Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif

Lesley J. Collins[1*], Thomas J. Macke[2] and David Penny[1]

[1] Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Private Bag 11222, Palmerston North, New Zealand.

[2] Department of Molecular Biology,
The Scripps Research Institute, La Jolla, CA 92037, USA

Summary

Non-coding RNAs (ncRNAs) contain both characteristic secondary-structure and short sequence motifs. However, "complex" ncRNAs (RNA bound to proteins in ribonucleoprotein complexes) can be hard to identify in genomic sequence data. Programs able to search for ncRNAs were previously limited to ncRNA molecules that either align very well or have highly conserved secondary-structure. The RNAmotif program uses additional information to find ncRNA gene candidates through the design of an appropriate "descriptor" to model sequence motifs, secondary-structure and protein/RNA binding information. This enables searches of those ncRNAs that contain variable secondary-structure and limited sequence motif information. Applying the biologically-based concept of "positive and negative controls" to the RNAmotif search technique, we can now go beyond the testing phase to successfully search real genomes, complete with their background noise and related molecules. Descriptors are designed for two "complex" ncRNAs, the U5snRNA (from the spliceosome) and RNaseP RNA, which successfully uncover these sequences from some eukaryotic genomes. We include explanations about the construction of the input "descriptors" from known biological information, to allow searches for other ncRNAs. RNAmotif maximizes the input of biological knowledge into a search for an ncRNA gene and now allows the investigation of some of the hardest-to-find, yet important, genes in some very interesting eukaryotic organisms.

1 Introduction

Non-coding RNAs (ncRNAs) make transcripts that function as RNA, rather than encoding proteins, the best-known examples being ribosomal-RNA (rRNA) and transfer-RNA (tRNA) [1]. Many ncRNAs form part of RNA-protein complexes (Ribonucleoproteins, RNPs) and play roles in cellular processes such as RNA processing and splicing. Some ncRNAs have catalytic functions e.g. RNaseP RNA, whereas others serve key structural roles in ribonucleoprotein complexes e.g. snRNAs [2]. Searching databases for homologues based on sequence similarity is only useful for the larger, more slowly evolving ncRNAs (such as ribosomal RNAs) and is less reliable for other ncRNAs. Sequence similarity methods may fail to find ncRNA gene candidates when there is a large evolutionary distance between the query species and the target genome being searched [3].

In the past, programs such as RNAMOT [4] and PatScan [5] were developed to define and search for RNA structures and these led to programs such as tRNAscan-SE [6], which were designed to look for specific kinds of structural RNA. Recent ncRNA search techniques (e.g.

* Corresponding author: L.J.Collins@massey.ac.nz

ERPIN [7] and RSEARCH [8]) take both sequence and structure into account but are unable to model small sequence and secondary-structure motifs that correspond to protein or RNA binding sites in the ncRNA. These programs rely on sequence and secondary-structure alignment, either between multiple ncRNA sequences (ERPIN) or between query and subject sequences (RSEARCH). Alignment is difficult for such ncRNAs as there is often little sequence homology between distantly related species. Although these RNAs have both a conserved secondary-structure and some highly conserved sequence motifs, they also contain some secondary-structure differences [9, 10].

The RNAmotif program [11] was developed from RNAMOT [4] and uses an expanded syntax for describing motifs along with an implementation of nearest-neighbor rules and other schemes for ranking hits. RNAmotif has previously been used to find two groups of ncRNAs; tRNA [12] and the Iron Response Element (IRE) [11, 13], both of which contain highly conserved secondary-structures. This program uses a user-defined "descriptor" as input, modelling allowable secondary-structure and sequence motifs. It also has a scoring section that assesses the different features of the match [11]. A criticism of RNAmotif software is the lack of any value of statistical significant attached to any returned sequences. This value can be easily calculated based on sequence and/or secondary-structure similarity but is difficult to compute based on a biologically-derived model. To overcome this hurdle, and until more sophisticated RNA-model comparison techniques become available, we introduce "positive and negative controls", a fundamental concept of molecular biology, to provide significance to the RNAmotif results. First a test database is constructed consisting of positive controls (sequences we expect to be returned with a descriptor) and negative controls (sequences we do not expect to be returned). A second testing phase tests the performance of a descriptor against genomic background noise and a third testing phase was to search a genome for its known ncRNA sequence, testing a descriptor against similar ncRNAs found in that genome.

This study also shows how the use of a user-defined scoring section, results filtering and parallel implementation reduce the problems associated with searches of both crown (animal, yeast and plants) and basal (protist) eukaryotic genomes. This resulted in the identification of candidates for both the U5snRNA and the RNaseP RNA, from *Giardia lamblia, Entamoeba histolytica* (*Ent. histolytica*) and the microsporidian, *Encephalitozoon cuniculi (*abbreviated here as *Ecz. cuniculi* to avoid confusion with *Ent. histolytica*), and the U5snRNAs from *Dictyostelium discoideum* and *Ciona intestinalis.*

The U5 snRNA molecule is part of the U5 snRNP ribonucleoprotein complex that is involved in the splicing of nuclear pre-messenger RNA [14]. U5snRNA has already been identified from a number of completely sequenced genomes including *Ecz. cuniculi* and *Plasmodium falciparum* making them ideal test subjects for this study. After the testing stage, the U5 descriptors were used to search other small eukaryotic genomes such as *G. lamblia*[15], *Dictyostelium discoideum,* [16], *Entamoeba histolytica* [17] and *Ciona intestinalis* [18].

The other ncRNA investigated here is Ribonuclease P (RNaseP) RNA, part of the ribonucleoprotein complex that cleaves 5'-leader sequences from precursor-tRNA to leave a mature tRNA molecule [19]. Apart from some short nucleotide motif sequences, eukaryotic RNaseP RNAs have little nucleotide sequence homology (except between closely related species) making this gene difficult to find in more distant species. RNaseP RNA contains features that make it more challenging to write an effective descriptor. We used an improved version of RNAmotif implementing parallel processing to search for the RNaseP RNA in the genomes mentioned above. A common criticism of software descriptions is often there is not enough detail on parameter-tuning to enable a researcher in the biological field to effectively use the program [20]. To this end, we provide a comprehensive explanation of the construction of the U5snRNA and RNaseP descriptors from known biological information

(i.e. RNA and protein binding sites), to enable researchers in the ncRNA field to design descriptors for their molecules of choice.

2 Methods

RNAmotif [11] is written in ANSI C and available as source code via 'anonymous ftp' from (ftp.scripps.edu/pub/macke/rnamotif-version.tar.gz where "version" is the version number, currently 3.0.0). RNAmotif supports parallel searches via an MPI based driver, called mrnamotif, which is included in the RNAmotif distribution. Parallel processing was done on the Helix Cluster, a distributed-memory Beowulf cluster with 65 nodes (128 processors) running the Linux RedHat (version 7.3) operating system and communicating with the MPI protocol (http://helix.massey.ac.nz). All nodes used in testing and searching with RNAmotif had AMD Athlon MP-2100 processors running at 1733.335 MHz. A Perl script used to split large databases into smaller units suitable for parallel processing is available from the corresponding author upon request.

The program "Getbest" (available from the corresponding author upon request) was incorporated into the RNAmotif searching technique filtering the results from each worker node to give a condensed results file. Getbest works by selecting only the best solution found at each position of the sequence being searched, in this case, the position with the lowest free energy (ΔG). As expected thermodynamic stabilities improve with length [21], the sequence with the lowest free energy will tend to have the longest sequence which is retained using Getbest.

2.1 Sequences and Genomes

U5snRNA sequences were downloaded from the Rfam database [22] and the databases at NCBI (http://www.ncbi.nlm.nih.gov/). The genomes of *Encephalitozoon cuniculi* [23], (AL391737 and AL590442-AL590451), *Ciona intestinalis* [18] (AABS00000000) and *Pyrococcus abyssi* (AL096836) were also downloaded from NCBI. The *Plasmodium falciparum* genome was downloaded from PlasmoDB [24, 25] http://plamodb.org). *Dictyostelium discoideum* (soil-living amoeba) [16] preliminary sequence data was obtained from The Wellcome Trust Sanger Institute (http://www.sanger.ac.uk). The *Entamoeba histolytica* genome sequencing data [17] was produced by the Sanger Institute Pathogen Sequencing Unit at the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/E-histolytica).

Early releases of the *Giardia lamblia* genome (WB strain, clone C6) was kindly provided by the *Giardia lamblia* Genome Project [15] is based at the Marine Biological Laboratory at Woods Hole, Massachusetts, U.S.A. (http://jbpc.mbl.edu/Giardia-HTML/index2.html). This "Whole Genome Shotgun" sequencing project has now been completed and deposited at DDBJ/EMBL/GenBank under the project accession AACB01000000.

RNaseP RNA sequences were downloaded from the RNaseP Database ([26], http://www.mbio.ncsu.edu/RNaseP/main.html) and NCBI. The RNaseP eukaryotic consensus secondary-structure was taken from Frank et al. 2000 [27].

2.2 RNAmotif Descriptors

2.2.1 Descriptor Design – U5snRNA

A descriptor is read by the RNAmotif program from the 5' end of the model to the 3' end, so both 'sides' of a helix must be represented in the code. For example, *h5 (tag = 'helix1', len = 4)* opens a helix of 4 base-pairs, and *h3(tag = 'helix1')* closes the helix. Single-stranded regions are

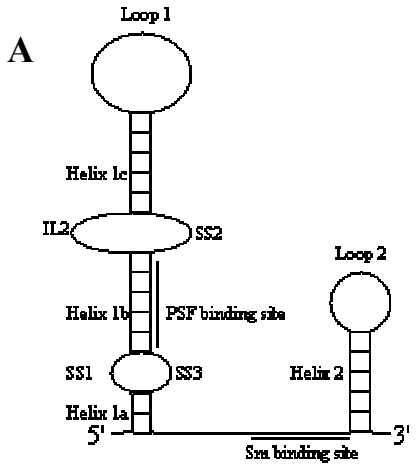
represented by *ss(tag = 'single_stranded 1')*. Both helices and single-stranded regions may contain length (*minlen* - minimum length; *maxlen* - maximum length), sequence, mismatch and mispairing parameters to allow for the small differences that are found in ncRNAs from different species. For all descriptors in this study, parameters were set to allow G:U pairing, and the folding of the structure to have a user-defined maximum energy level (emax).

The scoring section was designed to allow the user to see at a glance the different type of motifs that have been added together to produce the final score. The absence of a motif recorded a '*' in the motif position in a string of motif characters. The presence of a motif changed this '*' into a letter designating the selected motif. This motif scoring visualization is useful when looking for an ncRNA that contains some, but not other, elements, yet can still be a legitimate candidate.

Three descriptors were constructed for the U5snRNA based on features found in different combinations of species. The U5snRNA consensus secondary-structure contains a Sm protein-binding site, a highly conserved loop of eleven nucleotides next to a helix of 6-8 base-pairs [9, 14] (Figure1A). Features that are not present in U5snRNA sequences from some species include a second helix-loop structure and a PSF/p54^{nrb} protein-binding site. Figure 1 shows the U5_A descriptor and lists the differences between that descriptor and the other two U5snRNA descriptors used in this study, U5_B and U5_C. Secondary structure regions either absent or extremely variable between species (e.g. Helix 1a) were not included in the descriptors or converted to single-stranded regions. Mispairing events (i.e. *mispair = 1*) were permitted in some of the helices to improve the range of sequences recovered during testing, however including these events increased processing time. Highly variable single-stranded regions such as "IL2" were given a wide length range (in this case, between 3 and 18 nucleotides to allow for an extra helix that is present in some yeast species).

Helix1c, Loop1 and the Sm-binding site are important biological features of the U5snRNA [9]. Helix1c in some species has an internal mispairing event (a G:A pairing) which was modelled differently in descriptors U5_A and U5_B. U5_A allowed a mispairing on either end of the helix as well as internally (mispair = 1, ends = 'mm') whereas U5_B used stricter settings with mispairing only permitted on the distal (farthermost from the loop) end (mispair = 1, ends = 'mp'). Loop1 consists of eleven nucleotides containing a highly conserved sequence motif [28]. Loop1 was modelled differently in the three descriptors as shown in Figure 1B to allow for differences from the consensus model shown in the few basal eukaryotic U5snRNAs available (e.g. *P. falciparum* and *L. collosoma*). U5_A allowed for proximal mispairing whereas the U5_B descriptor was again stricter. The sequence within loop1 was scored the same between U5_A and U5_B but an alternative "less-strict" scoring scheme was used in U5_C. The Sm-protein binding site provides an example of how a sequence can be used for selection (in the "descr" section) then have viable sequence alternatives scored against in the scoring section. The Sm-binding sequence was also anchored to the end of the single-stranded region (using $). This greatly improved processing speed (non-anchored sequence, "ayuuuung" = 4 minutes, anchored sequence "ayuuuung$" = 33 seconds) and lowered the number of redundant hits and the output file-size.

Optional sequence motifs (e.g. The PSF-binding site on the 3' side of helix 1b) can be scored against in the scoring section but not included in the "descr" section as this would make this motif inclusion compulsory.



B

```
# U5snRNA Descriptor (U5_A)
# Maximum score is 6.
parms
    wc += gu;
    emax = -5;
descr
  h5(tag='1b', minlen=3, maxlen=10,mispair = 1)
  ss(tag='IL2',minlen=3, maxlen=18)
  h5(tag='1c',minlen=6, maxlen=9, mispair=1, ends='mm')
1 ss(tag='loop1', minlen=7,maxlen=12)
3 2 h3(tag='1c')
  ss(tag='ss2',minlen=0, maxlen=9)
  h3(tag='1b')
  ss(tag='Sm',minlen=15, maxlen= 34,mismatch = 1,seq = "ayuuuung$")
  h5(tag=2',minlen = 0,maxlen=9)
  ss(tag='loop2', minlen=3, maxlen=7)
  h3(tag='2')
score
 {
 s1 = 0.0;
 x = '*'; y = '*'; t = '*'; z = '*'; q = '*';
 if(length(h5(tag='1c')) == 8)  {s1 += 1.0; x = 'A';}
 if(length(ss(tag='loop1')) == 11)  {s1 += 1.0; y = 'B';}
4 if(ss(tag='loop1') =~ "cgccuuuuacu") {s1 += 2.0; t = 'a';}
  else if(ss(tag='loop1') =~ "gccuuuuac") {s1 += 1.5; t = 'b';}
  else if(ss(tag='loop1') =~ "ccuuuuac") {s1 += 0.5; t = 'c';}
  else if(ss(tag='loop1') =~ "cuuuu") {s1 += 0.5; t = 'd';}
 if(h3(tag='1b') =~ "ggagaggaac") {s1 += 1.0; r = 'a';}
 else if(h3(tag='1b') =~ "gaggaa") {s1 += 1.0; r = 'b';}
 else if(h3(tag='1b') =~ "ggaga") {s1 += 0.5; r = 'c';}
 if(ss(tag='Sm') =~ "auuuuuug$") {s1 += 1.0; q = 'a';}
 else if(ss(tag='Sm') =~ "ayuuuung$") {s1 += 0.5; q = 'b';}
 else if(ss(tag='Sm') =~ "auuuuuug$") {s1 += 0.6; q = 'c';}
 else if(ss(tag='Sm') =~ "ayuuuuug$") {s1 += 0.5; q = 'd';}
 if(s1 < 2.5) REJECT;
 s2 = efn( h5[1], h3[NSE] );
 if( s2 > emax )
    #REJECT;
    s1 -=0.01;
 SCORE = sprintf( "%s,%s,%s,%s,%s,%s,%1.2f,%1.3f", x, y, t, z, q, s1,s2 );
}
```

Figure 1: U5 snRNA model and RNAmotif descriptor. **A**: Consensus model of the U5snRNA secondary-structure based on the model of Peng et al. 2002. Loop2 is not present in the U5snRNAs of some species. **B**: U5_A descriptor used to search for potential U5snRNA genes. Other U5snRNA descriptors were based on this descriptor but contained the following differences:

U5_B descriptor

Point 1, h5(tag='1c',minlen=6, maxlen=8, mispair=1, ends='mp');

Point 2, ss(tag='loop1', len = 11) ;

Point 3, deleting these two lines thus not scoring on the length of helix 1c and loop1.

U5_C descriptor

Point 4, _the sequences scored upon were as follows:

```
if(ss(tag='loop1') =~ "ygccuuuuacu") {s1 += 1.0;t = 'a';}
else if(ss(tag='loop1') =~ "ccuuuuac") {s1 += 0.7;t = 'b';}
else if(ss(tag='loop1') =~ "ccuuuu") {s1 += 0.5;t = 'c';}
```

The scoring for option d is the same as in U5_A.

In the scoring section for each descriptor the individual letters i.e. 'A', 'B', 'x', 'y', 't', 'r' and 'q' represent the option scored for each sub-element allowing the user to see in the output the makeup of the overall score.

2.2.2 Descriptor Design – RNase P

The eukaryotic RNaseP RNA has a generally conserved secondary-structure (Figure 2A) with parts of the structure highly conserved while other parts contain variability in both sequence and secondary-structure [10, 29]. Thus the RNaseP RNA descriptor had to take into account both conserved and variable features. It was found that descriptors designed for the full eukaryotic secondary-structure were computationally-prohibitive, requiring weeks to search the simplest databases and often failed to return accurate, if any, results (data not shown). Using a descriptor for part of the RNaseP RNA structure allowed sequences to be returned that were then analyzed further for essential downstream elements (e.g. CRV pseudoknot pairing). Descriptors were designed (Figure 2E) for the P3-CRI-P7-P10 section of the complete secondary-structure (Figure 2B) and designated P7_A, P7_B and P7_C.

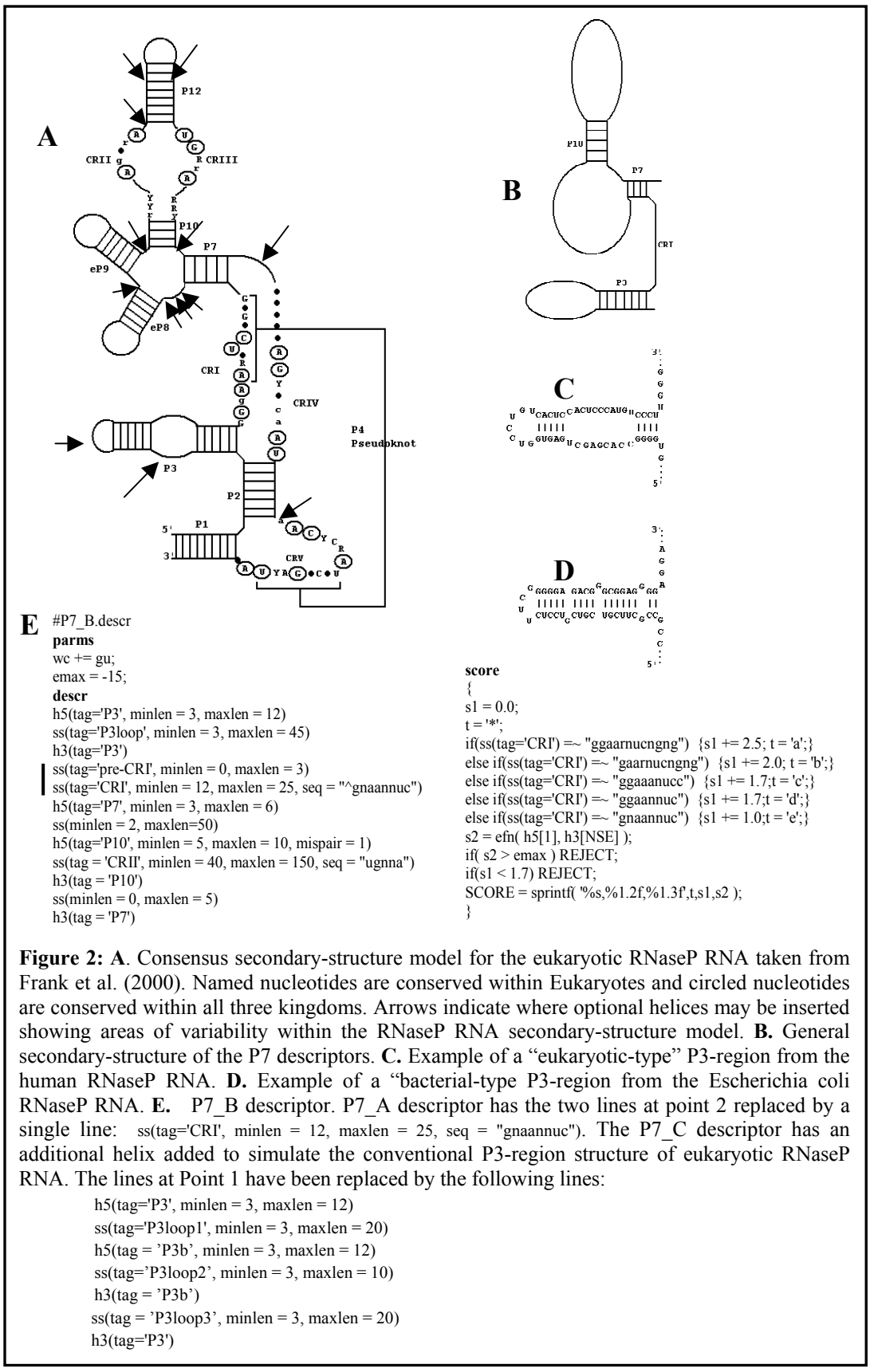


Figure 2: A. Consensus secondary-structure model for the eukaryotic RNaseP RNA taken from Frank et al. (2000). Named nucleotides are conserved within Eukaryotes and circled nucleotides are conserved within all three kingdoms. Arrows indicate where optional helices may be inserted showing areas of variability within the RNaseP RNA secondary-structure model. **B.** General secondary-structure of the P7 descriptors. **C.** Example of a "eukaryotic-type" P3-region from the human RNaseP RNA. **D.** Example of a "bacterial-type P3-region from the Escherichia coli RNaseP RNA. **E.** P7_B descriptor. P7_A descriptor has the two lines at point 2 replaced by a single line: ss(tag='CRI', minlen = 12, maxlen = 25, seq = "gnaannuc"). The P7_C descriptor has an additional helix added to simulate the conventional P3-region structure of eukaryotic RNaseP RNA. The lines at Point 1 have been replaced by the following lines:

```
h5(tag='P3', minlen = 3, maxlen = 12)
ss(tag='P3loop1', minlen = 3, maxlen = 20)
h5(tag = 'P3b', minlen = 3, maxlen = 12)
ss(tag='P3loop2', minlen = 3, maxlen = 10)
h3(tag = 'P3b')
ss(tag = 'P3loop3', minlen = 3, maxlen = 20)
h3(tag='P3')
```

The P3-region consists in archaea and bacteria of one helix-loop structure, but in eukaryotes has a large bulge in the middle of both the 5' and 3' strands of the helix forming two stacked helices (shown in Figures 2C and D) thus it was decided to allow for both types of structures in the RNaseP descriptors. Descriptors P7_A and P7_B code for a single helix-loop structure

but allow a large loop length to compensate for any second helix. Descriptor P7_C codes for a second helix with a minimum length of 3 base-pairs and single-stranded regions on either side. Adding this second helix both dramatically increased the processing time (from 3 minutes to 35mins) and lowered the specificity of the descriptor; however, most folding energies were improved with P7_C for each sequence region recovered. Designating the second helix optional (i.e. setting the minimum length of helix 1b to 0) recovered those sequences not detected with the P7_C descriptor, but increased the processing time tenfold. The lesson learnt here was to keep the number of helices to the minimum required for the desired species specificity.

The CRI region is the most highly conserved sequence motif in the RNaseP RNA and is critical for selection as an RNaseP RNA gene candidate. The CRI-minimal sequence motif (found in all RNaseP RNA sequences from all three kingdoms) was made mandatory for selection by including it in the CRI-motif parameter settings. This is the only sequence motif scored in the RNaseP descriptors as other CR-regions in the descriptor area (CR-II and CR-III) were too general to be useful. The CRI-minimal sequence motif was anchored in descriptor P7_B by adding a separate single-stranded region (*tag = 'preCRI'*) before this sub-element (*tag = 'CRI'*).

The P10 region contains allowances for the characteristic mispairing event that occurs in bacteria and some archaea but does not occur in sequences from crown eukaryotes (the situation in basal eukaryotes is unknown). The CRII-P12-CRIII single-stranded region (tag = 'CRII') is highly variable in length, and the number and position of helices between species and kingdoms [27]. This region was set to a single-stranded loop with a large range in length (between 40 and 150 nucleotides) to allow for this variability.

2.2.3 Descriptor Testing

Known U5snRNA and RNase P sequences were downloaded from the Rfam database [22] and from the NCBI databases. Test databases, TestDatabaseA (Table1) and Pdatabase (Table 2) were constructed containing sequences that were expected to be returned with the descriptors (positive controls), and sequences from other ncRNAs that were not expected to be returned (negative controls). Descriptors were tested against test databases to understand the performance of different variations of the descriptor.

Other ncRNAs contain some of the motifs found in our descriptors and thus representative sequences from all these ncRNAs were included in these test databases. It is unnecessary to construct overly large test databases (such as downloading the complete Rfam database) as long as appropriate positive and negative controls, usually determined with the biological knowledge of the ncRNA, have been included. If appropriate controls cannot be selected for a desired ncRNA then the Rfam database, although large, will be a reasonable alternative. Scoring cutoffs for each descriptor are selected after the analysis of positive and negative control results and determine the selectivity of the descriptor.

In order to test the level of background from other ncRNAs, three U5snRNA sequences, human (M23822) *Schizosaccharomyces pombe* (X15504) and *Caenorhabditis elegans* (Z69665) and the human RNaseP RNA (X15624) were randomly inserted into the *Pyrococcus abyssi* genome. To date there have been no U5snRNA sequences described for any archaeal species so it was expected that the inserted human, *C. elegans* and *S. pombe* sequences would be recovered with higher scores than any 'native' *P. abyssi* sequences. *P. abyssi* contains its own RNaseP but as the RNaseP descriptor was designed primarily for eukaryotic RNaseP, it was expected that the human RNaseP would be recovered with higher scores than the native *P. abyssi* RNaseP.

Table1	Species	U5	U1	U2	U4	U6	U11	U12	RNaseP
Plants	Arabidopsis thaliana	X13012	X53175	X06474	X67146	X52527			
	Rice	AC104179	AC025783	AF106845	AB026295	AC079128			
	Pea	X15934	X15926	X15936	X15931				
Fungi	Aspergillus nidulans	AC004395		AL683874		AY136823			
	Saccharomyces cerevisiae	M16510	M17411	M14625	U18778	Z73279			M27035
	Schizosaccharomyces pombe	X15310	X55773	X55772	X15491	M55650			X04013
	Encephalitozoon cuniculi	AL590450				AL590448			
Animal	Human	M77840	AC097369	X59360	X59361	AC114982	X13707	L43846	X15624
	Mouse	M10336	M14121	K00027	M10328	AC116657			
	Zebrafish	AL591593	AL929029	AL92108		AL929029			
	Frog	X06020	K02698	K02457					AF044330
	Caenorhabditis elegans	Z68215	Z81556	X51372	X51382	Z22178			
	Drosophila melanogaster	AC099022	X02136	X04241	K03095	M24605			
Basal	Entosiphon sulcatum	AF09539+		AF095839		AF095841			
	Leptomonas collosoma	AF006632		X56453	AF204671	X79014			
	Plasmodium falciparum	AE014823		AE014841	Z98547				
	Tetrahymena thermophila	X63789	X58845	X63786	X58844	X63790			
	Chlamydomonas reinhardtii	X67000	X70869	X71483		X71486			
	Trypanosoma brucei			X04678	M25777	X13017			

Table 1: Accession numbers of the sequences contained in the test database "TestDatabaseA" used in the evaluation of the U5snRNA descriptors. An empty cell indicates that this sequence was not available for inclusion in this database. Theoretically all U5snRNAs should be returned with the U5snRNA descriptors and thus be positive controls and all non-U5snRNAs be negative controls. For practical reasons only some sequences were selected to be specific positive and negative controls in the testing of the U5snRNA descriptors. Blue indicates a positive control and red indicates a negative control.

Table 2	Species	RNaseP	RNase MRP
Eukaryote	Saccharomyces cerevisiae	M27035	Z14231
	Schizosaccharomyces pombe	X04013	AL009197 (31216-31615)
	Homo sapiens	X15624	X51867
	Mus musculus	L08802	J03151
	Danio rerio	U50408	
	Xenopus laevis	AF044330	Z11844
	Drosophila melanogaster	AF434763	
	Arabidopsis thaliana		X65942 (34)
	Bos Taurus (Bovine)		Z25280
	Nicotiana tabacum (Tobacco)		K
	Rattus norvegicus (Rat)		J05014
Archaea	Pyrococcus abyssi	AJ248283	
	Sulfolobus acidocaldarius	L13597	
	Methanobacterium thermoautotrophicum	U42986	
	Methanococcus vannielii	AF192357	
	Archaeoglobus fulgidus	AE000782	
	Halobacterium cutirubrum	U42983	
	Aeropyrum pernix	AP000060	
Bacteria	Escherichia coli	V00338	
	Bacillus subtilis	M13175	
	Thermus aquaticus	Z15006	
	Streptomyces lividans	M64552	
	Agrobacterium tumefaciens	M59352	

Table 2: Sequences in the Pdatabase used for evaluation of the RNaseP descriptors. Included in this database are RNaseP RNA sequences from Eukaryotic, Archaeal and Bacterial species and RNase MRP sequences from Eukaryotic species. Note that RNase MRP has not been found in any Archaeal or bacterial species to date. An empty cell indicates that this sequence was not available for inclusion in this database. Theoretically all RNaseP RNAs should be returned with the RNaseP descriptors; however certain sequences are selected to be specific positive and negative controls. Blue indicates a positive control and red indicates a negative control. K -From Kiss and Filipowicz (1992)

Another testing stage is to run a descriptor against a genome in which the ncRNA has already been characterized. This could be done easily for the U5snRNA descriptors as the *Ecz. cuniculi* and *P. falciparum* U5snRNAs have already been identified and are available from the Rfam database (AL590450 and AE014823 respectively). This could not be done, however, for

the RNaseP descriptors as to date there have been no RNaseP RNAs characterized in any of the small eukaryotic genomes that were available.

All genomes used in this study were appended to contain positive controls (a file of U5snRNA and RNaseP sequences attached to the end of the genome file), so that if there were no sequences returned with a search, it could be determined that the program had run to completion successfully.

3 Results

3.1 U5snRNA

3.1.1 Descriptor Testing Results

RNAmotif searches against TestDatabaseA with each U5snRNA descriptor indicated their sensitivity (U5snRNA sequences from which species were returned) and their specificity (which ncRNAs other than the U5snRNA were returned). Results are shown in Table 3. All three descriptors returned all the designated positive controls (U5snRNAs from human, *Ecz. cuniculi*, *P. falciparum* and *Entosiphon sulcatum*). The looser U5_A and U5_C descriptors detected other ncRNAs with scores lower than 3.0, determining this number as the minimum score cutoff for subsequent genomic searches. The tighter U5_B descriptor did not return any other ncRNAs from TestDatabaseA and also failed to detect some of the known basal eukaryotic U5snRNAs.

Table 3	**U5_A**	**U5_B**	**U5_C**
Test Database Testing			
Processing Time	39.5 seconds	5.9 seconds	31.9 seconds
Output File Size	69 KB	33 KB (40 KB)	56KB
	Highest Scores	**Highest Scores**	**Highest Scores**
Human U5	4.49	1.50 (3.5)	3.99
Drosophila melanogaster U5	3.99	-	2.99
Caenorhabditis elegans U5	4.49	2.49 (4.49)	3.49
Arabidopsis thaliana U5	3.49	2.49 (3.49)	2.49
Oryza sativa U5	4.50	2.50 (4.50)	3.50
Schizosaccharomyces pombe U5	3.49	-	2.69
Tetrahymena themophila U5	4.00	-	3.00
Encephalitozoon cuniculi U5	4.00	2.00 (4.00)	3.50
Plasmodium falciparum U	3.99	2.00 (4.00)	3.00
Pysarum polycephalum U5	4.99	3.00 (5.00)	3.99
Entosiphon sulcatum U5	4.49	2.50 (4.50)	3.49
Human U11	2.99	-	2.99
Mouse U12	2.49	-	2.49
Caenorhabditis elegans U4	2.49	-	2.49
Genome Searches	**Processing Time**	**Processing Time**	**Processing Time**
Ecz. cuniculi genome	26 min 49 sec	5 min 35 sec	26 min 50 sec
"P. abyssi" genome	19 min 43 sec	3 min 12 sec	19 min 46 sec
G. lamblia genome	17 min 33sec	3 min 19 sec	17 min 41 sec
P. falciparum genome	738min 41sec	132min 14sec	746min 58sec
Ent. histolytica genome	640 min 0 sec	106min 39sec	Not run
D. discoideum genome	1014min 52sec	366min 30sec	Not run
C. intestinalis genome	495min 35sec	83min 14sec	Not run

Table 3: Evaluation results for the U5snRNA descriptors. Representative results from searches of TestDatabaseA are shown although other similar sequences that were in this database were also returned. All descriptors had "emax = 5". Scores below this threshold were not rejected during descriptor testing but had "0.01" subtracted from their overall score for indicative purposes. Descriptors U5_A and U5_C had score cutoffs set to 2.5. U5_B was run with two variations, the first containing scoring for the length of helix1c and loop1, and the second without this scoring. The scoring cutoff was set lower at 1.5 for latter run to compensate for the lessened maximum possible score. Timing differences between the two runs were the same so only one set of timing results are given. The *P. abyssi* genome has a number of U5snRNA sequences inserted for testing purposes. '=' indicates that this sequence was not detected with this descriptor. Species in blue were positive control for testing the U5snRNA descriptors.

RNAmotif searches of the 'doctored' *P. abyssi* genome with the U5 descriptors recovered all three inserted U5snRNA sequences above the cutoff score. The *S. pombe* U5snRNA was recovered with a lower score than some native sequences, indicating that with the parameters set in these descriptors, yeast-like U5snRNA sequences could not be expected to be recovered reliably above background noise.

In genomic testing of the U5snRNA descriptors, the known *Ecz. cuniculi* U5snRNA sequence (AL590450) was successfully recovered from its genome as the only top scoring hit with all three descriptors. The known *P. falciparum* U5snRNA (AE014823) was also easily recovered from the *P. falciparum* genome by all three U5 descriptors with the highest score. Recovery of their known U5snRNA sequences from the *Ecz. cuniculi* and *P. falciparum* genomes indicated that it was possible with the U5snRNA descriptors to distinguish between the U5snRNA and other closely related ncRNAs in their own genomes.

Other small eukaryotic genomes (*C. intestinalis*, *G. lamblia*, *Ent. histolytica* and *D. discoideum*) were then searched with the U5snRNA descriptors. A prior BLAST search of these genomes with all known U5snRNA sequences returned no significant results. With RNAmotif and the U5 descriptors, five candidate sequences were returned from the *C. intestinalis* genome, all of them contained the consensus loop1 sequence and could be folded into the consensus U5snRNA secondary-structure (Scaffold112:58432-58331;Scaffold71:40056-39955;Scaffold1849:5070-5172;Scaffold1028:15981-15885;Scaffold108:18565-18656). An alignment of these candidate sequences show that they are extremely similar to each other, with only a few nucleotide differences between them. Subsequent analysis showed that the *C. intestinalis* U5snRNA candidates showed similarity to other vertebrate U5snRNAs including human and mouse U5snRNAs. The proposed secondary structure for one of these sequences is shown in Figure 3E.

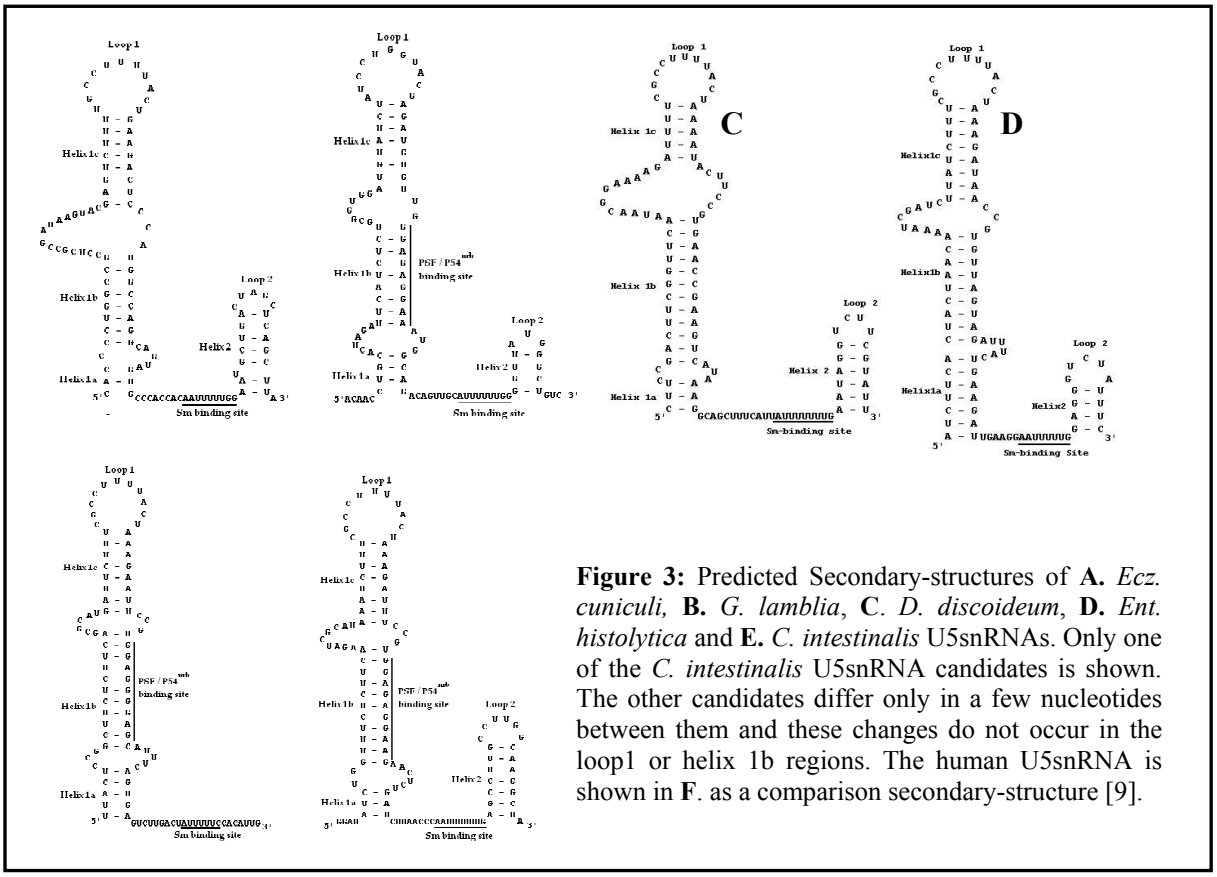


Figure 3: Predicted Secondary-structures of **A.** *Ecz. cuniculi,* **B.** *G. lamblia*, **C**. *D. discoideum*, **D.** *Ent. histolytica* and **E.** *C. intestinalis* U5snRNAs. Only one of the *C. intestinalis* U5snRNA candidates is shown. The other candidates differ only in a few nucleotides between them and these changes do not occur in the loop1 or helix 1b regions. The human U5snRNA is shown in **F**. as a comparison secondary-structure [9].

RNAmotif searches against the *G. lamblia* genome with the U5_A and U5_B descriptors (the U5_C descriptor failed to recover any clear candidate sequence) recovered a candidate sequence (AACB01000156: 17548-17456) that could be folded into the consensus U5snRNA secondary-structure. The sequence in loop1 does not conform entirely to the consensus loop1 motif (constructed from crown and basal eukaryotic U5snRNA loop1 sequences, [28]. The loop1 sequence in particular, affects splice site selection, particularly for introns with non-ideal 5' splice sites [30], and thus the differences in the loop1 sequences of the *G. lamblia* U5snRNA candidate may be a reflection of a difference in the splicing mechanism of this organism. The *G. lamblia* U5snRNA candidate has been shown to be expressed using RT-PCR (data not shown) and its sequence has been confirmed. The proposed secondary-structure of the *G. lamblia* U5snRNA is shown in Figure 3B.

An U5snRNA candidate was also recovered from the *D. discoideum* genome (Figure 3C). Helix1c is shorter than those found in other species, having only 5bp but allowing some non-canonical base-pairing i.e. G-A pairing could lengthen this helix. As the candidate *D. discoideum* U5snRNA sequence was recovered from only preliminary contig data, more work needs to be done to establish its viability when the genome has been more fully sequenced and assembled. However, a search of the Rfam database with the candidate *D. discoideum* U5snRNA sequence returned the U5snRNA alignment increasing the validity of this candidate.

Searches of the *Ent. histolytica* genome also produced a candidate U5snRNA sequence (Figure 3D). Again this candidate U5snRNA sequence was recovered from preliminary sequencing data and will require more investigation once the complete genome has been sequenced. As with the *D. discoideum* candidate, the *Ent. histolytica* U5snRNA returned the U5snRNA alignment with a search against the Rfam database.

3.2 RNaseP RNA

Testing against the Pdatabase (results shown in Table 4) showed that although the RNaseP descriptors covered only part of the total RNaseP RNA secondary-structure, they were still able to detect RNaseP RNA sequences from all three kingdoms. Descriptor P7_A showed the greatest ability not only to recover RNaseP RNA sequences from all three kingdoms, but to distinguish between them using CRI-motif scoring. RNaseMRP sequences were selected as negative controls during RNaseP descriptor testing because the RNaseMRP CRI-regions are similar to the RNaseP CRI-regions, with the expectation that the RNaseP descriptors should distinguish against the two different ncRNAs. RNaseMRP sequences were detected with the RNaseP descriptors at the lowest level (CRI-motif = ‘e’; the consensus CRI motif common to all three kingdoms), indicating that results returned with this motif may not be specific to RNaseP.

The RNaseP descriptors were also tested against TestDatabaseA to determine if other ncRNAs were also detected. Some U4 snRNA sequences were returned with the lowest scoring motif (CRI-motif = ‘e’) but no other snRNAs were detected above this level. Thus, future genomic searches used score cutoffs above this level. The P7_B returned the same sequences from both the Pdatabase and TestDatabaseA with less than half the processing time but did not distinguish clearly between the RNase MRP and the archaeal and bacterial RNaseP sequences. The P7_C descriptor took longer than the other two descriptors and did not return some of the archaeal and bacterial sequences, but gave much higher folding scores for the eukaryotic sequences.

Table 4	**P7_A**	**P7_B**	**P7_C**
Pdatabase Testing			
Processing Time	7 min 54 sec	3 min 12 sec	35 min 28 sec
Output File Size	47 KB	36 KB	39KB
	CRI motif	**CRI motif**	**CRI motif**
Human RNaseP	a (-15.03)	a (-13.81)	a (-20.43)
S. cerevisiae RNaseP	b (-15.05)	a (-15.05)	a (-19.75)
S. pombe RNaseP	e (-14.16)	e (-15.05)	-
P. abyssi RNaseP	d (-22.43)	d (-24.81)	c (-27.41)
S. acidocaldarius RNaseP	d (-13.52)	c (-16.25)	-
M. thermoautotrophicum RNaseP	d (-17.15)	c (-24.08)	d (-20.38)
A. fulgidus RNaseP	d (-26.70)	-	-
H. cutribricum RNaseP	c (-25.53)	c (-29.33)	-
E. coli RNaseP	b (21.46)	a (-21.46)	a (-20.46)
B. subtilis RNaseP	e (-15.27)	c (-15.27)	-
A. tumefaciens RNaseP	c (24.26)	a (-26.89)	a (-25.87)
S. lividans RNaseP	c (-31.73)	d (-25.63)	c (-27.93)
A. thaliana RNase MRP	e (-19.95)	c (-19.95)	c (-26.15)
N. tabacum RNase MRP	e (-18.92)	c (-18.92)	c (-22.96)
Human RNase MRP	e (-14.74)	c (-13.43)	c (-21.74)
S. pombe RNase MRP	e (-12.10)	-	c (-16.51)
S. cerevisiae RNase MRP	-	-	-
Other snRNAs that occur with descriptor:	L.collosoma_U4 (e) T. brucei_U4 (e)	L.collosoma_U4 (e)	L.collosoma_U4 (d) T. brucei_U4 (d)
Processing Time			
TestDatabaseA (Used for U5 testing)	2 min 54 sec	1 m 18 sec	8 min 23 sec
Ecz. cuniculi genome	866 min 25 sec	290 min 26 sec	1619 min 9 sec
P. falciparum genome	Not run	617min 27sec	Not run
G. lamblia genome	508min 52 sec	174 min 48 sec	Not run
Ent. histolytica genome	Not run	1627 min 37sec	Not run
D. discoideum genome	Not run	1093 min 26 sec	Not run

Table 4: Evaluation results from the RNaseP descriptors. Indicative results are shown. For processing times all descriptors had the following settings "emax = 15" with scores below this threshold being rejected and score cutoffs at 1.7. Detection of sequences with an "e" motif was achieved by setting the emax to -12 and the score cutoff to 1.0. The best folding energy scores are shown in brackets next to the CRI motif. Searches of the *P. falciparum*, *G. lamblia*, and *D, discoideum* genomes used parallel processing with emax at -20 and the score cutoff at 2.0. Searches of the *G. lamblia* genome were run with 8 nodes. Searches of the *P. falciparum*, *D. discoideum* and *Ent. histolytica* genomes were run with 16 nodes. '-' indicates that this sequence was not detected with this descriptor. CRI motif a = "ggaarnucngng" (Eukaryotic consensus CRI with first "g"); CRI motif b = "gaarnucngng" (Eukaryotic consensus CRI without first "g"); CRI motif c = "ggaaanucc" (Bacterial consensus CRI); CRI motif d = "ggaannuc" (Archaeal consensus CRI); CRI motif e = "gnaannuc" (Universal consensus CRI)

To date there have been no RNaseP RNA sequences described for *G. lamblia, Ecz. cuniculi, Ent. histolytica, D. discoideum, P. falciparum* and *C. intestinalis*. BLAST searches of these genomes with all known RNaseP RNA sequences failed to find any significant sequences. Candidate sequences were required to contain the CRIV, CRII and CRIII consensus regions in expected places and the CRI-region had to contain conserved nucleotides present in RNaseP RNAs from all three kingdoms.

RNAmotif searches against the *Ecz. cuniculi* genome recovered an RNase P candidate (Figure 4C) (Chromosome VII – start position 87184) which also contains some sequence similarity to RNaseP RNA sequences from rat and mouse. Another candidate was recovered from the *Ecz. cuniculi* genome (Figure 4A) with a proposed secondary structure that fits the general eukaryotic consensus secondary-structure; except for the P3-region which is more bacterial-like (compared with the example structures shown in Figures 2D and E).

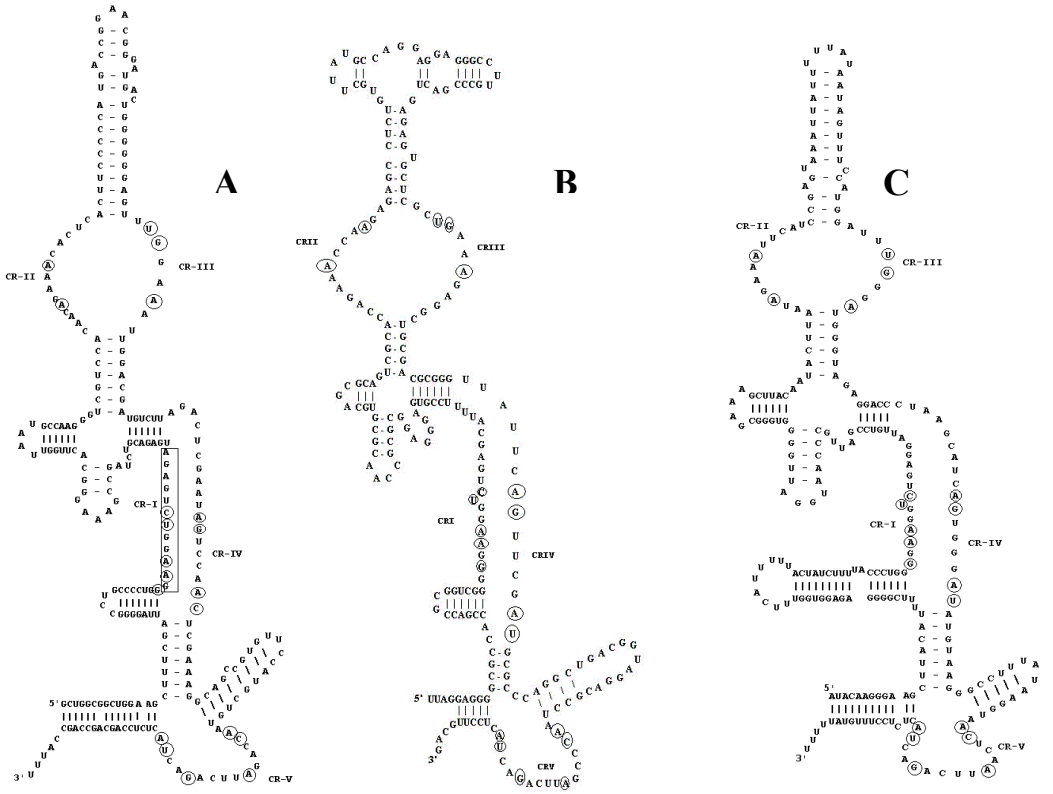


Figure 4: Putative RNaseP RNA secondary-structures of **A:** *Ecz. cuniculi*, **B:** *G. lamblia* and **C:** *Ent. histolytica* RNaseP RNA candidates. Circled nucleotides are those that are conserved between all three kingdoms (refer Figure 2) and must be present in these locations for the candidate to be considered a potential RNaseP RNA.

RNAmotif searches against *G. lamblia* using P7_A and P7_B descriptors (P7_C was not used due to its much longer processing time, even with parallel processing) returned a candidate RNase P sequence (Figure 4B)(AACB01000012: 65152-64918). As with the *Ecz. cuniculi* candidate, the P3-region was bacterial-like but the rest of the structure fitted the eukaryotic model. RT-PCR has shown that this sequence is expressed in *G. lamblia* and the sequence has been confirmed by sequencing.

RNAmotif searches against the *Ent. histolytica* genome also recovered a candidate RNaseP RNA sequence (Figure 4C) (*Ent1359g08.p1k:*355-440). The P3-region was longer than that found in the *Ecz. cuniculi* and *G. lamblia* RNaseP candidates, and although there was a two-nucleotide bulge in the 3' side of the helix, the *Ent. histolytica* P3-region still resembled the bacterial-like P3-region as opposed to the eukaryotic model. As this sequence was obtained from preliminary sequence data, further investigation will be required when the genome sequencing has been completed.

RNAmotif searches against *P. falciparum* failed to find any viable RNaseP RNA candidates, which may be due in part to its high A+T content. RNAmotif searches against the *D. discoideum* genome also failed to find any RNaseP RNA candidate. The genomic data for this genome is still preliminary and it is possible that the region containing an RNaseP RNA has not yet been sequenced. New releases of this genome will be screened in the future for possible RNaseP RNA candidates.

The RNaseP RNA candidates found in *Ecz. cuniculi*, *Ent. histolytica* and *G. lamblia* contain all the features that are expected in an RNaseP RNA, including nucleotides that have been shown to be conserved between RNaseP RNAs from all three kingdoms [27]. The Rfam

database was searched with the three RNaseP candidate sequences from *G. lamblia*, *Ent. histolytica* and *Ecz. cuniculi* but returned no hits. As a test, the RNaseP sequence from the fruitfly, *D. melanogaster* (the only eukaryotic RNaseP RNA not found in Rfam, AF434763), also failed to return any hits from Rfam. The eukaryotic RNaseP RNA is notoriously difficult to align which may account to some extent to the lack of any positive results from Rfam for any of our RNaseP candidates.

4 Discussion

Non-coding RNA genes are hard to find in genomic data. Previously, RNAmotif has been used to find ncRNAs with highly conserved secondary-structure, using very "tight" and efficient descriptors [11, 12]. This technique has now been taken to the next level, integrating sequence and secondary-structure sub-elements (representing protein and RNA binding sites) to search for ncRNA genes in eukaryotic genomic data.

A potential criticism of the RNAmotif software is that it cannot yet give a value of statistical significance on each returned sequence. Here positive and negative controls are used instead, because they can compensate at least partially. The concept of positive and negative controls is fundamental in molecular biology (as many of its techniques also lack statistical analysis). However, a measure of statistical significance would be desirable to compare whole ncRNA sequences. Sequence and secondary-structure alone may not be enough to produce a true representation of the features required by an ncRNA molecule in order to retain biological function. Algebraic dynamic programming techniques that can model the complete structure in a similar way to RNAmotif, are being developed [31]. Such methods at present are mathematically challenging but in future may result in a statistical evaluation method that will integrate well with RNAmotif and other ncRNA-associated software.

The design of appropriate descriptors is presently not a simple task. Testing has shown as the "descr" section of the descriptor becomes more complicated (i.e. more helices, sequences and parameters), processing time increases dramatically. Ideally the simplest possible descriptor should be designed to search for a particular ncRNA. However, this may not be appropriate when searching for ncRNAs for which little information is available. In these cases, 'atypical' elements found in some species but not in others may have to be included to facilitate sequence recovery. An example of this is the specific descriptor U5_A which was more accurate in finding candidates in "typical" eukaryotic organisms such as *C. intestinalis*, but did not work as well as the less-specific descriptor (U5_B) for searching basal eukaryotic genomes. By having descriptors with differing stringency and related in different ways, candidates could be recovered from distant genomes. At present, there is still a balance that must be achieved between a descriptor that can run in a realistic time-frame and one that allows enough variability for a search of a distantly-related genome.

Testing of the RNaseP descriptors showed that the complete ncRNA need not be modelled, and areas of high variability between species can be described as single-stranded regions, aiding both processing performance and species detection. Analysis of regions "downstream" of the descriptor region was done manually in this study, but could also be automatically incorporated into future RNAmotif releases.

The RNAmotif program is still a program under development. The largest genome size that we used here was that of the sea-squirt *C. intestinalis* (155Mbases). It is feasible to search larger genomes using parallel processing and descriptors, "fine-tuned" for performance efficiency. However, when searching into a little known genome such as those basal eukaryotes, many descriptor parameters cannot be completely optimized for performance without running the risk of not recovering the ncRNA of choice. Future RNAmotif releases

may include the incorporation of other biological information, such as whether a resulting sequence is within an open reading frame, or ways to compensate for AT-rich target genomes. However, RNAmotif now offers a realistic and biologically-orientated way to search for other non-coding RNAs within the increasingly wide range of sequenced genomes.

5 Acknowledgements

Many thanks to the administrators of the Helix parallel processing facility at Massey University, Albany, New Zealand for their help and advice. Many thanks to Mitchell L. Sogin, and Andrew G. McArthur and their teams at the *Giardia lamblia* Genome Project, (funded by the NIAID/NIH under cooperative agreement AI 043273), Marine Biological Laboratory at Woods Hole, for access to non-public data. Thanks also to Anu Idicula, Alicia Gore and Trish McLenachan for the RT-PCR and sequencing of some of the ncRNA gene candidates. This work was supported by the NZ Marsden Fund.

6 References

[1] S. R. Eddy. Non-coding RNA genes and the modern RNA world. Nat Rev Genet, 2(12): 919-29., 2001.

[2] G. Storz. An expanding universe of noncoding RNAs. Science, 296(5571): 1260-3, 2002.

[3] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2(1): 8, 2001.

[4] A. Laferriere, D. Gautheret and R. Cedergren. An RNA pattern matching program with enhanced performance and portability. Comput Appl Biosci, 10(2): 211-2, 1994.

[5] M. Dsouza, N. Larsen and R. Overbeek. Searching for patterns in genomic data. Trends Genet, 13(12): 497-8, 1997.

[6] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 25(5): 955-64, 1997.

[7] D. Gautheret and A. Lambert. Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles. J Mol Biol, 313(5): 1003-11., 2001.

[8] R. J. Klein and S. R. Eddy. RSEARCH: Finding homologs of single structured RNA sequences. BMC Bioinformatics, 4(1): 44, 2003.

[9] R. Peng, B. T. Dye, I. Perez, D. C. Barnard, A. B. Thompson and J. G. Patton. PSF and p54nrb bind a conserved stem in U5 snRNA. Rna, 8(10): 1334-47., 2002.

[10] L. J. Collins, V. Moulton and D. Penny. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. J Mol Evol, 51(3): 194-204., 2000.

[11] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res, 29(22): 4724-35., 2001.

[12] V. Tsui, T. Macke and D. A. Case. A novel method for finding tRNA genes. Rna, 9(5): 507-17, 2003.

[13] G. B. Fogel, V. W. Porto, D. G. Weekes, D. B. Fogel, R. H. Griffey, J. A. McNeil, E. Lesnik, D. J. Ecker and R. Sampath. Discovery of RNA structural elements using evolutionary computation. Nucleic Acids Res, 30(23): 5310-7., 2002.

[14] A. J. Newman. The role of U5 snRNP in pre-mRNA splicing. Embo J, 16(19): 5797-800., 1997.

[15] A. G. McArthur, et al. The Giardia genome project database. FEMS Microbiology Letters, 189(2): 271-273, 2000.

[16] L. Eichinger and A. A. Noegel. Crawling into a new era-the Dictyostelium genome project. Embo J, 22(9): 1941-6, 2003.

[17] B. J. Mann. Entamoeba histolytica Genome Project: an update. Trends Parasitol, 18(4): 147-8, 2002.

[18] P. Dehal, et al. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science, 298(5601): 2157-67, 2002.

[19] S. Xiao, F. Scott, C. A. Fierke and D. R. Engelke. Eukaryotic Ribonuclease P: A Plurality of Ribonucleoprotein Enzymes. Annu Rev Biochem, 71: 165-89, 2002.

[20] C. Mathe, M. F. Sagot, T. Schiex and P. Rouze. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res, 30(19): 4103-17, 2002.

[21] E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics, 16(7): 583-605., 2000.

[22] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna and S. R. Eddy. Rfam: an RNA family database. Nucleic Acids Res, 31(1): 439-41, 2003.

[23] M. D. Katinka, et al. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature, 414(6862): 450-3., 2001.

[24] M. J. Gardner, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature, 419(6906): 498-511, 2002.

[25] A. Bahl, et al. PlasmoDB: the Plasmodium genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). Nucleic Acids Res, 30(1): 87-90., 2002.

[26] J. W. Brown. The Ribonuclease P Database. Nucleic Acids Res, 27(1): 314, 1999.

[27] D. N. Frank, C. Adamidi, M. A. Ehringer, C. Pitulle and N. R. Pace. Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. RNA, 6(12): 1895-904., 2000.

[28] A. Szkukalek, E. Myslinski, A. Mougin, R. Luhrmann and C. Branlant. Phylogenetic conservation of modified nucleotides in the terminal loop 1 of the spliceosomal U5 snRNA. Biochimie, 77(1-2): 16-21., 1995.

[29] S. Xiao, F. Houser-Scott and D. R. Engelke. Eukaryotic ribonuclease P: increased complexity to cope with the nuclear pre-tRNA pathway. J Cell Physiol, 187(1): 11-20., 2001.

[30] T. S. McConnell and J. A. Steitz. Proximity of the invariant loop of U5 snRNA to the second intron residue during pre-mRNA splicing. Embo J, 20(13): 3577-86., 2001.

[31] C. Meyer and R. Giegerich. Matching and Significance Evaluation of Combined Sequence-Structure Motifs in RNA. Z. Phys. Chem., 216: 193-216, 2002.

[32] D. Penny and A. Poole. The nature of the last universal common ancestor. Curr Opin Genet Dev, 9(6): 672-7., 1999.