciens*. The verified AbcR1 target atu2422, here on rank 32 (not shown in table 5), is an ABC transporter.

hIntaRNA score	NC_003062	Annotation
3.473395e-06	atu3487	ABC transporter substrate binding protein (sugar)
3.484135e-06	atu5280	Hydrolase
1.810383e-05	atu3076	uracil transport protein
4.233513e-05	atu2737	Oxidoredutase
8.595159e-05	atu2493	forms a tetramer composed of 2 alpha subunits and 2 beta subunits
1.129300e-04	atu1174	pyrophosphate-energized proton pump
1.719447e-04	atu3198	ABC transporter substrate binding protein (ribose)
1.753479e-04	atu3485	short chain dehydrogenase
1.798424e-04	atu4550	LacI family transcriptional regulator
1.964779e-04	atu5071	ABC transporter substrate binding protein (dipeptide)
2.050593e-04	atu4123	ABC transporter substrate binding protein (branched amino acid)
2.222805e-04	atu4537	ABC transporter membrane spanning protein (amino acid)
2.244232e-04	atu3253	ABC transporter substrate binding protein
2.894201e-04	atu3114	ABC transporter substrate binding protein (sugar)
3.412708e-04	atu2296	2-dehydro-3-deoxygluconokinase

Table 5: hIntaRNA top 15 predicted targets for AbcR1 sRNA in *Agrobacterium tumefaciens*.

Due to the fact that sRNAs are often included into specific regulatory networks and affect mRNA targets with similar protein products, it is very striking that eight of the top 15 predictions for AbcR1 are also ABC transporters. This hints at the quality of the prediction for AbcR1, not only with respect to many of the targets belonging to the same group of genes (i.e. being functionally enriched) but also with respect to an already verified target (atu2422) also being an ABC transporter. Furthermore DAVID functional enrichment reported enrichment of genes related to periplasmatic space, cell envelope and external encapsulating structure for the top 50 hIntaRNA predictions with an enrichment score of 2.91. An enrichment of the top 50 predictions for the *Agrobacterium tumefaciens* AbcR2 sRNA rendered a cluster with genes related to the same cellular structures, but with a reduced enrichment score of 1.34.

The hIntaRNA prediction for the *Vibrio harveyi* Qrr1sRNA ranks the known target *luxO* (vibhar_02959) on rank two, while *luxR* (vibhar_00157), also a known Qrr1 target, is not correctly predicted and is ranked on position 1119. Functional enrichment of the top 50 resulted in strong representation of sugar metabolism with an enrichment score of 2.49. Cell membrane, envelope and transport related genes were also enriched, yielding an enrichment score of 1.09.

The *Anabena sp.* 7120 NsiR1 top 50 of the prediction yielded an overrepresentation of vitamin and nitrogen metabolic process associated genes with an enrichment score of 0.76, while the NsiR3 prediction showed stronger enrichment of peptidoglycan related genes with an enrichment score of 1.69. Furthermore, *patU3* (alr0101) and *invB* (alr0819) were highly ranked in the NsiR1 prediction.

4.2.3 STM1530 is a novel RybB target in *Salmonella*

The STM1530::GFP fusion experiment (performed by Dr. Kai Papenfort from Uni Würzburg) verified the interaction between the *Salmonella* STM1530 mRNA and the RybB sRNA. The western blot (Fig. 20) clearly reveals that the regulation is negative and based on RybB presence, as no down regulation of GFP expression is visible in the RybB mutant (RybB-M2).

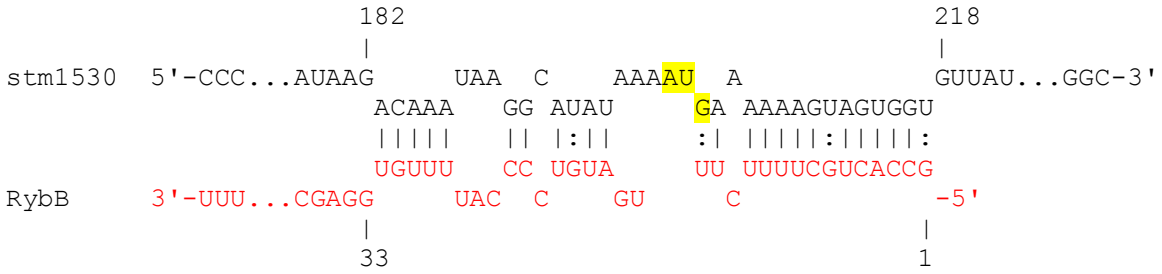


Figure 19: Interaction of STM1530 mRNA (black) and RybB sRNA (red) as predicted by IntaRNA (energy = -16.438 kcal*mol^{-1}). AUG (yellow box) located at 201-203 on the mRNA.

Furthermore the western blot shows that the central interacting region of RybB (R16TOM) is the same region, which interacts with the other, already earlier verified (Papenfort et al., 2010) RybB targets, even though the down regulation does not seem to be as strong as with the full RybB sequence. These authors suggested that the 3' flanking region of the interaction site on the target sequence, in most cases, initially consists of an adenosine. This

does not hold true for STM1530, where the interaction is flanked by a 3' guanine on the mRNA (Fig. 19).

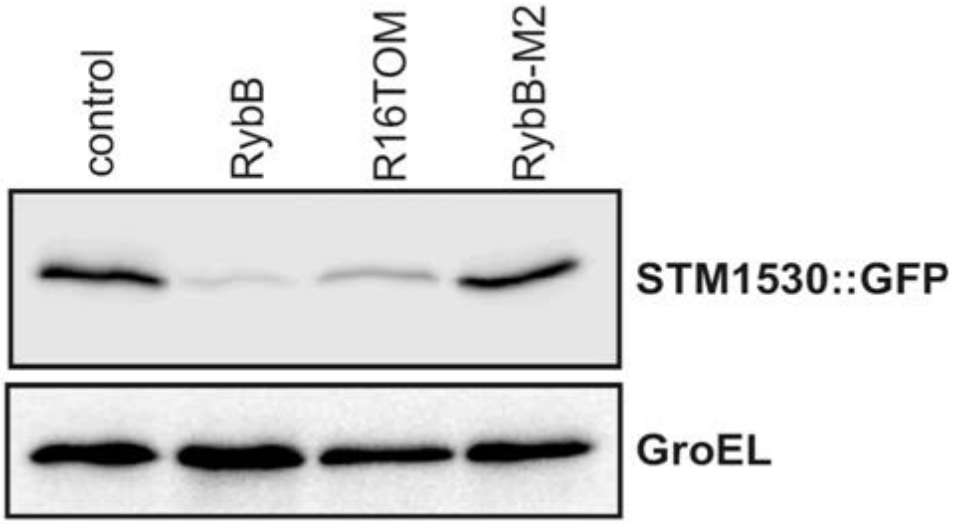


Figure 20: Western blot of the STM1530::GFP fusion experiment. RybB showing the GFP amount under RybB expression, R16TOM showing the GFP amount if only the R16TOM region (Papenfort et al. 2010) of RybB is expressed and RybB-M2 showing the GFP amount in a RybB mutant sequence. GroEL is a protein used as standard to show the average protein abundance. The experimental procedures (Urban and Vogel, 2007, Towbin et al., 1979) were done externally, in collaboration with Dr. Kai Papenfort (Würzburg).

5 Discussion

5.1 Discussion of technical results

This part of the discussion analyzes and interprets the results from the technical analyses, leading to the final implementation of hIntaRNA.

5.1.1 Initial p-values and run time

Uniformity of initial p-values did not greatly differ, especially between the methods in which the p-values were derived from fitting functions to the experimental (i.e. IntaRNA energies) data. In fact, the finding that empiric p-values generated on shuffled data show inferior uniform distributions compared to the p-values generated on the extreme value distribution fits was greatly helpful in reducing the run time of hIntaRNA. Ten times shuffled data increases the run time tenfold due to the increased sequence quantity on which IntaRNA predictions have to be carried out on, leading to reduced user practicability. Consequently, fitting an extreme value distribution to the IntaRNA energies is a statistically sound method for p-value deduction which furthermore shows acceptable run time.

The results of the comparison of EVD parameters between shuffled and unshuffled fits yielded a curious result. While location and scale are linearly correlated, the shape parameter shows two clusters: A top cluster and a bottom cluster, with reference to their location in figure 15. Interestingly, dinucleotide shuffling seems to change an attribute in certain target mRNAs, distributing the data points of the top cluster across the entire x axis, while the parameters on the y axis remain the same. With the shape parameter being zero, the unshuffled data for this cluster follows a Gumbel EVD, while the data shifts toward general EVD behavior for the shuffled data. Closer investigation of the sRNAs participating in the respective clusters, revealed two classes of sRNAs participating in the clusters. The top cluster consists of ArcZ, ChiX, GcvB, MicA, OxyS, RyhB and SgrS while the bottom cluster consist of CyaR, MicC, FnrS, MicF, OmrA, RprA, RybB, Spot42 and SyR1. It is tempting to conclude that the shift from Gumbel EVD to general EVD is based on specific sequence patterns being blended by the dinucleotide shuffling. Possibly the interactions for the sRNAs in the top cluster have stronger sequence motif dependence than those present in

the bottom cluster. However, this remains speculative and exceeds the boundaries of this project. If anything, this supports the use of the unshuffled data as background model as it gives a better description of the actual EVD.

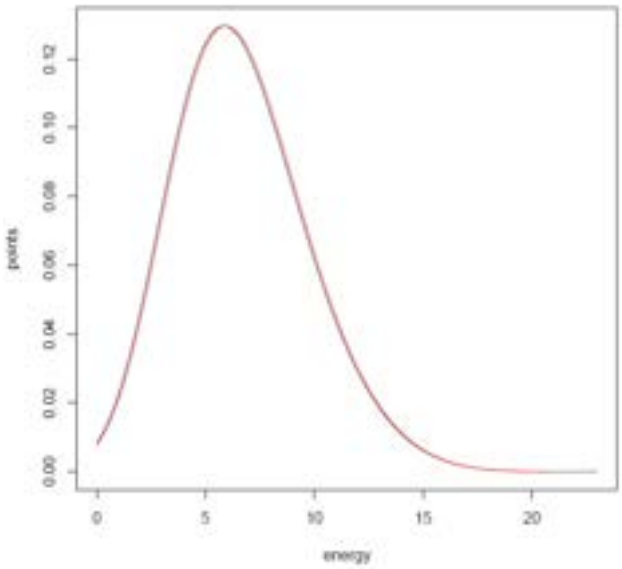


Figure 21: General EVD fits for the *Salmonella* (NC_003197) CyaR sRNA. (red: fit on unshuffled data, black: fit on shuffled data)

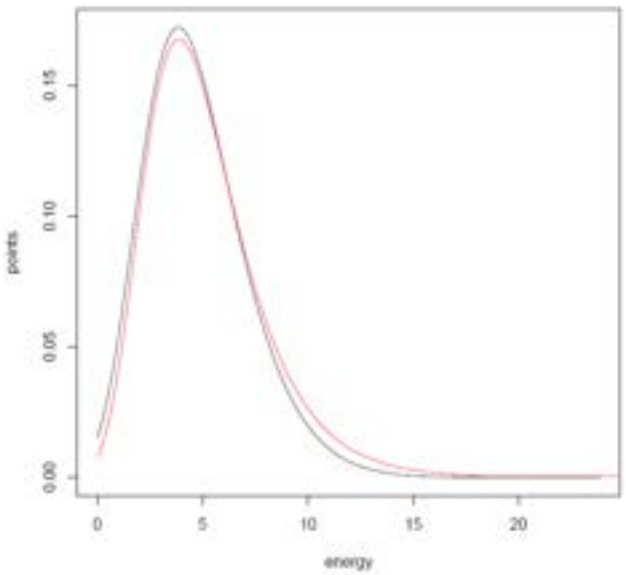


Figure 22: General EVD fits for the *Salmonella* (NC_003197) ChiX sRNA. (coloring as in figure 23)

Figures 21 and 22 depict the shift between the two clusters graphically. While CyaR is associated with the bottom cluster and shows general EVD behavior for shuffled (black) and unshuffled (red) data, the two plots for ChiX clearly deviate from each other.

5.1.2 Benchmark

Being the core result of this project, the benchmark is of increased interest. The results can be viewed from two angles. Firstly the general prediction performance and secondly the performance compared to the underlying algorithm can be analyzed. Comparing the IntaRNA and hIntaRNA predictions for the benchmarking dataset outlines the advantage of hIntaRNA and highlights its increased reliability compared to IntaRNA and consequently to other sRNA target prediction algorithms which are inferior to IntaRNA (Busch et al., 2008). Good predictions, of verified sRNA targets, in IntaRNA (i.e. top 50) yield better predictions in hIntaRNA for all examples except for three (Spot42 – b4311, OmrA – b0565, ChiX – b0619, yellow in supplementary Table 1) and average IntaRNA predictions (i.e. 50 < rank < 300) are often elevated into the top 50 in hIntaRNA (green in supplementary table 1). In the benchmarking data this is the case for 15 targets, while the improvements for RybB-stm1995 (rank 450 → 6) and SyR1-sll1578 (rank 61 → 1, i.e. Appendix 7.1) are particularly impressive. Yet, the elevation of stm1995 has to be put into perspective as the interaction employed in the clustering, is that of RybB and stm1530 which clusters with stm1995 in the MBGD cluster.tab but has a better p-value than stm1995 and is therefore chosen for p-value combination. Stm1995 is subsequently added to the results and acquires the same rank as stm1530. This kind of improvement is certainly not the standard but in this case a convenient addition. Also, the interaction predicted for stm1995 would definitely be considered when looking for new RybB targets, as stm1995 is an outer membrane protein and the interacting region on RybB is the 5'end which is central for the RybB-target interplay (Fig. 23, Papenfort et al., 2010).

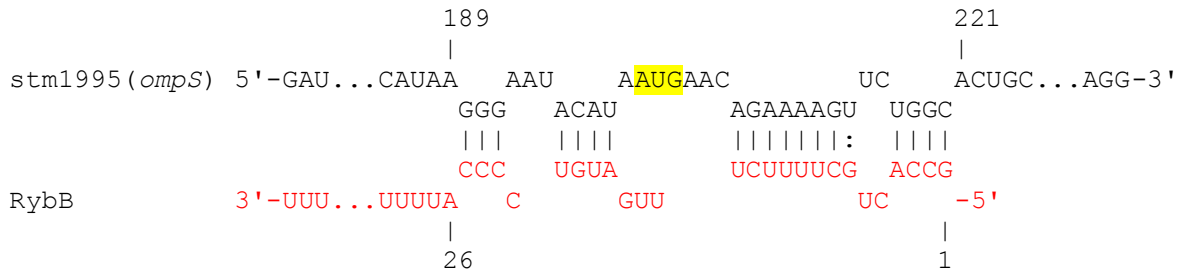


Figure 23: Interaction of STM1995 (*ompS*) mRNA (black) and RybB sRNA (red) as predicted by IntaRNA (energy = -12.026 kcal*mol^{-1}). AUG (yellow box) located at 201-203 on the mRNA.

The significant rank increase for ChiX-b0619 can be explained by a lack of conservation on the regulatory side. Most likely the regulation of this cluster of homologous genes is not conserved across all five organisms taking part in the analysis. *E. coli* b0619 has the best p-value by far with 0.0058 compared to the other four organisms which acquire p-values between 0.13 and 0.54. Furthermore they show entirely different interaction sites when compared to the *E. coli* interaction site (Fig. 24). In fact the alignment in figure 24 shows higher conservation in all four other organisms with strong overrepresentation of thymine stretches. This also points at a lack of regulatory conservation because these conserved regions lead to arbitrary IntaRNA predictions. The suboptimal IntaRNA prediction for b0619 (i.e. b0619subop in Fig. 24) also shows thymine stretches, additionally pointing at the arbitrary character of the predictions in *Citrobacter koseri*, *Citrobacter rodentium*, *Salmonella* and *Escherichia fergusonii*.

Figure 24: Alignment of ChiX target regions for b0619 homologs and the best suboptimal interaction of ChiX and b0619. Further specifications are as in figure 18.

The impairment for the OmrA-b0565 and Spot42-b4311 predictions is not as grave as in ChiX-b0619 but generally the same scheme holds. The only difference is that not all organisms in the analysis have a b0565 or b4311 homolog. These three examples outline the downside of hIntaRNA. Prediction reliability only increases if the regulation is conserved for a specific target in at least some of the participating organisms. However, this is often given and in any case the IntaRNA single organism predictions are also available and can be viewed in the context of functional enrichment together with the hIntaRNA results.

Assuming predictions with ranks smaller or equal to 20 being correct, this means that against the background of the benchmark, the likelihood of a correct prediction is 43%. The choice of rank 20 may seem arbitrary and it is certainly advised to choose a final score cutoff over a rank threshold but it can be hypothesized that in most cases the final score threshold will be affiliated with a higher than rank than rank 20. Furthermore, a plausible value for a final score cutoff has not yet been assessed.

Even though the general result of the benchmark is positive, it is also clear that not every target is correctly predicted at this time. This can either be due to IntaRNA not identifying an interaction or due to a lack of regulatory conservation. Both cases lead to predictions beyond the scope of consideration even for an expert user (i.e. blue in supplementary table 1). In the case of the benchmarking dataset, this counts for 27% of all predictions. This outlines that refinements are still necessary, especially in the base line prediction algorithm IntaRNA.

Preprocessing by clipping p-values higher or equal to 0.8 of single organism IntaRNA prediction p-values was tested but did not yield significantly improved results and is therefore not a part of the hIntaRNA implementation. It can be hypothesized that the negative effects of pre-processing the data in this way may even have the negative effect of increasing the false positive rate. This is supported by the fact that most hIntaRNA predictions which yield possibly interesting results (i.e. top 50) stay the same or worsen in all cases of the predictions with preprocessing.

Generally the hIntaRNA method is, compared to comparative approaches such as presented in RNAplex (Tafer et al., 2011, see paragraph 2.7), simple and unrestrictive. The only constraint is the conservation of a target gene in at least two of the participating organisms, while the type or region of an sRNA-mRNA interaction are not restricted. This leads to a reduced count of false negatives, while it most likely increases the rate of false positives. However, the benchmarking results strongly suggest that the negative effects are outweighed by the positive effects of reduced restriction.

5.2 Discussion of hIntaRNA application results

This part discusses the results of the hIntaRNA application to selected sRNAs. New targets and regulatory networks these sRNA may or do belong to are suggested and expanded.

5.2.1 GcvB, Spot42, RyhB and RybB

Functional enrichments of the top 50 *E. coli* and *Salmonella* GcvB targets yield compelling results. Not only that the enrichment score for the GcvB top 50 in *E. coli* is the best amongst all analyzed sRNAs in this project, but also the abundance of correct predictions in the

upper ranks is high. This supports the notion of GcvB acting as a major player in the amino acid metabolism network (Sharma et al., 2011) and also extends the scope of its influence by ten putative targets in *Salmonella* and one putative target in *E. coli*.

Sugar transport associated genes are enriched in the Spot42 top 50 and hereby validate the role of Spot42 in sugar metabolism (Beisel and Storz, 2011) for *E.coli*. An especially interesting result is the prediction of four additional (beside *gltA*) citric acid cycle associated genes (Fig. 25). It has been reported that Spot42 posttranscriptionally regulates citric acid associated gene *gltA* negatively (Beisel and Storz, 2011). The citrate synthetase, which is the protein product of *gltA*, plays a central role in the citric acid cycle, in that it catalyzes the reaction of oxaloacetic acid and acetyl-CoA to citric acid (Wiegand and Remington, 1986). The novel putative targets also take central roles in the citric acid cycle (Fig. 25).

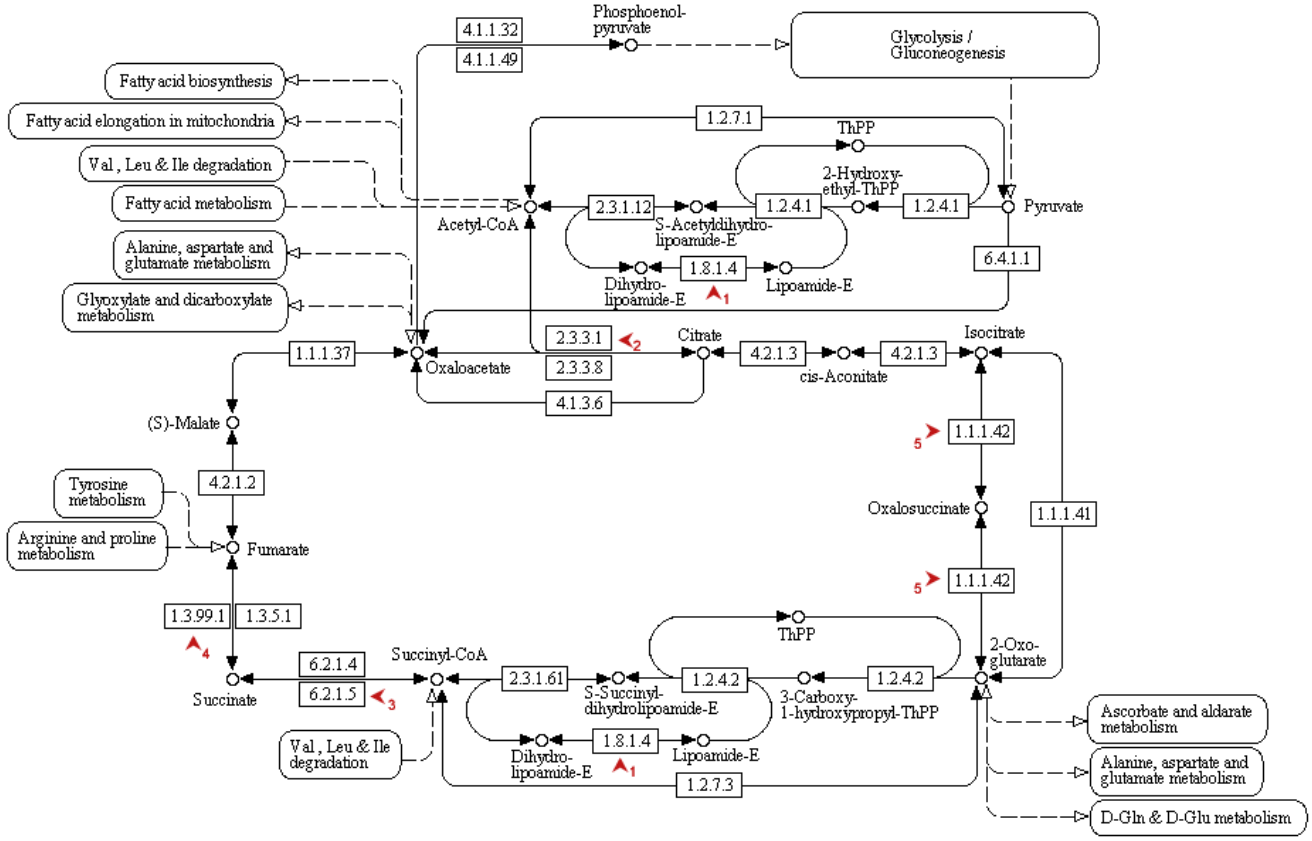


Figure 25: Marked 1-5, KEGG (Kanehisa et al., 2011) map of the citric acid cycle with putative and verified Spot42 targets. 1: *lpd* (dihydrolipoamide dehydrogenase), 2 (verified): *gltA* (citrate synthase), 3: *sucC* (succinyl-CoA synthetase beta subunit), 4: *sdhC* (succinate dehydrogenase flavoprotein subunit), 5: *icd* (isocitrate dehydrogenase)

Isocitrate dehydrogenase (*icd*) catalyzes the formation of α-ketoglutarate (2-oxoglutarate) from isocitrate while succinyl-CoA synthetase (*sucC*) aids in synthesis of succinate from succinyl-CoA. Succinate dehydrogenase (*sdhC*) catalyzes fumarate creation from succinate, whereas dihydrolipoamide dehydrogenase (*lpd*) is not directly incorporated into the citric acid cycle, but takes part in a parallel metabolic process, which withdraws 2-oxoglutarate from the citric acid cycle and consequently influences its reaction equilibriums. In all cases, the predicted interaction site in the reported Spot42 structure (Møller et al., 2002) is one of the three single stranded regions central to Spot42 interactions. This supports the notion of these targets being regulated by Spot42 due to the predicted interaction being functionally sound. Should all four additional targets be verified, this would assign a key role to Spot42 in the regulation of the citric acid cycle. By modulating availability of central enzymes, Spot42 could work as a performance regulator of the citric acid cycle throughput. Previous work suggests that Spot42 actually negatively regulates the *sdhC* mRNA (number 4 in Fig. 25) as overexpression of Spot42 reduces *E. coli* growth rate on succinate medium (Rice and Dahlberg, 1982). Furthermore, the succinate dehydrogenase mRNA is also a reported target of the sRNA RyhB (Massé and Gottesman, 2002) and overexpression of RyhB also interferes with cell growth on succinate as carbon source, illuminating that *sdhC* is prone to sRNA regulation and its repression leads to problems in cell growth on succinate.

It has been reported that Spot42 levels are high when glucose concentration is high (Sahagan and Dahlberg, 1979a, 1979b) but it remains delusive why exactly the citric acid cycle should be down regulated under glucose presence, as glucose indirectly injects the citric acid cycle with acetyl-CoA via glycolysis. A possible explanation is that Spot42 may have a stabilizing effect on the reaction equilibriums within the citric acid cycle. Especially the intracellular levels of citric acid seem important with respect to the activity of the enzyme phosphofructokinase-1, which catalyzes the formation of fructose-6-phosphate to fructose1,6-bisphosphate and is allosterically inhibited by citric acid. Spot42 possibly aids in keeping citric acid levels stable, and by this prevents overrepression of the phosphofructokinase-1. Similar theories may hold for the four additional suggested putative Spot42 target mRNAs with products participating in the citric acid cycle. All together Spot42 might achieve a steady and constant flow between glycolysis and the citric acid cycle. The fact that overexpression of Spot42 impairs growth on succinate medium supports this theory, in the case that *sdhC*, is indeed a Spot42 target. Overexpression shifts the Spot42 concentrations out of their regular magnitude and would consequently lead to an

abolishment of the regular role of Spot42 as a stabilizer and change its role to a master regulator completely switching translation of all existing target mRNAs off.

Conversely to the current notion (Richards and Vanderpool, 2011), the data collected for RyhB suggests that the scope of RyhB regulated RNAs is beyond mRNAs coding for iron associated proteins only. It seems like RyhB has a more global role, and might regulate metal associated proteins in general. The mRNAs of b3365 (*nirB*), b0156 (*erpA*) and b3867 (*hemN*) are particularly interesting, as the predicted interactions in all cases occlude the SD-sequence and the start codon, which is a known method by which RyhB down-regulates its target mRNAs (Richards and Vanderpool, 2011). Masking of the start codon and SD-sequence for these three mRNAs is not only the predicted interaction mechanism in *E. coli*, but also in all of the other genes of the three respective clusters. This hints at functionally correct prediction of conserved regulation by RyhB in all three mRNAs across species boundaries. Furthermore, there is evidence that *nirB*, *hemN* and *erpA* are non essential genes (Gerdes et al., 2003), which fits the scheme of RyhB regulation (Richards and Vanderpool, 2011), as it mainly down-regulates transcription of mRNAs that code for non-essential proteins that need iron as co-factor if iron is limited.

A curious observation is the putative linkage of the RyhB and Spot42 regulatory networks by the citric acid cycle. If the theory proposed for Spot42 regulation of the citric acid cycle is true, the effects of the two respective sRNAs on the citric acid cycle are different from the angle of their impact on the translation of their targets. While Spot42 would act as a fine tuning device, RyhB would rather act as master regulator under iron limiting conditions and repress translation of citric acid cycle, metal associated target mRNA products.

The target prediction for RybB and subsequent experimental testing of the predicted target stm1530 revealed a new, yet unknown outer membrane protein mRNA target. Stm1530 ranks on position six in the hIntaRNA prediction, while it only ranked on position 74 with regular IntaRNA. Consequently, stm1530 is the first new target suggested by hIntaRNA, which was afterwards verified experimentally. Strangely, this RybB target was not previously investigated as it is similar to the other already verified outer membrane protein targets of RybB and the interaction is also mainly the same. However, the interaction is not flanked by a 3’ adenosine and does not fit into the current notion of RybB-target interactions (Papenfort et al., 2010). This new data suggests a re-evaluation of the “3’ adenosine”-theory and its potential in aiding sRNA target predictions.

5.2.2 AbcR1 & 2, Qrr1, NsiR1 & 3

The hIntaRNA results for the AbcR1 and AbcR2 sRNAs stress the previously suggested notion of these two sRNAs having global roles in ABC transporter regulation in *Agrobacterium tumefaciens* (Wilms et al., 2011). The type of ABC transporters that seem to be regulated by AbcR1 do not seem to be restricted to one type of transporter, but rather a variety of sugar, peptide and amino acid transporters appear to be targeted. In total, the results suggest at least seven additional ABC transporter mRNA targets for AbcR1.

Evaluating the results for the *Vibrio harveyi* Qrr1 prediction is not trivial. Obviously sugar processing genes are strongly enriched. However, this is hard to synchronize with currently verified Qrr targets. In any case, the regulation, at least for vibhar_01575 and vibhar_01345 seems to be negative as the start codon is occluded by the interaction with the sRNA. Unfortunately the results for Qrr1 are very inconclusive, but this is strongly related to the lack of known targets for this sRNA.

Interestingly, the heterocyst-associated genes alr0101 (*patU3*) (Zhang et al., 2007) and alr0819 (*invB*) (Vargas et al., 2011) are highly ranked in the hIntaRNA prediction for NsiR1. NsiR1 is present in heterocysts (Mitschke et al., 2011b) as are alr0101 and alr0819. Together this evidence suggests a direct regulation of these targets by NsiR1. The enrichment for NsiR3 also hints at heterocyst-specific tasks.

Finally, it must be stressed that all new target candidates, except for stm1530, presented in the paragraphs 5.2.1 and 5.2.2 are solely suggestions. This outlines the problem of pure predictive approaches, as all results remain speculative. However, experimental testing of these target site suggestions is highly likely to verify the predictions in many cases.

Furthermore, functional enrichment scores are not the gold standard for the evaluation of an sRNA target prediction. This has two reasons. First of all an sRNA may only have a small and limited amount of targets leading to a low abundance of the metabolic processes they participate in amongst the top hits, and secondly the quality of an enrichment is strongly based on the quality of functional annotation of a genome. If a genome is poorly annotated, functional enrichment will not work regardless of the sRNA target prediction quality. Yet, a good enrichment score is a strong indicator of prediction validity if it surpasses a certain threshold especially if regulatory networks, which have already reported targets for a specific sRNA, are significantly enriched.

6 Outlook

The future for sRNA research is very bright. While research progresses, it is becoming clearer that sRNAs and RNAs in general have global roles in regulation. In line with this, state of the art research in regulative processes will no longer be able to neglect RNA contributions. *Trans*-acting sRNAs may soon be, similar to proteins acting as global transcription factors, regarded as global translation factors and will receive more attention in research, as well as in text books conveying basic knowledge of genetics. Two pages refer to *trans*-acting sRNAs in the 2007 edition of the text book Genes IX (Lewin, 2007), while only one page is devoted to sRNAs in Genes VIII (Lewin, 2003). sRNA regulation also opens new doors for biotechnology and medicine. Especially RNA-based medication seems very promising, when targeting pathogens, which are not responsive to common antibiotics. The function can be very specifically adjusted by the nucleotide sequence and may in this way be able to specifically target pathogens, which are at the present time very difficult to antagonize. In biotechnology, knockouts of specific sRNAs could lead to higher yields of certain desired products such as bio fuels, or introduction of engineered or ectopically expressed sRNAs could change metabolic processes in favorable ways.

Sequencing technologies are progressing towards the third generation. Due to these technological advances, the cost of sequencing transcriptomes will be further reduced and the quality of sequencing results will increase, simplifying the detection of new sRNAs and introducing new material for sRNA research.

As for sRNA target prediction, the future is also very good. Algorithmic approaches will improve, making it continuously easier to predict sRNAs and their targets. The more sRNA-target interactions are verified, the more base material to abstract from will be available. While single prediction software will improve, the hIntaRNA concept will improve alongside, as the general idea is not restricted to IntaRNA. However, hIntaRNA is still open to improvements. Methodology for identifying homologous genes between species can most certainly be refined, and automatic gene ontology clustering should be added. Furthermore, assessment of final score thresholds must be a center of future research. Generating scores, that yield a clear measure of reliability, is important, especially for inexperienced users. Detaching hIntaRNA from Linux command line and establishing an easy to use webserver application is planned.

7 Appendix

7.1 hIntaRNA an example – the SyR1 target *cpcA*

In a hIntaRNA analysis of potential SyR1 targets in five distinct cyanobacteria, initially the genomic regions of interest (in this case 200 upstream and 100 downstream of the start codon) are parsed from the RefSeq files and the single organism IntaRNA predictions are carried out.

In the following step the homologous genes are clustered. In the case of *cpcA*, all five organisms contain a *cpcA* gene, consequently leading to a cluster of five (i.e. sll1578, mae_24460, cce_2652, pcc7427_0160, cyan8802_3045 with respective single IntaRNA p-values of 0.0258; 0.0001; 0.0000119; 0.00031; 0.397).

1	2	3	4	5	organism
0.00	7.19	10.93	11.76	12.45	NC_010546 1
	0.00	11.04	13.05	12.34	NC_013161 2
		0.00	14.00	13.83	NC_011729 3
			0.00	12.19	NC_000911 4
				0.00	NC_010296 5

Table 6: Distance matrix created for NC_010546, NC_013161, NC_011729, NC_000911, NC_010296 16s-linker-23s sequence, using the Jukes-Cantor correction method.

The combination of p-values, using the distance matrix in table 3, for this specific example is executed as follows:

$$0.0258^{51/237.56} * 0.0001^{50.81/237.56} * 0.0000119^{42.33/237.56} * 0.00031^{49.8/237.56} * 0.397^{43.62/237.56} = \underline{0.00131}$$

Then the k_{eff} for the clusters of five is assessed (i.e. part 3.6), in this case yielding a k_{eff} of 4. The combined p-value is powered by four, which produces the final score for the *cpcA* cluster as follows:

$$(0.00131)^4 = \underline{2.945 * 10^{-12}}$$

Overall, this leads to an elevation of the SyR1 - *cpcA* target prediction from rank 61 to rank one.

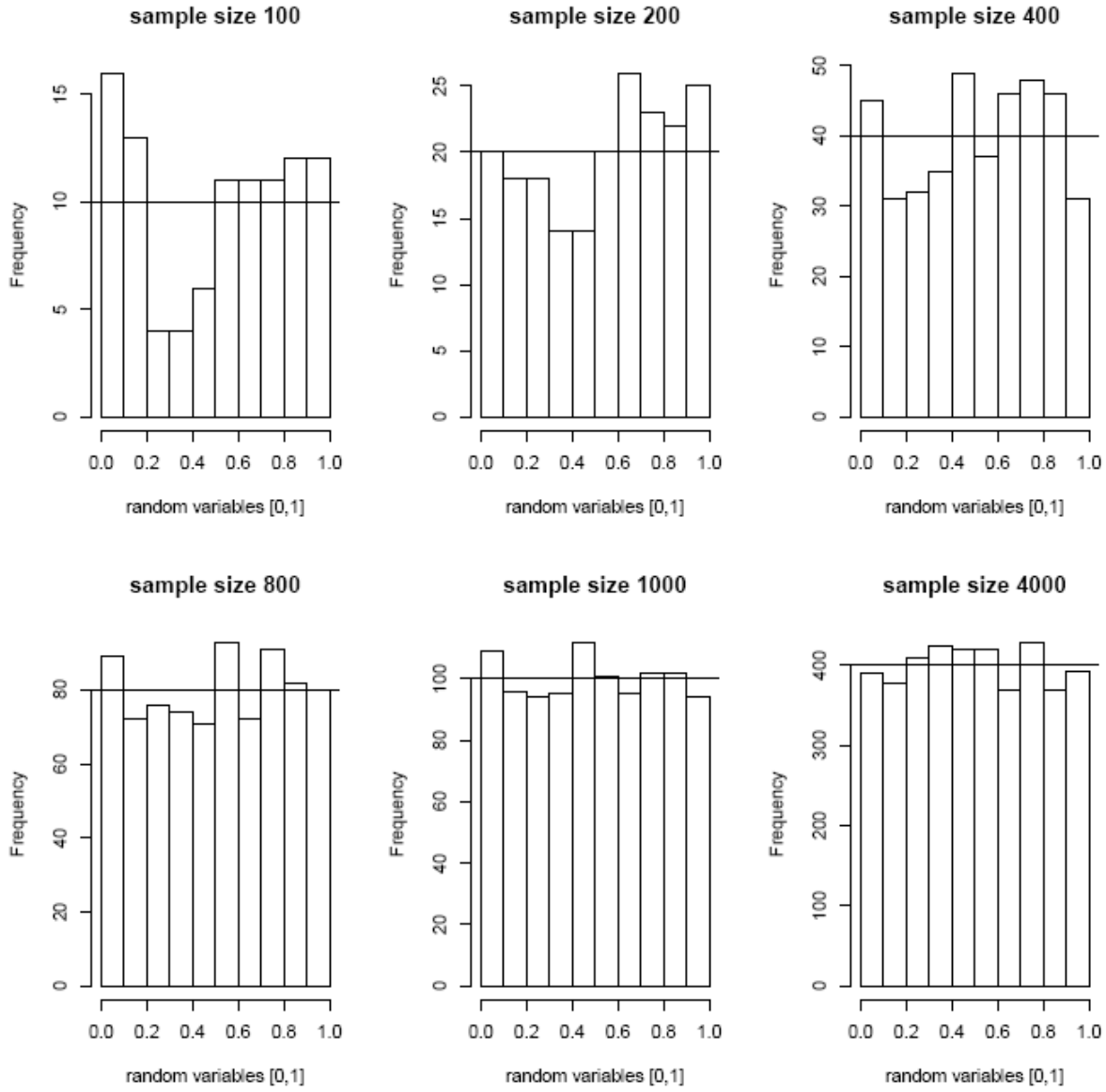
7.2 Benchmark table

ncRNA	Target	IntaRNA.pos	homology.IntaRNA.pos	Organism	improvement	clip 0.8.pos
ArcZ	b2741	247	520	NC_000913	0	396
ArcZ	stm2970	4214	1828	NC_003197	1	N/A
ArcZ	stm3216	2969	3215	NC_003197	0	3328
ArcZ	stm1682	3976	1034	NC_003197	1	931
ChiX	b1737	2	2	NC_000913	0	2
ChiX	b0619	6	190	NC_000913	0	212
ChiX	b0681	4	3	NC_000913	1	3
ChiX	stm1313	3	2	NC_003197	1	2
ChiX	stm0687	9	3	NC_003197	1	3
CyaR	b2687	405	97	NC_000913	1	101
CyaR	b1740	1606	1611	NC_000913	0	2244
CyaR	b0814	46	5	NC_000913	1	7
CyaR	b2666	491	532	NC_000913	0	562
CyaR	stm0833	26	5	NC_003197	1	7
DsrA	b1237	6	3	NC_000913	1	3
DsrA	b2741	1	2	NC_000913	0	2
FnrS	b2153	414	448	NC_000913	0	317
FnrS	b2303	3776	2063	NC_000913	1	2141
FnrS	b0755	1086	574	NC_000913	1	701
FnrS	b1479	264	51	NC_000913	1	58
FnrS	b1656	705	76	NC_000913	1	92
GcvB	stm2355	45	12	NC_003197	1	12
GcvB	stm4398	13	4	NC_003197	1	4
GcvB	stm3630	14	9	NC_003197	1	9
GcvB	stm0665	106	6	NC_003197	1	8
GcvB	stm3567	68	205	NC_003197	0	233
GcvB	stm3564	43	3	NC_003197	1	3
GcvB	stm1746.s	284	20	NC_003197	1	24
GcvB	stm4351	490	1394	NC_003197	0	1556
GcvB	stm3909	6	5	NC_003197	1	5
GlmZ	b3729	410	134	NC_000913	1	153
InvR	stm1572(stm1530)	3419	159	NC_003197	1	159
MicA	b0957	64	11	NC_000913	1	11
MicA	b1130	71	34	NC_000913	1	35
MicA	stm4231	143	15	NC_003197	1	15
MicC	b2215	3	1	NC_000913	1	1
MicC	stm1572(stm2267)	271	1	NC_003197	1	1
MicC	stm2267	1	1	NC_003197	0	1
MicF	b0929	10	7	NC_000913	1	7
OmrA	b2155	949	1042	NC_000913	0	1141
OmrA	b1040	51	21	NC_000913	1	25
OmrA	b3405	71	46	NC_000913	1	52
OmrA	b0565	37	89	NC_000913	0	99

OmrB	b2155	904	431	NC_000913	1	391
OmrB	b1040	27	2	NC_000913	1	2
OmrB	b3405	222	19	NC_000913	1	20
OmrB	b0565	131	629	NC_000913	0	695
OxyS	b2731	954	4014	NC_000913	0	511
RprA	b2741	19	1	NC_000913	1	1
RybB	stm2391	2664	400	NC_003197	1	247
RybB	stm1070	695	101	NC_003197	1	195
RybB	stm2267(stm1530)	247	6	NC_003197	1	6
RybB	stm1572(stm1530)	483	6	NC_003197	1	6
RybB	stm0999	597	597	NC_003197	0	597
RybB	stm1473(stm1530)	59	6	NC_003197	1	6
RybB	stm1995(stm1530)	450	6	NC_003197	1	6
RybB	stm1732	1632	126	NC_003197	1	156
RybB	stm0413	377	796	NC_003197	0	954
RybB	stm0687	671	217	NC_003197	1	266
RyhB	b3607	89	65	NC_000913	1	70
RyhB	b2530	154	117	NC_000913	1	126
RyhB	b0683	2834	1451	NC_000913	1	1642
RyhB	b1981(b4111)	1879	125	NC_000913	1	135
RyhB	b1656	465	62	NC_000913	1	66
SgrS	b1817	1380	410	NC_000913	1	488
SgrS	b1101	4	4	NC_000913	0	5
Spot42	b0757	1	1	NC_000913	0	1
Spot42	b0720	343	6	NC_000913	1	7
Spot42	b4311	38	82	NC_000913	0	121
Spot42	b2702	474	543	NC_000913	0	704
Spot42	b3962	279	1690	NC_000913	0	1711
Spot42	b3566	53	29	NC_000913	1	34
SyR1	sll1578	61	1	NC_000911	1	1
SyR1	slr1655	1	2	NC_000911	0	2

Supplementary table 1: Benchmarking results for 74 experimentally verified 5'UTR interactions. The last column (clip 0.8.pos) shows the results for the Benchmark on the same dataset with preprocessing by elimination of all predictions with p-values < 0.8. 1 and 0 replace yes and no respectively. Locus tags in parentheses are the genes of which the single organism p-values were used for clustering instead of the actual gene (locus tag preceding the parentheses), due to grouping of homologous genes in one organism. (green: improvement into top 50, yellow: dropped out of top 50, blue: beyond rank 200 in IntaRNA and hIntaRNA)

7.3 Random variables between [0,1] – different sample sizes



Supplementary Figure 1: Experiment in R-statistics showing that random variables between [0,1] are approximately uniformly distributed at a sample size of 1000. Sample sizes below 800 show strong deviance from uniformity. The horizontal line shows the theoretically ideal uniform distribution of each respective plot. Samples were created in R-statistics with the runif() command.

7.4 hIntaRNA user manual – Standard operating procedure (SOP)

1. Installing dependencies
    - Install Vienna package (Lorenz at al., 2011)
    - Install IntaRNA (Busch et al., 2008)
    - Install EMBOSS package (Rice et al., 2000)
    - Install Perl with modules (Statistics::R, List::Util, Getopt::Long, Bio::SeqIO, List::MoreUtils, Parallel::ForkManager, Math::Combinatorics)
    - Install R-statistics with package evir (R development core team, 2011, McNeill and Stevenson, 2011)
2. Preparing the data
    - Download the RefSeq files for the organisms the analysis shall run on. The final file names must contain the RefSeq ID (i.e. NC_000000) and cannot contain commas. (i.e. Synechocystis_PCC6803_NC_000911.gb)
    - Create an MBDG cluster table at http://mbgd.genome.ad.jp/ with the organisms in your analysis. The file name must remain cluster.tab.
    - Extract (for example with Artemis genome browser) the 16s-linker-23s region from the RefSeq files and create a fasta with these. The fasta header must only contain RefSeq IDs. (i.e. >NC_000000)
    - Create a fasta with the homologous sRNA sequences. The header must contain the exact sRNA name for each sequence and the affiliated RefSeq ID. (i.e. >SyR1_NC_000000)
    - Avoid RefSeq IDs in the non RefSeq files. (i.e. no match with regular expression _NC_\d{6})
3. Running the program
    - Add all previously created files to a folder with the Perl scripts (14 scripts + csvheader.head).
    - In the Linux terminal, change to the directory containing the files the analysis shall run on.
    - Call the script as follows: perl homology_intaRNA.pl [ncRNA.fa] [#nt upstream start codon] [#nt downstream start codon or stopcodon depending on specification of 5utr or cds respectively] [16s-linker-23s.fa] [5utr/cds] [chromosome1_org1,chromosome2_org1,plasmid1_org1] [chromosome1_org2,plasmid1_org2]

(Concrete example:

perl homology_intaRNA.pl 0680a.fasta 200 100 16S-linker-23S.fasta 5utr Paracoccus_denitrificans_PD1222_chromosome_1_ _NC_008686.gb Rhodobacter_sphaeroides_2.4.1_chromosome_1_NC_007493.gb,Rhodobacter_sphaeroides_2.4.1_chromosome_2_complete_sequence_NC_007494.gb Rhodobacter_sphaeroides_ATCC17025 _genome_NC_009428.gb Sinorhizobium_meliloti_1021_complete_genome_NC_003047.gb Roseobacter_denitrificans_OCh114_complete_genome_NC_008209.gb)

- The organism of interest must be supplied first in the chromosome enumeration.
- The final output can be found in the file ncrnaname_hIntaRNA.csv in the same directory.

4. Common errors
    - Formatting of the input files is wrong.
    - cluster.tab lacks organisms.
    - Not all RefSeq files are supplied in the command line input.
    - Locus tags in cluster.tab and RefSeq files differ (i.e. maybe annoted as old_locus_tag).

7.5 The hIntaRNA output explained

Supplementary Table 2 shows the hIntaRNA output for one organism (NC_003197). Due to space reasons only the column for one organism is displayed. The actual output shows a column for each organism participating in the hIntaRNA prediction. Each line refers to a cluster of homologous genes. The column on the left contains the hIntaRNA final scores for each cluster. These scores are essential for the ranking of predictions. The lower this score is, the better. The organism columns (right column) hold the individual information for each gene. The scheme is as follows:

> locus_tag(gene name|IntaRNA energy score|single prediction p-value|target RNA start|target RNA stop|sRNA start|sRNA stop|Entrez GeneID).

In some cases a cell can be blank. This means that no significant IntaRNA prediction could be made for a target sequence or that no homologs of the specific cluster are present.

hIntaRNA score	NC_003197
6.036877e-04	stm1228(N/A|-18.99440|0.004751815|119|157|1|38|GeneID:1252746)
7.808366e-04	stm1338(pheT|-21.33600|0.0007468356|243|299|1|51|GeneID:1252856)
1.232213e-03	stm0123(murE|-20.79860|0.001187421|185|214|8|36|GeneID:1251641)
1.304160e-03	stm1984(yodD|-20.09570|0.002097539|200|246|2|44|GeneID:1253505)
1.512205e-03	stm2089(rfbJ|-20.31350|0.001765979|42|75|2|37|GeneID:1253610)
1.525463e-03	stm0273(N/A|-20.53670|0.001474757|257|296|6|51|GeneID:1251791)
1.646917e-03	stm3380(accC|-20.86670|0.001121280|110|166|1|54|GeneID:1254903)
1.705872e-03	stm0362(N/A|-20.30350|0.001780130|153|196|2|44|GeneID:1251881)
2.231687e-03	stm1821(yoaA|-21.30740|0.0007660151|198|233|11|40|GeneID:1253340)
2.828964e-03	stm0944(clpS|-19.86400|0.002508946|119|154|13|54|GeneID:1252463)
3.183355e-03	stm4524(hsdS|-16.50220|0.02322836|192|230|2|44|GeneID:1256050)
3.627325e-03	stm1728(yciG|-19.17480|0.00417936|179|208|14|37|GeneID:1253247)
4.422385e-03	stm1772(kdsA|-19.12410|0.004333797|207|261|1|57|GeneID:1253291)
4.627570e-03	stm0550(fimY|-16.39360|0.02471802|127|161|12|39|GeneID:1252070)
4.774965e-03	stm4511(yjiE|-18.98810|0.004772988|214|274|3|53|GeneID:1256037)
5.934468e-03	stm2541(iscA|-18.70010|0.00583341|228|260|10|37|GeneID:1254063)
6.214461e-03	stm0760(aroG|-18.40270|0.007139089|138|185|17|59|GeneID:1252280)
6.355115e-03	stm1883(purT|-16.95560|0.01781538|243|283|1|37|GeneID:1253404)
6.452811e-03	stm3601(N/A|-18.70010|0.00583341|194|238|1|56|GeneID:1255124)
6.530804e-03	stm4272(N/A|-18.55710|0.00643252|158|205|7|51|GeneID:1255798)
7.223460e-03	stm0773(galM|-18.59520|0.006267868|187|244|9|57|GeneID:1252293)
7.641175e-03	stm1714(topA|-18.32310|0.007529144|147|198|2|44|GeneID:1253233)
8.806392e-03	stm0425(thiI|-18.34390|0.007425467|65|106|1|37|GeneID:1251944)
8.896687e-03	stm4460(pyrB|-18.09040|0.0087781|187|210|12|32|GeneID:1255986)

Supplementary Table 2: hIntaRNA table of top 24 predictions for *Salmonella* (NC_003197) InvR sRNA.

8 References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).

Argaman, L. & Altuvia, S. fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *Journal of Molecular Biology* **300**, 1101-1112 (2000).

Ashburner, M. *u. a.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

Aziz, R.K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

Backofen, R. & Hess, W.R. Computational prediction of sRNAs and their targets in bacteria. *RNA Biol* **7**, 33-42 (2010).

Bailey, T.L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48-54 (1998).

Bartel, D.P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215-233 (2009).

Beisel, C.L. & Storz, G. The base-pairing RNA spot 42 participates in a multioutput feedforward loop to help enact catabolite repression in Escherichia coli. *Mol. Cell* **41**, 286-297 (2011).

Benito, Y. *et al.* Probing the structure of RNAIII, the Staphylococcus aureus agr regulatory RNA, and identification of the RNA domain involved in repression of protein A expression. *RNA* **6**, 668-679 (2000).

Bernhart, S.H., Hofacker, I.L. & Stadler, P.F. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22**, 614-615 (2006).

Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R. & Stadler, P.F. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**, 474 (2008).

Breaker, R.R. Riboswitches and the RNA World. *Cold Spring Harbor Perspectives in Biology* (2010).

Busch, A., Richter, A.S. & Backofen, R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **24**, 2849 - 2856 (2008).

Cao, Y. *et al.* sRNATarget: a web server for prediction of bacterial sRNA targets. *Bioinformation* **3**, 364-366 (2009).

Chen, C.-L., Perasso, R., Qu, L.-H. & Amar, L. Exploration of pairing constraints identifies a 9 base-pair core within box C/D snoRNA-rRNA duplexes. *J. Mol. Biol.* **369**, 771-783 (2007).

Darwin, C. *The origin of species*. (London, John Murray: 1859).

Eddy, S.R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2**, 919-929 (2001).

Eggenhofer, F., Tafer, H., Stadler, P.F. & Hofacker, I.L. RNApredator: fast accessibility-based prediction of sRNA targets. *Nucleic Acids Res.* **39**, W149-154 (2011).

Gerdes, S.Y. *et al.* Experimental Determination and System Level Analysis of Essential Genes in Escherichia coli MG1655. *Journal of Bacteriology* **185**, 5673 -5684 (2003).

Gerlach, W. & Giegerich, R. GUUGle: a utility for fast exact matching under RNA complementary rules including G–U base pairing. *Bioinformatics* **22**, 762 -764 (2006).

Gilbert, W. Origin of life: The RNA world. *Nature* **319**, 618 (1986).

Gottesman, S. & Storz, G. Bacterial Small RNA Regulators: Versatile Roles and Rapidly Evolving Variations. *Cold Spring Harbor Perspectives in Biology* (2010).

Gumbel, E.J. *Statistics of extremes*. (Courier Dover Publications: 2004).

Hofacker, I.L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie Chemical Monthly* **125**, 167-188 (1994).

Horler, R.S.P. & Vanderpool, C.K. Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res* **37**, 5465-5476 (2009).

Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

Johansen, J., Rasmussen, A.A., Overgaard, M. & Valentin-Hansen, P. Conserved small non-coding RNAs that belong to the sigmaE regulon: role in down-regulation of outer membrane proteins. *J. Mol. Biol.* **364**, 1-8 (2006).

Jukes, T. & Cantor, C. Evolution of protein molecules. *Mammalian protein metabolism* **III**, 21-132 (1969).

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40**, D109-D114 (2011).

Kang, Z., Wang, X., Li, Y., Wang, Q. & Qi, Q. Small RNA RyhB as a potential tool used for metabolic engineering in Escherichia coli. *Biotechnol. Lett.* **34**, 527-531 (2012).

Larkin, M. a. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947 -2948 (2007).

Lewin, B. *Genes VIII*. (Benjamin Cummings: 2003).

Lewin, B. *Genes IX*. (Jones & Bartlett Publ.: 2007).

Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).

Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T. & Gottesman, S. DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc Natl Acad Sci U S A* **95**, 12462-12467 (1998).

Corcoran, C. P., Papenfort K. & Vogel J. Chapter 2: Hfq-associated regulatory small RNAs. Marchfelder, A. and Hess, W. *Regulatory RNAs in Prokaryotes*, 15-50 (Springer: 2011).

Massé, E. & Gottesman, S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc Natl Acad Sci U S A* **99**, 4620-4625 (2002).

McNeil, A. & Stephenson, A. evir: Extreme Values in R. R package version 1.7-2. http://CRAN.R-project.org/package=evir (2011).

Mitrophanov, A.Y. & Borodovsky, M. Statistical significance in biological sequence analysis. *Briefings in Bioinformatics* **7**, 2 -24 (2006).

Mitschke, J. *et al.* An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803. *Proc Natl Acad Sci U S A* **108**, 2124-2129 (2011).

Mitschke, J., Vioque, A., Haas, F., Hess, W.R. & Muro-Pastor, A.M. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in Anabaena sp. PCC7120. *Proceedings of the National Academy of Sciences* **108**, 20130 - 20135 (2011).

Modi, S.R., Camacho, D.M., Kohanski, M.A., Walker, G.C. & Collins, J.J. Functional characterization of bacterial sRNAs using a network biology approach. *Proceedings of the National Academy of Sciences* **108**, 15522 -15527 (2011).

Møller, T. *et al.* Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol. Cell* **9**, 23-30 (2002).

Mückstein, U. *et al.* Thermodynamics of RNA–RNA binding. *Bioinformatics* **22**, 1177 - 1182 (2006).

Murdoch, D.J., Tsai, Y.-L. & Adcock, J. P-Values are Random Variables. *The American Statistician* **62**, 242-245 (2008).

Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453 (1970).

Negrete, A., Majdalani, N., Phue, J.-N. & Shiloach, J. Reducing acetate excretion from E. coli K-12 by over-expressing the small RNA SgrS. *New Biotechnology* (2011).

Papenfort, K., Bouvier, M., Mika, F., Sharma, C.M. & Vogel, J. Evidence for an autonomous 5' target recognition domain in an Hfq-associated small RNA. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 20435-20440 (2010).

Papenfort, K. *et al.* σE-dependent small RNAs of Salmonella respond to membrane stress by accelerating global omp mRNA decay. *Mol Microbiol* **62**, 1674-1688 (2006).

Peer, A. & Margalit, H. Accessibility and evolutionary conservation mark bacterial small-rna target-binding regions. *J. Bacteriol.* **193**, 1690-1701 (2011).

Pfeiffer, V., Papenfort, K., Lucchini, S., Hinton, J.C.D. & Vogel, J. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nat. Struct. Mol. Biol.* **16**, 840-846 (2009).

Pichon, C. & Felden, B. Small RNA gene identification and mRNA target predictions in bacteria. *Bioinformatics* **24**, 2807 -2813 (2008).

Pulvermacher, S.C., Stauffer, L.T. & Stauffer, G.V. The role of the small regulatory RNA GcvB in GcvB/mRNA posttranscriptional regulation of oppA and dppA in Escherichia coli. *FEMS Microbiol. Lett.* **281**, 42-50 (2008).

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (2011).

Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507-1517 (2004).

Rice, J.B. & Vanderpool, C.K. The small RNA SgrS controls sugar-phosphate accumulation by regulating multiple PTS genes. *Nucleic Acids Res.* **39**, 3806-3819 (2011).

Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

Rice, P.W. & Dahlberg, J.E. A gene between polA and glnA retards growth of Escherichia coli when present in multiple copies: physiological effects of the gene for spot 42 RNA. *J Bacteriol* **152**, 1196-1210 (1982).

Richards, G.R. & Vanderpool, C.K. Molecular call and response: the physiology of bacterial small RNAs. *Biochim. Biophys. Acta* **1809**, 525-531 (2011).

Richter, A.S. & Backofen, R. Accessibility and conservation in bacterial small RNA-mRNA interactions and implications for genome-wide target predictions. *Proceedings of the German Conference on Bioinformatics* **(GCB 2011)**, 2011.

Romby, P., Vandenesch, F. & Wagner, E.G.H. The role of RNAs in the regulation of virulence-gene expression. *Curr. Opin. Microbiol.* **9**, 229-236 (2006).

Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945 (2000).

Sahagan, B.G. & Dahlberg, J.E. A small, unstable RNA molecule of Escherichia coli: spot 42 RNA. I. Nucleotide sequence analysis. *J. Mol. Biol.* **131**, 573-592 (1979).

Sahagan, B.G. & Dahlberg, J.E. A small, unstable RNA molecule of Escherichia coli: spot 42 RNA. II. Accumulation and distribution. *J. Mol. Biol.* **131**, 593-605 (1979).

Schulz, B. Signifikanz von RNA-RNA Interaktionen und von RNA-Sequenz-Strunktur-Slignments, Bachelor Thesis at the University of Freiburg (2009).

Sharma, C.M. *et al.* Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. *Mol. Microbiol.* **81**, 1144-1165 (2011).

Shimoni, Y. *et al.* Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol Syst Biol* **3**, (2007).

Shine, J. & Dalgarno, L. The 3′-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *Proc Natl Acad Sci U S A* **71**, 1342-1346 (1974).

Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197 (1981).

Storz, G., Vogel, J. & Wassarman, K.M. Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Molecular Cell* **43**, 880-891 (2011).

Tafer, H., Amman, F., Eggenhofer, F., Stadler, P.F. & Hofacker, I.L. Fast accessibility-based prediction of RNA–RNA interactions. *Bioinformatics* **27**, 1934 -1940 (2011).

Tafer, H. & Hofacker, I.L. RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics* **24**, 2657 -2663 (2008).

Thompson, K.M., Rhodius, V.A. & Gottesman, S. σE Regulates and Is Regulated by a Small RNA in Escherichia coli. *J Bacteriol* **189**, 4243-4256 (2007).

Tjaden, B. *et al.* Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res* **34**, 2791-2802 (2006).

Towbin, H., Staehelin, T. & Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 4350-4354 (1979).

Tu, K.C., Waters, C.M., Svenningsen, S.L. & Bassler, B.L. A small-RNA-mediated negative feedback loop controls quorum-sensing dynamics in Vibrio harveyi. *Mol. Microbiol.* **70**, 896-907 (2008).

Uchiyama, I., Higuchi, T. & Kawai, M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Research* **38**, D361-D365 (2009).

Udekwu, K.I. *et al.* Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev* **19**, 2355-2366 (2005).

Urban, J.H. & Vogel, J. Translational control and target recognition by Escherichia coli small RNAs in vivo. *Nucleic Acids Res* **35**, 1018-1037 (2007).

Vargas, W.A., Nishi, C.N., Giarrocco, L.E. & Salerno, G.L. Differential roles of alkaline/neutral invertases in Nostoc sp. PCC 7120: Inv-B isoform is essential for diazotrophic growth. *Planta* **233**, 153-162 (2011).

Vogel, J. & Luisi, B.F. Hfq and its constellation of RNA. *Nat Rev Micro* **9**, 578-589 (2011).

Voss, B., Georg, J., Schön, V., Ude, S. & Hess, W.R. Biocomputational prediction of non-coding RNAs in model cyanobacteria. *BMC Genomics* **10**, 123 (2009).

Wadler, C.S. & Vanderpool, C.K. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proceedings of the National Academy of Sciences* **104**, 20454 -20459 (2007).

Wassarman, K.M. 6S RNA: a small RNA regulator of transcription. *Curr. Opin. Microbiol.* **10**, 164-168 (2007).

Waters, L.S. & Storz, G. Regulatory RNAs in bacteria. *Cell* **136**, 615-628 (2009).

Wiegand, G. & Remington, S.J. Citrate Synthase: Structure, Control, and Mechanism. *Annual Review of Biophysics and Biophysical Chemistry* **15**, 97-117 (1986).

Wilms, I., Voss, B., Hess, W.R., Leichert, L.I. & Narberhaus, F. Small RNA-mediated control of the Agrobacterium tumefaciens GABA binding protein. *Molecular Microbiology* **80**, 492-506 (2011).

Yakovchuk, P., Protozanova, E. & Frank-Kamenetskii, M.D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res* **34**, 564-574 (2006).

Zhang, W. *et al.* A gene cluster that regulates both heterocyst differentiation and pattern formation in Anabaena sp. strain PCC 7120. *Mol. Microbiol.* **66**, 1429-1443 (2007).

Zuker, M. Prediction of RNA Secondary Structure by Energy Minimization. *Computer Analysis of Sequence Data* 267-294 (1994).