INDIAN INSTITUTE OF SCIENCE AND RESEARCH MOHALI

Structural & Functional Analysis of mutated s2m mobile genetic element present in human and bat/pangolin coronaviruses

---

Author:
Akhil Pratap
(MS16087)

Supervisor:
Dr. Monika Sharma

A dissertation submitted in fulfillment of the requirements
for the degree of BS-MS in Chemical Sciences

April 2021

Certificate of Examination

This is to certify that Akhil Pratap (Reg. No. MS16087) submitted the dissertation titled "**Structural & Functional Analysis of mutated s2m mobile genetic element present in human and bat/pangolin coronaviruses**" for the BS-MS dual degree programme in Chemical Sciences to the Institute. The thesis committee, hereby named by the Institute, has examined it. The committee considers the candidate's work to be acceptable and suggests that the work be approved.

Dr.Sabyasachi Rakshit | Dr.Shashi Bhushan Pandit | Dr. Monika Sharma (Supervisor)

Declaration

I carried out research presented in this dissertation at the Indian Institute of Science Education and Research in Mohali, under the supervision of Dr. Monika Sharma. This thesis has not been applied in part or in full to any other university or institute for a degree, certificate, or fellowship. Where other people's contributions are involved, every attempt is made to explicitly demonstrate this, with due credit given to collective analysis and discussions. This thesis is a direct record of my original work, and the bibliography lists all of the references cited inside.

Akhil Pratap
(Candidate)

Date:12-4-2021

In my capacity as the candidate's project supervisor, I certify that the candidate's above statements are genuine to the best possible standard.

Dr. Monika Sharma
(Supervisor)

Acknowledgements

First and foremost, I want to thank my supervisor, Dr. Monika Sharma, for her constant support of my MS thesis and research, as well as her patience, encouragement, enthusiasm, and extensive knowledge.

Apart from my supervisor, I'd like to express my gratitude to the rest of my thesis committee: Dr. Shashi Bhushan Pandit and Dr. Sabyasachi Rakshit for their useful suggestions.

Also, I'd like to express my gratitude to my friends Mr. Aditya Mishra and Mr. Shivam Kumar for their unconditional support during my dissertation, as well as Mr. Aritra Bhattacharya for his in-depth knowledge of MD simulations.

I'd like to express my beloved family and IISER Mohali for their unwavering support during my thesis.

Notation

PDB	Protein data bank
MFE	Minimum free Energy
Pf	Partition function
UTR	Untranslated region
S2M	Stem-loop II motif
Pp	base-pair probabilities
Pe	Positional entropy

List of figures

Figure 1.1. Distribution of the whole genome of Sars-CoV2…….....16
Figure 1.2. s2m secondary structure (Minimum Free Energy structure drawing encoding base-pair probabilities).....17
Figure 1.3. 3d Structure of s2m motif (PDB:1XJR).....18
Figure 2.1. 17 different s2m Sequences found in various h-COV19 sequences obtained fromGISAID/NCBI……...……...20
Figure 2.2. The standard reference sequence for GISAID (EPI ISL 402124) is hCoV-19/Wuhan/WIV04/2019 (WIV04).....21
Figure 2.3.1. Representation of vector bit 0 and 1…….....22
Figure 2.3.2. Each nucleotide takes on a distinct form as a result of this mapping……23
Figure 3.1 Mapping of all 18 nucleotides with reference 402124 (Reference genome for human) and PDB: 1XJR (Experimental structure).....26
Figure 3.2.1 MFE and Centroid Secondary structure of motif 402124……...28
Figure 3.2.2 MFE and Centroid Secondary structure of motif MT308984.1……..29
Figure 3.2.3 MFE and Centroid Secondary structure of motif KY417152.1 & MT072865.1…….30
Figure 3.2.4 MFE and Centroid Secondary structure of motif 862313……...31
Figure 3.2.5 MFE and Centroid Secondary structure of motif 1396884…….32
Figure 3.2.6 MFE and Centroid Secondary structure of motif 1390981…….33
Figure 3.2.7 MFE and Centroid Secondary structure of motif 1391006.……..….34
Figure 3.2.8 MFE and Centroid Secondary structure of motif 1391015.……..….35
Figure 3.2.9 MFE and Centroid Secondary structure of motif 1391017.……..…....36
Figure 3.2.10 MFE and Centroid Secondary structure of motif 1397948.……...…...37
Figure 3.2.11 MFE and Centroid Secondary structure of motif 1300659.……...…...38
Figure 3.2.12 MFE and Centroid Secondary structure of motif 1292989.……...…...39
Figure 3.2.13 MFE and Centroid Secondary structure of motif 1273404.……...…...40
Figure 3.2.14 MFE and Centroid Secondary structure of motif 414480.……...….41
Figure 3.2.15 MFE and Centroid Secondary structure of motif 862510.……...….42
Figure 3.2.16 MFE and Centroid Secondary structure of motif 1319040.…….….43
Figure 3.2.17 MFE and Centroid Secondary structure of motif MW883290.1……..……...44
Figure 3.2.18 MFE and Centroid Secondary structure of PDB:1XJR……...45

Figure 3.2.1 3D structure of 402124 Left (optimal structure) and Right (Centroid Structure) motif.……47
Figure 3.3.2 3D structure of MT308984.1 Left (optimal structure) and Right (Centroid Structure) motif.……...48
Figure 3.3.3 3D structure of KY417152.1 & MT072865.1 Left (optimal structure) and Right (Centroid Structure) motif.……....48
Figure 3.3.4 3D structure of 862313 Left (optimal structure) and Right (Centroid Structure) motif.……...…….49
Figure 3.3.5 3D structure of 1396884 Left (optimal structure) and Right (Centroid Structure) motif.……...…….49
Figure 3.3.6 3D structure of 1390981 Left (optimal structure) and Right (Centroid Structure) motif.……...…….50
Figure 3.3.7 3D structure of 1391006 Left (optimal structure) and Right (Centroid Structure) motif.……...…….50
Figure 3.3.8 3D structure of 1391015 Left (optimal structure) and Right (Centroid Structure) motif.……...…….51
Figure 3.3.9 3D structure of 1391017 Left (optimal structure) and Right (Centroid Structure) motif.……...…….51
Figure 3.3.10 3D structure of 1397948 Left (optimal structure) and Right (Centroid Structure) motif.……...……...…….52
Figure 3.3.11 3D structure of 1300659 Left (optimal structure) and Right (Centroid Structure)motif.……...…….53
Figure 3.3.12 3D structure of 1292989 Left (optimal structure) and Right (Centroid Structure) motif.……...……...…….53
Figure 3.3.13 3D structure of 1273404 Left (optimal structure) and Right (Centroid Structure)motif.……...…….54
Figure 3.3.14 3D structure of 414480 Left (optimal structure) and Right (Centroid Structure) motif.……...…….54
Figure 3.3.15 3D structure of 862510 Left (optimal structure) and Right (Centroid Structure) motif.……...…….55
Figure 3.3.16 3D structure of 1319040 Left (optimal structure) and Right (Centroid Structure) motif.……..……..55
Figure 3.3.17 3D structure of MW883290.1 Left (optimal structure) and Right (Centroid Structure) motif.……..…..56

Figure3.3.18 3D structure of PDB: 1XJR Left (optimal structure) and Right (Centroid Structure) motif.

List of tables

Table 2.1.1. 18 Different s2m motifs.

Table 3.6. Seventeen different S2M Motifs. (a) For sequence representation input for ClustalW multiple sequence alignment trees and stem-loop structure representation, alignment of stem-forming elements and columns for each genotype. (b) Display of S2M coronavirus ClustalW numerous sequence alignment trees.

Contents

Certificate of Examination ii

Declaration of Authorship iv

Acknowledgement vi

Notation vii

List of figures viii

List of Tables xi

Abstract xv

1 **Introduction**……....16

1.1 s2m motif…….....17

2 **Materials, Methodology and Computational logs**……...19

2.1 System studied……...20

2.2. Occupancy and Allele Frequency Calculation……..22

2.3. System Visualization……....22

2.4. Quantifying the Similarities between Motifs, Secondary Structures and Tertiary structure……24

2.4.1. Partition function foldings and dot plots……24

2.4.2. Minimum free Energy Secondary Structure……...24

2.4.3. Centroid Secondary structure…….25

2.4.4.Dot-Bracket Energy notation……..25

2.5. Guided Trees…….25

2.6. Molecular Dynamics Simulation……...25

2.6.1 Energy Minimization of 3D structures……....26

3 **Results**…….27

3.1. System Visualization in comparison with s2m motif(1XJR) and Reference genome WIV04/402124……...27
3.2. MFE and Centroid Secondary structure……....29
3.3. Tertiary Structures…….48
3.4 Interaction between S2M motifs and human miRNAs…….58
3.5 Thermodynamic ensemble results of various motifs compared with the reference genome…….58
3.7 Guided Trees & phylogenetic trees……...….61
4 **Conclusion**……...64

Abstract

The current COVID-19 pandemic highlighted the role of gaining a better understanding of the coronavirus life cycle. SARS-CoV-2, the disease's causative agent, is being extensively studied structurally in order to gain insight into key molecular mechanisms needed for its survival.Various conserved stem-loop elements are found in the untranslated regions (UTR) of the SARS-CoV-2 genome and are thought to be involved in RNA replication, viral protein translation, and discontinuous transcription.There is a lack of information about SARS-CoV-2 mutations and the effect of these polymorphisms on viral transmission load.The genomic sequence of SARS-CoV-2 was analysed to find variants in the 3'UTR region of its cis-regulatory RNA elements in this analysis. There was discovered a 43-nucleotide genetic element with a highly conserved stem-loop II-like motif (S2M).As compared to the S2M structures of bat/pangolin models, these polymorphisms tend to make the S2M secondary and tertiary structures in human SARS-CoV-2 models less stable. This gives RNA structures more mobility, which may be one of its defensive mechanisms against host defences or make it easier for it to enter host proteins and enzymes. While this S2M sequence may not be found in all human SARS-CoV-2 models, when it does, it is always highly conserved. It may be used as a target for vaccines and therapeutic agents in the future for mutated strains.

Chapter 1

Introduction

The emergence of new viral pathogens is a public health concern (1). The Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-1), Middle East Respiratory Syndrome Coronavirus (MERS-CoV), and a recently described zoonotic coronavirus (SARS-CoV-2) are zoonotic viruses that use bats as their natural reservoir. They are then passed on to humans through intermediate hosts (2). The strain on SARS-host CoV-2's selection in human models would have an effect on the long-term conservation of mutations that increase the virus's transmissibility.

Standard structural genomics researchers based on obtaining the three-dimensional structures of proteins located within a genome and defining unknown protein functions based on three-dimensional homology to recognise protein structures. The three-dimensional structures of non-protein gene products, such as the various RNAs necessary for mRNA processing, protein synthesis, and other cellular functions, must also be identified and elucidated (3).

Through SARS-related beta coronaviruses and currently available SARS-CoV-2 sequences on GISAID, we look for RNA sequence conservation and conserved 3'-UTR regions that include s2m in each sequence set; these intervals can be used to develop diagnostic and antiviral strategies (4). To find organized structured regions, we use structural ensemble predictions for each RNA to predict structures with the highest expected accuracy across conserved regions and report the support for these single structures.

Like a typical mRNA, the SARS-CoV-2 genome has a 5′-cap, a 5′UTR, a 3′-UTR, and a poly-A tail
(1) 16 Nonstructural protein (NSP1 -NSP16) position 266 to 21555,
(2) S Protein (Spike glycoprotein) position 21563 to 25384,
(3) ORF3a (NS3a) position 25393 to 26220,
(4) ORF3b (NS3b) position 25765 to 26220,
(5) E (Envelope protein) position 26245 to 26472,
(6) M (Membrane protein) position 26523 to 27191,

(7) ORF6 (Nonstructural protein NS6) position 27202 to 27387,
(8) ORF7a (Nonstructural protein NS7a) position 27394 to 27759,
(9) ORF7b (Nonstructural protein NS7b) position 27756 to 27887,
(10) ORF8 (Nonstructural protein NS8) position 27894 to 28259,
(11) N (Nucleocapsid protein) position 28274 to 29533,
(12) ORF10 (Nonstructural protein NS10) position 29558 to 29674

While the existence of s2m near the 3′ end of specific coronavirus genomes has been previously reported, this analysis is an update on the current status of s2m in this virus family due to the exponential growth of gene sequence data available through public databases(GISAID,NCBI,etc) (5).

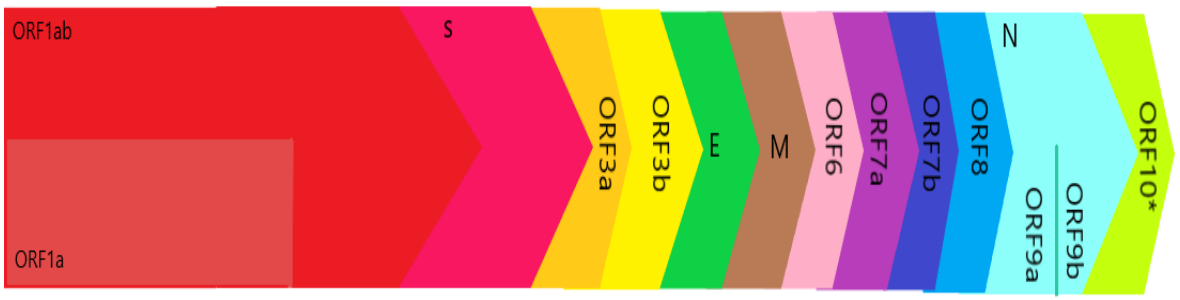


Figure 1.1. Distribution of the whole genome of Sars-CoV2.

1.1. s2m (motif)

SARS and related coronaviruses and astroviruses all possess a motif at the 3' end of their RNA genomes, called the s2m, whose pathogenic importance is inferred from its strict sequence conservation in an otherwise rapidly mutable RNA genome. We discovered that the necessity to form a highly organized RNA with a unique tertiary structure that includes a sharp 90-degree kink of the helix axis and several novel longer-range tertiary interactions explains this intense conservation. The tertiary base interactions form a negative-charged tunnel that runs perpendicular to the central helical axis and connects two magnesium ions.Near the 3' end of members of the Astroviridae family, a 43 base pair conserved

sequence motif was discovered . The genetic element was called s2m because it corresponded to the second most 3' stem-loop structure (stem-loop II) in human astroviruses. Later, the sequence motif was discovered in three more virus families: Caliciviridae, Picornaviridae, and Coronaviridae (5). s2m appears to be restricted to positive-sense, single-stranded RNA (+ssRNA) viruses, and the factor is often found near the genome's 3' end.These unusual characteristics are most likely used to shape interaction surfaces with conserved host cell components or other reactive sites necessary for virus function. We propose that these distinctive structural features in the s2m RNA element are attractive targets for the design of antiviral therapeutic agents based on their conservation in viral pathogen genomes and their absence in the human genome.

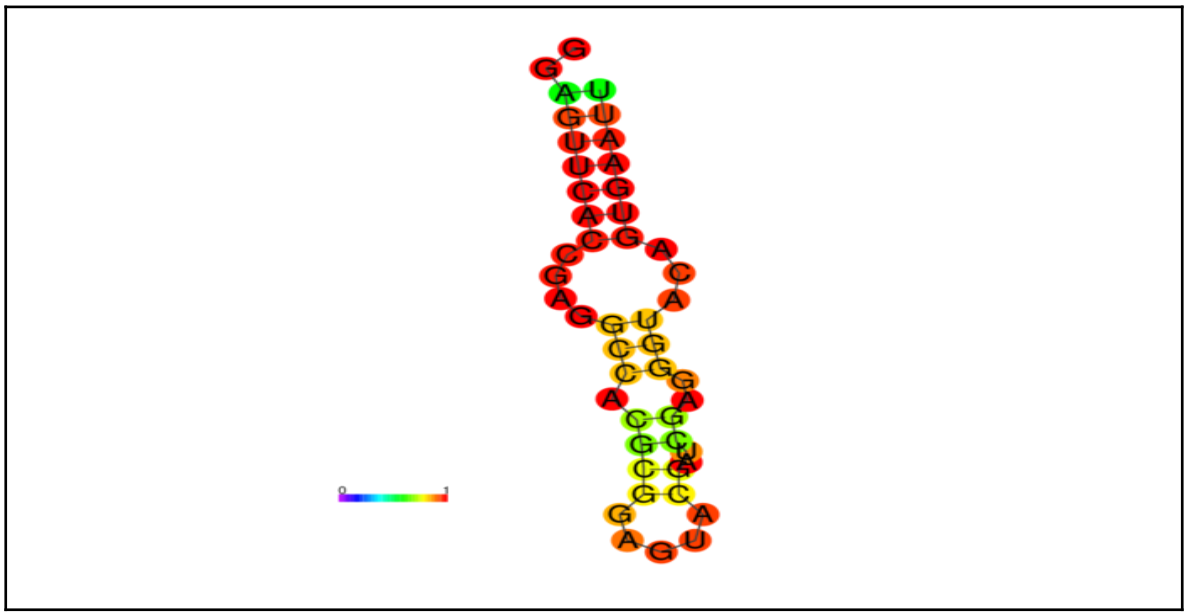


Figure 1.2. s2m secondary structure (Minimum Free Energy structure drawing encoding base-pair probabilities).

Results for thermodynamic ensemble prediction for the Experimental Structure:

The free energy of the thermodynamic ensemble is -9.99 kcal/mol.

The frequency of the MFE structure in the ensemble is 38.59 %.

The ensemble diversity is 4.63 .

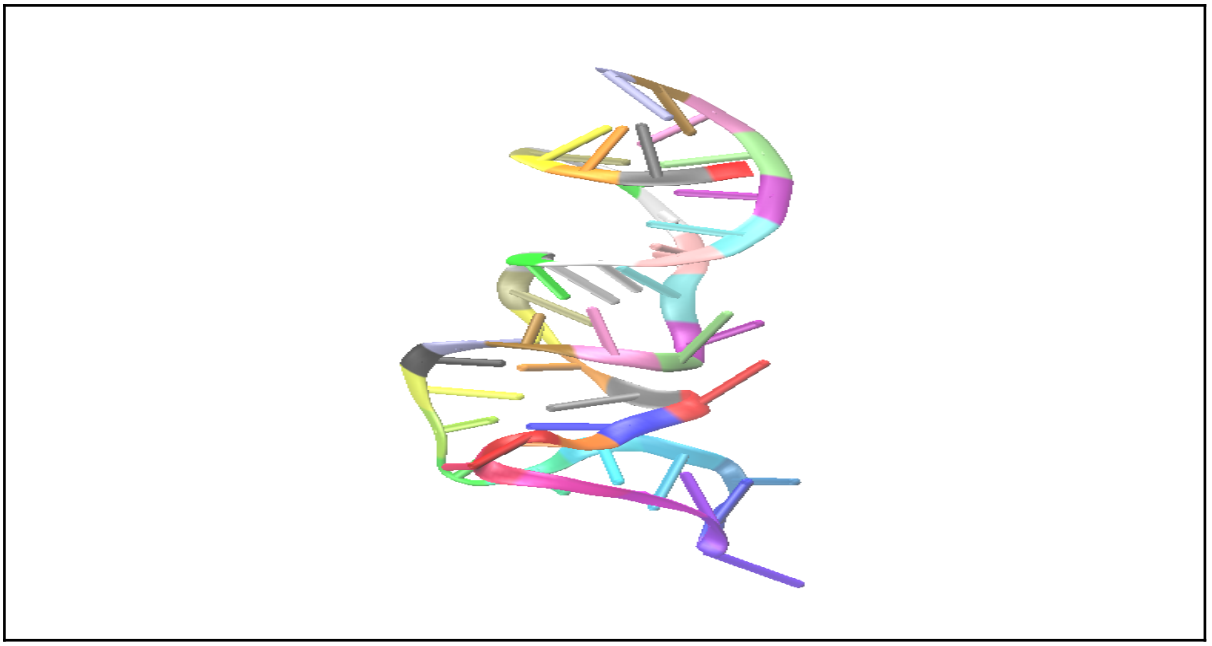


Figure 1.3. 3d Structure of s2m motif (PDB:1XJR)

Chapter 2

Materials, Methodology and Computational logs

2.1. System studied

In my thesis, I have taken many different genomic Sequences of hCoV-19 covering Australia, India, USA, Canada, etc. All Studied s2m-motif-containing Sequences are listed below are mutated at some location in the 43-nucleotide genetic element of positive-sense ssRNA viruses. Accession number(NCBI) or GISAID(Accession ID) is provided with the Sequences.

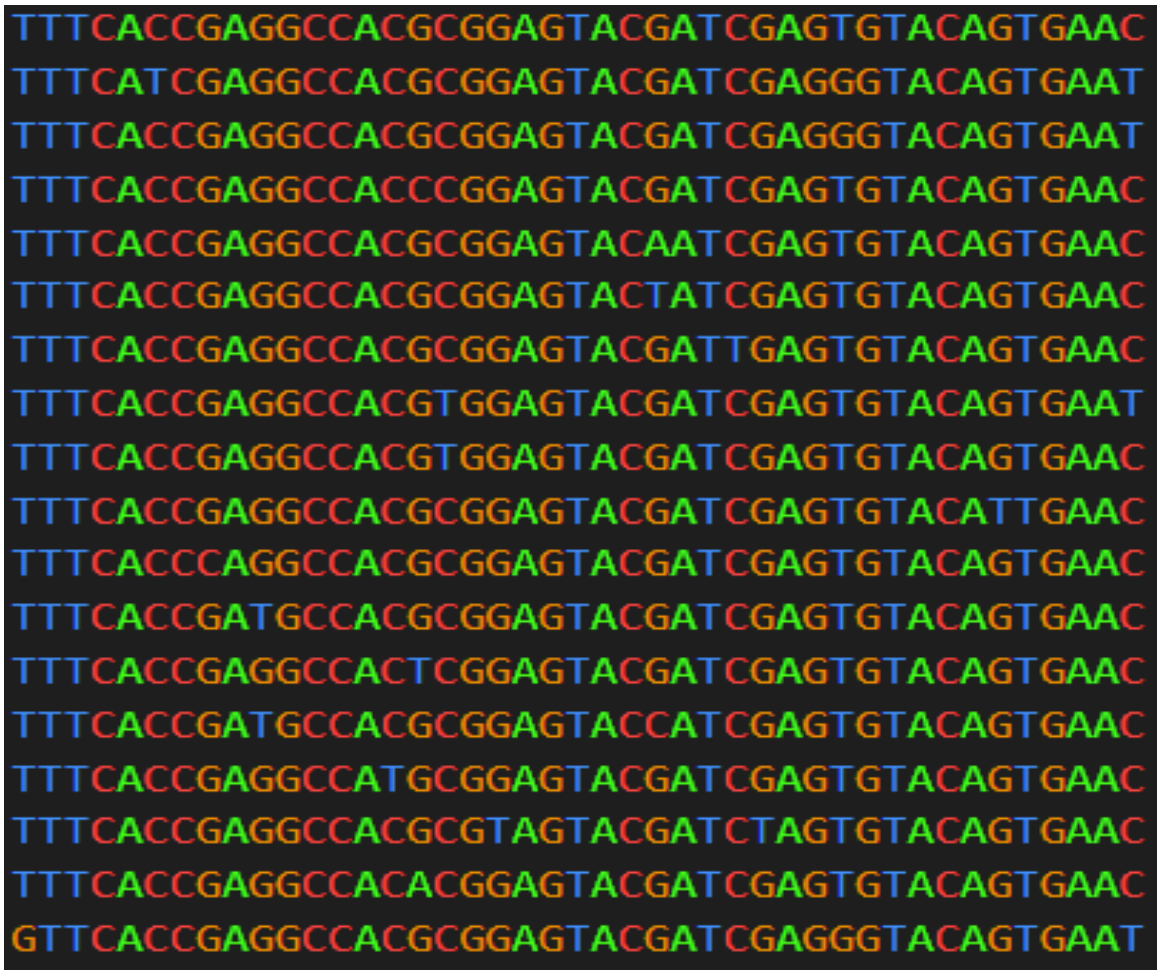


Table 2.1.1. 17 Different s2m motifs.

```
a> hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30
TTTCACCGAGGCCACGCGGAGTACGATCGAGTGTACAGTGAAC
b> MT308984.1 Mutant SARS coronavirus Urbani clone
TTTCATCGAGGCCACGCGGAGTACGATCGAGGGTACAGTGAAT
c> KY417152.1 Bat SARS-like coronavirus & MT072865.1 Pangolin coronavirus
TTTCACCGAGGCCACGCGGAGTACGATCGAGGGTACAGTGAAT
d> hCoV-19/India/AP-CS0319/2020|EPI_ISL_862313|2020-07-01
TTTCACCGAGGCCACCCGGAGTACGATCGAGTGTACAGTGAAC
e> hCoV-19/France/IDF-HMN-21032100486/2021|EPI_ISL_1396884|2021-03-09
TTTCACCGAGGCCACGCGGAGTACAATCGAGTGTACAGTGAAC
f> hCoV-19/Canada/QC-HCLM-7056140333/2020|EPI_ISL_1390981|2020-11-12
TTTCACCGAGGCCACGCGGAGTACTATCGAGTGTACAGTGAAC
g> hCoV-19/Canada/QC-CHIC-R4180563/2020|EPI_ISL_1391006|2020-10-18
TTTCACCGAGGCCACGCGGAGTACGATTGAGTGTACAGTGAAC
h> hCoV-19/Indonesia/NT-NIHRD-C002128132/2021|EPI_ISL_1391015|2021-02-11
TTTCACCGAGGCCACGTGGAGTACGATCGAGTGTACAGTGAAT
i> hCoV-19/Indonesia/JB-NIHRD-C002130218/2021|EPI_ISL_1391017|2021-03-04
TTTCACCGAGGCCACGTGGAGTACGATCGAGTGTACAGTGAAC
j> hCoV-19/USA/NY-Wadsworth-21030106-01/2021|EPI_ISL_1397948|2021-03-07
TTTCACCGAGGCCACGCGGAGTACGATCGAGTGTACATTGAAC
k> hCoV-19/Bosnia and Herzegovina/KCUS21066/2021|EPI_ISL_1300659|2021-02-24
TTTCACCCAGGCCACGCGGAGTACGATCGAGTGTACAGTGAAC
l> hCoV-19/USA/CT-Yale-1770/2021|EPI_ISL_1292989|2021-02-19
TTTCACCGATGCCACGCGGAGTACGATCGAGTGTACAGTGAAC
m> hCoV-19/Lesotho/N3421/2021|EPI_ISL_1273404|2021-01-18
TTTCACCGAGGCCACTCGGAGTACGATCGAGTGTACAGTGAAC
n> hCoV-19/USA/CruiseA-21/2020|EPI_ISL_414480|2020-02-21
TTTCACCGATGCCACGCGGAGTACCATCGAGTGTACAGTGAAC
o> hCoV-19/India/AP-CS0314/2020|EPI_ISL_862510|2020-06-29
TTTCACCGAGGCCATGCGGAGTACGATCGAGTGTACAGTGAAC
p> hCoV-19/South Korea/KDCA2188/2021|EPI_ISL_1319040|2021-03-06
TTTCACCGAGGCCACGCGTAGTACGATCTAGTGTACAGTGAAC
q> MW883290.1 Severe acute respiratory syndrome coronavirus 2
TTTCACCGAGGCCACACGGAGTACGATCGAGTGTACAGTGAAC
r>1XJR_1|Chain A|s2m RNA|null
GTTCACCGAGGCCACGCGGAGTACGAUCGAGGGTACAGTGAAT
```

Figure 2.1. 18 different s2m Sequences found in various h-COV19 sequences obtained from GISAID/NCBI.

2.2. Occupancy and Allele Frequency Calculation

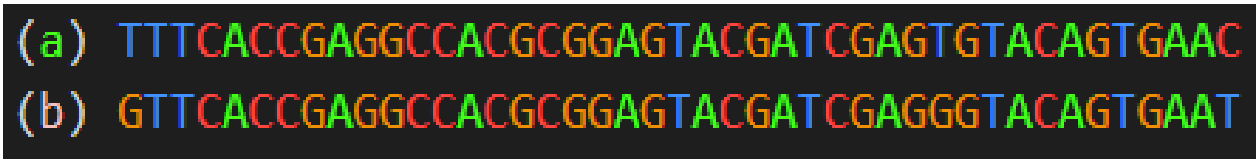


Figure 2.2. (a) The standard reference sequence for GISAID (EPI ISL 402124) is hCoV-19/Wuhan/WIV04/2019 (WIV04) human genome. (b) Experimental structure observed for s2m element.

The WIV04(402124) was selected because of its high-quality genome sequence and because it represented the accord of a few early COVID-19 betacoronavirus submissions.The S2M motifs (43 nucleotides long) were correlated with bat/pangolin models (Fig. 2.1(c) and human SARS-COVs models (Fig. 2.1(b)). A G to T amino acid transfer at position 32 and G to C/A/T amino acid transfer at position 16 have been discovered in human SARS-CoV-2 . A C to T amino acid transfer at position 6 (Fig. 2.1(b)). A G to C mutation happens at position 8(Fig. 2.1(k)). A G to T amino acid transfer at position 10 (Fig. 2.1(l), 2.1(n)). A C to T mutation was observed at position 15 and position 17(Fig. 2.1(o), 2.1(h), 2.1(i)).A G to T amino acid transfer at position 19 (Fig. 2.1(p)). G to C/A/T amino acid transfer at position 25 (Fig. 2.1(n), 2.1(e), 2.1(f)). A C to T mutation at position 28 (Fig. 4(g)). A G to T amino acid transfer at position 29 and position 38 (Fig. 2.1(p), 2.1(j)).

2.3. System Visualization

For Visualizing 19 different sequences with respect to the reference genome (402124). UCSC.2bit format and the Qi et al. Huffman coding system are used in Squiggle's DNA and RNA visualization method. A DNA sequence is translated to binary using a two-bit encoding scheme that maps T to 00, C to 01, A to 10, and G to 11 (6).

The following vectors are placed end to end, starting at the origin, for each bit:

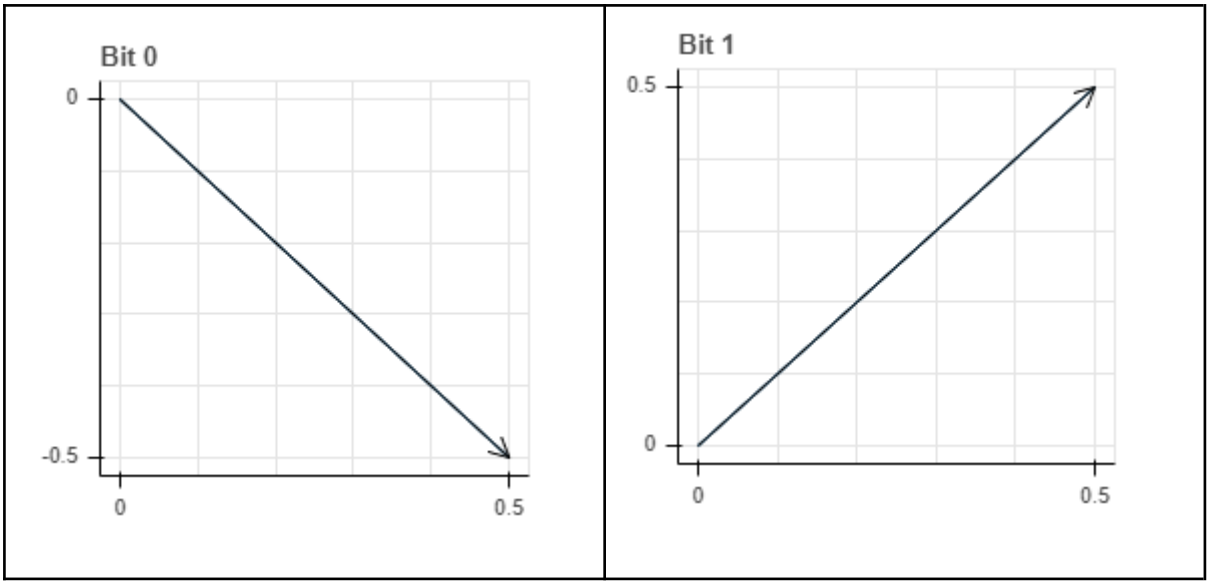
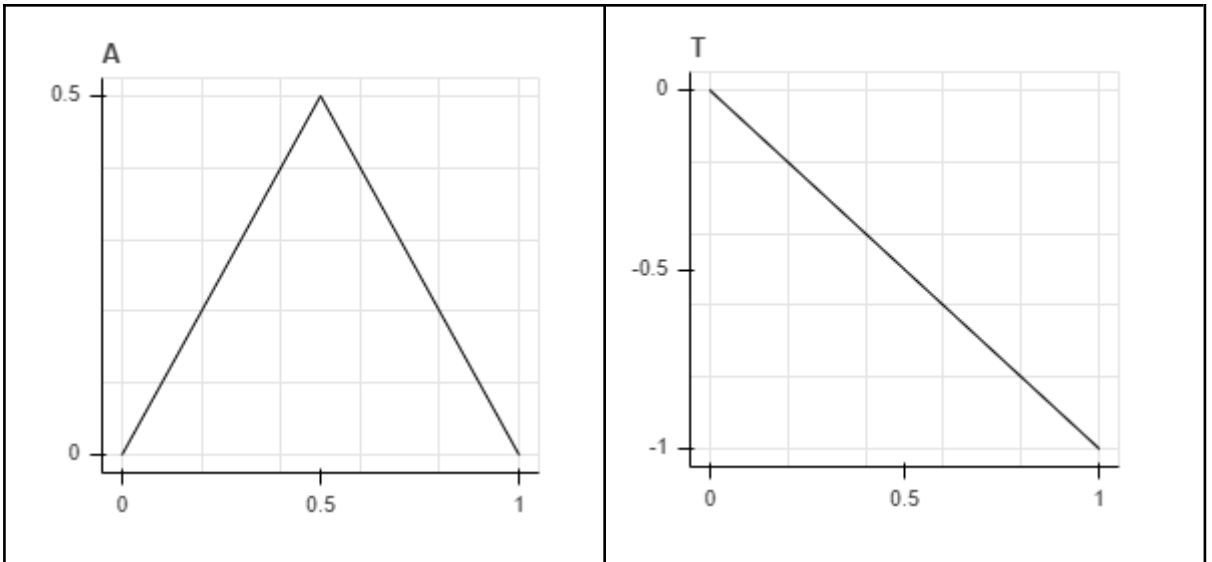


Figure 2.3.1. Representation of vector bit 0 and 1.

This encoding approach is a form of advantages in several manners.At a glance, the total GC content can be deduced based on if the graph's endpoint is above or below zero.The nucleotide position is directly represented on the x-axis.



A to 10 is encoded using a two-bit encoding scheme	T to 00 is encoded using a two-bit encoding scheme

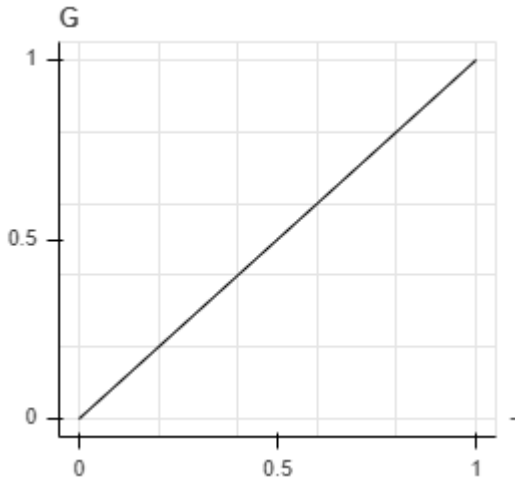
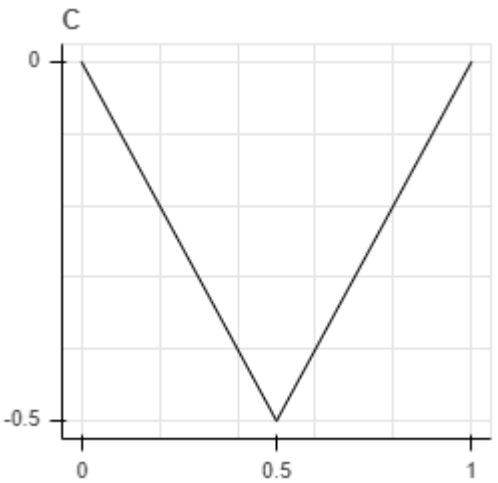
	
G to 11 is encoded using a two-bit encoding scheme	C to 01 is encoded using a two-bit encoding scheme

Figure 2.3.2. Each nucleotide takes on a distinct form as a result of this mapping.

2.4. Quantifying the Similarities between Motifs, Secondary Structures and Tertiary structures

2.4.1. Partition function foldings and dot plots:

The Boltzmann ensemble S contains the probability of the occurrence of a secondary structure. Provided an RNA sequence, the partition function Q sums all Boltzmann weighted free energies of all secondary systems that are possible. This allows base-pairing probabilities to be calculated for each potential pair of bases. As a result, the Vienna RNA Web servers use the algorithms in the Vienna RNA Package to generate a base-pairing probability matrix, also known as a dot plot.

2.4.2. Minimum free Energy Secondary Structure:

An RNA sequence's MFE structure is the secondary structure that contributes the least amount of free energy. A loop-based energy model and the dynamic programming algorithm proposed by Zuker et al. (7) are used to predict this structure. The loop-based energy model treats the free energy F(s) of an RNA secondary structure s as the number of the contributing free energies Fl of the loops L found in s since an RNA secondary structure can be uniquely decomposed into circles and external bases. The secondary structure that

minimizes F(s) is computed using the chosen energy parameter set and a given temperature (defaults to 37 °C).

2.4.3. Centroid Secondary structure:

An RNA sequence's centroid structure is the secondary structure in the Boltzmann ensemble with the least base-pair length to all other secondary structures.

2.4.4.Dot-Bracket Energy notation:

The Vienna RNA set's dot-bracket notation for an RNA secondary structure can classify pseudo-knot-free secondary structures. A string of matching brackets and dots of equal length is used to describe an RNA structure on an n-length sequence. Dots represent unpaired bases, while base pairs are identified by a '(' at a position, I and a ')' at position j.

3'UTR (29,543–29,891) RNA sequences were aligned with S2M motif queries in a BLASTn scan of the NCBI database (8) and also for GISAID. The RNAfold web server (http://rna.tbi.univie.ac.at/forna/) (9) was used to evaluate the 3' UTR stem-loop structures. Additionally, the S2M Motif PDB structure (PDB ID: 1XJR) was collected from the Protein Data Bank (http://www.rcsb.org/pdb). VMD was used to visualize all of the structures, and analysis was carried out as described in previous studies.The three-dimensional (3D) RNA structures were modeled using RNAComposer.

2.5. Guided Trees

For guided Phylogenetic trees, Clustal Omega 2.1was used to apply mBed algorithms. Multiple sequence alignments were carried out using this mBed algorithm.

2.6. Molecular Dynamics Simulation

NAMD 2.14 (10) for Win64-multicore-CUDA has been used in combination with the all-atom CHARMM36 All-Hydrogen Nucleic Acid Topology File(.rtf) force field across all MD simulations. I
started with the Tertiary structures(PDB) of all the 17 sequences and 1XJR(s2m motif). A PSF file, also recognized as a protein structure file, stores all of the molecule-specific data

needed to apply a force field to all the molecular systems created using the CHARMM36 All-Hydrogen Nucleic Acid Topology File with the help of VMD (11).

2.7..Energy Minimization of 3D structures

Energy minimization of 3d structures was done with CHARMM36 All-Hydrogen Nucleic Acid Parameter File(.prm) and minimization Config file (.namd). A log and dcd file was generated after the system's energy minimization simulation, and then it is further used for the analysis of simulation.

Chapter 3

Results

3.1. System Visualization in comparison with s2m motif(1XJR) and Reference genome WIV04/402124

Squiggle's plot was generated for each sequence in comparison with the s2m motif (PDB:1XJR). It was just plotted to see the mutation point in all 18 sequences.

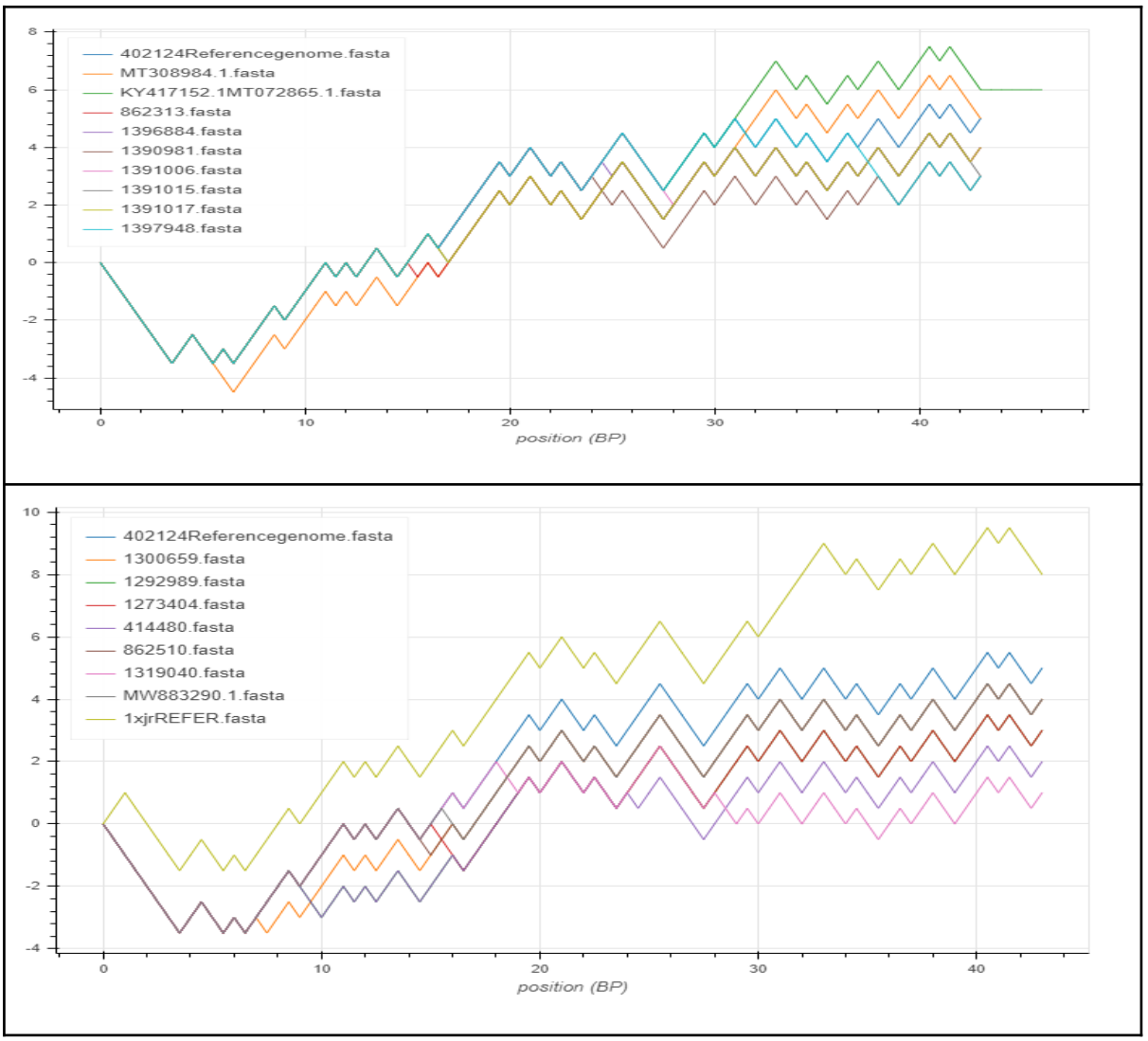


Figure 3.1 Mapping of all 18 nucleotides with reference 402124 (Reference genome for human) and PDB: 1XJR (Experimental structure)..

1XJR motif has G at position 1, G at position 32, and T at position 43 whereas, 402124 have T at position 1, T at position 32, and C at position 43.The 402124 motif has C at position 6, T at position 32, and C at position 43 whereas, MT308984.1 has T at position 6, G at position 32, and T at position 43.KY417152.1 and MT072865.1 have identical sequences. The 402124 motif has T at position 32 and C at position 43, whereas KY417152.1 and MT072865.1 have G at position 32 and T at position 43.The 402124 motif has G at position 16, whereas 862313 has C at position 16.The 402124 motif has G at position 25, whereas 1396884 has A at position 25.The 402124 motif has G at position 25, whereas 1390981 has T at position 25. The 402124 motif has G at position 28, whereas 1391006 has T at position 28.The 402124 motif has C at position 17 and C at position 43 whereas 1391015 has T at position 17 AND T at position 43.The 402124 motif has C at position 17 whereas 1391017 has T at position 17. The 402124 motif has G at position 38 whereas 1397948 has T at position 38.The 402124 motif has G at position 8 whereas 1300659 has C at position 8.The 402124 motif has G at position 10 whereas 1292989 has T at position 10.The 402124 motif has G at position 16 whereas 1273404 has T at position 16.The 402124 motif has G at position 10 and G at position 25 whereas 414480 has T at position 10 and C at position 25.The 402124 motif has C at position 15 whereas 862510 has T at position 15.The 402124 motif has G at position 19 and G at position 29 whereas 1319040 has T at position 19 and T at position 29.The 402124 motif has G at position 16 whereas MW883290.1 has A at position 16.

3.2. MFE and Centroid Secondary structures

An RNA sequence's MFE structure is the secondary structure that contributes the least amount of free energy. The centroid structure of an RNA sequence is the secondary structure in the Boltzmann ensemble with the shortest base pair gap to all other secondary structures. The Vienna Program is used to color base pairs based on base-pair probabilities computed with thermodynamic parameters. Colors ranging from red (low entropy, well-defined) to violet (high entropy, well-defined) are used to code the positional (high entropy, ill-defined) (14).

1.402124

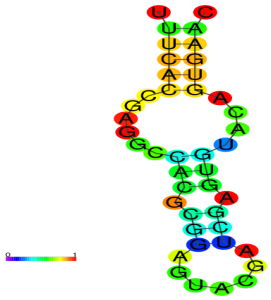
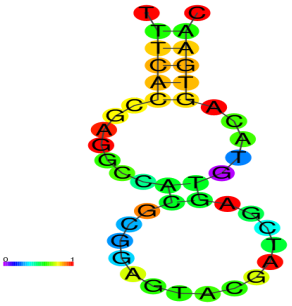
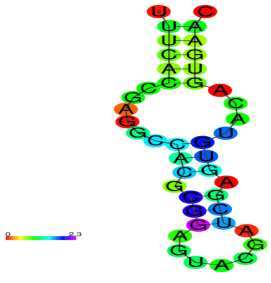
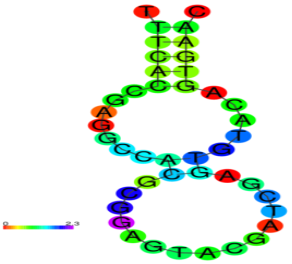
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.1 MFE and Centroid Secondary structure of motif 402124.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.98 kcal/mol.

The frequency of the MFE structure in the ensemble is 5.59 %.

The ensemble diversity is 10.24 .

2.MT308984.1

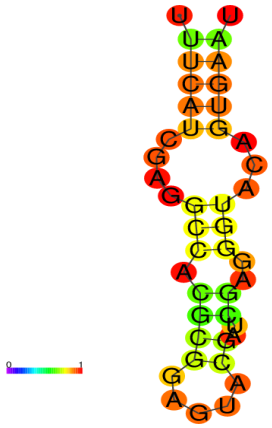
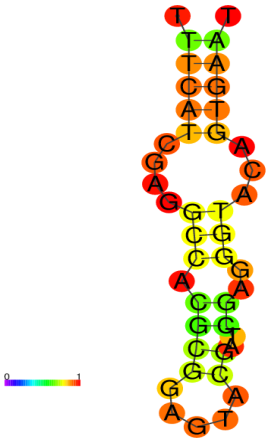
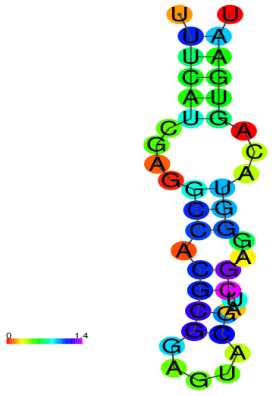
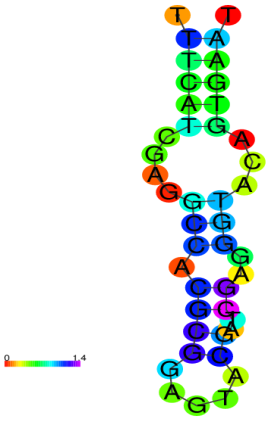
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.2 MFE and Centroid Secondary structure of motif MT308984.1.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.95 kcal/mol.

The frequency of the MFE structure in the ensemble is 34.83 %.

The ensemble diversity is 6.45 .

3.KY417152.1 & MT072865.1

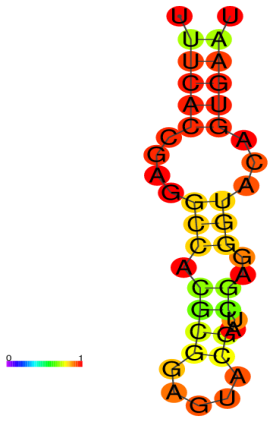
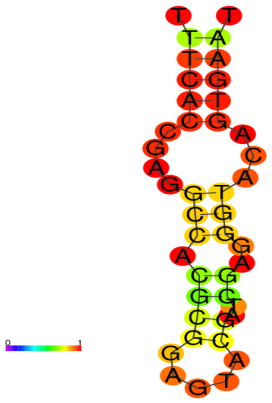
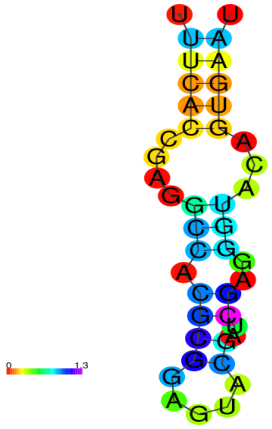
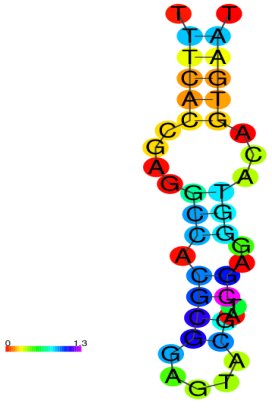
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.3 MFE and Centroid Secondary structure of motif KY417152.1 & MT072865.1.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -9.37 kcal/mol.

The frequency of the MFE structure in the ensemble is 39.51 %.

The ensemble diversity is 4.78 .

4.862313

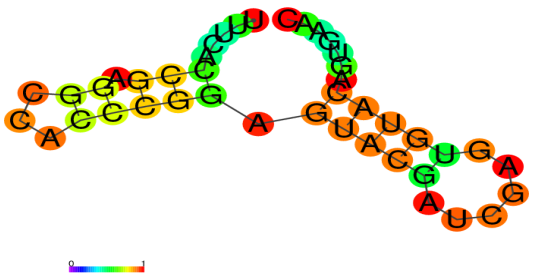
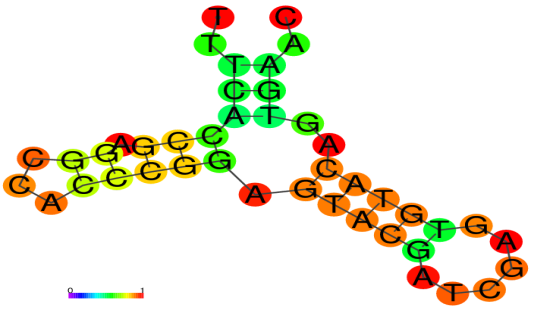
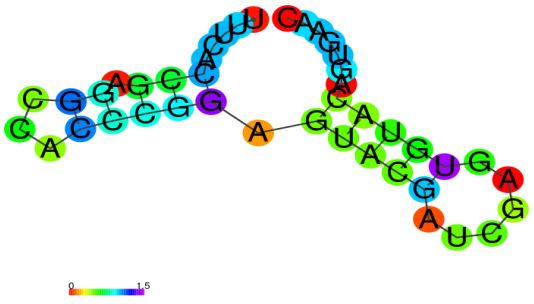
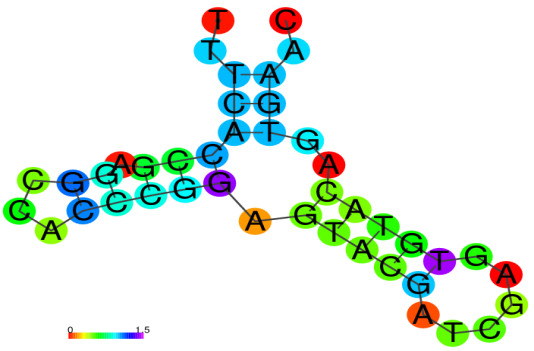
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.4 MFE and Centroid Secondary structure of motif 862313.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -8.61 kcal/mol.

The frequency of the MFE structure in the ensemble is 22.98 %.

The ensemble diversity is 6.88 .

5.1396884

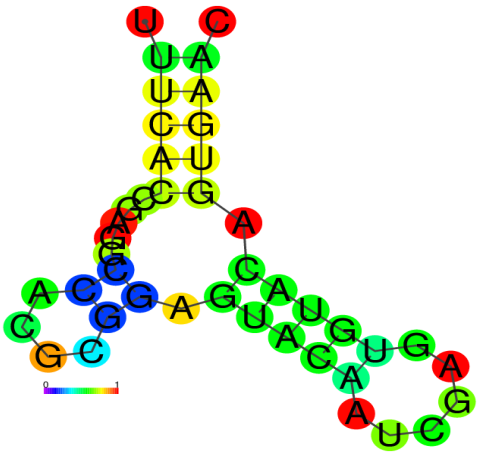
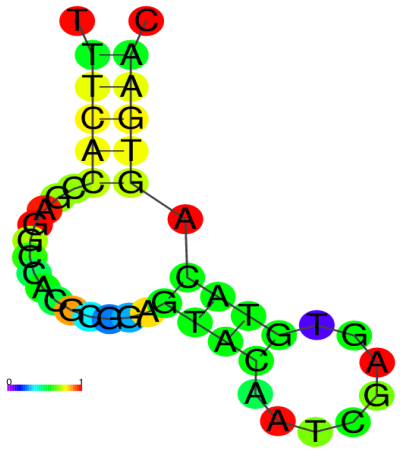
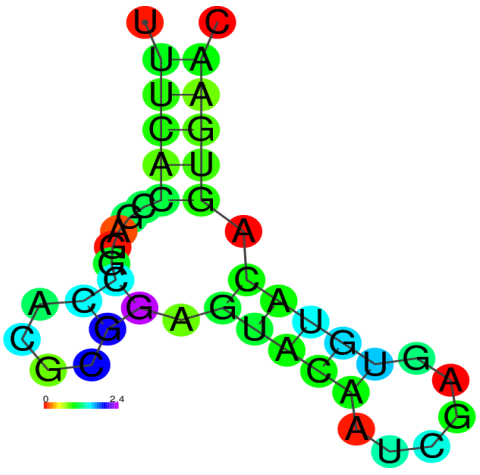
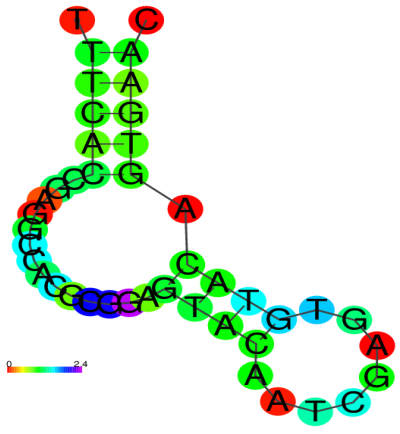
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.5 MFE and Centroid Secondary structure of motif 1396884.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -8.21 kcal/mol.

The frequency of the MFE structure in the ensemble is 11.96 %.

The ensemble diversity is 10.51 .

6.1390981

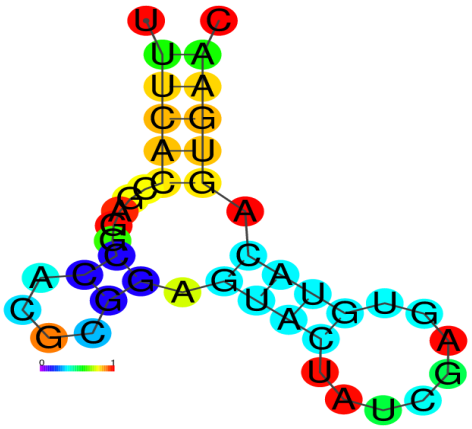
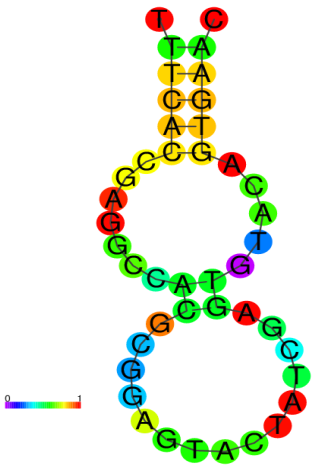
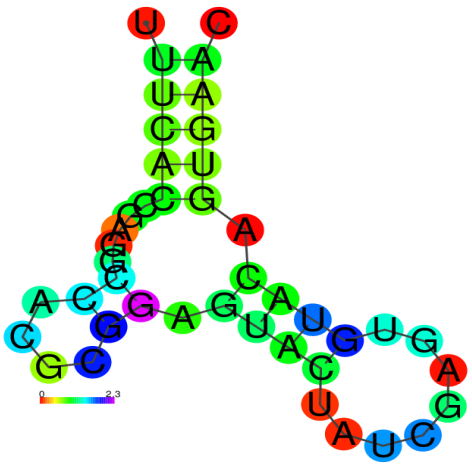
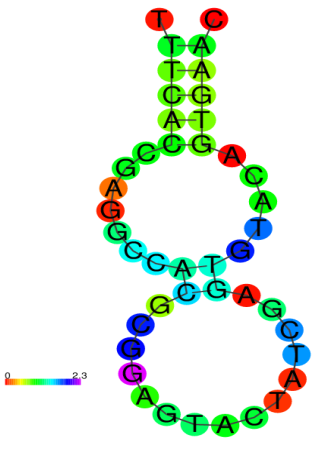
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.6 MFE and Centroid Secondary structure of motif 1390981.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.97 kcal/mol.

The frequency of the MFE structure in the ensemble is 9.14 %.

The ensemble diversity is 9.95 .

7.1391006

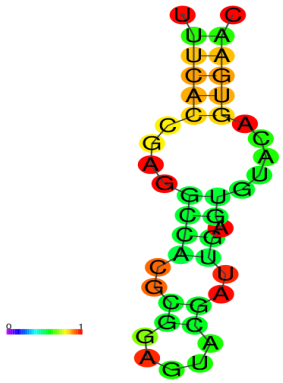
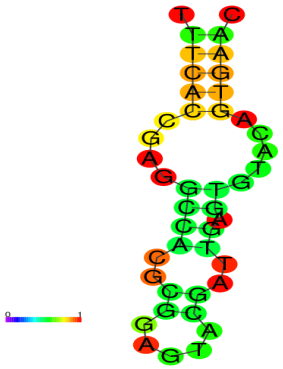
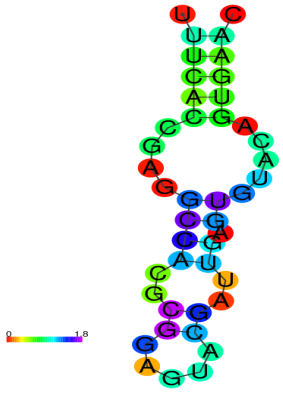
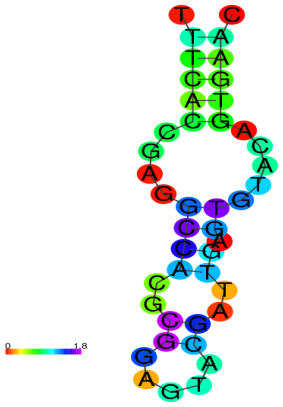
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.7 MFE and Centroid Secondary structure of motif 1391006.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.93 kcal/mol.

The frequency of the MFE structure in the ensemble is 36.01 %.

The ensemble diversity is 9.00 .

8.1391015

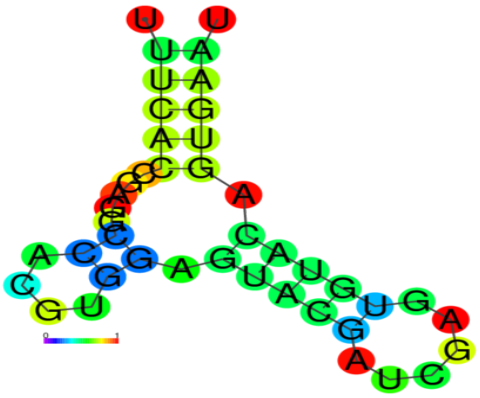
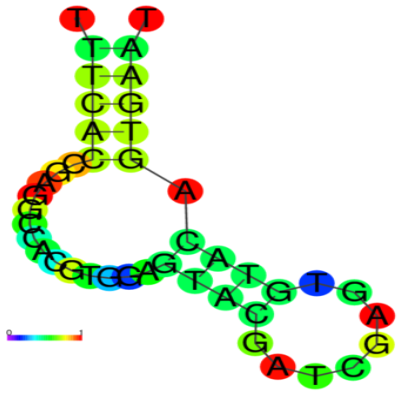
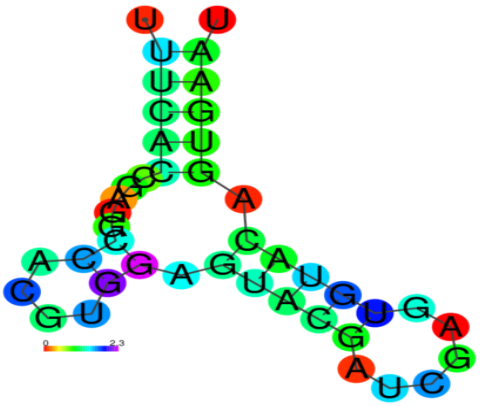
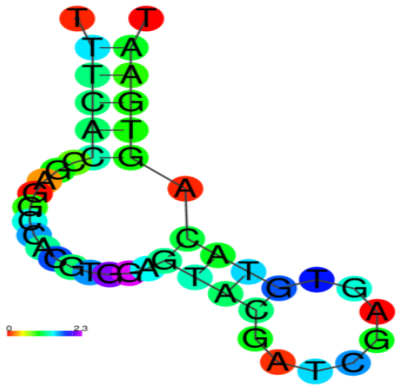
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.8 MFE and Centroid Secondary structure of motif 1391015.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.65 kcal/mol.

The frequency of the MFE structure in the ensemble is 11.15 %.

The ensemble diversity is 11.72 .

9.1391017

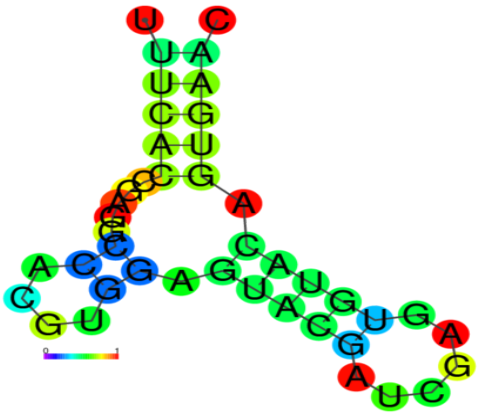
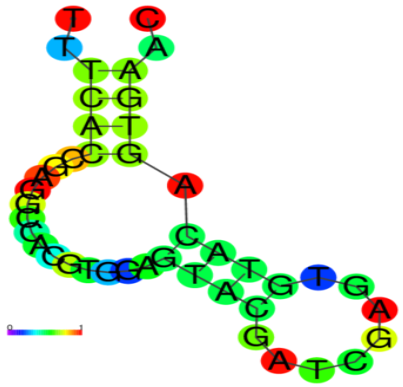
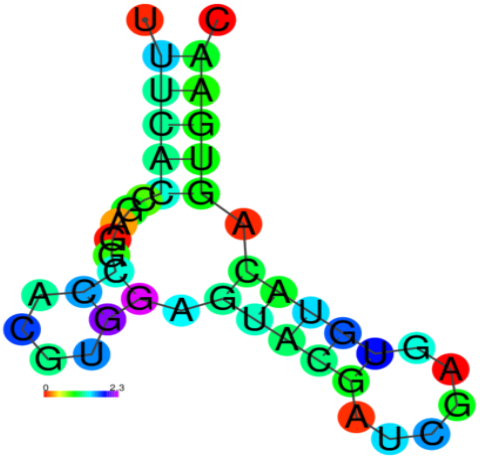
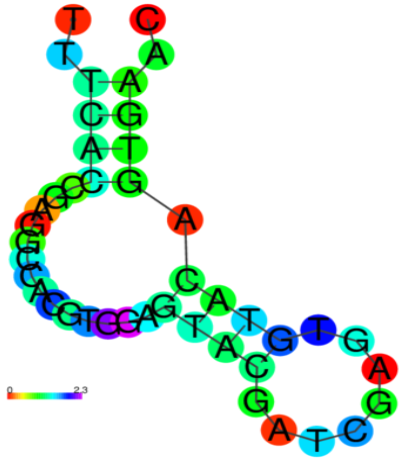
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.9 MFE and Centroid Secondary structure of motif 1391017.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.60 kcal/mol.

The frequency of the MFE structure in the ensemble is 10.30 %.

The ensemble diversity is 11.90 .

10.1397948

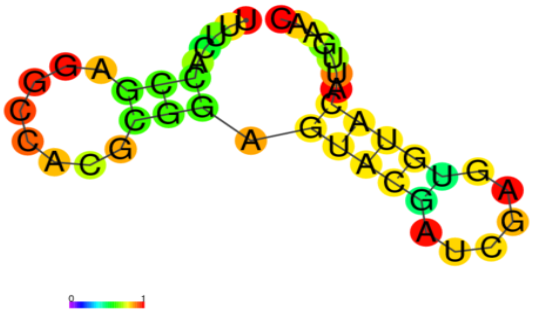
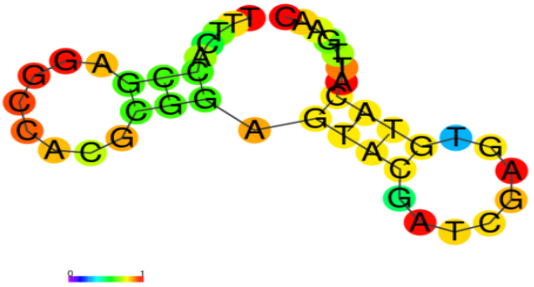
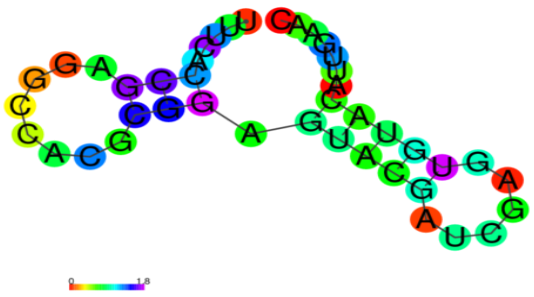
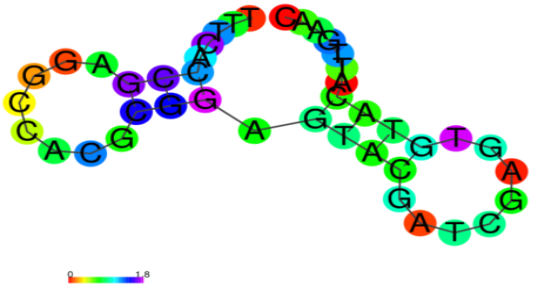
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.10 MFE and Centroid Secondary structure of motif 1397948.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -6.96 kcal/mol.

The frequency of the MFE structure in the ensemble is 24.66 %.

The ensemble diversity is 8.30 .

11.1300659

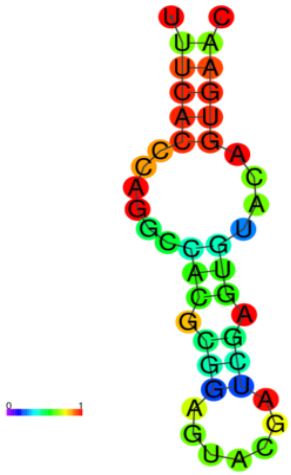
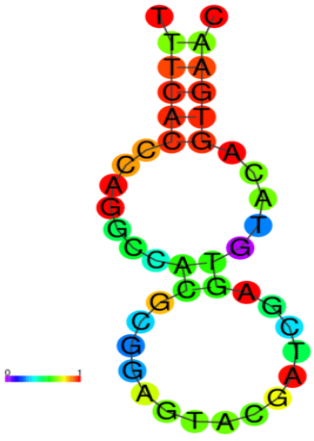
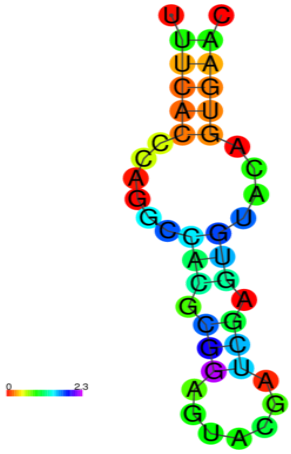
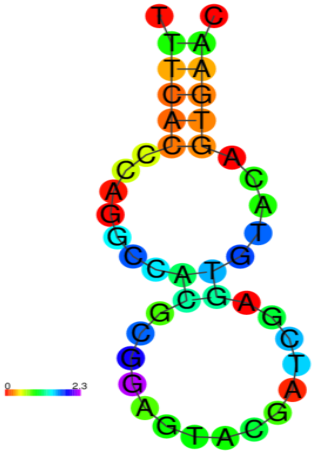
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.11 MFE and Centroid Secondary structure of motif 1300659.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.92 kcal/mol.

The frequency of the MFE structure in the ensemble is 6.15 %.

The ensemble diversity is 8.91 .

12.1292989

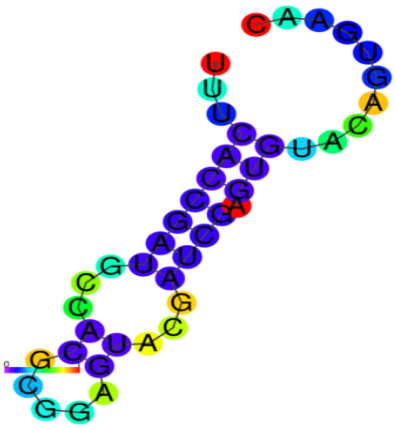
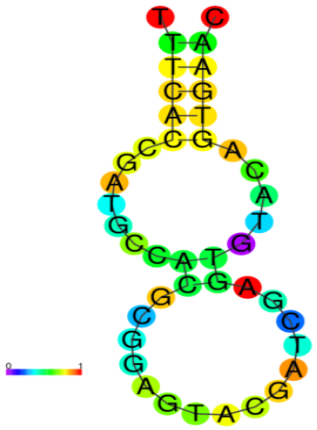
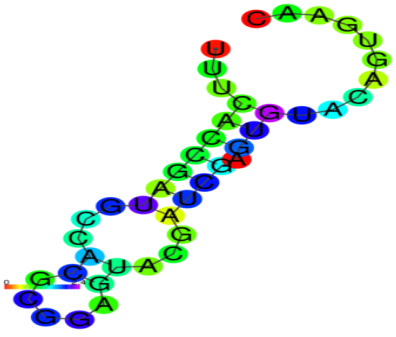
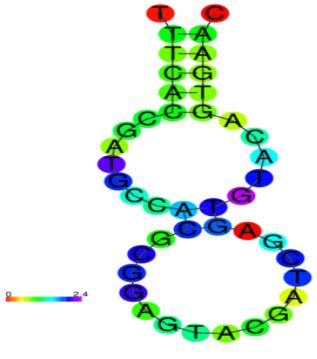
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.12 MFE and Centroid Secondary structure of motif 1292989.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -8.38 kcal/mol.

The frequency of the MFE structure in the ensemble is 7.70 %.

The ensemble diversity is 12.43 .

13.1273404

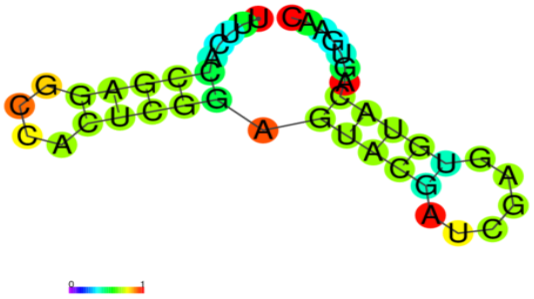
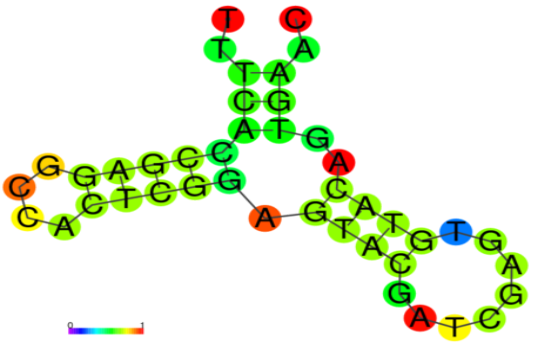
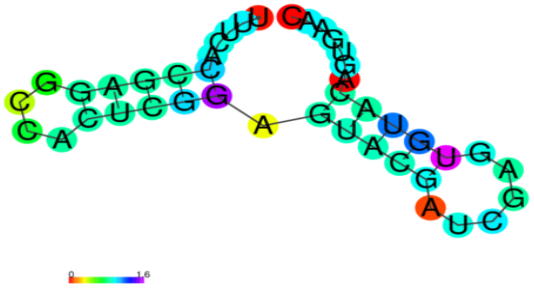
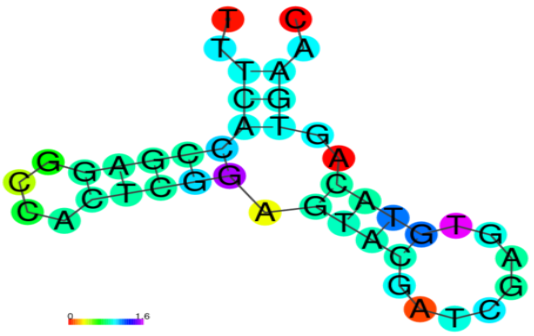
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.13 MFE and Centroid Secondary structure of motif 1273404.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -12.63 kcal/mol.

The frequency of the MFE structure in the ensemble is 22.16 %.

The ensemble diversity is 10.04 .

14.414480

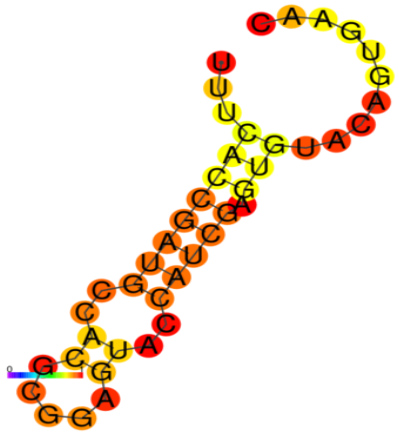
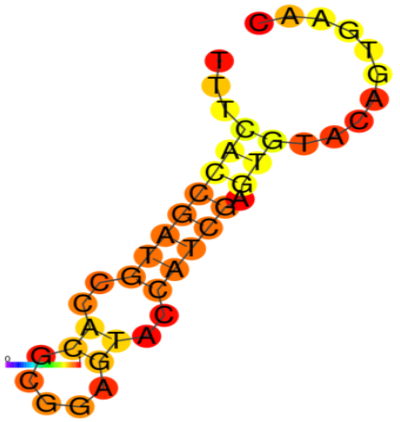
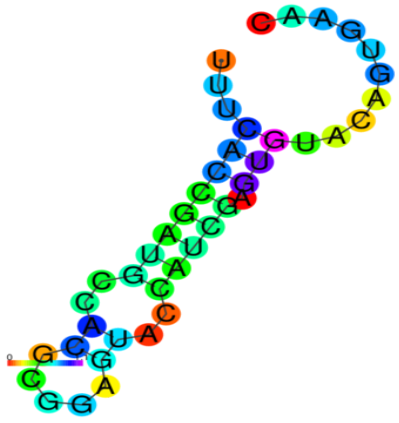
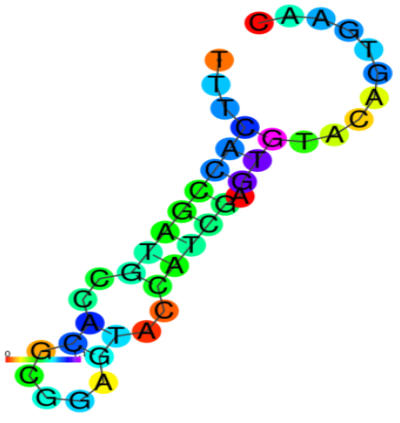
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.14 MFE and Centroid Secondary structure of motif 414480.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -9.64 kcal/mol.

The frequency of the MFE structure in the ensemble is 68.07 %.

The ensemble diversity is 5.43 .

15.862510

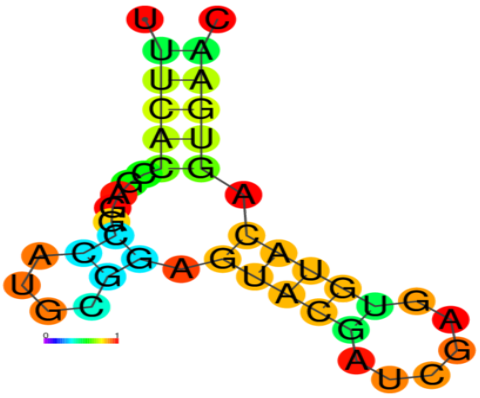
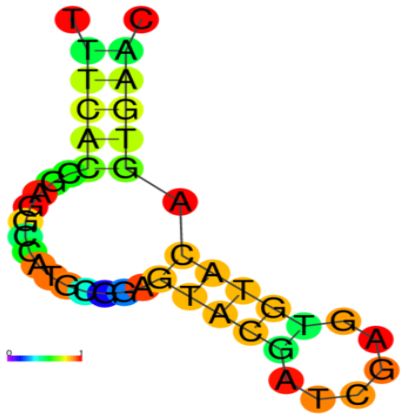
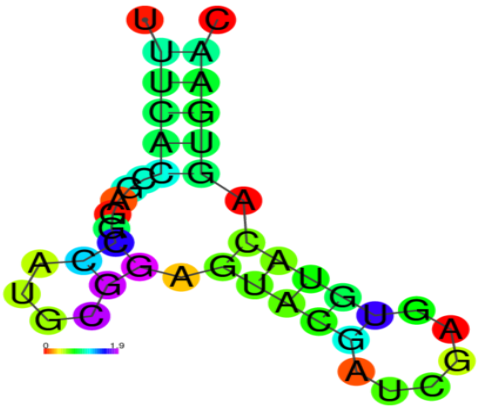
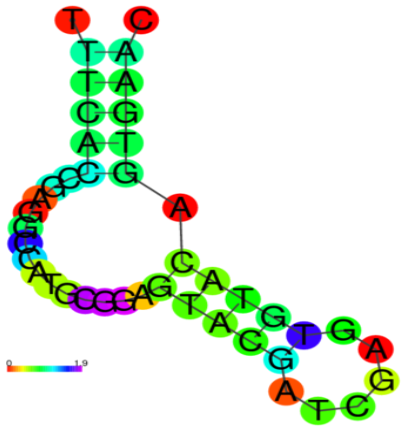
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.15 MFE and Centroid Secondary structure of motif 862510.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.39 kcal/mol.

The frequency of the MFE structure in the ensemble is 14.49 %.

The ensemble diversity is 7.88 .

16.1319040

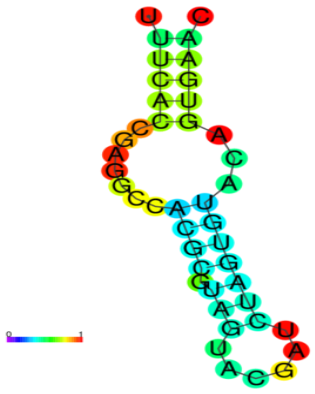
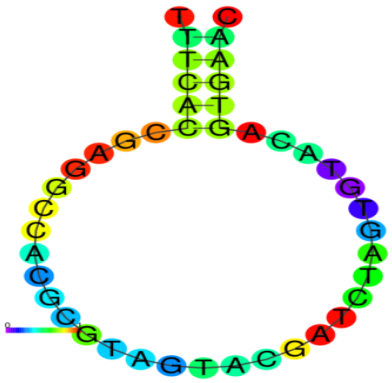
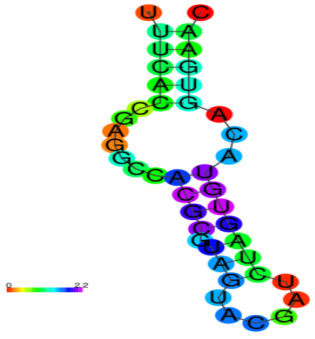
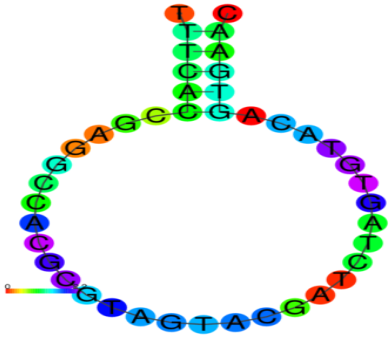
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.16 MFE and Centroid Secondary structure of motif 1319040.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.19 kcal/mol.

The frequency of the MFE structure in the ensemble is 20.09 %.

The ensemble diversity is 12.70 .

17.MW883290.1

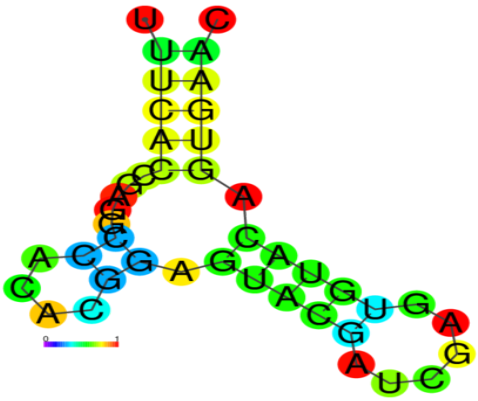
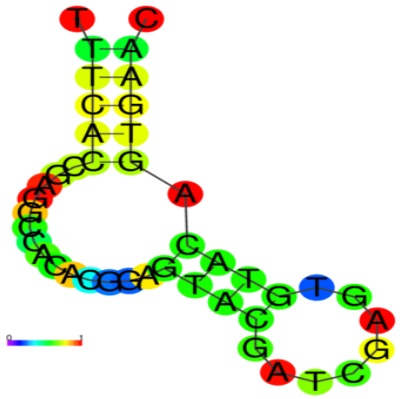
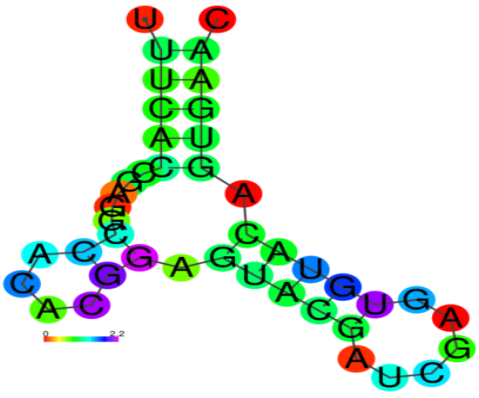
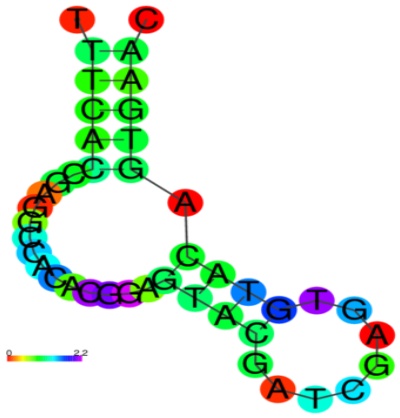
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.17 MFE and Centroid Secondary structure of motif MW883290.1.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -7.49 kcal/mol.

The frequency of the MFE structure in the ensemble is 12.43 %.

The ensemble diversity is 10.56 .

18. PDB: 1XJR (Experimental structure of s2m motif)

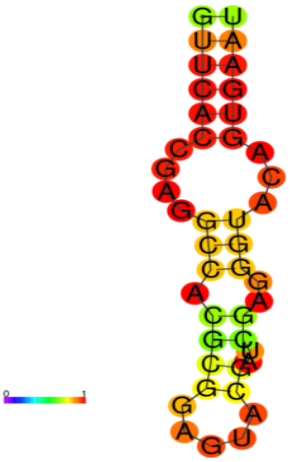
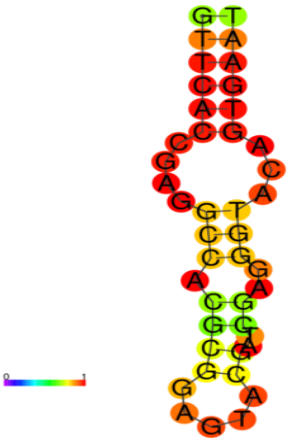
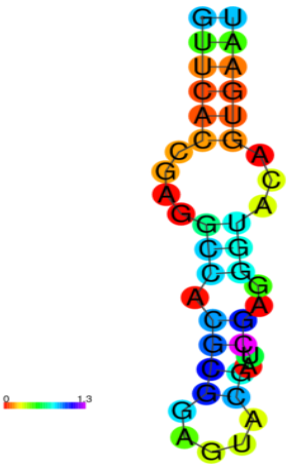
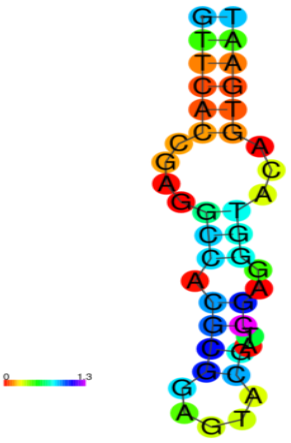
	
Drawing of an MFE structure encoding base-pair probabilities	Drawing a centroid structure to encode base-pair probabilities
	
Positional entropy is encoded by MFE structure drawing.	Positional entropy is encoded by drawing a centroid structure.

Figure 3.2.18 MFE and Centroid Secondary structure of PDB:1XJR.

Results for thermodynamic ensemble prediction:

The free energy of the thermodynamic ensemble is -9.99 kcal/mol.

The frequency of the MFE structure in the ensemble is 38.59 %.

The ensemble diversity is 4.63 .

3.3. Tertiary Structures

RNAcomposer was used to model 3D Structures of all 18 motifs. The Dot-bracket notation was given to specify sequence and secondary structure. The secondary structure is divided into fragments of overlapping canonical base pairs, according to the graph representation. Those more minor secondary structure elements are linked to the 3D structure elements that they belong to.

1.402124

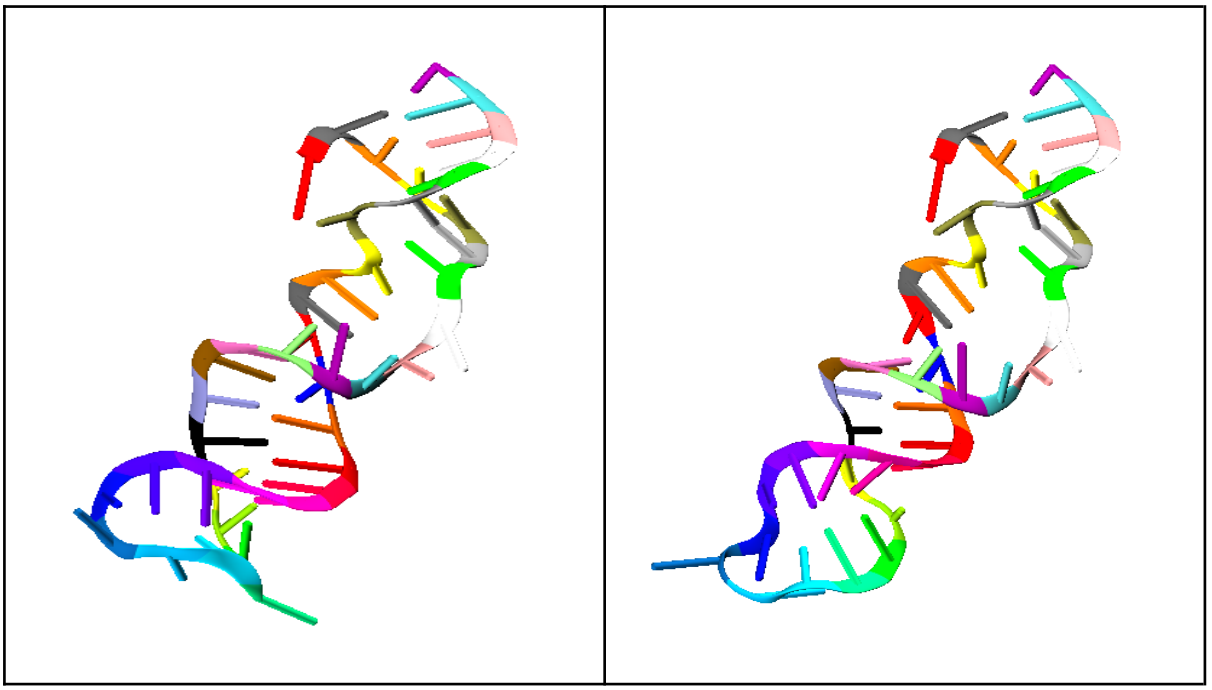


Figure 3.3.1 3D structure of 402124 Left (optimal structure) and Right (Centroid Structure) motif.

2.MT308984.1

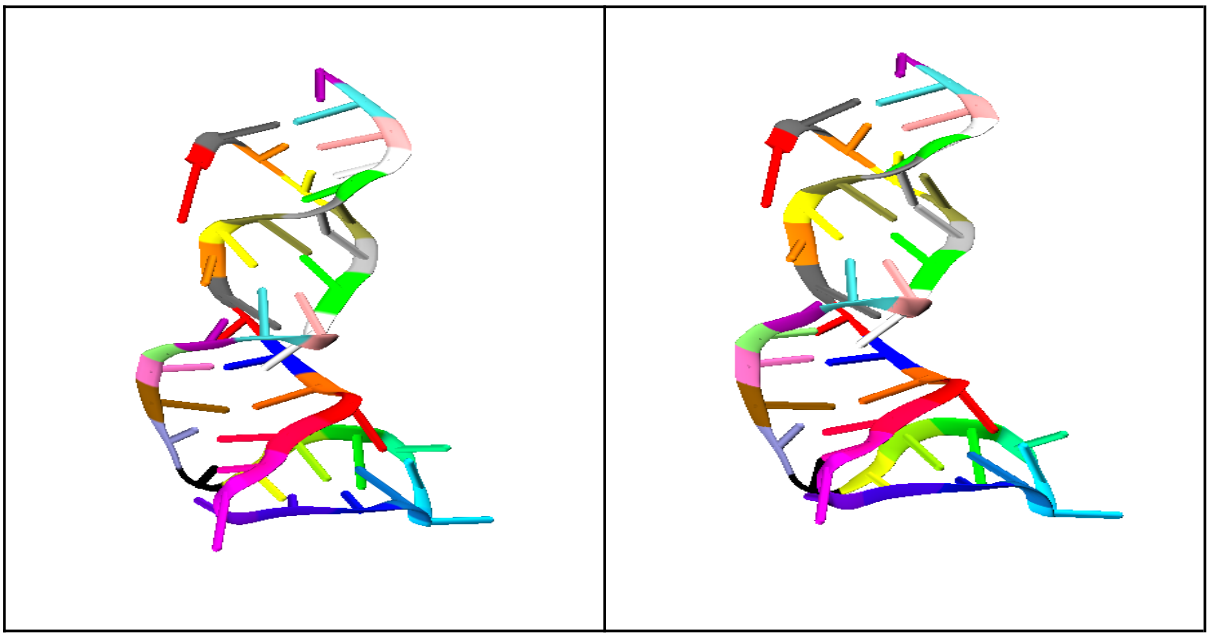


Figure 3.3.2 3D structure of MT308984.1 Left (optimal structure) and Right (Centroid Structure) motif.

3.KY417152.1 & MT072865.1

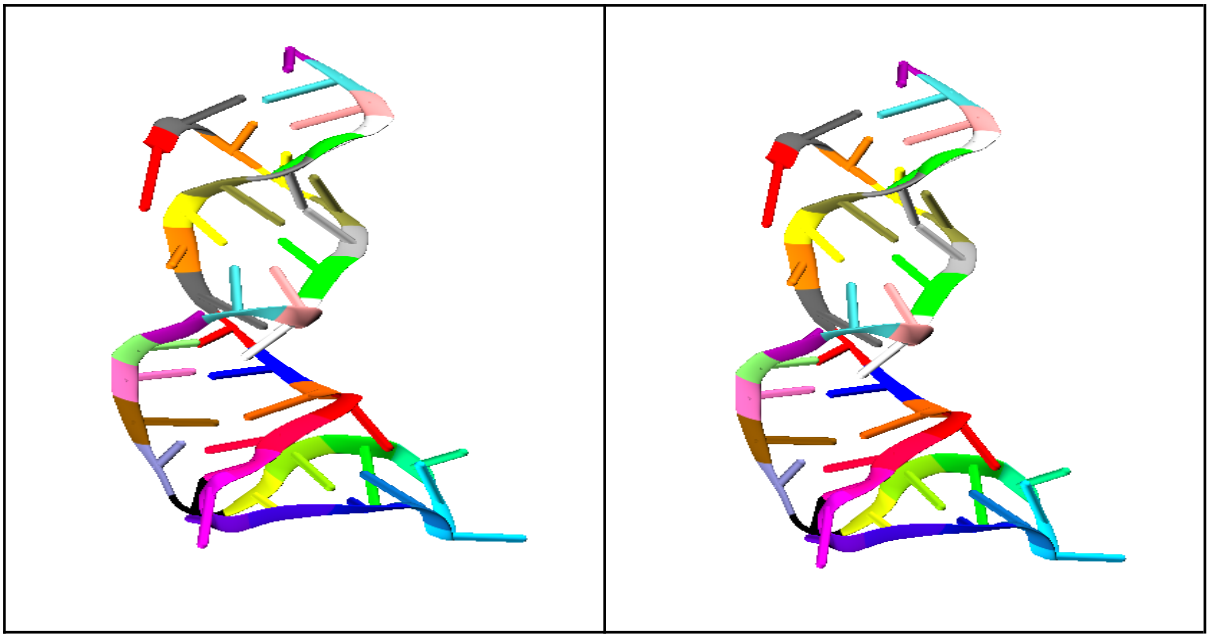


Figure 3.3.3 3D structure of KY417152.1 & MT072865.1 Left (optimal structure) and Right (Centroid Structure) motif.

4.862313

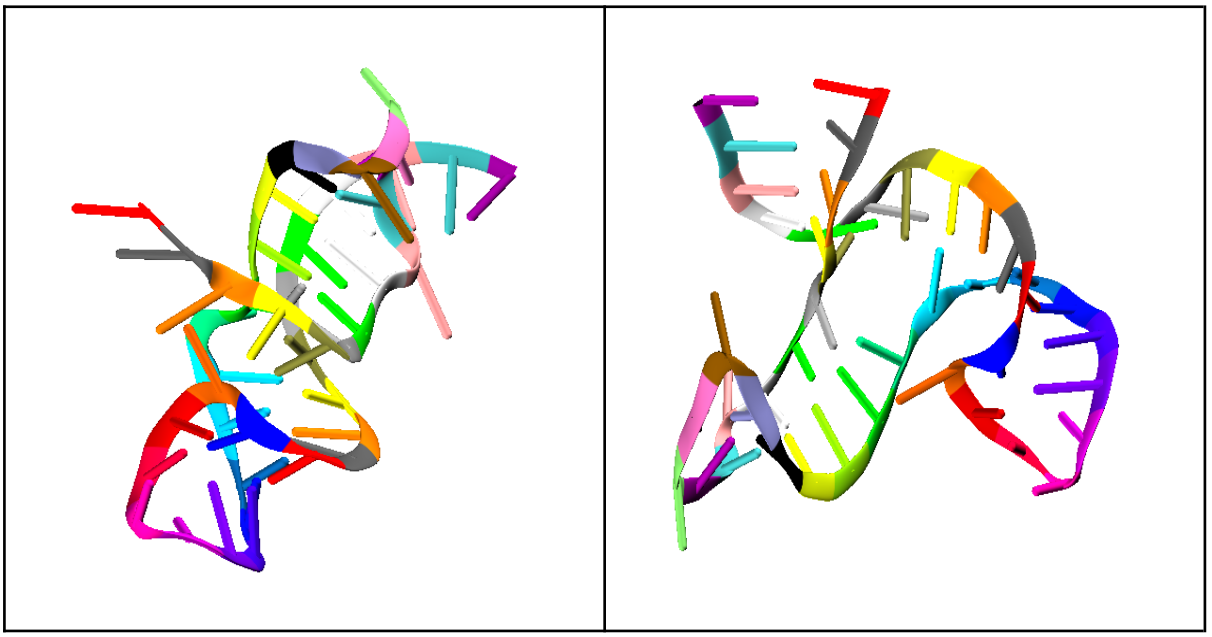


Figure 3.3.4 3D structure of 862313 Left (optimal structure) and Right (Centroid Structure) motif.

5.1396884

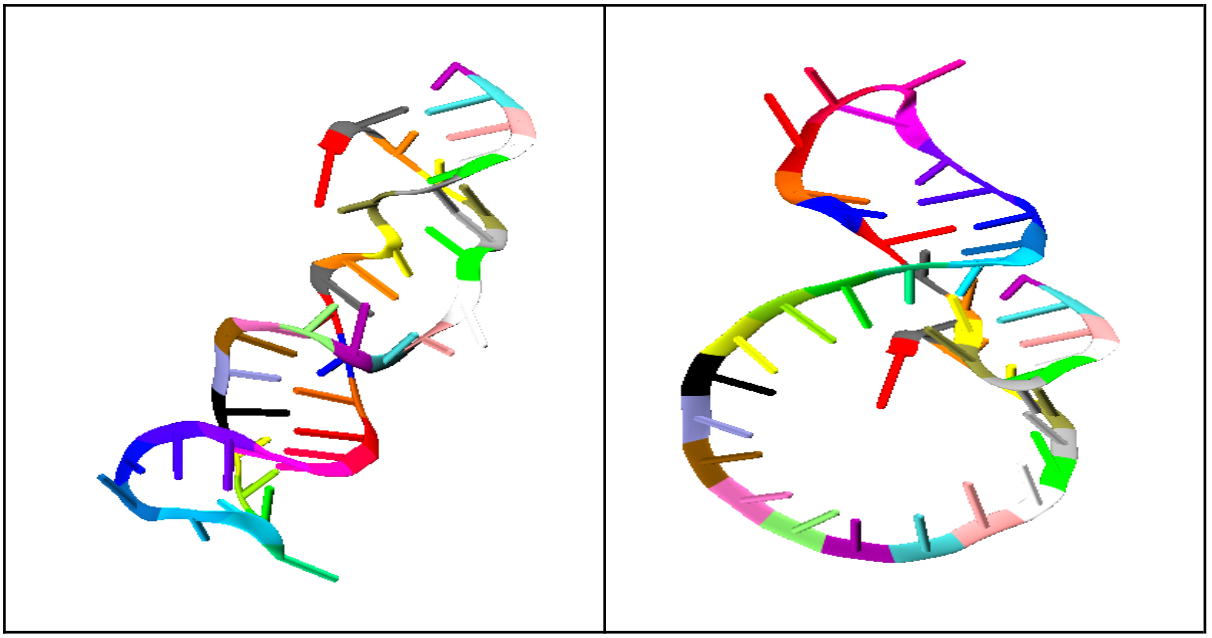


Figure 3.3.5 3D structure of 1396884 Left (optimal structure) and Right (Centroid Structure) motif.

6.1390981

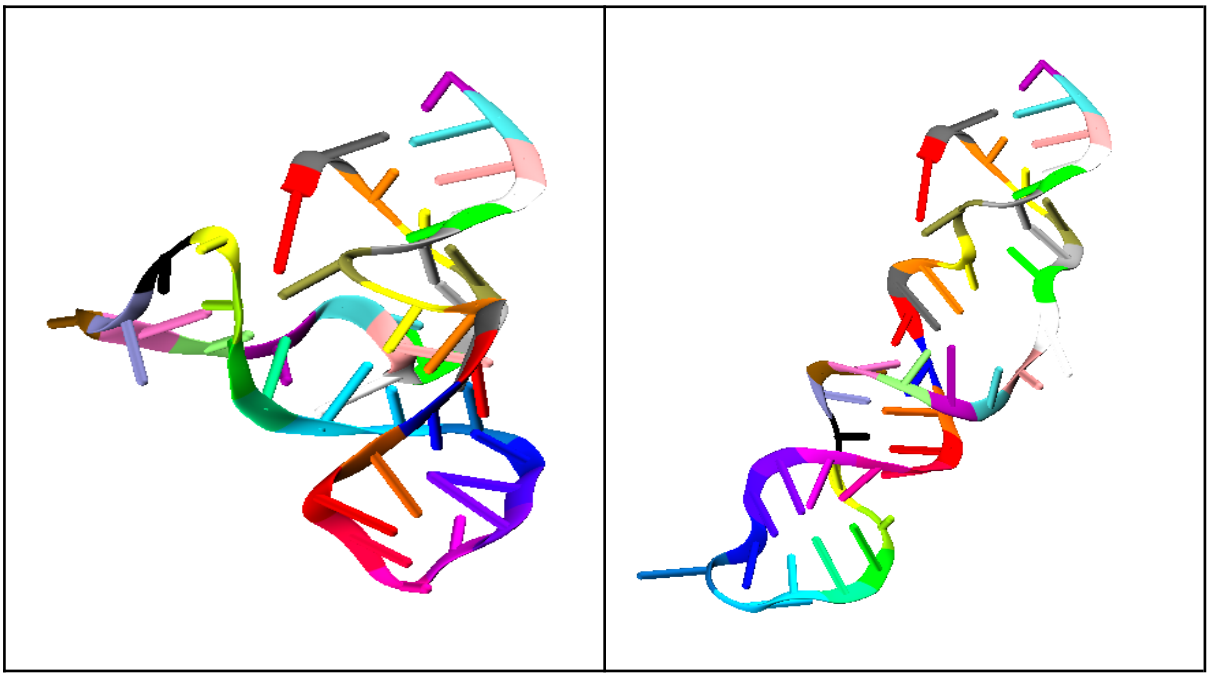


Figure3.3.6 3D structure of 1390981 Left (optimal structure) and Right (Centroid Structure) motif.

7.1391006

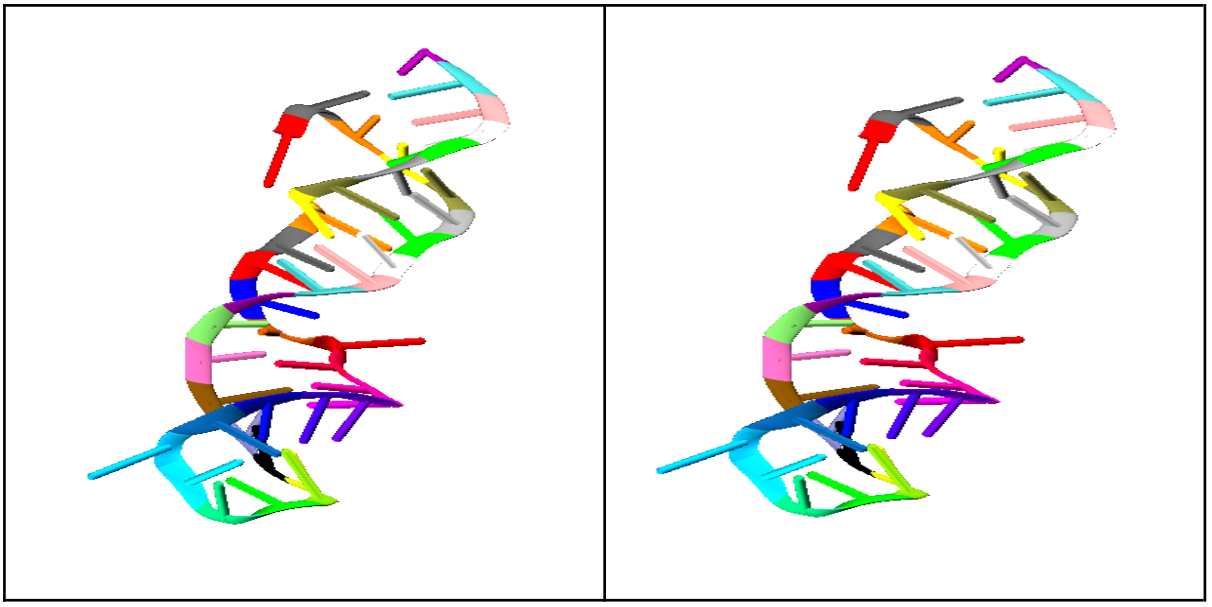


Figure3.3.7 3D structure of 1391006 Left (optimal structure) and Right (Centroid Structure) motif.

8.1391015

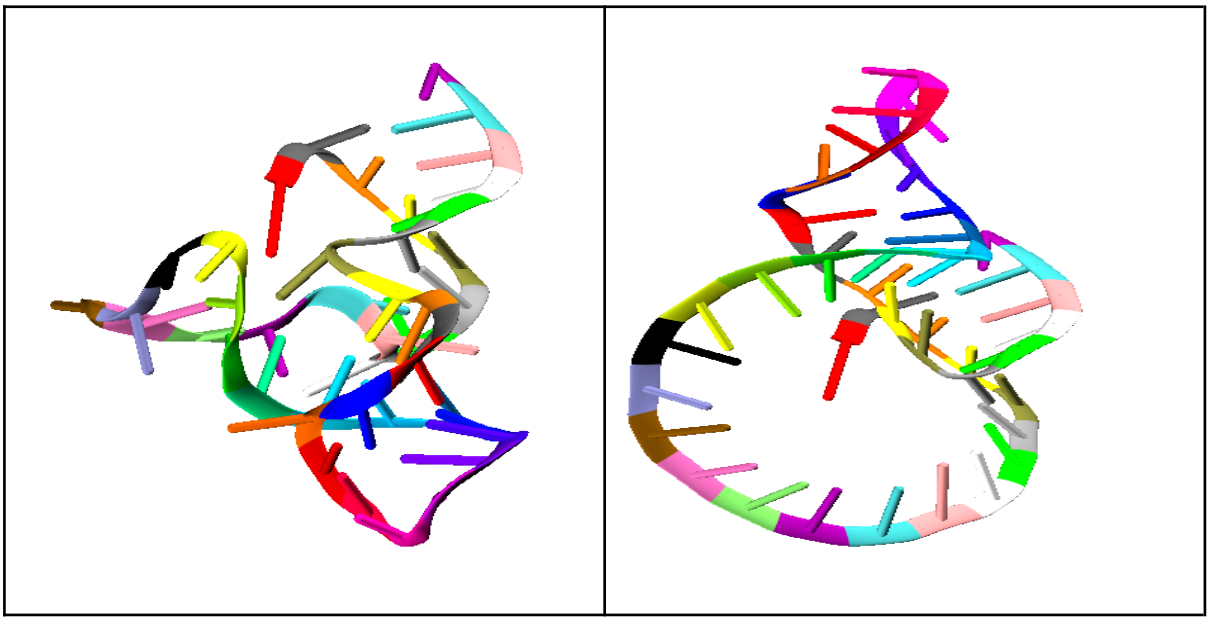


Figure3.3.8 3D structure of 1391015 Left (optimal structure) and Right (Centroid Structure) motif.

9.1391017

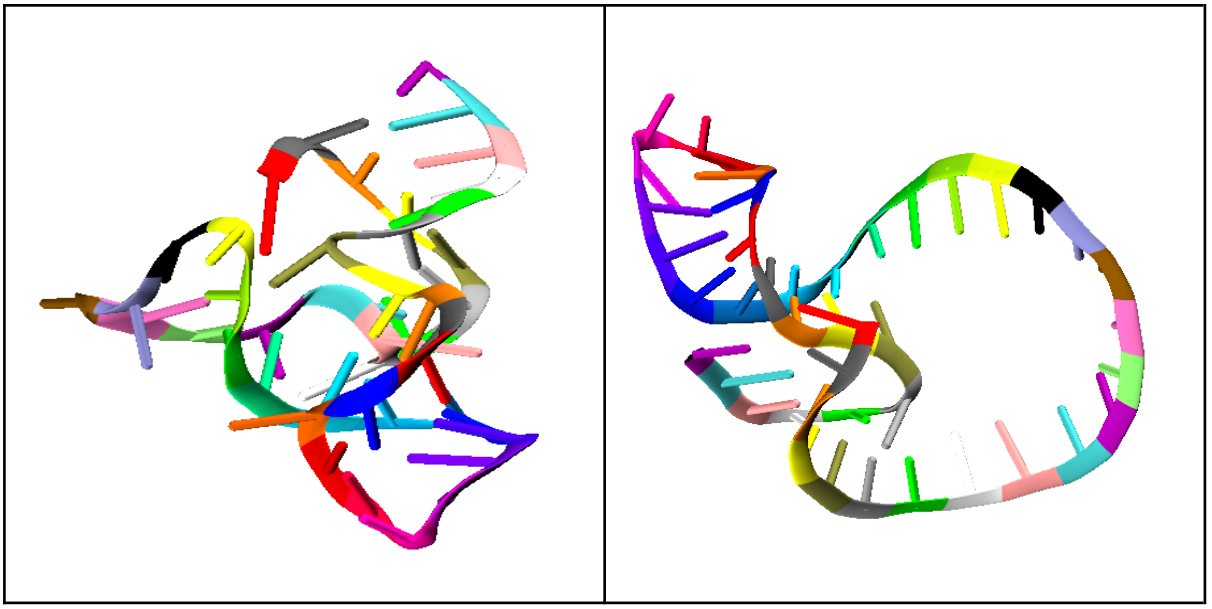


Figure3.3.9 3D structure of 1391017 Left (optimal structure) and Right (Centroid Structure) motif.

10.1397948

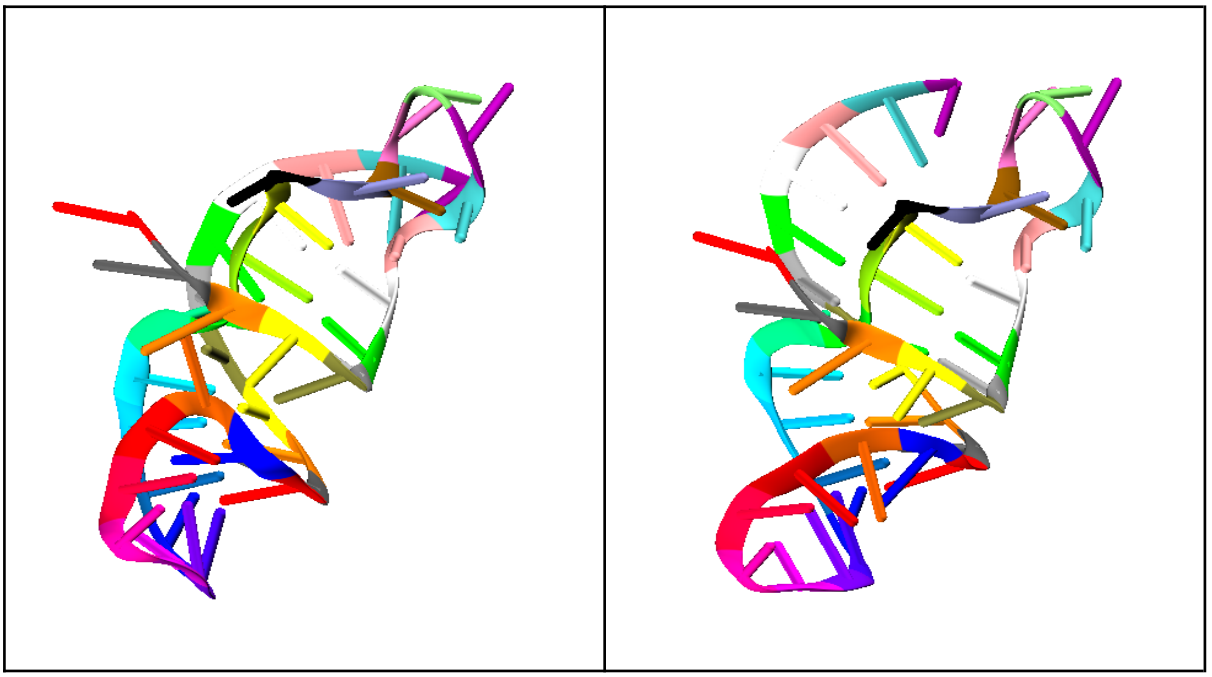


Figure3.3.10 3D structure of 1397948 Left (optimal structure) and Right (Centroid Structure) motif.

11.1300659

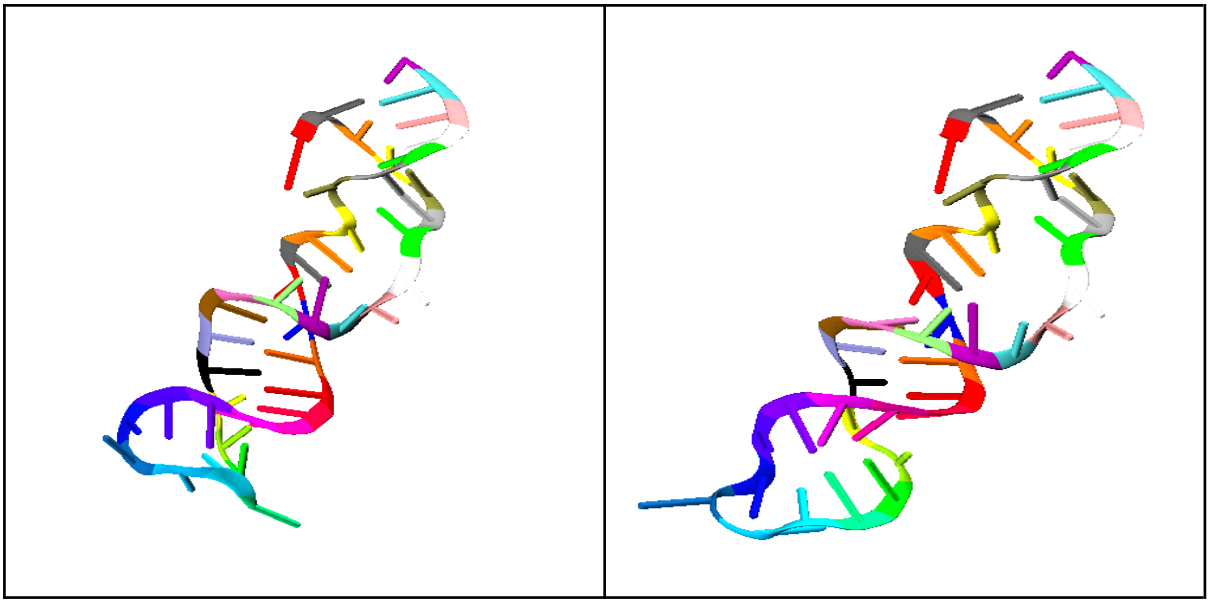


Figure3.3.11 3D structure of 1300659 Left (optimal structure) and Right (Centroid Structure) motif.

12.1292989

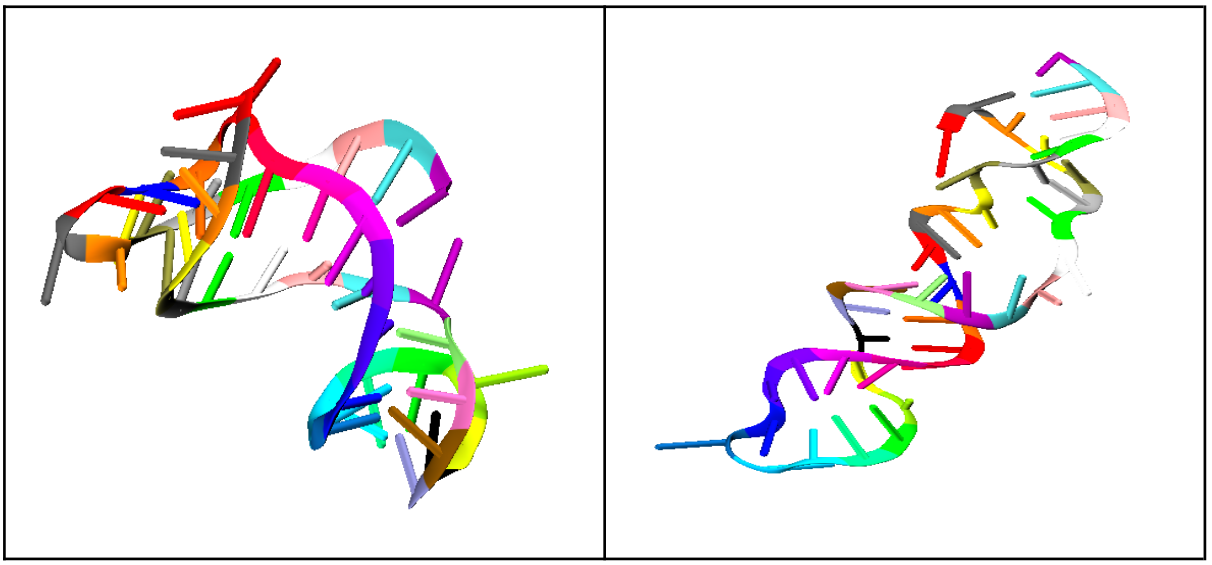


Figure3.3.12 3D structure of 1292989 Left (optimal structure) and Right (Centroid Structure) motif.

13.1273404

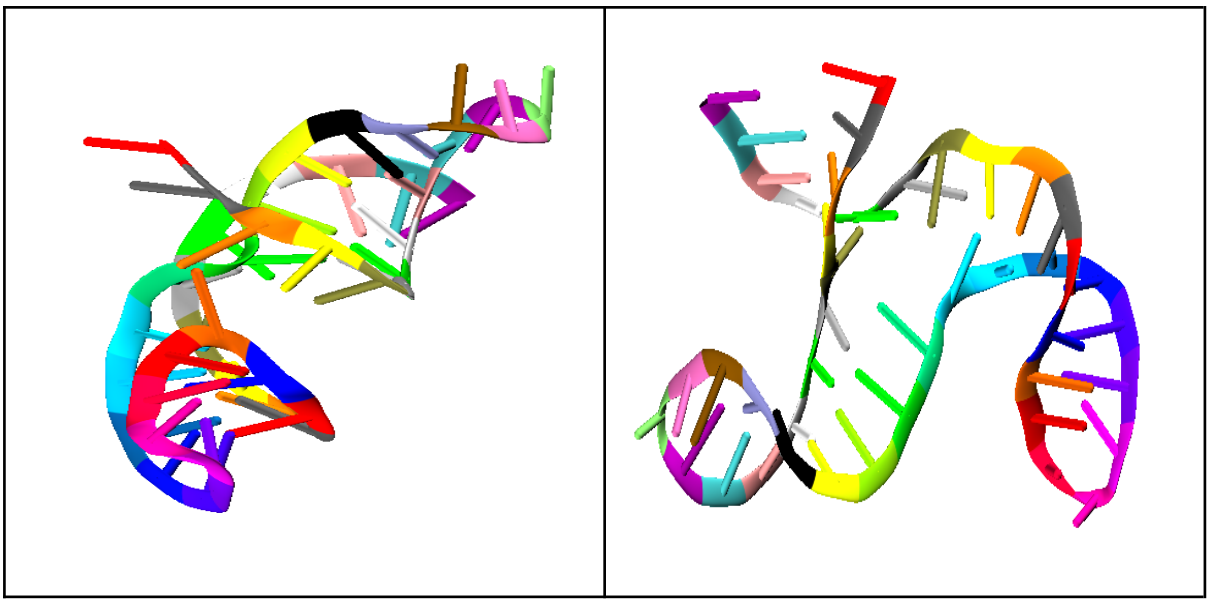


Figure3.3.13 3D structure of 1273404 Left (optimal structure) and Right (Centroid Structure) motif.

14.414480

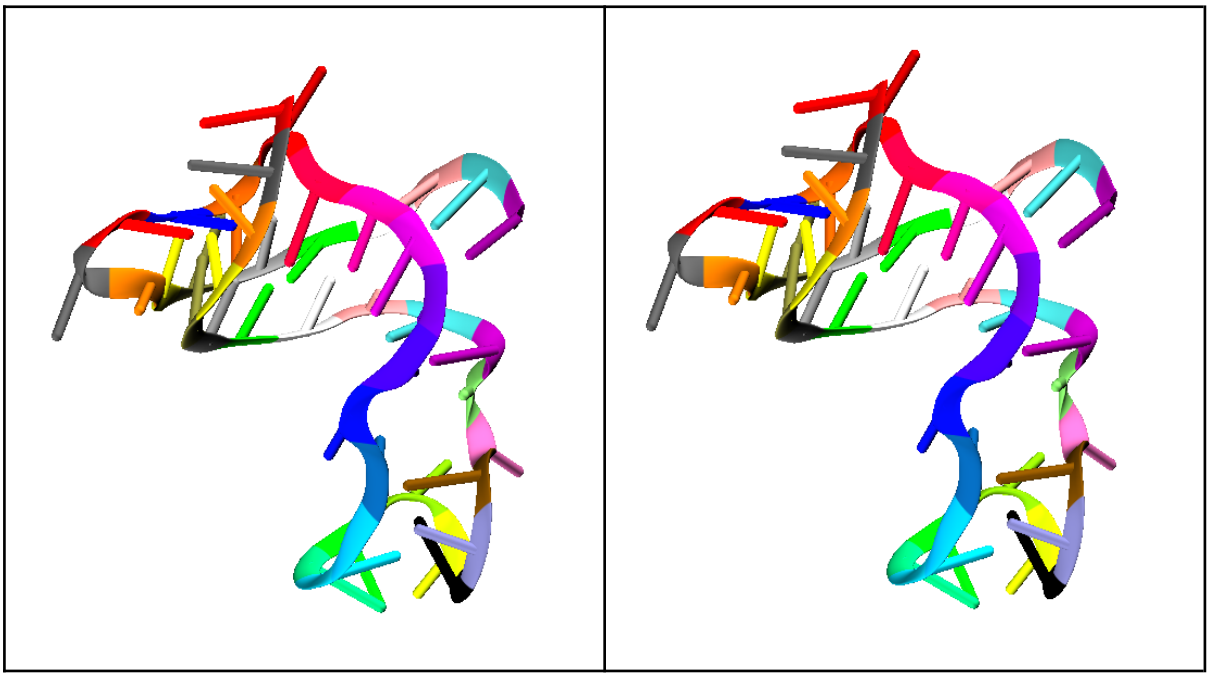


Figure3.3.14 3D structure of 414480 Left (optimal structure) and Right (Centroid Structure) motif.

15.862510

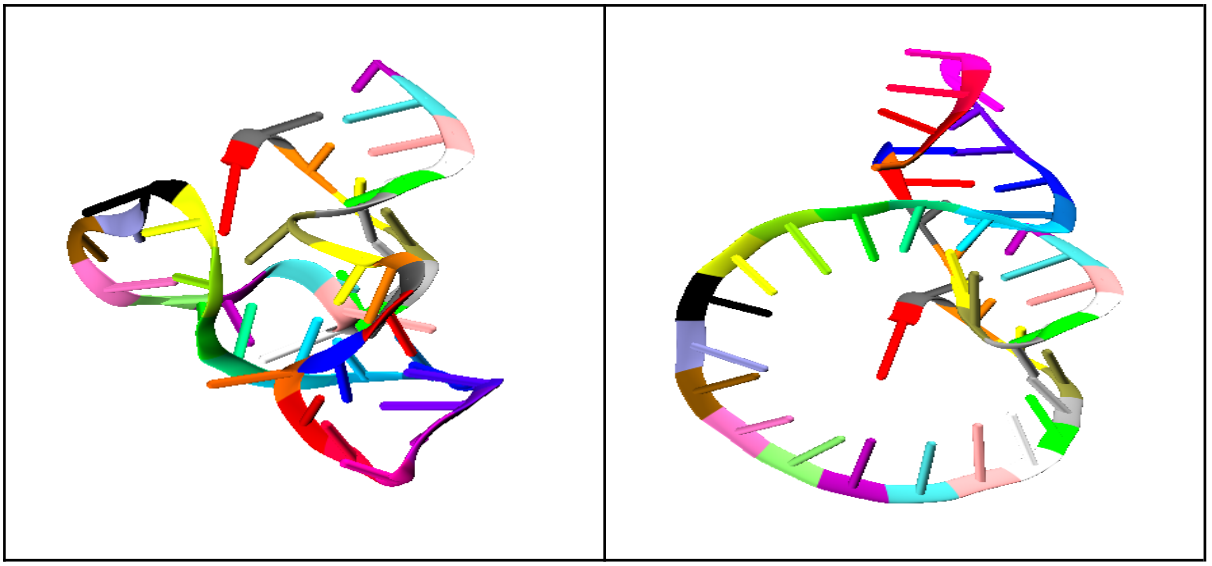


Figure3.3.15 3D structure of 862510 Left (optimal structure) and Right (Centroid Structure) motif.

16.1319040

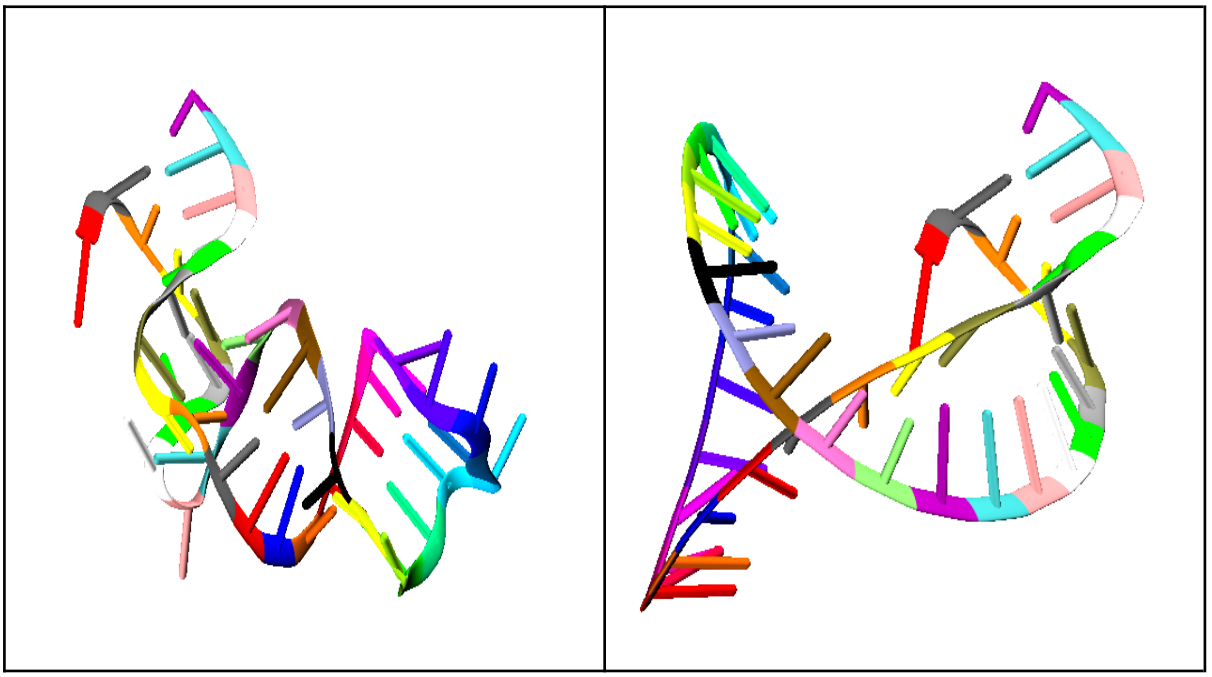


Figure3.3.16 3D structure of 1319040 Left (optimal structure) and Right (Centroid Structure) motif.

17.MW883290.1

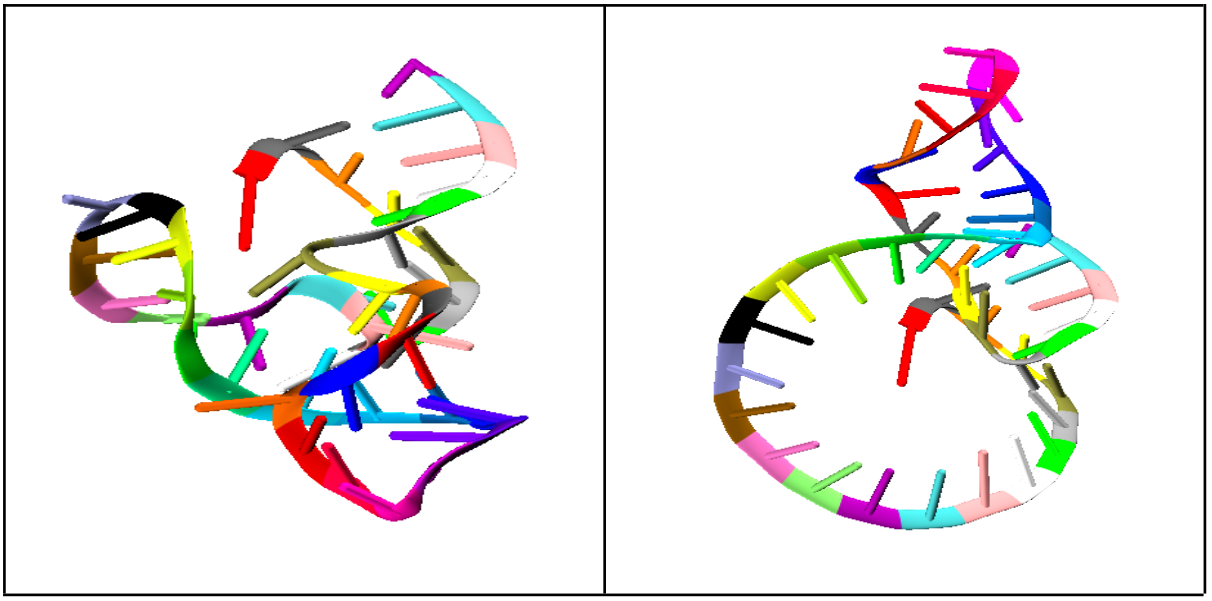


Figure3.3.17 3D structure of MW883290.1 Left (optimal structure) and Right (Centroid Structure) motif.

18. PDB : 1xjr Experimental Structure

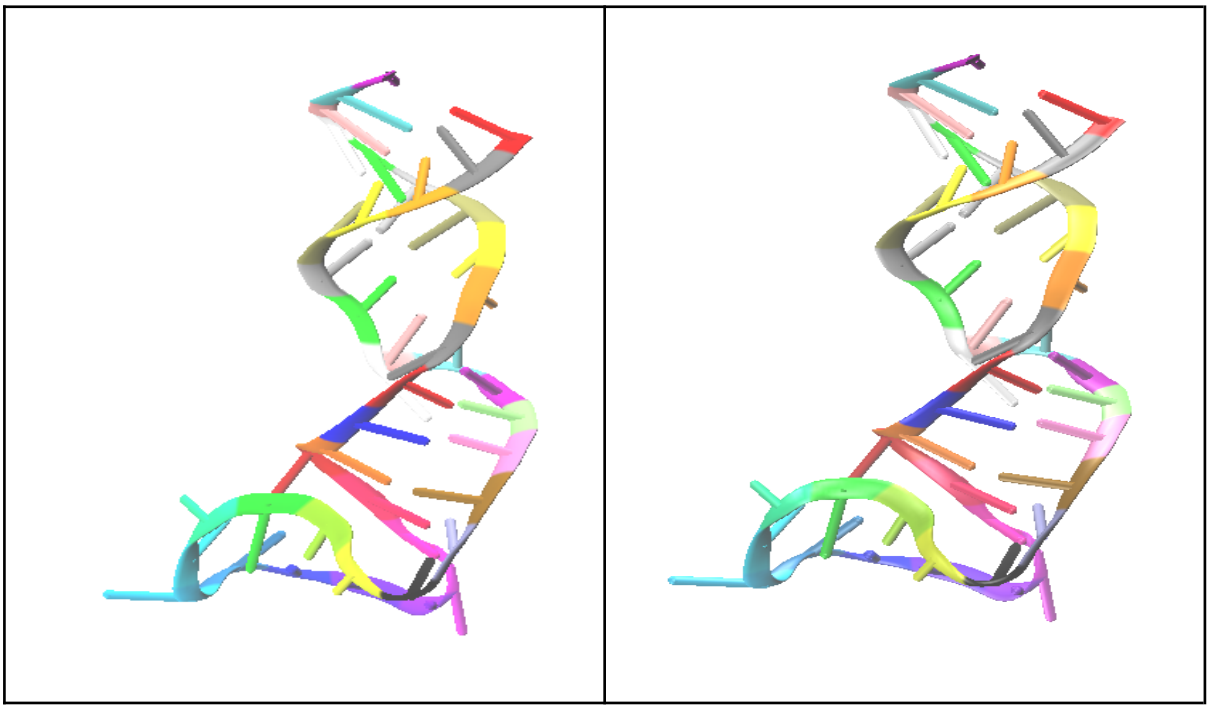


Figure3.3.18 3D structure of PDB: 1XJR Left (optimal structure) and Right (Centroid Structure) motif.

3.4 Interaction between S2M motifs and human miRNAs

Human miRNA that could target S2M sequences was screened using Mirbase. Additional attention was paid to miRNAs, which have been implicated in antiviral miRNA-mediated defense [28]. hsa-miR-1304-3p and hsa-miR-1307-3p were discovered as possible binding sites within the S2M sequences of bat/pangolin and SARS-CoV-1. Only one possible binding site was found within the S2M of Australian and Iranian further, we have observed in our sample 1273404 (Africa / Lesotho / Quthing) & MW883290.1, which have the same secondary structure . SARS-CoV-2 samples: hsa-miR-1307-3p.

3.5 Thermodynamic results of various motifs compared with the reference genome (402124 / human genome)

1.402124

When using RNAfold measurements, decreased RNA structural stability for both MFE and centroid secondary structure which is -6.20 kcal/mol
and -0.80 kcal/mol.

2.MT308984.1

When using RNAfold measurements, the effect of 32 G>U mutations increased RNA secondary structure stability in both MFE (-7.30 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-7.30 kcal/mol vs. -0.80 kcal/mol).

3.KY417152.1 & MT072865.1

When using RNAfold measurements, the effect of 32 G>U mutation and 43 T>C increased RNA secondary structure stability in both MFE (-8.80 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-8.80 kcal/mol vs. -0.80 kcal/mol).

4.862313

When using RNAfold measurements, the effect of 16 C>G mutations increased RNA secondary structure stability in both MFE (-7.70 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-6.20 kcal/mol vs. -0.80 kcal/mol).

5.1396884

When using RNAfold measurements, the effect of 25 A>G mutations increased RNA secondary structure stability in both MFE (-6.90 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-3.60 kcal/mol vs. -0.80 kcal/mol).

6.1390981

When using RNAfold measurements, the effect of 25 T>G mutations increased RNA secondary structure stability in MFE (-6.50 kcal/mol vs. -6.20 kcal/mol) and same in centroid structures (-0.80 kcal/mol vs. -0.80 kcal/mol).

7.1391006

When using RNAfold measurements, the effect of 28 T>G mutations increased RNA secondary structure stability in both MFE (-7.30 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-7.30 kcal/mol vs. -0.80 kcal/mol).

8.1391015

When using RNAfold measurements, the effect of 17 T>C mutation and 43 T>C increased RNA secondary structure stability in both MFE (-6.30 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-3.70 kcal/mol vs. -0.80 kcal/mol).

9.1391017

When using RNAfold measurements, the effect of 17 T>C mutations increased RNA secondary structure stability in both MFE (-7.70 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-6.20 kcal/mol vs. -0.80 kcal/mol).

10.1397948

When using RNAfold measurements, the effect of 38 T>G mutations decreased RNA secondary structure stability in MFE (-6.10 kcal/mol vs. -6.20 kcal/mol) and increased in centroid structures (-5.80 kcal/mol vs. -0.80 kcal/mol).

11.1300659

When using RNAfold measurements, the effect of 8 C>G mutations on RNA secondary structure stability in MFE (-6.20 kcal/mol vs. -6.20 kcal/mol) is the same and increased in centroid structures (-3.00 kcal/mol vs. -0.80 kcal/mol).

12.1292989

When using RNAfold measurements, the effect of 10 T>G mutations increased RNA secondary structure stability in MFE (-6.80 kcal/mol vs. -6.20 kcal/mol) and same in centroid structures (-0.80 kcal/mol vs. -0.80 kcal/mol).

13.1273404

When using RNAfold measurements, the effect of 16 T>G mutations increased RNA secondary structure stability in both MFE (-11.70 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-9.90 kcal/mol vs. -0.80 kcal/mol).

14.414480

When using RNAfold measurements, the effect of 10 T>G mutation and 25 C>G increased RNA secondary structure stability in both MFE (-9.40 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-9.40 kcal/mol vs. -0.80 kcal/mol).

15.862510

When using RNAfold measurements, the effect of 15 T>C mutations increased RNA secondary structure stability in both MFE (-6.20 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-3.90 kcal/mol vs. -0.80 kcal/mol).

16.1319040

When using RNAfold measurements, the effect of 19 T>G mutation and 29 T>G increased RNA secondary structure stability in both MFE (-6.20 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-1.07 kcal/mol vs. -0.80 kcal/mol).

17.MW883290.1

When using RNAfold measurements, the effect of 16 A>G mutations increased RNA secondary structure stability in both MFE (-6.20 kcal/mol vs. -6.20 kcal/mol) and centroid structures (-3.60 kcal/mol vs. -0.80 kcal/mol).

3.6 Guided trees

```
>hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30
TTTCACCGAGGCCACGCGGAGTACGATCGAGTGTACAGTGAAC
>MT308984.1 Mutant SARS coronavirus Urbani clone
TTTCATCGAGGCCACGCGGAGTACGATCGAGGGTACAGTGAAT
>KY417152.1 Bat SARS-like coronavirus & MT072865.1 Pangolin coronavirus
TTTCACCGAGGCCACGCGGAGTACGATCGAGGGTACAGTGAAT
>hCoV-19/India/AP-CS0319/2020|EPI_ISL_862313|2020-07-01
TTTCACCGAGGCCACCCGGAGTACGATCGAGTGTACAGTGAAC
>hCoV-19/France/IDF-HMN-21032100486/2021|EPI_ISL_1396884|2021-03-09
TTTCACCGAGGCCACGCGGAGTACAATCGAGTGTACAGTGAAC
>hCoV-19/Canada/QC-HCLM-7056140333/2020|EPI_ISL_1390981|2020-11-12
TTTCACCGAGGCCACGCGGAGTACTATCGAGTGTACAGTGAAC
>hCoV-19/Canada/QC-CHIC-R4180563/2020|EPI_ISL_1391006|2020-10-18
TTTCACCGAGGCCACGCGGAGTACGATTGAGTGTACAGTGAAC
>hCoV-19/Indonesia/NT-NIHRD-C002128132/2021|EPI_ISL_1391015|2021-02-11
TTTCACCGAGGCCACGTGGAGTACGATCGAGTGTACAGTGAAT
>hCoV-19/Indonesia/JB-NIHRD-C002130218/2021|EPI_ISL_1391017|2021-03-04
TTTCACCGAGGCCACGTGGAGTACGATCGAGTGTACAGTGAAC
>hCoV-19/USA/NY-Wadsworth-21030106-01/2021|EPI_ISL_1397948|2021-03-07
TTTCACCGAGGCCACGCGGAGTACGATCGAGTGTACATTGAAC
>hCoV-19/Bosnia and Herzegovina/KCUS21066/2021|EPI_ISL_1300659|2021-02-24
TTTCACCCAGGCCACGCGGAGTACGATCGAGTGTACAGTGAAC
>hCoV-19/USA/CT-Yale-1770/2021|EPI_ISL_1292989|2021-02-19
TTTCACCGATGCCACGCGGAGTACGATCGAGTGTACAGTGAAC
>hCoV-19/Lesotho/N3421/2021|EPI_ISL_1273404|2021-01-18
TTTCACCGAGGCCACTCGGAGTACGATCGAGTGTACAGTGAAC
>hCoV-19/USA/CruiseA-21/2020|EPI_ISL_414480|2020-02-21
TTTCACCGATGCCACGCGGAGTACCATCGAGTGTACAGTGAAC
>hCoV-19/India/AP-CS0314/2020|EPI_ISL_862510|2020-06-29
TTTCACCGAGGCCATGCGGAGTACGATCGAGTGTACAGTGAAC
>hCoV-19/South Korea/KDCA2188/2021|EPI_ISL_1319040|2021-03-06
TTTCACCGAGGCCACGCGTAGTACGATCTAGTGTACAGTGAAC
>MW883290.1 Severe acute respiratory syndrome coronavirus 2
TTTCACCGAGGCCACACGGAGTACGATCGAGTGTACAGTGAAC
>1XJR_1|Chain A|s2m RNA|null
GTTCACCGAGGCCACGCGGAGTACGAUCGAGGGTACAGTGAAT
```

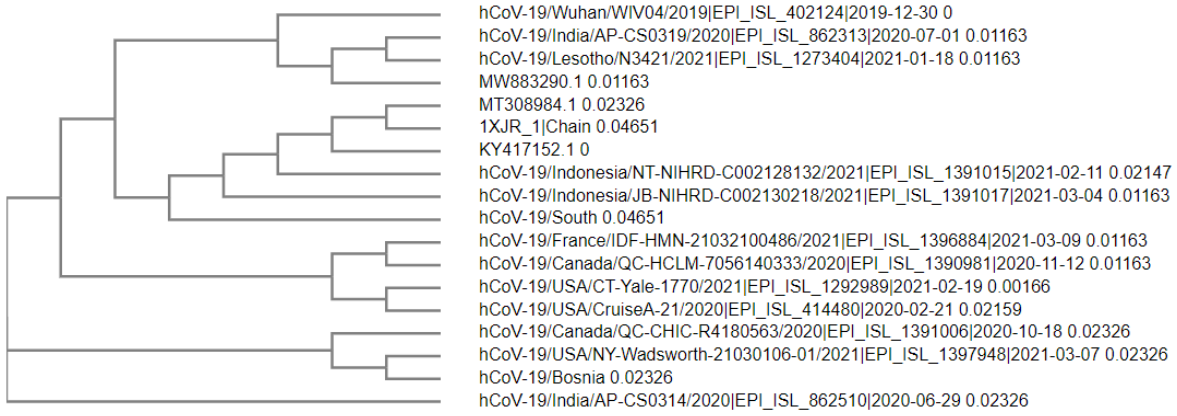


Table 3.6

Eighteen different S2M Motifs. (a) For sequence representation input for ClustalW multiple sequence alignment trees and stem-loop structure representation, alignment of stem-forming elements and columns for each genotype. (b) Display of S2M coronavirus ClustalW numerous sequence alignment trees.

SARS-CoV-2 is thought to have originated in bats due to its close genetic resemblance to bat coronaviruses (96 percent). While the virus shares up to 92 percent similarities with pangolin coronaviruses, there is no clear evidence that another host served as a reservoir for the virus until it was transmitted to humans. SARS-CoV-2 is believed to be bat-borne due to its high homology with bat-coronaviruses. The function of a suspected intermediate host reservoir, pangolins, is still unknown.

Chapter 4

Conclusion

1. Previous study discovered a consistent G>U mutation at position 32 of SARS-43-nucleotide CoV-2's long S2M sequence. This mutation has not been detected in any CoV strains from bats or pangolins. We concluded that this mutation was linked to bat/pangolin transmissibility in humans.

2. In contrast to bat/pangolin coronavirus sequences, the MFE and centroid secondary structures in SARS-CoV-2 sequences were found to be substantially different. This appears to be the product of a mutation in S2M at position 32. Between the different human SARS-COV-2 S2M sequences, there were also essential variations in the MFE and centroid structures. In S2M bat/pangolin coronavirus and SARS-COVs, the MFE and centroid secondary structures were found to be identical.

3. From the 17 different genomes, we had found mutation at positions 6, 8, 10, 15, 16, 17, 19, 25, 28, 29, 32, 38, and 43, which are responsible for more hairpin loop-like secondary structures.The S2M sequences of (+) ssRNA viruses are possibly still involved and will continue to influence the evolution of these viruses.

4. These structural changes make SARS CoV-2's structure less stable, giving it more freedom, which may be one of the virus's escape mechanisms from host defenses or make it easier for it to enter host proteins and enzymes. 32G>U mutations may affect.

5. .S2M probably works via an RNA interference-like mechanism, targeting homologous sequence loci in infected species, based on structural similarities between micro RNA hairpins implicated in gene regulation.

6. Two target sequences were discovered within the bat/pangolin and SARS-CoV-1 S2M, and one within the SARS-CoV-2 S2M when searching for human miRNA targets on S2M motifs. The existence of host miRNA targets within S2M motifs may

be essential for host selection and anti-viral miRNA defense. SARS-S2M CoV-2's can increase the virus's viability and infectivity.

7. S2M probably works via an RNA interference-like mechanism, targeting homologous sequence loci in infected species, based on structural similarities between micro RNA hairpins implicated in gene regulation.

8. The S2M sequences of (+) ssRNA viruses are possibly still involved and will continue to influence the evolution of these viruses (5). The 16 G>U/A and 32 G>U nucleotide changes in SARS-S2M CoV-2's sequence make it a candidate for hsa-miR-1307-3p, a human miRNA. Two human miRNAs, hsa-miR-1307-3p and hsa-miR-1304-3p, will target bat/pangolin coronavirus and SARS-CoV-1 S2M sequences. As a result, only one human miRNA can influence SARS-CoV-2 (1273404 & MW883290.1) viral replication by targeting the S2M series.

Bibliography

1. Weiss RA, and McMichael AJ. Social and environmental risk factors in the emergence of infectious diseases. Nat Med. 2004;10(12 Suppl):S70–6.

2. Totura AL, and Baric RS. SARS coronavirus pathogenesis: host innate immune responses and viral antagonism of interferon. Curr Opin Virol. 2012;2(3):264–75.

3. Tengs T, and Jonassen CM. Distribution and Evolutionary History of the Mobile Genetic Element s2m in Coronaviruses. Diseases. 2016;4(3).

4. Tengs T, Kristoffersen AB, Bachvaroff TR, and Jonassen CM. A mobile genetic element with unknown function found in distantly related viruses. Virol J. 2013;10:132.

5. Robertson MP, Igel H, Baertsch R, Haussler D, Ares M, Jr.., and Scott WG. The structure of a rigorously conserved RNA element within the SARS virus genome. PLoS Biol. 2005;3(1):e5.

6. Lee, B. D. (2018). Squiggle: a user-friendly two-dimensional DNA sequence visualization tool. Bioinformatics. doi:10.1093/bioinformatics/bty807.

7. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots.(https://doi.org/10.1016/S0166-218X(00)00186-4).

8. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov.

9. Kerpedjiev P, Hammer S, Hofacker IL (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 31(20):3377-9.

10.Kalé L.V., Bhatele A., Bohm E.J., Phillips J.C. (2011) NAMD (NAnoscale Molecular Dynamics). In: Padua D. (eds) Encyclopedia of Parallel Computing. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09766-4_505.

11. Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.

12. R. Lorenz, S.H. Bernhart, C. Hoener zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler and I.L. Hofacker (2011), "ViennaRNA Package 2.0", Algorithms for Molecular Biology: 6:26.

13. J.S. McCaskill (1990), "The equilibrium partition function and base pair binding probabilities for RNA secondary structures", Biopolymers: 29, pp 1105-1119

14. Elbe S, and Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017;1(1):33–46.
15. Antczak, M., Popenda, M., Zok, T., Sarzynska, J., Ratajczak, T., Tomczyk, K., Adamiak, R.W., Szachniuk, M. New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure, Acta Biochimica Polonica, 2016, 63(4):737-744 (doi:10.18388/abp.2016_1329).
16. Popenda, M., Blazewicz, M., Szachniuk, M., Adamiak, R.W. RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. Nucleic Acids Research, 2008, 36:D386-D391 (doi:10.1093/nar/gkm786).
17. miRBase: annotating high confidence microRNAs using deep sequencing data. Kozomara A, Griffiths-Jones S. Nucleic Acids Res 2014 42:D68-D73

Appendix

A.1 Mountain plot

In a plot of height versus location, the Mountain plot depicts secondary structures such as MFE, the thermodynamic ensemble of RNA(pf), and centroid structures as an x-y graph. MFE stands for Minimum free energy structure, pf stands for partition, and centroid stands for centroid structure. The number of base pairs enclosed at position k determines the height m(k). The MFE structure (red), pairing probabilities (blue), and positional entropy curve are all seen (green). Low entropy identifies well-defined areas .

1.402124

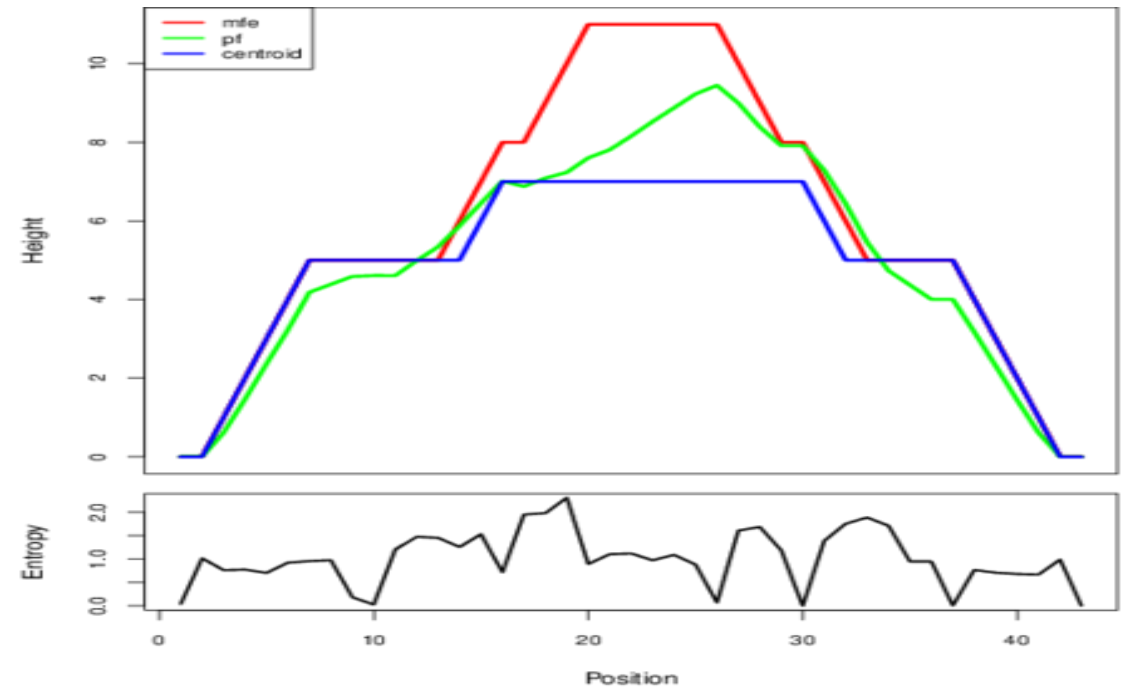


Figure A.1 Mountain plot of height versus the position of motif 402124.

2.MT308984.1

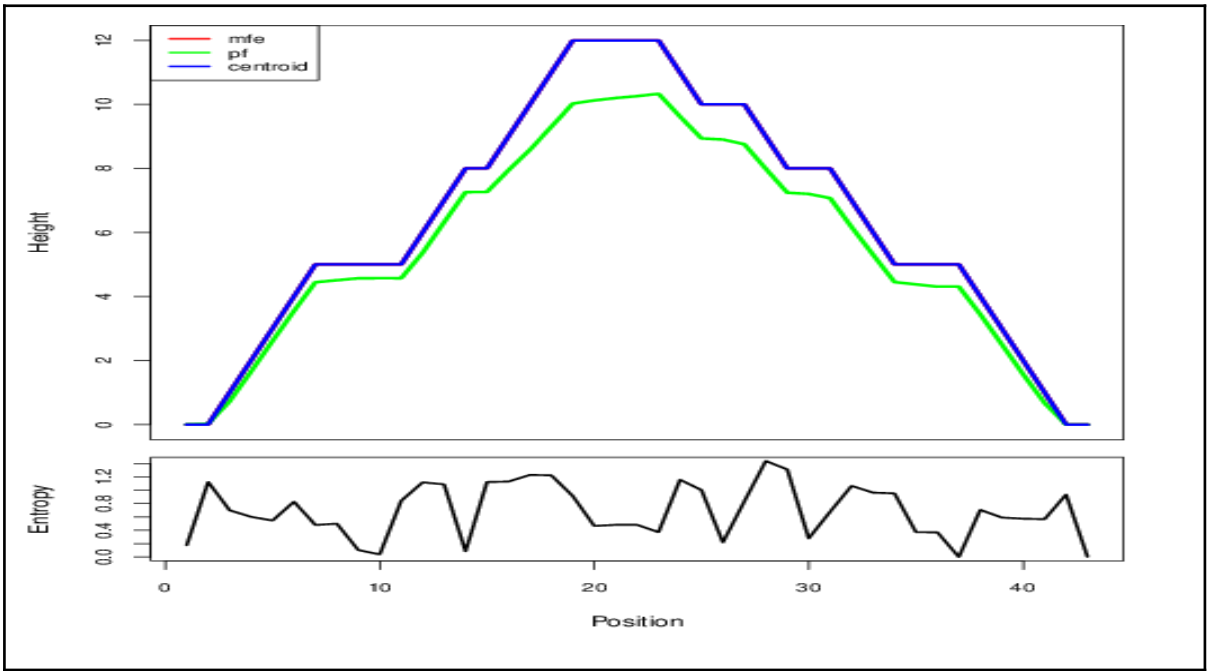


Figure A.2 Mountain plot of height versus the position of motif MT308984.1.

3.KY417152.1 & MT072865.1

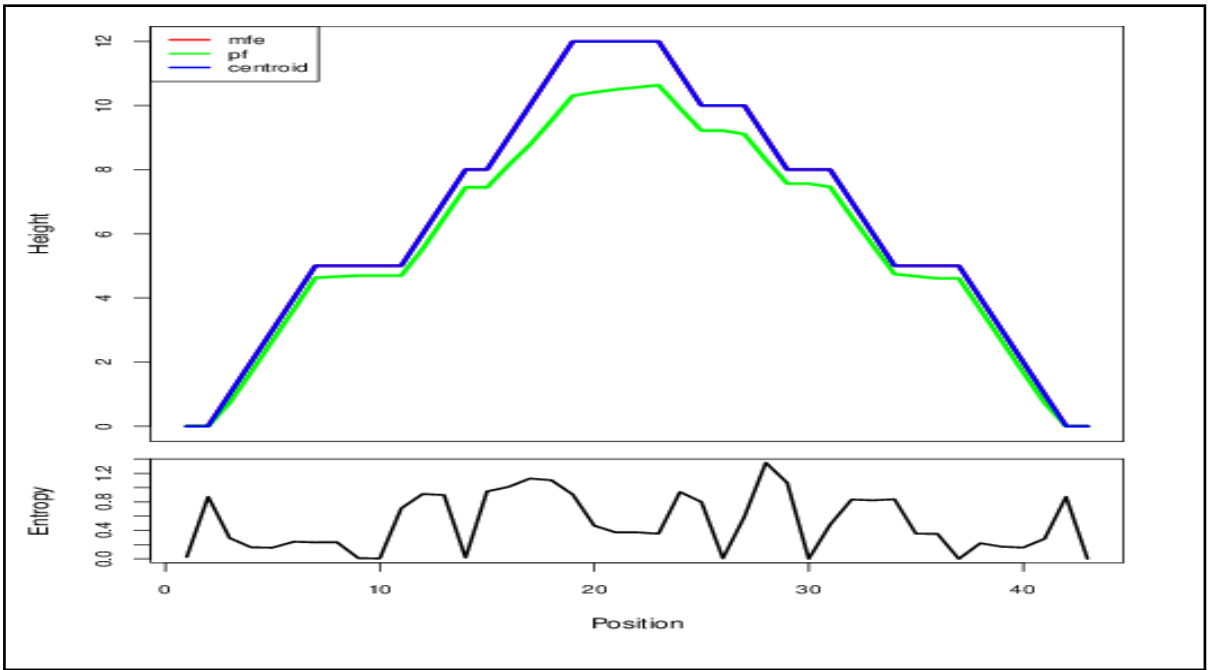


Figure A.3 Mountain plot of height versus the position of motif KY417152.1 & MT072865.1.

4.862313

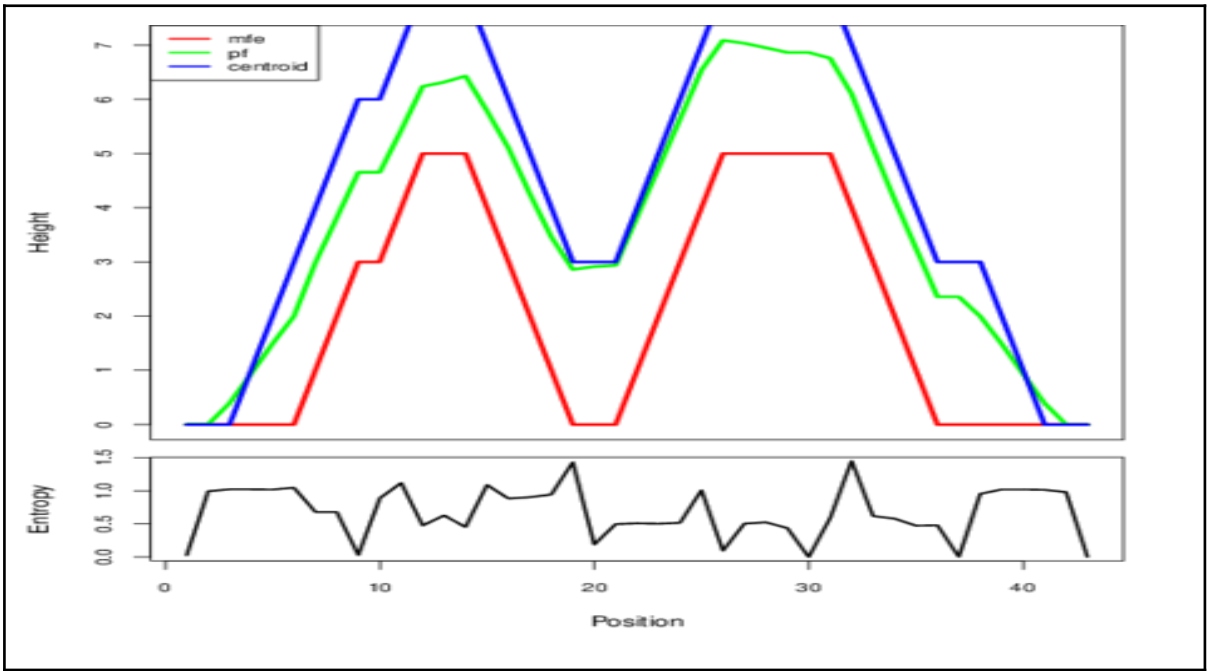


Figure A.4 Mountain plot of height versus the position of motif 862313.

5.1396884

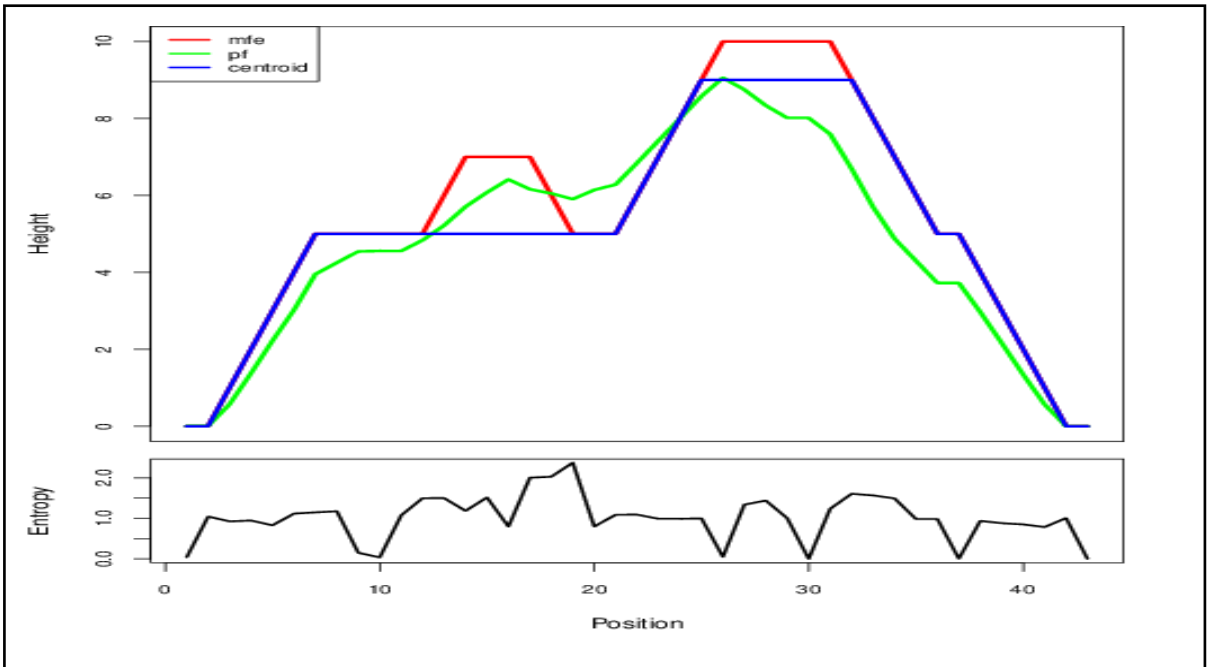


Figure A.5 Mountain plot of height versus the position of motif 1396884.

6.1390981

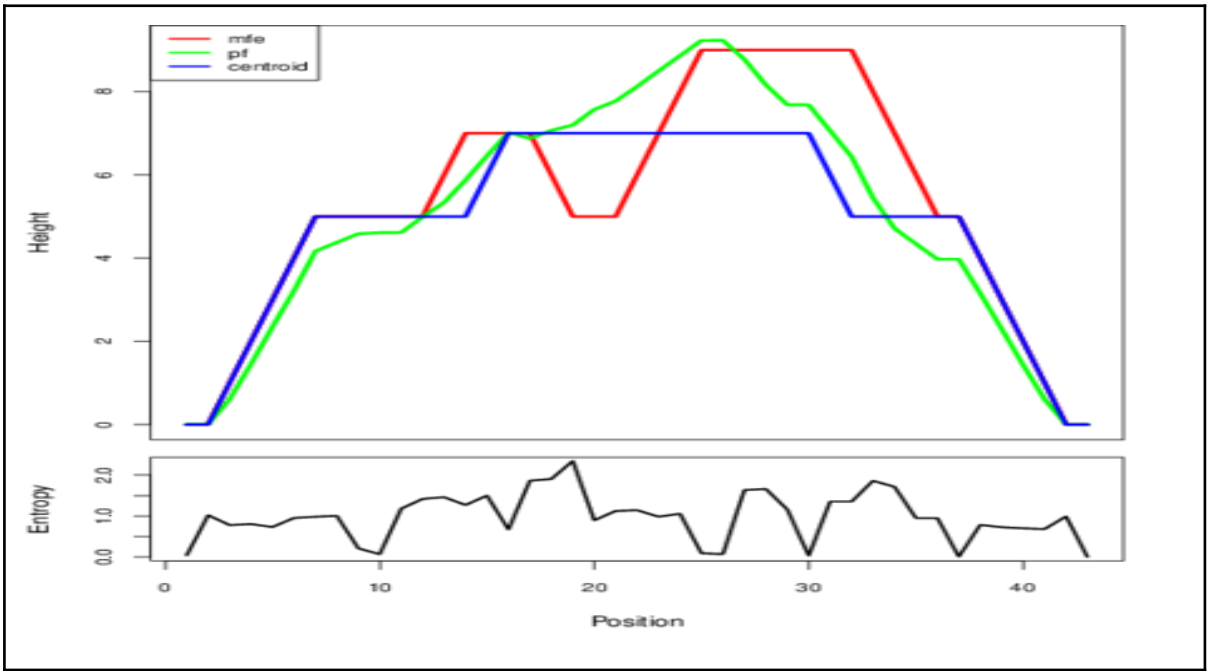


Figure A.6 Mountain plot of height versus the position of motif 1390981.

7.1391006

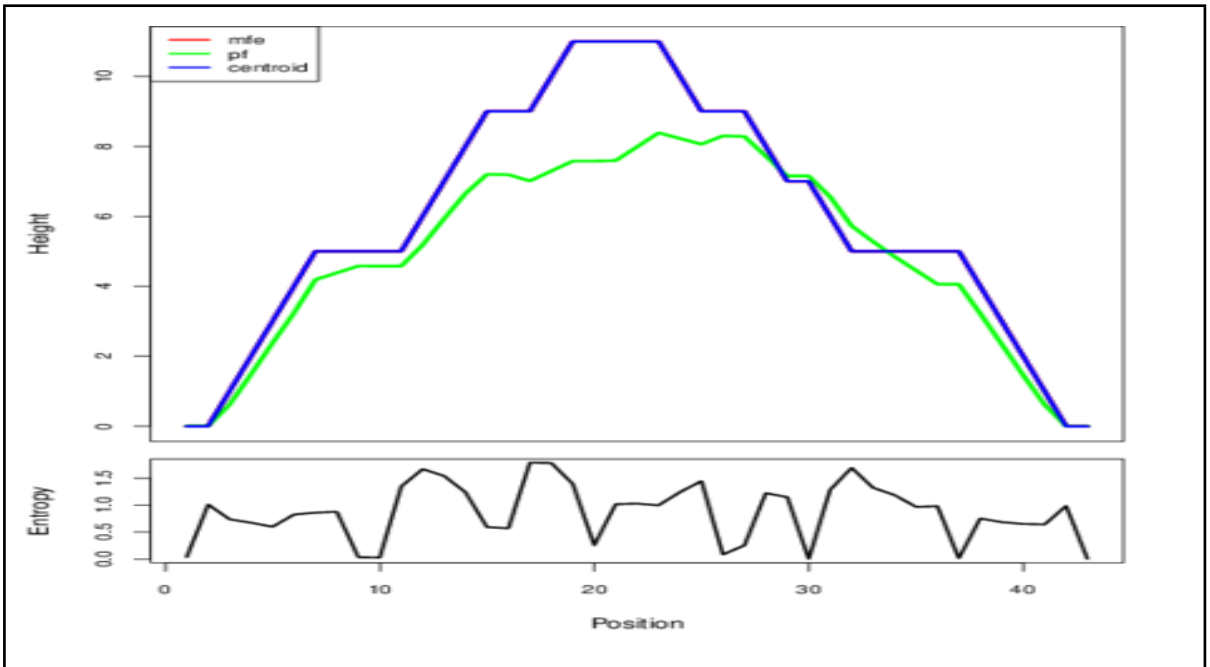


Figure A.7 Mountain plot of height versus the position of motif 1391006.

8.1391015

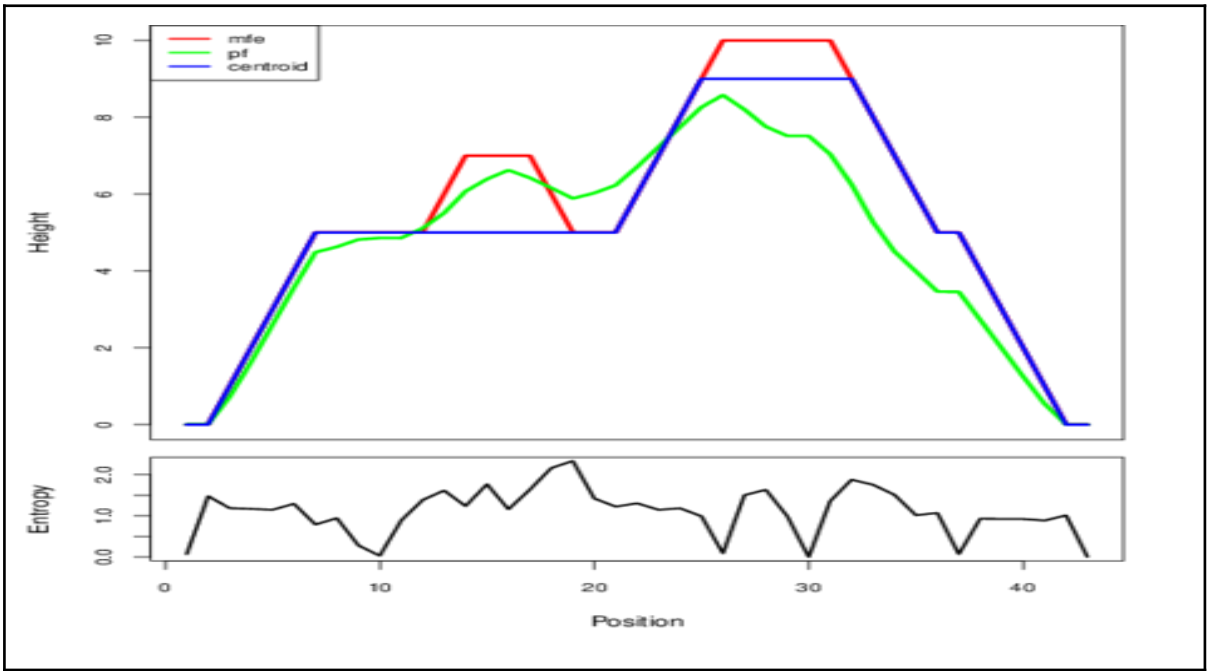


Figure A.8 Mountain plot of height versus the position of motif 1391015.

9.1391017

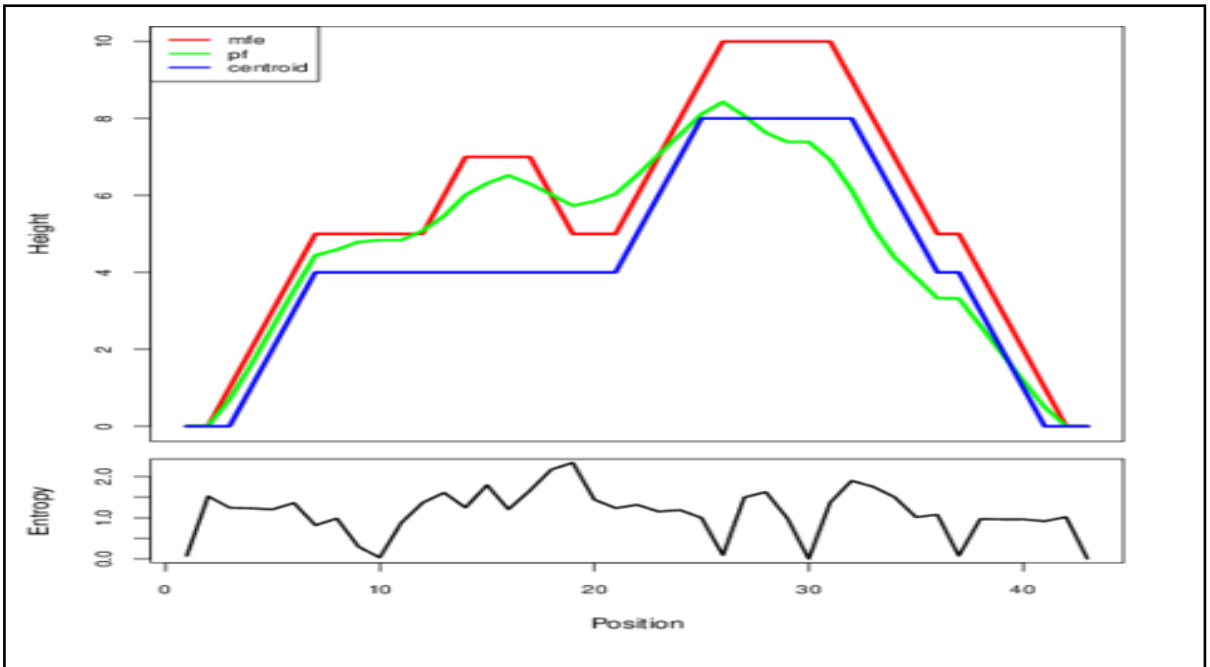


Figure A.9 Mountain plot of height versus the position of motif 1391017.

10.1397948

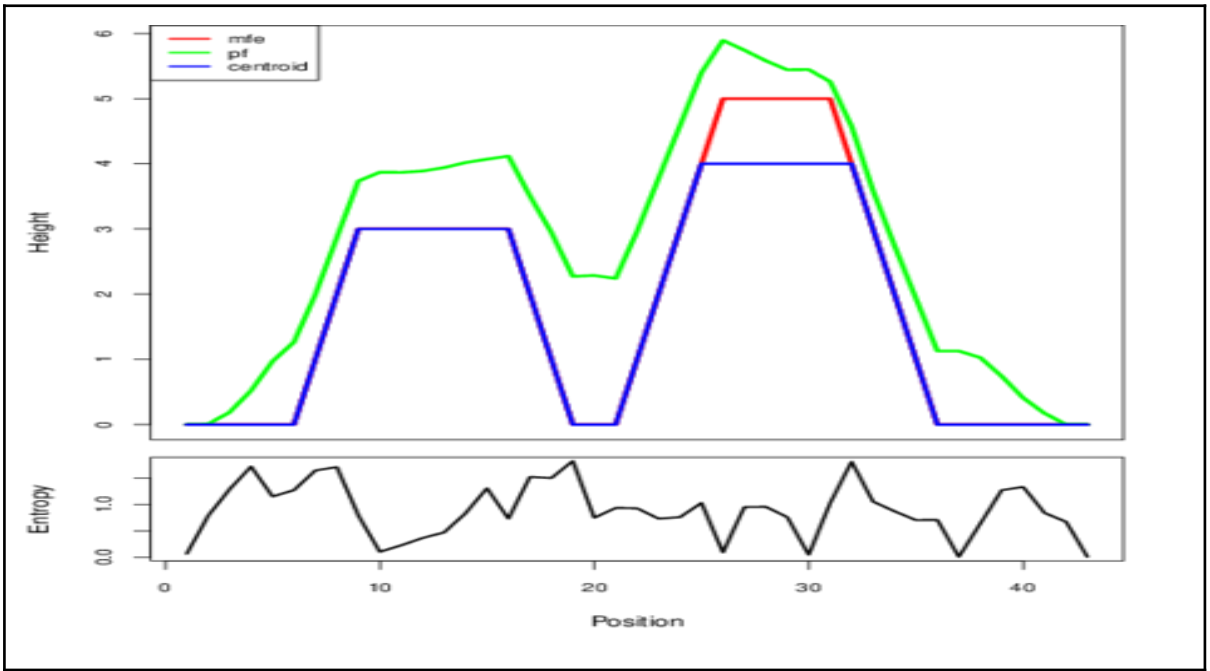


Figure A.10 Mountain plot of height versus the position of motif 1397948.

11.1300659

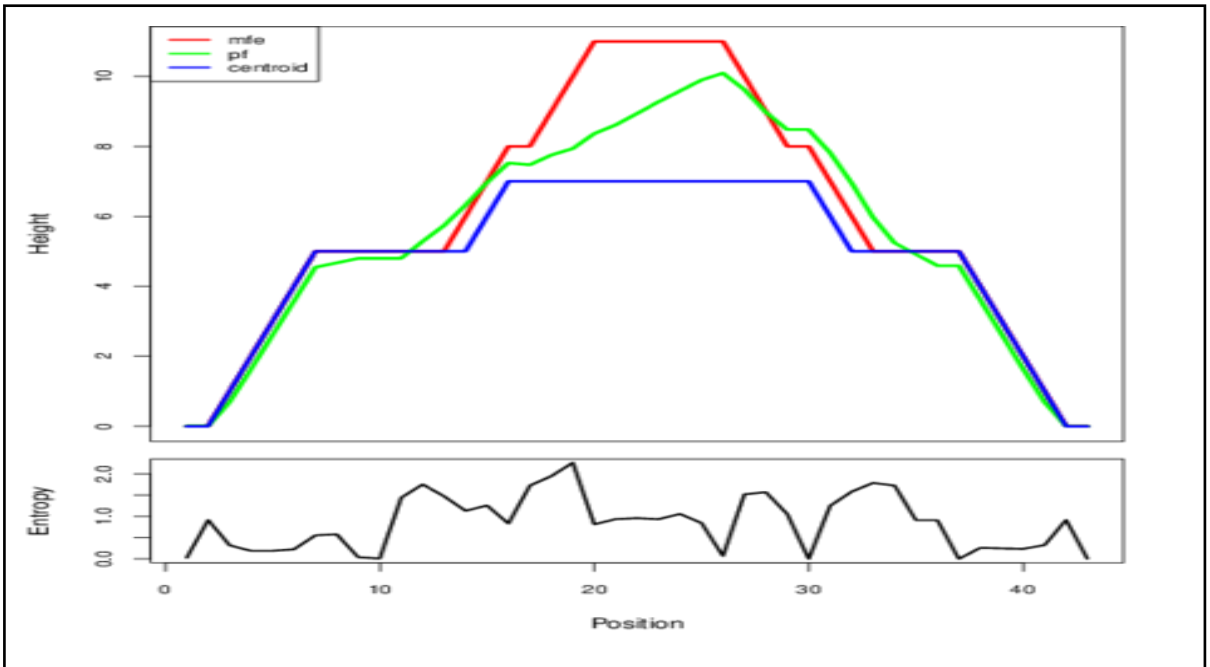


Figure A.11 Mountain plot of height versus the position of motif 1300659.

12.1292989

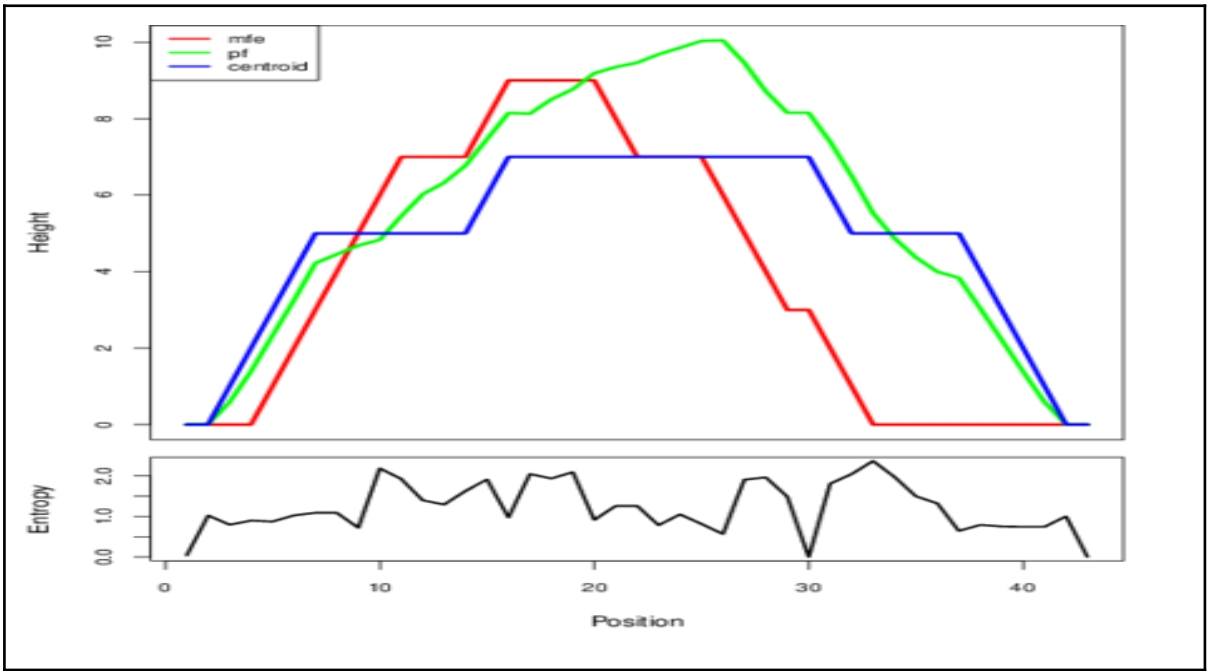


Figure A.12 Mountain plot of height versus the position of motif 1292989.

13.1273404

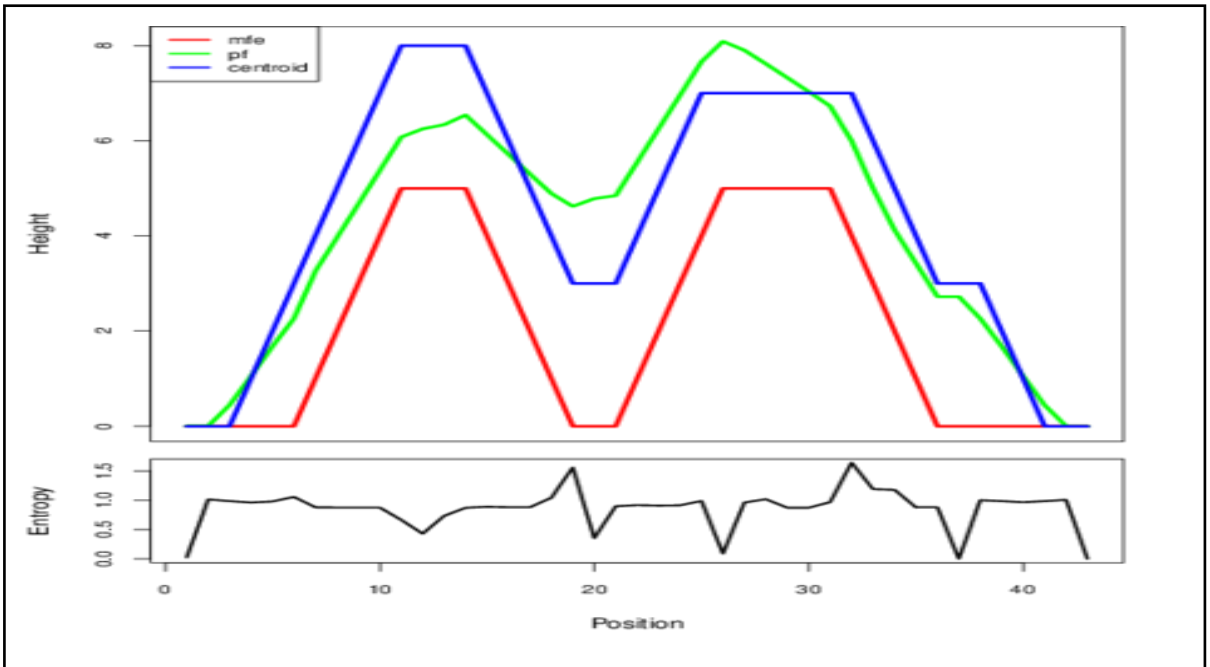


Figure A.13 Mountain plot of height versus the position of motif 1273404.

14.414480

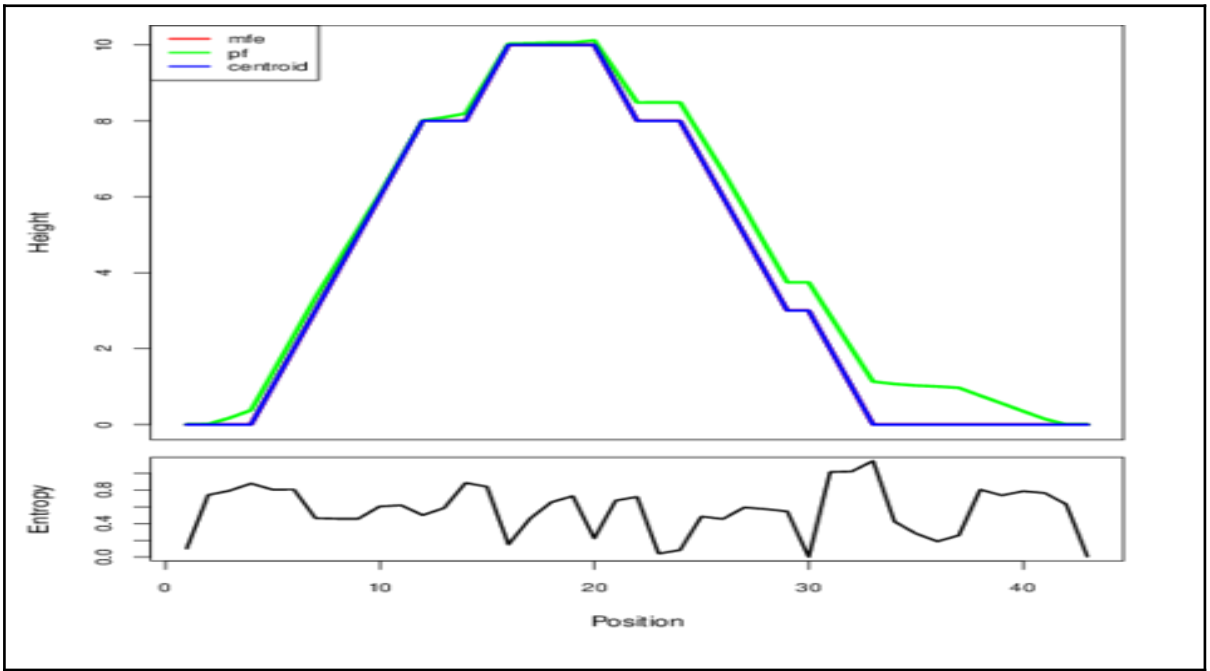


Figure A.14 Mountain plot of height versus the position of motif 414480.

15.862510

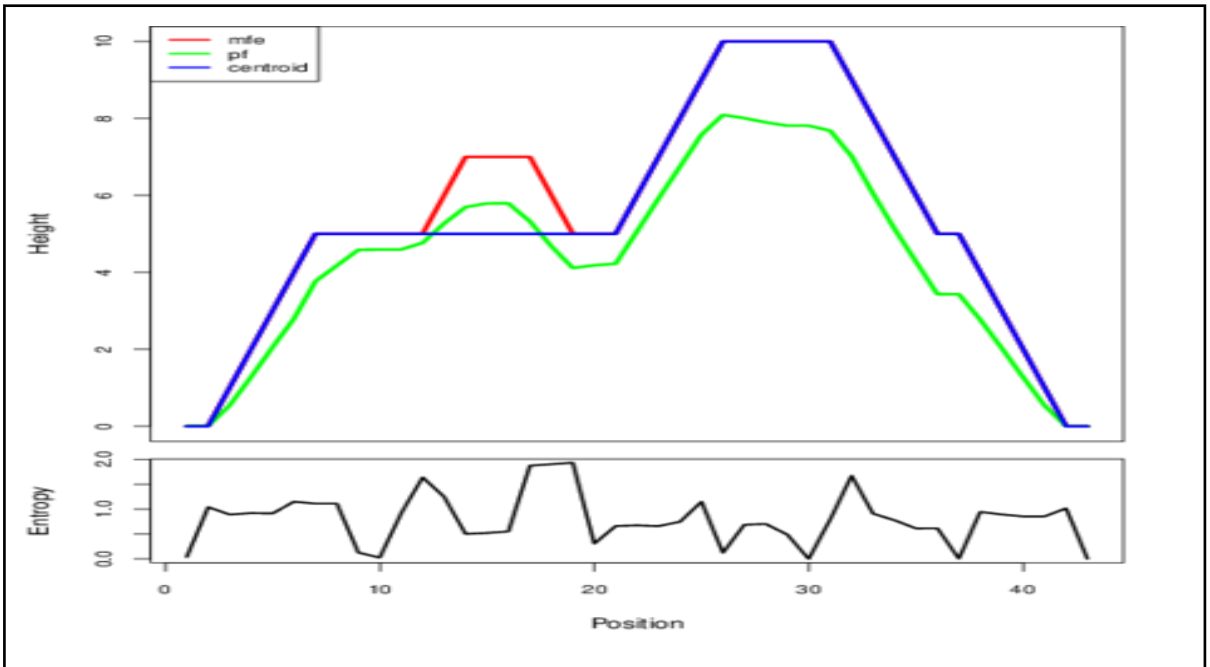


Figure A.15 Mountain plot of height versus the position of motif 862510.

16.1319040

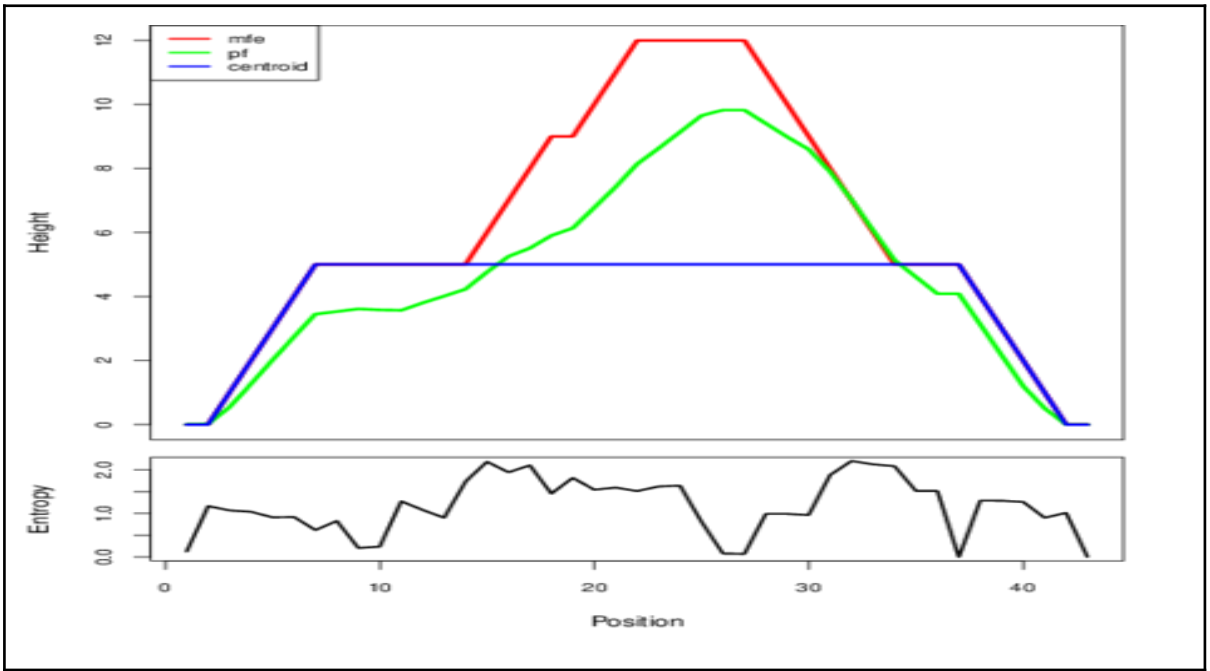


Figure A.16 Mountain plot of height versus the position of motif 1319040.

17.MW883290.1

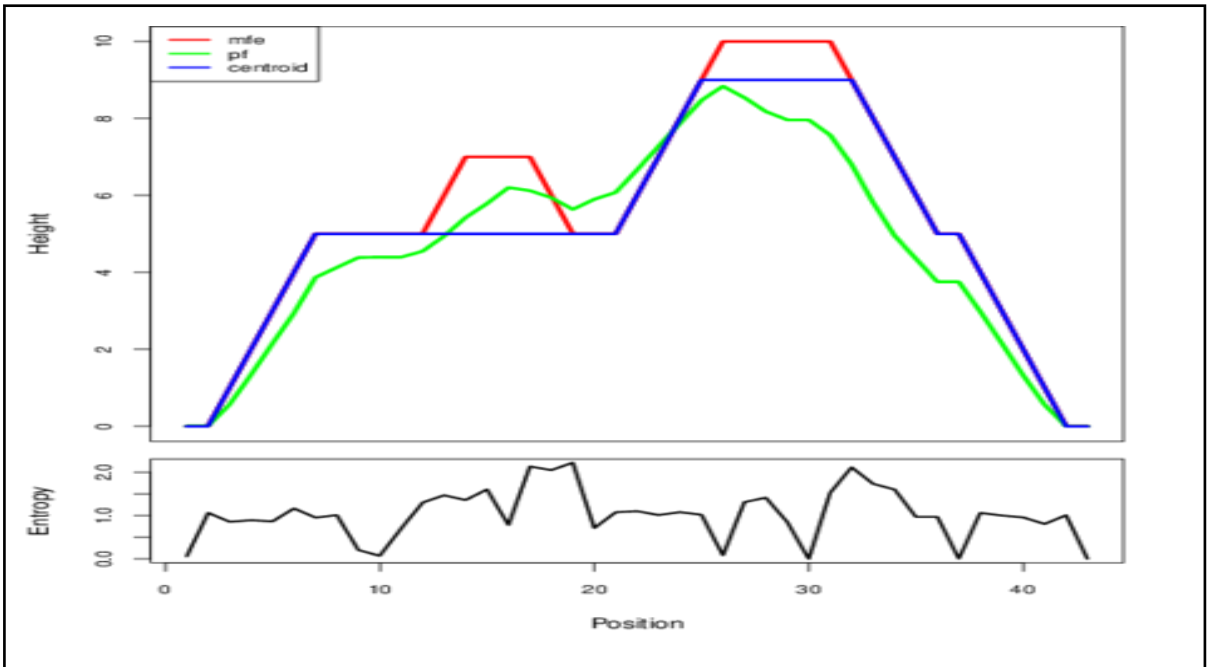
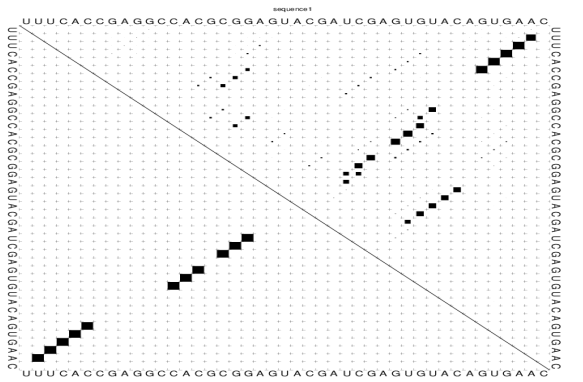


Figure A.17 Mountain plot of height versus the position of motif MW883290.1.

A.2. Dot plot

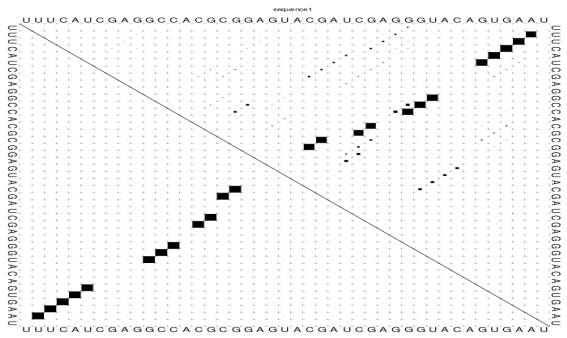
RNAfold (12) was used to create a dot plot. In the lower-left triangle, the MFE structure is depicted. The top right triangle shows the likelihood of all possible base pairs. The region of each dot is proportional to the base pair pairing probability. The size of the boxes, on the other hand, is proportional to the base-pairing likelihood, with small boxes indicating a low chance of forming a base pair and big boxes indicating a high chance.

1.402124

	**Figure A.2** Dot plot containing the base pair probabilities of motif 402124.

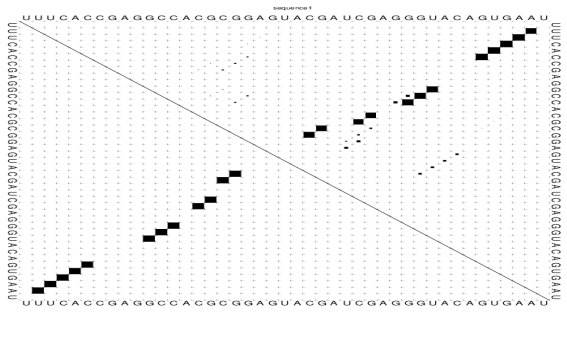
UUUCACCGAGGCCACGCGGAGUACGAUCGAGUGUACAGUGAAC .(((((......(((.(((.......))).))).....))))).	UUUCACCGAGGCCACGCGGAGUACGAUCGAGUGUACAGUGAAC .(((((.......((..............)).....))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.20 kcal/mol.	The centroid secondary structure has a minimum free energy of -0.80 kcal/mol in dot-bracket notation.

2.MT308984.1

	**Figure A.2** Dot plot containing the base pair probabilities of motif MT308984.1.

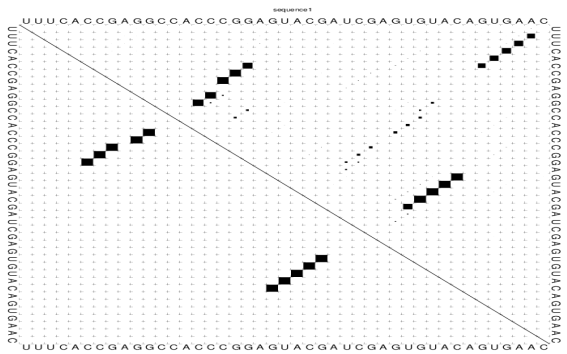
UUUCAUCGAGGCCACGCGGAGUACGAUCGAGGGUACAGUGAAU .(((((....(((.((((.....))..))..)))...))))).	UUUCAUCGAGGCCACGCGGAGUACGAUCGAGGGUACAGUGAAU .(((((....(((.((((.....))..))..)))...))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -7.30 kcal/mol.	The centroid secondary structure has a minimum free energy of -7.30 kcal/mol in dot-bracket notation.

3.KY417152.1 and MT072865.1

	**Figure A.3** Dot plot containing the base pair probabilities of motif KY417152.1 and MT072865.1.

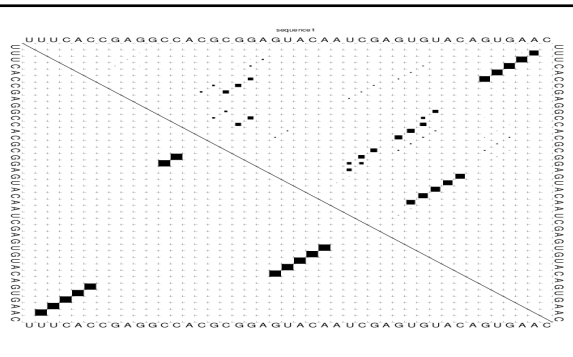
UUUCACCGAGGCCACGCGGAGUACGAUCGAGGGUACAGUGAAU .(((((....(((.((((.....))..))..)))...))))).	UUUCACCGAGGCCACGCGGAGUACGAUCGAGGGUACAGUGAAU .(((((....(((.((((.....))..))..)))...))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -8.80 kcal/mol.	The centroid secondary structure has a minimum free energy of -8.80 kcal/mol in dot-bracket notation.

4.862313

	**Figure A.4** Dot plot containing the base pair probabilities of motif 862313.

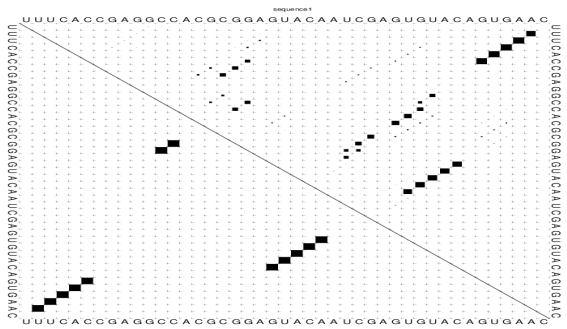
UUUCACCGAGGCCACCCGGAGUACGAUCGAGUGUACAGUGAAC .....(((.((...))))).(((((......))))).....	UUUCACCGAGGCCACCCGGAGUACGAUCGAGUGUACAGUGAAC ..((((((.((...))))).(((((......)))))..))))..
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -7.70 kcal/mol.	The centroid secondary structure has a minimum free energy of -6.20 kcal/mol in dot-bracket notation.

5.1396884

	**Figure A.5.** Dot plot containing the base pair probabilities of motif 1396884.

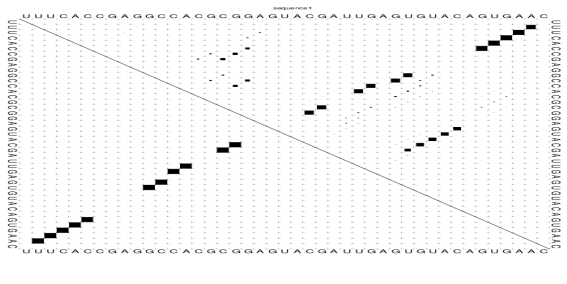
UUUCACCGAGGCCACGCGGAGUACAAUCGAGUGUACAGUGAAC .(((((.....((....)).(((((......))))).))))).	UUUCACCGAGGCCACGCGGAGUACAAUCGAGUGUACAGUGAAC .(((((..............((((........)))).))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.90 kcal/mol.	The centroid secondary structure has a minimum free energy of -3.60 kcal/mol in dot-bracket notation.

6.1390981

	**Figure A.6.** Dot plot containing the base pair probabilities of motif 1390981.

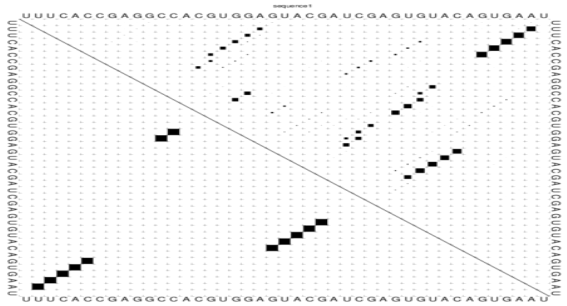
UUUCACCGAGGCCACGCGGAGUACUAUCGAGUGUACAGUGAAC .(((((.....((....)).((((........)))).))))).	UUUCACCGAGGCCACGCGGAGUACUAUCGAGUGUACAGUGAAC .(((((.......((..............)).....))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.50 kcal/mol.	The centroid secondary structure has a minimum free energy of -0.80 kcal/mol in dot-bracket notation.

7.1391006

	**Figure A.7.** Dot plot containing the base pair probabilities of motif 1391006.

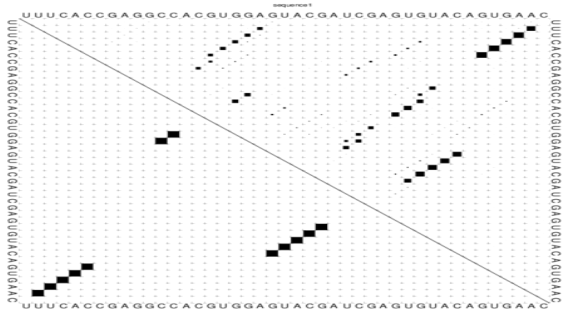
UUUCACCGAGGCCACGCGGAGUACGAUUGAGUGUACAGUGAAC .(((((....((((..((.....))..)).))).....))))).	UUUCACCGAGGCCACGCGGAGUACGAUUGAGUGUACAGUGAAC .(((((....((((..((.....))..)).))).....))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -7.30 kcal/mol.	The centroid secondary structure has a minimum free energy of -7.30 kcal/mol in dot-bracket notation.

8.1391015

	**Figure A.8.** Dot plot containing the base pair probabilities of motif 1391015.

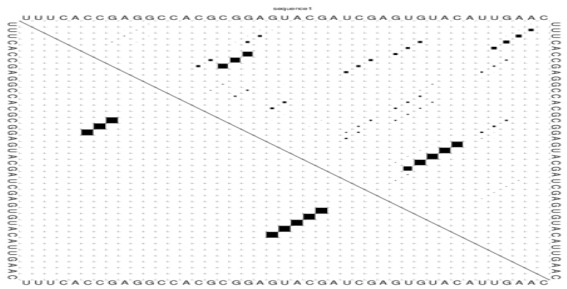
UUUCACCGAGGCCACGUGGAGUACGAUCGAGUGUACAGUGAAU .(((((.....((....)).(((((......))))).))))).	UUUCACCGAGGCCACGUGGAGUACGAUCGAGUGUACAGUGAAU .(((((..............((((........)))).))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.30 kcal/mol.	The centroid secondary structure has a minimum free energy of -3.70 kcal/mol in dot-bracket notation.

9.1391017

	**Figure A.9** Dot plot containing the base pair probabilities of motif 1391017.

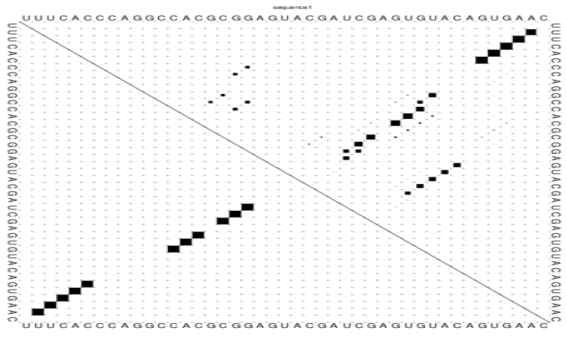
UUUCACCGAGGCCACCCGGAGUACGAUCGAGUGUACAGUGAAC .....(((.((...))))).(((((......))))).....	UUUCACCGAGGCCACCCGGAGUACGAUCGAGUGUACAGUGAAC ..((((((.((...))))).(((((......)))))..))))..
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -7.70 kcal/mol.	The centroid secondary structure has a minimum free energy of -6.20 kcal/mol in dot-bracket notation.

10.1397948

	**Figure A.10** Dot plot containing the base pair probabilities of motif 1397948.

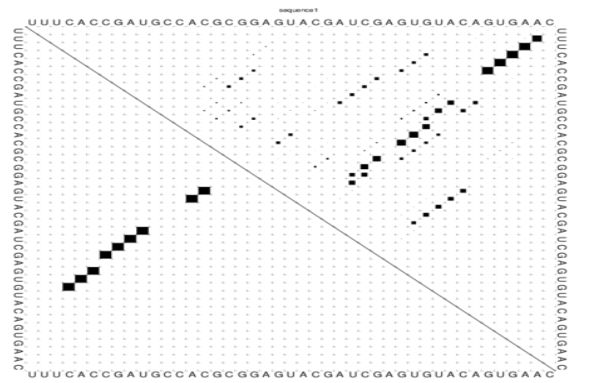
UUUCACCGAGGCCACGCGGAGUACGAUCGAGUGUACAUUGAAC .....(((.........))).(((((.....))))).....	UUUCACCGAGGCCACGCGGAGUACGAUCGAGUGUACAUUGAAC .....(((.........))).((((.........)))).....
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.10 kcal/mol.	The centroid secondary structure has a minimum free energy of -5.80 kcal/mol in dot-bracket notation.

11.1300659

	**Figure A.11.** Dot plot containing the base pair probabilities of motif 1300659.

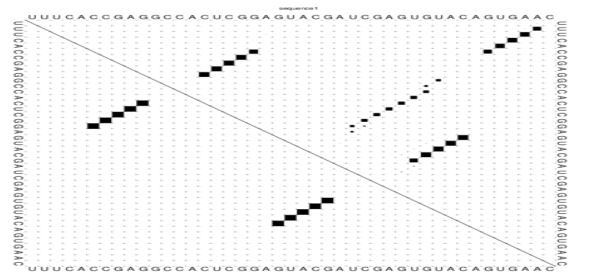
UUUCACCGAGGCCACGUGGAGUACGAUCGAGUGUACAGUGAAC .(((((.....((....)).(((((.....))))).))))).	UUUCACCGAGGCCACGUGGAGUACGAUCGAGUGUACAGUGAAC ..((((.............((((.........)))).)))))..
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.20 kcal/mol.	The centroid secondary structure has a minimum free energy of -3.00 kcal/mol in dot-bracket notation.

12.1292989

	**Figure A.12** Dot plot containing the base pair probabilities of motif 1292989.

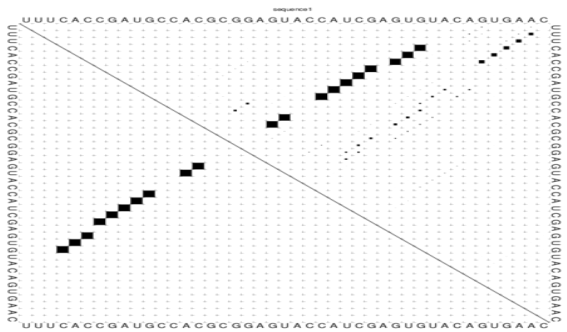
UUUCACCGAUGCCACGCGGAGUACGAUCGAGUGUACAGUGAAC ...(((((((...((.....)).....)))).))).....	UUUCACCGAUGCCACGCGGAGUACGAUCGAGUGUACAGUGAAC .(((((.......((..............)).....))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.80 kcal/mol.	The centroid secondary structure has a minimum free energy of -0.80 kcal/mol in dot-bracket notation.

13.1273404

	**Figure A.13.** Dot plot containing the base pair probabilities of motif 1273404.

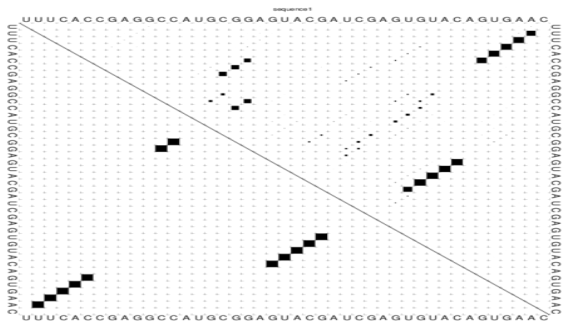
UUUCACCGAGGCCACUCGGAGUACGAUCGAGUGUACAGUGAAC .....(((((....))))).(((((......)))))....	UUUCACCGAGGCCACUCGGAGUACGAUCGAGUGUACAGUGAAC ..(((((((((....))))).((((........))))..)))..
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -11.70 kcal/mol.	The centroid secondary structure has a minimum free energy of -9.90 kcal/mol in dot-bracket notation.

14.414480

	**Figure A.14.** Dot plot containing the base pair probabilities of motif 414480.

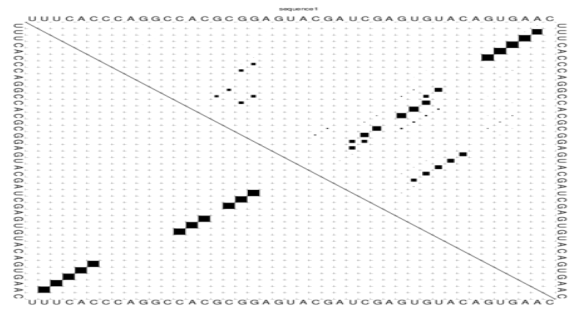
UUUCACCGAUGCCACGCGGAGUACCAUCGAGUGUACAGUGAAC ...(((((((((..((.....))..)))))).))).....	UUUCACCGAUGCCACGCGGAGUACCAUCGAGUGUACAGUGAAC ...(((((((((..((.....))..)))))).))).....
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -9.40 kcal/mol.	The centroid secondary structure has a minimum free energy of -9.40 kcal/mol in dot-bracket notation.

15.862510

	**Figure A.15** Dot plot containing the base pair probabilities of motif 862510.

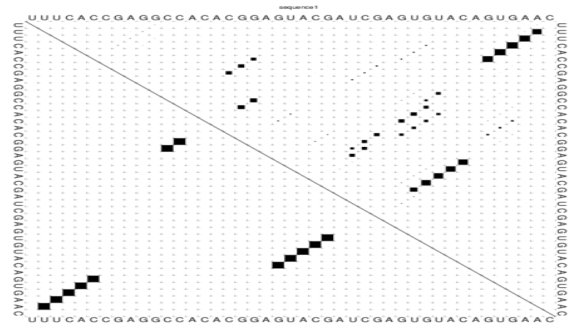
UUUCACCGAGGCCACGUGGAGUACGAUCGAGUGUACAGUGAAC .(((((.....((....)).(((((.....))))).))))).	UUUCACCGAGGCCACGUGGAGUACGAUCGAGUGUACAGUGAAC ..((((..............((((........)))).))))..
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.20 kcal/mol.	The centroid secondary structure has a minimum free energy of -3.90 kcal/mol in dot-bracket notation.

16.1319040

	**Figure A.16.** Dot plot containing the base pair probabilities of motif 1319040.

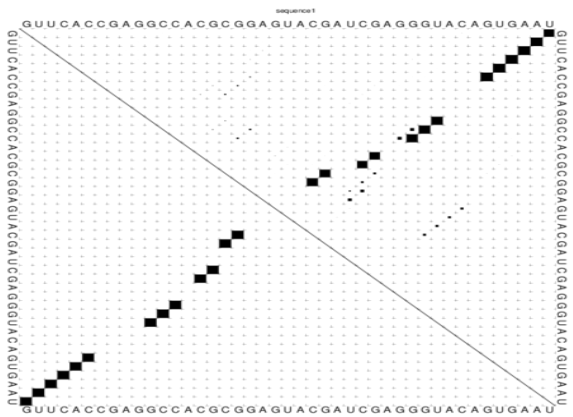
UUUCACCGAGGCCACCCGGAGUACGAUCGAGUGUACAGUGAAC .....(((.((...))))).(((((......))))).....	UUUCACCGAGGCCACCCGGAGUACGAUCGAGUGUACAGUGAAC ..((((((.((...))))).(((((......))))).))))..
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.20 kcal/mol.	The centroid secondary structure has a minimum free energy of --1.07 kcal/mol in dot-bracket notation.

17.MW883290.1

	**Figure A.17** Dot plot containing the base pair probabilities of motif MW883290.1.

UUUCACCGAGGCCACACGGAGUACGAUCGAGUGUACAGUGAAC .(((((.....((....)).(((((......))))).))))).	UUUCACCGAGGCCACACGGAGUACGAUCGAGUGUACAGUGAAC .(((((..............((((........)))).))))).
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -6.20 kcal/mol.	The centroid secondary structure has a minimum free energy of -3.60 kcal/mol in dot-bracket notation.

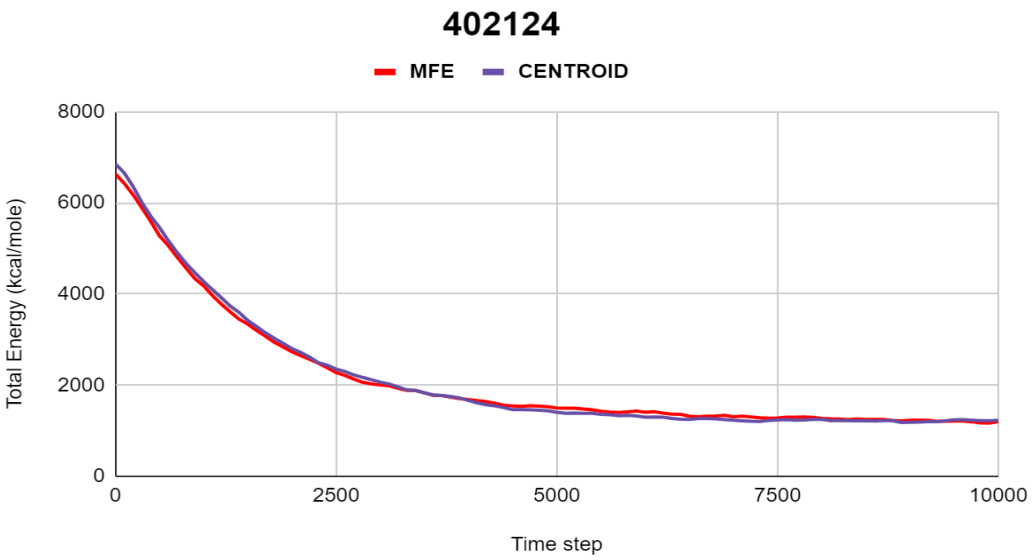
17. PDB :1XJR

	**Figure A.17** Dot plot containing the base pair probabilities of PDB:1XJR.

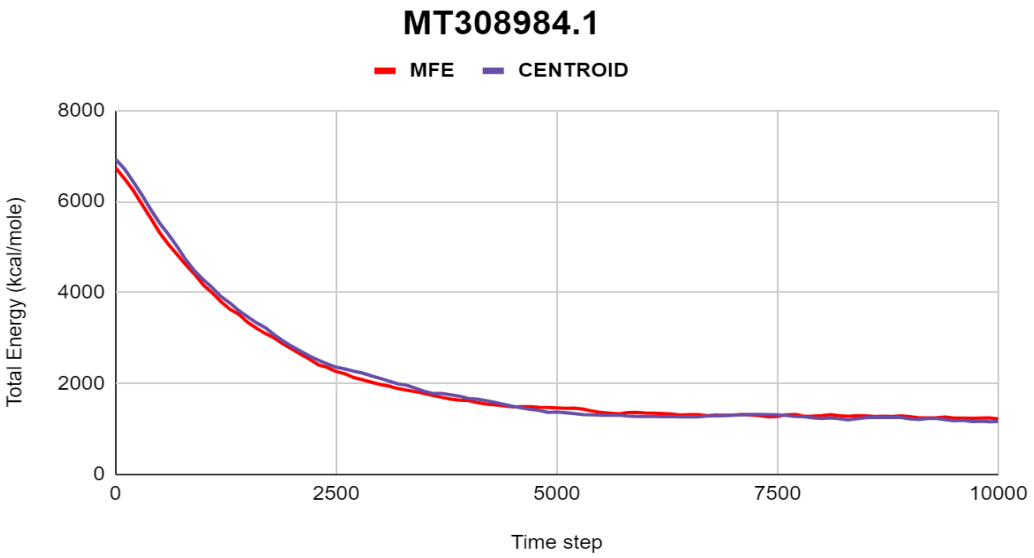
GUUCACCGAGGCCACGCGGAGUACGAUCGAGGGUACAGUGAAU ((((((....(((.((((.....))..))..)))...))))))	GUUCACCGAGGCCACGCGGAGUACGAUCGAGGGUACAGUGAAU ((((((....(((.((((.....))..))..)))...))))))
In dot-bracket notation, the optimum secondary structure has a minimum free energy of -9.40 kcal/mol.	The centroid secondary structure has a minimum free energy of -9.40 kcal/mol in dot-bracket notation.

A.3 Energy Minimization

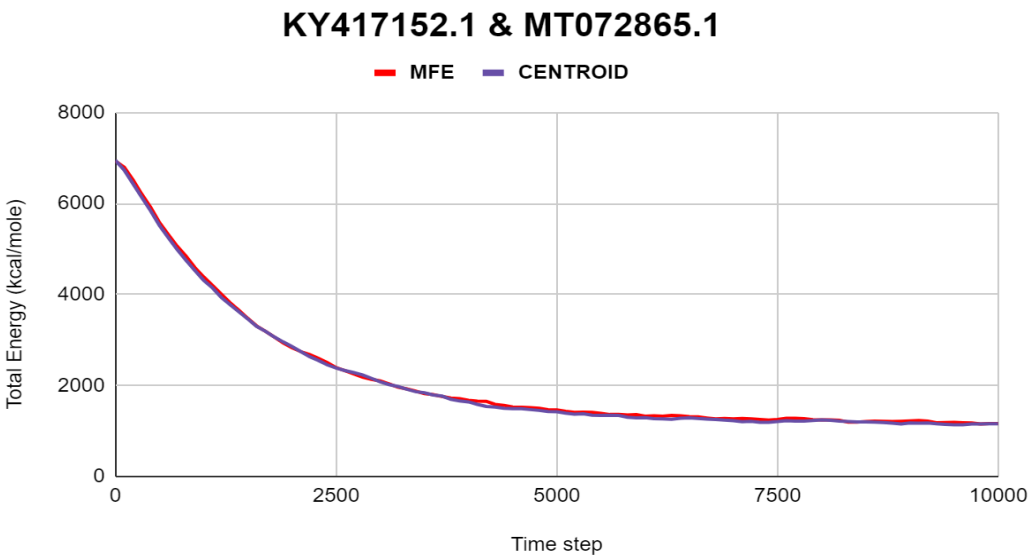
1.402124



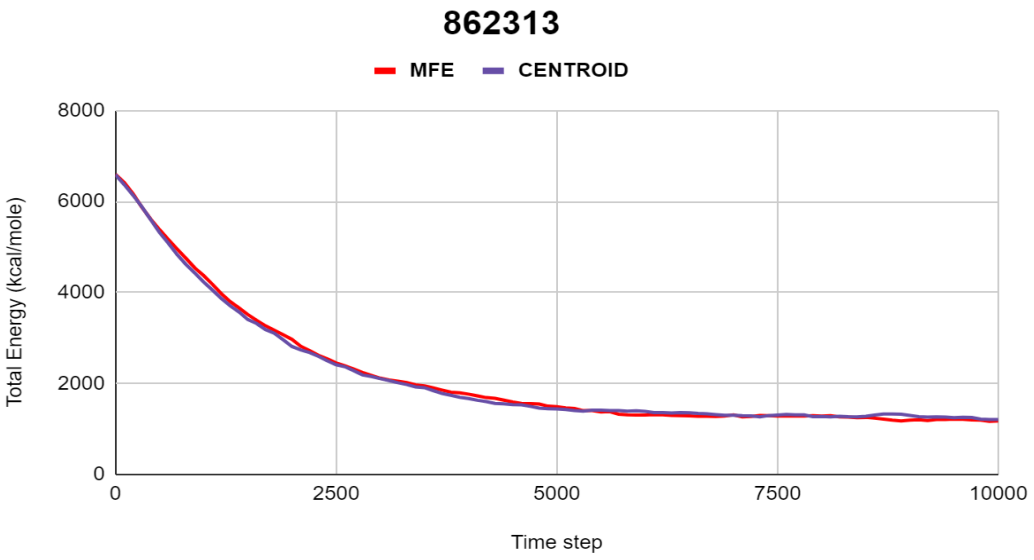
2.MT308984.1



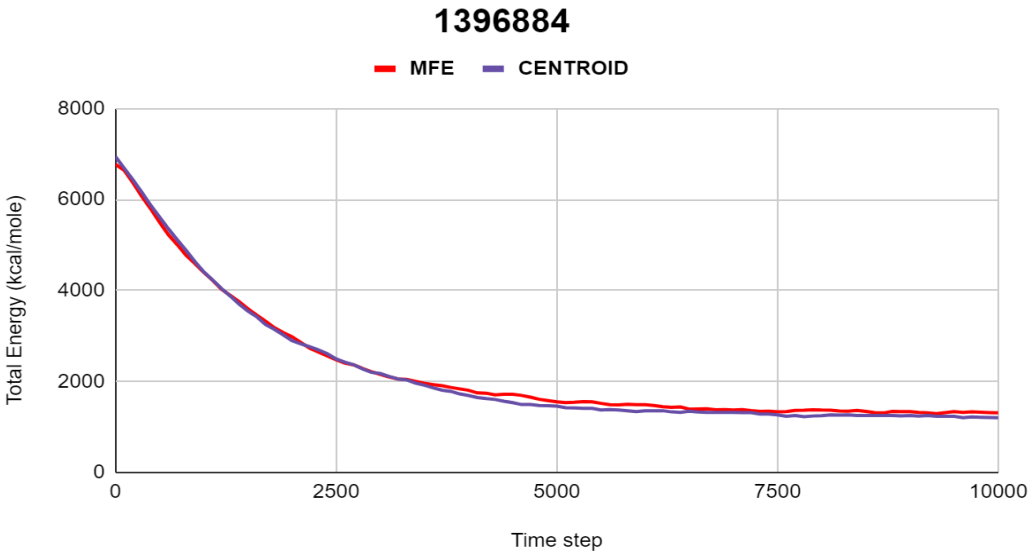
3.KY417152.1 & MT072865.1



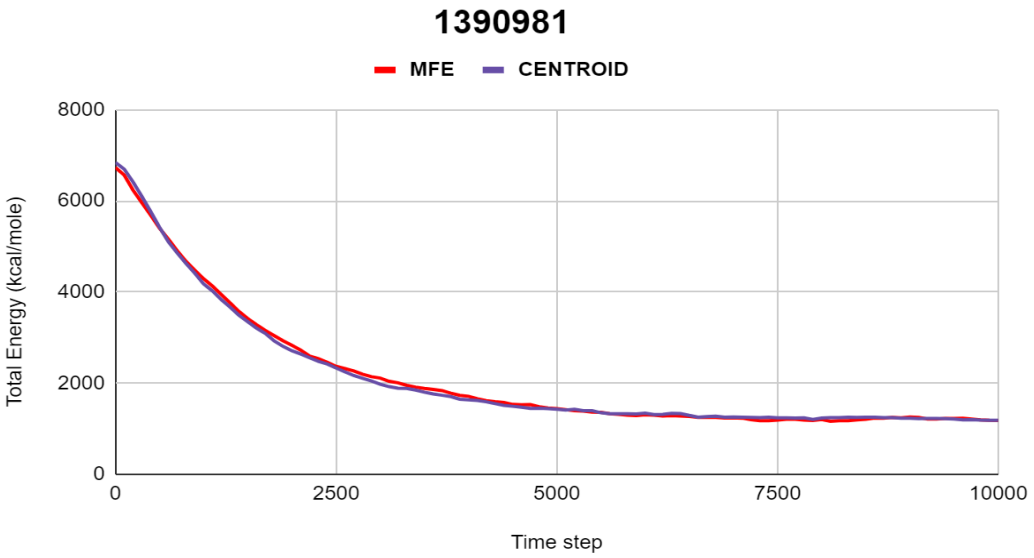
4.862313



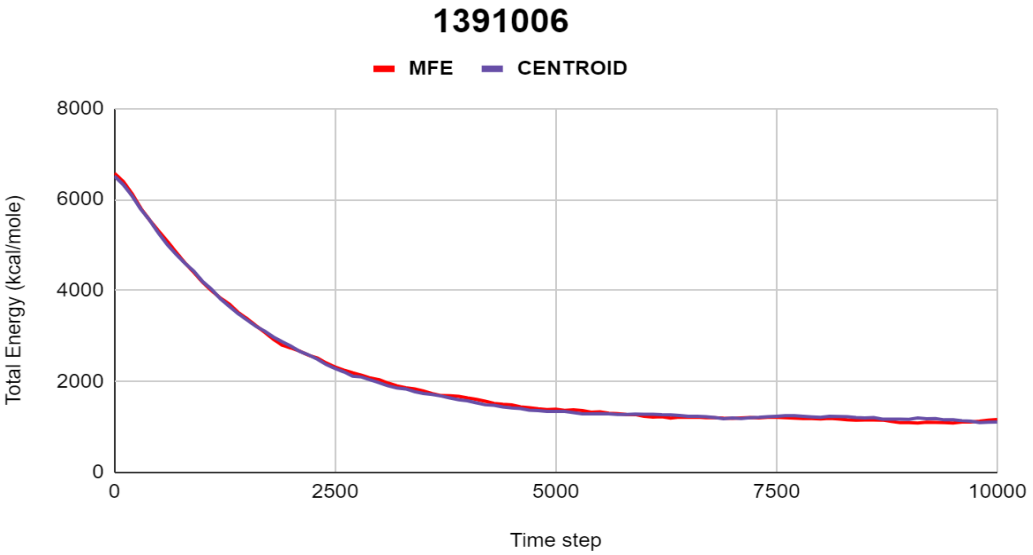
5.1396884



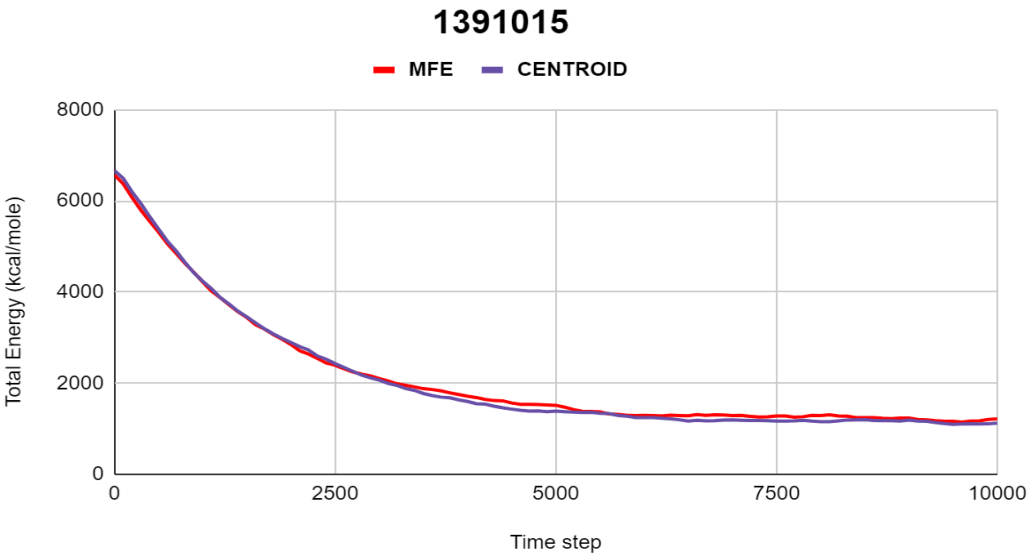
6.1390981



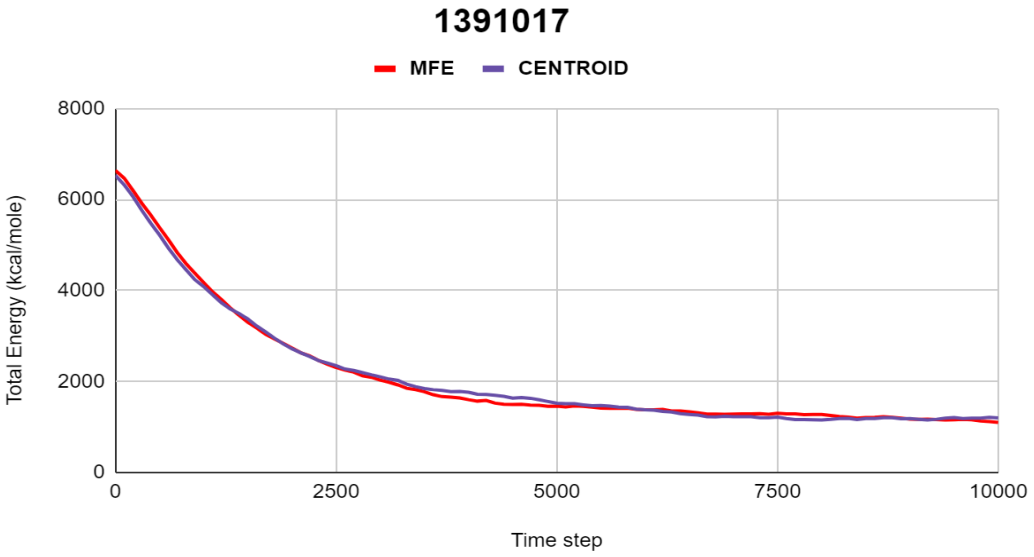
7.1391006



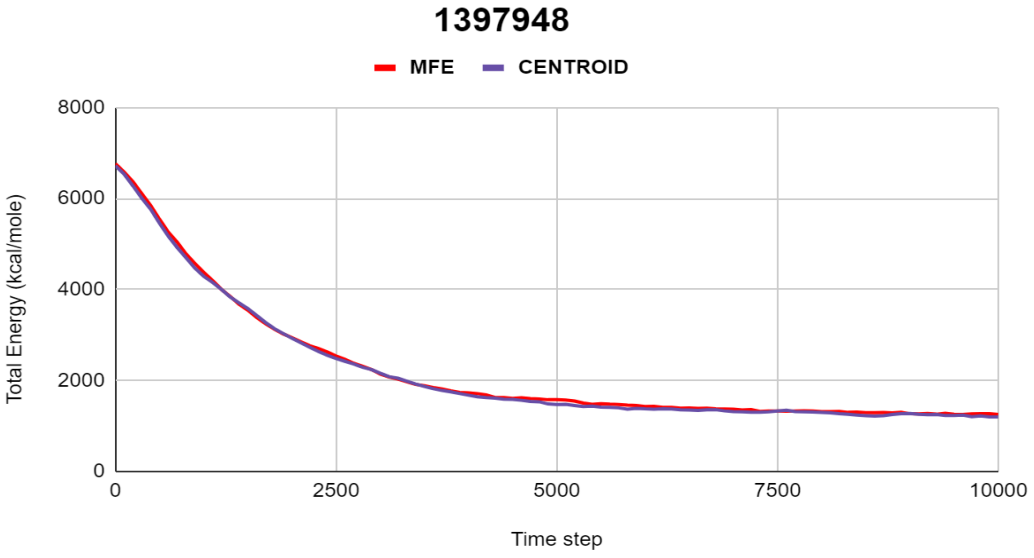
8.1391015



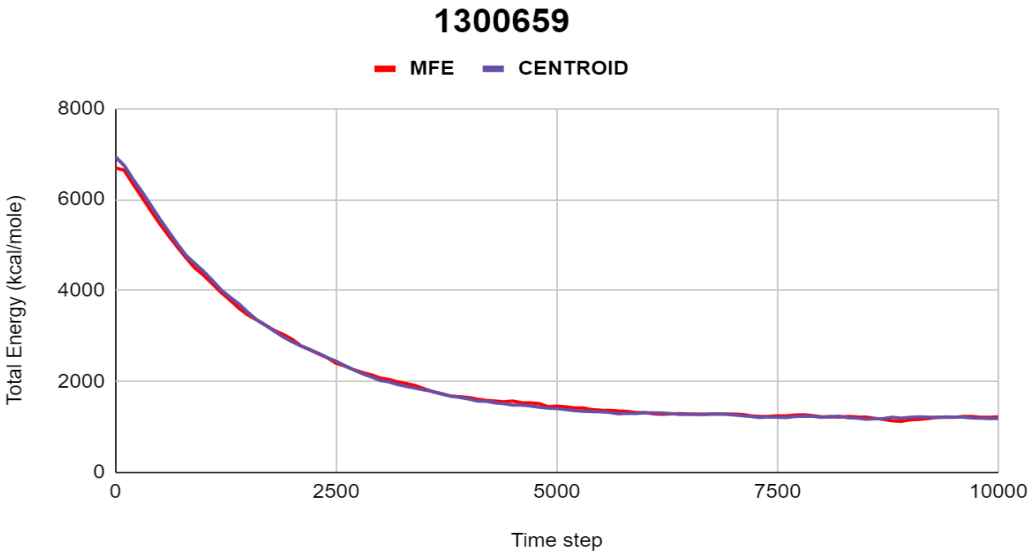
9.1391017



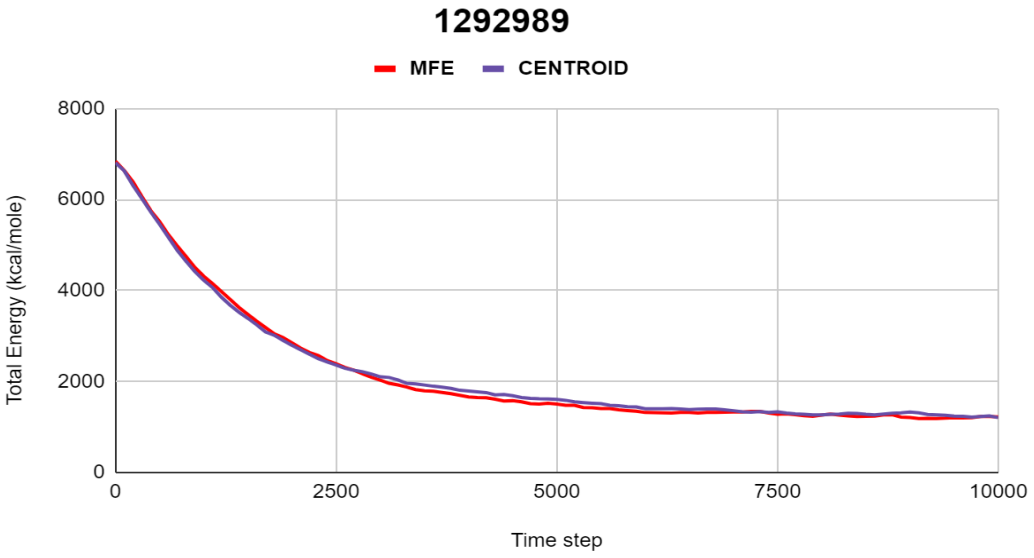
10.1397948



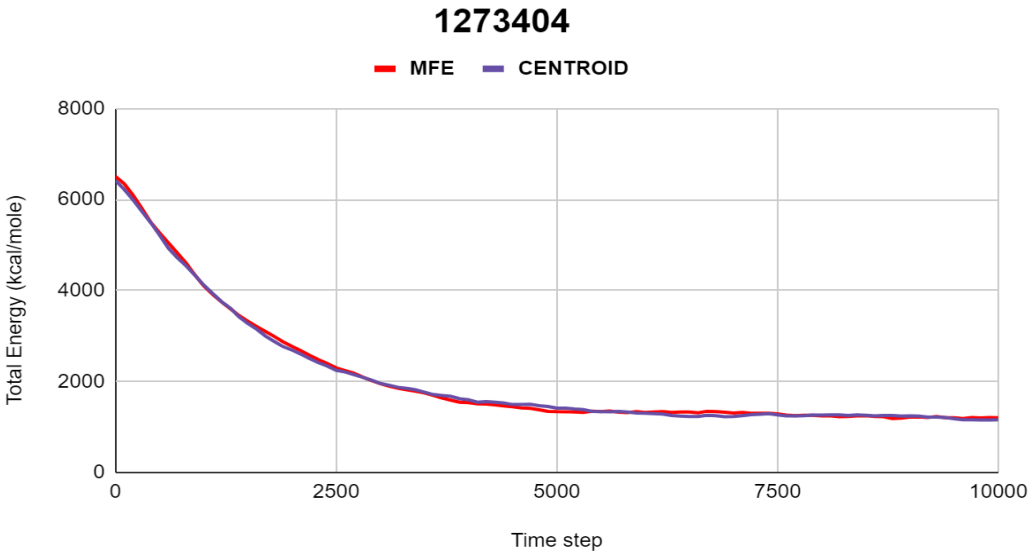
11.1300659



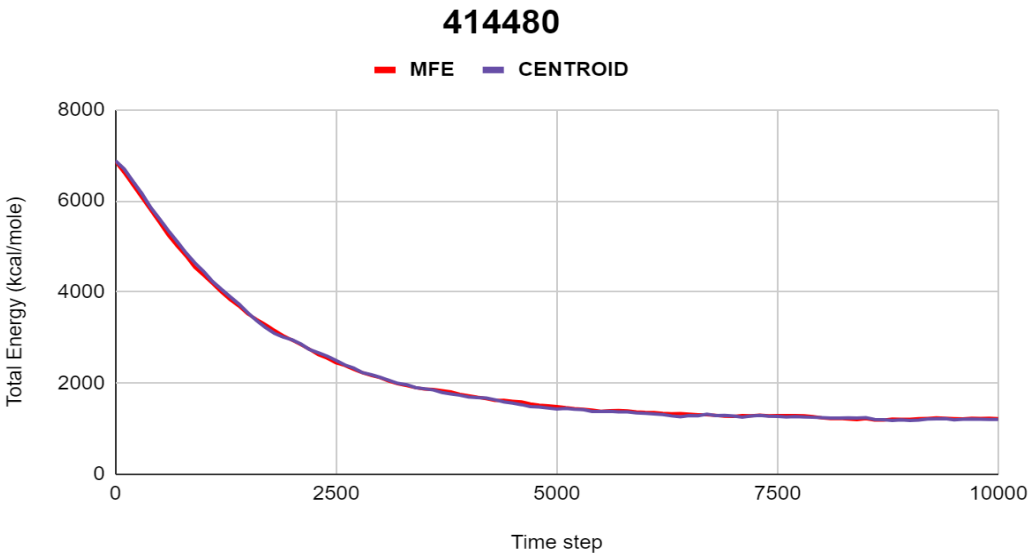
12.1292989



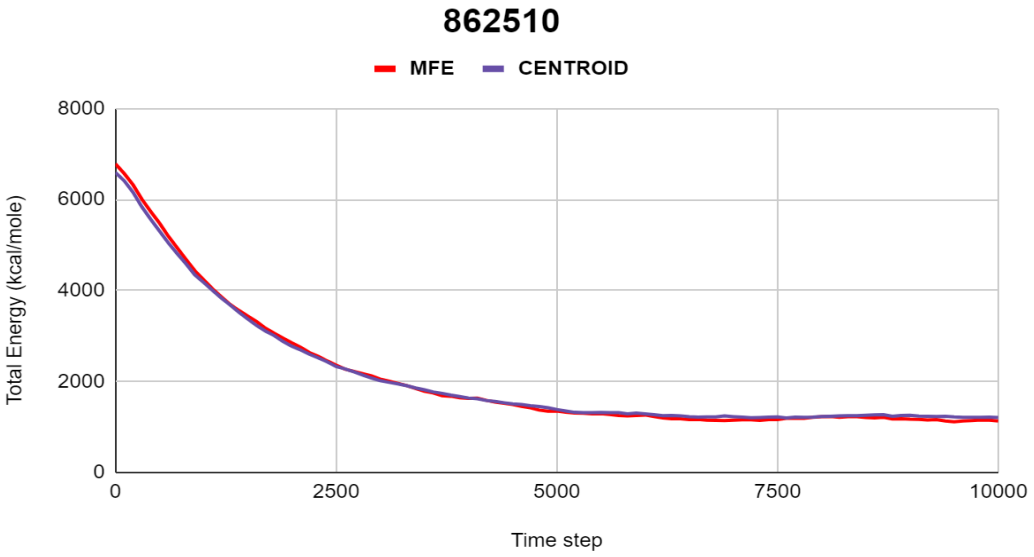
13.1273404



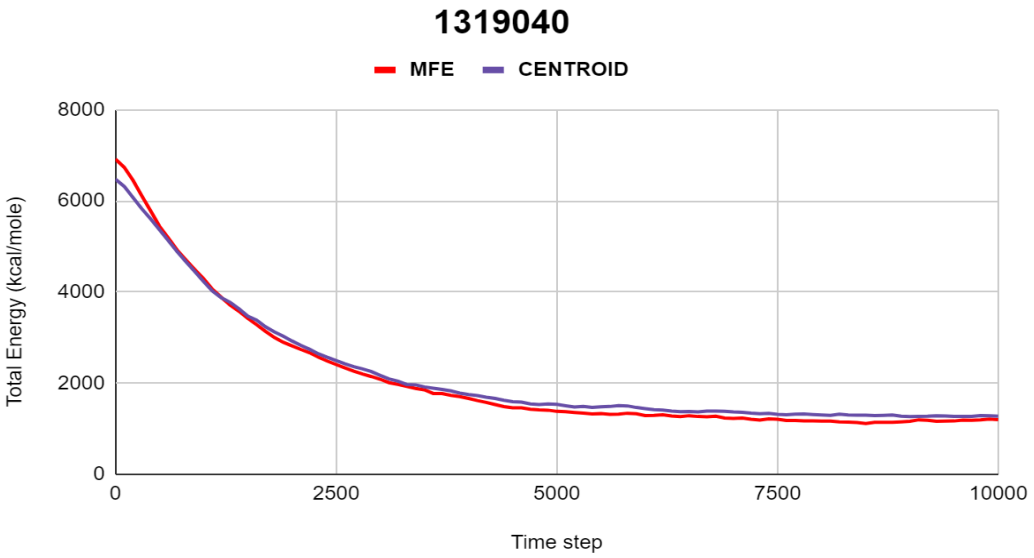
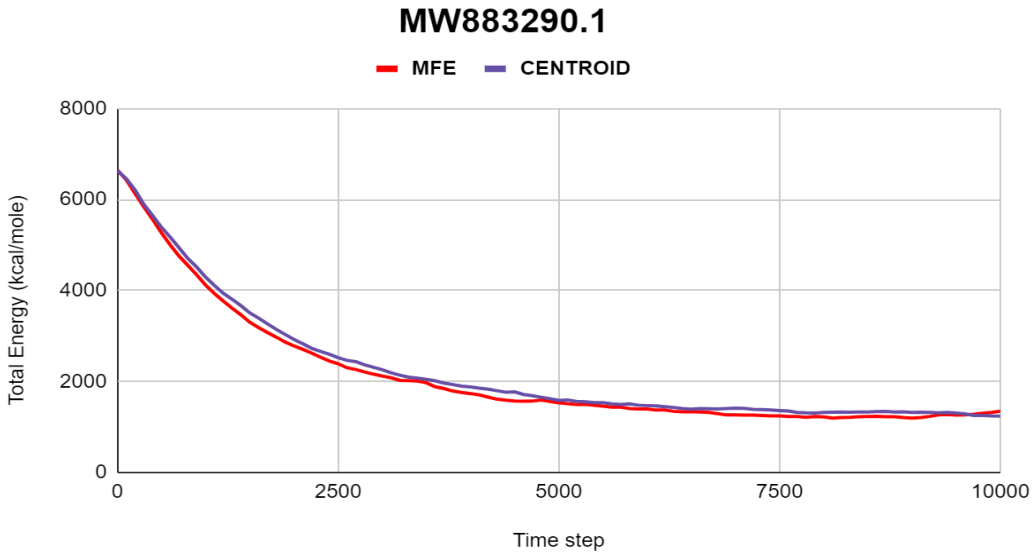
14.414480



15.862510



16.1319040



17.MW883290.1