2015

Evolution of Eukaryotic Ribosomes at Atomic Resolution

Agni Kumar

Abstract

The goal of this research is to develop and test a model of ribosomal expansion, in which given sites of eukaryotic expansion have highly conserved structure even if the sequences are widely divergent. I have focused on rRNA expansion segments, which have been added over evolution to the conserved core through elongation and insertion into pre-existing helices, working toward developing a method to predict and calculate the secondary structures of various eukaryotic ribosomes. While deducing detailed rRNA evolutionary patterns based on distinctive insertion fingerprint evidence, I attempted to extrapolate backwards through time and gain an understanding about how the ribosome has evolved into what it is today.

Through both computational and experimental methods, rRNA fragments from the most variable part of the ribosome – Helix 25 Expansion Segment 7 – were successfully isolated for multiple complex eukaryotes. I analyzed the secondary structures of these rRNA fragments through SHAPE, dideoxy sequencing, and capillary electrophoresis. It was concluded that by carefully parsing the phylogenetic tree and experimentally determining a small subset of secondary structures, one can interpolate and obtain credible computational estimates of large numbers for which only the rRNA sequences are known. As the NASA Astrobiology Institute has calculated (through crystallization and decomposition) only six ribosomal secondary structures over the past six years, this research may serve as a foundation for the development of more cost-effective prediction methods. The results have also revealed new targets for drug design, as rRNA is widely used as the target of several clinically relevant antibiotics.

Introduction

The ribosome was largely formed at the time of the last universal common ancestor, LUCA, around 3–4 billion years ago and is common to all living species. Ribosomes are cytoplasmic particles containing ribosomal RNA (rRNA) and ribosomal proteins. Ribosomes are processed and assembled in the nucleolus of a eukaryotic cell and form two subunits, the large ribosomal subunit (LSU) and the small ribosomal subunit (SSU). The LSU rRNA acts as a ribozyme, catalyzing peptide bond formation, while the SSU decodes the mRNA.[1]

The process of *translation* is responsible for synthesizing all coded protein in living systems. Functioning as the final piece of the central dogma to molecular biology, it converts a given messenger RNA (mRNA) sequence passing through the ribosome into a string of amino acids that form proteins. The understanding of the translation system and the function of rRNA within various living organisms has increased over the last decade through advancements in sequence databases and three-dimensional structure determination by X-ray diffraction.

The unique characteristics of rRNA are important in medicine and evolution. Over the past few decades, rRNA has been widely used as the target of several clinically relevant antibiotics such as streptomycin, chloramphenicol, erythromycin, etc. As rRNA is one of only a few gene products present in all cells, nucleotides that encode the rRNA can be sequenced to identify an organism's taxonomic group, refer to related groups, and estimate rates of species divergence. As a result, many thousands of rRNA sequences are known and stored in specialized databases.[2]

In molecular biology, rRNA sequences are widely used for working out evolutionary relationships among organisms, since they are ancient in origin and are found in all known forms of life. However, while ribosomes are universal particles, they are also variable – ribosomes increase in size and vary in different patterns as species become more complex.[3] As shown in

Fig. 1 below, the changes in size of the rRNAs are observed in the large subunit, with bacteria having the smallest LSU, followed by archaea and then eukarya, which is variable. Any given species on the *phylogenetic tree of life* has evolved from common ancestors, beginning with LUCA and branching out into the kingdoms of bacteria, archaea, and eukarya. It can be observed that ribosomes increase in complexity and size in eukaryotes, especially in metazoans. *H. sapiens* have the largest known ribosomes.[4]

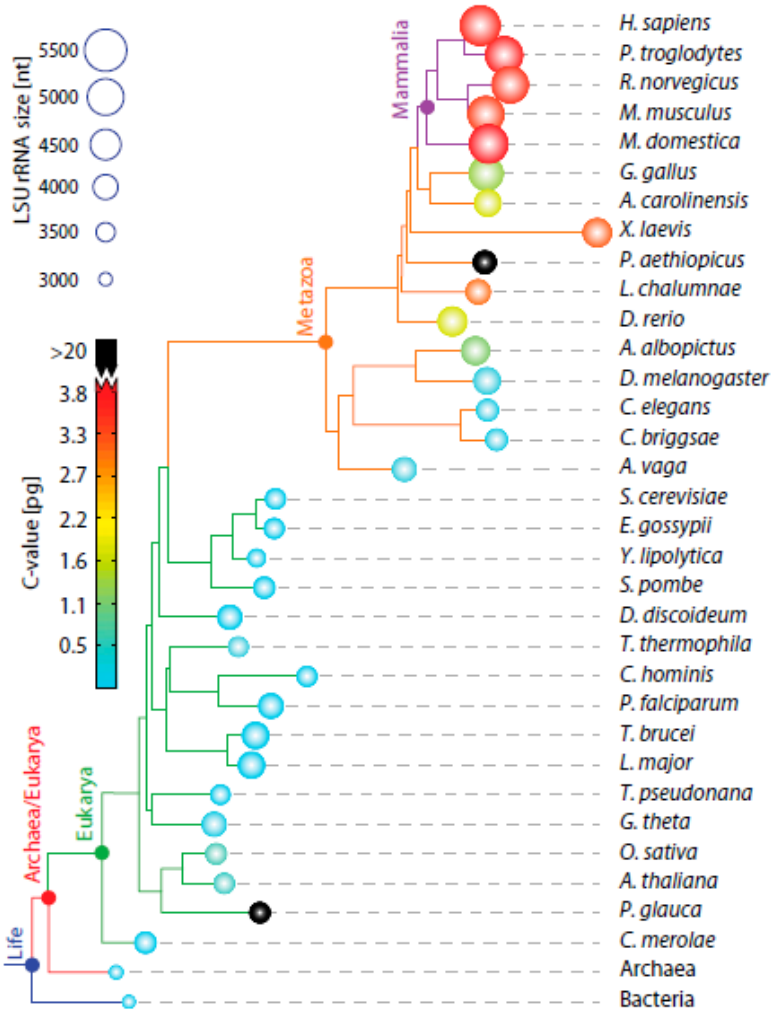


Figure 1. The phylogenetic tree of life showing the size evolution of the ribosome through 25 eukaryotes and their relationship to average archaea and bacteria. Circle radii are proportional to total length of the large subunit of a given ribosome, which is measured in nucleotides.[4]

It has been found that there is a *common core* of the ribosome that is preserved throughout the phylogenetic tree, and that rRNA expansion segments have been added to rRNAs without

perturbing the preexisting core, leaving behind "insertion fingerprints" that are evident at atomic resolution, as shown in **Fig. 2** below.

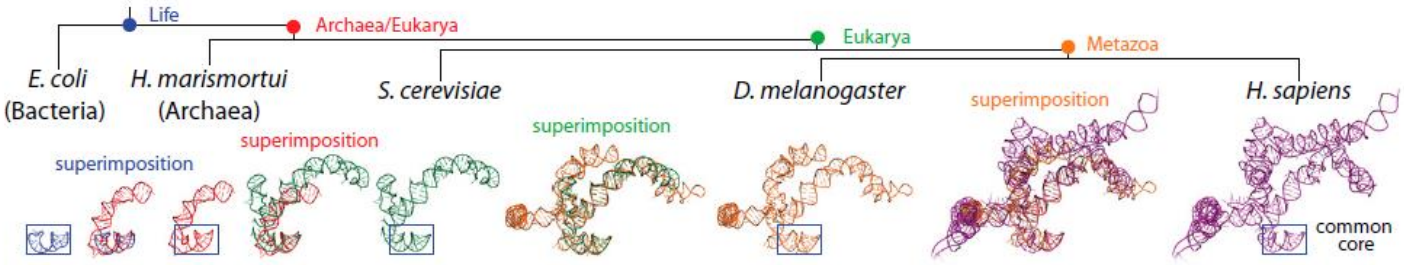


Figure 2. In a recent study, scientists compared the three-dimensional structures of ribosomes from a variety of species on the phylogenetic tree, showing the serial accretion of rRNA onto a common core. The simplest form of this core was present in bacteria (LUCA), and segments were added to this core throughout the domain eukarya.[4]

Processes of rRNA expansion can be further observed by comparing the secondary and three-dimensional rRNA structures of various eukaryotes to each other, as illustrated in **Fig. 3**. S*econdary structures* are representations that reduce the dimensionality from three to two, assuming that a given base is in one of two discrete states: paired or non-paired.[5] A paired base is involved in secondary interactions. Following Levitt[6], helices are defined as three consecutive base-paired nucleotides bounded by non-paired nucleotides. Secondary structures contain stems (helices), loops, and pseudoknots.

X-ray crystallography and cryoelectron microscopy have provided three-dimensional structures of multiple ribosomes from all three domains of life.[4] While such structures are diverse and seemingly complex, they are actually composed of repeating, easily recognizable molecular building blocks that correlate well with RNA sequence, unlike ribosomal secondary structures. Currently, the RiboEvo[7] and RiboVision[8] database, partnered with the NASA Astrobiology Institute, contains only six rRNA secondary structures – those of *E. coli* (bacteria), *H. marismortui* (archea), *S. cerevisiae* (yeast), *D. melanogaster* (fruit fly), *T. thermophila* (protozoan), and *H. sapiens* (human) – that were obtained through an extensive, expensive process of ribosomal crystallization followed by decomposition of the structure from three

dimensions to two dimensions. In **Fig. 3** below, the *E. coli* secondary structure and two of the four known eukaryotic secondary structures are placed beside their corresponding three-dimensional structures. rRNA expansion in the eukaryotic rRNAs is evident.[4]

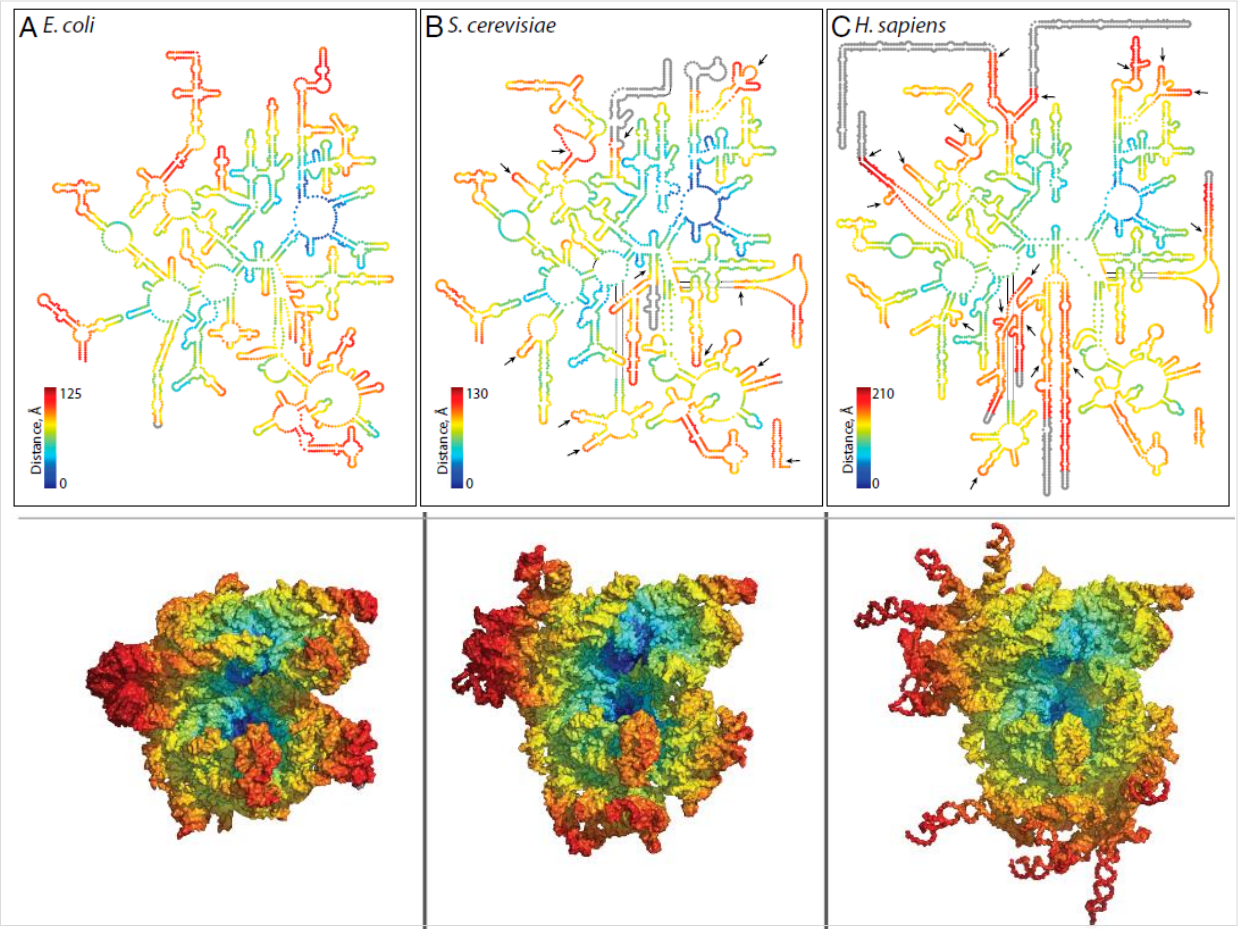


Figure 3. LSU rRNA secondary structures of (A) *E. coli*, (B) *S. cerevisiae*, and (C) *H. sapiens*. The color indicates the three dimensional proximity to the origin (blue is close to the origin, red is remote from the origin). The sites of rRNA expansion from *E. coli* to *S. cerevisiae* and from *S. cerevisiae* to *H. sapiens* are marked by arrows. For clarity, proteins are omitted from the structures.[4]

The main objective of this research project is to prove that there exists an expansion-on-common-core evolutionary relationship between all eukaryotic rRNA secondary structures. Through this project, I also worked toward developing a novel method to predict and calculate the secondary structures of the remaining eukaryotes on the phylogenetic tree, using the four known secondary structures to deduce detailed rRNA evolutionary patterns based on distinctive insertion fingerprint evidence. With these long-buried fingerprints, I attempted to work

backwards through time and gain knowledge about how the ribosome has evolved into what it is today. It was hypothesized that the secondary structure is more conserved than rRNA sequence throughout the phylogenetic tree, and that rRNA expansion sequences have been added through elongation and insertion into pre-existing helices. Using both computational and experimental methods, evidence of striking relationships was shown between the secondary structures of various eukaryotic organisms located close together on the phylogenetic tree. For the experiments, the most variable part of the ribosome – Helix 25 Expansion Segment (ES) 7 – was successfully isolated for multiple complex eukaryotes, on which testing was done. It was reasoned that if the evolution of *this* part of the ribosome showed serial accretion of rRNA onto a conserved core, that the other less-variable parts of the ribosome would behave in a similar manner was, for all intents and purposes, a given.

Materials and Methods

The methodology of this project lies in two main parts.

Part 1: Computational Analysis. The computational analysis part of this project served to use known secondary structures, aligned sequences, and folding algorithms to predict secondary structures of uncharacterized expansion elements (specifically Helix 25/ES 7) of various eukaryotic ribosomes. The Mfold[9] program was used to predict the secondary structures and to estimate their stabilities. Fitting parameters (level for prediction, scoring model, refinement, and positive weight for the true base pairs at each level) were varied to maximize the fit of the predicted secondary structure to homologous known secondary structures. For this project, the parameters of Level 1 (pseudoknot free), McCaskill model or CONTRAfold model, with refinement, and default positive weight levels were used. Predicted secondary structures from Mfold[9] were compared and analyzed to the four known secondary structures obtained from

RiboEvo[8]. Similarities in secondary structure were characterized by overall topology, and by numbers and dispositions of helices and loops.

Part 2: SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) and dideoxy sequencing. Among the results gathered from the Part 1 investigation, it was found that the predicted secondary structure of *S. cerevisiae* (yeast) showed strong resemblance to its known secondary structure, while the predicted secondary structure of a more complex eukaryote, *D. melanogaster*, did not. In order to better the secondary structure predictions of the complex eukaryotes, SHAPE experiments were conducted on the ribosomes of two closely related species, *D. melanogaster* (fruit fly) and *A. albopictus* (mosquito), with a hypothesis that their secondary structures – even in the most variable part of the ribosome – should be almost identical despite their very different rRNA sequences (with *D. melanogaster* being rich in A (adenine) and U (uracil) nitrogenous bases, and *A. albopictus* being rich in G (guanine) and C (cytosine) nitrogenous bases).

SHAPE is a nucleic acid structure probing technique that takes advantage of reagents that are able to modify the backbone of RNA in structurally flexible regions. Reagents such as N-methylisatoic-anhydride (NMIA) undergo hydrolysis to form adducts on the 2'-hydroxyl of the RNA backbone. Adduct formation is quantified for each nucleotide in a given RNA by extension of a complementary DNA primer with reverse transcriptase and comparison of the resulting fragments with those from an unmodified control. The results of the SHAPE experiments did not actually give the means to exactly calculate, quantitatively, the two secondary structures; rather, a series of graphs was provided qualitatively reporting on a portion of the rRNA secondary structure.[10]

The initial stage of the Part 2 investigation was to isolate the Helix 25/ES 7 (H25/ES7)

fragments needed to begin the SHAPE experiments. To begin, DNA genes in plasmids were designed and ordered to express the *D. melanogaster* and *A. albopictus* H25/ES7 rRNAs (**Fig. 4**). Plasmids were linearized with HindIII and transcribed the RNAs using T7 RNA polymerase. The RNAs were then purified and checked on a gel.

Original gene, Extra base pairs extending helix, T7 promoter, EcoR1, Stability Bases, Weeks, HindIII

Drosophila melanogaster

GTGGGAATTCTAATACGACTCACTATAGGGGCGCCCGATGAACCTGAATATCCGTTATGGAAAATTCATCA
TTAAAATTGTAATATTTAAATAATATTATGAGAATAGTGTGCATTTTTTCCATATAAGGACATTGTAATCTA
TTAGCATATACCAAATTTATCATAAAATATAACTTATAGTTTATTCCAATTAAATTGCTTGCATTTTAACAC
AGAATAAATGTTATTAATTTGATAAAGTGCTGATAGATTTATATGATTACAGTGCGTTAATTTTTCGGAATT
ATATAATGGCATAATTATCATTGATTTTTGTGTTTATTATATGCACTTGTATGATTAACAATGCGAAAGATT
CAGGATACCTTCGGGCGCTCGATCCGGTTCGCCGGATCCAAATCGGGCTTCGGTCCGGTTCAAGCTTGGTG

Aedes albopictus

GTGGGAATTCTAATACGACTCACTATAGGGGCGCCCGTTGAACTCGATTATCCGAGCGGAGACATTCACCT
GCGGTTGGCCGGCGACGGCACGGCCGCAGGGCACTTGTCTCTCGCACTAGCCAAGAGGACACTGCGATCC
ATTACGAAACAGCTTTCGCGCAAGCGCAAGGTCGCCCGACAACTGCCCCCTGGTGCTGGTTGCTTGCCCCA
CAGTAGCGACGCTCAGTTCTGAAGGCCTGTGCCGCGAGGTGGGGCTTACTGCACGTGGTGTTCTAGCAGTC
GGGCGCGTGATGGATTCCCCCGGACACCGGGTGGTCTTCCCGTAAGGGGCCACCGGACTGTCGATCGGCAG
TGAAAGAATCGAGGTACCTTCGGGCGCTCGATCCGGTTCGCCGGATCCAAATCGGGCTTCGGTCCGGTTCA
AGCTTGGTG

Figure 4. Gene designs for *D. melanogaster* and *A. albopictus* Helix 25/ES 7 fragment.

To begin the SHAPE experiment[11], each RNA sample was diluted in water (~17 pmol). A folding buffer[11] was then added to renature the RNA to a biologically relevant conformation. See **Fig. 5** below for the RNA "recipes".

D. melanogaster

Need 32 uL of 68.75 ng/uL RNA in flat-lid 0.2 mL PCR tubes for each sample.

# of samples	Total Vol req'd (uL)	Total RNA req'd (ng)	[RNA] stock (ng/uL)	Vol of stock needed (uL)
2	66	4537.5	2200	2.0625
RNA recipe				
RNA stock (uL	2.0625			
10x folding	8.25			
dH2O (uL)	55.6875			
Total	66			

10x folding buffer: 2 M NaOAc (sodium acetate), 500 mM NaHEPES (folding buffer) pH 8

A. albopictus

Need 32 uL of 69.5 ng/uL RNA in flat-lid 0.2 mL PCR tubes for each sample.

# of samples	Total Vol req'd (uL)	Total RNA req'd (ng)	[RNA] stock (ng/uL)	Vol of stock needed (uL)
2	66	4587	1600	2.866875
RNA recipe				
RNA stock (uL	2.866875			
10x TE (uL)	8.25			
dH2O (uL)	54.883125			
Total	66			

Figure 5. *D. melanogaster* and *A. albopictus* stock solution RNAs had to be prepared in water before they could be folded.

All samples were heated and cooled in a thermocycler to allow the RNA to achieve an appropriate conformation prior to NMIA modification. 32 uL of RNA solution were added to 4 uL of 10x additive (10 mM DCTA, samples 1&3; 100 mM $MgCl_2$, samples 2&4) in a 0.2 mL PCR tube, and the contents of the tubes were mixed well by pipetting. The samples were incubated at 85˚C for 30 sec then cooled linearly (~1.5˚C/min) to 30˚C in a thermocycler. The RNAs were then modified with NMIA[11]. First, the NMIA solution was dissolved in DMSO (dimethyl sulfoxide) by adding 65 mM NMIA (0.057g NMIA) to 5000 uL of DMSO. The RNA was then divided into (+) and (-) samples, each with 18 uL of the folded RNA solution in 0.2 mL PCR tubes. 2 uL NMIA solution were added to the (+) samples, and 2 uL DMSO was added to the (-) samples (**Fig. 6**).

Sample #	Condition	RNA
1+	DCTA	Dm
1-	DCTA	Dm
2+	Mg	Dm
2-	Mg	Dm
3+	DCTA	Aa
3-	DCTA	Aa
4+	Mg	Aa
4-	Mg	Aa

Figure 6. Breakdown of folding solutions added to each of the 8 (+) and (-) samples.

The samples were then incubated (for 1 hour at 37˚C in a thermocycler with heated lid) until essentially all of the NMIA had either reacted with the RNA or had degraded due to hydrolysis with water (**Fig. 7**).

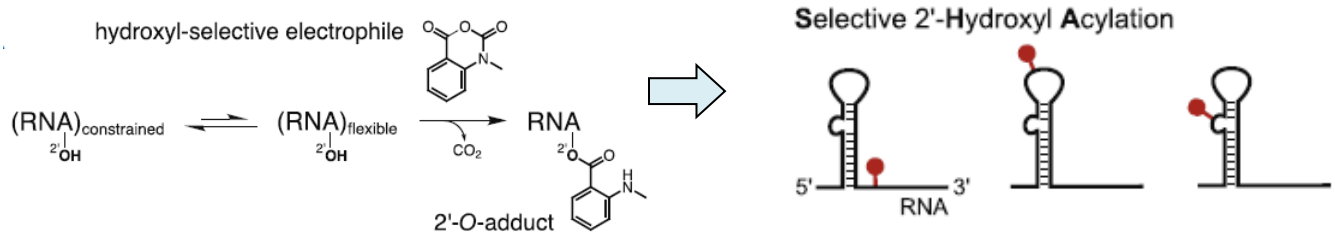


Figure 7. NMIA reacts with the 2'hydroxyl group in flexible nucleotides to form a stable 2'-*O*-adduct. This reaction occurs more readily in single-stranded regions (e.g. those in loops or bulges) and not in double-stranded ones, which eventually provided many clues about the two eukaryotes' secondary structures.[12]

After the NMIA-modification process, the RNA was precipitated with ethanol[11] and purified using the *Zymo Research RNA Clean & Concentrator-25 kit*[13]. Next, the RNA was suspended in 27 uL of 1xTE buffer (10 mM Tris + 1 mM EDTA) at 65˚C. The TE was left to sit on column for about 5 minutes before spinning to elute.

In this experiment, NMIA-modified positions were detected by primer extension[11] using target-specific primers and the process of reverse transcription, a technique commonly used to detect RNA expression levels (**Fig. 8**).

Figure 8. Since the reverse transcriptase enzyme cannot pass by 2-*O*-modified sites in RNA, the lengths of the resulting cDNA pieces correspond to the distance between the primer binding and the sites of the SHAPE modifications.[12]

The reverse transcription was conducted as per the protocol provided in the *Thermo Maxima H minus reverse transcription kit*[14]. 20 uL of purified RNA was added to 15 uL of 0.8 pmol/uL (12 pmol) of FAM-labeled reverse transcription primer (florescent-labeled primer) to each sample, then mixed by pipetting. To anneal the primers, the samples were heated for 1 minute at 95˚C in a thermocycler with heated lid, then cooled to 65˚C for 3 minutes. Next, 14 uL reverse transcriptase (RT) mix were added to each sample, mixed by pipetting (**Fig. 9**).

RT mix recipe		
# of aliquots req'd		
8		
	Vol/aliquot (uL)	Total Volume (uL)
5x Maxima RT buffer	10	90
DTT	0	0
dNTPs (40 mM)	3.125	28.125
nuc-free dH2O	0.875	7.875
Total	14	126

Figure 9. "Recipe" for reverse transcriptase mix added to each of the 8 samples.

After that, 1 uL (200 U) Maxima H Minus reverse transcriptase enzyme mix was added to each sample, mixed by pipetting. The samples were then heated for 2 hours at 55˚C in a thermocycler, then to 85˚C for 5 minutes to inactivate the reverse transcriptase. The samples

were then frozen in preparation for capillary electrophoresis, in which material elutes from the capillary array over time and passes through a fluorescence detector.

After the SHAPE experiment, dideoxy sequencing[11] of the expansion segments was conducted. Sequencing data is generated in order to confirm that the primer anneals and is able to be extended along its target RNA, and is used as a reference in order to inform where exactly the SHAPE modifications are occurring. First, ~3 pmol RNA was prepared in 1xTE buffer; it was calculated that 20 uL RNA needed to be placed in each of five 0.2 mL PCR tubes for both *D. melanogaster* and *A. albopictus* (**Fig. 10**).

D. melanogaster

# of samples	Total Vol req'd (uL)	Total RNA req'd (ng)	[RNA] stock (ng/uL)	Vol of stock needed (uL)
5	110	2120.57175	2200	0.96389625
RNA recipe				
RNA stock (uL)	0.96389625		$MW_{RNA\ (g/mol)}$	128,519.50
10x TE (uL)	11		Mass of 3 pmol (g)	3.85559E-07
dH2O (uL)	98.03610375		Mass of 3 pmol (ng)	385.5585
Total	110		$[RNA]_{desired}$ (ng/uL)	19.277925

A. albopictus

# of samples	Total Vol req'd (uL)	Total RNA req'd (ng)	[RNA] stock (ng/uL)	Vol of stock needed (uL)
5	110	2157.5895	1600	1.348493438
RNA recipe				
RNA stock (uL)	1.348493438		$MW_{RNA\ (g/mol)}$	130,763.00
10x TE (uL)	11		Mass of 3 pmol (g)	3.92289E-07
dH2O (uL)	97.65150656		Mass of 3 pmol (ng)	392.289
Total	110		$[RNA]_{desired}$ (ng/uL)	19.61445

Figure 10. RNA "recipes" for *D. melanogaster* and *A. albopictus* in preparation for the sequencing experiment

Using the same procedure as conducted in the SHAPE experiment, and after annealing the primers in a similar fashion, RT components (dNTPs, ddNTPs, reverse transcription buffer, and reverse transcriptase) were added in the same quantities as SHAPE to each reaction.

During the setup for capillary electrophoresis (CE)[11], a 96-well plate (**Fig. 11**) was prepared with a mixture of 0.3 uL ROX-labeled DNA ladder and 8.7 uL of HiDi-formamide in each well. 9 uL of the ROX-HiDi mix was put into each well. Next, 1 uL of each sample (or dilution) was added to each well of the plate. After the samples were mixed, the plate was sealed

with adhesive film and spun in a plate centrifuge from 1 minute at 900 rpm. Under the guidance of a graduate student, hybrid duplexes were denatured by heating the plate to 95˚C for 5 minutes in a thermocycler with heated lid, then allowed to cool to ~30˚C before removing and spinning again. After removing the adhesive seal, a 96-well septum was placed over the plate and loaded it into a CE cassette inserted into a tray. After a new protocol was created in the CE control software and linked to the plate, the capillary electrophoresis process began (~45 min/4 samples). Data collected was analyzed in MATLAB[15].

	1	2	3	4	5	6	7	8	9	10	11	12
A	1+	Dm ddA	Aa ddT									
B	1-	Dm ddC	Aa (-)									
C	2+	Dm ddG										
D	2-	Dm ddT										
E	3+	Dm (-)										
F	3-	Aa ddA										
G	4+	Aa ddC										
H	4-	Aa ddG										

Figure 11. 96-well plate prepared for capillary electrophoresis (18 samples)

Results and Discussion

Results from Part 1: As a positive control, the known secondary structure of the ES7 from *S. cerevisiae* was successfully replicated by the folding algorithm. Selected H25/ES7 secondary structures, predicted by the algorithm, are shown in **Fig. 12 and 13** below.

Similarities in secondary structure were characterized by overall topology, and by numbers and dispositions of helices and loops. The calculated secondary structures of *Y. lipolytica* and *E. gossypii* were highly similar to the known, experimentally determined *S. cerevisiae* structure. The calculated secondary structures of all three species had similar topologies, helices, and loops. Differences were at the level of single nucleotide bulges. These similarities suggest that one can computationally fold ES7 from *Y. lipolytica* and *E. gossypii*, two

species located close to *S. cerevisiae* on the phylogenetic tree. To my knowledge this is the first proposal to calculate the secondary structures of ES7 for these two species. The results confirm the close sequence-based evolutionary relationship between these three species, and demonstrate that secondary structural similarity can be a tool for establishing phylogenetic relationships.

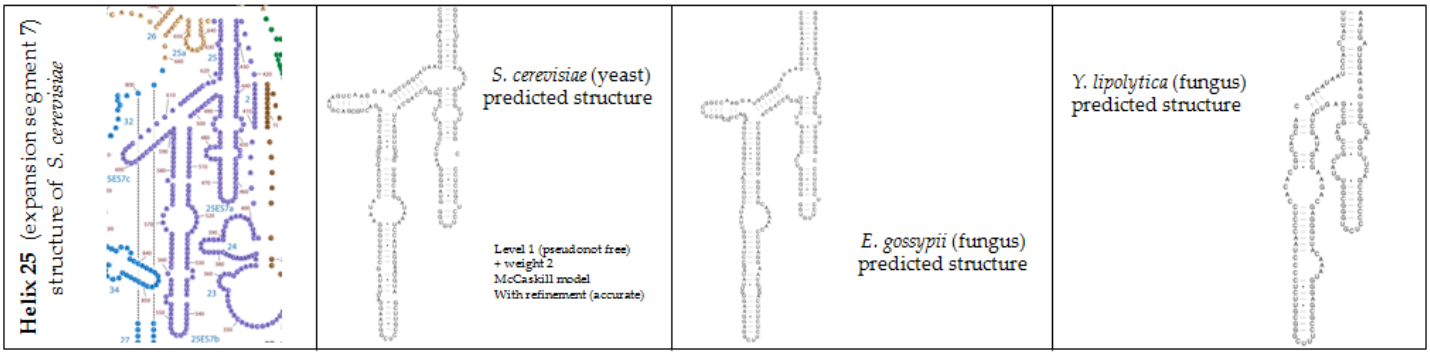


Figure 12. Known and predicted secondary structures of closely-related species on the phylogenetic tree. The observed secondary structure of *S. cerevisiae* (left) was successfully predicted by the folding algorithm, as indicated by similar topologies, helices, and loops. The calculated secondary structures of *Y. lipolytica* and *E. gossypii* were also very similar to the known *S. cerevisiae* structure. [9]

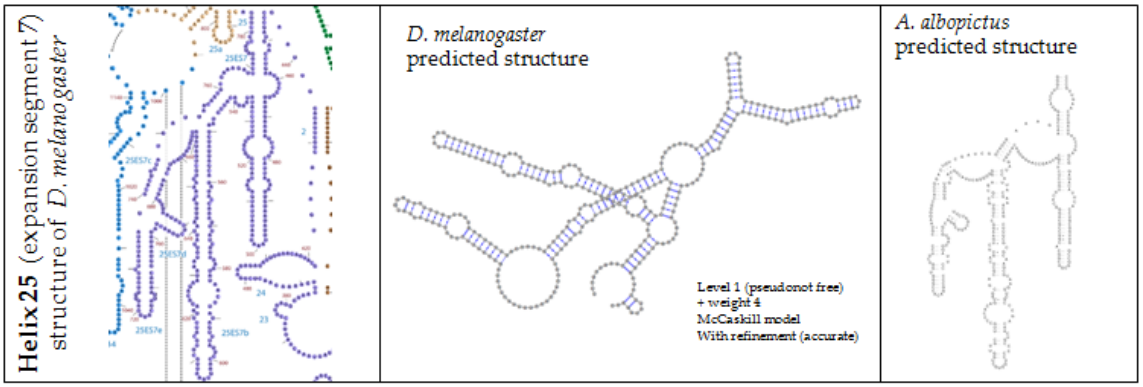


Figure 13. Shown here are the predicted secondary structures of *D. melanogaster* (fruit fly) and *A. albopictus* (mosquito) as compared to the crystallized rRNA secondary structure of *D. melanogaster*.[9]

By contrast, the calculated secondary structure of ES7 of *D. melanogaster*, with a sequence rich in A-U base pairs, was divergent from its known secondary structure. However, with other eukaryotes on the phylogenetic tree (like *T. thermophila* and *A. thaliana*), reasonable similarities were observed between their predicted secondary structures and the closest known secondary structure in aspects of topologies, and number and disposition of helices and loops. Thus, it can be concluded from our results that *D. melanogaster* may be an outlier. Its secondary

structure was not correctly predicted through computation, and for these purposes the rRNA is not an accurate comparison model for other secondary structure predictions. It is astonishing that one of the most broadly used model systems in biological sciences (*D. melanogaster*) shows has an anomalous relationship between rRNA sequence and secondary structure, at least in ES7. This deficiency was further investigated through experimentation.

Results from Part 2: The graphs obtained from the capillary electrophoresis are provided below (dideoxy sequencing graphs in **Fig. 14 and 15**, and SHAPE graphs in **Fig. 16**).

During capillary electrophoresis, material elutes from the capillary array over time. This material passes through a fluorescence detector that scans separately for the FAM and ROX emission wavelengths. Smaller fragments of DNA will elute earlier in the run (less retarded by the polymer matrix), while larger fragments elute later in the run. In this way, one can consider the x-axis of the graphs as time, since the instrument scans happen at set intervals throughout the capillary electrophoresis run. The y-axis of a sequencing or SHAPE graph represents fluorescence intensity. This is a relative fluorescence value reported by the fluorescence detector at the end of the capillary array. In both types of graphs, the green trace represents the ROX ladder that was loaded with the samples in order to easily align the data later, and the blue represents the fluorescent primer extension products.

In a sequencing experiment, the fluorescent primer is extended by the reverse transcriptase enzyme, using the RNA as a template, until a dideoxy NTP (ddNTP) is incorporated. At that point, that strand of DNA is no longer chemically capable of being extended. The peaks in sequencing data represent pieces of DNA where a ddNTP has been incorporated, terminating the reverse transcription of that piece of DNA, and indicating the identity of the corresponding base at that location in the RNA (i.e. if it was a ddCTP, then it must have been paired across from a G

in the RNA). Peaks that appear towards the start of the run (left side) indicate locations of the corresponding base close to the beginning of the primer, whereas peaks that appear further in the run (proceeding to the right) indicate larger fragments, which represent locations of the corresponding base increasingly further from the primer.

Figure 14. Dideoxy sequencing of *D. melanogaster* H25/ES7 rRNA (using Thermo Maxima RT) [15]

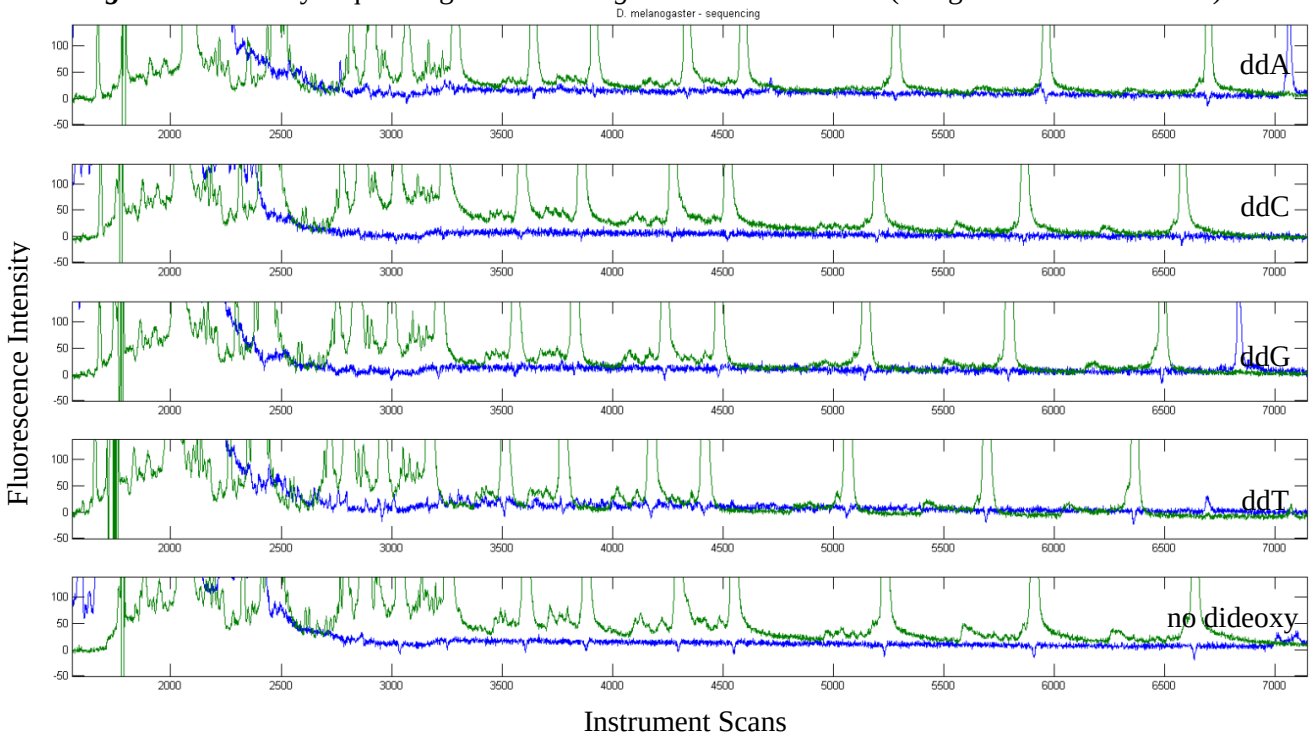
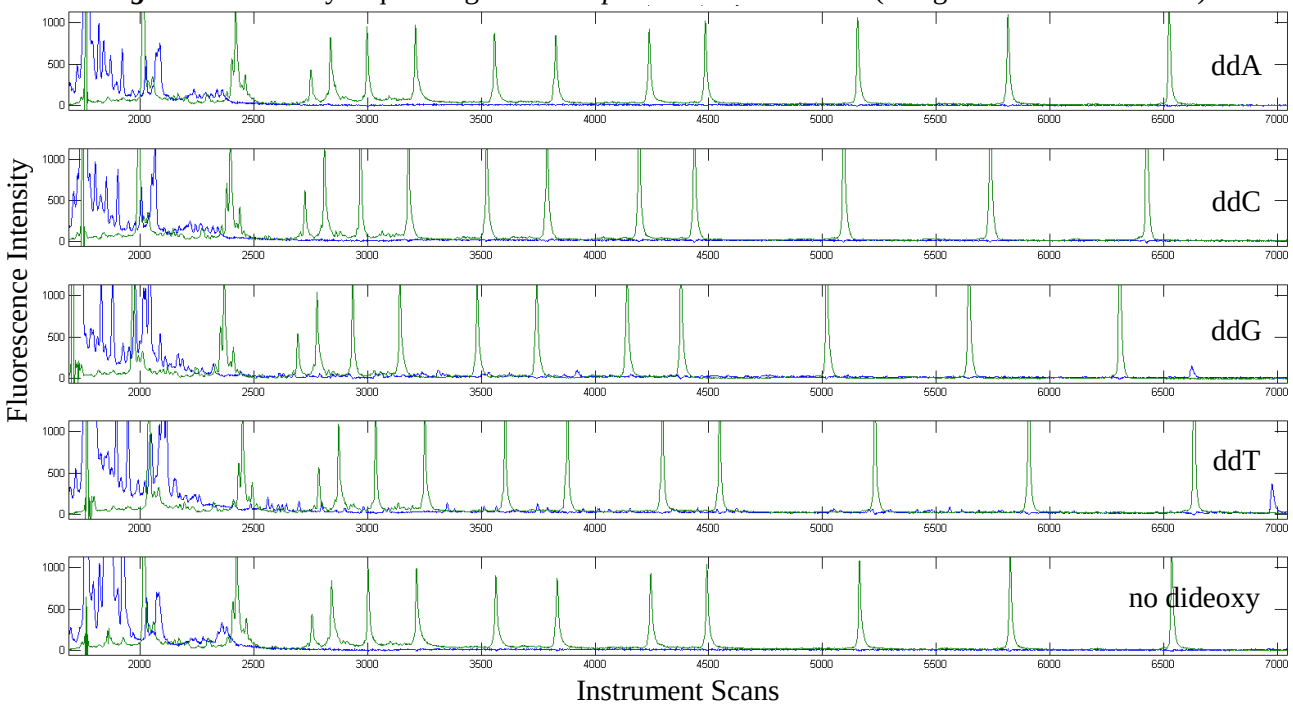


Figure 15. Dideoxy sequencing of *A. albopictus* H25/ES7 rRNA (using Thermo Maxima RT)[15]



From the data collected in the dideoxy sequencing experiment, it was evident that the capillary electrophoresis itself had worked, because the ROX ladder (green) eluted as expected. However, the fact that there was very low/no signal in the sequencing data (blue) indicates that the sequencing reaction did not work very well. The full-length product peaks visible in a couple of the sequencing experiments (towards the right end of the graphs at ~7000 in the ddT lane for *A. albopictus* and in the ddA, ddG, and ddT lanes of *D. melanogaster*) indicate that the reverse transcription worked somewhat, because some of the fluorescent primers were extended all the way to the end of the RNA. However, the signal in the sequencing data was low overall, possibly due to low amounts of RNA, or a low proportion of the primer successfully binding to the RNA. To get rid of ambiguous peaks in the graphs, the sequencing will need to be repeated in the future.

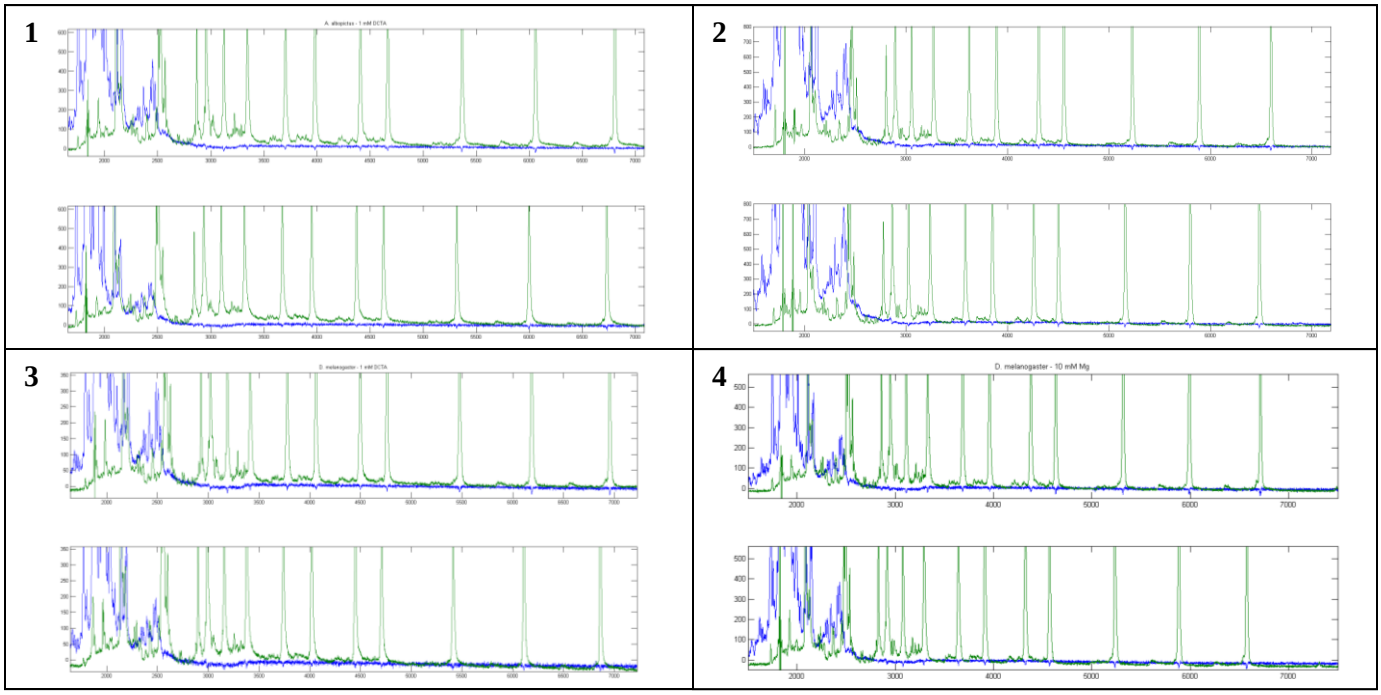
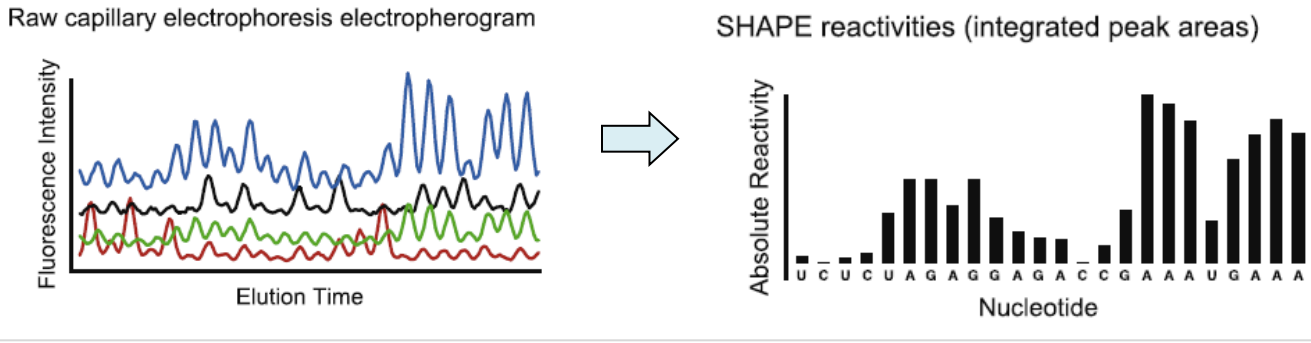


Figure 16. Raw (unprocessed) SHAPE results of (1) *A. albopictus* DCTA, (2) *A. albopictus* Mg, (3) *D. melanogaster* DCTA, and (4) *D. melanogaster* Mg. Top graph in each box represents (+) NMIA and bottom graph is (-) NMIA.[15]

In a SHAPE experiment, the reverse transcription is terminated by the chemical modifications made to the RNA during the NMIA reaction step. When the reverse transcriptase enzyme encounters one of these modifications, it is unable to read any further on the RNA

template strand, and the extension of the fluorescent primer is therefore terminated at that position, providing a collection of cDNA pieces whose various sizes indicate the positions of the various SHAPE modifications in the RNA. A large peak indicates that many primer extensions were terminated at that position, whereas a small peak indicates very few primer extensions were terminated at that position. The raw SHAPE results in **Fig. 16** were then analyzed using the software program ShapeFinder[16] and custom MATLAB scripts, which processed these data to yield normalized SHAPE reactivity values (**Fig. 17**).

Figure 17. Example SHAPE reactivity graph (right) generated from raw SHAPE results[12]



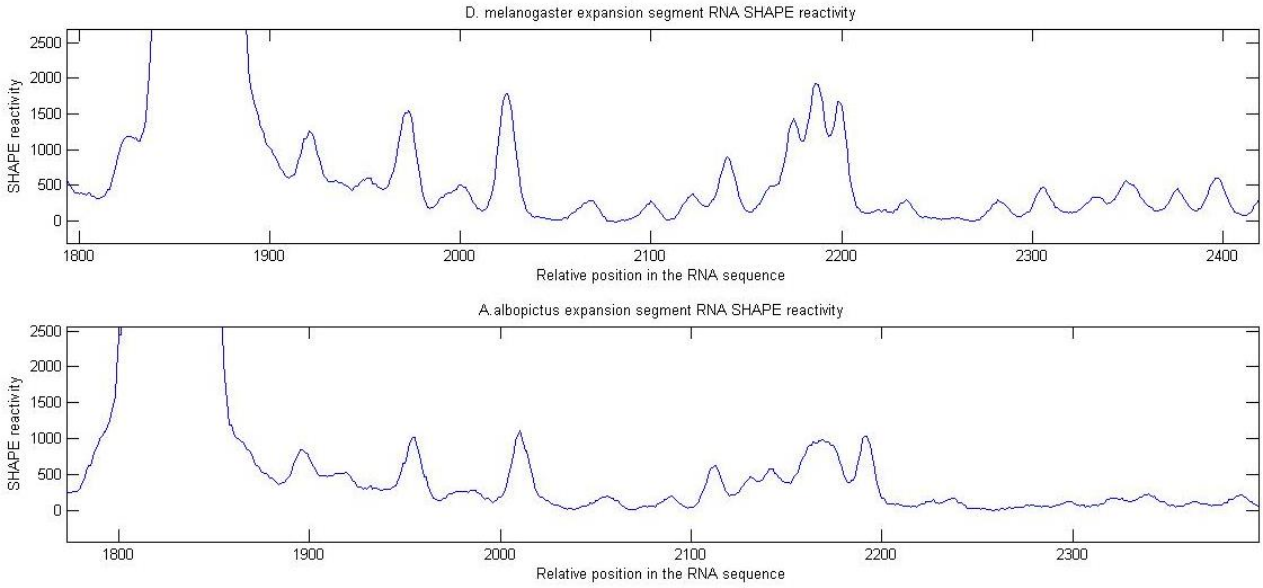


Figure 18. Comparing the SHAPE reactivity of a portion (the 3' end, toward the end of the RNA sequence) of the expansion segment RNA of *D. melanogaster* and *A. albopictus*. As the sequencing data contained several ambiguous peaks, SHAPE reactivity cannot be tied to particular nucleotides at the time.[15]

While comparing the SHAPE reactivity graphs in **Fig. 18** above, an obvious pattern was observed in the peaks that are nearly the same between *D. melanogaster* and *A. albopictus*. From this, it was inferred that the base-pairing pattern is the same between the two species. Therefore, the two species' secondary structures can be considered (in this section of the RNA) very similar as well, perhaps with a minor exception in the range between 2100 and 2200. Overall, the data thus far is consistent with a model in which these two RNAs fold in the same way. Given this information, one may be able to use known secondary structures from other organisms to inform, algorithmically, the secondary structures of the portions that are not known.

Conclusions and Future Work

The results obtained from this research project have brought us closer to the goal of proving the existence of rRNA serial accretion onto a common core in eukaryotic rRNA secondary structures, by allowing for testing of the prediction that the divergent sequences of ES7 of *D. melanogaster* and *A. albopictus* yield similar secondary structures. Both the computational results and SHAPE experiments support the conclusion that rRNA secondary structure can be conserved even when sequence is not, even in the most variable part of the ribosome.

From the sequencing and SHAPE experiments that used the successfully isolated H25/ES7 eukaryotic rRNA, there was strong evidence of striking similarities in the SHAPE profiles of certain portions of the secondary structures of *D. melanogaster* and *A. albopictus* ribosomes. This research can aid in filling in the gaps of secondary structures throughout the phylogenetic tree, based on newer, stronger evolutionary relationships that have been proven. I have shown that by carefully parsing the phylogenetic tree and experimentally determining a small subset of secondary structures, one can interpolate and obtain credible computational estimates of large

numbers for which there are sequences but not experimental data on secondary structure. The results can also be used to help observe the effects of diseases on organisms. Organisms with similar genomes can be compared in their response to different chemicals and conditions. For this reason, this project can contribute to advances in healthcare and medicine by comparing the reactions of organisms to various treatments.

The future for this work will include repeating sequencing and SHAPE in an attempt to increase the capillary electrophoresis signal (e.g., peak height) and length of the read along the gene. A new method has been proposed to predict and calculate any given eukaryotic secondary structure, and to implement this method, work will be done toward creating a database of observed ribosomal evolutionary patterns, in order to determine the exact sites of insertion and elongation occurring in an unknown eukaryotic secondary structure as compared to a closely related one.

Works Cited

[1] Rodnina, M. V., Beringer, M., & Wintermeyer, W. (2007). How ribosomes make peptide bonds. *Trends in Biochemical Sciences,* 32(1), 20-26.

[2] Wiśniewski, M., & Długosz, E. (2006). Molecular diagnostic of parasites using rRNA gene sequence. *PubMed*, 52(4), 263-9.

[3] Gerbi, S.A. (1996). Expansion segments: Regions of variable size that interrupt the universal core secondary structure of ribosomal RNA. In R. A. Zimmermann & A. E. Dahlberg (Eds.), Ribosomal RNA—Structure, evolution, processing, and function in protein synthesis (pp. 71–87). Boca Raton, FL: CRC Press.

[4] Petrov, A. S., Bernier, C. R., Hsiao, C., Norris, A. M., Kovacs, N. A., Waterbury, C. C., . . . Williams, L. D. (2014). Evolution of the Ribosome at Atomic Resolution. Proc Natl Acad Sci USA, submitted.

[5] Leontis, N. B., Stombaugh, J., & Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Research, 30(16), 3497-3531.

[6] Sim, A. Y., & Levitt, M. (2011). Clustering to identify RNA conformations constrained by secondary structure. Proc Natl Acad Sci USA, 108(9), 3590-3595. doi: 10.1073/pnas.1018653108

[7] Center for Ribosomal Origins and Evolution, Department of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA.

[8] Bernier, C., Petrov, A.S., Waterbury, C., Jett, J., Li, F., Freil, L.E., Xiong, b., Wang, L., Le, A., Milhouse, B.L., Hershkovitz, E., Grover, M., Xue, Y., Hsiao, C., Bowman, J.C., Harvey, S.C., Wartel, J.Z., and Williams, L.D. (2014). Ribovision: Visualization and Analysis of Ribosomes. Faraday Discuss, 169,1-12.

[9] Markham, N. R., Wright, A., Zuker, L. S., & Zuker, M. (n.d.). The mfold Web Server (Version 2.3) [Computer software]. Retrieved from http://mfold.rna.albany.edu/?q=mfold

[10] Merino, E.J., Wilkinson, K.A., Coughlan, J.L., Weeks, K.M. (2005). RNA Structure Analysis at Single Nucleotide Resolution by Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE). J. Am. Chem. Soc, 127,4223-4231.

[11] Wilkinson, K.A., Merino, E.J., & Weeks, K.M. (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols, 1*(3), 1610-1616. doi :10.1038/nprot.2006.249

[12] Low, J.T. & Weeks, K.M. (2010). SHAPE-directed RNA secondary structure prediction. *Methods, 52,* 150-158. doi: 10.1016/j.ymeth.2010.06.007

[13] *Zymo Research RNA Clean & Concentrator-25 kit.* ZymoResearch, 2013.

[14] *Thermo Maxima H minus reverse transcription kit.* Thermo-scientific Bio, 2013.

[15] MATLAB version 8.3. Natick, Massachusetts: The MathWorks Inc., 2014.

[16] ShapeFinder version 2.0. Stemmer Imaging, 2014.

Annotated Bibliography

[1] Rodnina, M. V., Beringer, M., & Wintermeyer, W. (2007). How ribosomes make peptide bonds. *Trends in Biochemical Sciences,* 32(1), 20-26.

This research report from the Royal Swedish Academy of Sciences, by the 2009 Nobel Prize in Chemistry awardees Venkatraman Ramakrishnan, Thomas Steitz, and Ada Yonath, details the structure and function of the ribosome. The report discusses the central dogma, components of the ribosome, initiation of protein synthesis, peptide elongation and translocation, and the path to obtaining high resolution crystal structures of ribosomal subunits.

[2] Wiśniewski, M., & Długosz, E. (2006). Molecular diagnostic of parasites using rRNA gene sequence. *PubMed*, 52(4), 263-9.

In their PubMed paper, Dlugosz and Wiśniewski focus on the genes that encode different rRNA molecules within eukaryotic ribosomes. The report summarizes the uses of internal transcribed spacers (ITS), or the regions within the ribosomal transcript that are degraded during maturation, for molecular diagnostic of parasite. As ITS sequences generally show more variation than ribosomal sequences, they are considered species-specific and are popular for phylogenetic analysis and identification of species or strains.

[3] Gerbi, S.A. (1996). Expansion segments: Regions of variable size that interrupt the universal core secondary structure of ribosomal RNA. In R. A. Zimmermann & A. E. Dahlberg (Eds.), Ribosomal RNA—Structure, evolution, processing, and function in protein synthesis (pp. 71–87). Boca Raton, FL: CRC Press.

In his paper, Gerbi introduced the concept of expansion segment rRNA, claiming that the length and sequence of these expansion segments varies considerably between organisms. Studying the 18s rRNA, Gerbi found evidence of twelve expansion segments, each with an average length of 250 nucleotides, and categorized them based on sequence variability and conservation.

[4] Petrov, A. S., Bernier, C. R., Hsiao, C., Norris, A. M., Kovacs, N. A., Waterbury, C. C., . . . Williams, L. D. (2014). Evolution of the Ribosome at Atomic Resolution. Proc Natl Acad Sci USA, submitted.

Here, members of the Williams lab detail the evolution of the ribosome through C-value, LSU rRNA size, and placement on the *phylogenetic tree of life*. The report investigates the role of "insertion fingerprints", available structures that allow one to make direct

comparisons of pre- and post-expanded rRNA, and to observe rRNA conformation at sites where expansion elements join common core rRNA. Williams and his team also examine the buildup of the peptidyl transferase center, responsible for peptide bond formation, over evolution.

[5] Leontis, N. B., Stombaugh, J., & Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Research, 30(16), 3497-3531.

This research report explores the underlying backbone structure of RNA; here, Leontis, Stombaugh, and Westhof propose a classification of RNA base pairs based on geometry. They claim that this approach is justified by the need to easily identify recurrent structural motifs in new crystal structures (i.e. three-dimensional RNA models) and to predict occurrences of motifs through comparative sequence analysis, with the assumption that secondary and tertiary structure is more conserved than sequence. In the paper, twelve such basic geometric families were identified, and geometric relationships between the 16 pair-wise combinations of the four standard bases, A, C, G, and U, were summarized in "isostericity matricies".

[6] Sim, A. Y., & Levitt, M. (2011). Clustering to identify RNA conformations constrained by secondary structure. Proc Natl Acad Sci USA, 108(9), 3590-3595. doi: 10.1073/pnas.1018653108

Working as a Professor at the Stanford School of Medicine, Levitt discusses secondary structures, which are representations that reduce the dimensionality from three to two, assuming that a given base is in one of two discrete states: paired or non-paired. A paired base is involved in secondary interactions. Among other topics discussed, the report focuses on the underlying RNA backbone structure that consists of helices, loops, and psedoknots. Following Levitt, we define helices as three consecutive base-paired nucleotides bounded by non-paired nucleotides.

[7] Center for Ribosomal Origins and Evolution, Department of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA.

This research was conducted at the Center for Ribosomal Origins and Evolution (Ribo Evo), partnered with the NASA Astrobiology Institute, at the Georgia Institute of Technology.

[8] Bernier, C., Petrov, A.S., Waterbury, C., Jett, J., Li, F., Freil, L.E., Xiong, b., Wang, L., Le, A., Milhouse, B.L., Hershkovitz, E., Grover, M., Xue, Y., Hsiao, C., Bowman, J.C., Harvey, S.C., Wartel, J.Z., and Williams, L.D. (2014). Ribovision: Visualization and Analysis of Ribosomes. Faraday Discuss, 169,1-12.

RiboVision is the large 3D graphics system that was used to visualize complex biomolecules. Here, expansion segments (in three-dimensional form) were matched with their corresponding positions on secondary structure diagrams. RiboVision aided greatly in mapping out the data collected.

[9] Markham, N. R., Wright, A., Zuker, L. S., & Zuker, M. (n.d.). The mfold Web Server (Version 2.3) [Computer software]. Retrieved from http://mfold.rna.albany.edu/?q=mfold

Markham et al. have created an algorithm by which to calculate ribosomal secondary structures based on inputted rRNA sequences through integer programming. Based on probabilities of most likely base-pairings occurring, a secondary structure is drawn on the screen. However, it was found that secondary structural prediction did not work, in general, for complex eukaryotic ribosomes.

[10] Merino, E.J., Wilkinson, K.A., Coughlan, J.L., Weeks, K.M. (2005). RNA Structure Analysis at Single Nucleotide Resolution by Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE). J. Am. Chem. Soc, 127,4223-4231.

This research paper introduces SHAPE (**s**elective 2'-**h**ydroxyl **a**cylation analyzed by **p**rimer **e**xtension), a nucleic acid structure probing technique that takes advantage of reagents that are able to modify the backbone of RNA in structurally flexible regions. Reagents such as N-methylisatioc-anhydride (NMIA) undergo hydrolysis to form adducts on the 2'-hydroxyl of the RNA backbone. Adduct formation is quantified for each nucleotide in a given RNA by extension of a complementary DNA primer with reverse transcriptase and comparison of the resulting fragments with those from an unmodified control. Though the results of my SHAPE experiments did not actually give the means to quantitatively calculate the eukaryotic secondary structures, I was provided with a series of graphs qualitatively reporting on a portion of the rRNA secondary structure.

[11] Wilkinson, K.A., Merino, E.J., & Weeks, K.M. (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols, 1*(3), 1610-1616. doi :10.1038/nprot.2006.249

In this report, Wilkinson, Merino, and Weeks outline the SHAPE protocol for experiments. The process includes the initial RNA design, RNA folding, primer extension, and dideoxy sequencing through capillary electrophoresis. The final result produced is a graph of normalized SHAPE reactivity values, which can be used to confirm/refine secondary structures, monitor RNA folding, propose new structures, and detect alternative conformations.

[12] Low, J.T. & Weeks, K.M. (2010). SHAPE-directed RNA secondary structure prediction. *Methods, 52*, 150-158. doi: 10.1016/j.ymeth.2010.06.007

Primarily, this third SHAPE paper by Kevin Weeks discusses the steps required to convert capillary electrophoresis electropherograms into quantitative reactivity measurements with the ShapeFinder software. An example of a "final product" secondary structure obtained through thermodynamic and SHAPE-derived free energy change contributions is summarized in the form of a simple hairpin loop with a successfully-predicted structure.

[13] *Zymo Research RNA Clean & Concentrator-25 kit.* ZymoResearch, 2013.

Procedures and materials contained in the *Zymo Research RNA Clean & Concentrator-25 kit* were used in the purification process of the RNA after it was purified with ethanol once the NMIA-modification process had completed.

[14] *Thermo Maxima H minus reverse transcription kit.* Thermo-scientific Bio, 2013.

Reverse transcription for the SHAPE experiments was conducted as per the protocol provided in the *Thermo Maxima H minus reverse transcription kit*. During this step, DNA primers are annealed to the RNA and then extended to sites of modification in the presence of dNTPs by the activity of reverse transcriptase, after which cDNA fragments with varying lengths are obtained and resolved through dideoxy sequencing.

[15] MATLAB version 8.3. Natick, Massachusetts: The MathWorks Inc., 2014.

Data collected from capillary electrophoresis was inputted into MATLAB to produce fluorescence intensity graphs, which were further combined and quantified into SHAPE reactivity graphs. In both the dideoxy sequencing and SHAPE graphs produced, the green traced represented the ROX ladder that was loaded with the samples in order to easily align the data later, and the blue represented the florescent primer extension products. Integrates computation, visualization, and programming into a single environment, MATLAB served as an effective tool in reading all graphs obtained.

[16] ShapeFinder version 2.0. Stemmer Imaging, 2014.