Universität Leipzig

Fakultät für Mathematik und Informatik
Institut für Informatik

Noncoding RNA Detection Using Comparative Genomics

Diplomarbeit

Aufgabenstellung und Betreuung:	vorgelegt von:
Prof. Peter F. Stadler	Dominic Rose
2. Mai 2006	Studiengang:
	Informatik
	Studienrichtung:
	Bioinformatik

Abstract

Recent research results concede a growing importance of noncoding RNA (ncRNA) to various cellular processes and regulatory functions. The reliable detection of ncRNAs using bioinformatic methods promises to improve the understanding of essential biologic processes and could help to economise time- and cost expensive experiments. A promising approach for the prediction of ncRNAs is the `RNAz` program.

In consideration of current knowledge about RNA and the fact that novel sequenced genomes are available, we started genomewide detection and annotation of structured ncRNAs in Trypanosoma and Leishmania taxa. We predict more than hundred structured ncRNAs among the genomes of *Trypanosoma brucei*, *Leishmania infantum* and *Leishmania major* using the young and promising `RNAz` prediction approach. We demonstrate how to predict ncRNAs genomewidely in automated large-scale analyses using comparative genomics.

Zusammenfassung

Nichtkodierender RNA (ncRNA) wird in jüngsten Forschungsergebnissen eine zunehmende Bedeutung in einer Reihe von zellulären Prozessen und regulatorischen Funktionen zugestanden. Der zuverlässige Nachweis von RNAs mit Hilfe bioinformatischer Methoden verspricht ein besserens Verstandnis grundlegender biologischer Prozesse und könnte helfen, zeit- und kostenintensive Experimente einzusparen. Ein vielversprechender Ansatz fur die Vorhersage von ncRNAs ist das Programm `RNAz`.

Wir haben, unter Berücksichtigung von aktuellem Wissen über RNA und der Tatsache, dass neue Genome als Datenbasis verfügbar sind, mit dem genomweitem Nachweis von ncRNAs in Trypanosomen und Leishmanien begonnen. Wir sind in der Lage über 100 in Struktur faltende RNAs in den Genomen von *Trypanosoma brucei*, *Leishmania infantum* and *Leishmania major* vorauszuberechnen. Wir zeigen mit Hilfe von vergleichender Genomik, wie ncRNAs automatisiert in großem Umfang genomweit vorhergesagt werden können.

Acknowledgment

At the beginning I wish to thank all involved people, who contributed to the success of this work.

First and foremost Peter Stadler, who helped me to achieve an attractive and diversified study with his impressive scientific knowledge and his openhearted nature. His prior confidence in my person enabled me to participate instructively at his working group, which led among other things to this diploma thesis.

My special thank goes to Kristin Missal for listening and answering my huge amount of questions strenuously and her willingness in implementing some of the scipts of our prediction-pipeline, especially the algorithm to combine neighboured `blast` hits. Moreover I thank Stefan Washietl for his exertion in developing `RNAz`, what in the end gave me the chance of writing this thesis. He never hesitated in answering my questions. Thank you for the profitable email correspondences and the fruitful discussions during our common time at the TBI Winterseminars in Bled. I thank both for giving me insights into their materials and the permission to reuse it.

Furthermore I wish to thank all the other coworkers, especially Jens Steuck, who helped me to guarantee the availability of the the computational environment and the utilized databases in general. Thanks for proofreading to Jana Hertel and Andrea Tanzer.

Particularly I am grateful to my parents, who supported my study unselfishly.

Danksagung

Zu Beginn möchte ich allen Beteiligten, die zum Gelingen dieser Arbeit beigetragen haben, meinen herzlichen Dank aussprechen.

Allen voran Peter Stadler, der mir durch seine beeindruckende wissenschaftliche Kompetenz und sein offenes Wesen den Weg zu einem attraktiven und abwechslungsreichen Studium ebnete. Sein mir im Vorfeld entgegengebrachtes Vertrauen ermöglichte mir eine lehrreiche Mitarbeit in seiner Arbeitsgruppe, die unter anderem in dieser Arbeit mündete.

Mein besonderer Dank gilt Kristin Missal, die stets ein offenes Ohr für mich hatte und mir jede meiner (vielen) Fragen unermüdlich beantwortete. Ausserdem programmierte sie einen Teil der Skripte unserer Vorhersage-Prozedur, insbesondere den Algorithmus zum Zusammenfassen benachbarter `blast` Hits. Ich bedanke mich bei Stefan Washietl für seine Arbeit und der Entwicklung von `RNAz`, was mir meine Diplomarbeit überhaupt erst ermöglichte. Er zögerte nie mir meine Fragen zu beantworten. Vielen Dank fuer die gewinnbringende email Korrespondenz und die förderlichen Diskussionen während unserer gemeinsamen Zeit in Bled bei den TBI Winterseminaren. Vielen Dank an Euch beide, dass ihr mir Einblicke in Eure Materialien gewährt habt und mir die Erlaubnis gegeben habt, sie weiter zu benutzen.

Zudem möchte ich mich bei allen weiteren Kolleginnen und Kollegen bedanken, besonders bei Jens

Steuck, mit dessen Hilfe eine hohe Verfügbarkeit der zum Einsatz gekommenen Datenbanken und der EDV gewährleistet werden konnte. Für das Korrekturlesen bedanke ich mich bei Jana Hertel und Andrea Tanzer.

Mein besonderer Dank gilt meinen Eltern, die mein Studium selbstlos unterstützten.

Abbreviations

These shortenings are used in the document:

DNA	**D**esoxyribo**n**ucleic **a**cid
RNA	**R**ibo**n**ucleic **a**cid
ncRNA	**N**on**c**oding **RNA**
mRNA	**M**essenger **RNA**
tRNA	**T**ransfer **RNA**
rRNA	**R**ibosomal **RNA**
snRNA	**S**mall **n**uclear **RNA**
snoRNA	**S**mall **n**ucle**o**lar **RNA**
slRNA	**S**pliced **l**eader **RNA**
gRNA	**G**uide **RNA**
miRNA	**Mi**cro **RNA**
tmRNA	**RNA** with dual **t**RNA-like and **m**RNA-like character
Tb	**T**rypanosoma **b**rucei
Tc	**T**rypanosoma **c**ongolense
Tv	**T**rypanosoma **v**ivax
Li	**L**eishmania **i**nfantum
Lm	**L**eishmania **m**ajor
nt	**N**ucleo**t**ide(s)
SCI	**S**tructure **c**onservation **i**ndex
SMN	**S**urvival Of **M**oto**n**euron
CM	**C**ovariance **m**odel
BC	**B**last**c**lust
SRP	**Single Recognition Particle**
LSU	**L**arge **S**ub**u**nit
SSU	**S**mall **S**ub**u**nit

Contents

Chapter 1

Introduction

1.1 Motivation

The central dogma of molecular biology postulates that DNA makes RNA makes protein (protein biosynthesis). Genes encode proteins and proteins are agents of cellular activity. This led to a common opinion that RNA mainly functions in gene expression, reducing RNA to its role as mRNA, tRNA and rRNA in the past. This point of view is not wrong but it underestimates the abilities RNA provides and thus became obsolete. Various research results of the last years concede RNA a more complex role in the cell and a growing importance in general[1],[2],[3]. A variety of transcripts were discovered that are not translated into protein, but instead are processed upon transcription leading to their actual products. Members of this class of molecules are called noncoding RNAs (ncRNAs) in the sense of "non-protein-coding". Many of these ncRNAs are believed to act as an additional layer controlling several cellular processes. The different groups of ncRNAs are classified by their function, e.g. small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs) or the relatively young micro RNAs (miRNAs, shortly introduced at section 2.1). However, before speculating about an RNA's function, it has to be identified and localized in a genomic sequence.

The determination of whole genome sequences is one of the most fascinating goals in science. A major step of genomics was the initial working draft sequence of the human genome published in 2001[4],[5]. Up to now sequencing techniques were improved and thus more and more genomes of a variety of taxa become available. The increase of high quality sequences offers new bioinformatic challenges for identification and characterization of functional elements in these datasets. The search for functional elements includes annotation of protein coding genes as well as non-coding genes, elements with influence on gene regulation or chromosomal structure, stability and dynamics and additionally all kind of undescribed elements. Unlike protein coding genes, ncRNA sequences do not exhibit a strong *common* statistical signal that separates them significantly from their genomic environment. For example they have no start or stop codon and they lack open reading frames, CpG islands or typical splicing signals. Individual families of ncRNAs exhibit evolutionarily very well-conserved secondary structures.

There are common experimental methods and also new computational approaches for detection of the elements mentioned above, but in general three main ways of discovering new noncoding RNA genes seem to be promising[1]. The first one uses specific amplification and cloning strategies to enrich ncRNAs. Here the spectrum reaches from simple cloning and sequencing of small RNAs in total RNA extracts[6] to immunoprecipitation with antibodies against proteins associated with specific ncRNA families[7] in order to localise ncRNAs within their subcellular compartements[8]. The second method uses microarray technology for probing entire genomes systematically[9]. The third strategy could be described as computational comparative genome analysis. With the help of fast and reliable algorithms ncRNA candidate detection could be done computational without time and cost expensive laboratory experiments. It seems to be a very fruitful way of identifying ncRNAs as some trailblazing screens with different genomes have successfully been performed[10],[11],[12],[13].

The structural conservation of ncRNAs can be understood as a consequence of stabilizing selection acting (predominantly) on the secondary structure whereas their sequences are often highly variable. This results in a substitution pattern that can be utilized to design a general-purpose RNA genefinder based on comparative genomics. This idea was first implemented in the tool QRNA[14], which is based on an SCFG (stochastic context free grammar) method to asses the probability that a pair of aligned sequences evolves under the constraint of preserving a secondary structure. RNAs that are under long-time selection for secondary structure can be expected to have sequences that are more resilient against mutations[15, 16], which in turn correlates with increased thermodynamic stability of the folded RNA. Indeed, it has been observed that functional RNAs are more stable than the structures formed by randomized sequences[17, 18, 19]. The program RNAz[20] combines both approaches. It uses a z-score measuring thermodynamic stability of individual sequences and a *structure conservation index* (SCI) obtained by comparing the folding energies of the individual sequences and the energy of the predicted consensus folding. Both values measure different aspects of stabilizing selection preserving RNA structure.

1.2 Subject of this thesis

This work is based on two projects I participated in prior to my diploma thesis, where we successfully detected ncRNAs in *Ciona intestinalis*[21] and *Caenorhabditits elegans*[22]. In this thesis I will reuse material already published in those two articles without further citation.

The core tool of the whole project, RNAz, allows fast and reliable prediction of ncRNAs[20]. Due to the fact that RNAz is a new tool, only a few surveys for RNAz based ncRNA detection have been made. In an exemplary study thousands of functional ncRNAs could be retrieved in human[23]. In order to extend our knowledge about ncRNAs to more basal organisms, we decided to investigate taxonomic families further down the root of life. One of the pre-requisites for comparative genome analysis using RNAz is a set of relatively closely related species, which nevertheless show enough sequence variation to yield reliable signals for conserved structures. Another aspect is the biological relevance. In contrast to metazoans and higher plants, the main targets of ncRNA research to date, little is known about functional RNAs in protista. Recent sequencing initiatives revealed members of fully sequenced genomes of *Trypanosoma sp.* and *Leishmania sp.*. Thus, we decided to screen these parasites for ncRNAs. We expect that RNAz prediction is possible among every taxon offering new ways of fast and reliable ncRNA detection.

Inspired by the idea of identifying functional elements computational, this thesis concentrates on ncRNA prediction by comparative genome analysis using modern ncRNA detection algorithms and annotation methods. The main goals are (i) to detect and to deliver a set of putative ncRNAs in the recently sequenced Trypanosoma and Leishmania genomes and (ii) to discuss the quality of the prediction. We expect to identify known ncRNA genes as well as novel (unseen) candidate ncRNAs.

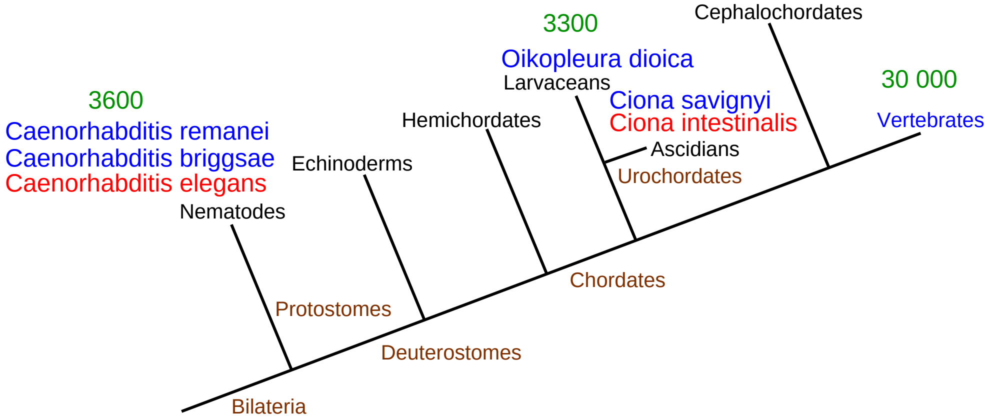


Figure 1.1: Overview on prior RNAz based ncRNA prediction screens. Vertebrates (~30,000 ncRNA candidates), nematodes (~3,600 ncRNA candidates) and urochordates (~3,300 ncRNA candidates) are covered. But how does the situation of ncRNA prediction looks like at less sophisticated organisms and their genomes nearby at the left side respectively the root of the phylogenetic tree?

Chapter 2

Background

2.1 Noncoding RNA variety

The modern ncRNA world consists of ncRNAs acting in well-adapted specialized biological processes, including translocation, transcriptional regulation, chromosome replication, RNA processing and modification, mRNA stability, protein translation and degradation[2].

Table 2.1 provides an overview on major biological processes affected by ncRNAs. It is only a selection to illustrate the variety of the ncRNA world, more details can be obtained from the `Noncode`[1] database[24]. In general `Noncode` contains thousands of sequences from hundreds of organisms covering all kingdoms of life. Some ncRNAs are induced or repressed by stress others are specific to diseases, imprinted domains, sex, tissue, or developmental stage[25]. The better we characterise ncRNA loci and the biological function of their RNA product the better we are able to understand the organism.

2.2 Objects of research - The organisms

Due to the fact that the underlying tools of this work are new and that there are only a few successful approaches in managing genomewide `RNAz` based ncRNA annotation (cp. section 1.2), we considered to analyse some of the most recently sequenced genomes from different species in an automated way. Table 2.2 presents a phylogenetic overview of the screened organisms. The aim of the next subsections is a brief introduction of those organisms. Overall, this thesis deals with *Trypanosoma brucei* (Tb), *Trypanosoma congolense* (Tc), *Trypanosoma vivax* (Tv), *Leishmania major* (Lm) and *Leishmania infestans* (Li).

[1] `http://noncode.bioinfo.org.cn/`

Classification	**Process**	**Example**	**Function**
6S RNA	Transcription	184 nt *E. coli* 6S	Involved in stationary phase regulation of transcription by the sigma70-holoenzyme, modulates promotor use.
XIST	Gene silencing	16,500 nt human *Xist*	Required for X chromosome inactivation.
Telomerase RNA	Replication	451 nt human telomerase RNA	A RNA component of ribonucleoprotein reverse transcriptase that synthesises telomeric DNA, core of telomerase and telomere template.
snRNA	RNA processing	186 nt human U2 snRNA	Small RNA molecules that are found within the nucleus of eukaryotic cells, involved in a variety of important processes such as RNA splicing, forming the core of the spliceosome.
RNase P	RNA processing	377 nt *E. coli* RNase P	Catalytic core of RNase P.
snoRNA	RNA modification	102 nt *S. cerevisiae* U18 C/D snoRNA	A class of small RNA molecules found within the nucleolus, involved in chemical modifications of rRNAs and other RNA genes by methylation or pseudouridylation, the example directs the 2'-O-ribose methylation of target rRNA.
gRNA	RNA modification	68 nt *T. brucei* gCYb gRNA	Function in RNA editing that has been found only in the mitochondria of kinetoplastids in which mRNAs are edited by inserting or deleting stretches of uridylates.
miRNA	mRNA translation	22 nt *C. elegans* lin-4 miRNA	Represses translation by pairing with 3' end of target mRNA
tmRNA	Protein stability	363 nt *E. coli* tmRNA	Directs addition of tag to peptides on stalled ribosomes
4.5S RNA	Protein translocation	114 nt *E. coli* 4.5S RNA	Integral component of signal recognition particle central to protein translocation across membranes

Table 2.1: Overview of ncRNA classes and their activities[2],[24]

PHYLUM	Euglenozoa	
CLASS	Zoomastigophora (Flagellata)	
ORDER	Kinetoplastida	
FAMILY	Trypanosomatidae	
GENUS	Trypanosoma	Leishmania
SPECIES	*Trypanosoma brucei*	*Leishmania major*
	Trypanosoma congolense	*Leishmania infantum*
	Trypanosoma vivax	

Table 2.2: Taxonomic overview of used Trypanosoma and Leishmania

2.2.1 Trypanosoma

Trypanosomes are unicellular, flagellated protozoan organisms of the genus Trypanosoma, which is part of the order Kinetoplastida. This order is characterised by the presence of one flagellum and a single mitochondrion containing the kinetoplast, a specialized DNA containing organelle. Acting as a parasite, they affect vertebrates, invertebrates or plants. Some are potentially pathogenic, causing disease in humans and their domestic animals. Amongst these are Tc, causative agent of Chagas' disease in South America or Tb causing African Trypanosomiasis (also known as Sleeping sickness). Transmission of the human-infective trypanosome is via blood feeding insects. Throughout its life cycle, the parasite alternates between development in mammalian tissue fluids and bloodstream as well as growth in its vector's (the Tsetse fly) midgut and salivary gland. Development is accompanied by changes in morphology, biochemistry, cell cycle stage and expression of major surface markers[26]. Figure 2.1 illustrates the common trypanosome Tb.

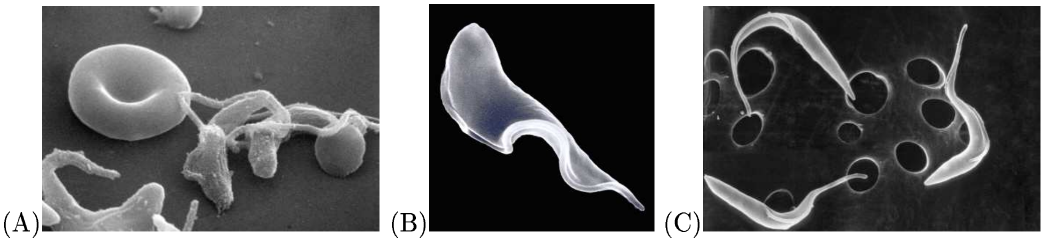


Figure 2.1: Electron micrographs of *T. brucei* (A) trypomastigotes, (B) short stumpy form and (C) in a bloodstream[27]

Untreated Human African Trypanosomiasis, caused by the subspecies *T. brucei gambiense* and *T. brucei rhodesiense*, fatally threatens over 60 million people with 500.000 deaths a year. Therefore expectations on the benefit of the trypanosomatid genome-sequencing projects are considerably high with the intent to improve our knowledge of parasite biology stepwise. Todays genomic research concentrates on aspects of these organism concerning cell differentiation, changes in metabolism or cell cycle control[27]. RNA analyses are somehow underrepresented and thus, we decided to use the Tb, Tc and Tv genomes for ncRNA annotation.

2.2.2 Leishmania

Leishmania are intracellular protozoic parasites propagating in white blood cells, especially macrophages and dendritic cells. Therefore, they use various mechanisms to foil mammalian humoral and cellular immunresponse[28]. They are flagellates belonging to the order of the Kinetoplastida and the family Trypanosomatidae. At least 20 species of Leishmania are recognised.

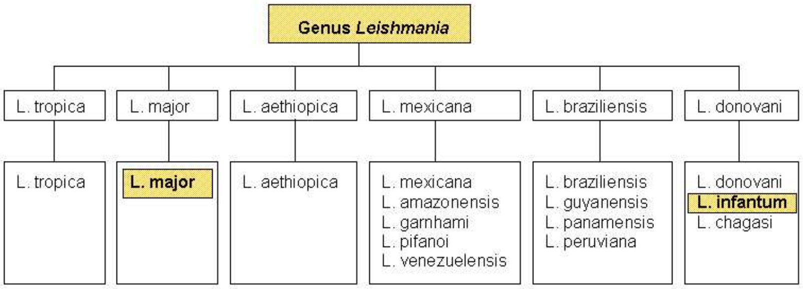


Figure 2.2: Pathogenic complexes and species of Leishmania

Their life cycle involves a vertebrate host (e.g. human) and a vector (a sand fly) that transmits the parasite between vertebrate hosts. They reproduce asexually in the gut of the vector and form a characteristic morphological state, called the promastigote. Promastigotes are injected into the vertebrate host during the bite of the vector. While entering the vertebrate cells they change into a form of life called the amastigote and start to reproduce. Eventually the host's cell dies and the amastigotes are released and get the ability to infect other cells. This leads to a disease called leishmaniasis[29].

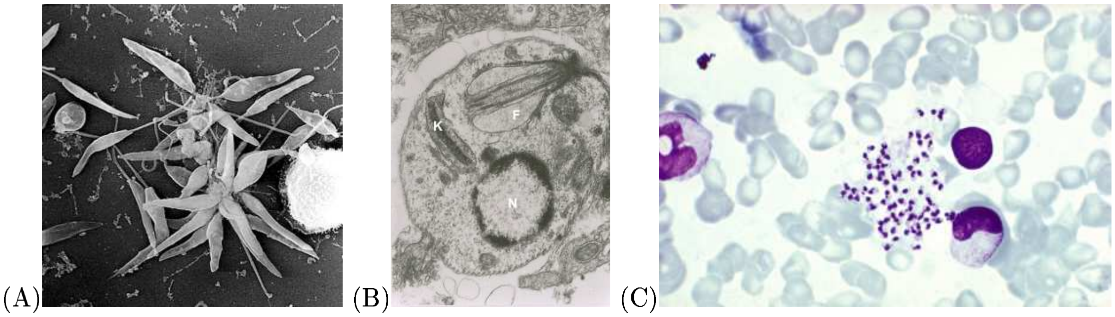


Figure 2.3: Electron micrographs of (A) Leishmania promastigotes, (B) Leishmania amastigotes with labeled nucleus(N), kinetoplast(K) and flagellum(F) and (C) Leishmania donovani infantum in a bone marrow smear of an Old World visceral leishmaniasis patient[27]

Leishmaniasis threatens 350 million people in 88 countries of the world. In general the disease is divided into two groups based on geographical distribution: Old World leishmaniasis threatens Asia, Africa and some mediterranean countries (66 nations), the New World leishmaniasis pertains Central and South America (22 countries). The disease subclassifies into visceral (VL) and cutaneous (CL) forms of leishmaniasis. Among others genomic research could potentially enable the identification of elementary chromosome loci controlling the susceptibility to the disease, new drug targets or the development of effective vaccines[27].

Out of the various species of Leishmania, the genome sequencing projects of Lm (which causes Old World CL) and Li (which causes Old World VL) are high in progress and so we decided to use them for further analyses.

Genome projects represent one of the major ressources for producing data to supplement our fundamental knowledge on related organisms. However, translation of this knowledge into an understanding of specific living beings biology takes time and care. The next subsections provide an insight on the genomes of the screened organisms and the data sources of this project.

2.3 Objects of research - The genomes

2.3.1 Trypanosoma

All trypanosomatid genomic sequences were retrieved from the public accessible ftp server of the Wellcome Trust Sanger Institute[2]. The Tb genome[3] was available in version 4 from July 2005. Both, the Tc[4] and the Tv[5] genomes, were from November 2004.

The Tb genome consists of 11 chromosomes, partitioned to 15 files and 27,736,938 nt in sum. The Tc genome was obtained in 4,676 contigs, containing 33,385,363 nt. The genomic sequence of Tv was structured to 7,366 contigs with 44,233,297 nt. Out of the three trypanosomatid species, Tb was the only one with a given annotation in EMBL file format[6], summarized at table 2.3. These datasets can be accessed and visualised with `Artemis`[7].

[2] `http://www.sanger.ac.uk/`
[3] `ftp://ftp.sanger.ac.uk/pub/databases/T.brucei_sequences/T.brucei_genome_v4/`
[4] `ftp://ftp.sanger.ac.uk/pub/databases/T.congolense_sequences/`
[5] `ftp://ftp.sanger.ac.uk/pub/databases/T.vivax_sequences/`
[6] `http://www.ncbi.nlm.nih.gov/collab/FT/`
[7] `http://www.sanger.ac.uk/Software/Artemis/`

Element	#	avg(len)
CDS	10,114	1,459
repeat_region	8,485	155
misc_feature	4,007	422
repeat_unit	1,658	73
variation	963	3
sig_peptide	372	73
snoRNA	353	83
rRNA	106	688
tRNA	65	73
misc_RNA	29	132
gap	29	100
3'UTR	9	294
5'UTR	6	91
snRNA	6	143
gene	2	1122
promoter	2	182
intron	1	330

Table 2.3: Statistical overview of the initially given *T. brucei* annotation (cp. figure 2.4); Several coding and repetitive elements are known, they can be used to define regions of noncoding DNA by excluding their loci from the underlying genome; Given RNAs can be used to validate the quality of the resulting nc predictions, hopefully we detect most of them.

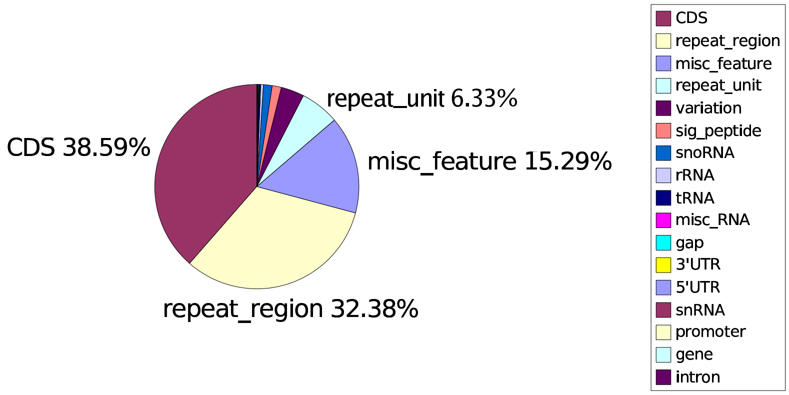


Figure 2.4: Graphical overview of the initially given *T. brucei* annotation (cp. table 2.3); The major part of the given annotation are coding and repetitive elements.

2.3.2 Leishmania

The Li genome[8] was downloaded in version 2.0 and the Lm genome[9] in version 5.2. Both were available in one single file. The genomic sequence of Li exists of 34,744,916 nt in sum and the sequence of Lm has a total length of 32,810,825 nt. Annotation data was given for both, similar to Tb (EMBL file format). Tables 2.4, 2.5 and figures 2.5, figure 2.6 describe the initial distribution of the given Leishmania annotations.

Element	#	**avg(len)**
misc_feature	17,921	210
CDS	8,173	1,850
repeat_region	3,655	208
rRNA	62	602
tRNA	62	73
snoRNA	43	97
misc_RNA	14	2,642
snRNA	7	123
repeat_unit	3	1153

Table 2.4: Statistical overview of the initially given *L. infantum* annotation (cp. figure 2.5); Less coding sequences and repeats are known for Li in comparison to the Tb annotation.

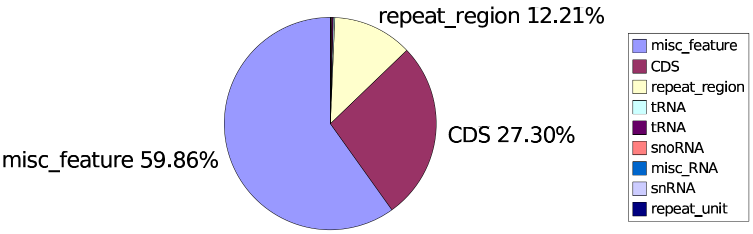


Figure 2.5: Graphical overview of the initially given *L. infantum* annotation (cp. table 2.4)

[8] ftp://ftp.sanger.ac.uk/pub/pathogens/L_infantum/
[9] ftp://ftp.sanger.ac.uk/pub/databases/L.major_sequences/

Element	**#**	**avg(len)**
misc_feature	19,909	291
repeat_region	16,389	95
CDS	8,311	1,899
snoRNA	693	88
tRNA	83	74
rRNA	63	637
misc_RNA	61	446
snRNA	7	101
repeat_unit	3	1,212
miscrecomb	1	7,615

Table 2.5: Statistical overview of the initially given *L. major* annotation (cp. figure 2.6); The Lm genome comprises considerably more repeats than Li.

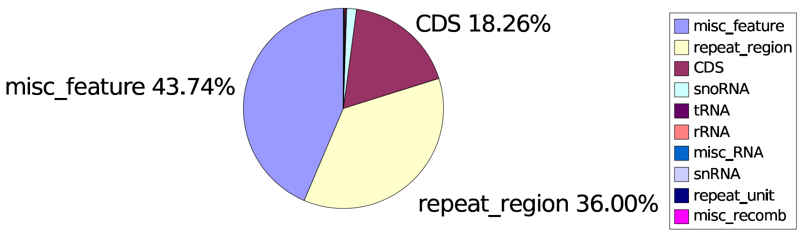


Figure 2.6: Graphical overview of the initially given *L. major* annotation (cp. table 2.5)

Chapter 3

Methods

Automation is of utmost importance while handling data sets covering complete genomes. The purpose of this chapter is to introduce the methodical steps that lead to ncRNA predictions out of blank sequence data.

3.1 Basic ideas and screen design

Each screen is based on up to three organisms. There is a target genome in which we try to annotate candidate ncRNAs primarily and other genomes which are used to identify conserved and thus putative functional elements. A simplified view of our procedure for retrieving ncRNA predictions is illustrated in figure 3.1. Details of each step, implemented algorithms and used tools follow in the upcoming subsections. An overview of our performed screens is given in table 3.1. Although the `RNAz` program is able to handle six-way alignments we choose three as an upper bound of aligned sequences because of coding effort, overall runtime complexity and the observation that it is possible to get significant noncoding signals with three genomic sequences per alignment only[21],[22].

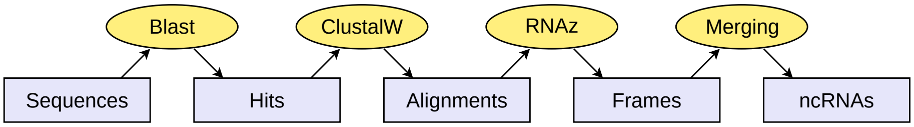


Figure 3.1: Very simplified view on the ncRNA prediction pipeline

3.1.1 Genome-wide alignments of noncoding DNA

At the beginning, we have to calculate nc regions, for they were not given with the initial annotation files explicitly. Due to the specific annotation formats of our target organisms (cp. subsections 2.3.1 and 2.3.2) we start with collections of all contiguous regions that are not annotated as either

Order	Screen description	Alignment
Kinetoplastida	**Tb**Tc	pairwise
	TbTcTv	three-way
	TbTcLi	three-way
	TbTcLm	three-way
	LiLm	pairwise
	LiLmTb	three-way
	LmLi	pairwise
	LmLiTb	three-way

Table 3.1: Overview of the performed screens. The two-letter code abbreviating the organisms names indicates the order of setting up pairwise and multiple alignments. Bolded letters indicate the target genomes. Screen design is due to the given amount of available data including genomic sequences and annotation sheets and the phylogenetic relation of the organisms.

"CDS", "repeat_region" or "repeat_unit" for the trypanosomatid genomes.

For each noncoding DNA interval within the target genome, we determine potentially homologous regions with the corresponding subject genome by pairwise `blast`[30] searches using an Expect value of $E < 10^{-3}$. This threshold is pretty low, though we want to ensure that nothing is missed. Regions separated by short distances ($\leq$ 30 nt) only are combined given the alignments pass the consistency checks outlined below.

Structured RNAs are less conserved in regions without base pair interactions, which might prevent `blast` from extending the sequence alignment into such regions. In order to ensure that a global alignment constitutes a complete ncRNA gene, `blast` hits with short distances between them are combined. But due to rearrangement, deletion, and duplication events during evolution, not all local alignments lead to a consistent global alignment. We therefore employed the following algorithm:

A global alignment is inconsistent if at least one region of sequence A is conserved with at least two regions of sequence B (duplication or deletion) or if at least two distinct regions of sequence A are conserved in different order in sequence B (rearrangement), cp. figure 3.2. It is useful to construct a graph G_S in the following way: Local alignments are the vertices, and there is an edge between two vertices if the distance of the corresponding alignments is less than a certain threshold value $\ell = 30\,nt$. Thus the connected components of G_S comprise sets of alignments with pairwise short distance; amon them, all combinations of consistent, global alignments have to be determined. In short, the first step checks whether each pair x and y of local alignments are consistent, in the sense that they can be derived from the same global alignment. Two further auxiliary graphs G_C and G_I store this consistency information. If x and y are consistent an edge in G_C is introduced, otherwise an edge in G_I is added between x and y. Finally, the graph G_F is constructed by inserting edges between the two nodes x and y if at least one path between x and y exists in G_C which does not contain pairs of nodes that are inconsistent, i.e., connected by an edge in G_I. Complete subgraphs of G_F correspond to local alignments which can be combined to a consistent global alignment. Only maximal local alignments, i.e., the maximal cliques of G_F, are of interest for our purpose. They can be computed efficiently e.g. by the program `cliquer`[31]. We remark that this approach is similar in spirit to the consistency checking algorithm implemented in the `tracker` algorithm for phylogenetic footprinting[32]. Figure 3.3 illustrates a simple example of checking three hypothetical `blast` hits.

These combined pairwise `blast` hits form the basis for a second `blast` search where we retrieve conserved regions within a third organism. The sequences should be aligned in order of their phylogenetic distance to avoid too restrictive searches, which would result in less conservation and therefore loss of signal. Hits obtained by the second search are treated with the same consistency checks as used in the first search. Global alignments of the resulting regions are then computed using `clustalw`[33] to improve the alignment quality. Alignments of both reading directions are produced. Finally, columns with gaps flanking the sequences are removed.

We tried to align the genomes in the order of their phylogenetic relation, starting with the closest ones and then adding the third. But those relations mostly are hard to enlighten and phylogenetic trees are often questionable. For the trypanosomes we found bootstrapped maximum parsimony, minimum evolution and quartet maximum likelihood 18S rRNA based trees[34],[35] supposing that *T. brucei* is closer to *T. congolense* than it is to *T. vivax.* For the two Leishmania genomes we simply performed two screens because both have a given annotation.

3.1.2 Noncoding RNA prediction using `RNAz`

The `clustalw` alignments described above are screened with `RNAz`[20] (version 0.1.1) to detect regions that are additionally conserved on the level of RNA secondary structure. Due to computational limitations and restrictions in the training set of the support vector machine (SVM) implemented in the `RNAz` program, alignments were scanned by moving a window of length 120 in steps of 50 nt. An SVM is a tool that first needs to be trained with positive and negative datasets and is then able to decide of an input data whether it belongs to the positive or negative group. In our case the decision is expressed in terms of the RNA classification probability $p > p_c$. Alignments with a total length smaller than 120 nt are screened directly with `RNAz`. We only scanned alignments of at least 40 nt length, for most known ncRNA families use not to be shorter. The `RNAz` algorithm evaluates the thermodynamic stability of RNA secondary structures (relative to an ensemble of randomized sequences) and quantifies the evidence for stabilizing selection by comparing the energy of a consensus structure with the ground-state energies of the individual structures. `RNAz` performs the classification by means of a support vector machine. The classification is based upon the folding energy z-score and the structure conservation index (SCI). The length and sequence divergence of the alignment and the number of aligned sequences (N) are used to normalise those descriptors.

$$\text{SCI} = \frac{\text{consensus MFE}}{\text{mean single sequence MFE}} \tag{3.1}$$

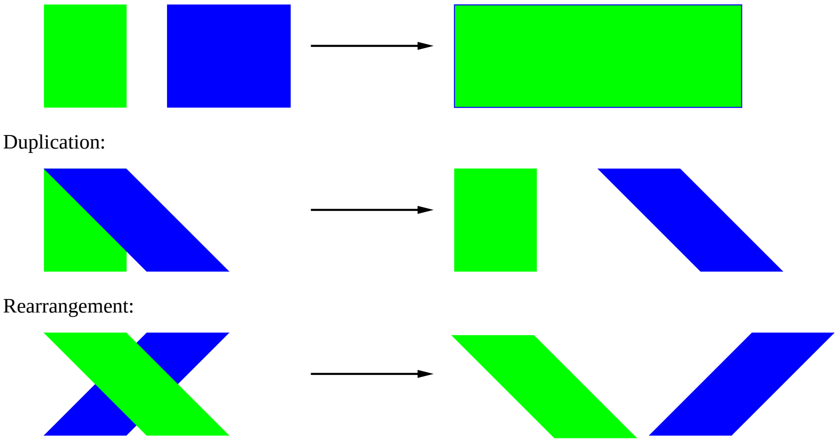


Figure 3.2: Local pairwise alignments will lead to an inconsistent global alignment in case of duplication, deletion or rearrangement events. They are combined to a global alignment only if they are consistent, otherwise there will be one alignment for each region.

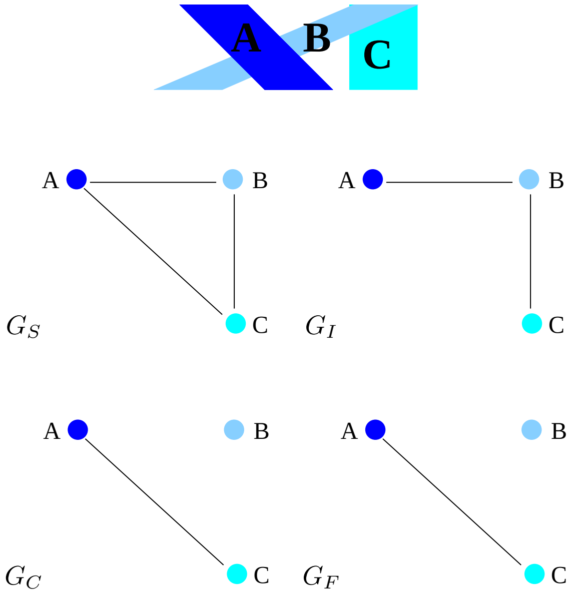


Figure 3.3: A simple example of consistency checks validating three `blast` hits A, B, C.
G_S: Edge between two vertices (local alignments) if they have a distance ≤ 30 nt $\rightarrow$ find all combinations of consistent global alignments
G_C: Edges between consistent pairs of local alignments
G_I: Edges between inconsistent pairs of local alignments
G_F: Edge between x and y if at least one path from x to y exists in G_C which does not contain vertices connected in G_I
Cliques in G_F are local alignments which can be combined to a consistent global alignment, in the example there is one single trivial clique, thus A and C are consistent and can be combined.

$$\text{z-score} = \frac{\sum \text{single sequence z-score}}{N} \qquad (3.2)$$

The significance of the classification is quantified as an RNA classification probability p. A value of $p > 0.5$ classifies the alignment as noncoding RNA with low significance, whereas $p > 0.9$ indicates a high significance for structured and thus functional RNA. For each global alignment, all possible reading directions are considered, because the classification of `RNAz` is based on the thermodynamic stability of the potentially transcribed RNA, which is inherently direction dependent.

For each alignment all overlapping frames classified with $p > 0.5$ are merged together. For some loci we obtain more than one alignment for the same query organisms sequence. This does not constitute a problem for the ncRNA detection, since we obtain essentially identical alignments with different paralogs. Two different alignments of the same reading direction were merged onto the same genomic loci if they overlap at least 90 % in a query genome. All such genomic regions are combined again if they overlap at least 90 % independent of the reading direction of their alignments. Putative ncRNA clusters in close vicinity might still cover a genomic region more than once. Of all merged regions that overlap more than 20 % all except one are discarded,leaving us with a unique genomic locus for each ncRNA gene. For each locus we choose the alignment with the maximal `RNAz` classification probability and the maximal length as the best representative. Hence, for all statistics reported below, each genomic location is represented in at most one structured RNA candidate. In the end we can provide an almost distinct and unique ncRNA annotation of loci where we observe overall noncoding signals ('adjusted region') and their best representative obtained by a single alignment ('best RNA') for each of the combined regions. It seems to be wrong to merge loci of different alignments because we would flout their biological context; the start and end position of a unique ncRNA may become unclear and we can not speak of conserved elements with a specific biological function.

3.1.3 Specificity, sensitivity and false positive rates

In order to estimate the specificity and the false positive rate of `RNAz`, we repeat each screen with shuffled input alignments. All alignments are shuffled with the `Perl` script `shuffle-aln.pl` published by Washietl *et al.* for randomization of multiple sequence alignments and the destruction of native secondary structures[18].

The specificity in terms of individual `RNAz` scanning windows is defined as

$$\text{specificity} := \frac{\text{number of shuffled scanning windows with } p \leq p_c}{\text{number of shuffled scanning windows}} \quad . \qquad (3.3)$$

In order to estimate the sensitivity of a screen we compare our ncRNA predictions to the initially given annotation data. We annotate a putative ncRNA candidate of our screen as known if its genomic locus overlaps at least 70 % with an ncRNA, that is already annotated in the corresponding organism, leading to the following definition of sensitivity:

$$s_{Ng} := \frac{N}{N_g} \quad . \tag{3.4}$$

Here N is the number of unique genomic loci, classified by RNAz, that can be identified as a known member of a specific ncRNA family (required overlap of 70 %) and N_g is the entire number of ncRNAs of this family in the genome. The sensitivity of a screen largely depends on the number of ncRNAs which have a conserved primary structure between the organisms. To state how many known ncRNAs can be detected by our screen in principle we also report the sensitivity of our alignment procedure defined as

$$s_{Na} := \frac{N}{N_a} \quad , \tag{3.5}$$

where N_a is the minimal unique number of sequence-conserved motifs in the alignment input set, identified by blast ($E < 10^{-3}$) and blat (default parameters). Additionally to the conservative approach of comparing positions we ask for common motifs of known ncRNAs and our ncRNA predictions. This defines sensitivity values s_{na} and s_{ng} as

$$s_{na} := \frac{n}{N_a} \tag{3.6}$$

and

$$s_{ng} := \frac{n}{N_g} \tag{3.7}$$

where n is the minimal number of unique hits of both tools obtained by individual searches of our ncRNA candidates against nucleotide blast databases of the given ncRNA sequences. Furthermore we try to define a measuring value in spirit of a “false positive rate” as

$$\text{false positive rate} := \frac{\text{number of shuffled scanning windows with } p > p_c}{\text{number of original scanning windows with } p > p_c} \quad . \tag{3.8}$$

3.1.4 Annotation of predicted ncRNAs

To interpret our own results, the large number of ncRNA candidates needs to be verified. This is either done by comparison of our hits with known ncRNAs given by external databases or by third party ncRNA detection tools. We compare our hits by local blast searches against known sequences from the Noncode[24] (Release 1.0), the Rfam[36] (Version 7.0, March 2005), the miRBase[37] (Release 8.0, February 2006) and the snoRNA-LBME[38] (Version 2) databases. These blast searches are performed with an E-value of $E < 10^{-10}$. Another possibility is to perform blast searches against sequences provided by the NCBI[1] databases. But we do not expect to obtain significant blast alignments with short query sequences like miRNAs. Specific ncRNA detection tools are tRNAscan-SE[39] for identification of tRNAs, Infernal/cmsearch to align our predictions with Rfam covariance models and RNAmicro for miRNA detection[40]. RNAmicro was called with window sizes of 70, 100 and 130 nt. Infernal/cmsearch produces many hits with low scores, thus a bitscore cutoff value is set at 10 (cp. the bitscore distributions 4.2, 4.7, 4.12, 4.17,

[1] http://www.ncbi.nlm.nih.gov/

4.23, 4.28) for all screens. By default we ignore hits with senseless models and only keep hits of biological mean (cp. table 3.2).

Accepted CM	**Ignored CM**
5S ribosomal RNA	Antizyme RNA frameshifting stimulation element
ctRNA	Cardiovirus cis-acting replication element (CRE)
Pyrococcus C/D box small nucleolar RNA	Coronavirus 3' stem-loop II-like motif (s2m)
Small nucleolar RNA U54	Hammerhead ribozyme (type I)
tRNA	Hepatitis C stem-loop IV
U2 spliceosomal RNA	Histone 3' UTR stem-loop
U6 spliceosomal RNA	Infectious bronchitis virus D-RNA
U7 small nuclear RNA	Iron response element
UPSK RNA	Renin stability regulatory element (REN-SRE)
	S-element
	Tymovirus/Pomovirus tRNA-like 3' UTR element
	UnaL2 line 3' element

Table 3.2: Overview of accepted and ignored `Rfam` covariance models. Only model descriptions with a hit during the various `Infernal` runs are listed. Maybe the list of accepted models should be handled more restrictive. Hits of accepted models with a bitscore higher than 10 are kept. A complete list of `Rfam` RNA families is available at `http://www.sanger.ac.uk/Software/Rfam/browse/old_index.shtml`.

`RNAmicro` is under continuous development at our group and works in spirit of `RNAz`. Similar to `RNAz` it uses a trained SVM to detect miRNA precursor sequences. We try to recognise SMN binding sites and critical sequence features involved in forming the SMN complex (survival of motonen) and the regarding assembly of an Sm core[41]. The RNA motif and pattern searcher `RNAbob`[2] is used for this purpose. We scan our predictions for matches of combinations of stem-loops and Sm sites derived from the motifs shown in figure 3.5. Explicitly we searched after the Sm sites listed in table 3.3.

Organism	**Sm site**
Common	`AUUUUUG`
Leptomonas seymouri	`AUUUUG`
Crithidia fasciculata	`AUAUUUUGA`
Trypanosoma brucei	`ACUUUG`

Table 3.3: Overview of Sm sites used for `RNAbob` runs. Sm sites are taken from the motifs listed in table 3.5.

As a last consistency check we try to cluster our predictions with `Blastclust`[30] to determine if our predictions are repetitive. First we cluster with default parameters and then less restrictive with -S 75 as similarity threshold and -L 0.5 as minimum length coverage. This results in clusters with higher cardinalities. Identical or even similar predictions are grouped together.

[2] http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#RNAbob

All figures of this thesis showing consensus secondary structures without any further citation are created with the `Vienna RNA package 1.6`[45],[46]. It is a set of tools for RNA secondary structure prediction and comparison[3]. Exemplary consensus structures are colored using the `ALIDOT` color scheme[4] regarding inconsistent sequences in the alignment by saturation and the occurring types of basepairing by color (cp. figure 3.4).

All comparisons of positions of resulting ncRNA predictions with known elements are done with the positions of the 'adjusted regions'. Everything dealing with primary sequence or alignments like `blast`/`blat` searches or consensus folding is done with 'best RNAs' to preserve the original biological context.

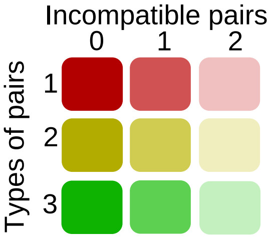


Figure 3.4: Coloring code of RNA consensus secondary structures. Inconsistency increases from the left to the right and differences in basepairing are marked by type of color. Thanks to S. Washietl for the permission to use this figure.

[3] `http://www.tbi.univie.ac.at/~ivo/RNA/`

[4] `http://www.tbi.univie.ac.at/RNA/ALIDOT/alidot-2.0.html#SEC6`

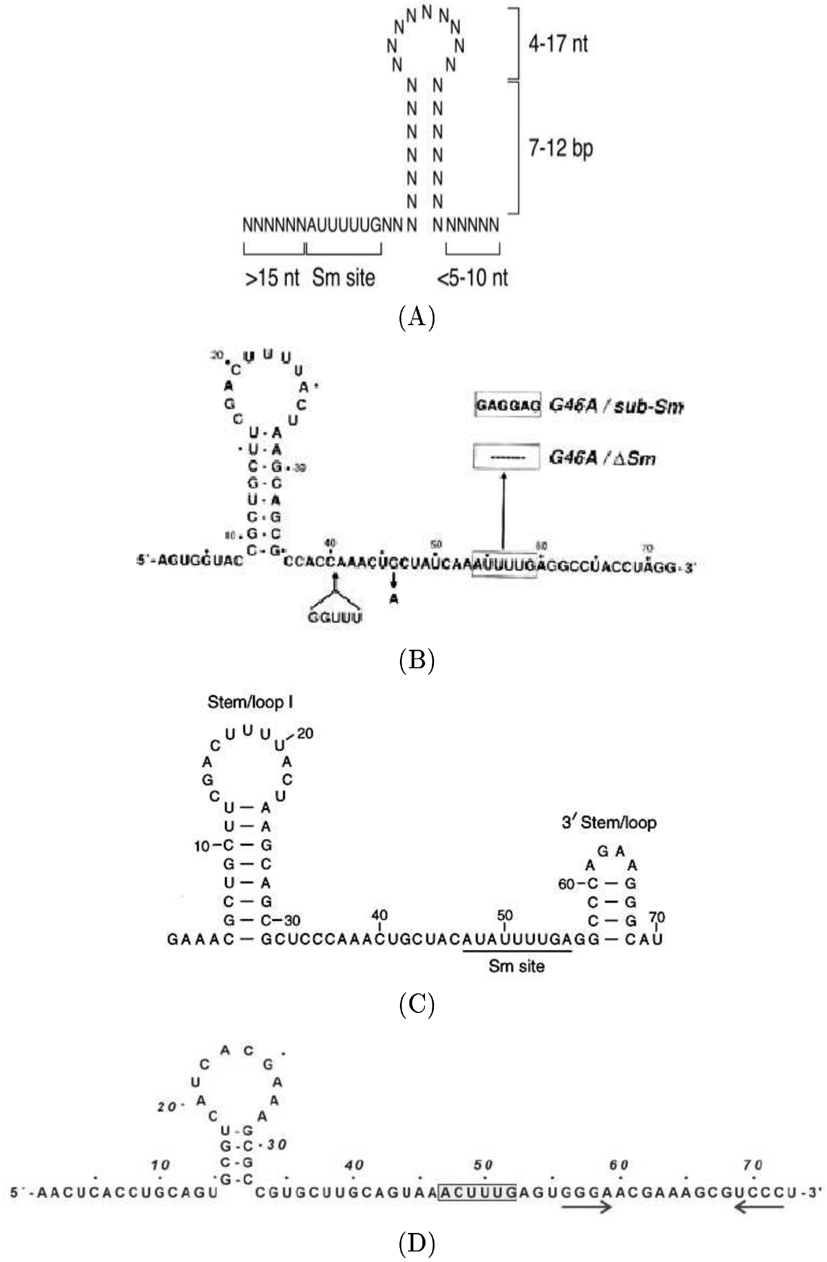

4-17 nt
7-12 bp
NNNNNNAUUUUUGNN N N NNNNN
>15 nt
Sm site
<5-10 nt
(A)
GAGGAG G46A / sub-Sm
G46A / ΔSm
5'-AGUGGUAC CCACCAAACUGCUAUCAAAUUUUGAGGCCUACCUAGG-3'
GGUUU
A
(B)
Stem/loop I
3' Stem/loop
GAAAC — GCUCCCAAACUGCUACAUAUUUUGAGG — CAU
Sm site
(C)
5'-AACUCACCUGCAGU CGUGCUUGCAGUAAACUUUGAGUGGGAACGAAAGCGUCCCU-3'
(D)

Figure 3.5: Critical RNA sequence features inducing SMN binding and Sm core formation; (A) Common critical RNA sequence features for SMN binding[41]; (B) *Leptomonas seymouri*, U5 snRNA[42]; (C) *Crithidia fasciculata*, U5 snRNA[43]; (D) *Trypanosoma brucei*, U1 snRNA[44]; Based on this motifs we set up descriptors allowing to search for putative SMN binding sites by RNAbob.

3.2 The supporting database system

Large-scale analyses imply to set up a computational environment where the huge amount of data can be managed efficiently. Processing of alignments of both reading directions, splitting them into RNAz scanning windows and genomewide annotations leads to fast growing data sets. With the advantages of an integrated relational database in mind, we set up a MySQL database server (http://www.mysql.com, version 4.1) for storing, processing and analysing the occurring data. We decide to use the MySQL database system due to its ability to handle character large objects (CLOBS) and the fact that build-in string functions work performantly on them. Moreover it has a high quality documentation and is open source.

The database model is shown in figure 3.6. It is divided into three main parts, indicated by three different background colors. The red area focuses on storing the raw source data including complete genomic sequences and initially given annotation. The yellow region handles results of processing steps, starting from the early blast searches up to the RNAz noncoding prediction. The green area provides tables and thus functionality for annotation purposes and result handling of additional tools. Therefore we are able to annotate predicted ncRNAs automatically. The central table POSITIONS is used to link all results together. It is the main annotation table where every genomic element including our ncRNA predictions is registered. Overall there are 22 tables providing data which we access either directly per command-line SQL statements or via a set of Perl scripts. The system can easily be adopted and upgraded with the release of new genomes and annotation data.

It is a general problem in Computer Science that programs are less worthy without documentation. Moreover it was a major part of this thesis to set up the computational environment. Therefore we decide to explain each table shortly. The description can be found in table 3.4.

Tablename	**Documentation**
POSITIONS	Central table concerning all kind of annotation data. Here, given and newly calculated positions of annotatable genomic elements are stored.
Handling sources	
ORGANISM	Each organism of the screen is entered here.
SCAFFOLDS	Stores all genomic sequences, no matter if they are chromosomes, scaffolds, contigs, shotgun-reads a.s.o. (The name of the table is a relict of our very first urochordate screen where the sequences were available as scaffolds).
TYPES	We have to specify what we want to annotate in general.
SUBTYPES	Subclasses of entries of the table TYPES
Noncoding RNA prediction	
SEARCHES	Specifies our different `blast` searches
SIGALIGN	Stores significant `blast` alignments
RNAZ	Provides data for each frame scanned with `RNAz`
CORRESPONDENCES	Think of a sequence with several `blast` hits. Each hit is used for further searches and may also lead to several hits. To avoid explosive data growth we only use one specific `blast` hit from the first search to perform the second.
RNAZCLUSTER	For each alignment all overlapping frames with $p > 0.5$ are merged together.
RNAZCLUSTERBRIDGE	We want to know which `RNAz` frame is used for which `RNAz` cluster.
RNAZANNOTATION	Clusters of the same reading direction overlapping more than 90 % are combined and build unadjusted ncRNA regions.
RNAZBRIDGE	We want to know which `RNAz` frames form which unadjusted region.
ADJUSTED	If there are regions overlapping more than 90 % of different reading directions we merge them and form adjusted regions by storing the best representative independent of the reading direction. Thereby we annotate a putative reading direction for the ncRNA candidate.
BESTRNA	Each adjusted region has one best representative in terms of combined `RNAz` frames of a specific alignment (`RNAz` clusters).
VBDOKU	Due to a restrictive merging procedure we still have multiple covered loci. Of all merged regions which overlap more than 20 % we discard all except one leaving us with a unique genomic locus for each ncRNA gene.
WINDOWS	We want to know which `RNAz` frames participate in setting up the ncRNA predictions. This table provides hard coded redundant information so that we do not have to calculate extensive table joins (Stepping back the processing pipeline would also reveal this information.).
Additional annotation	
ANNOTYPES	The different kind of annotation approaches for analysing the ncRNA predictions are listed here.
COMMENTS	What we can say about each prediction.
BCRUN	Each `Blastclust` run is listed here.
BCLUST	Each observed `Blastclust` cluster is stored here.
BMEM	Which ncRNA prediction participates in which cluster.

Table 3.4: Brief documentation of the database system

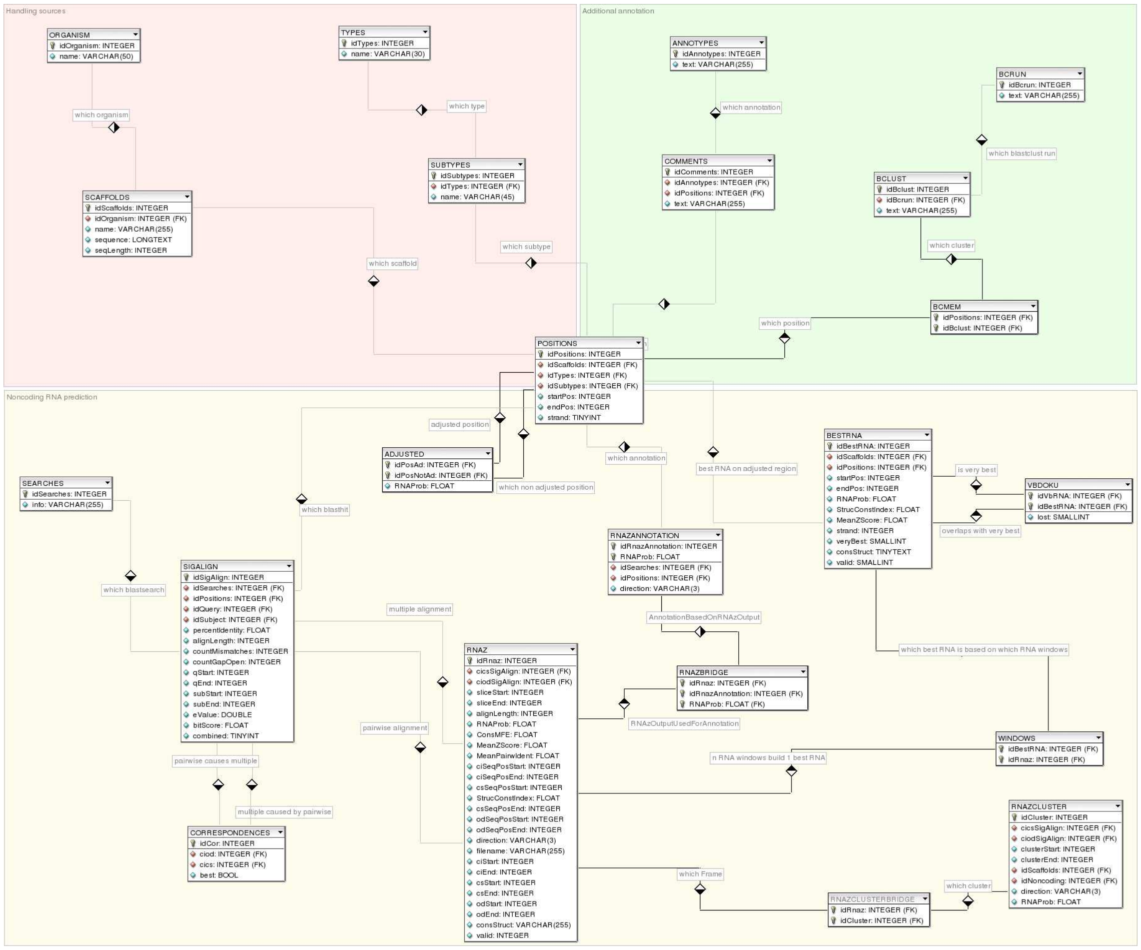


Figure 3.6: Database model used for automated ncRNA prediction.

Chapter 4

Noncoding RNA predictions

We performed several prediction screens (section 3.1, table 3.1). This chapter presents the obtained results. The complete output of the major processing tools is stored at several integrated databases allowing fast assaying. Currently we set up 5 databases for the true data sets and 5 for shuffled alignment approaches. The databases handling the true data sets contain up to $360,000$ (TbTc, TbTcTv), $92,000$ (TbTcLi), $97,000$ (TbTcLm), $7,800,000$ (LiLmTb) and $2,500,000$ (LmLiTb) records providing putative annotations of novel noncoding RNAs.

4.1 Novel ncRNA candidates of the genus Trypanosoma

In this section we want to report about our predicted ncRNA candidates of the *Trypanosoma brucei* genome. The numbers of observed ncRNA signals are listed in table 4.1.

p	**TbTc**	**TbTcTv**	**TbTcLi**	**TbTcLm**
> 0.50	290	140	122	117
> 0.90	136	71	66	70
> 0.98	84	24	37	38
> 0.99	68	18	24	30

Table 4.1: Overview of the numbers of *Trypanosoma brucei* ncRNA predictions. Obviously there are more hits among the screens using trypanosomes only, but the predictions including Li and Lm seem to have higher p values.

Out of the given annotation we calculated 17,485 regions of Tb noncoding DNA. These elements comprise 11,819,781 nt what is nearly 43 % of the total Tb genomic sequence.

4.1.1 The TbTc approach

Based on pairwise alignments of Tb with Tc we detect 290 structured RNA signals ($p > 0.5$, Tab. 4.1) of which 192 are annotatable. Due to the definitions of subsection 3.1.3 we obtain the specifity values and false positive rates shown in table 4.2, the resulting sensitivity values are shown in table 4.3. Observed detection rates are shown in table 4.4.

p	> 0.50	> 0.90	> 0.98	> 0.99
False positive rates	63 %	57 %	56 %	54 %
	(1,842/2,912)	(754/1,323)	(377/678)	(294/540)
Specificity per test	0.939	0.975	0.988	0.990

Table 4.2: False positive rates and specifities of the TbTc screen; With higher p value the false positive rate does not decrease significantly.

type	N	n	N_a	N_g	s_{Na}	s_{Ng}	s_{na}	s_{ng}
rRNA	82	83	92	106	0.89	0.77	0.90	0.78
tRNA	30	32	56	65	0.54	0.46	0.57	0.49
misc_RNA	25	1	29	29	0.86	0.86	-	-
snRNA	3	4	5	6	0.60	0.50	0.80	0.67
snoRNA	7	7	27	353	0.26	0.02	0.26	0.02

Table 4.3: Estimated sensitivities of the TbTc screen; Given misc_RNAs are actually slRNAs. The positions of our ncRNA predictions seem to be woolly, however. 25 out of 29 slRNAs map to our predictions if we search for matches with an overlap of 70 %. Only one matches if we search with the absolute positions (required overlap 100 %).

	per 1 mb alignment		per 1 mb nc region	
p	normal	shuffled	normal	shuffled
> 0.50	5,428.68	3,433.94	246.37	155.84
> 0.90	2,466.40	1,405.64	111.93	63.79
> 0.98	1,263.96	702.82	57.36	31.90
> 0.99	1,006.69	548.09	45.69	24.87

Table 4.4: Noncoding RNA detection rates of the TbTc screen.

We obtain 8,439 `blast` hits with an average length of ~64 nt between the Tb noncoding DNA and the Tc genome. 4,523 of them made it into the `RNAz` input set. Their average length is $\sim 95\,nt$. In consideration of all possible reading directions the total alignment input set comprises 9,046 alignments. Overall there are 30,162 `RNAz` frames. At the $p > 0.5$ level we count 2,912 and with $p > 0.9$ we observe 1,323. Figure 4.1 illustrates the complete distribution of structured `RNAz` frames.

The tRNA finder `tRNAscan-SE` produced 24 tRNA predictions, which all decode standard amino acids. Their average length is 73 nt. A `blast` search of the Tb predictions against the complete

Noncode database affects 5 of our ncRNAs (34 hits). Overall a run of the miRNA detection tool RNAmicro reveals 348 hits. 12 of them are classified with $p > 0.5$. But a blast search against the miRBase produced no hit. A blast search against the Rfam leads to 1,199 hits affecting 62 of our ncRNA candidates. A search after RNAs homologous to the Rfam covariance models produced 319 hits covering 224 of our ncRNA predictions. The distribution of the obtained bitscores is shown in figure 4.2. After filtering (cp. restrictions mentioned at 3.2) we keep 32 annotations. Furthermore we obtain no blast hits with sequences of the snoRNA-LBME-db. Results of the RNAbob searches are listed in table 4.5. Exemplary secondary structures of conserved RNAbob hits are illustrated in figure 4.3.

The ncRNA candidates are groupable into 148 cluster by the first and into 120 cluster by the second Blastclust run. The maximal observed cardinality is 24 in both runs indicating that we detect the same or in fact very similar ncRNAs about 20 times in the genome (cp. figure 4.4). For both searches the cluster with the maximal cardinality comprise the slRNAs. The next clusters are a M4 rRNA (large subunit epsilon/zeta) and a M1 rRNA cluster (large subunit gamma).

Finally we present some exemplary ncRNA consensus structures of the TbTc screen at figure 4.5.

descriptor	**# hits**	**# conserved hits**	**# unique predictions**
Common	0	0	0
L. seymouri	0	0	0
C. fasciculata	0	0	0
T. brucei	2	2	2

Table 4.5: Number of ncRNA candidates for SMN binding of the TbTc screen. Conserved in this context indicates that we count hits where the motif is found in the majority of the aligned sequences. Unique indicates the number of involved ncRNA predictions with different secondary structure.

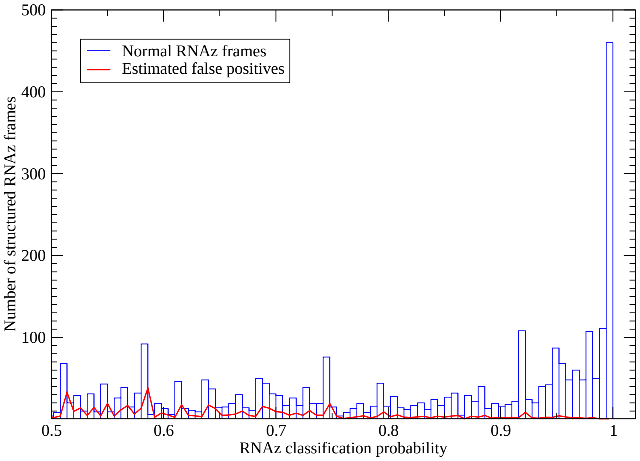


Figure 4.1: Histogram of the `RNAz` classification probability of the TbTc screen. The distribution looks nice, with higher p value we detect more structures.

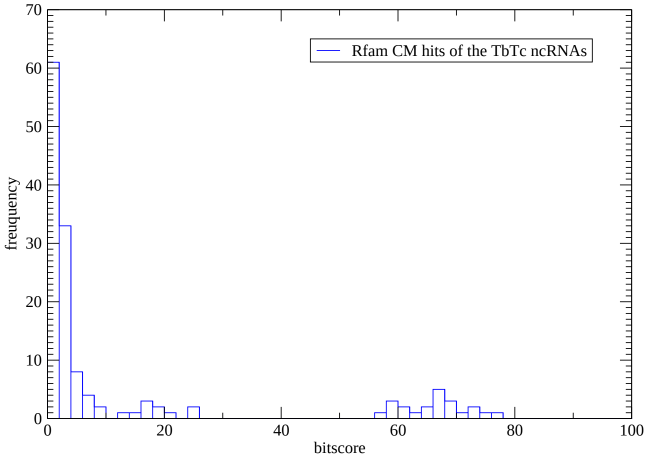


Figure 4.2: Bitscores obtained by `Infernals cmsearch` of the TbTc screen. The bitscore is problematic, there are so many hits with low score, we set a cutoff value at 10.

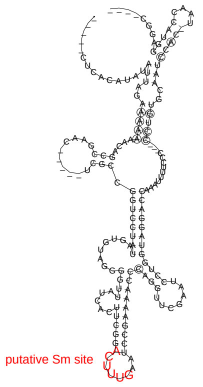


Figure 4.3: Exemplary `RNAbob` hit using the *T. brucei* SMN descriptor (TbTc screen); Illustrated structure: 44432.

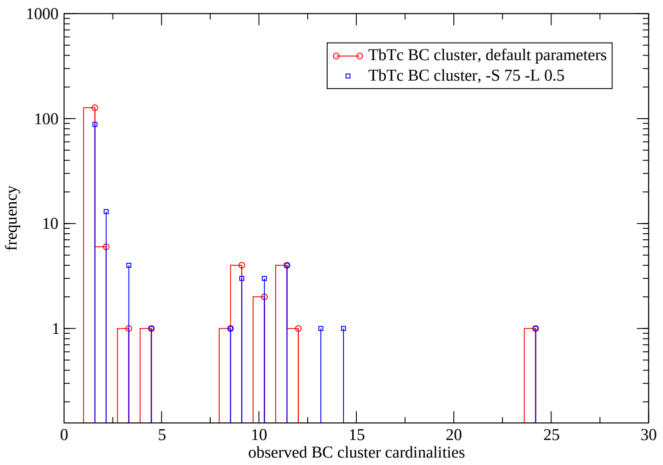
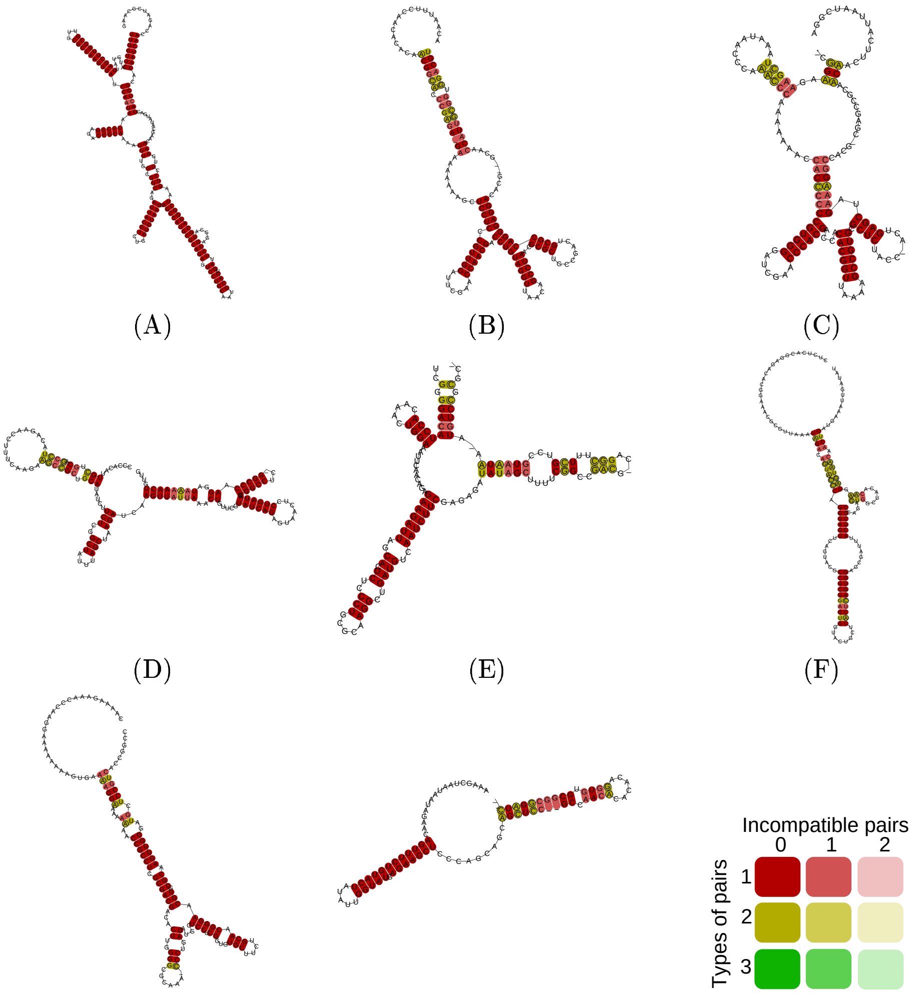


Figure 4.4: Histogram of the obtained `Blastclust` cardinalities of the TbTc ncRNAs. The y-axis is scaled logarithmically. The major fraction of the predicted ncRNAs with low cluster cardinalities seems to be non-repetitive, but we observe some significant repetitive hits with cardinalities of more than 20.



(A) (B) (C)

(D) (E) (F)

(G) (H)

Figure 4.5: Exemplary consensus structures of the TbTc ncRNAs.
(A) Structure 47594, p=0.999, known as rRNA cluster of small subunit 18S rRNA
(B) Structure 47800, p=0.816, known as asparagine tRNA
(C) Structure 47803, p=0.959, known as lysine tRNA
(D) Structure 47667, p=0.886, known as U2 snRNA
(E) Structure 47742, p=0.685, known as U6 snRNA
(F) Structure 47865, p=0.921, known as H/ACA snoRNA Tb10Cs2H2
(G) Structure 47807, p=0.506, known as 7SL, SRP RNA
(H) Structure 47917, p=0.943, known as slRNA

4.1.2 The TbTcTv approach

Based on three-way alignments of Tb, Tc and Tv we detect 140 structured RNA signals ($p > 0.5$, Tab. 4.1) of which 86 are annotatable. Due to the definitions of subsection 3.1.3 we obtain the specifity values and false positive rates shown in table 4.6, the resulting sensitivity values are shown in table 4.7. Observed detection rates are shown in table 4.8.

p	> 0.50	> 0.90	> 0.98	> 0.99
False positive rates	57 %	61 %	67 %	71 %
	(2,241/3,911)	(653/1,072)	(247/370)	(144/202)
Specificity per test	0.959	0.988	0.995	0.997

Table 4.6: False positive rates and specifities of the TbTcTv screen; The increase of the FPR with higher p value has never been observed in prior screens. This screen should be doubted, the results of the SVM classification seem to be uncertain due to the phylogenetic range of the genomic sequences. The average SVM classification criteria of resulting ncRNAs of this screen (p=0.85, z=-1.52, SCI=0.74) tend to get outperformed by the comparable TbTcLi (cp. 4.1.3, p=0.87, z=-1.59, SCI=0.83) and TbTcLm (cp. 4.1.4, p=0.90, z=-1.68, SCI=0.82) ncRNA candidates. The high false positive rate implies repetitive elements among the true data set. Shuffling did not destroy the structures, they remain detectable in the shuffled set. To enlighten those assumptions we manually curated the annotation of each ncRNA prediction by `blast` searchs against the `NCBI`. We are able to annotate 101 predictions by hand. With 24 the most often occurring hit class is spliced leader mini-exon and they indeed appear repetitive in the trypanosomes. But they are not present in the other three-way screens (and especially not in their shuffled counterpart) and thus mislead TbTcTv false positive statistics.

type	N	n	N_a	N_g	s_{Na}	s_{Ng}	s_{na}	s_{ng}
rRNA	33	45	51	106	0.65	0.31	0.88	0.43
tRNA	17	20	47	65	0.36	0.26	0.43	0.31
misc_RNA	25	28	28	29	0.89	0.86	1.00	0.97
snRNA	2	2	1	6	1.**	0.34	1.**	0.34
snoRNA	0	0	4	353	0.00	0.00	0.00	0.00

Table 4.7: Estimated sensitivities of the TbTcTv screen; Given misc_RNAs are actually slRNAs; ** indicate improper ratios. Only one of the given snRNAs in the alignment input set is detectable by `blast` or `blat`, but we have 2 of the given snRNAs in our result set (namely U2, U3). The U5 snRNA is annotatable by `NCBI` `blast` (cp. figure 4.9). In conclusion we detect all of the given snRNAs of the input set. It is correct to have a smaller input set in comparison to the TbTc screen (two genomes) because the RNAs have to be conserved in all three genomes. In percentages we notably perform worse detecting tRNAs, but `tRNAscan-SE` is also not able to detect more of them.

Prior pairwise `blast` hits are used for global alignments with Tv. This resulted in 13,648 `blast` hits with an average length of ~66 nt. 11,107 of them made it into the `RNAz` input set. Their average length is ~113 nt. In consideration of all possible reading directions the total alignment input set comprises 22,214 alignments. Overall there are 54,554 `RNAz` frames. At the $p > 0.5$ level we count 3,911 and with $p > 0.9$ we observe 1,072. Figure 4.6 illustrates the complete distribution of structured `RNAz` frames.

	per 1 mb alignment		per 1 mb nc region	
p	normal	shuffled	normal	shuffled
> 0.50	7,291.06	4,177.77	330.89	189.60
> 0.90	1,998.47	1,217.35	90.70	55.25
> 0.98	689.77	460.47	31.30	20.90
> 0.99	376.58	268.45	17.09	12.18

Table 4.8: Noncoding RNA detection rates of the TbTcTv screen.

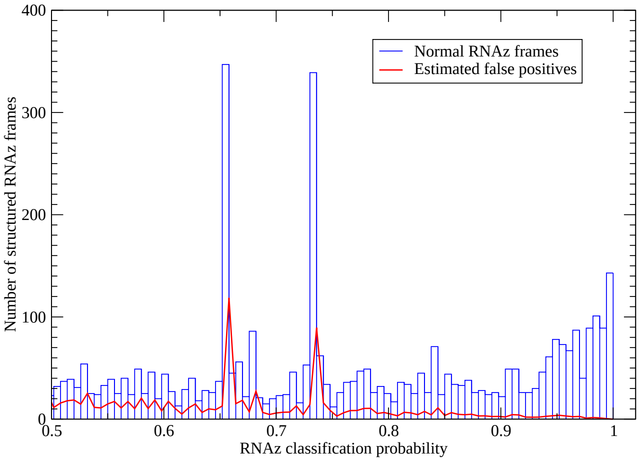


Figure 4.6: Histogram of the RNAz classification probability of the TbTcTv screen. The two peaks at around $p = 0.66$ and $p = 0.73$ destroy the significance of the entire screen. There is a considerable fraction of low scoring RNAz frames, we assume that the phylogenetic range is too close or an abundance of unannotated repeats.

The tRNA finder tRNAscan-SE produced 13 tRNA predictions, which all decode standard amino acids. Their average length is 74 nt. A blast search of the Tb predictions against the complete Noncode database affects 2 of our ncRNAs (21 hits). Overall a run of the miRNA detection tool RNAmicro reveals 288 hits. Still 2 of them are classified with $p > 0.5$. But a blast search against the miRBase produced no hit. A blast search against the Rfam leads to 1,102 hits affecting 43 of our ncRNA candidates. A search after RNAs homologous to the Rfam covariance models produced 149 hits covering 113 of our ncRNA predictions. The distribution of the obtained bitscores is shown in figure 4.7. After filtering (cp. restrictions mentioned at 3.2) we keep 18 annotations. Furthermore we obtain no blast hits with sequences of the snoRNA-LBME-db. Results of the RNAbob searches are listed in table 4.9.

The ncRNA candidates are groupable into 74 cluster by the first and into 58 cluster by the second Blastclust run. The maximal observed cardinality is 23 (first run) and 25 (second run) indicating that we detect the same or in fact very similar ncRNAs about 20 times in the genome. For both

descriptor	# hits	# conserved hits	# unique predictions
Common	0	0	0
L. seymouri	14	0	0
C. fasciculata	0	0	0
T. brucei	6	0	0

Table 4.9: Number of ncRNA candidates for SMN binding of the TbTcTv screen. Conserved in this context indicates that we count hits where the motif is found in the majority of the aligned sequences. Unique indicates the number of involved ncRNA predictions with different secondary structure. It is surprising that we do not get a conserved hit with the Tb descriptor because we have hits with ncRNA predictions of the TbTcLi and TbTcLm screen (cp. 4.13, 4.17), although Tv is phylogenetically closer to Tb than it is Li or Lm.

searches the cluster with the maximal cardinality comprise the slRNAs. The next clusters are a M4 rRNA (large subunit epsilon/zeta) and a M2 rRNA cluster (large subunit delta).

Finally we present some exemplary ncRNA consensus structures at figure 4.9.

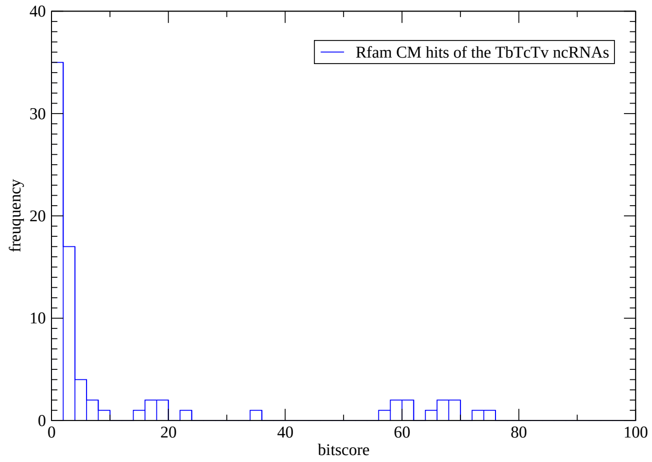


Figure 4.7: Bitscores obtained by `Infernals` `cmsearch` of the TbTcTv screen. The bitscore is problematic, there are so many hits with low score, we set a cutoff value at 10.

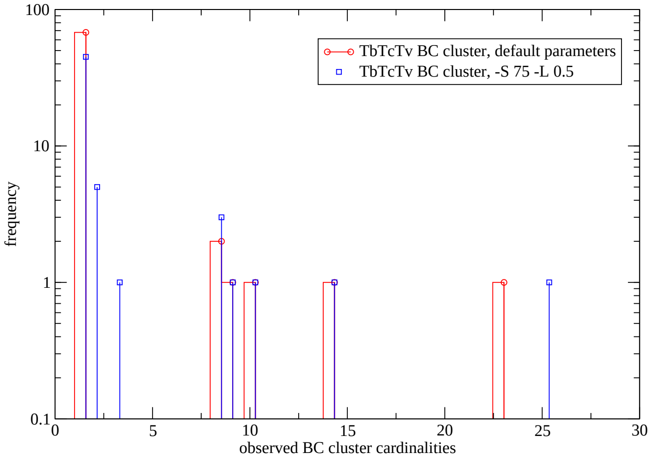
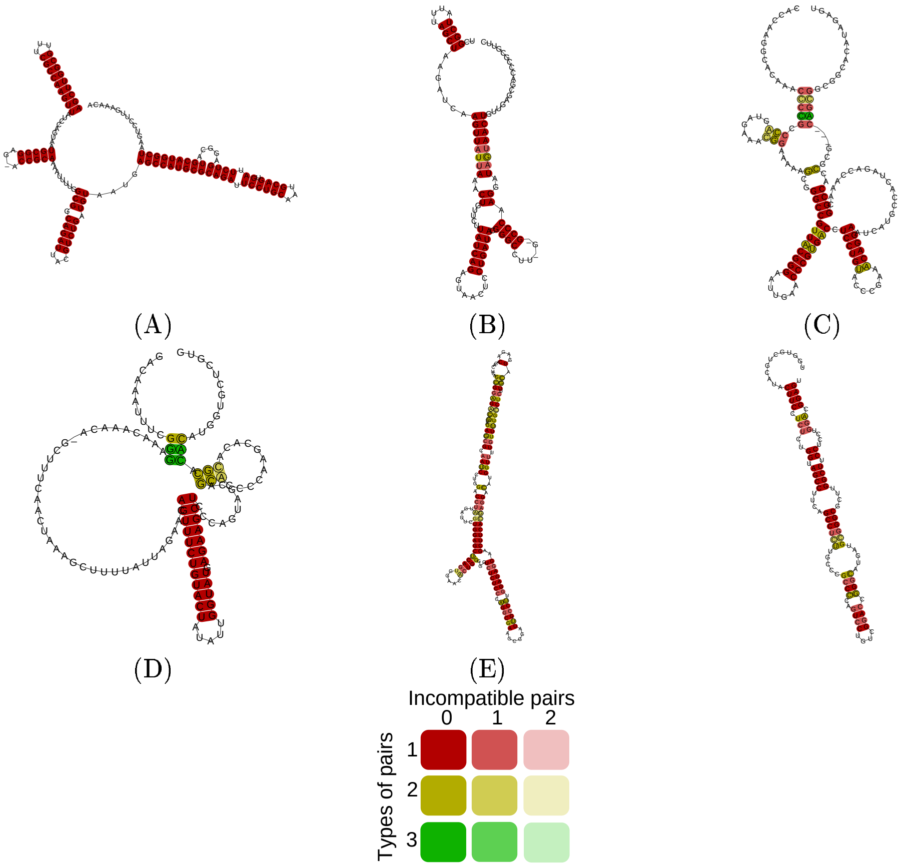


Figure 4.8: Histogram of the obtained `Blastclust` cardinalities of the TbTcTv ncRNAs. The y-axis is scaled logarithmically. The major fraction of the predicted ncRNAs with low cluster cardinalities seems to be non-repetitive, but there are some significant repetitive hits among our predictions with cluster cardinalities of more than 20.


(A)

(B)

(C)

(D)

(E)

Figure 4.9: Exemplary consensus structures of the TbTcTv ncRNAs. Obviously they are similar to the TbTc structures, so we illustrate some of the unannotatable ncRNA candidates.
(A) Structure 46758, p=0.965, known as rRNA cluster and 18S rRNA
(B) Structure 46779, p=0.848, known as U2 snRNA
(C) Structure 46807, p=0.999, overlaps with known tRNA, a `NCBI blast` search indicates the U5 snRNA of *T. brucei* and *L. seymouri*
(D) Structure 46889, p=0.843, known as slRNA, we did not match the slRNA completely and the consensus structure confirms this
(E) Structure 46811, p=0.963, putative novel ncRNA, yet unannotated, a `NCBI blast` search revealed no further annotation
(F) Structure 46917, p=0.968, putative novel ncRNA, yet unannotated, a `NCBI blast` search revealed no further annotation

4.1.3 The TbTcLi approach

Based on three-way alignments of Tb, Tc and Li we detect 122 structured RNA signals ($p > 0.5$, Tab. 4.1) of which 117 are annotatable. Due to the definitions of subsection 3.1.3 we obtain the specifity values and false positive rates shown in table 4.10, the resulting sensitivity values are shown in table 4.11. Observed detection rates are shown in table 4.12.

p	> 0.50	> 0.90	> 0.98	> 0.99
False positive rates	18 %	20 %	14 %	13 %
	(64/362)	(29/147)	(11/79)	(8/60)
Specificity per test	0.974	0.988	0.996	0.997

Table 4.10: False positive rates and specifities of the TbTcLi screen; The false positive rate improves considerably in comparison to the prior Trypanosoma only screens.

type	N	n	N_a	N_g	s_{Na}	s_{Ng}	s_{na}	s_{ng}
rRNA	61	69	83	106	0.74	0.58	0.83	0.65
tRNA	31	35	54	65	0.57	0.48	0.65	0.54
misc_RNA	0	0	0	29	0.00	0.00	0.00	0.00
snRNA	2	2	2	6	1.00	0.34	1.00	0.34
snoRNA	0	0	0	353	0.00	0.00	0.00	0.00

Table 4.11: Estimated sensitivities of the TbTcLi screen; Members of the misc_RNA family are 7 sl SRP RNA and other slRNAs; They miss the input set and thus are not detectable; We match the known U2 and U6 snRNA, but miss the given U1, U3, U4 and U5 snRNAs; snoRNAs also are not in the input set.

	per 1 mb alignment		per 1 mb nc region	
p	normal	shuffled	normal	shuffled
> 0.50	674.86	119.31	30.63	5.41
> 0.90	274.04	54.06	12.44	2.45
> 0.98	147.28	20.51	6.68	0.93
> 0.99	111.85	14.91	5.08	0.68

Table 4.12: Noncoding RNA detection rates of the TbTcLi screen.

Prior pairwise `blast` hits are used for global alignments with Li yielding to 23,844 `blast` hits with an average length of ~30 nt. Only 453 of them made it into the `RNAz` input set. Their average length is ~128 nt. In consideration of all possible reading directions the total alignment input set comprises 906 alignments. Overall there are 2,460 `RNAz` frames. At the $p > 0.5$ level we count 362 and with $p > 0.9$ we observe 147. Figure 4.10 illustrates the complete distribution of structured `RNAz` frames.

`tRNAscan-SE` produced 29 tRNA predictions. 28 of them decode for standard amino acids and one is classified as pseudogene. Their average length is 73 nt. A `blast` search of the Tb predictions against the complete Noncode database affects 2 of our ncRNAs (23 hits). Overall a run of RNAmicro reveals 193 hits. Still 8 of them are classified with $p > 0.5$. But a blast search against the miRBase produced no hit. A blast search against the Rfam leads to 998 hits affecting 65 of our ncRNA candidates. A search after RNAs homologous to the Rfam covariance models produced 180 hits covering 100 of our ncRNA predictions. The distribution of the obtained bitscores is shown in figure 4.12. After filtering (cp. restrictions mentioned at 3.2) we keep 43 annotations. Furthermore we obtain no blast hits with sequences of the snoRNA-LBME-db. Results of the RNAbob searches are listed in table 4.13. Exemplary secondary structures of conserved RNAbob hits are illustrated in figure 4.11.

The ncRNA candidates are groupable into 54 cluster by the first and into 41 cluster by the second Blastclust run. The maximal observed cardinality is 11 (first run) and 13 (second run) indicating that we detect the same or in fact very similar ncRNAs about 10 times in the genome. The cluster with maximal cardinalities comprise rRNAs in both searches. From the top to the bottom we observe 5.8S rRNA (M3), 28S LSU alpha, and M2 LSU delta clusters by the first BC run and M2, M4 and M3 clusters by the second one.

Finally we present some exemplary ncRNA consensus structures of the TbTcLi screen at figure 4.14.

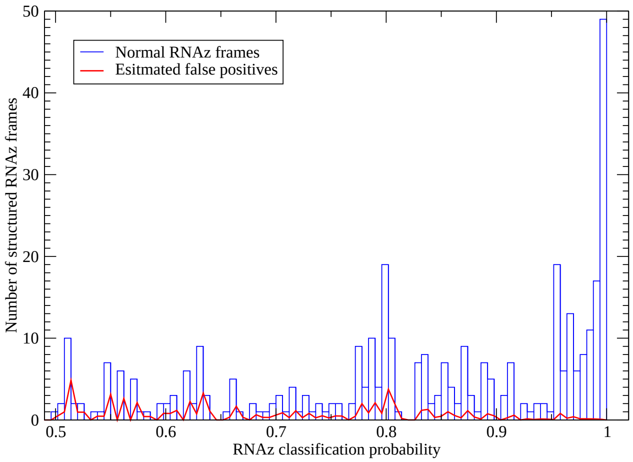


Figure 4.10: Histogram of the `RNAz` classification probability of the TbTcLi screen. The distribution looks decently, there are many frames with high p values.

descriptor	# hits	# conserved hits	# unique predictions
Common	0	0	0
L. seymouri	11	0	0
C. fasciculata	0	0	0
T. brucei	16	11	3

Table 4.13: Number of ncRNA candidates for SMN binding of the TbTcLi screen. Conserved in this context indicates that we count hits where the motif is found in the majority of the aligned sequences. Unique indicates the number of involved ncRNA predictions with different secondary structure.

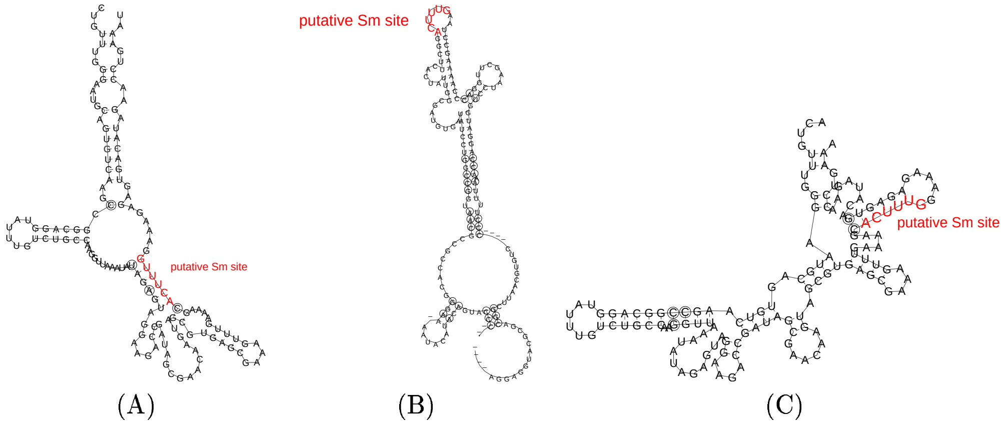


Figure 4.11: Exemplary `RNAbob` hits using the *T. brucei* SMN descriptor (TbTcLi screen); (A) structure: 44270; (B) structure: 44360; (C) structure: 44390. Altogether they do not look very stable.

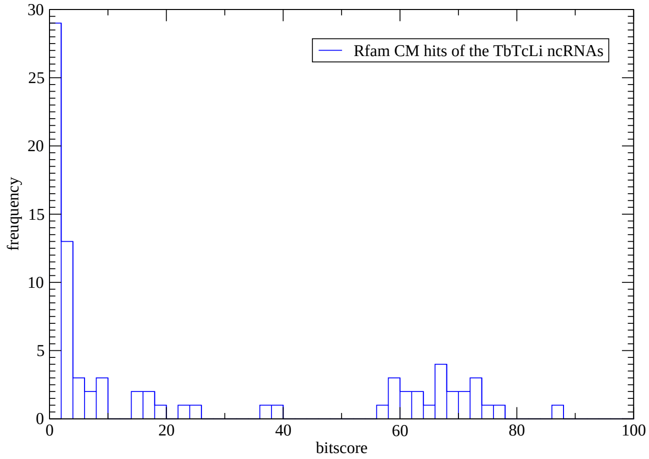


Figure 4.12: Bitscores obtained by `Infernals` `cmsearch` of the TbTcLi screen. The bitscore is problematic, there are so many hits with low score, we set a cutoff value at 10.

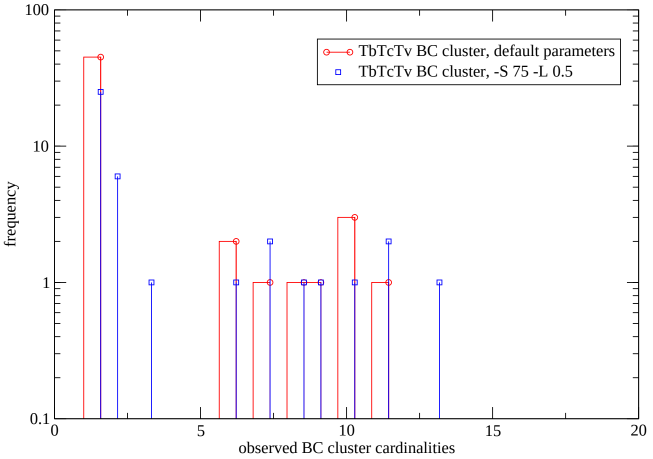
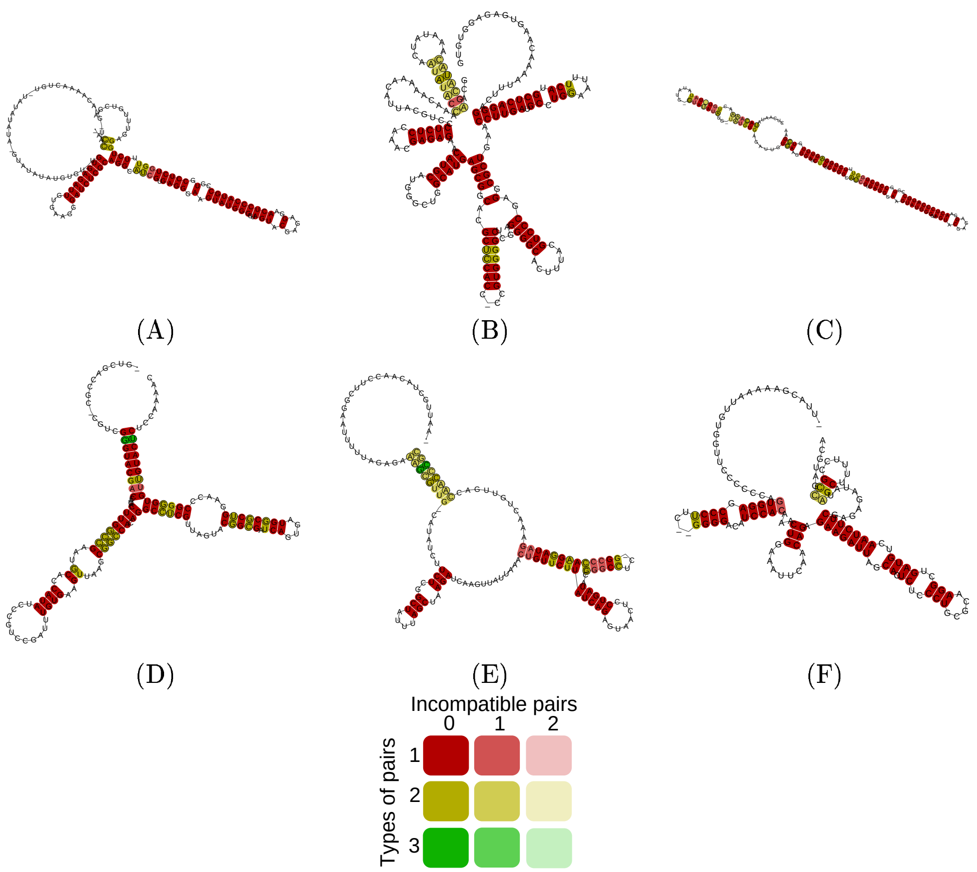


Figure 4.13: Histogram of the obtained `Blastclust` cardinalities of the TbTcLi ncRNAs. The y-axis is scaled logarithmically. The major fraction of the predicted ncRNAs with low cluster cardinalities seems to be non-repetitive, but we observe some significant repetitive hits with cardinalities of more than 10.



(A) (B) (C)

(D) (E) (F)

Figure 4.14: Exemplary consensus structures of the TbTcLi ncRNAs.
(A) Structure 44288, p=0.776, known as M4 LSU rRNA
(B) Structure 44338, p=0.956, known as M2 rRNA
(C) Structure 44350, p=0.799, known as M2 LSU rRNA delta, we matched the second half of the known structure
(D) Structure 44351, p=0.632, known as 5S M5 rRNA
(E) Structure 44289, p=0.992, known as U2 snRNA
(F) Structure 44329, p=0.632, known as U6 snRNA

4.1.4 The TbTcLm approach

Based on three-way alignments of Tb, Tc and Lm we detect 117 structured RNA signals ($p > 0.5$, Tab. 4.1) of which 112 are annotatable. Due to the definitions of subsection 3.1.3 we obtain the specifity values and false positive rates shown in table 4.14, the resulting sensitivity values are shown in table 4.15. Observed detection rates are shown in table 4.16.

p	> 0.50	> 0.90	> 0.98	> 0.99
False positive rates	21 %	16 %	13 %	9 %
	(259/1210)	(78/475)	(28/208)	(13/152)
Specificity per test	0.966	0.990	0.996	0.998

Table 4.14: False positive rates and specifities of the TbTcLm screen; The false positive rate improves considerably in comparison to the prior Trypanosoma only screens.

type	N	n	N_a	N_g	s_{Na}	s_{Ng}	s_{na}	s_{ng}
rRNA	56	60	83	106	0.68	0.53	0.72	0.57
tRNA	33	37	54	65	0.61	0.31	0.69	0.57
misc_RNA	0	0	0	29	0.00	0.00	0.00	0.00
snRNA	1	1	1	6	0.17	0.17	0.17	0.17
snoRNA	0	0	0	353	0.00	0.00	0.00	0.00

Table 4.15: Estimated sensitivities of the TbTcLm screen; Members of the misc_RNA family are 7 sl SRP RNA and other slRNAs; They miss the input set and thus are not detectable; We match the known U6 snRNA, but miss the given U1, U2, U3, U4 and U5 snRNAs; snoRNAs also are not in the input set.

	per 1 mb alignment		per 1 mb nc region	
p	normal	shuffled	normal	shuffled
> 0.50	2,255.74	482.84	102.37	21.91
> 0.90	885.52	145.41	40.19	6.60
> 0.98	387.76	52.20	17.60	2.37
> 0.99	283.37	24.24	12.86	1.10

Table 4.16: Noncoding RNA detection rates of the TbTcLm screen.

Prior pairwise `blast` hits are used for global alignments with Lm yielding to 20,970 `blast` hits with an average length of ~40 nt. Only 1,225 of them made it into the `RNAz` input set. Their average length is ~190 nt. In consideration of all possible reading directions the total alignment input set comprises 2,450 alignments. Overall there are 7,737 `RNAz` frames. At the $p > 0.5$ level we count 1,210 and with $p > 0.9$ we observe 475. Figure 4.15 illustrates the complete distribution of structured `RNAz` frames.

`tRNAscan-SE` produced 30 tRNA predictions, 29 of them decode for standard amino acids and one is classified as pseudogene. Their average length is 72 nt. A `blast` search of the Tb predictions against the complete Noncode database affects 2 of our ncRNAs (7 hits). Overall a run of RNAmicro reveals 196 hits. Still 8 of them are classified with $p > 0.5$. But a blast search against the miRBase produced no hit. A blast search against the Rfam leads to 922 hits affecting 56 of our ncRNA candidates. A search after RNAs homologous to the Rfam covariance models produced 164 hits covering 96 of our ncRNA predictions. The distribution of the obtained bitscores is shown in figure 4.17. After filtering (cp. restrictions mentioned at 3.2) we keep 33 hits. Furthermore we obtain no blast hits with sequences of the snoRNA-LBME-db. Results of the RNAbob searches are listed in table 4.17. Exemplary secondary structures of conserved RNAbob hits are illustrated in figure 4.16.

The ncRNA candidates are groupable into 54 cluster by the first and into 40 cluster by the second Blastclust run. The maximal observed cardinality is 13 (first run) and 14 (second run) indicating that we detect the same or in fact very similar ncRNAs about 10 times in the genome. The cluster with maximal cardinalities comprise rRNAs in both searches. From the top to the bottom we observe M4 LSU epsilon/zeta, 24S LSU alpha and 5.8S M3 rRNA clusters by the first BC run and M4, M2, 24S LSU alpha rRNA clusters by the second one.

Finally we present some exemplary ncRNA consensus structures of the TbTcLm screen at figure 4.19.

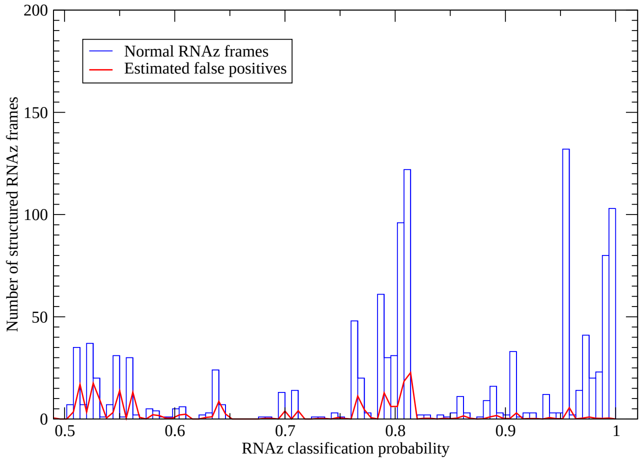


Figure 4.15: Histogram of the **RNAz** classification probability of the TbTcLm screen. The distribution implies an abundance of repetitive structures ($p = 0.8$).

descriptor	# hits	# conserved hits	# unique predictions
Common	2	0	0
L. seymouri	15	0	0
C. fasciculata	0	0	0
T. brucei	16	12	3

Table 4.17: Number of ncRNA candidates for SMN binding of the TbTcLm screen. Conserved in this context indicates that we count hits where the motif is found in the majority of the aligned sequences. Unique indicates the number of involved ncRNA predictions with different secondary structure.

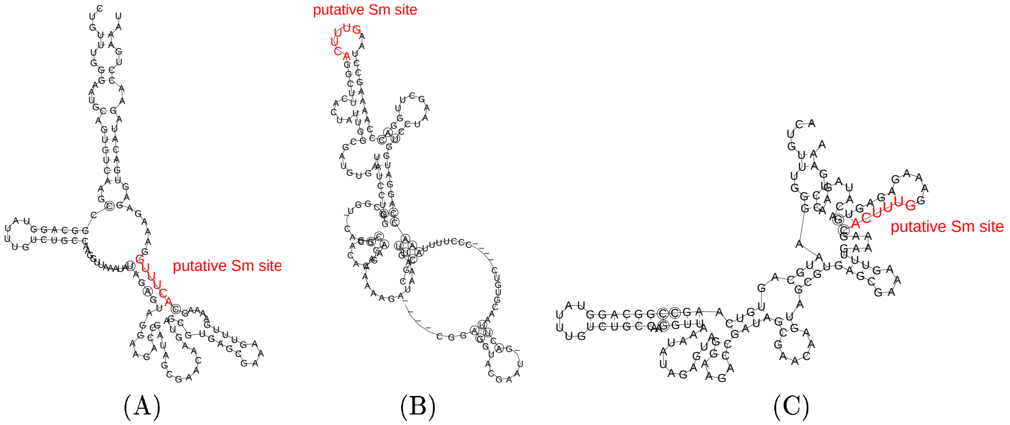


Figure 4.16: Exemplary **RNAbob** hits using the *T. brucei* SMN descriptor (TbTcLm screen); (A) structure: 44042; (B) structure: 44128; (C) structure: 44155. The stability of the illustrated structures is questionable.

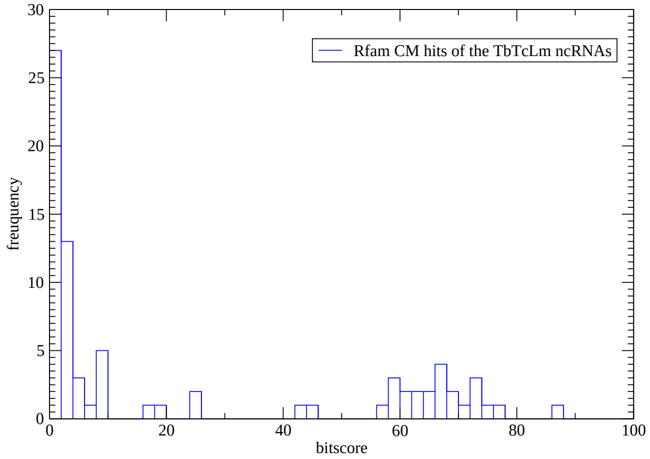


Figure 4.17: Bitscores obtained by `Infernals` `cmsearch` of the TbTcLm screen. The bitscore is problematic, there are so many hits with low score, we set a cutoff value at 10.

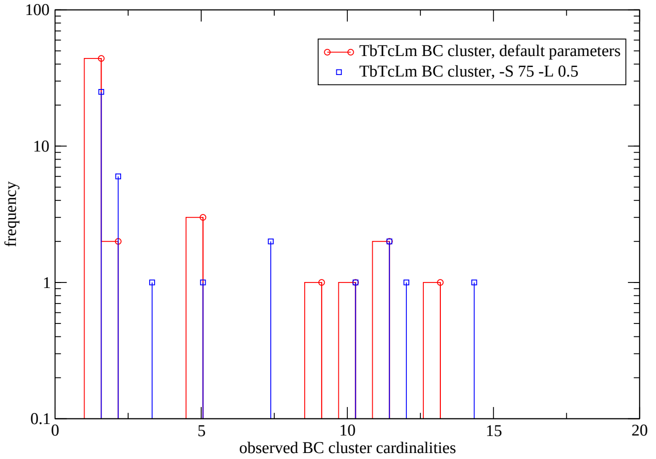


Figure 4.18: Histogram of the obtained `Blastclust` cardinalities of the TbTcLm ncRNAs. The y-axis is scaled logarithmically. The major fraction of the predicted ncRNAs with low cluster cardinalities seems to be non-repetitive, but we observe some significant repetitive hits with cardinalities of more than 10.

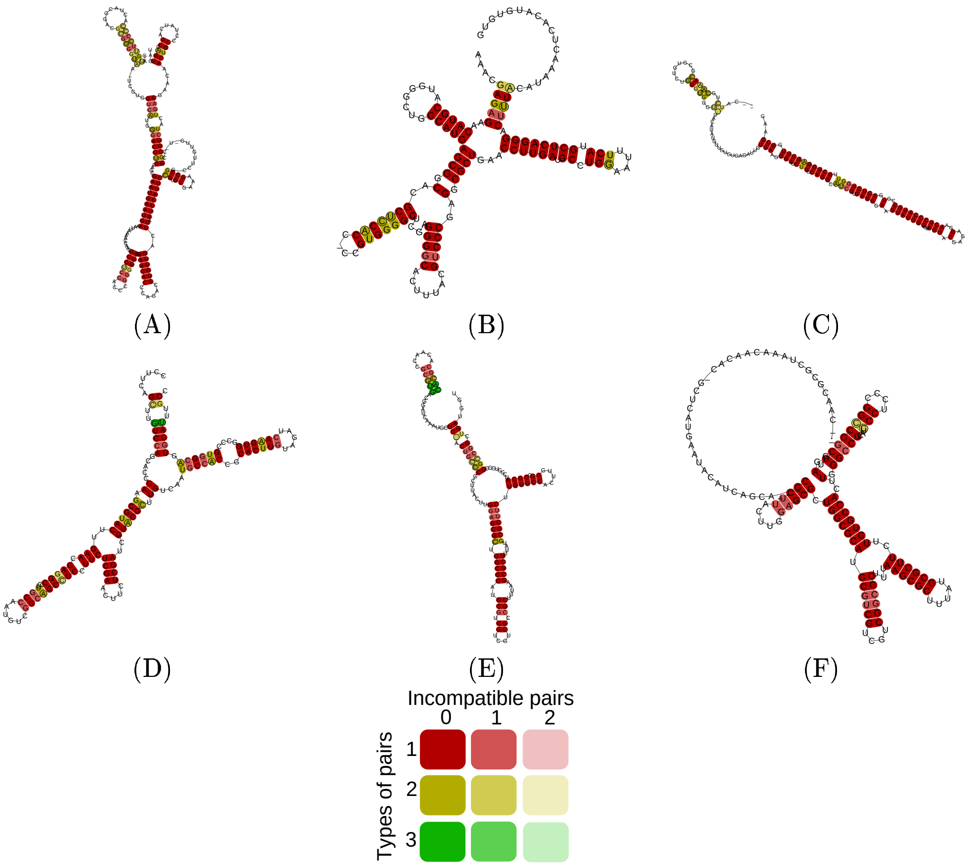
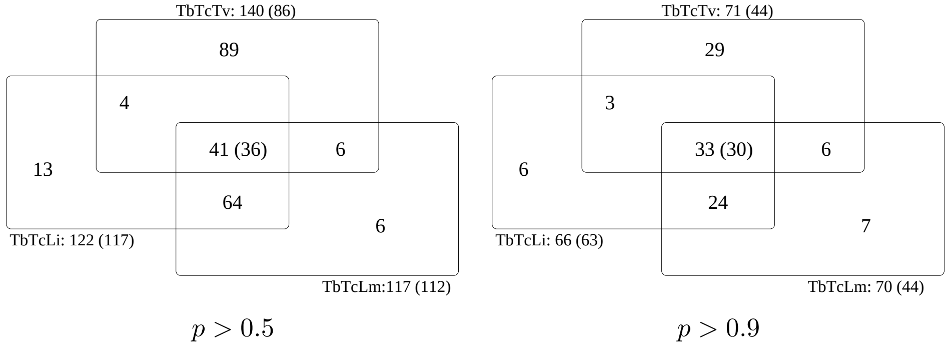


Figure 4.19: Exemplary consensus structures of the TbTcLm ncRNAs.
(A) Structure 44039, p=0.999, known as 18S SSU rRNA
(B) Structure 44110, p=0.888, known as M4 LSU epsilon rRNA
(C) Structure 44113, p=0.939, known as M2 LSU delta rRNA
(D) Structure 44142, p=0.695, `NCBI` blast search indicates *T. brucei* elongation factor 1-alpha, so it belongs to the false positive set.
(E) Structure 44136, p=0.867, putative novel ncRNA, yet unannotated, a `NCBI blast` search revealed no further annotation
(F) Structure 44137, p=0.997, putative novel ncRNA, yet unannotated, a `NCBI blast` search revealed no further annotation

4.1.5 Common ncRNA signals of the Trypanosoma screens

It is straight forward to be interested in novel ncRNAs which appear in all screens. RNAs conserved through the three-way alignment approaches imply functionality and thus promise to be the most important ones. We note that 41 ncRNAs are conserved between the three-way Trypanosoma screens. We are able to annotate 36 out of these 41 predictions. In this context we speak of annotatable if the related candidates of all three screens are annotatable. Figure 4.20 illustrates the resulting sets of the comparison at the $p > 0.5$ and $p > 0.9$ level. These well conserved ncRNAs are listed at [1].



$p > 0.5$ $p > 0.9$

Figure 4.20: Venn diagrams illustrating the commonalities of the Trypanosoma screens; Numbers in brackets indicate the count of annotatable predictions.

4.2 Novel ncRNA candidates of the genus Leishmania

In this section we report about our predicted ncRNA candidates of the *Leishmania infantum* and *Leishmania major* genomes. The numbers of observed ncRNA signals are listed in table 4.18

p	LiLm	LiLmTb	LmLi	LmLiTb
> 0.50	113,575	92	30,762	131
> 0.90	60,506	50	15,632	78
> 0.98	18,239	29	8,633	30
> 0.99	13,456	18	6,667	21

Table 4.18: Overview of the numbers of Leishmania ncRNA prediction. The pairwise screens became meaningless because of the high hit rates. The genomes are related to close and we assume that not all repeats are annotated completely. We do not care about these pairwise screens in further analyses because of the insignificant overrepresentation of ncRNAs. Although *T. brucei* is close related to the Leishmania species, three-way alignments reduce the amount of hits dramatically and induce more reliability.

The Li noncoding DNA used for the initial `blast` search comprises 19,065,849 nt what is nearly 55% of the total genomic sequence. The Lm noncoding DNA consists of 15,669,214 nt what is less than

[1] `http://www.bioinf.uni-leipzig.de/~dominic/projects/tryp/index.php?id=results`

48% of the total genome. The pairwise predictions seem to be irredeemably overestimated. The high numbers of ncRNA candidates are due to the fact that the two *Leishmania* species are related too close phylogenetically and the existence of putative unannotated repeats. Annotation and evaluation of the pairwise approaches would be expensive because of the many ncRNA candidates and it even seems to be unsure because of the missing third genome providing essential data. So we decide to concentrate on Leishmania screens with three-way alignments only. Those predictions are definitely more reliable.

4.2.1 The LiLmTb approach

Based on three-way alignments of Li, Lm and Tb we detect 92 structured RNA signals ($p > 0.5$, Tab. 4.18) of which 76 are annotatable. Due to the definitions of subsection 3.1.3 we obtain the specifity values and false positive rates shown in table 4.19, the resulting sensitivity values are shown in table 4.20. Observed detection rates are shown in table 4.21.

p	> 0.50	> 0.90	> 0.98	> 0.99
False positive rates	26 %	17 %	17 %	15 %
	(133/506)	(34/200)	(14/81)	(8/52)
Specificity per test	0.984	0.996	0.998	0.999

Table 4.19: False positive rates and specifities of the LiLmTb screen

type	N	n	N_a	N_g	s_{Na}	s_{Ng}	s_{na}	s_{ng}
rRNA	21	35	44	62	0.47	0.33	0.80	0.56
tRNA	38	43	57	62	0.67	0.61	0.75	0.69
misc_RNA	0	0	0	14	0.00	0.00	0.00	0.00
snRNA	1	1	1	7	0.14	0.14	0.14	0.14
snoRNA	0	0	0	43	0.00	0.00	0.00	0.00

Table 4.20: Estimated sensitivities of the LiLmTb screen; Out of the rRNAs we hit the known 5.8S, 18S and 28S; Among others the given misc_RNAs are pseudogenes (so it is good not to find them).

	per 1 mb alignment		per 1 mb nc region	
p	normal	shuffled	normal	shuffled
> 0.50	12.79	3.36	26.54	6.98
> 0.90	5.06	0.86	10.49	1.78
> 0.98	2.05	0.35	4.25	0.73
> 0.99	1.31	0.20	2.73	0.42

Table 4.21: Noncoding RNA detection rates of the LiLmTb screen.

Prior pairwise `blast` hits are used for global alignments with Lm yielding to 14,580 `blast` hits with an average length of ~37 nt. Only 2,504 of them made it into the RNAz input set. Their average length is ~71 nt. In consideration of all possible reading directions the total alignment input set comprises 5,008 alignments. Overall there are 8,544 RNAz frames. At the $p > 0.5$ level we count 506 and with $p > 0.9$ we observe 200. Figure 4.21 illustrates the complete distribution of structured RNAz frames.

tRNAscan-SE produced 56 tRNA predictions. 54 of them decode for standard amino acids and 2 are classified as pseudogenes. Their average length is 72 nt. A blast search of the Li predictions against the complete Noncode database affects 2 of our ncRNAs (22 hits). Overall a run of RNAmicro reveals 111 hits. Still 2 of them are classified with $p > 0.5$. But a blast search against the miRBase produced no hit. A blast search against the Rfam leads to 298 hits affecting 65 of our ncRNA candidates. A search after RNAs homologous to the Rfam covariance models produced 135 hits covering 81 of our ncRNA predictions. The distribution of the obtained bitscores is shown in figure 4.23. After filtering (cp. restrictions mentioned at 3.2) we keep 62 annotations. Furthermore we obtain no blast hits with sequences of the snoRNA-LBME-db. Results of the RNAbob searches are listed in table 4.22. Exemplary secondary structures of conserved RNAbob hits are illustrated in figure 4.22.

The ncRNA candidates are groupable into 79 cluster by the first and into 64 cluster by the second Blastclust run. The maximal observed cardinality is 2 (first run) and 4 (second run) indicating that the predictions are less similar on the level of primary sequence than they are at the Trypanosoma screens. A specific annotation for each cluster seems complicated. For example Blastclust combines predictions that we have annotated as tRNAs, known Noncode smnRNAs of *Leishmania tarentolae* and 7SL SRP rRNA of *Leptomonas collosoma* into one cluster.

Finally we present some exemplary ncRNA consensus structures of the LiLmTb screen at figure 4.25. Its caption is shown here because of space problems:

(A) Structure 40687, p=0.975, putative unannotated tRNA in Li, blast searches against the Rfam and the NCBI indicate valine tRNA (*L. tarentolae*)
(B) Structure 40688, p=0.779, putative unannotated tRNA in Li, blast searches against the Rfam and the NCBI indicate valine tRNA (*L. tarentolae*)
(C) Structure 40700, p=0.634, matches with 5S rRNA by NCBI blast
(D) Structure 40741, p=0.831, known as 28S LSU alpha rRNA
(E) Structure 40748, p=0.999, miRNA candidate found by RNAmicro(p=0.988); We present the complete RNAz hit, the flanking regions should be trimmed.
(F) Structure 40745, p=0.830, putative novel ncRNA, looks like tRNA, but it is not annotatable
(G) Structure 40702, p=0.878, NCBI blast search indicates *T. cruci* histon H3 gene, so it belongs to the false positive set.
(H) Structure 40705, p=0.508, NCBI blast search indicates *L. donovani* S11 mRNA gene and *T. brucei*/*T. cruci* 40S ribosomal protein, so it belongs to the false positive set.
(I) Structure 40782, p=0.973, NCBI blast search indicates *L. braziliensis* elongation factor 2, so it belongs to the false positive set.
(J) Structure 40783, p=0.777, NCBI blast search indicates *L. major* or *T. cruci* elongation factor 2, so it belongs to the false positive set.

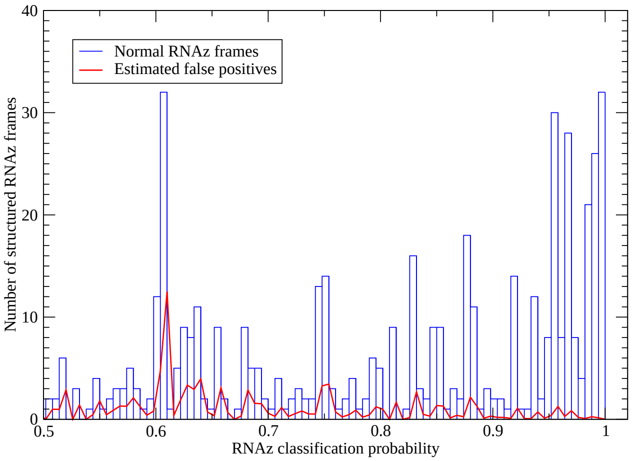


Figure 4.21: Histogram of the `RNAz` classification probability of the LiLmTb screen. The peak at $p = 0.6$ implies repetitive structures.

descriptor	# hits	# conserved hits	# unique predictions
Common	1	1	1
L. seymouri	2	2	2
C. fasciculata	0	0	0
T. brucei	9	1	1

Table 4.22: Number of ncRNA candidates for SMN binding of the LiLmTb screen. Conserved in this context indicates that we count hits where the motif is found in the majority of the aligned sequences. Unique indicates the number of involved ncRNA predictions with different secondary structure.

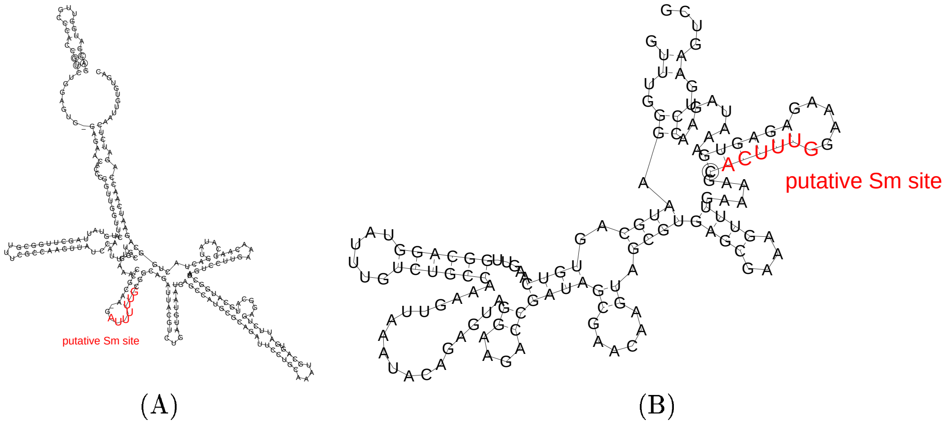


Figure 4.22: Exemplary `RNAbob` hits of the LiLmTb screen; (A) structure: 40735, found with the common and the *L. seymouri* descriptor; (B) structure: 40738, found with the *T. brucei* descriptor;

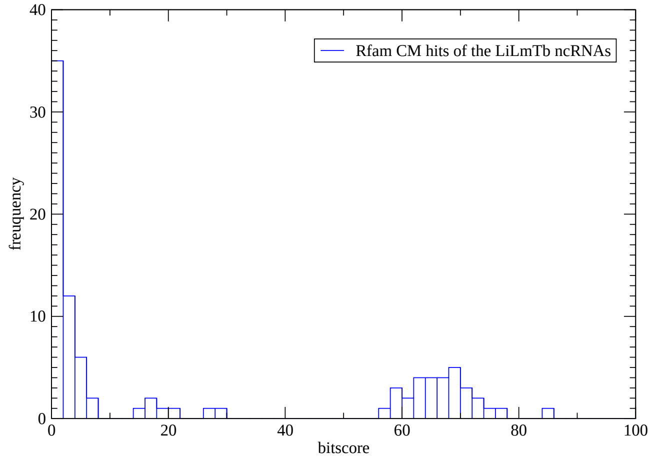


Figure 4.23: Bitscores obtained by `Infernals` `cmsearch` of the LiLmTb screen. The bitscore is problematic, there are so many hits with low score, we set a cutoff value at 10.

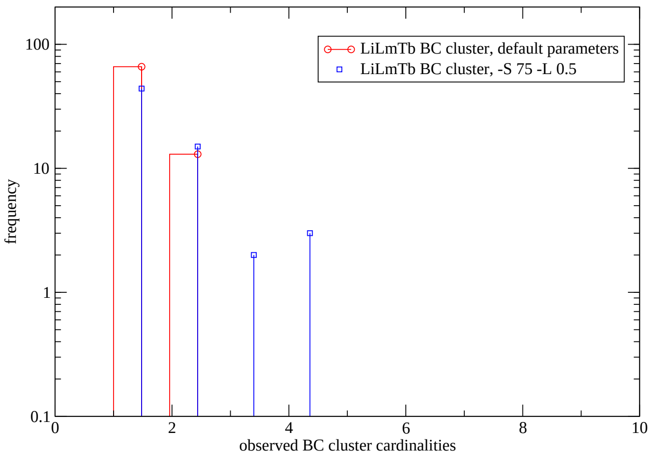


Figure 4.24: Histogram of the obtained `Blastclust` cardinalities of the LiLmTb ncRNAs. The y-axis is scaled logarithmically. The major fraction of the predicted ncRNAs with low cluster cardinalities seems to be non-repetitive. Although there are some clusters with more than one member, the screen is less repetitive on the level of primary sequence than the prior screens of Trypanosoma are.

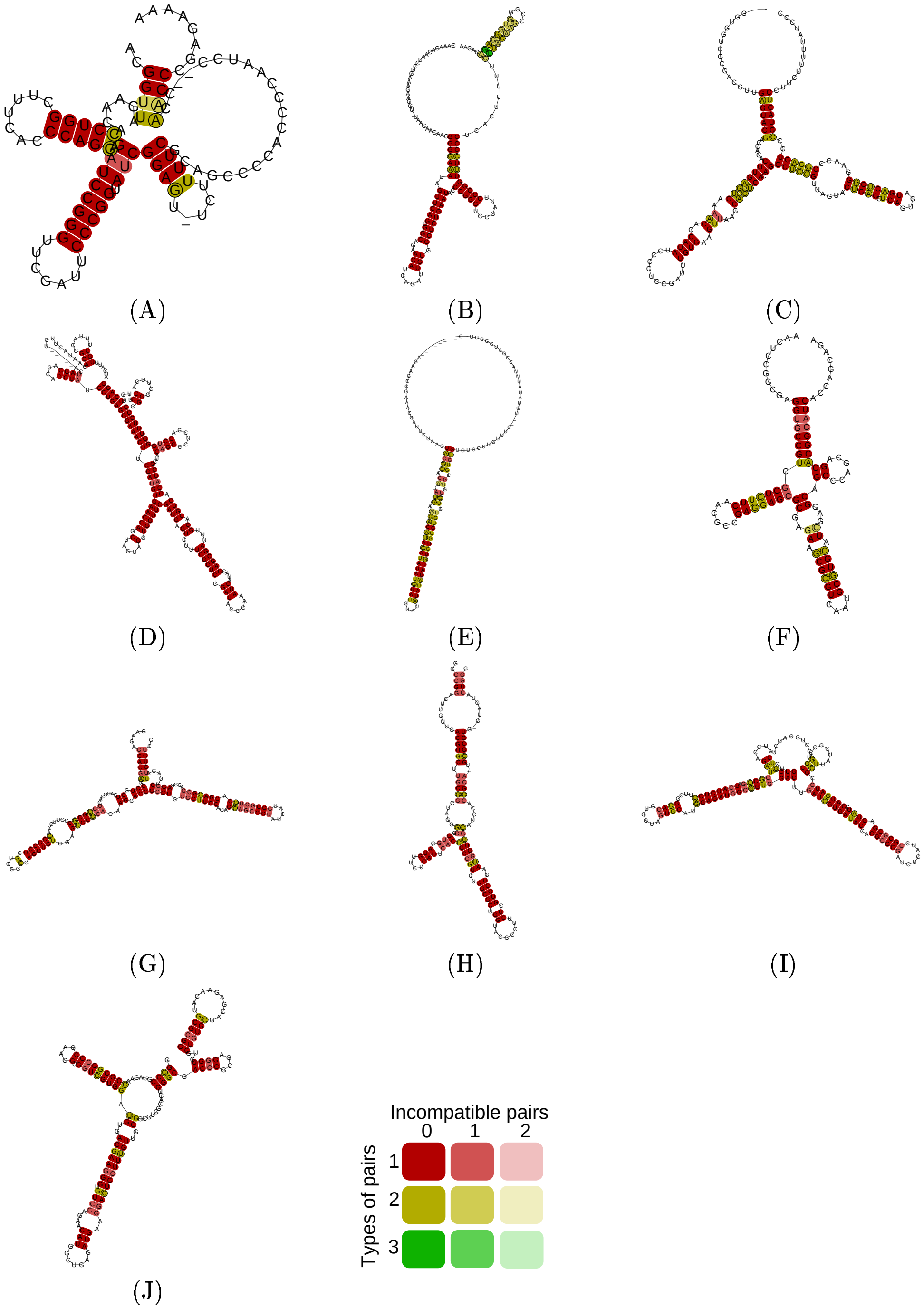


Figure 4.25: Exemplary consensus structures of the LiLmTb ncRNAs. Further details are given at the end of section 4.2.1 because of space problems.

4.2.2 The LmLiTb approach

Based on three-way alignments of Lm, Li and Tb we detect 131 structured RNA signals ($p > 0.5$, Tab. 4.18) of which 116 are annotatable. Due to the definitions of subsection 3.1.3 we obtain the specifity values and false positive rates shown in table 4.23, the resulting sensitivity values are shown in table 4.24. Observed detection rates are shown in table 4.25.

p	> 0.50	> 0.90	> 0.98	> 0.99
False positive rates	33 %	24 %	29 %	17 %
	(497/1498)	(128/523)	(33/114)	(16/93)
Specificity per test	0.961	0.990	0.997	0.999

Table 4.23: False positive rates and specifities of the LmLiTb screen

type	N	n	N_a	N_g	s_{Na}	s_{Ng}	s_{na}	s_{ng}
rRNA	28	34	46	63	0.61	0.44	0.74	0.54
tRNA	54	60	82	83	0.66	0.65	0.73	0.72
misc_RNA	0	0	0	61	0.00	0.00	0.00	0.00
snRNA	2	2	2	7	1.00	1.00	0.29	0.29
snoRNA	0	0	0	693	0.00	0.00	0.00	0.00

Table 4.24: Estimated sensitivities of the LmLiTb screen; Out of the rRNAs we hit the known 18S and 28S rRNAs, but we missed the known 5.8S rRNAs; Among others the given misc_RNAs are mostly mini-exon genes (so it is good not to find them).

	per 1 mb alignment		per 1 mb nc region	
p	normal	shuffled	normal	shuffled
> 0.50	52.60	17.45	95.60	31.72
> 0.90	18.37	4.49	33.38	8.17
> 0.98	4.00	1.16	7.28	2.11
> 0.99	3.27	0.56	5.94	1.02

Table 4.25: Noncoding RNA detection rates of the LmLiTb screen.

Prior pairwise `blast` hits are used for global alignments with Li yielding to 3,817 `blast` hits with an average length of ~107 nt. Only 1,570 of them made it into the `RNAz` input set. Their average length is ~188 nt. In consideration of all possible reading directions the total alignment input set comprises 3,140 alignments. In comparison to the LiLmTb screen the number of unfiltered `blast` hits is pretty higher (14,580 at LiLmTb and 3,817 at LmLiTb). This is due to the heavy impact of 16,392 annotated Lm repeats ('repeat_region', 'repeat_unit'). That is about 450 % of known Li repeats (3,658). We only cut repeats and coding sequences of the reference genome. The subject genome of the `blast` searches remains untouched. Overall there are 12,879 `RNAz` frames. At the $p > 0.5$ level we count 1,498 and with $p > 0.9$ we observe 523. Figure 4.26 illustrates the complete distribution of structured `RNAz` frames.

`tRNAscan-SE` produced 49 tRNA predictions. 48 of them decode for standard amino acids and one is classified as pseudogene. Their average length is 72 nt. A `blast` search of the Lm predictions against the complete `Noncode` database affects 2 of our ncRNAs (22 hits). Overall a run of `RNAmicro` reveals 231 hits. Still 6 of them are classified with $p > 0.5$. But a `blast` search against the `miRBase` produced no hit. A `blast` search against the `Rfam` leads to 298 hits affecting 62 of our ncRNA candidates. A search after RNAs homologous to the `Rfam` covariance models produced 163 hits covering 100 of our ncRNA predictions. The distribution of the obtained bitscores is shown in figure 4.28. After filtering (cp. restrictions mentioned at 3.2) we keep 56 annotations. Furthermore we obtain no `blast` hits with sequences of the `snoRNA-LBME-db`. Results of the `RNAbob` searches are listed in table 4.26. Exemplary secondary structures of conserved `RNAbob` hits are illustrated in figure 4.27.

The ncRNA candidates are groupable into 74 cluster by the first and into 54 cluster by the second `Blastclust` run. The maximal observed cardinality of both runs is 10 indicating that we detect the same or in fact very similar ncRNAs not more than 10 times per genome. The cluster with the maximal cardinalities comprise rRNAs for both BC runs. The largest groups contain 28S LSU epsilon (M4) and LSU alpha.

Finally we present some exemplary ncRNA consensus structures of the LmLiTb screen at figure 4.30.

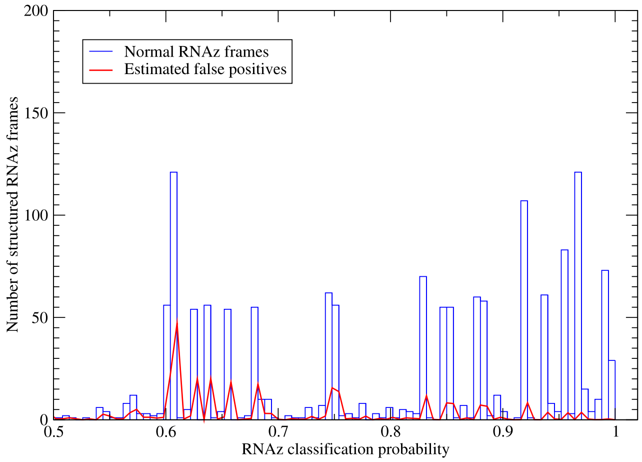


Figure 4.26: Histogram of the `RNAz` classification probability of the LmLiTb screen. Similar to the LiLmTb screen we observe a peak at $p = 0.6$ inducing repetitive structures.

descriptor	# hits	# conserved hits	# unique predictions
Common	6	6	1
L. seymouri	6	6	1
C. fasciculata	0	0	0
T. brucei	14	6	1

Table 4.26: Number of ncRNA candidates for SMN binding of the LmLiTb screen. Conserved in this context indicates that we count hits where the motif is found in the majority of the aligned sequences. Unique indicates the number of involved ncRNA predictions with different secondary structure.

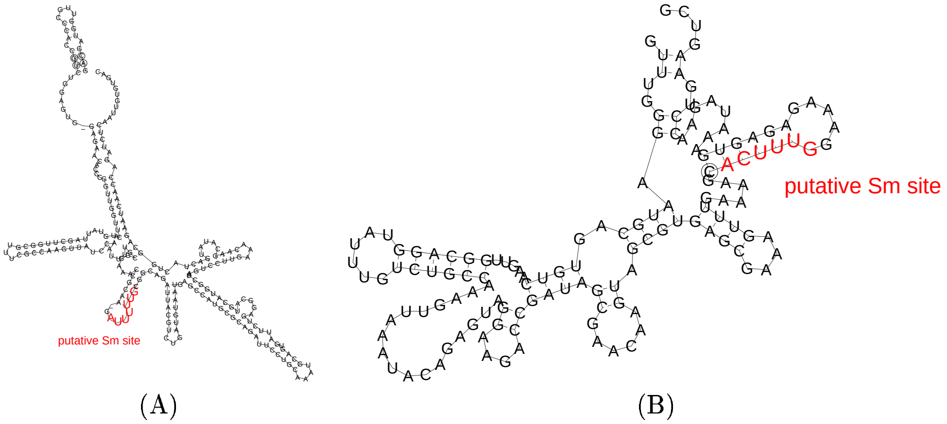


Figure 4.27: Exemplary `RNAbob` hits of the LmLiTb screen; (A) structure: 68483, found with the common and the *L. seymouri* descriptor; (B) structure: 468485, found with the *T. brucei* descriptor;

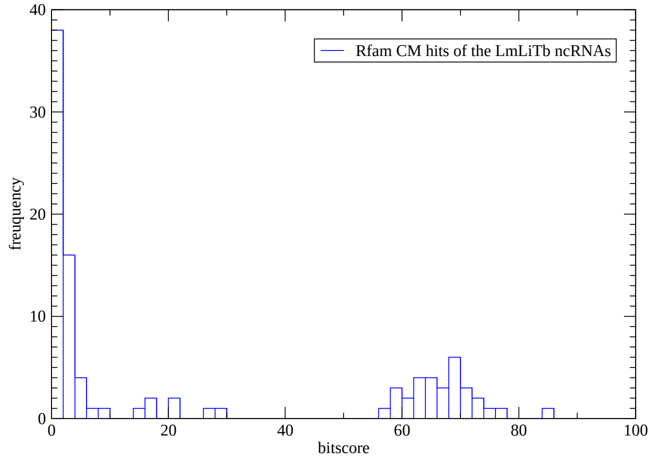


Figure 4.28: Bitscores obtained by `Infernals cmsearch` of the LmLiTb screen. The bitscore is problematic, there are so many hits with low score, we set a cutoff value at 10.

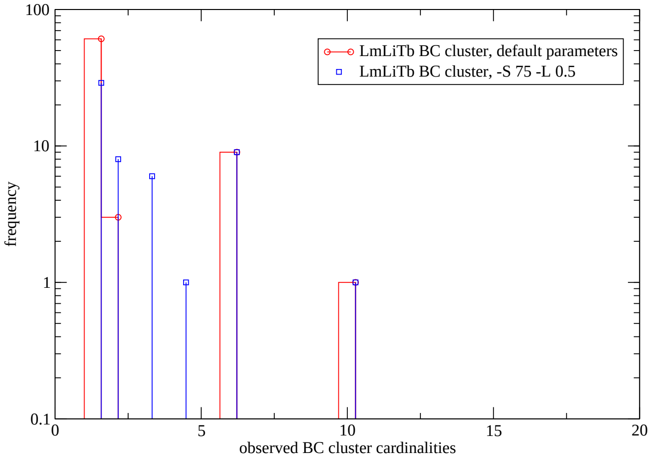


Figure 4.29: Histogram of the obtained `Blastclust` cardinalities of the LmLiTb ncRNAs. The y-axis is scaled logarithmically. The major fraction of the predicted ncRNAs with low cluster cardinalities seems to be non-repetitive. Contrary to the LiLmTb screen, we now observe cluster cardinalities of more than 10. Thus the genome of *Leishmania major* consists of more repetitive elements than *Leishmania infantum*.

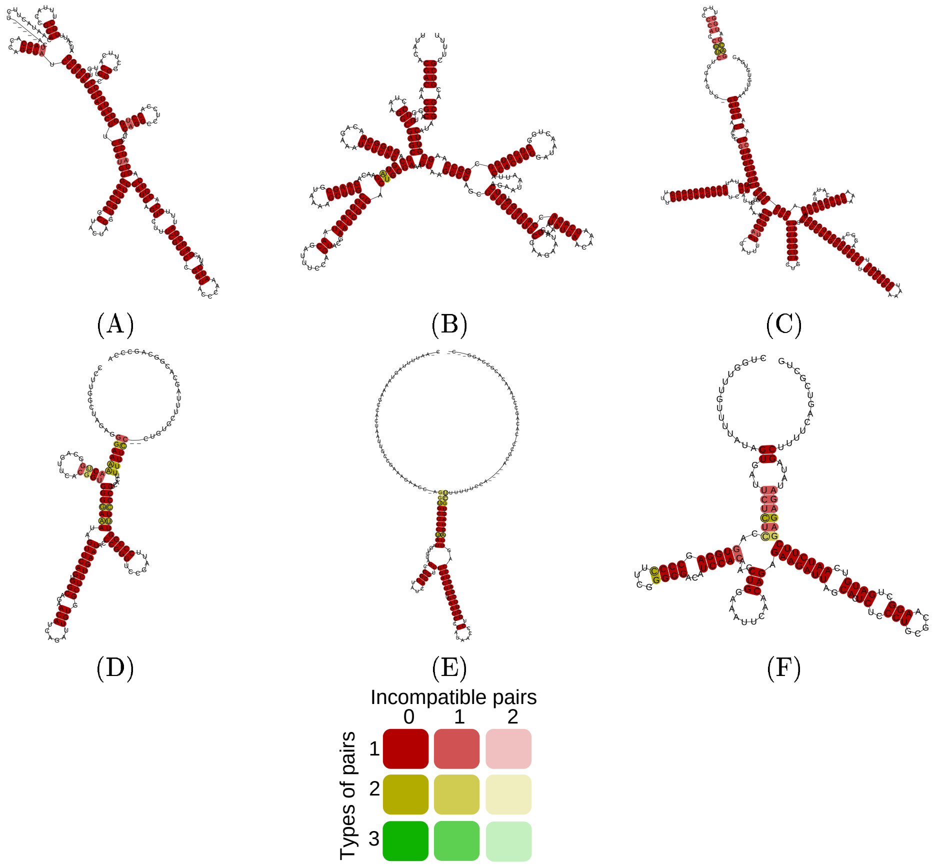


Figure 4.30: Exemplary consensus structures of the LmLiTb ncRNAs. The unpaired nucleotides at the ends of the structures (for example D, E and F) indicate that we do have problems with assigning the correct RNA positions absolutely. But this issue is due to the decision of aligning conserved regions with some flanking nucleotides. The borders are too wide but the RNA structure is not cut and thus completely detected.
(A) Structure 68502, p=0.831, known as 28S LSU alpha rRNA
(B) Structure 68491, p=0.935, known as 28S LSU beta rRNA
(C) Structure 68497, p=0.970, known as 5.8S SSU rRNA
(D) Structure 68441, p=0.925, known as His tRNA
(E) Structure 68594, p=0.999, known as Gln tRNA
(F) Structure 68477, p=0.920, known U6 snRNA

4.2.3 Common ncRNA signals of the Leishmania screens

It is straight forward to be interested in novel ncRNAs which appear in all screens. RNAs conserved through the three-way alignment approaches imply functionality and thus promise to be the most important ones. If we bring the three-way Leishmania screens together by combination of ncRNAs which appear in both screens we note 131 cluster. This is done by a comparison of conserved Lm ncRNAs between the screens. It is interesting to notice that all predictions of the LiLmTb screen have their corresponding prediction in the LmLiTb screen. Figure 4.31 illustrates the resulting sets. Related ncRNAs are listed at [2].

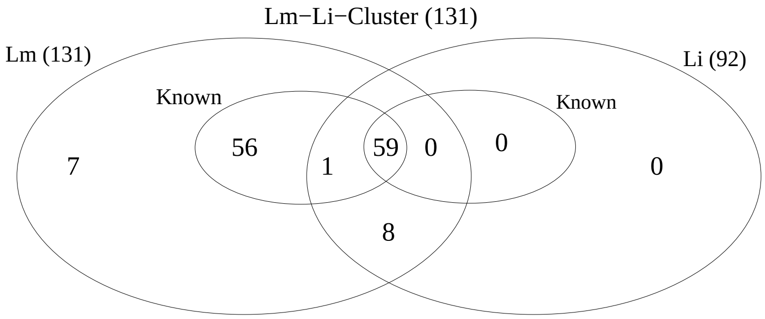


Figure 4.31: Venn diagram illustrating the commonalities of the Leishmania screens; All 92 Lm sequences of the LiLmTb screen can be grouped together with Lm sequences of the LmLiTb screen.

[2] http://www.bioinf.uni-leipzig.de/∼dominic/projects/leish/index.php?id=results

Chapter 5

Discussion

5.1 Promising results

The reliable prediction of ncRNAs is a challenging task and it is very hard to proof those results, e.g. in terms of mathematical evidence. Regardless of this problem we try to validate our predictions by functional assignments. Figure 5.1 summarizes the numbers of our predicted ncRNAs and their annotatable fractions. The majority of the ncRNA candidates is annotatable by automated methods and we conclude in consideration with these results to have shown promising possibilities of computational ncRNA detection. The functions of only a minor part remains unidentifiable. The prediction recovers the majority of conserved known noncoding elements which allows the conclusion that our procedure works quite good among trypanosomatid species. Considering the overall reliability of noncoding predictions and the false positive rates we do not perform worse within trypanosomes than we did in our prior screens of *Ciona intestinalis* or *Caenorhabditis elegans*.

`tRNAscan-SE` has established through the years and is a common, appropriate and very fast tRNA detection tool. This tRNA finder has a complimentary detection rate of 99-100 %[39]. This sounds nice and comparing the observed hit rates of our candidates with known tRNAs we can state that, overall, we do not perform worse (cp. figure 5.2) than `tRNAscan-SE` in sum.

A major fraction of our candidates refers to known elements pleading the prediction procedure, but there are also novel up to now unassigned elements among the result set. Our screens include structured RNAs with putative or even unknown function but also missclassified coding elements. As mentioned above, manual interference like `blast` searches against the `NCBI` help to validate the predictions. Table 5.1 provides an overview of the extant predictions without an automated function-assignment.

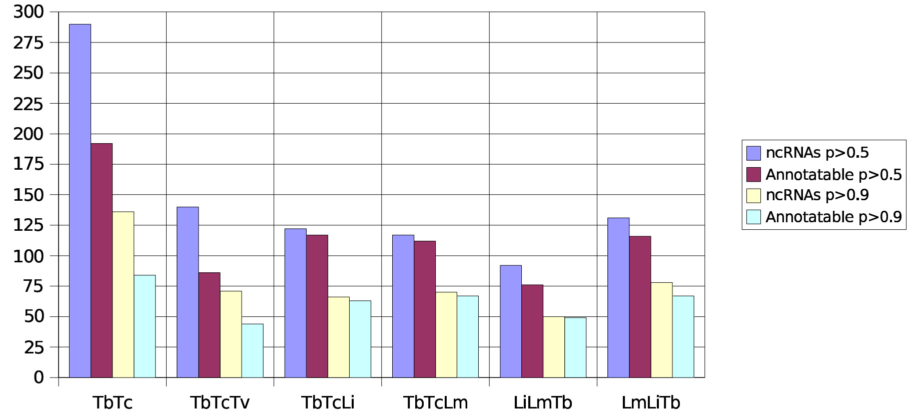


Figure 5.1: Summary of predicted ncRNAs. From the left to the right we summarise the amount of predicted ncRNAs and the number of annotatable ones at the $p > 0.5$ and the $p > 0.9$ level for each screen. For annotation purposes of this diagram we only considered automated methods like blast searches against ncRNA databases like `Noncode` and `Rfam`. Manual assignments, for example **`blast`** searches against the `NCBI`, are not taken into account here because we want to keep the whole procedure as much automated as possible. The difference of the numbers of predicted RNAs and their annotatable subset is much bigger for the TbTcTv screen than it is among the other three-way alignment approaches. This effect is caused by the mentioned repetitive mini-exon genes conserved among the trypanosomes (cp. table 4.6).

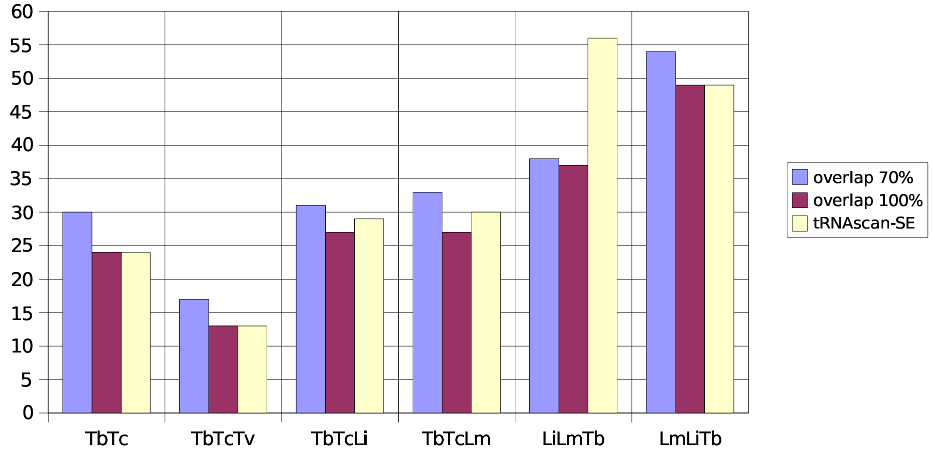


Figure 5.2: `RNAz` challenges `tRNAscan-SE`: We mention to identify almost the same number of known tRNAs with our `RNAz` prediction pipeline than `tRNAscan-SE` reveals in our candidate sets. For each screen `tRNAscan-SE` was run on the `RNAz` ncRNA candidate set and our predictions were compared with given known tRNAs. The comparison was done by calculating percentages of overlapping genomic loci. `RNAz` hit rates increase if we allow an overlap of 70 % indicating that we do not match the known RNAs perfectly due to the `RNAz` sliding window mechanism, but we are able to detect signaling conserved subregions.

screen	TbTc	TbTcTv	TbTcLi	TbTcLm	LiLmTb	LmLiTb
#	98 (38 %)	54 (39 %)	5 (4 %)	5 (4 %)	16 (17 %)	15 (12 %)
coding			2	2	7	6
noncoding			1	1	3	6
unknown			2	2	6	3

Table 5.1: Overview of ncRNA candidates that lack automated function-assignment. Predictions of the most reliable three-way alignment screens without an annotation due to automated procedures are curated manually by `blast` searches against the `NCBI` databases. The hit set can be grouped into coding elements, noncoding RNAs and structures without a clear homology to known sequences. The coding set and thus the false positive set comprises hypothetical proteins, elongation factors, some ribosomal proteins and other coding genes (e.g. rod proteins or histone H3 genes). The `blast` results indicate that the major fraction of those coding signals is trypanosomatid specific (the phylogenetic range of the hit set is almost limited to trypanosomatid taxa). This could be a hint why `RNAz` missclassified them (`RNAz` mostly is trained with vertebrate data).

5.2 Limitations, improvements and future work

Although we got positive feedback by the successful mapping of predicted ncRNAs to known elements we strongly emphasize that we only present methods for ncRNA prediction and in the end it is hard to say what are true ncRNAs. Verifying predictions with other prediction methods only increases certainty if they are completely reliable, but which method features this undoubtedly? The comparison of our candidate set with known annotated RNAs is reliable if the annotation is confirmed and thus reliable, but even our starting data sets are full of predictions.

`Blast` alignments of candidate sequences with ncRNA databases are a common method for functional assignment. The unpublished `RNAmicro` is considered to be a qualified miRNA detector although allover `RNAmicro` experience is comparatively low because it is a new tool. The `Infernal` runs contain only little information because the provided bitscore is hard to interpret. There is no valid distribution or scoring scheme recognizable. Thus we tried to exalt the quality of obtained hits by setting bitscore cutoff values and neglecting improper `Rfam` model families. We did not use the number of `RNAbob` hits for annotation statistics. This pattern matching approach is very quick, but seems to be dirty, too. The possibilities of formulating applicable RNA class descriptors are limited, however. It is allegeable that we do not get any hit with the `snoRNA-LBME-db` and the `miRBase`. On the one hand the `snoRNA-LBME-db` only contains sequences of human H/ACA and C/D box snoRNAs and the phylogenetic distance between human and trypanosomatid species seems to be too large. On the other hand the miRNA sequences are too short to be identified significantly with `blast`.

Functions of unassigned predictions remain unclear without strong similarities to known ncRNAs. We even notice `RNAz` hits without a significant `NCBI` `blast` hit to known elements, whether they belong to coding or noncoding classes. However, we also know that `blast` can not find everything and there are false positives in the result set. The next step of identification and thus verification of ncRNAs could be performing lab experiments like reverse transcriptase-dependent PCR, microarray or northern blot analyses[47],[48]. Unfortunately it is difficult to run them in large-scale and genomewide contexts as they are definitely more expensive than computational approaches. Their benefit is discussable if computational gained results are handled carefully and with a substantially portion of mistrust.

A major problem of our method is that we do not match known RNAs perfectly and our annotation differs in some nucleotides from the positions of known elements (cp. table 5.2). As a consequence of the RNAz sliding window mechanism, we sometimes could hit specific subregions instead of complete RNAs. Even though we annotate those partial hits as their known counterpart. Furthermore, we have the problem of providing two start and end positions for each detected structure. After various merging steps of combining single RNAz frames we annotate a segment ("adjusted region") as ncRNA summarizing all signals pointing to this locus, but it is more precise (more "biological") to annotate loci retrieved by a single alignment. We therefore tried to choose a best representative of a single alignment ("best RNA") for each "adjusted region". It is straight forward that a "best RNA" can be shorter than its "adjusted region" (cp. figure 5.3). Comparisons of positions of our ncRNAs to known elements are done with the positions of the "adjusted regions". Additionally consensus folding is done with "best RNAs" to preserve their biological context (cp. chapter 3). This could result in some "unexpected" consensus structures, however. Although we annotate an ncRNA candidate as "tRNA" for example, parts of the consensus structure do not need to look like a tRNA.

screen	TbTc	TbTcTv	TbTcLi	TbTcLm	LiLmTb	LmLiTb
rRNA	56/82	23/33	28/61	31/56	20/21	12/28
tRNA	24/30	13/17	27/31	27/33	37/38	49/54

Table 5.2: The annotation of ncRNA predictions differs with corresponding known elements. For every screen we list the number of predictions that correspond to known rRNAs and tRNAs. The first number describes the number of matches with known RNAs given an overlap of 100 % is required, the second one indicates the amount with an overlap of only 70 %. Obviously the second number is higher than the first concluding that we do not match known elements completely but major signalling subregions remain detectable.

RNAz itself is limited to six-way alignments. In our case we did not exhaust this issue, we used three genomes per screen at the most. This is a practicable compromise between data availability, data quality and computing time. More and more sequencing projects are ongoing but for prediction purposes we need several progressed genomes out of specific taxons. The phylogenetic distance between the underlying organisms has to be taken into account. If they are too closely related, everything aligns with everything and a reliable SVM classification is not possible. This issue is especially observable within the pairwise alignment screens. If the organisms are related too distantly the signal is lost and ncRNA detection is problematic. Prediction becomes uncertain if we screen close related organisms only. This is the case for TbTc, TbTcTv, LiLm and LmLi. About 60 % of the putative ncRNAs of the TbTc and TbTcTv screen are annotatable. If we concentrate on sequences conserved between Trypanosoma and Leishmania we note a considerable increase of annotatable predictions. There we note that 96 % of the TbTcLi and the TbTcLm candidates and 83 % respectively 89 % of the LiLmTb and LmLiTb screen are annotatable. However, we could miss species specific elements if we align distantly related sequences. Computation time may become

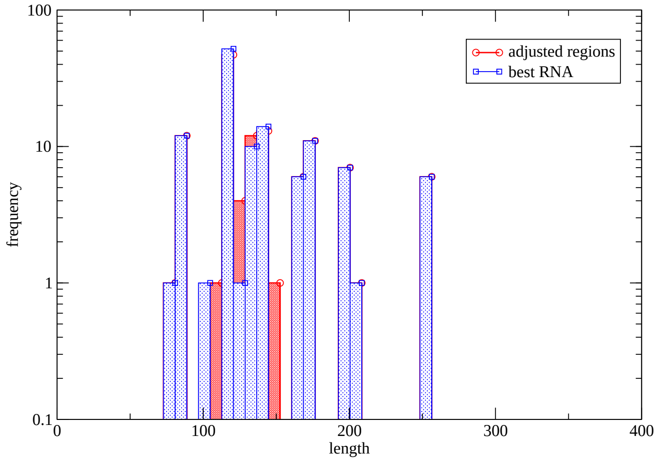


Figure 5.3: Exemplary length distribution of TbTcLi ncRNA candidates. The lengths of the "adjusted regions" and their "best RNAs" do not differ significantly. Nevertheless it is precise to distinguish between "adjusted region" and "best RNA" to preserve the biological context of a single alignment. The peak at 120 nt is due to the screening window size of 120 nt.

a problem if large genomes are at hand. Repetitive hits have to be curated immediately during the creation of the alignments. Although we tried to exclude annotated repeats, we can not be sure that all repeats are masked completely in the given annotation. Usually one query sequence is found several times in genomewide `blast` searches. This multiplicative effect grows with the number of genomes per screen and complicates the whole process of setting up multiple `RNAz` input alignments. Moreover, we screened alignments consisting of every combination of possible reading directions (there are 2^n possibilities with n genomes). This considerably increases the amount of the `RNAz` input data and the overall computation time.

It is obvious that predictions depend on the initial screen design and underlying source data. We set up two individual Leishmania screens such that we looked at conserved loci in two ways. On the one hand we started with Li and on the other hand we started with Lm noncoding regions. We expected to see similar `blast` hits and in the end similar ncRNA candidates. However, the starting genomes and their underlying annotation differ which results in some differences between the two screens. Anyway, all LiLmTb ncRNAs predictions have their counterpart within the LmLiTb screen. In addition, we observe some further signals in the LmLiTb screen. We conclude that the quality of the whole procedure simply gets worse with the abundance of repeats. In case of repetitive hits it is difficult to decide if valid and functional RNAs have been observed or if only relicts of duplication events have been detected. The `Blastclust` runs confirm repetitive hits among our predictions.

Indeed, we observe a considerable amount of structured RNAs in the shuffled test sets. The number of positive classified `RNAz` frames in the shuffled approaches (false positives) can not be argued away. The more repetitive a data set is, the higher the false positive rate will be. With the increase of repeats in the dataset it becomes more difficult to destroy every structural signal by shuffling the alignments. Columnwisely randomised alignments may not be the best what we can do, but it is a first step in creating some sort of false positive rate. Figure 5.4 illustrates some discrepancies of SVM classification results on native data sets in comparison with shuffled alignments. The rating of the predictions and the creation of a reliable false positive rate remains a open problem.

As mentioned above, we only performed screens with three organisms in this work. It is obvious that it is useful to have a general set of tools at hand to handle the RNAz maximum of six organisms in an automatic way. Up to now we did not implement this general computational environment although the basis with a well normalized database scheme and related input and output perl modules is ready. The currently implemented procedure meets all requirements to handle three organism, but performance improvements are still conceivable, especially at the creation of the input alignments, using specialised multiple alignment approaches like the threaded blockset aligner TBA[49] or LAGAN (Multi-LAGAN)[50]. This will automate the procedure promisingly once again, but less influence on the inputset are a consequence. In the last days of this work RNAz 1.0 was released. It partly supports those features in ready-made perl scripts.

Although we know that our implemented procedure works quite fine among different taxonomic classes, we decide not to publish the complete system. Integration of all scripts to a general tool of ncRNA prediction is not practicable because of missing file format standards and changing requirements. We therefore refer to the new RNAz 1.0 release. The above mentioned perl scripts allow large-scale genomic screens from the scratch (without database support).

Automation is the key word if we ask for ncRNA prediction at taxons where the amount of available genomes and annotation data is high, however. Among others this is the case for a lot of bacterial organisms. However, we do not know of comprehensive and published RNAz ncRNA predictions affecting Eubacteria, for instance, up to now. Perhaps there are also possibilities of ncRNA prediction in taxons where the situation of available genomes looks even worse. Imagine genomes with high occurrences of gene duplication events. It has been shown that a major fraction of the genome of *Arabidopsis thaliana* consists of paralogous genes that probably originated through one or more ancient large-scale gene or genome duplication events[51] and moreover that duplicated RNA copies acquire new functionality as they evolve[52],[53]. What happens if we align the genome to itself? First identical hits have to be excluded and then setting up multiple alignments as RNAz input may also be possible.

A next step in the RNAz based ncRNA prediction could be to extend the range of the SVM classification. Beside “valid ncRNA” and “other” it rather would be interesting to have a specialised classification into resulting ncRNA types (tRNA, rRNA, miRNA, snRNA. snoRNA, ...) implemented in one single program. We admit that this is not yet possible and new algorithms factoring more ncRNA characteristics have to be developed to reach this pretentious goal.

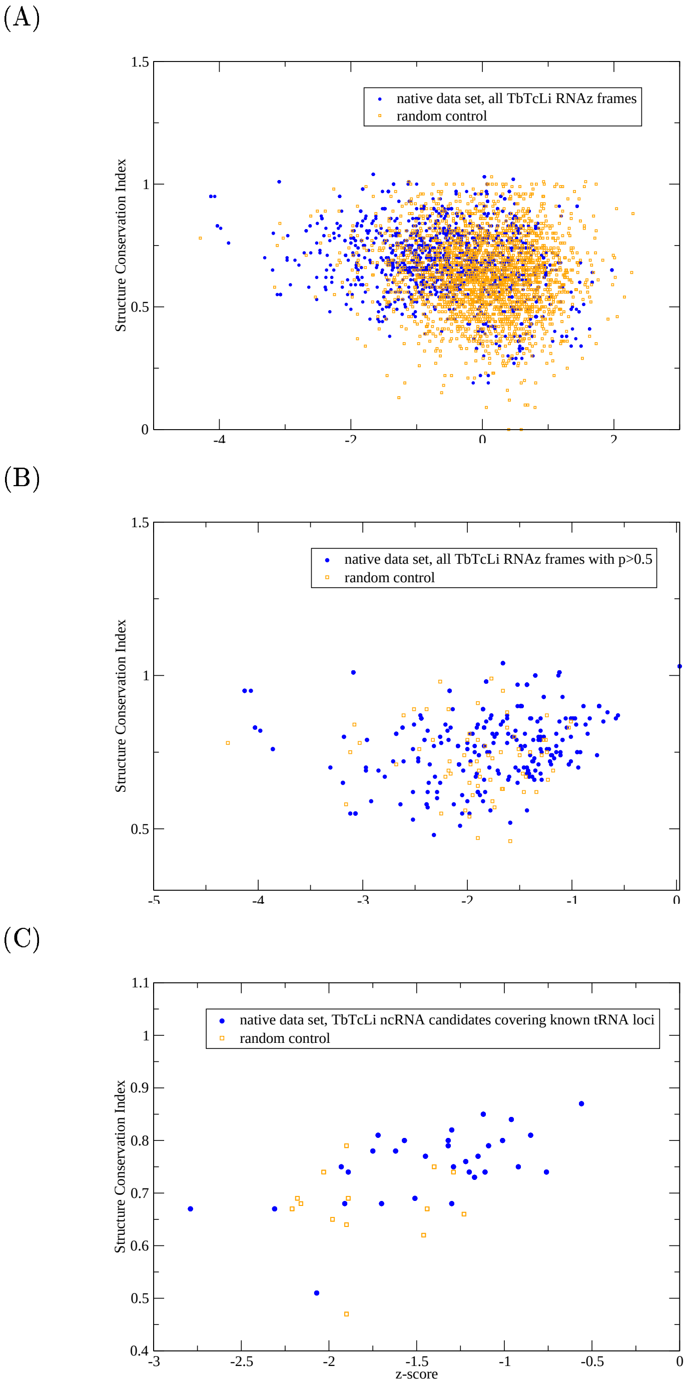


Figure 5.4: Comparison of SVM classification results of the TbTcLi screen. The z-score of a valid ncRNA should be as negative (-3) and the SCI as positive (+1) as possible. Thus it is expected to obtain two separated clusters, the native data set in the upper left and the randomised set in the lower right corner of the plot.

(A) One may divine a separation of the native set comprising all scored `RNAz` frames and the corresponding random control.

(B) A separation of `RNAz` frames with $p > 0.5$ is not explicitly observable considering z-score and SCI, but we know that SVM classification yieldes to 342 `RNAz` windows for the native set in contrast to 64 frames for the shuffled set at the $p > 0.5$ level.

(C) However, a comparison of z-score and SCI for ncRNA candidates covering known tRNA loci reveals a separation. We count 31 tRNA predictions and only 14 in the shuffled set. On the one hand the z-score is pretty high for the random set, but on the other hand we observe that the native predictions tend to have a superiorly conserved structure.

5.3 Conclusion

We showed a promising way of ncRNA prediction by comparative genome analyses. Starting with blank nucleotide sequences we implemented a methodology for the identification of structured ncRNAs. We demonstrated how to receive novel unknown ncRNA loci from current genomes. We assume that `RNAz` based ncRNA prediction works very well for trypanosomatids. We recommend to use database systems for efficient handling of the huge amount of data retrieved by large-scale genomewide analyses. As a consequence, most of the processing steps are automated. Considerable effort has to be taken into account for the validation of those predictions. For our purposes it is not necessary to reveal every biological function of predicted loci in detail but we are able to enlighten the general function of a major fraction of predicted ncRNA genes via homology based approaches. Furthermore, we assume to have helped in gaining more experience with `RNAz` and ncRNA prediction at all. Hopefully we motivated future developments of ncRNA prediction approaches.

Computational ncRNA prediction remains a challenging field of bioinformatic research - and the very first steps towards reliable predictions already have been taken.

5.4 Supplement

We set up summarising webpages of the predicted ncRNAs for every three-way alignment screen. All candidate sets and their current annotation are listed there. Furthermore the predictions can be downloaded in `fasta` and `EMBL` format (viewable with `Artemis`). We added a unofficial self-defined "ncRNA" feature key to the `EMBL` files to address the ncRNAs.

- Trypanosoma (TbTcTv, TbTcLi, TbTcLm):
  `http://www.bioinf.uni-leipzig.de/~dominic/projects/tryps/`
- Leishmania (LiLmTb, LmLiTb):
  `http://www.bioinf.uni-leipzig.de/~dominic/projects/leish/`

List of Tables

List of Figures

Bibliography

[1] S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2(12):919–929, Dec 2001.

[2] Gisela Storz. An expanding universe of noncoding RNAs. *Science*, 296(5571):1260–1263, May 2002.

[3] John S Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10):930–939, Oct 2003.

[4] et al. J. C. Venter. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[5] et. al E. S. Lander. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[6] A. Huttenhofer, M. Kiefmann, S. Meier-Ewert, J. O'Brien, H. Lehrach, J. P. Bachellerie, and J. Brosius. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J*, 20(11):2943–2953, Jun 2001.

[7] A. D. Omer, T. M. Lowe, A. G. Russell, H. Ebhardt, S. R. Eddy, and P. P. Dennis. Homologs of small nucleolar RNAs in Archaea. *Science*, 288(5465):517–522, Apr 2000.

[8] Z. Kiss-Laszlo, Y. Henry, J. P. Bachellerie, M. Caizergues-Ferrer, and T. Kiss. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, 85(7):1077–1088, Jun 1996.

[9] K. M. Wassarman, F. Repoila, C. Rosenow, G. Storz, and S. Gottesman. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*, 15(13):1637–1651, Jul 2001.

[10] L. Argaman, R. Hershberg, J. Vogel, G. Bejerano, E. G. Wagner, H. Margalit, and S. Altuvia. Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol*, 11(12):941–950, Jun 2001.

[11] E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol*, 11(17):1369–1373, Sep 2001.

[12] Ilka M Axmann, Philip Kensche, Jorg Vogel, Stefan Kohl, Hanspeter Herzel, and Wolfgang R Hess. Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol*, 6(9):R73, 2005.

[13] John P McCutcheon and Sean R Eddy. Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. *Nucleic Acids Res*, 31(14):4119–4128, Jul 2003.

[14] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.

[15] A. Wagner and P. F. Stadler. Viral RNA and evolved mutational robustness. *J Exp Zool*, 285(2):119–127, Aug 1999.

[16] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A*, 96(17):9716–9720, Aug 1999.

[17] Eric Bonnet, Jan Wuyts, and Pierre Rouzand Yves Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–2917, Nov 2004.

[18] Stefan Washietl and Ivo L Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*, 342(1):19–30, Sep 2004.

[19] Peter Clote, Fabrizio Ferr, Evangelos Kranakis, and Danny Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591, May 2005.

[20] Stefan Washietl, Ivo L Hofacker, and Peter F Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, Feb 2005.

[21] Kristin Missal, Dominic Rose, and Peter F Stadler. Non-coding RNAs in Ciona intestinalis. *Bioinformatics*, 21 Suppl 2:ii77–ii78, Sep 2005.

[22] Kristin Missal, Xiaopeng Zhu, Dominic Rose, Wei Deng, Geir Skogerb, Runsheng Chen, and Peter F Stadler. Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. *J Exp Zoolog B Mol Dev Evol*, Jan 2006.

[23] Stefan Washietl, Ivo L Hofacker, Melanie Lukasser, Alexander Huttenhofer, and Peter F Stadler. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol*, 23(11):1383–1390, Nov 2005.

[24] Changning Liu, Baoyan Bai, Geir Skogerbo, Lun Cai, Wei Deng, Yong Zhang, Dongbo Bu, Yi Zhao, and Runsheng Chen. NONCODE: An integrated knowledge database of non-coding RNAs. *Nucleic Acids Res*, 33(Database issue):D112–D115, Jan 2005. noncode.

[25] Bioinformatics Research group, cas - The Noncode Database.
`http://noncode.bioinfo.org.cn/index.htm`
Latest visit: 03/01/2006.

[26] The Wellcome Trust Sanger Institute: The Trypanosoma brucei Genome Project.
`http://www.sanger.ac.uk/projects/t_brucei/`
Latest visit: 06/11/2005.

[27] WHO/TDR, Wellcome Trust and Genedb 'Tritryp' sequencing consortium et al.: Trypanosomatids: Genomes and Biology. Two disc CD-rom set.

[28] T. j. Naucke: Leishmanien - Die Parasiten.
`http://leishmaniose.de/leishmania.html`
Latest visit: 01/11/2005.

[29] Graphic Images of Parasites: Leishmania spp. (leishmaniasis).
`http://www.biosci.ohio-state.edu/ parasite/leishmania.html`
Latest visit: 28/12/2005.

[30] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[31] Patric R. J. Ostergaard. A fast algorithm for the maximum clique problem. *Discrete Appl. Math.*, 120(1-3):197–207, 2002.

[32] Sonja J Prohaska, Claudia Fried, Christoph Flamm, Gnter P Wagner, and Peter F Stadler. Surveying phylogenetic footprints in large gene clusters: applications to Hox cluster duplications. *Mol Phylogenet Evol*, 31(2):581–604, May 2004.

[33] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.

[34] J. R. Stevens and W. Gibson. The molecular evolution of trypanosomes. *Parasitol Today*, 15(11):432–437, Nov 1999.

[35] Austin Hughes and Helen Piontkivska. Molecular phylogenetics of Trypanosomatidae: contrasting results from 18S rRNA and protein phylogenies. *Kinetoplastid Biol Dis*, 2(1):15, Oct 2003.

[36] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R Eddy, and Alex Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue):D121–D124, Jan 2005.

[37] Sam Griffiths-Jones, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–D144, Jan 2006.

[38] Laurent Lestrade and Michel J Weber. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, 34(Database issue):D158–D162, Jan 2006.

[39] T. M. Lowe and S. R. Eddy. `tRNAscan-SE`: A program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, 25:955–964, 1997.

[40] Peter F. Stadler Jana Hertel. Hairpins in a haystack: Recognizing microrna precursors in comparative genomics data. *submitted*, 2006.

[41] T J et al. Golembe. Specific sequence features, recognized by the smn complex, identify snrnas and determine their fate as snrnps. *Mol. Cell. Biol*, 25:10989–11004, 2005.

[42] M Bell and A Bindereif. Cloning and mutational analysis of the leptomonas seymouri u5 snrna gene: function of the sm site in core rnp formation and nuclear localization. *Nucleic Acids Res.*, 27:3986–3994, 1999.

[43] M. N. Schnare and M. W. Gray. Structural conservation and variation among U5 small nuclear RNAs from trypanosomatid protozoa. *Biochim Biophys Acta*, 1490(3):362–366, Feb 2000.

[44] Zsofia Palfi, Bernd Schimanski, Arthur Günzl, Stephan Lücke, and Albrecht Bindereif. U1 small nuclear RNP from Trypanosoma brucei: A minimal U1 snRNA with unusual protein components. *Nucleic Acids Res*, 33(8):2493–2503, 2005.

[45] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of rna secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[46] Ivo L Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, Jul 2003.

[47] Timothy Ravasi, Harukazu Suzuki, Ken C Pang, Shintaro Katayama, Masaaki Furuno, Rie Okunishi, Shiro Fukuda, Kelin Ru, Martin C Frith, M. Milena Gongora, Sean M Grimmond, David A Hume, Yoshihide Hayashizaki, and John S Mattick. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res*, 16(1):11–19, Jan 2006.

[48] Alexander Huttenhofer and Jorg Vogel. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res*, 34(2):635–646, 2006.

[49] Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F A Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, David Haussler, and Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–715, Apr 2004.

[50] Michael Brudno, Chuong B Do, Gregory M Cooper, Michael F Kim, Eugene Davydov, Eric D Green, Arend Sidow, Serafim Batzoglou, and N. I. S. C. Comparative Sequencing Program. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4):721–731, Apr 2003.

[51] Jeroen Raes, Klaas Vandepoele, Cedric Simillion, Yvan Saeys, and Yves Van de Peer. Investigating ancient duplication events in the Arabidopsis genome. *J Struct Funct Genomics*, 3(1-4):117–129, 2003.

[52] F. Barneche, C. Gaspin, R. Guyot, and M. Echeverria. Identification of 66 box C/D snoRNAs in Arabidopsis thaliana: extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2'-O-methylation sites. *J Mol Biol*, 311(1):57–73, Aug 2001.

[53] Christopher Maher, Lincoln Stein, and Doreen Ware. Evolution of Arabidopsis microRNA families through duplication events. *Genome Res*, Mar 2006.

Affirmation

Hereby I explain to have written this work independently and only to have used the sources and aids stated in the bibliography.

Place and date DD/MM/YYYY

Signature

Erklärung

Hiermit erkläre ich diese Arbeit selbstständig angefertigt und nur die im Literaturverzeichnis angeführten Quellen und Hilfsmittel benutzt zu haben.

Ort und Datum

Unterschrift