Universiteit Leiden
The Netherlands

Mining structural patterns in a non-coding RNA related to influenza virus infection and cancer development.

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)

Bioinformatics
Menco Cheung 1654829

Supervisor:
Dr. A. P. Goultieav
a.p.goultiaev@liacs.leidenuniv.nl

Second reader:
Dr. K.J. Wolstencroft
k.j.wolstencroft@liacs.leidenuniv.nl

Abstract

LncRNAs play important roles in many different biological processes and have been found in every branch of life. The particular mechanics of how each lncRNA achieve these functions differ per lncRNA and are not yet known for many lncRNAs. For some the higher order structure of the lncRNA causes the specific function while for others the act of transcription is of more importance. PSMB8-AS1 is a lncRNA that plays a role in influenza virus replication and cancer development. The intricacies of how it achieves these roles is not known. In this study we sought to determine if the structure is of importance of the function of PSMB8-AS1. Two approaches were taken. In the first, PSMB8-AS1 orthologues were used together with RNAz to find regions that showed a possibility conservation and predict their structures. These structures were then analysed using multiple alignment to see if conserved structures were present. A total of four regions were found that showed signs of conservation. These four where then analysed with other orthologues. We found that out of the four only one showed possible conservation. The possible conservation in this region was strongest in primates, but also present in other mammals. The other three regions showed no significant signs of conservation. In the second approach, the orthologues of three of the individual regions were found and analysed. One of these regions showed significant signs of conservations. Both approaches have shown that there is an evolutionary pressure for structural conservation in the second exon of PSMB8-AS1. Therefore, our study concludes that the higher order structure is of importance to the functions of the second exon of PSMB8-AS1.

Contents

Introduction

Long non-coding RNA

It has become well-known that most of the transcribed genes are not translated into proteins. Out of the 85% of the human genome that is transcribed, only 2% is described as protein-coding genes (Hangauer et al, 2013). The remainder of the transcriptomes are classified as non-coding RNA (ncRNA). NcRNAs are divided into two main groups, small ncRNA and long ncRNA (lncRNA). In this article we will mainly focus on long ncRNAs.

LncRNA are loosely defined by RNA that have a length of 200 nucleotides or longer and show no apparent protein-coding role (Quinn & Chang, 2016). LncRNAs have been found in every branch of life (Quinn & Chang, 2016), in which they play a pivotal role in many different biological processes such as transcription, splicing, translation, the cell cycle and apoptosis, protein localization, imprinting or stem cell pluripotency (Bryzghalov et al, 2019). LncRNAs can achieve these functions in different ways, including direct RNA:RNA interaction, miRNA sponge activity, nucleosome repositioning, histone modifications, DNA methylation or binding the transport factors to inhibit the nuclear localization of specific transcription factors (Bryzghalov et al, 2019). There have also been lncRNAs that have been linked to human diseases such as cancer development and growth, influenza and Alzheimer's disease (More et al, 2019: Novikova et al, 2012; Faghihi et al, 2008). Some cases have shown that the higher-order structure of lncRNA allow for its specific functionality, albeit several studies have shown that for some lncRNA the act of transcription seems to be of more importance than the transcript and its structure. In 2016 scientists conducted a study on the functional importance of secondary RNA structures. They genetically manipulated twelve genomic loci that produce lncRNA to find that five of the loci influenced the expression of neighbouring genes. They found that none of the five loci required the specific lncRNA transcripts and that instead the processes associated with the transcription were needed (Engreitz et al, 2016). Natural antisense transcripts (NATs) are RNAs that regulate the expression of their sense partners (Katayama et al, 2005; Lehner et al, 2002). Up to 70% of the human genes show evidence of antisense transcription. There are multiple ways the task of a NAT can be achieved. However, the prevailing mechanism is the recruitment of complex epigenetic machinery that mediates histone modifications, which leads to transcriptional deregulation of target genes (Kaikkonen et al, 2011). Little to no sequence specificity is required during this process; thus, no constraint upon sequence conservation is present (Bryzghalov et al, 2019).

lncRNA PSMB8-AS1

In this work, we studied conserved structures in the lncRNA PSMB8 antisense RNA 1 (PSMB8AS1). PSMB8-AS1 is a lncRNA that is found in humans and was found to play a role in the regulation of influenza virus replication. In an article published in RNA Biology (More et al, 2019), it was found that repression of PSMB8-AS1 using CRISPR interference reduced viral mRNA and protein levels as well as the release of progeny influenza virus particles (More et al, 2019). PSMB8-AS1 has also been shown to play a role in cancer development, with a significantly higher expression level in pancreatic ductal adenocarcinoma tumours than in normal samples (Guilletti et al, 2018). PSMB8-AS1 has four transcripts, the accession numbers for these transcripts are: NR_037173.1, NR_03714.1, NR_037175.1 and NR_037176.1. NR_037173.1 was chosen for this study, because it was the longest of the four transcripts. The longer transcripts give more potential regions of conservation in different organisms. NR_037173.1 has a total length of 1498 nucleotides(nt) and consists of three exons; exon 1: nucleotides 1 through 189, exon 2: nucleotides 190 through 415 and lastly exon 3: nucleotide 416 through 1498.

RNA Structure

The functionality of some lncRNAs depend on the higher-order structure, in particular the two-dimensional structure of lncRNAs. Similar to DNA, RNA is more stable when the nucleotides are paired with other nucleotides. However, RNA does not have a complementary strand like DNA. Therefore, the nucleotides of the RNA strands form pairs with nucleotides within its own strand giving the RNA a stable two-dimensional structure. The pairs made in RNA are very similar to
DNA, like in DNA the most stable pairs are cytosine (C) with guanine (G) and adenine (A) with uracil (U), instead of thymine (T) in DNA. RNA has, besides the two standard pairs, a third pair that does not follow the Watson and Crick base pair rules (Campbell, 2015). In RNA, guanine and uracil can form a stable base pair known as a wobble pair. The wobble pair is possible because two locations where hydrogen bonds are formed in guanine are complementary to two locations in uracil allowing for a stable bond (Kuchin, 2011).
To asses if the function and structure of PSMB8-AS1 are linked we searched for conserved secondary structures. Conserved structures are two-dimensional structures within an RNA molecule that show high similarity in orthologues of the RNA. The amount of conservation of a specific two-dimensional structure gives an argument for the relation between structure and function, with lncRNA being a highly heterogeneous class. Functional studies on lncRNAs are quite challenging with only a small fraction of lncRNAs being characterized (Bryzghalov et al, 2019). Analysis of conserved structures can partially mitigate this issue. First, by using the conservation of structures to indicate whether an RNA is functional or not. Second, knowledge of the level of lncRNA conservation and lncRNA orthologues helps to characterize lncRNAs and assign each to their hypothetical functional domains (Bryzghalov et al, 2019).

Structure Computation

To search for conserved structures within PSMB8-AS1 orthologues a structural analysis will be done. Many tools exist for the use of studying RNA structures, but in this study, we mostly used the tools provided by the Vienna package (Hofacker et al, 1994). Specifically, we will be using RNAz (Gruber et al, 2010). Another tool we will be using is the RNA mfold web server application (Zuker, 2003), both algorithms compute the minimum free energy (MFE) of a given RNA sequence to determine possible structures.
A problem with a free energy minimization of a single sequence is that the algorithm mainly uses the minimum free energy to compute the structure; however, this poses limitations for the accuracy of the structure because the structure with the lowest free energy is more often than not incorrect. Cell environments are very diverse and for each different environment different folding rules apply (Zhu et al, 2018). Furthermore, RNA binding proteins (RBPs) can also influence the two-dimensional structure of the RNA (Sasse et al, 2018). To minimize these problems, we searched for conserved structures within PSMB8-AS1 orthologues.

Research Plan

We searched for PSMB8-AS1 orthologues in eleven different organisms using BLAST. Out of the organisms five were primates and the remaining six were other mammals that were no primates, two of the other mammals we grouped in distant relatives 1 and the other four in distant relatives 2. Using RNAz we searched for conserved structures between different sets of organisms plus humans. Then we did a multiple alignment of the conserved structure location and assessed if this structure was possible in other organisms. A second analysis will be done, by using BLAST to search for each probable conserved structure individually and construct datasets of the possible conserved regions in other organisms which we will analyse using multiple alignment.

Materials & Methods

Dataset

For this research we started with the long non-coding RNA region PSMB8-AS1 PSMB8 antisense RNA 1. This RNA region has four transcripts with the accession numbers: NR_037173.1, NR_03714.1, NR_037175.1 and NR_037176.1. Out of the four we chose NR_037173.1. This decision was made because NR_037173.1 is the longest of the four transcripts. NR_037173.1 consists of three exons; Exon 1: nucleotides 1 through 189, exon 2: nucleotides 190 through 415 and lastly exon 3: nucleotide 416 through 1498. Homologues were found by using BLAST (Basic Local Alignment Search Tool). BLAST is a program that finds regions of similarity between biological sequences. BLAST can be used for different biological sequences, consisting of RNA, DNA and proteins. We have specifically used RNA to RNA BLAST for our research and the parameters were kept to the BLASTN default settings. When searching for the whole sequence of NR_037173.1 the results were not straight forward. There were no significant matches for RNA sequences that showed complete homologues of the NR_037173.1 sequence. Therefore, we decided to search for each exon separate. When an exon homologue was identified the other exons were searched for in the same organism. If all exons were present, they would be fused into a putative transcript similar to the human NR_037173.1. only exons that had an alignment score of 80 or higher were chosen. In total, eleven homologous transcripts in different organisms were identified. These organisms were split into three groups: Primates, distant relatives 1 and distant relatives 2. The organisms for each group are shown in table 1.

Table 1. Organisms that were used for the PSMB8-AS1 orthologues divided into their respective categories.

Primates	**Distant relatives 1**	**Distant relatives 2**
Chimpanzee (XM_009450991.3)	Horse (LT745777.1)	Pig (CU633196.10)
Gorilla (AC270182.1)	Cattle (AY957499.1)	Cat (EU153401.1)
Common marmoset (AC242643.3)		False killer whale (AB989436.1)
Rhesus Macaque (KT332315.1)		Wild yak (CP027091.1)
Gelada (XR_003119153.1)		

Structure prediction

To predict the structure of the lncRNA, we first had to find the sequence regions that showed a high potential of structure conservation. To find these regions we used RNAz. RNAz is an algorithm that uses a multiple sequence alignment to find conserved and thermodynamically stable structures (Gruber et al, 2010). The thermodynamic stability is expressed by the number of standard deviations by which the MFE of a structure deviates from the mean MFE of a set of randomized sequences that have the same length and base composition (Gruber et al, 2010).

RNAz evaluates the conserved RNA structures in terms of the structure conservation index (SCI). A consensus structure is predicted using the RNAalifold algorithm. RNAalifold is an algorithm that predicts a secondary RNA structure of multiple sequences, with the constraint that all sequences must

sfold into a common structure (Gruber et al, 2010). By calculating the ratio of the consensus folding energy to the unconstrained folding energies of the single sequences RNAz measures structural (Gruber et al, 2010). RNAz is based on multiple alignments of the sequences. For the multiple alignment, Clustal Omega was used. Clustal Omega is a program that is used for the multiple sequence alignment of three or more sequences (Sievers & Higgins, 2014). Clustal Omega is the current standard version of the Clustal software, with the first being released in 1988 and created by Desmond G. Higgins (Higgins & Sharp, 1988; Sievers & Higgins, 2014). All Clustal tools create a multiple sequence alignment following three main steps:

1. Do a pairwise Alignment using the progressive alignment method

2. Create a guide tree

3. Use the guide tree to carry out a multiple alignment

Clustal Omega tool is currently available on the European Bioinformatics Institute (EBI).

In RNAz we searched for conserved structures between humans and primates as well as humans and distant relatives 1.

Structure analysis

For the conserved structure obtained from humans with the distant relatives 1 sequences, we checked the conservation of this structure by using mfold on the marmoset and chimp sequences to produce a similar result. The mfold web server describes a number of software applications used for the prediction of the secondary structure of single stranded nucleic acids (Zuker, 2003). The structure is predicted mainly by free energy minimization. Mfold tries to reduce the limitations of free energy minimization structure predictions by generating suboptimal structures with a similar low free energy. Mfold also allows the user to force certain regions into base pairs of prohibit the base pairs. This allowed us to see if the conserved structure could also be formed in chimps and marmosets. The structures predicted in chimps and marmosets using mfold were then analysed in the remaining organisms of primates and distant relatives 2 dataset to see if the structure was possible for multiple other organisms. Again, multiple alignment was used to analyse the structures. In the case of the conserved structure in human and primate region 1040-1160 nt, the use of mfold was not necessary. For the region that consisted of multiple adjacent windows we used mfold to generate a structure of the full conserved structure by taking parts of the smaller conserved structured predicted by RNAz.

BLAST search individual conserved regions

Another BLAST search was done on the individual regions that showed possible conservation. The regions chosen were loc1 (800-920, figure 1), loc2 (520-800, figure 2) and loc3 (1040-1160, figure 2). The regions were searched using BLASTN on the default parameters, and the sequences that were chosen showed an alignment score of 80 or higher and the default e-value threshold. Loc2 showed many more similar sequences than loc1 and loc2 therefore we decided to limit the sequences chosen by stopping at an alignment score of 329. Furthermore, the dataset for loc2 was divided into primates and non-primates to avoid an overly large multiple alignment result. Table 2 shows the lists of organisms used for each dataset and the specific accession numbers.

Table 2 Lists of organisms and the accession numbers found using BLAST for each region that showed a possibility of conservation.

Loc1	Loc2 primates	Loc2 non-primates	Loc3
Pongo abelii (XR_654256.2)	Piliocolobus tephrosceles (XR_003307278.2)	Camelus ferus (XM_032462778.1)	Pongo abelii (XR_654256.2)
Nomascus leucogenys (XR_004028058.1)	Pongo abelii (XR_654256.2)	Camelus dromedarius (XM_010994582.2)	Nomascus leucogenys (XR_004028058.1)
Hylobates Moloch (XR_004245799.1)	Rhinopithecus roxellana (XR_004056806.1)	Camelus bactrianus (XM_010949116.1)	Piliocolobus tephrosceles (XR_003307278.2)
Aotus nancymaae (XR_001104500.1)	Pan paniscus (XM_003808603.3)	Urocitellus parryii (XM_026382584.1)	Hylobates moloch (XR_004245799.1)
Piliocolobus tephrosceles (XR_003307278.2)	Nomascus leucogenys (XR_004028058.1)	Ictidomys tridecemlineatus (XM_005318839.3)	Macaca nemestrina (XR_003016375.1)
Rhinopithecus bieti (XR_001880766.1)	Rhinopithecus biet (XR_001880766.1)	Ceratotherium simum simum (XM_014782846.1)	Macaca fascicularis (KT331252.1)
Rhinopithecus roxellana (XR_004056806.1)	Hylobates moloch (XR_004245799.1)	Marmota flaviventris (XM_027943695.1)	Chlorocebus aethiops (AC241599.3)
Chlorocebus aethiops (AC241599.3)	Macaca nemestrina (XR_003016375.1)	Marmota marmota marmota (XM_015487514.1)	Rhinopithecus roxellana (XR_004056806.1)
Papio anubis (XR_002521520.2)	Macaca fascicularis (KT331252.1)	Neophocaena asiaeorientalis asiaeorientalis (KT804704.1)	Rhinopithecus bieti (XR_001880766.1)
Macaca fascicularis (KT330814.1)	Chlorocebus aethiops (AC241599.3)	Vicugna pacos (XM_006202144.3)	Aotus nancymaae (XR_001104500.1)
Macaca nemestrina (XR_003016375.1)	Papio anubis (XR_002521520.2)	Physeter catodon (XM_007107875.3)	Neophocaena asiaeorientalis asiaeorientalis (KT804704.1)
Neophocaena asiaeorientalis asiaeorientalis (KT804704.1)	Colobus angolensis palliates (XM_011944813.1)	Equus asinus (XM_014846360.1)	
Pongo abelii (XR_654256.2)	Chlorocebus sabaeus (XM_007972988.1)	Equus przewalskii (XM_008511509.1)	
	Cercocebus atys (XM_012036412.1)	Orcinus orca (XM_004267762.2)	
	Mandrillus leucophaeus (XM_011968949.1)	Puma concolor (XM_025925628.1)	
	Aotus nancymaae (XR_001104500.1)	Phocoena sinus (XM_032648863.1)	
	Cebus capucinus (XM_017522724.1)	Monodon monoceros (XM_029243039.1)	
	Sapajus apella (XM_032253345.1)	Lagenorhynchus obliquidens (XM_027126906.1)	
	Saimiri boliviensis boliviensis (XM_003926393.2)	Tursiops truncatus (XM_004326005.2)	
	Galeopterus variegatus (XM_008592456.1)		

Results

PSMB8-AS1 has four transcripts, of the four NR_037173.1 was chosen because it was the longest of the transcripts. Using BLAST, we constructed a dataset of NR_037173.1 orthologues from different organisms. The organisms were divided into three groups: primates, distant relatives 1 and distant relatives 2. The group of primates consisted of: chimpanzees, gorilla, common marmosets, geladas and rhesus macaques. The distant relatives 1 consisted of cattle and horses. Lastly, distant relatives 2 consisted of: Pigs, wild yaks, cat and false killer whales. Using Clustal Omega (Sievers et al, 2011), we constructed two multiple alignments: the first for the human NR_037173.1 sequence with the sequences of primates and the second for the human RNA with distant relatives 1. These multiple alignments were then used to compute conserved structures using RNAz (Gruber et al, 2010). Predictions yielded for complementary RNA were not considered, because these do not correspond to PSMB8-AS1 transcripts.

Figure 1 shows the regions of potential conserved structures in the dataset of humans and primates. There are multiple locations of interest for conserved structures. The first location is region corresponding between 520 and 800 nt. This region shows multiple adjacent windows that have a window size of 120 nt. The overlap in the predicted regions indicates that the region 520 – 800 in actuality consists of one larger conserved structure instead of multiple smaller conserved structures. Besides the region 520 – 800 nt, there are two remaining regions of interest, the region between 1040 and 1160 nt and lastly the region between 1320 and 1440 nt. these regions show a possible conserved structure however with a lower p value. What is noticeable is that all the regions for conserved structures are situated within the third exon which is nucleotides 416 through 1498.

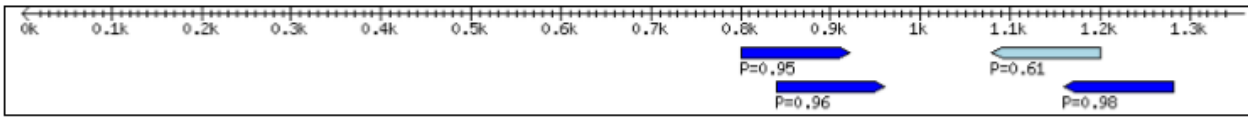

Figure 2. RNAz computed regions of potential conserved structures between human NR_037173.1 and distant relatives 1 orthologues. With window of conserved structure being 120 nt. And with distant relatives 1 database consisting of horse and cattle. The shades of blue show the possibility of the predicted conserved structures. Darker blue has a higher possibility of being a conserved structure. The possibility is also shown in a P-value below the predicted conserved regions.

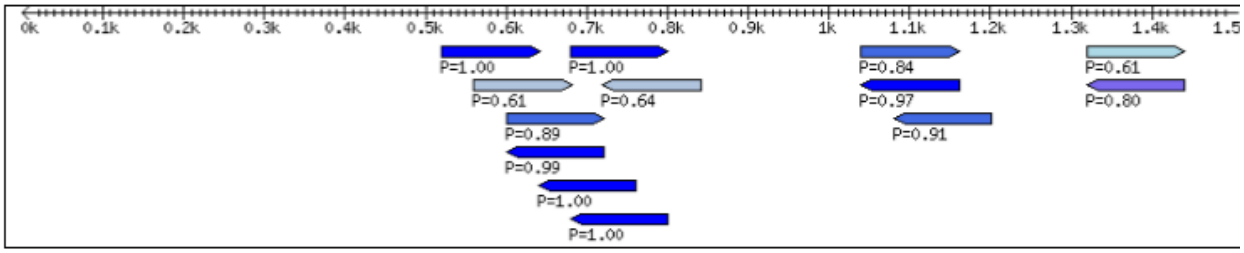

Figure 1. RNAz computed regions of potential conserved structures between human NR_037173.1 and primate orthologues. With window of conserved structure being 120 nt. And with primates being: Chimpanzee, Gorilla, Rhesus macaque, Common marmoset and Gelada. The shades of blue show the possibility of the predicted conserved structures. Darker blue has a higher possibility of being a conserved structure. The possibility is also shown in a P-value below the predicted conserved regions.

Figure 2 shows the potential conserved structures detected in the dataset of humans and distant relatives 1. The conserved structures between humans and distant relatives 1 sequences shows fewer conserved structures than the conserved structures between humans and primates. Two regions are present in figure 1: the first is 800 – 940 nt, the same as region 520 -800 nt in humans and primates. The second region is 1080 – 1280 nt, this region is in antisense and therefore not considered. Similar to figure 1, no potential conserved structures in exon 1 or in exon 2 were detected between human

and distant relatives 1 sequences. Also, there was no similarity discovered between predictions of the datasets: human/distant relatives and human/primates.

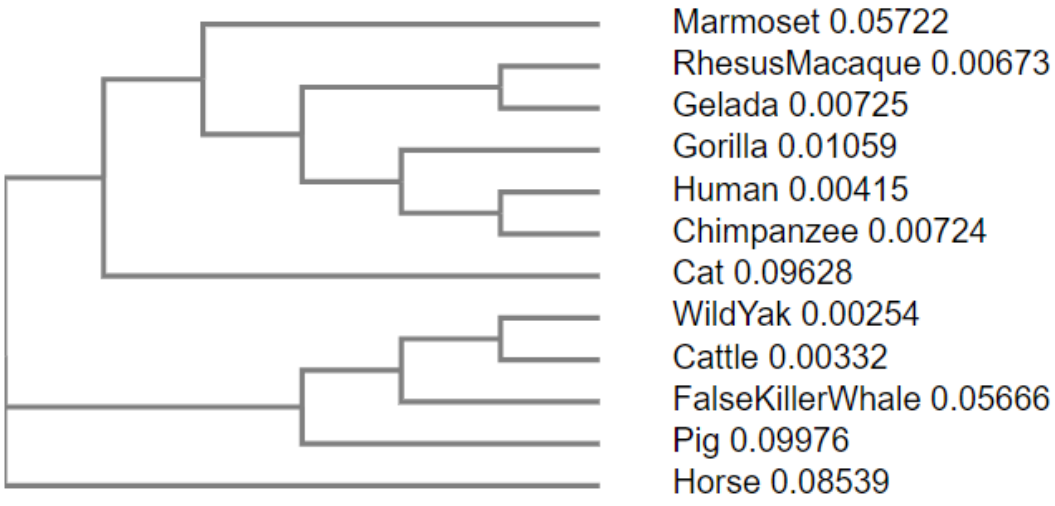

Figure 3 Phylogenetic tree constructed using Clustal Omega Multiple sequence alignment tool. Dataset used: all organisms and all three exons.

Figure 3 shows a phylogenetic tree for all exons of all organisms constructed using Clustal Omega. As expected, the primates are a separate subtree where humans and chimps are the most distantly related organisms. It is surprising to see cats so relatively close to the primate subtree.

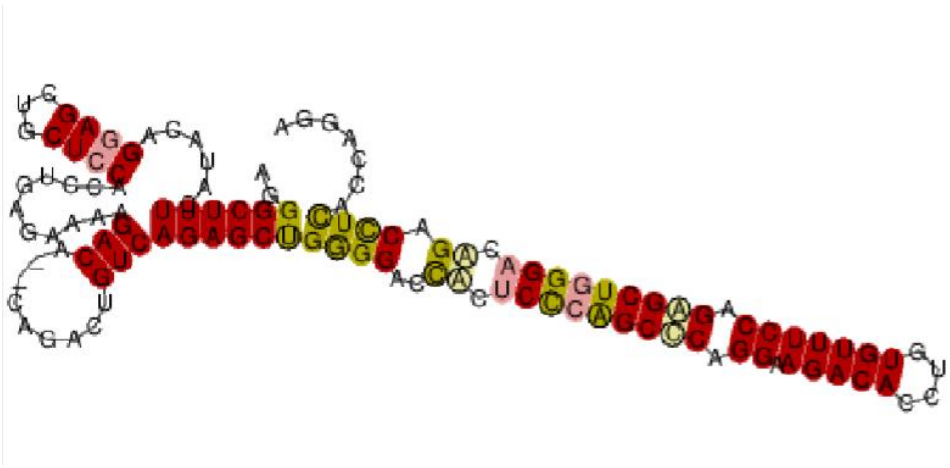

Figure 4. RNAz predicted structure, conserved in the human sequence and distant relatives 1. Region is between nucleotides 800 and 920 of human RNA.

The predicted regions and structures computed using RNA were further used for building structure models and testing the extent of their conservation. The predicted structures for the region between 800 and 920 nucleotides in de human/distant relatives 1 dataset was forced in marmoset and chimp RNA of the same region (fig 5 and fig 6).

Figure 4 shows the predicted conserved structure in region 800- 920 nt between humans and distant relatives 1. This structure was used as a basis to see if this structure could be stable in the primate sequences.

We decided to use the chimp and the marmoset sequence in order to check the conservation of the structure shown in figure 4. Chimp and marmoset were chosen because of the similarity in sequence while still having some differences. By forcing certain base pairs, we computed the structure shown in figure 5 for chimpanzees. And the same was don for the marmoset sequence giving us the structure shown in figure 6. Between these three structures there are two visible similarities. The first similarity is the hammerhead-like structure all three figures have around the 15 – 45 bases. The second similarity is the hairpin with an interior loop around the 70 – 80 bases. These similarities suggest that there is a certain extent of conservation of RNA structures present. While there are similarities, there are also differences between the structures, most notably is that the marmoset structure has a second hammerhead-like structure while an extender stem-loop with interior loop structure seem to be more stable in human and chimp sequence (fig 5,6). However, this difference only concerns specific predictions. When investigating the sequences, we can see that both structures can be formed in all of the sequences.

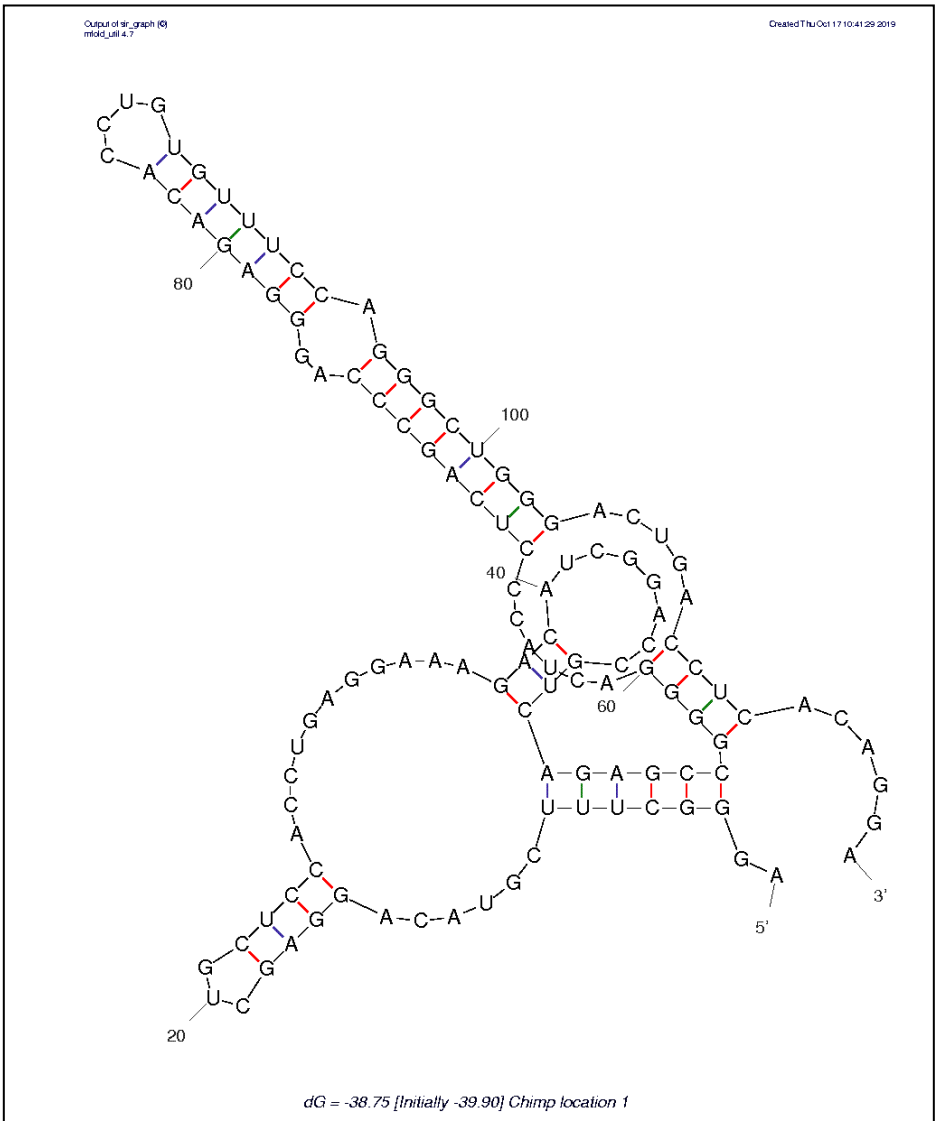

Figure 5. Structure prediction by folding the chimp RNA sequence corresponding with region 800 – 920 nt in human RNA using mfold. The structure in fig 4 was used as a basis to obtain this structure. Forcing bases 4 – 8 together with 51 – 55 and bases 37 – 39 together with bases 48 – 50 and selecting the first result. This structure had an initial ΔG = -39.90 kcal/mol.

Because both structures (interior loop and hammerhead) can be formed in the sequences, we decided to see which of the two is more likely to be formed by analysing the rest of the sequences (figure 7). The cattle, horse, human and chimp sequences all showed the interior loop structure. In these sequences, the parts needed for the interior loop are coloured red. The marmoset sequence showed a possibility of a hammerhead-like formation. The parts needed for this are coloured blue and yellow. Looking at the chimp sequence we can see that the first hairpin of the hammerhead(blue) can also be formed in the chimp sequence, here it is possible for the GGG to pair with the CCC. Furthermore, the second hairpin (labelled yellow) can also be formed. Here the CUG can pair with the GAC, for human the same applies. If we compare the marmoset to the chimp and human sequences, we can see that the interior loop can also be formed in the marmoset. CUGGCACUGACUUC can pair with GACUCCC- - -AGGGG. Some differences are shown between the cattle sequence and the marmoset sequence. In the cattle sequence the first hairpin loop (blue) contains a stem consisting of CAG and CUG while the marmoset contains GGG and CCC, this shows a change either that an AG pair mutated into an CG pair of vice versa. Even though a mutation occurred the pair remained complementary. This gives a more compelling argument towards the hammerhead-like structure. A similar situation is observed at the second hairpin: the marmoset sequence consists of CUG and GAC, the cattle sequence however consists of CUC and GAC. Again, these sequences in both organisms differ however the sequences are still complementary to each other, giving more arguments towards the hammerhead-like structure. In horses, the first hairpin can also be formed because the GGG and the CUC can form pairs. For the second hairpin, the formation is less likely because CAG and GAC do not form a stable structure. Lastly for the rhesus macaque and gelada, a large gap is observed for a large part of the first

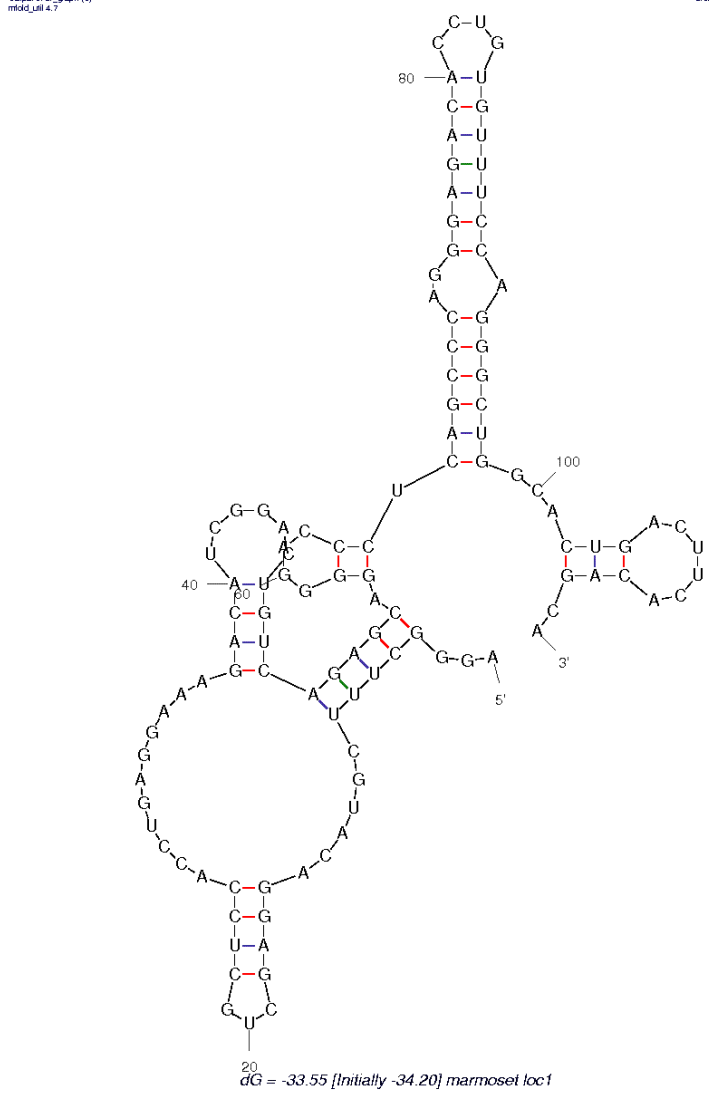

Figure 6. Structure prediction by folding the marmoset RNA sequence corresponding with region 800 – 920 nt in human RNA using mfold. The structure in fig 4 was used as a basis to obtain this structure. Forcing bases 15-19 with 21-25 and bases 4-8 with bases 51-55. This structure had an initial ΔG = -34.20 kcal/mol.

hairpin and the interior loop structure (figure 7). The interior loop cannot be formed due to this gap. In both the rhesus macaque as well as in the gelada there is a complementary sequence present directly downstream from the gap. In the rhesus macaque and gelada sequence the GGG sequence of the first hairpin can pair with the CCC sequence found behind the gap giving a stable stem for the first hairpin. The second hairpin can be formed in both organisms.

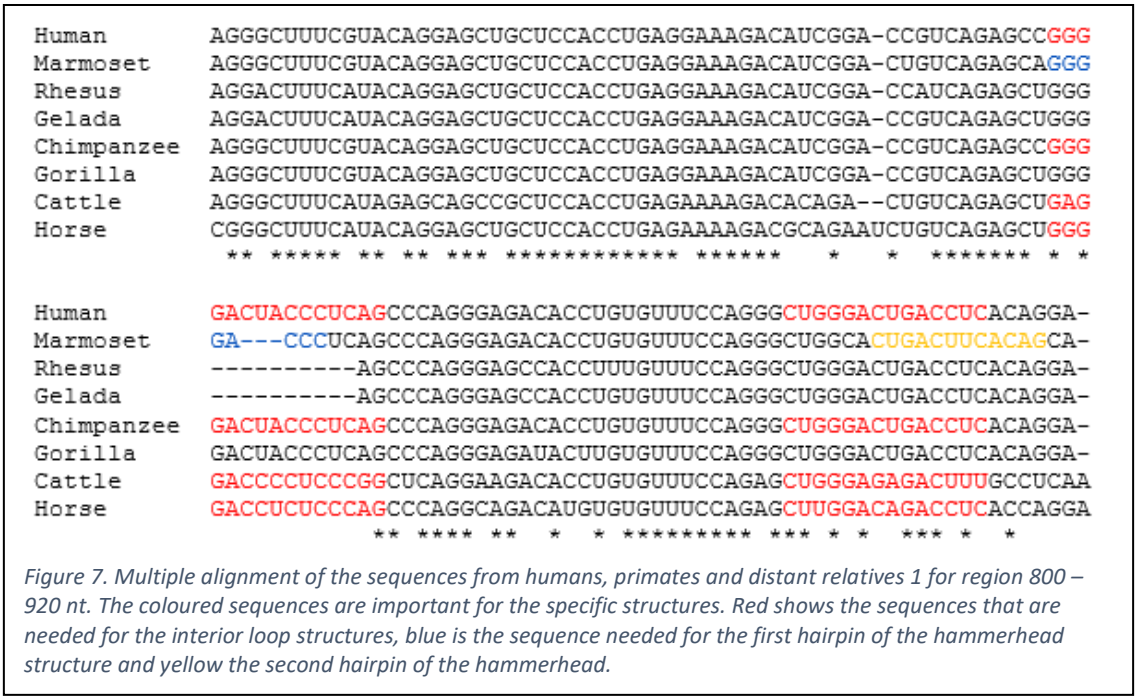

Figure 7. Multiple alignment of the sequences from humans, primates and distant relatives 1 for region 800 – 920 nt. The coloured sequences are important for the specific structures. Red shows the sequences that are needed for the interior loop structures, blue is the sequence needed for the first hairpin of the hammerhead structure and yellow the second hairpin of the hammerhead.

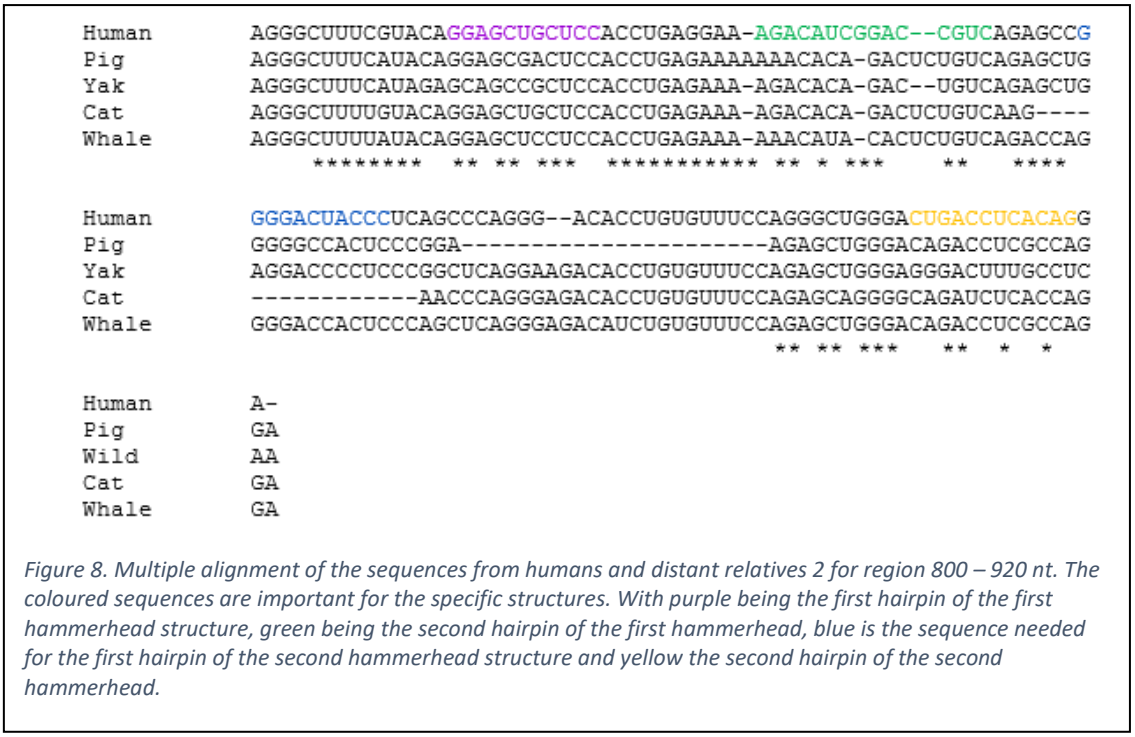

Figure 8. Multiple alignment of the sequences from humans and distant relatives 2 for region 800 – 920 nt. The coloured sequences are important for the specific structures. With purple being the first hairpin of the first hammerhead structure, green being the second hairpin of the first hammerhead, blue is the sequence needed for the first hairpin of the second hammerhead structure and yellow the second hairpin of the second hammerhead.

Figure 8 shows the multiple alignment for region 800 – 920 nt in humans and the distant relatives 2 dataset. The regions that are needed for the certain structures are coloured. When looking at the sequence for the first hairpin of the first hammerhead structure (shown in purple) all sequences, but the wild yak, have GGAG on the 5' side and CUCC on the 3' side of the RNA sequence. These sequences are complementary giving a stable structure. The wild yak has a different sequence for the 5' side, it starts with GCAG. Here a guanine had been mutated into a cytosine. However, a hairpin can still be formed with AG and CU, though this will be less stable. For the second hairpin (green), GAC and GUC is present for the begin and end respectively in all sequences except the pig. The pig sequence can still form a leFss stable stem with AC and GU. The first hairpin of the second hammerhead (shown in blue) in humans has GGG on the 5' side and CCC on the 3' side. In pigs and false killer whales the CCC is mutated to CUC, this can still form a stable structure because guanine and uracil can make a semi-stable bond. In yaks, both sides are mutated. The GGG is mutated to GAG and the CCC is mutated to CUC, this shows that there is a possibility that there is evolutionary pressure to mutate a less stable U-G bond into a more stable U-A bond or G-C bond. This hairpin is completely missing for cats. The region between the second hammerhead's first hairpin (blue) and the second hammerhead's second hairpin (yellow) there also is a longer hairpin with interior loop. A notable result is that pigs are completely missing this region. The last hairpin can only be stably formed in yaks where the sequences are GGG and CCU.

For the region 520 – 800 nt we had to combine the RNAz results for individual windows into a larger structure. To do this we used Mfold, so we could force parts of the smaller structures into the full-length structure. This result is shown in figure 9.

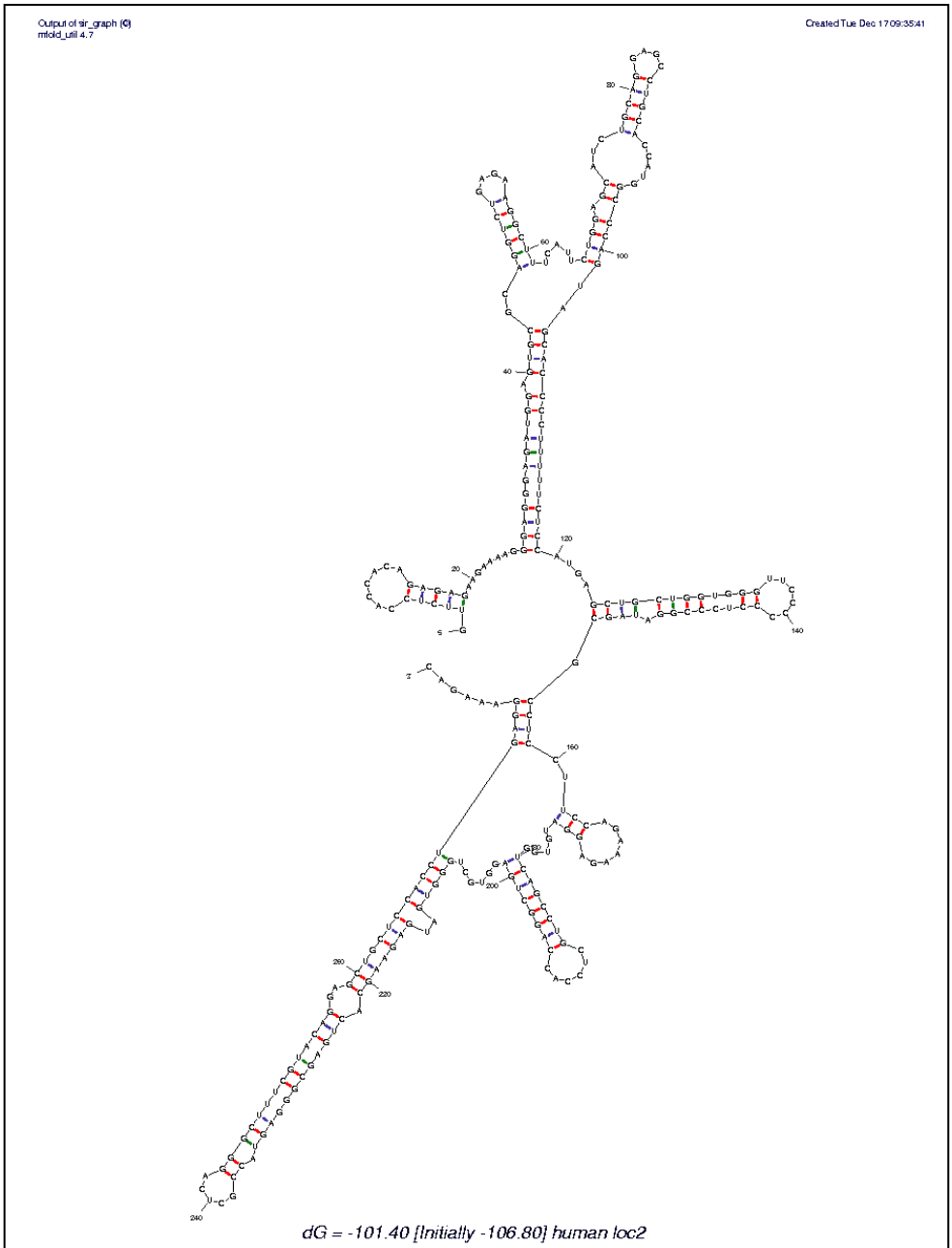

Figure 9. Structure prediction by folding the human sequence region 520 -800 using mfold. And forcing bases 2 – 7 together with bases 14 – 18, bases 181 – 186 together with 195 – 200 and 208 – 211 together with 267 – 270. The structure had an initial free energy of -106.80.

In the structure prediction for region 520 – 800 nt there are some structures of interest, mainly the hairpin structures. There are seven hairpins, some of the stems of these hairpins seem to be relatively stable. The sequence between 180 nt and 200 nt is an example of where this stem shows complete complementary sequences. Figure 10 shows the multiple alignment of PSMB8-AS1 in diverse species, where each hairpin is assigned a different colour.

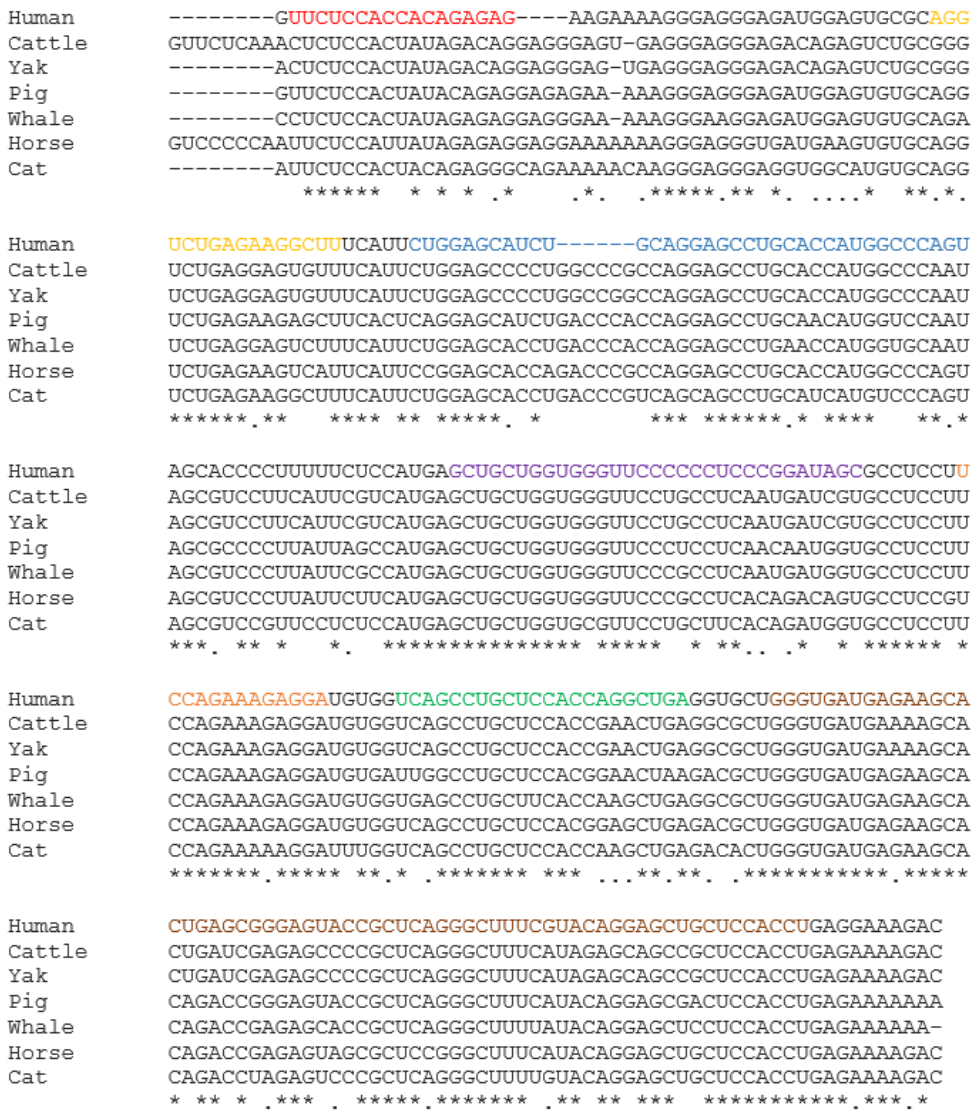

Figure 10. Multiple alignment of the sequences from humans, distant relatives 1 and mammal orthologues for region 520 – 800. Each colour represents a hairpin in the two-dimensional structure.

The first hairpin (red) has the sequences UUCUC and GAGAG that can pair in the human RNA. Horses show no differences compared to humans. The results in cats show have a difference, the complementary sequence in cats is GAGGG. Even though the difference, pairs can still be formed in this hairpin. In false killer whales the first uracil in UUCUC is changed into a cytosine. This allows for a more stable bond to be formed with the second sequence of the stem. In pigs the second sequence of the stem is changed, the GAGAG is changed into a UAGAG. Yaks and cattle both have the same differences when compared to the human sequences. In yaks and cattle, the first uracil is changed into a cytosine. Also, in the second sequence of the stem the second guanine is changed into a cytosine.
In the second hairpin (yellow) the stem consists of AGGUCU and AGGCUU. The sequences for cattle and wild yak are homologous to teach other but different from the human sequence. The sequences

in cattle and wild yak are GGGUCU and AGUGUU, resulting in a mismatch. A similar mismatch is also present in the pig sequence though to a lesser extent in pigs the sequence has a mismatch in G-G. Pigs however, have another difference in their sequence. The fourth pair is a U-A pair instead of a U-G pair. In false killer whales there is again a double mismatch in the third (A C) and fourth (U U) pair. In horses there are two mismatches in the second (G A) pair and the fourth (U U) pair. Lastly, cats have no differences compared to the human sequence. The next hairpin (blue) is a hairpin that contains an interior loop. The human sequence contains an insert that the other organisms do not have. In the sequences from the other organisms, this extra sequence starts with either one or two stable base pairs and is followed by an interior loop until the pairs C-G and A-U. Giving the structure two interior loops instead only the one that the human sequence has. The other sequences, except the one of cats, have mismatches in the first few base pairs. Usually these mismatches are present in the first and/or the second base pairs; however, in false killer whales there is a mismatch in the fourth base pair.

The fourth hairpin (purple) is a hairpin that has a mismatch in its stem (U C), on base pair nine and a single adenine on position five. In cattle and wild yak there are four mismatches, these are the base pairs six through nine. In pigs, base pairs five through nine are mismatched. In false killer whales again sig through nine and in horses and cats, base pairs eight and nine. What is interesting is that all sequences have consecutive mismatches. The differences in sequence, besides these mismatches, all allow for stable base pairs to be formed.

The fifth hairpin (orange) is almost completely unchanged. There is one change in cats where a G is replaced with an A; however, this nucleotide resides in a loop, thus is not important for the formation of base pairs.

The sixth hairpin (green) in humans has quite a large stem with eight base pairs. All the sequences, besides humans, have between one and three mismatched pairs. Some of the mismatched are due to the change of one nucleotide. In the case of the pig sequence, in the second base pair both bases are different from the human sequence but remain a stable pair. There is also a lot of variation in the seventh base pair where the difference lies in a U-A pair or a U-G pair. For the last hairpin (brown) the alignment shows that there are not as many mutations between the sequences, large parts of the sequences remain the same.

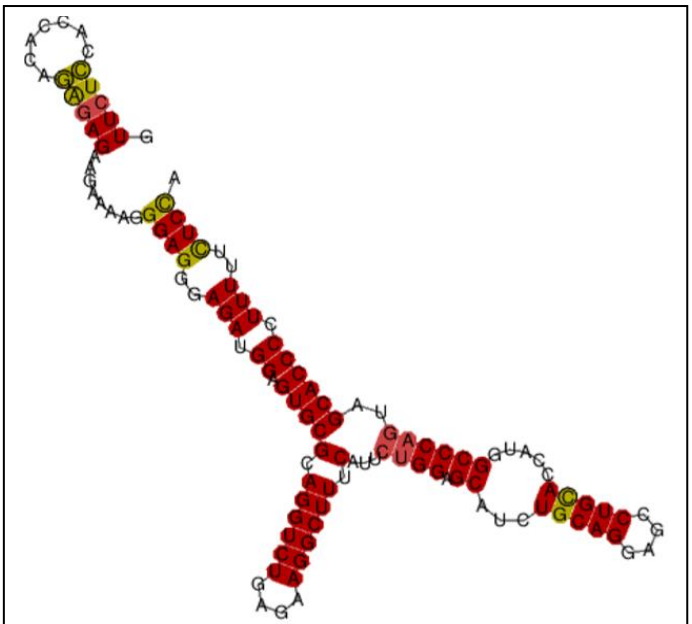

Figure 11. RNAz predicted structure, conserved in the human sequence and the primate orthologues. Region is between nucleotides 520 and 640 of human RNA.

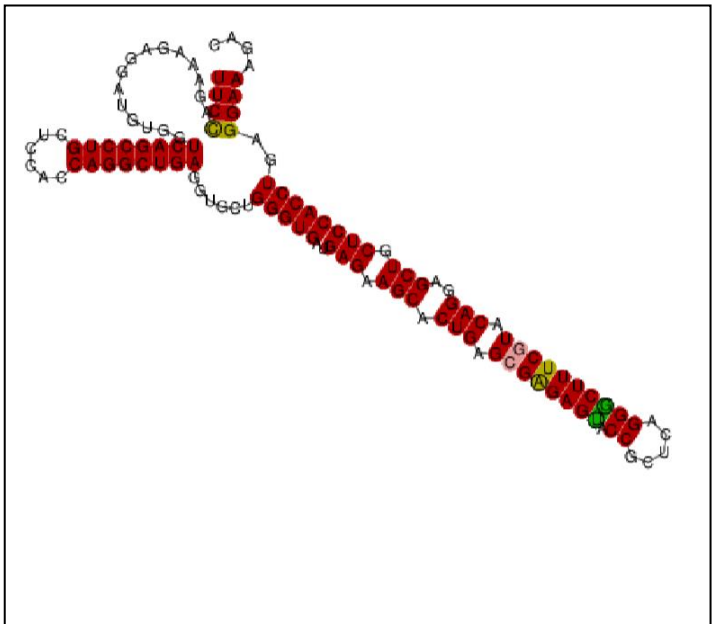

Figure 12. RNAz predicted structure, conserved in the human sequence and the primate orthologues. Region is between nucleotides 680 and 800 of human RNA.

The RNAz prediction for the region 520-800 of the human sequence with the primate orthologues consisted in actuality of two smaller predictions. The first was the region from 520-640 and the second was the region from 680-800. These two structures were also analysed with figure 11 showing the structure for region 520-640 and figure 11 showing the structure for region 680-800.

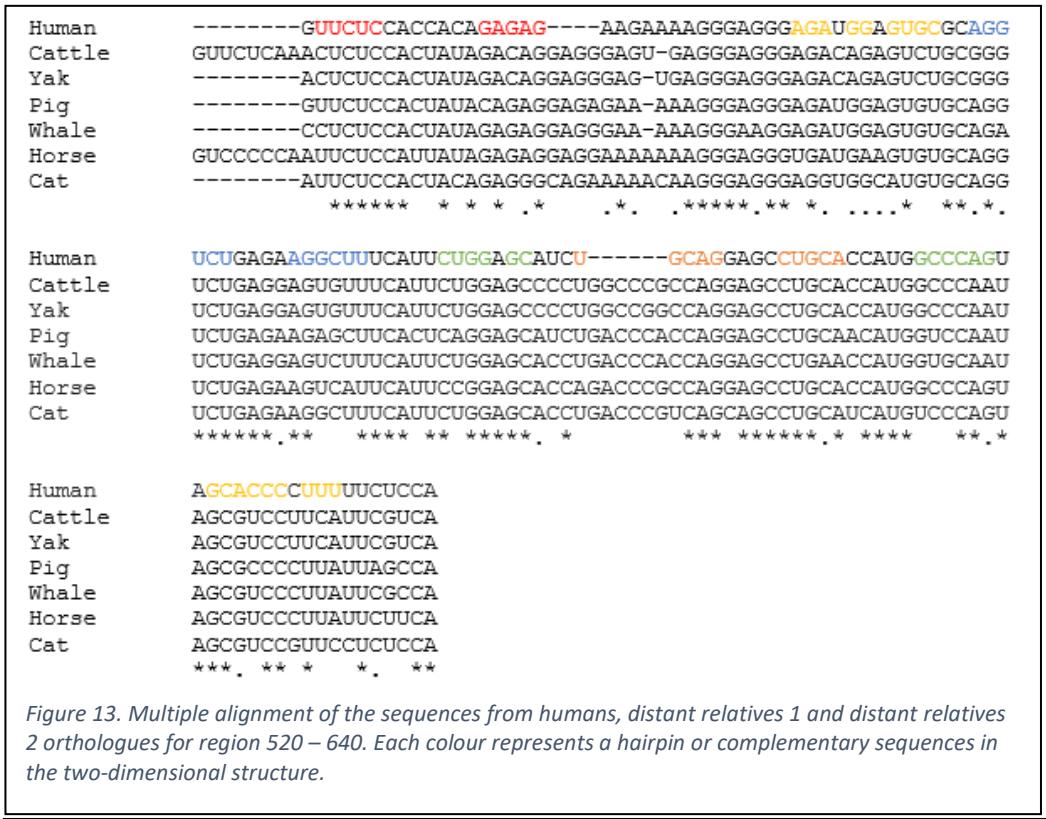

Figure 13. Multiple alignment of the sequences from humans, distant relatives 1 and distant relatives 2 orthologues for region 520 – 640. Each colour represents a hairpin or complementary sequences in the two-dimensional structure.

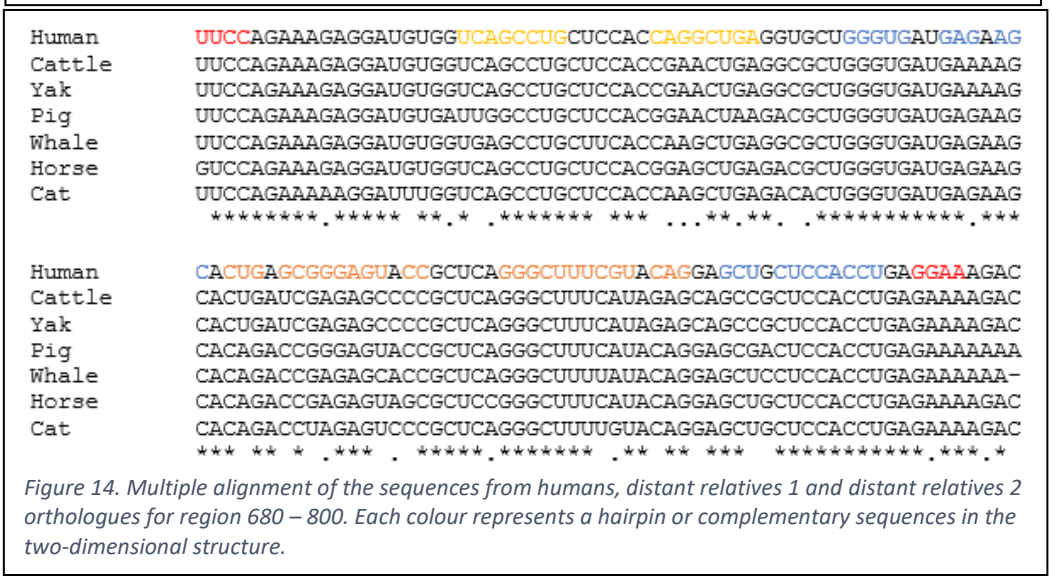

Figure 14. Multiple alignment of the sequences from humans, distant relatives 1 and distant relatives 2 orthologues for region 680 – 800. Each colour represents a hairpin or complementary sequences in the two-dimensional structure.

Multiple alignment was done for region 520-640 (figure 13) and for region 680-800 (figure 14). The sections of interest were coloured. Starting with region 520-640 (figure 13), the first coloured sections (shown in red) show minimal differences between orthologues. Humans, pigs, horses and cats all have the sequence UUCUC while in cattle, yaks and whales this first uracil is a cytosine. The second section also show some differences. Humans, whales and horses contain the sequence GAGAG. Cattle and yaks have the sequence GACAG. Pigs contain the sequence CAGAG. And finally, cats have the sequence GAGGG. Humans, horses, whales and cats show no constrictions in the ability to form a stem using these sequences. In cattle and yaks a mismatch occurs in a C-C pair on the third base. However, for

cattle and yaks it is possible to form a stem with a region downstream from the second red section. Both the cattle and yak orthologues contain a GAGGG sequence which can make a more stable stem. Pigs have a mismatch in the last base as a C-C pair. However, the same as with yaks and cattle a more stable stem can be made with a sequence a bit downstream. Pigs contain the sequences GAGAG which has no mismatches with the first section of this stem.

The next section of importance is marked yellow (figure 13). This section is split into a three base stem and a six base pair stem. Starting with the three base pair stem, in cattle and yaks the first base of the second section is different from humans in humans this whole sequence is UUU while in cattle and yaks this sequence is ACU, meaning an A-A mismatch in the first bases. Pigs and whales also have this mismatch however, the second base in the second section is the same the human sequence. Both bases of the first base pair are different from humans. In horses the first base pair is a U-A pair instead of the A-U pair humans have. Cats the bases are AGG for the first section and CUU for the second. Giving am A-C mismatch in the first base, however stable pairs in the other two bases.

For the six base pair stem more differences are observed. Cattle and yaks have multiple differences from the human sequence. The changes result in a mismatch in the first base (A-C pair) and the second to last base (C-C pair). There are also changes that keep stable pairs. In the second section the third and the fourth bases are G and U respectively which are complementary to U and G in the first section. In the first section of the pig and whale sequence the last base is a U this does not result in any mismatched pairs. In the second part of whales the third to last is a G which also does not result in any mismatched pairs. In cats there are no mismatched however there are differences that keep stable pairs. The third and fourth pair are different from the human sequence.

The next section of interest (shown in blue) the first section has a few differences in the orthologues. The first section in pigs, horses and cats are identical to the human sequence. While cattle and yaks the first base is a G. In whales the third base is an A. In the second section of cattle and yaks the third and fourth bases are different from the human sequence, this means that there is a mismatch in the third and fourth base pairs. Pigs have a mismatch in the fourth base pair, however there are also differences that keep stable pairs. In whales the third and fourth base pair are mismatches. Horses have mismatches in the third and fifth base pairs. Cats lastly have no differences from the human sequence.

The region coloured green has multiple differences from the human sequence. Cattle and yaks have a mismatch in the first sequence. Pigs also have a mismatch in the first pair, furthermore pigs have mismatch in the second pair but also a difference from the human sequences that does not result in a mismatch in the fifth pair. Whales also have mismatch in the first base pair and the fourth pair. In horses the second pair differs from the human sequence but is still a stable pair. But the second to last pair is a mismatch. Cats do not have any mismatches. The second pair is different but still stable.

Looking at the last marked sequence (orange) in the multiple alignment of the region 520-640 (figure 13). The most noticeable feature is that the human sequence has a gap within the first section. This gap causes a mismatch in the first base pair in all the orthologues. The second base pair also is a mismatch in all the orthologues.

The multiple alignment for the region 680-800 (figure 14) has four smaller regions that were analysed. The first is the region coloured red. For all sequences beside the human sequence the second third base of the second sequence is an A instead of a G. This difference would cause a mismatch in the third sequence however it is possible for the UUC of the first section to pair with the AAG of the second part. Giving a three base pairs stem instead of a four. Horses have another difference. The first base of the first section is a G causing a mismatch. What is possible for this section is that the GUC can pair with the last three bases of the whole 680-800 sequence which are CAG.

The second region (marked in yellow) for cattle and yaks have two mismatched pairs, the fifth and sixth pairs are mismatched. The seventh pair is different from humans however does not cause a mismatch. The pig sequence has quite a lot of mismatches. In pigs the fifth, sixth and eighth pair all are mismatches. While the second pair is a completely different pair from humans. Whales have mismatch in the second and sixth pair. Horses have mismatches in the sixth and eighth pair. The

seventh pair is different from humans however does not cause a mismatch. Cats only have a mismatch in the sixth pair and have no further differences.
Next is the third region (marked in blue). This region is mostly the same in all orthologues. However, there are some differences. In cattle, yaks and pigs the third pair are mismatched. Cattle and yaks also have an extra mismatch in the fourth pair.
The final region (marked in orange) also had differences between orthologues. In cattle and yaks the third through fifth pairs all are mismatches. The seventh pair is different from humans but still complementary. Pigs have mismatches in the second, fourth and fifth pairs and no further differences. Whales the same as pigs also have mismatches in the second, fourth and fifth pairs. Whales however have more differences than pigs. The sixth and seventh pair are different from humans but still complementary. Horses, the same as pigs and whales, have mismatches in the second, fourth and fifth pair but also another mismatch in the twelfth pair. The eight pair is different from humans but still make a stable pair. Lastly, cats have mismatched in the second, fourth and sixth pairs. There are also differences that keep stable pairs in the seventh pair.

The predicted structure for the region 1040-1160 nt for the alignment of human with primates is shown in figure 15. This structure is extended and has very little mismatched pairs.

Figure 16 shows the multiple alignment of the human sequence with both distant relatives datasets. The regions that are complementary are coloured. This region was not present in the false killer whale sequence. Starting with the sequence labelled in blue. In all sequences, besides humans, the third base pair is a mismatch. In horses the first base pair is also mismatched. In cattle, wild yak and cats the final base pair of this sequence is mismatched in an A-A pair. Cats have a third mismatch in that the fourth to last pair is mismatched. Other changes are A-U to G-U differences. The green sequence has a mismatch in the first sequence for pigs, cats, wild yaks, cattle and horses. Besides horses al other sequences have a mismatch in the second to last base pair.

The sequence of horses is the only one that is different and has a mismatch in the second pair.

The orange sequence also has a few differences. Part of the sequences in cattle and wild yaks are missing and there are no other regions that have complementary sequences. The only difference in the cats is that the U-A pair is switched to a U-G pair. Pigs have a single mismatch in the last pair and horses have two mismatched in the first and last pair.
For the yellow labelled sequence. Pigs, wild yaks and cattle are fully missing the first sequence. Cats and horses have similar mismatches in the third and last pair, but cats have an extra mismatch in the
first pair. In the last sequence (marked red) there are a few minor changes, however none of these inhibit the forming of pairs.

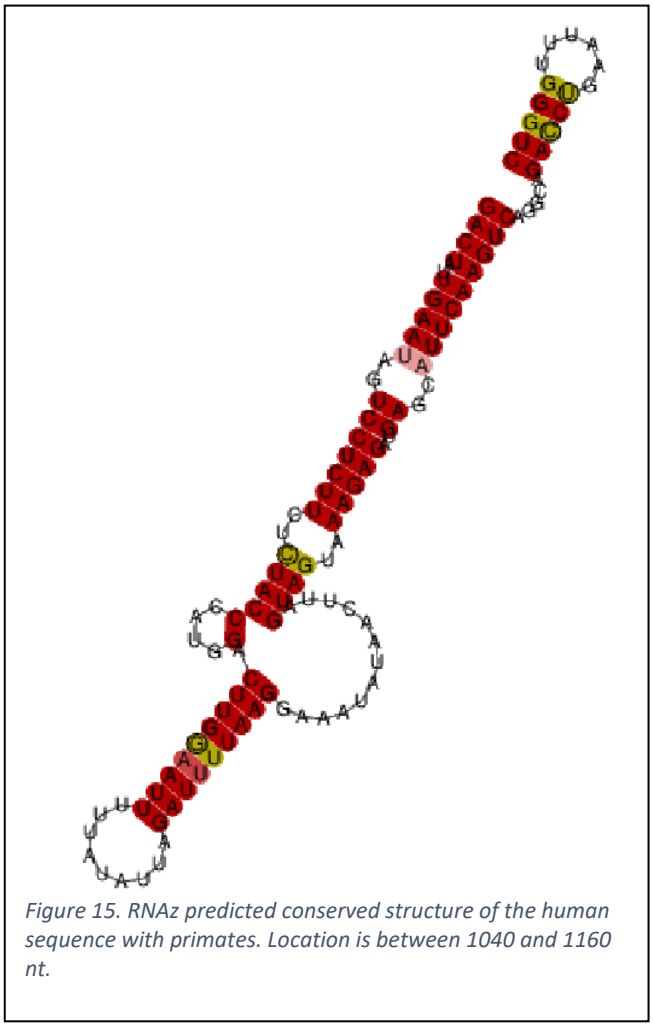

Figure 15. RNAz predicted conserved structure of the human sequence with primates. Location is between 1040 and 1160 nt.

Figure 16. Multiple alignment of the sequences from humans, distant relatives 1 and distant relatives 2 for region nucleotides 1040 – 1160. The colours represent the sequences that are complementary.

The last conserved region found in the RNAz prediction of humans with primates was the region from nucleotides 1320 to 1440. This region is not present in all the other sequences in distant relatives 1 and 2.

Next, we used BLAST to search for the individual regions that showed potential conservation the regions were loc1 (800-920, figure 1), loc2 (520-800, figure 2) and loc3 (1040-1160, figure 2).

Figure 17. Multiple alignment of the blast search for loc1 (800-920, figure 1) sequences used are shown in table 2 under loc1. The coloured sections represent the sequences of interest.

When looking at the multiple alignment of Loc1 (figure 17), we can see that a large part of the sequences are unchanged in the different organisms. Even in the narrow-ridged finless porpoise, which is the only non-primate most of the sequences is the same as the sequences found in the different primates. If we look at the coloured sections, we can see that the purple section is conserved in its entirety. The section marked green is mostly the same over all sequences. There are differences in single sequences. The sequence most different from the rest is the porpoise which is as expected seeing that it is the only non-primate. Only the crab-eating macaque and the southern pig-tailed macaque show changes that would inhibit bonds to be formed in the third pair. In the porpoise and night monkey can form an extra bond that humans and the rest of the sequences cannot form, in the fourth pair. For the section marked blue a lot of the sequences have a similar gap that rhesus macaques and geladas have. The last section (marked yellow) show no differences from the human sequence except for the porpoise. The porpoise could still form a hairpin using this section by forming bonds using the GGG in a bit earlier in the sequence to bind with the CUC at the end.

Loc2 had a lot of sequences therefore it was decided to limit the amount of sequences in the BLAST search. It was also decided to divide the dataset into two datasets, one for primates and the other for non-primates. For primates we expected to have a lot of similarities among sequences.
We started with the multiple alignment of loc2 for primates (figure 18, 19). The section marked red showed that the sequences for night monkeys, black-capped squirrel monkeys and both capuchins had an insert consisting of GAG. This insert caused a mismatch in the fourth base pair however it was possible for the first two bases to pair with the AA directly behind the red section which would form a stable stem consisting of 6 base pairs. The capuchin sequences have a change where a C is replaced with a U, which keeps the stable bond. The rest of the changes are in the loop. The section marked yellow has a few differences the Sunday flying lemur has two differences the first is on the sixth base which is a C in the lemur and the thirteenth base which is a U in the lemur both differences cause mismatches in their respective base pairs. The last difference is in the Angola colobus in the fourth base the colobus has an A, however this change did not cause a mismatch. In the blue section the sequences for the Sunda flying lemur and olive baboon have a difference in the fifth to last base. In both sequences this base is a U, which pairs with a G resulting in no mismatch. For the sequences found in night monkeys, black-capped squirrel monkeys and both capuchins the last base is an A instead of a U. This difference results in a mismatch in the first base pair. Another important difference is found in the twenty third base almost half of the sequences have a U instead of a C, this difference does not cause a mismatch. What is notable is that other differences that do cause mismatches are only present in one or two sequences. For example, the sequences for night monkey, black capped squirrel monkeys and the capuchins have different differences in the twelfth to fourteenth bases which cause mismatches, but these changes are contained to one or two sequences and not multiple sequences. The purple section has the same phenomena as the previous section in the purple section the last base can either be a C or a U, both bases can stably bind with G. In the Sunda flying lemur sequences both the seventh base and the ninth to last base are different which cause a mismatch in the seventh and eighth base pairs. the seventh to last base pair also differs however this does not cause a mismatch. The orange section only shows one differences in the whole section; this change however is not in the stem but in the loop of the hairpin loop structure. The green section only shows two differences. The first is found in the Sunda flying lemur where the sixth to last base is an A instead of a G, this change does not cause a mismatch. the second difference is found in black-capped squirrel monkeys and both capuchins, the seventh base is a C in these sequences causing a mismatch. Lastly the brown section this section shows multiple cases where a base is either a G or a A in different sequences, examples are the twenty third, thirty eighth and the forty fourth bases. The twenty third base and thirty eighth base are shown to bond with a U in the model shown in figure 8. The forty fourth base binds with a C as shown in the model. There are other differences but these cause mismatches and are only present in one or two sequences.

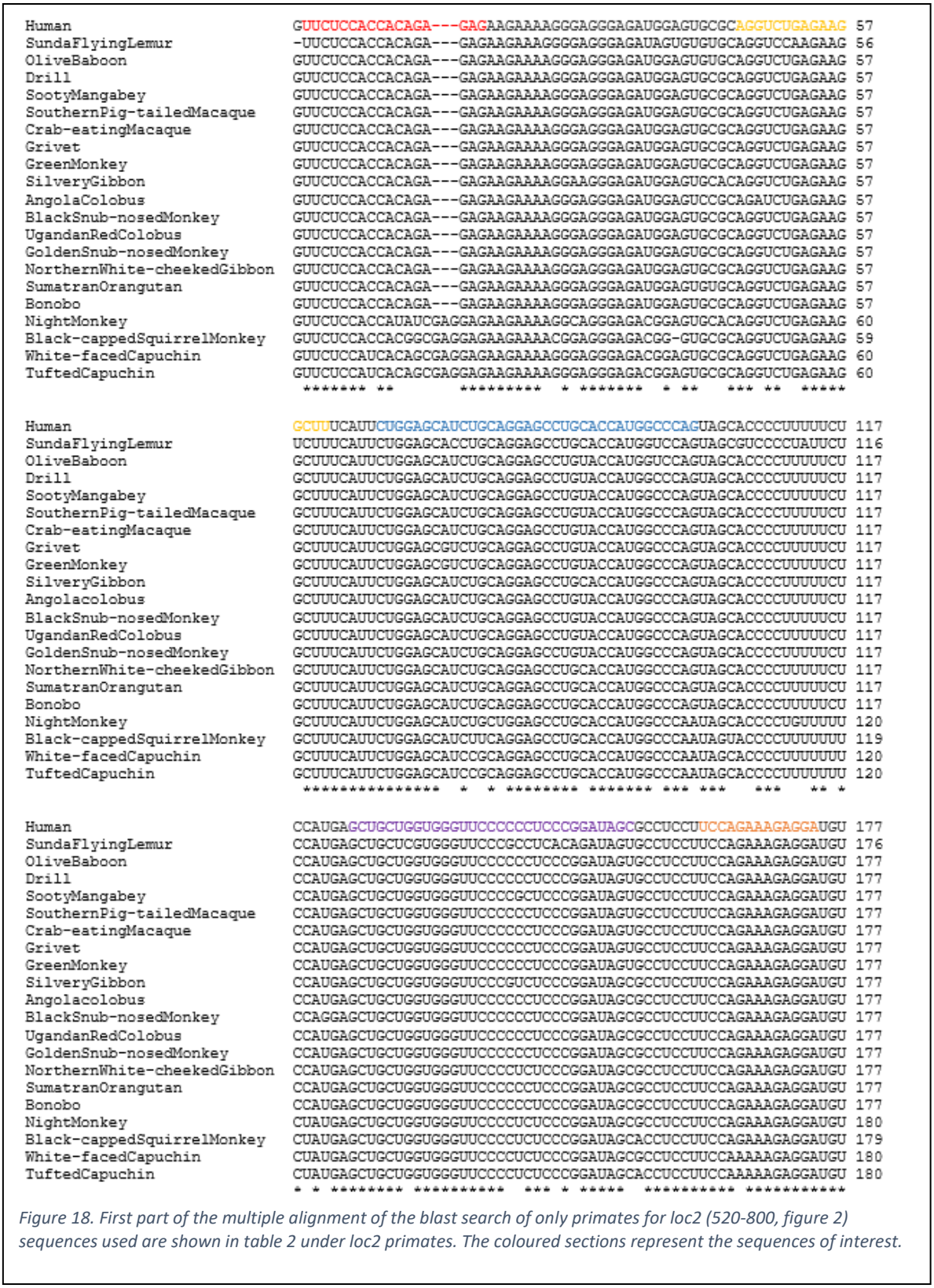

Figure 18. First part of the multiple alignment of the blast search of only primates for loc2 (520-800, figure 2) sequences used are shown in table 2 under loc2 primates. The coloured sections represent the sequences of interest.

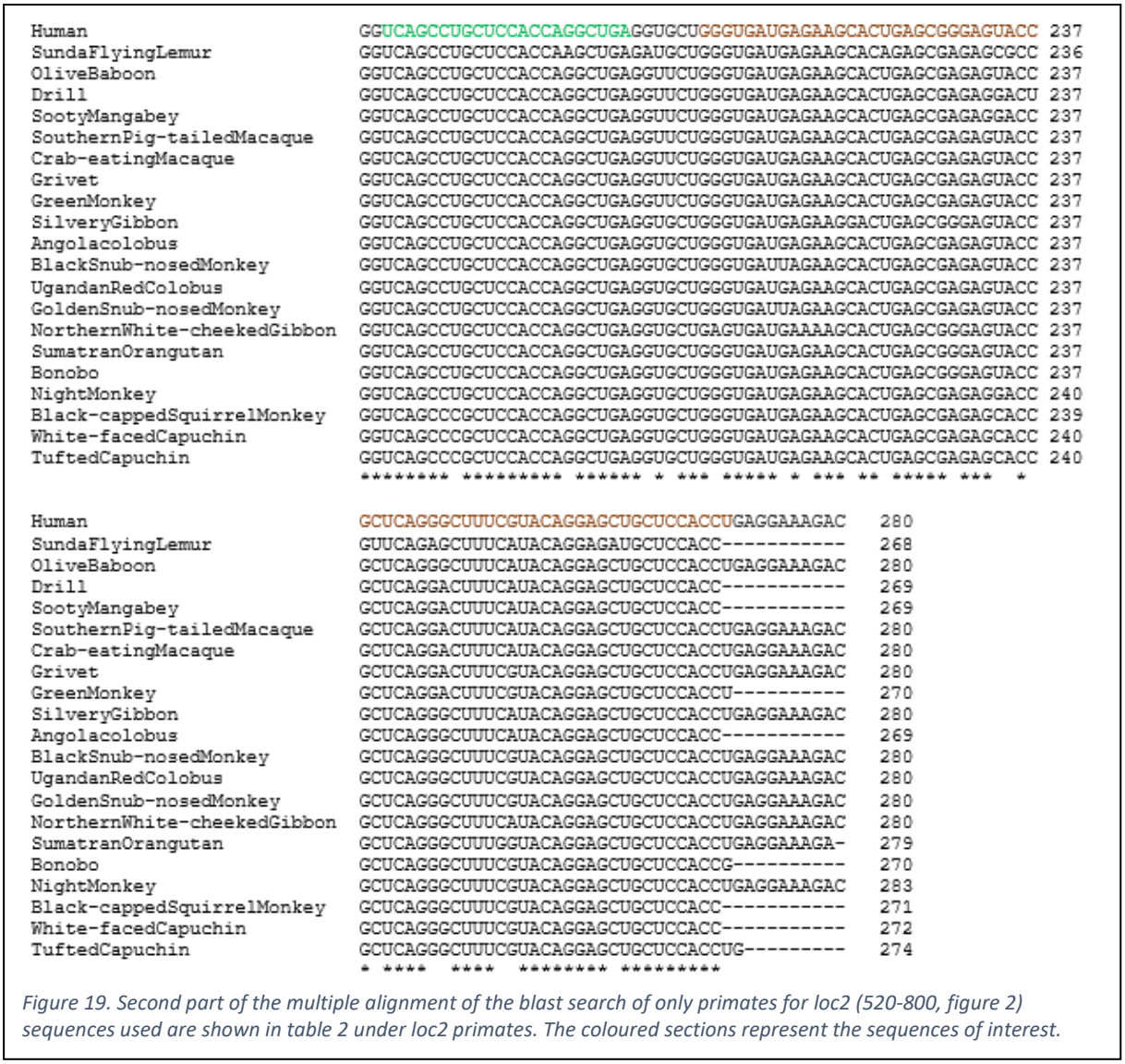

Figure 19. Second part of the multiple alignment of the blast search of only primates for loc2 (520-800, figure 2) sequences used are shown in table 2 under loc2 primates. The coloured sections represent the sequences of interest.

```
Human                        GUUCUCCACCACAGAGAGAAGA----AAAGGGAGGGAGAUGGAGUGCGCAGGUCUGAGAA 56
Yellow-belliedMarmot         GUUCUCCACCACAGAGAACAGAAUA-AAAUGGAGGGAAAGGUAAUGUGCAGGUA-GAGAA 58
AlpineMarmot                 GUUCUCCACCACAGAGAACAGAAGA-AAAUGGAGGGAAAGGUAAUGUGCAGGUA-GAGAA 58
ArcticGroundSquirrel         GUUCUCCACCACAGAGAACAGAAGA-AAAUGGAGGGAAAGGUAAUGUGCAGGUA-GAGAA 58
Thirteen-linedGroundSquirrel GUUCUCCACCACAGAGAACAGAAGA-AAAUGGAGGGAAAGGUAAUGUGCAGGUA-GAGAA 58
Cougar                       -UUCUCCACUACAGAGGGCAGAAAAACAAGGGAGGGAGGUGGCGUGUGCAGGUCUGAGAA 59
Asinus                       -UUCUCCAUUAUAGAGAGGAGGAAAAAAAGGGAGGGUGAUGAAGUGUGCAGGUCUGAGAA 59
Przewalski'sHorse            -UUCUCCAUUAUAGAGAGGAGGAAAAAAAGGGAGGGUGAUGAAGUGUGCAGGUCUGAGAA 59
WhiteRhinoceros              -UUCUCCACUAUAGAGAAGAGAAAA-AAAGGGAGGGAGAGGGAGGGUGCAGGUCUGAGAA 58
WildBactrianCamel            --UCUCCACUAUAGAGAGAAGGUAA-AAAGGGAGGGAGAUGGAGUGUGCAGGUCUGAGAA 57
Dromedary                    --UCUCCACUAUAGAGAGAAGGUAA-AAAGGGAGGGAGAUGGAGUGUGCAGGUCUGAGAA 57
BactrianCamel                --UCUCCACUAUAGAGAGAAGGUAA-AAAGGGAGGGAGAUGGAGUGUGCAGGUCUGAGAA 57
Alpaca                       --UCUCCACUAUAGAGAGAAGGUAA-AAAGGGAGGGAGAUGGAGUGUGCAGGUCUGAGAA 57
SpermWhale                   --UCUCCACUAUAGAGAGGAGGGAA-AAAGGGAGAGAGAUGGAGUGUGUAGGUCUGAGAA 57
Narwhal                      --UCUCCACUAUAGAGAGGAGGGAA-AAAGGGAAGGAGACGGAGUGUGCAGAUCUGAGGA 57
Narrow-ridgedFinlessPorpoise --UCUCCACUAUAGAGAGGAGGGAA-AAAUGGAAGGAGACGGAGUGUGCAGAUCUGAGGA 57
Vaquita                      --UCUCCACUAUAGAGAGGAGGGAA-AAAUGGAAGGAGACGGAGUGUGCAGAUCUGAGGA 57
CommonBottlenoseDolphin      --UCUCCACUAUAGAGAGGAGGGAA-AAAGGGAAGGAGAUGGAGUGUGCAGAUCUGAGGA 57
KillerWhale                  --UCUCCACUAUAGAGAGGAGGGAA-AAAGGGAAGGAGAUGGAGUGUGCAGAUCUGAGGA 57
PacificWhite-sidedDolphin    --UCUCCACUAUAGAGAGGAGGGAA-AAAGCGAAGGAGAUGGAGUGUGCAGAUCUGAGGA 57
                               ******  * ****   **      **  **  *    *    * * ** *  *** *

Human                        GGCUUUCAUUCUGGAGCAUCUG------CAGGAGCCUGCACCAUGGCCCAGUAGCACCCC 110
Yellow-belliedMarmot         GUCUUUCAUUCUGGAGCACCUG------CAGGAGCCUGCACCAUGUCGAAGUAGCAUCCC 112
AlpineMarmot                 GUCUUUCAUUCUGGAGCACCUG------CAGGAGCCUGCACCAUGUCGAAGUAGCAUCCC 112
ArcticGroundSquirrel         GUCUUUCAUUCUGGAGCACCUG------CAGGAGCCUGCACCAUGUCGAAGUAGCAUCCC 112
Thirteen-linedGroundSquirrel GUCUUUCAUUCUGGAGCACCUG------CAGGAGCCUGCACCAUGUCAAAGUAGCAUCCC 112
Cougar                       GGCUUUCAUUCUGGAGCACCUGACCCGUCAGCAGCCUGCAUCAUGUCCCAGUAGCGUCCG 119
Asinus                       GUCAUUCAUUCCGGAGCACCAGACCCGCCAGGAGCCUGCACCAUGGCCCAGUAGCGUCCC 119
Przewalski'sHorse            GUCAUUCAUUCCGGAGCACCAGACCCGCCAGGAGCCUGCACCAUGGCCCAGUAGCGUCCC 119
WhiteRhinoceros              GUCUUUCAUUCCGGAGCGCCUGACCCGCCAGGAGCCUGCACCAUGGCCCAGUAGCGUCCC 118
WildBactrianCamel            GUCUUUCAUUCCGGAGCACCAGACCCACCAGGAGCCUGCACCAUGGUACAAUAGCGUCCC 117
Dromedary                    GUCUUUCAUUCCGGAGCACCAGACCCACCAGGAGCCUGCACCAUGGUACAAUAGCGUCCC 117
BactrianCamel                GUCUUUCAUUCCGGAGCACCAGACCCACCAGGAGCCUGCACCAUGGUACAAUAGCGUCCC 117
Alpaca                       GUCUUUCAUUCCGGACCACCAGACCCACCAGGAGCCUGCACCAUGGUACAAUAGCGUCCC 117
SpermWhale                   GUCUUUCAUUCUGGAGCACCUGAUCCACCAGGAGCCUGAACCAUGGUCCAAUAUCGUCCC 117
Narwhal                      GUCUUUCAUUCUGGAGCACCUGACCCACCAGGAGCCUGAACCAUGGUGCAAUAGCGUCCC 117
Narrow-ridgedFinlessPorpoise GUCUUUCAUUCUGGAGCACCUGACCCACCAGGAGCCUGAACCAUGGUGCAAUAGCGUCCC 117
Vaquita                      GUCUUUCAUUCUGGAGCACCUGACCCACCAGGAGCCUGAACCAUGGUGCAAUAGCGUCCC 117
CommonBottlenoseDolphin      GUCUUUCAUUCUGGAGCACCUGACCCACCAGGAGCCUGAACCAUGGUGCAAUAGCGUCCC 117
KillerWhale                  GUCUUUCAUUCUGGAGCACCUGACCCACCAGGAGCCUGAACCAUGGUGCAAUAGCGUCCC 117
PacificWhite-sidedDolphin    GUCUUUCAUUCUGGAGCACCUGACCCACCAGGAGCCUGAACCAUGGUGCAAUAGCGUCCC 117
                             * * ******* *** *  * *      *** ****** * ****    * ** *  **

Human                        UUUUUCUCCAUGAGCUGCUGGUGGGUUCCCCCCUCCCGGAUAGCGCCUCCUUCCAGAAAG 170
Yellow-belliedMarmot         CUGUUCUCCAUGAGCUGCUGGUGGGUUCCCUCCUCACAGAUGGUGCCUCCUUUCAGAAAG 172
AlpineMarmot                 CUAUUCUCCAUGAGCUGCUGGUGGGUUCCCUCCUCACAGAUGGUGCCUCCUUUCAGAAAG 172
ArcticGroundSquirrel         CUAUUCUCCAUGAGCUGCUGGUGGGUUCCCUCCUCACAGAUGGUGCCUCCUUUCAGAAAG 172
Thirteen-linedGroundSquirrel CUAUUCUCCAUGAGCUGCUGGUGGGUUCCCUCCUCACAGAUGGUGCCUCCUUUCAGAAAG 172
Cougar                       UUCCUCUCCAUGAGCUGCUGGUGGGUUCCUGCUUCACAGAUGGUGCCUCCUUCCAGAAAA 179
Asinus                       UUAUUCUUCAUGAGCUGCUGGUGGGUUCCCGCCUCACAGACAGUGCCUCCGUCCAGAAAG 179
Przewalski'sHorse            UUAUUCUUCAUGAGCUGCUGGUGGGUUCCCGCCUCACAGACAGUGCCUCCGUCCAGAAAG 179
WhiteRhinoceros              UUAUUCUCCAUGAGCUGCUGGUGGGUUCCCGCCUCACAGAUGAUGCCUCCUUCUAGAAAG 178
WildBactrianCamel            UUAUUCUCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCCUUCAAGAAAG 177
Dromedary                    UUAUUCUCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCCUUCAAGAAAG 177
BactrianCamel                UUAUUCUCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCCUUCAAGAAAG 177
Alpaca                       UUAUUCUCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCCUUCAAGAAAG 177
SpermWhale                   UCAUUCGCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCUUUCCAGAAAG 177
Narwhal                      UUAUUCGCCAUGAGCUGUUGGUGGGUUCCCGCCUCAACGAUGGUGCCCCCUUCCAGAAAG 177
Narrow-ridgedFinlessPorpoise UUAUUCGCCAUGAGCUGCUGGUGGGUUCCCGCCUCAGUGAUGGUGCCUCCUUCCAGAAAG 177
Vaquita                      UUAUUCGCCAUGAGCUGCUGGUGGGUUCCCGCCUCAGUGAUGGUGCCUCCUUCCAGAAAG 177
CommonBottlenoseDolphin      UUAUUCGCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCCUUCCAGAAAG 177
KillerWhale                  UUAUUCGCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCCUUCCAGAAAG 177
PacificWhite-sidedDolphin    UUAUUCGCCAUGAGCUGCUGGUGGGUUCCCGCCUCAAUGAUGGUGCCUCCUUCCAGAAAG 177
                                 **  ********* ***********  * **   **    *** *  *  *****
```

Figure 20 First part of the multiple alignment of the blast search of only non-primates for loc2 (520-800, figure 2) sequences used are shown in table 2 under loc2 non-primates. The coloured sections represent the sequences of interest.

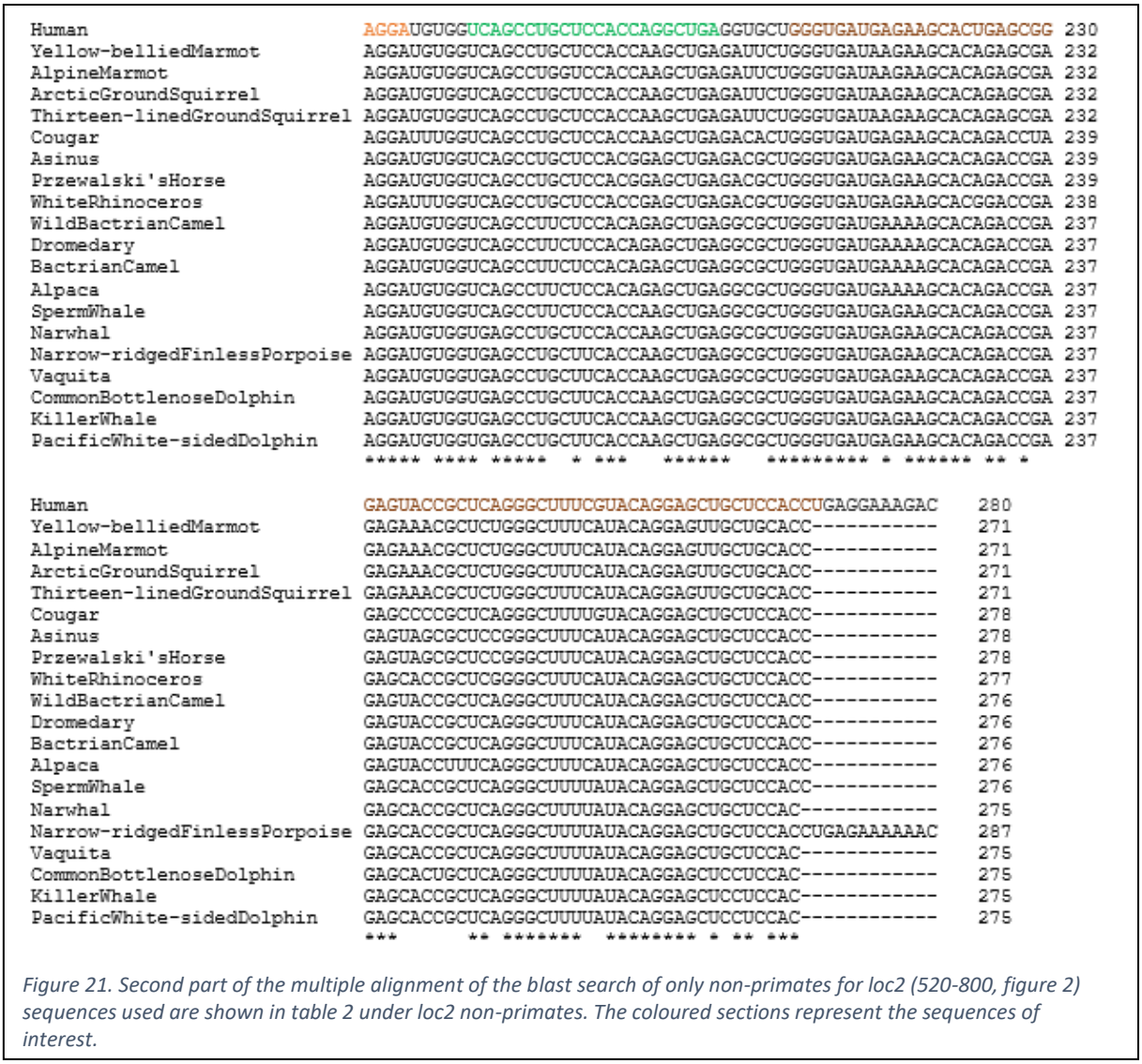

Figure 21. Second part of the multiple alignment of the blast search of only non-primates for loc2 (520-800, figure 2) sequences used are shown in table 2 under loc2 non-primates. The coloured sections represent the sequences of interest.

Next, we look at the multiple alignment for the non-primate sequences (figures 20, 21). Starting with the section marked red. We can see that for a large part of the sequences the first base is missing, which results in one less base pair to be able to be formed. Other differences are A and G switches in pairings with U which do not inhibit base pairs to be formed. The section marked yellow also has an G and A switch in the third base for the sequences found in narwhals, narrow-ridged finless porpoise, vaquita, killer whales and both dolphin sequences. However, this difference is with a G-C pairing which when the G is swapped for an A would be a mismatch. Both marmot sequences and squirrel sequences have the same difference. In these sequences the fifth base is different, and the sixth base is missing. Both changes cause mismatches. In the section marked blue the third and fourth to last bases in the both marmot sequences and both squirrel sequences are different causing mismatches. These sequences together with the cougar sequence also have a difference in the sixth to last sequences where they have a U also causing a mismatch. This section shows more cases of having changes that are sustained throughout multiple different sequences. However, contrary to what was found in the loc2 multiple alignment of primates where the differences that were present in multiple organisms mostly were a part of stable pairings, in this section the bases were a part of mismatched pairs. Most of the differences, were found in only one sequence or very close relatives. In the purple section the last base for all sequences except the human sequences is a U instead of a C which can pair with the first base, which is a G. the same which the third to last base where almost all sequences have an G

instead of an A, thus pairing stably with the U. The ninth to last base is different for all sequences and causes a mismatch and the eighth to last base is different in more than half of the sequences and also causes a mismatch. Looking at the orange marked sequence the second base is a U for all sequences except the human sequence. Both U and C can form stable bonds with the G. The third base is different for the camels, alpaca and dromedary. The change in these sequences do cause a mismatch. The section marked green has changes that do result in mismatches. But are only present in close relative sequences. In the seventh to last base some sequences have a G instead of an A, which both can form stable bonds with the U. This change is present in more sequences than only close relatives. Lastly the section marked brown. The seventeenth base is a U in humans but in the other sequences it is an A, these bases do not have an overlapping base which they can form stable bonds thus resulting in a mismatch. Similarly as with the primate multiple alignment the twenty third and the forty fourth bases are all an A instead of a G.

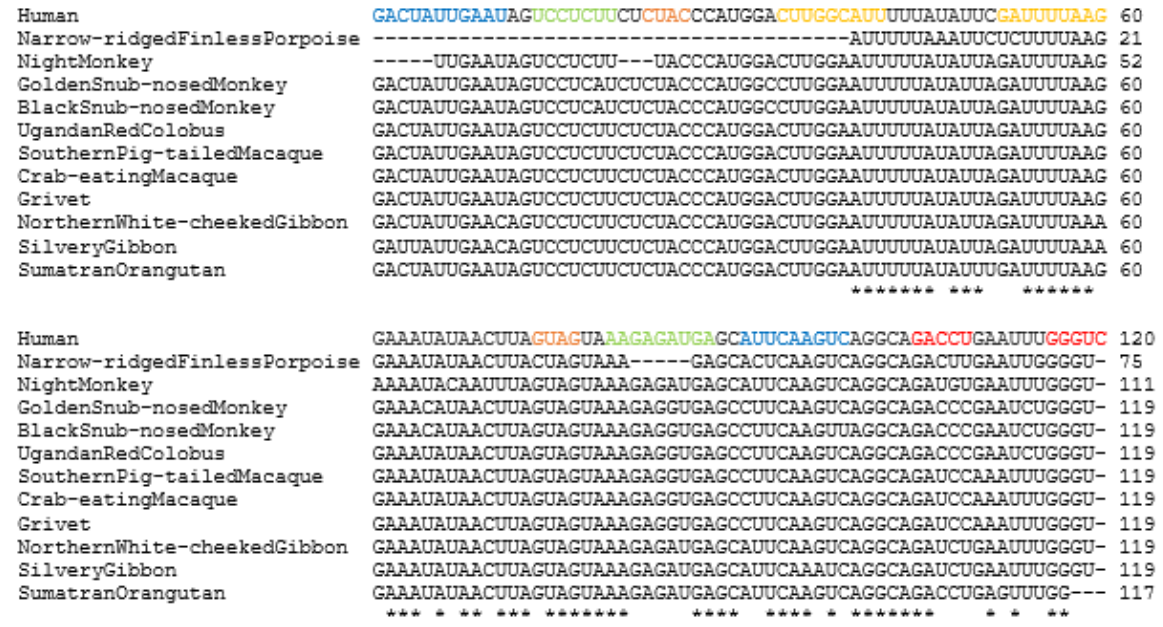

Figure 22. Multiple alignment of the blast search of only for loc3 (1040-1160, figure 2) sequences used are shown in table 2 under loc3. The coloured sections represent the sequences of interest.

As is visible in the multiple alignment of loc3 (figure 22)., the porpoise is missing almost the complete first half of the sequence. The only sections where both sides of the stem are present are the sections marked red which could still form the stem. The sections marked blue is mostly unchanged in the in the sequences the night monkey sequence is missing the first five bases. Another difference is present in both gibbon sequences. In these sequences the last base is a C instead of a U which would result in a mismatched pair. The second part of the blue section shows a difference. Here the first base is a C for some sequences instead of an A that humans have. This difference also causes mismatched pairs. And lastly the silvery gibbon has a difference that does not result in a mismatch. In the second section the third to last base is an A, which can still make a stable pair because the third base for the first section has changed into a U. For the sections marked green there are only two differences. The first is a U and A difference in the second to last base of the first section for golden snub-nosed monkey and black snub-nosed monkey these differences cause mismatches. The second difference is a difference in the fourth to last base in the second section for both snub-nosed monkeys, both macaques, the colobus and grivet. This difference did not cause a mismatch seeing that it is an A-U pair and G-U pair difference. The orange sections have a single difference in the night monkey where the first base is missing. In the sections marked yellow there are two differences. For the first difference humans are the outlier. For the fourth from last base for the first section is a C only in the

human sequence for the other sequences except the porpoise this base is an A. in humans this base causes a mismatch however for the other sequences it is a stable base pair. The second difference is the last base of the second section for both gibbon sequences is an A instead of a G, thus causing a mismatched pair. Lastly the sections marked red in these sections the last base is missing for all sequences besides humans. The orangutan is missing the last three. There are also multiple differences in the last three bases of the first section. Only the orangutan has the same three bases as humans all other sequences differ. Night monkeys have UGU, both snub-nosed monkeys and colobus have CCC, both macaques and the grivet have UCC and the gibbons have UCU.

Discussion

PSMB8-AS1 role in influenza replication (More et al, 2019) and cancer development (Guilletti et al, 2018) highlight the importance of lncRNAs in diseases and biological processes. The mechanisms that allow for the specific functions a lncRNA can have differ between lncRNAs. Some lncRNAs have shown to be functional because of their higher order structure, while for other lncRNAs the act of transcription allowed for their specific function (Engreitz et al, 2016). In this study we sought to find whether PSMB8-AS1's structure was of importance to its specific function. Using RNA structure prediction tools, we identified regions in the PSMB8-AS1 sequence that showed a high potential of RNA structure conservation in orthologues. Structural models were made based on the predicted region and re-evaluated to evaluate if the structure conservation was significant.

Using RNAz, two full sequence predictions were done to find regions in the PSMB8-AS1 sequence that showed a possibility of conservation. The first RNAz prediction that was done, was human together with the primate orthologues. The second was human with horse and cattle orthologues. What was remarkable was that RNAz only predicted conserved structures in the third exon, in the first and the second exon no conservation was predicted. There are two possibilities for this: the first possibility is that the first and second exon do not have any conserved structures, the second possibility is that the exons were too short and the window, with which RNAz sought, was too large. Both exon one and exon two are about 200 nt long, while exon three is 1083 nt long. It is less likely to find a 120 nt conserved structure in 200 nt than in 1000 nt.

Next, we looked at the predicted conserved structure models. The model predicted for region 800 – 920 using RNAz with PSMB8-AS1 and the orthologues found in cattle and horses (figure 4) showed a high possibility of conservation within primate orthologues. The model structure was successfully computed in the chimpanzee and common marmoset orthologues using Mfold. The structures computed with Mfold for chimpanzee (figure 5) and common marmoset (figure 6) were different from each other. The structure computed with the marmoset sequence had a double hammerhead-like structure, while the structure computed using the chimpanzee sequence had a structure more similar to the structure predicted in the RNAz. The chimpanzee structure only showed one hammerhead-like structure and a hairpin structure with interior loops. Both the structure shown in the chimpanzee as well as in the common marmoset sequence were possible structures. Therefore, we analysed the multiple alignment of human PSMB8-AS1 with the primate and distant relative orthologues to determine which of the computed structures were more likely.

When looking at the multiple alignment (figure 7), it shows that the double hammerhead-like structure is the more likely structure. There are multiple cases where both sides of a base pair are mutated and remain complementary. Furthermore, a gap in the rhesus and gelada orthologue sequences make the interior loop structure not possible whereas the hammerhead like structure is still possible.

A multiple alignment analysis was also done on the human PSMB8-AS1 sequence region 800-920 nt together with the distant relatives 2 dataset orthologues (figure 8). The distant relatives 2 dataset showed more differences in the sequence to the human sequence as opposed the primate orthologues, which was as expected seeing that primates were more closely related to humans than the other mammals. More differences in the sequence would also result in more differences in the pairs present in the structure. However, the first hammerhead-like structure has several differences comparing to the human sequence. The structure can still be formed in all orthologues. The second hammerhead-like structure also has differences comparing to the human sequence. In the first hairpin of the second hammerhead-like structure of yaks, an CG pair mutated into an AU pair showing a possible evolutionary pressure to keep this structure. Cats, however, seem to be fully missing this hairpin. The second hairpin of the second hammerhead-like structure is completely missing in pigs and can only be stably formed in yaks. This shows that maybe the hammerhead like structure is not the correct structure but something more like a structure with a single hairpin instead of two.

For the region for 520 – 800 nt predicted in RNAz using PSMB8-AS1 and primate dataset we had to combine multiple of the predicted RNAz results using Mfold. We analysed this structure in using a

multiple alignment of human PSMB8-AS1 and cattle, horse and the distant relatives 2 dataset orthologues (figure 10). The multiple alignment showed more mismatches in the orthologues, contrary to the predicted structure for the 800 – 920 nt region (figure 8). The region 800 – 920 nt showed mutations where the pairs and stems remain stable or even got more stable. Whereas the structure computed for region 520 – 800 nt showed more mutations that resulted into mismatched pairs. This could mean that this region does not show any form of structural conservation. Another possibility is that the computed structure model is not the correct structure.
We also analysed the smaller structures that were predicted using RNAz, region 520-640 (figure 11 and figure 13) and region 680-800 (figure 12 and figure 14). These structures were analysed to see if maybe the conserved structures were lost when computing the larger 520-800 region structure. When looking at the multiple alignment of both structure we can see the all the orthologues had multiple differences with the human sequence in the marked regions. The orthologue sequences show in each marked region differences with the human sequence. These differences caused a loss of stability in the stems when looking at the orthologues. There are some differences in the sequences where pairs remain complementary however these differences were less common than the cases where mismatches occurred. The differences in the sequence for these structures do not show any evolutionary pressure to be conserved.
The region predicted on nucleotides 1040 -1160 nt for the PSMB8-AS1 with primate orthologues was not present in false killer whales. The other orthologues showed multiple differences from the human sequence. Some of these differences resulted into mismatches. Pigs, wild yaks and cattle are missing a section that is of importance for the predicted structure. There are no cases of both nucleotides mutating to keep a certain pair. This likely means that this region is not a conserved structure and that this particular structure is not of importance.

In the BLAST search results we can see that in all regions there is quite a lot of conservation between sequences most differences are only single mutations found in one sequence or sequences of close relatives. It was expected that most of the loc2 primate (figure 18, 19) multiple alignment would show a similarity. When looking at loc1 (figure 17) and loc3 (figure 22) this can seem like this is because most of the sequences are sequences found in primates, however when looking at loc2 non-primates (figure 20, 21) we can see that even in a dataset consisting of humans and only non-primates, even though less than with primates, a significant part of the sequence is unchanged.
Loc1 and loc3 showed similar results, most of the differences in sequence we single point mutations and there were no real significant changes that showed evidence for the conservation of a structure or sequence. Both regions showed similar sequences in different organisms.
However, when looking at loc2 there are some instances that would suggest possible conservation of structure. In the multiple alignment results for loc2 there are cases where multiple sequences share the same change, most of these changes consists of G and A changes and C and U changes. Where either G is swapped for an A or A is swapped for a G, and the same with C and U. Both G and A can stably bind with U, C and U can bind stably with G. The suggestion that these kinds of changes are more likely to remain in the sequences in different organisms show that there is an evolutionary pressure to keep a pair bonded. Which would mean that there is a structural importance to this region. Some of the instances that showed G to A or U to C changes were shown in the model (figure 9) as bonded with C for G to A and bonded with A for U to C. These cases show that the model is most likely not correct either because parts of the sequence are missing or a wrong pairing in sections.

When looking at the regions we analysed using the primates, distant relatives 1 and 2 datasets only one showed a possibility of having a form of conservation. The region from 800 to 920 showed a high possibility of conservation in primates. In the distant relatives 2 sequences this region also showed some conservation however this could also be a result of the random mutations. The other predicted structures did not show any significant conservation in the orthologue sequences. The differences that showed an increase of stability or a conservation of base pairings in specific locations seemed to be more of a random occurrence than as a form of conservation. However, when looking at the BLAST

search results, we can see that loc2 showed a high possibility of conservation. First, loc2 was present in many different organisms both primates and non-primates. Second, loc2 showed signs of having base pairings to be conserved in specific locations by keeping certain mutations that would allow for the pairings. Other mutations that would inhibit these pairings were only present in very close relatives or a single organism. The significance of this specific region for the whole functionality of PSMB8-AS1 is still not sure. Only the second exon of PSMB8-AS1 sequence has shown signs of conservation. In the other exons no regions that showed conservation were found.

References

(1) Higgins D.G., Sharp P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 73, Dublin, pp 237-244

(2) Sievers F., Higgins D.G. (2014) Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. In: Russell D. (eds) Multiple Sequence Alignment Methods.
Methods in Molecular Biology (Methods and Protocols), vol 1079. Humana Press, Totowa, NJ

(3) Hofacker I.L., Fontana W., Stadler P.F., Bonhoeffer L.S., Tacker L., Schuster P. (1994) Fast folding and comparison of RNA secondary structures. , Volume 125, Issue 2, pp 167-188

(4) Lorenz R., Bernhart S.H., Höner zu Siederdissen C., Tafer H., Flamm C., Stadler P.F., Hofacker I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* 6, 26

(5) Gruber A.R., Findeiß S., Washietl S., Hofacker I.L., Stadler P.F. (2010) RNAz 2.0: Improved noncoding RNA detection. Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany.

(6) Hangauer M. J., Vaughn I. W., McManus M. T., Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic
Noncoding RNAs, PLoS Genet., vol. 9, no. 6, 2013

(7) Dhanoa J.K., Sethi R.S., Verma R., Arora J.S., Mukhopadhyay C.S. Long non-coding RNA:
its evolutionary relics and biological implications in mammals: a review. *J. Anim. Sci. Technol.* 2018; 60:25.

(8) Bryzghalov O., Wojciech Szcześniak M., Makałowska I. SyntDB: defining orthologues of human long noncoding RNAs across primates, *Nucleic Acids Research*, 2019

(9) Holley R.W., Apgar J., Everett G.A., Madison J.T., Marquisee M., Merril S.H., Penswick J.R., Samir A. (1965). Structure of ribonucleic acid. *Science*. 147 (3664): 1462–5.

(10) More S., Zhu Z., Lin K., Huang C., Pushparaj S., Liang Y., Sathiaseelan R., Yang X., Liu L.,
Long non-coding RNA PSMB8-AS1 regulates influenza virus replication. RNA biology, Volume 16, 2019, pp 340-353

(11) Kuchin S. (2011). "Covering All the Bases in Genetics: Simple Shorthands and Diagrams for Teaching Base Pairing to Biology Undergraduates". Journal of Microbiology & Biology Education.

(12) Johnsson P., Lipovich L., Grandér D., Morris K.V., Evolutionary conservation of long noncoding RNAs; sequence, structure, function. Biochim Biophys Acta. 2014, 1840(3):106371

(13) Quinn J.J., Chang H.Y., Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016 jan;17(1):47-62

(14) Novikova I.V., Hennelly S.P., Sanbonmatsu K.Y., Sizing up long non-coding RNAs. Bioarchitecture. 2012 nov;2(6):189-199

(15) Zuker M., Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research. 2003 Jul ;31(13):3406-15

(16) Faghihi M.A., Modarresi F., Khalil A.M., Wood D.E., Sahagan B.G., Morgan T.E., Finch C.E., St Laurent G. 3rd, Kenny P.J., Wahlestedt C., Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase expression. Nature medicine. 2008 Jul;14(7):723-30.

(17) Reece J.B., Urry L.A., Cain M.L., Wasserman S.A., Minorsky P.V., Campbell Biology (tenth edition). 2015, pp 345-347

(18) Engreitz J., Haines J., Perez E., Munson G., Chen J., Kane M., McDonal P.E., Guttman M., Lander E.S., Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539,** 452–455 (2016)

(19) Giulietti M., Righetti A., Principato G., Piva F., LncRNA co-expression network analysis reveals novel biomarkers for pancreatic cancer. Carcinogenesis. 2018 Jul 30;39(8):10161025

(20) Lehner B., Williams G., Campbell R.D., Sanderson C.M., Antisense transcripts in the human genome. Trends in Genetics. 2002 Feb;18(2):63-5

(21)Katayama S., Tomaru Y., Kasukawa T., Waki K., Nakanishi M., Nakamura M., Nishida H., Yap C.C., Suzuki M., Kawai J., Suzuki H., Carninci P., Hayashizaki Y., Wells C., Frith M., Ravasi T., Pang K.C., Hallinan J., Mattick J., Hume D.A., Lipovich L., Batalov S., Engström P.G., Mizuno Y., Faghihi M.A., Sandelin A., Chalk A.M., Mottagui-Tabar S., Liang Z., Lenhard B., Wahlestedt C.; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium., Antisense transcription in the mammalian transcriptome. Science. 2005 Sep 2;309(5740):1564-6

(22)Kaikkonen M.U., Lam M.T., Glass C.K., Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Research. 2011 Jun 1;90(3):430-40

(23)Zhu Y., Xie Z., Li Y., Zhu M., Chen Y.P., Research on folding diversity in statistical learning methods for RNA secondary structure prediction. *Int J Biol Sci*. 2018;14(8):872–882.

(24)Sasse A., Laverty K.U., Hughes T.R., Morris Q.D., Motif models for RNA-binding proteins. Current Opinion in Structural Biology. 2018 Dec;53:115-123

(25)Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D., Higgins D.G., Fast, scalable generation of highquality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011 Oct 11;7:539