Dedicated to the ever bright memory of my parents.

PhD Thesis

Genomic Sequence Analysis of Noncoding RNA

Submitted to the

Cochin University of Science And Technology

in partial fulfilment of the

requirement for the award of the degree of

Doctor of Philosophy

Under the faculty of Technology

by

TINA P G

Under the supervision of

Prof. Dr. Tessamma Thomas

Department of Electronics

Cochin University of Science And Technology

Cochin

December 2017

Genomic Sequence Analysis of Noncoding RNA

Ph.D Thesis in the field of Genomics Signal Processing

Author

Tina P G

Research Scholar

Department of Electronics

Cochin University of Science And Technology

Cochin – 682022

Kerala, India

email : ptinageorge@gmail.com

Supervising Guide

Dr. Tessamma Thomas

Professor

Department of Electronics

Cochin University of Science And Technology

Cochin – 682022

Kerala, India

email : tessamma1@gmail.com

CERTIFICATE

This is to certify that the work presented in the PhD Thesis titled **"*Genomic Sequence Analysis of noncoding RNA*"** is a bonafide record of the of the original research carried out by Mrs. Tina PG at the Department of Electronics, Cochin University of Science and Technology, under my supervision.

Kochi
29/12/2017

Dr. Tessamma Thomas
(Supervising guide)
Department of Electronics
Cochin University of
Science And Technology
Kochi - 22

Certificate

This is to certify that all the relevant corrections and modifications suggested by the audience during the pre-Synopsis seminar and recommended by the Doctoral Committee of Mrs. Tina PG has been incorporated into the PhD Thesis titled "**Genomic Sequence Analysis of noncoding RNA**" which is being submitted by her.

Kochi
29/12/2017

Dr. Tessamma Thomas
(Supervising guide)
Department of Electronics
Cochin University of Science And Technology
Kochi - 22

DECLARATION

I, Tina PG, hereby declare that the work presented in this PhD thesis titled **"*Genomic Sequence Analysis of noncoding RNA*"** is based on the original research done by me, under the supervision of Dr. Tessamma Thomas, in the Department of Electronics, Cochin University of Science and Technology and that this work did not form any part of any dissertation submitted for the award of any degree, diploma, associate-ship or any other title or recognition from any other University/Institution.

CUSAT
26/12/2017

Tina PG
Research Scholar
Department of Electronics
Cochin University of Science
And Technology
Kochi - 22

Acknowledgement

I bow before The Almighty, my strength and my fortress without whose Grace my very existence is meaningless.

I express my deepest gratitude to my Research Supervisor, Dr. Tessamma Thomas for her guidance and encouragement throughout the period of this work. I deem it my proud privilege for being able to work under an eminent academician like her. I express my heartfelt thankfulness to Dr. Supriya M H, Professor, and Head of the Department of Electronics, CUSAT for being supportive of all my activities during the period of my research. I thank Prof. Dr. K Vasudevan from the bottom of my heart for timely advises and for being a constant source of encouragement to me. I am extremely grateful to Prof. Dr. C K Aanandan for the help and encouragement I have received from him. I place on record my thankfulness to Prof. Dr. PRS Pillai, Prof. Dr. P Mohanan, Prof. Dr James Kurien, Dr. Bijoy A Jose, Mr. Arun A Balakrishnan, Mr. Mithun Haridas T P, for their help and support. I also place on record my earnest thankfulness towards all the technical, non-technical and administrative staff of the Department of Electronics, CUSAT for their help and support.

I thank from the bottom of my heart Ms. Suja S, Ms. Sangeetha R, Ms. Deepthi, Ms. Binthiya Suny Gabriel, Ms. Mini P P, and Mr. Anu Sabarish, my co-researchers and friends at the AIRL, for the help, encouragement, and support I have received from them. I thank Mr. Aji George, Mr. Mithun M S, Ms. Navya Mohan Research Scholars, IMSL and Mr. Lindo Ouseph, Research Scholar, RCS Lab, for the help I have received from them. I thank Mr. Suraj Kamal, Research Scholar, INSPIRE, for his valuable comments and suggestions. I thank all my fellow researchers at the Department of Electronics, CUSAT for the help, support, and encouragement received.

I also express my sincere thanks to The Director, CAPE, Govt. of Kerala and The Principal, College of Engineering, Kidangoor for granting me the leave to complete my research.

Words fall short when it comes to thanking the most precious people. But I am making an effort to. I thank my husband, Dr. Joseph Edison with all my heart for meticulously copy-editing my thesis with utmost patience and for being there for me, always. I express the most loving appreciation I have for my boys Reuben and George for being very understanding and very supportive.

I thank all my friends, family, and well-wishers whose encouragement made this period a lot easier for me.

List of publications

Tina P George, Tessamma Thomas. **Exon mapping in long noncoding RNAs using digital filters.** Genomic Insights. 2017;10:1-12. Available at : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624354/

Tina P George, Tessamma Thomas. **Novel Approach to analyzing MFE of Noncoding RNA Sequence**s. Genomics Insights. 2016;9:41–49 doi:10.4137/GEI.S39995. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5029481/

Tina P George**,** Tessamma Thomas. **Discrete Wavelet Transform De-noising In Eukaryotic Gene Splicing**, BMC Bioinformatics 2010. 11(Suppliment 1) S50,10.1186/1471-2105-11-S50.Available at : http://www.biomedcentral.com/content/pdf/1471-2105-11-s1-s50.pdf

Abstract

The genetic code for every organism is stored in bio-molecules called nucleic acids. There two types of nucleic acids, the deoxyribonucleic acid (DNA) and the ribonucleic acid (RNA). In higher organisms DNA is found inside the nucleus and RNA outside the nucleus. DNA has two regions – the gene and the non-gene. Gene is the portion of the DNA that is directly responsible for coding of proteins which are needed by the organism. RNA has many support functions. In the early days of genome studies, DNA was thought to be of prime importance and all of the studies were DNA-centric. The non-gene portions of the genome as well as the RNA were ignored to a large extent in genome studies. RNA molecules were thought to be intermediary entities in the formation of protein from DNA. But systematic screening of the genome of various organisms has shown that there are set of noncoding RNA sequences encoded by the noncoding region of the genome. There are yet another range of noncoding RNAs which have been called long ncRNA (lncRNA). Although two thirds of the human genome gets transcribed, only 2% of the transcribed genome encodes proteins. It has been found that the remaining gets converted into long ncRNA molecules too, among other ncRNAs. They were thought of as "transcriptional noise" even in the genomic era. However, they have assumed prime importance in molecular biology in the current decade after their varied functions have been unveiled. Epigenetic regulation, chromatin modelling, gene transcription, protein transport, protein trafficking, cell differentiation, organ or tissue development, cellular transport, metabolic processes and chromosome dynamics are just a few examples of lncRNA functions.

This thesis is contains the results of the studies based on digital signal processing techniques done on noncoding RNA (ncRNA) sequences taken from bench-marked, public databases. Four classes of sequences of noncoding RNA molecules are studies here. snRNA (small nuclear RNA), snoRNA (small nucleolar RNA), miRNA (micro RNA) and rRNA (ribosomal RNA). Each of which have specific functions in various stages of protein formation and gene expression. They play vital roles in the formation of protein and expression /suppression of genes. The function of these ncRNA is decided by its secondary structure, and across organisms, the secondary structure is more conserved than the sequence itself. In the first part of this work, the optimal secondary structure or the minimum free energy (MFE) structure of a sample of around 2500 sequences of non-coding RNA molecules belonging to four above mentioned different classes is found out based on the thermodynamic nearest neighbour model. Mathematical models linking MFE to the signal properties are found out for each of the four classes of ncRNA analyzed. It is seen that found that the MFE values computed with the proposed models are in concordance with those obtained with the standard web servers. 95% of the sequences analyzed had deviation of MFE values within +/-15% relative to those obtained from standard web servers.

The second part of this work analyses sequences which are called long noncoding RNAs (lncRNA). These molecules have assumed prime importance in genome studies in the recent two decades after the discovery that the play crucial roles in almost all stages of biological regulation. These molecules have been implicated in various diseases and play vital roles in various developmental processes. Recent studies emphasise the need for in-depth study of the sequences, their structural features, and genomic architecture. Here in this work, we perform mapping of exons of human lncRNA sequences taken from NCBI GenBank, making use of digital filers. Anti-notch filters are used to locate exons. The period 3 property which is an established indicator for locating exons in genes is used here. The discrete wavelet transform filter bank is used to de-noise the exon plots. In an earlier work, a quadratic filter was successfully used by the authors to bring down the spectral noise while mapping exons of coding regions. However, it is found that this quadratic function introduces additional spectral noise when used with lncRNA sequences. This indicates that the sequence spectrum of lncRNAs cannot be amply represented by the A-T spectra alone as in protein coding genes. As reported in literature, G-C concentration in lncRNA sequences is seen to be less than 50%, which is much lower than that found in coding regions. It is seen that none of the sequences analysed have STOP codons although different START codon patterns are found in them. The exon maps show exon locations that conform to the ranges specified in GenBank. The spectral noise in the exon map of lncRNA occupies the same frequency ranges as that of coding regions. From this we can conclude that the period 3 property and the de-noising techniques used for exon prediction in genes can be extended to lncRNAs too.

GENOMIC SEQUENCE ANALSIS OF NONCODING RNA

APPENDIX

Appendix – I

1. List of tables

2. List of figures

3. List of abbreviations

DFT	-	Discrete Fourier transform
DNA	-	Deoxyribonucleic acid
DSP	-	Digital signal processing
DWT	-	Discrete wavelet transform
ENCODE	-	Encyclopaedia of DNA Elements
HGP	-	Human Genome Project
NHGRI	-	National Human Genome Research Institute (US)
NIH	-	National Institute of Health (US)
NLM	-	National Library of Medicine (US)
RNA	-	Ribonucleic acid
NHGR I Institute (US)	-	National Human Genome Research
NCBI Information (US)	-	National Centre for Biotechnology
ncRNA	-	noncoding RNA
lncRNA	-	long non coding RNA
STFT	-	Short time Fourier transform
DWT	-	Discrete wavelet transform
ORF	-	Opening reading frame

Chapter 1

Introduction

1.1. Genetics and Genomics

Genetic studies started in 1856 with the work of Gregor Mendel who, with his experiments on peas discovered that certain traits are passed on from the parent to the offspring through entities called ‘genes’. Genes were later on discovered to be located within the chromosomes which were known to contain DNA (de-oxyribonucleic acid) and proteins. Initially, proteins were mistaken to contain genes. The fact that genes are found within the DNA was discovered much later.

The helical structure of DNA as we know it today was revealed in 1953 by the work of Watson & Crick. This discovery of the double-helical structure of DNA is the milestone in the history of natural science and gave rise to modern molecular biology. There has been dramatic progress in genomics in the last seven decades. We are now in the genomic era, with the human genome project completed in 2003. Today large amounts of genomic and proteomic data are available in the public domain and it is to be processed in ways which are beneficial to mankind. Genomic signal processing is primarily the processing of DNA sequences, RNA sequences, proteins and other forms of genomic data viz. DNA microarray images, fluorescent in-situ hybridization (FISH) images etc. Traditional as well as modern signal processing methods find wide application in this area.

A DNA sequence is made from an alphabet of four elements, namely *A, T, C,* and *G* which represent the four bases or the four nucleotides (Adenine, Thymine, Cytosine and Guanine). For RNA sequences, T is replaced with U, as Uracyl is the nucleotide base present in RNA sequences in the place of Thymine [Watson 2007], [Alberts 2007] Since DNA contains the genetic information of living organisms, we see that life is governed by a code comprised of an alphabet of four letters (A, T/U, C, G). Another example of discrete-alphabet sequences in life forms is protein which controls a large number of functions in living organisms. A protein can be regarded as a sequence of amino acids. There are twenty distinct amino acids and therefore, a protein can be regarded as a sequence defined on an alphabet of size twenty.

Genomic information available in public data sources are in the form of character strings. Appropriate mapping of these letter strings into numerals is mandatory in order to apply digital signal processing methods to analyse them [Voss 1992], [Anastassiou 2001(1)]. Genomic information is digital in a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, like DNA, RNA and proteins, can be easily represented as mathematical sequences. If we assign appropriate numerical values to the four letters in the DNA sequence, the genomic sequences become conducive to signal processing.

1.2. Non coding RNA and its relevance

In simple terms, DNA sequence contains two regions, the gene and the non-gene [Watson 2007], [Alberts 2007]. In the early days of the genomic era, which was the last decade of the previous century, much of the genome studies were DNA and gene centric, as DNA was found to be the carrier of genetic information. The region of the DNA namely, the non-gene was largely ignored. But with discoveries that the non-gene contains valuable information that codes for other regulatory nucleic acids, namely noncoding RNAs (ncRNA) the focus has shifted to the noncoding region of the genome [Eddy 2001], [Gisela 2002], [Mattick 2006], [Ponting 2010], [Kung 2013], [Chen 2017]. A more detailed discussion of this background is given in Chapter 3. RNA molecules can interact with DNA and other RNA molecules in diverse ways which make them vital to development of the organism as a whole [Eddy 2001]. It has been discovered that non coding RNAs are extensively involved in wide range of regulatory mechanisms and are also implicated in the development of diseases [Erdmann 2001], [Gottesman 2002]. Noncoding RNA molecules have been found to be involved in almost all stages of cell biogenesis. This work analyses noncoding RNA sequences using established digital signal processing methods.

1.3. Numerical representation of the genomic code

As seen, the DNA/RNA sequence is made up of four letters of the alphabet viz., A, T/U, C, G. The letters are replaced with appropriate numerals. Currently there are many techniques to perform the mapping of the genomic sequences into numbers. The earliest and the most popular of the techniques is to represent the genomic sequence using four binary indicator sequences $x_A(n), x_T(n), x_C(n), x_G(n)$ each of which are binary sequences of length ‘n’ with values 1 or 0 depending on whether the corresponding nucleotide base is present or absent at the location ‘n’ [Voss 1992]. ‘n’ would be the length of the genomic sequence considered. Such that,

$$x_A(n) + x_T(n) + x_C(n) + \ x_G(n) = 1 \qquad (1.1)$$

Scaling of the indicator sequences, $a. x_A(n), t. x_T(n), c. x_C(n), g. x_G(n)$ with appropriately chosen a, t, c, g is also done [Anastassiou 2001(1)]. The mapping can be done to the Electron Ion Interaction Potential (EIIP) of the different bases too [Novysh 1997], [Nair 2006].

As explained in Chapter 7 of this thesis, exons present in the long noncoding RNA sequences (which are coded from the non gene region of DNA sequences) have been mapped out. Binary indicator sequences have been used to represent the long noncoding RNA sequences [George 2017]. A novel model to compute the thermodynamic entity, minimum free energy (MFE) [Xia 1998], [Tinoco 1999], [Zuker 2000], [Trotta 2014] of noncoding RNA sequences from their signal properties has also been developed in this work [George 2016]. For which, the convention of complex notation [Cristea 2002] has been used to represent RNA sequences. Genomic sequence x(n) is converted into the four indicator sequences as seen above. The bases for the RNA sequences are A, U, C and G. The indicator sequences would be $x_A(n), x_U(n), x_C(n), x_G(n)$ respectively. Multipliers a, u, c, g are selected such that the RNA sequence $x(n)$ can be expressed as,

$$x(n) = a. x_A(n) + u. x_u(n) + c. x_C(n) + \ g. x_G(n) \qquad (1.2)$$

And the multipliers are, $a = 1 + j,\ u - 1 - j,\ c = -1 + j,\ g = -1 - j.$ [Cristea 2006]

1.4. MFE, Length, Standard deviation of spectral coefficient matrix of ncRNA sequences

Minimum Free Energy (MFE) is a thermodynamic feature of ncRNA sequences which decides the optimal secondary structure and hence their function [Clote 2005], [Hofacker 2002]. The parameters sequence length and MFE have been used in analyzing RNA from a very early time [Grüner 1996], [Galzitskaya 1998]. There have been studies which explore the influence of length and MFE on sequence stability [Pervouchine 2003], [Trotta 2014]. In this work, MFE of ncRNA sequences was analyzed with respect to its relationship to the sequence length and the standard deviation (SD) of spectral coefficients.

The noncoding RNA sequences analysed in this work are ribosomal RNA (rRNA), micro RNA (miRNA), small nuclear RNA (snRNA) and small nucleolar (snoRNA). The sequences are analysed and related to the signal properties of these sequences namely length and the standard deviation of the spectral coefficient matrix of these sequences. It was found that MFE could be linearly related to both the length as well as the standard deviation of the spectral coefficient matrix of the sequences for all four classes of the ncRNA analysed. Making use of this association of MFE with length and SD of the spectral coefficient matrix of the sequences analysed, a model to evaluate MFE from the signal parameters has been developed for each class of the ncRNA analysed.

1.5. Statistical tools

The mathematical relationship between MFE – sequence length and MFE – SD of spectral coefficient matrix of ncRNA sequences analysed is arrived at by making use of simple linear regression (SLR) analysis [Kirchner 2001], [Montgomery 2006].

Equations of the form

$$y = m_1x_1 + c_1 \tag{1.3}$$

$$y = m_2x_2 + c_2 \tag{1.4}$$

linking MFE (y), sequence length (x_1) and SD of spectral coefficient matrix (x_2) are developed making use of multiple linear regression (MLR) analysis. This analysis was done for a sample space comprising of a total of 120 specimen, 30 from each class. As the relationship was found true for all the 4 classes of ncRNA analysed, the analysis was extended for a larger sample space for each of the 4 classes of ncRNA. A mathematical model was developed using multiple linear regression (MLR) [George 2016]. A total of 2656 sequences were analysed of which 902 were snRNA, 805 rRNA, 376 miRNA and 573 snoRNA. The model for MFE (y) arrived at from signal length (x_1) and SD of the spectral coefficient matrix (x_2) has the following format.

$$y = m_1x_1 + m_2x_2 + b \tag{1.5}$$

1.6. Frequency spectrum analysis of genomic signals

Signal processing methods have been used successfully on genomic sequences, especially the coding region of the genome [Anastassiou 2001], [Anastassiou 2000], [Kakumani 2008], [Vaidyanathan 2004]. It could be said that frequency spectrum analysis has assumed primary importance in the study of genomic sequences, because of certain properties which they possess have enabled better understanding of the genome outside the wet-labs [Tiwari 1997], [Vaidyanathan 2002], [Anastassiou 2000], [Anastassiou 2001(1)], [George 2010]. It was found right from the early days of genome studies that the frequency spectrum of DNA sequences has special properties [Voss 1992], [Tiwari 1997]. Voss calculated the deterministic auto-correlation for the indicator sequences of the bases. The autocorrelation coefficient $r_A(k),$ for the indicator sequence $x_A(n),$ for nucleotide A is given as

$$r_A(k) = \sum_n x_A(n)\, x_A(n-k) \qquad (1.6)$$

The Discrete Fourier Transforms of the indicator sequences, computed as per the classical equations [Proakis 2006], [Oppenheim 2009] can be indicated by $X_A(k), X_T(k), X_C(k), X_G(k).$

The DFT of any discrete N point sequence x(n) is another discrete N point sequence given by

$$X(k) = \sum_{k=0}^{N-1} x(n) e^{-jk2n\pi/N} \qquad (1.7)$$

1.7. Period 3 property of the spectrum of genomic sequences.

It has been noticed from very early days [Triffanov 1980] that the protein coding regions (genes) of DNA sequences have a period 3 component. And this property has been effectively used to locate exons [Anastassiou 2001(1)], [Vaidyanathan 2002] within the genes and also to locate genes themselves [Tiwari 1997].

Let S(k) be defined as

$$S(k) = |X_a(k)|^2 + |X_t(k)|^2 + |X_t(k)|^2 + \left|X_g(k)\right|^2 \qquad (1.8)$$

Where $X_A(k), X_T(k), X_C(k), X_G(k)$ represent the short time Fourier transforms (STFT) of the four binary indicator sequences taken such that N is a multiple of three. In such a situation if the length of the sequence considered x(n) contains exons, we would be able to see a peak in the plot of S(k) at k=N/3 which would correspond to discrete frequency $\omega = 2\pi/3^C$, remembering the fact that the range of discrete frequency ω^C in any DFT.

1.8. Noise removal in genomic signal spectrum

The signal peak in the power spectrum of gene regions of the DNA sequences can be picked out using a filter that has maximum gain at the frequency $\omega = 2\pi/3^C$ and very high attenuation in the remainder of the frequencies. That is, a filter with the inverse properties of a notching filter or in other words single peaking filter at $\omega = 2\pi/3^C$. This filter is also called an anti-notch filter [Vaidyanathan 2002]. The filter used here is the IIR single peaking filter with high attenuation in the stop band and very high gain in the passband. Representation of the magnitude response of a single peaking filter at $\omega = 2\pi/3^C$ is shown in Figure 1.1 below. The filter used in this work is of Direct form II implementation.

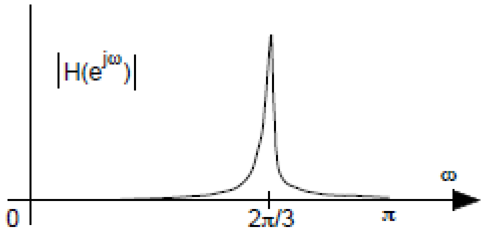


Figure 1.1. Magnitude response of the single peaking filter

The exon plot obtained is further refined using the Discrete Wavelet Transform filter bank which performs sub-band coding [Soman 2004]. Filtering of a signal x(n) with a filter of impulse response h(n) is given by the following time domain equation.

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k].h[n-k] \tag{1.9}$$

The DWT is equivalent to two bands of filtering, a lower sub-band and an upper sub-band. Thus the signal is split into two frequency bands.

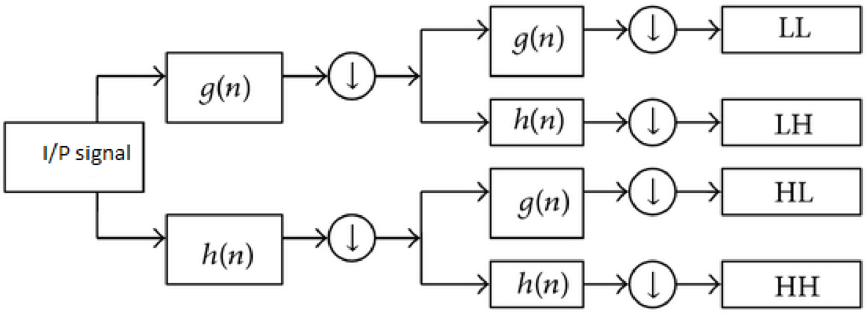


Figure 1.2. Sub-band coding using DWT

The upper band filter g(n) removes all the lower half of the frequencies and gives the upper band as output, and the lower band filter h(n) removes all the upper half of the frequencies contained in the signal being treated with the DWT. Mathematically, we can represent this with a couple of equations as follows.

$$y_{high}[k] = \sum_n x[n]\,.\,g[2k-1] \tag{1.10}$$

$$y_{low}[k] = \sum_n x[n]\,.\,h[2k-1] \tag{1.11}$$

This is decimation or down-sampling. Noise removal is achieved by avoiding the sub-band which contains the noise frequencies while reconstructing or upsampling. Effective noise removal from the exon plots can be achieved using DWT by employing the apt wavelet and the appropriate number stages of decimation and re-construction [George 2010], [George 2017].

1.9. Motivation and objectives of the work

The research reported in this thesis is in the area called Genomics Signal Processing which is the application of Digital signal Processing techniques to genomic signals.

Ever since the double-helical structure of DNA was revealed in 1937, there has been dramatic advancements in molecular biology and related studies. All characteristics of living beings are determined by the gene sequences. This includes biogenesis and pathogenesis of all living organisms. Thus the study of genome data is undoubtedly much beneficial to humanity.

Noncoding RNA molecules were ignored for a long time in genome studies. But they have come to be of vital importance in both molecular biology as well as in genome studies as it has become evident that they play vital roles in many biological processes. The work presented in this thesis analyses noncoding RNA sequences. Computational methods have been used widely in the study of both protein coding and noncoding regions of the genome. However, analysis of the noncoding genome using DSP methods has not been reported in literature. In this work, signal processing techniques are made use of to analyze the noncoding portion of the genome. In the first part of this work, MFE, the thermodynamic energy which decides the secondary structure and thereby the function of small noncoding RNA is analysed. In the latter part of the work, molecules called long noncoding RNA are analysed. Long noncoding RNA sequences have been found to possess exonic regions and in this work, the exons in these sequences are mapped out using digital filtering techniques.

1.10. Organization of the Thesis

This thesis is organized into eight chapters as follows.

Chapter 1 : Introduction

Chapter 1 gives an overall introduction to the work presented in the thesis. The molecular biology background over which this work was conceived is mentioned. A brief overview of the digital signal processing methods and the statistical techniques used in the work is also given.

Chapter 2 : Literature Review

A brief look at the classical computational methods used for biomolecular sequence analysis is presented in chapter 2. The existing methods, both DSP based and computational techniques which are relevant to this study are also presented.

Chapter 3 : Molecular Biological background of the study

Fundamentals of Molecular Biology which forms the backdrop of the study is presented in this Chapter. The relevance of studying the sequences selected in this work, namely non coding RNA sequences is also presented.

Chapter 4 : MFE based Prediction of ncRNA Secondary Structure

Chapter 4 introduces the reader to the concept of secondary structure of RNA and its importance. The results of secondary structure prediction of a small sample noncoding RNA sequences using the classical thermodynamic nearest neighbour model and the MFE computed are presented.

Chapter 5 : Novel relationship between MFE and Signal Parameters of the ncRNA Sequences

In this Chapter, the relationship between the Minimum Free Energy (MFE) of the RNA secondary structure and the signal parameters of the

RNA sequence viz., the length of the sequence and its spectral coefficients are explored. The spectral coefficients of the nucleotide sequence are obtained by making use of the Discrete Fourier Transform via the FFT algorithm. A novel linear relationship has been arrived at, between the values of MFE and nucleotide length, MFE and the standard deviation of spectral coefficient matrix of the sequences analysed using simple linear regression.

Chapter 6 : Novel mathematical model for MFE

Chapter 6 presents how a novel mathematical model for MFE of noncoding RNA sequences is arrived at, from their signal properties. The model is developed using the statistical analysis tool, multiple linear regression. The models developed are made use of to evaluate MFE from signal parameters and the correctness of the models is checked with webservers RNAfold and RNAstructure.

Chapter 7 : Exon mapping in lncRNA using digital filters

In Chapter 7, exon mapping of human lncRNA sequences (taken from NCBI GenBank) using digital filters is presented. During the initial days of genome studies and even up till the last decade, long noncoding RNAs (lncRNA) were dismissed as “transcriptional noise”. However, they have become a vital area of study from the beginning of this decade after their roles in biological regulation in various developmental processes and diseases were discovered.

Chapter 8 : Conclusion and future of the study

This chapter concludes the thesis and includes the future scope of this study.

Chapter 2

Literature Survey

2.1. Introduction

This chapter presents the review of literature done for this work in relation to the analysis of genomic data with special importance to the area of Genomics Signal Processing. In this literature review the background of genetic and genomic studies is briefed so that the reader understands the relevance of this study and why the investigations done in this study were carried out.

2.2. History of genetic studies

Genetic studies began with the work of Gregor Mendel in 1856. Gregor Mendel was an Austrian monk who discovered the basic principles of heredity through experiments with pea plants in his garden [Watson 2007], [Alberts 2007]. Mendel is known as “the father of modern genetics”. With his experiments, Gregor Mendel proved that certain “factors” are inherited by the offspring from the parent. Almost half a century later, Walter Sutton and T H Morgan of the Columbia University discovered that the key to genetic inheritance was present in chromosomes which were found in the nucleus. Chromosomes were known to contain proteins and DNA molecules. In 1930 DNA was found to be a long molecule made of nitrogenous bases, Adenine (A), Thymine (T), Cytosine (C), and Guanine (G) [Watson 2007]. Initially proteins were thought to be the entities that carried genetic information. The experiments of OT Avery in 1944 proved that it was DNA which carried the genetic traits and not proteins. The historic discovery of the double-helical structure of DNA was made by James Watson and Francis Crick in 1953. This discovery revolutionized science and marked the birth of modern Molecular Biology [Pray 2008].

The growth of research in genetics lead to the idea of the Human Genome Project (HGP) to be conceived and carried out. The HGP was intended to map out the entire genome of human beings. The HGP could be thought of as the natural culmination of genetic research which started as early as 1853. The human genome project [NLM Website#hgp] is an international, collaborative research program, the goal of which was the complete comprehension of all the genes of human beings. The human genome project started in the year 1990 and completed in 2003 and now the complete map of the human genome is available. When the HGP was completed, it was found that the number of human genes were only around 20,500 as against the earlier estimates of around 60,000. The tools created through HGP help in characterizing the genomes of other organisms like mice, fruit fly etc. used extensively in biological research.

Subsequent to the HGP, the National Human Genome Research Institute (NHGRI) a division of the National Institutes of Health, USA launched a public consortium, The ENCODE Project, with an aim to identify all the functional elements in the human genome sequence. The ENCODE project unveiled more secrets about the human genome [NHGRI Website]. It was found that there were more to the non-coding genome than what was thought about it [Harrow 2012], [Kapustha 2014]. The Gencode version7 release contains 20,687 protein-coding genes, 9640 noncoding RNA loci and 33,977 coding transcripts that were not represented in popular genome databases [Harrow 2012]. Subsequent to this, much work was carried out in the analysis of the noncoding portion of the genome and this had lead us to a point where the noncoding genome is called the “iceberg” [Kapustha 2014]. The current statistics available as per the latest release Gencode v27 is represented in the pie-chart in Figure 3.16.

2.3. Genomics Signal Processing

Following the discovery of the double helical structure of the DNA molecule by Watson and Crick, there has been enormous progress in the area of genomics. The complete set of DNA of an organism is called its genome. Genomics is a branch of molecular biology which is concerned with the study of structure, function, and evolution of genomes [NIH Website]. We are now in the genomic era where much research in the disease-drug area is carried out at the molecular level [Esau 2007], [Chen 2017], [Esteller 2011].

The genome as we know it, can be represented using an alphabet comprising of four letters, A, T, C and G. Making use of this representation, genomes have been analysed using computational methods as well as digital signal processing methods. The concepts of Digital Signal Processing are finding increased applications in molecular biology. With enormous amount of genomic data available in the public domain, it is mandatory that new methods for its use be put forth so that the information is useful to mankind. Genomics Signal Processing is the analysis, processing, and use of genomic signals for gaining knowledge and translation of that knowledge into systems-based applications [Anastassiou 2001(2)], [Dougherty 2005], [Vaidyanathan 2004]. It has evolved into a discipline of Engineering that studies the processing of genomic signals employing the concepts of digital signal processing.

2.4. Sequences analysed in early days of genome studies

In the early days of genome studies, much of the research done was DNA centric, the reason being the Central Dogma of Molecular Biology [Watson 2007], [Alberts 2007]. This controlled our understanding of the genome in the early days of genome studies. The central dogma states that the flow of genetic information in an organism happens as: DNA – RNA – protein. RNA was thought to be an intermediate element between DNA and protein and the relevance of studying RNA was inadvertently overlooked. This dogma also excludes the involvement of any other type of molecule in any other role in the process. Closely following this trend in molecular biology, genomics studies which made use of computational and DSP methods also centred around the coding region of the genome and ignored the noncoding region. The studies undertaken using DSP techniques mainly dealt with detection of exons within the coding DNA sequences, analysis of the process of coding or conversion of DNA into proteins, detection of genes etc. [Tiwari 1997], [Nair 2006], [Yoon 2007].

A digital signal filtering approach to the process of translation of nucleic acids into proteins is described by Anastassiou [Anastassiou 2001(1)]. Prediction of exons within gene regions of DNA is described by Vaidyanathan and Yoon [Vaidyanathan 2002]. In their paper, the authors also extend the use of digital filters in detecting the genes within DNA sequences. The role of digital filters in identification of exons is clearly manifested by Vaidyanathan and Yoon [Vaidyanathan 2004]. An anti-notch filter which is a single peaking filter is used to pick out exons from the gene regions of DNA sequences. The concept of long-range correlation between the base-pairs in DNA sequences is also analyzed in the above mentioned work.

Biomolecular sequences have been extensively analysed in the frequency domain [*Anastassiou* 2000], [Fox 2004]. Rao and Swamy present a method of identification of active sites or hotspots in protein sequences making use of a continuous wavelet transform method which uses the modified Morlet wavelet [Rao 2008]. These are but a few examples which show that much of the work in Genomics Signal Processing in the early decades of this century were centred around the coding DNA. RNA and the noncoding regions of the genome were ignored at large, because of the prevailing understanding of the genome was based on the central dogma.

2.5. The importance of noncoding RNA studies. Small ncRNA, long ncRNA

Only the three entities mentioned in the central dogma are explicitly involved in the formation of the protein i.e. during transcription and translation [Lodish 2000], [Nature Website], [Watson 2007], [Alberts 2007]. Almost all genome studies were gene and protein centric, and the coding region of the genome alone was considered relevant. RNA was seen as a passive intermediary that bridges the gap between DNA and protein [Watson 2007]. The only RNA molecules that were known to be “functional” and did not directly take part in protein formation were the transfer RNA (tRNA) and the ribosomal RNA (rRNA). These are often termed the “classical functional RNA” molecules [Washietl 2005].

But with systematic screening of the genome of various organisms, it was discovered that there is more to the genome than just the coding region of the DNA [Erdmann 2001], [Eddy 2002]. Though much was known about these functional RNAs (tRNA and rRNA) which are highly evolutionarily conserved and occur in almost all forms of life [Dinger 2008], [Morris 2015], little was known about other functional RNAs. The human genome project which concluded in 2003 followed by the studies of the GENCODE project consortium [Harrow 2012] threw light on various facets of the human genome which was not known till then.

In the last decade of the 20th century systematic screening of various genomes identified myriads of noncoding RNAs. Many functional RNA molecule do not directly take part in protein coding, but have other regulatory functions [Yoon 2007], [Eddy 2001]. The regions of the genome that were thought to be “junk” were found to hold the keys to the functions that are vital to life including alternative splicing, control of epigenetic variations and so forth [Yoon 2007], [Dinger 2008], [Morris 2015]. It was thought that the human genome contains around 60,000 genes. But the predicted number of protein-coding genes has come down and it is now clear that some of them were wrongly annotated earlier and they in fact represent non-protein-coding transcripts [Ponting 2008].

Some of the RNAs in this group of noncoding RNAs are termed small noncoding RNAs [Eddy 2001], [Boon 2016], and the others long noncoding RNAs [Brosnan 2009], [Kapustha 2014] based on the number of nucleotides in these molecules. Some authors opine that this classification based entirely on the number of nucleotides in these molecules is very inaccurate and gross [Ponting 2009]. However the alternative option of differentiating RNA molecules based on their protein-coding capabilities is deemed equally difficult as it does not consider the other functions of these molecules [Dinger 2008], [Ponting 2009].

2.6. Analysis of small noncoding RNA

Two types of noncoding RNA sequences are studied in this work; the ones which are called long noncoding RNA (lncRNA) and those that are named small noncoding RNA (small ncRNA). As already mentioned, small noncoding RNA is the generic term given to noncoding RNA sequences that are lesser than 200 nucleotides in length [Eddy 2001], [Carninci 2009].

The past two decades have witnessed steep rise in the study of the non-coding RNA. Systematic screening of various genomes has brought to light a completely new knowledge database of the noncoding RNA [Eddy 2001], [Gisela 2002], [Mattick 2006]. One of the most important recent advancements in molecular biology has perhaps been the discovery that noncoding region of the genome can regulate transcription, translation and gene expression. It was discovered that functions of ncRNA include translocation, RNA processing and modification, chromosome replication, to name a few [Garst 2011], [Cech 2014]. These factors emphasize the need to study small noncoding RNAs.

Many well proven computational methods have been developed over this decade and in the previous one, for the analysis of noncoding RNA. Tran et. al. present an algorithm for the prediction of novel noncoding RNA genes making use of features derived from the sequences and structures of known noncoding RNA genes in comparison to decoys [Tran 2009]. These features were made use of to train a neural network-based classifier which the authors claim gave an average prediction sensitivity and specificity of 68% and 70% respectively in Escherichia coli (E coli). Yoon and Vaidyanathan make use of an improvised version of hidden Markov model (HMM) namely, the context sensitive HMM for the prediction of secondary structure of noncoding RNA [Yoon 2004].

An efficient method for detecting noncoding RNAs is described by Washietl [Washietl 2005]. The author describes his approach as one which combines comparative sequence analysis and structure prediction approaches and is suitable for a large genomic screen. Gardner and Giegerich [Gardner

2004] present a comparison of comparative methods of RNA structure prediction. A review of the various popular computational approaches that analyze noncoding RNA is presented by Washietl et.al. [Washietl 2012]. The above mentioned works are but a few examples of the computational techniques used in the analysis of noncoding RNA. However, DSP based methods that analyze the noncoding genome were not found in literature.

2.6.1. Relevance of secondary structure and MFE

Structure of bio-molecules governs their function [Tinoco 1999], [Pederson 2000] [Washietl 2012], and many functional RNAs have well conserved structures across species [Eddy 2014]. RNA is a single stranded molecule, which folds onto itself due to nucleotide pairing via hydrogen bonds between the bases. RNA involves in complementary base-pairing via hydrogen bonds (A-U, C-G, Watson-Crick/canonical base-pairing) in the same strand [Eddy 2001], [Gisela 2002]. The folded structure thus obtained is the secondary structure of the RNA molecule. RNA secondary structure is seen to influence every step in gene expression [Wan 2011].

RNA molecules could further fold into 3D tertiary structure which decides many of its functions [Tinoco 1999]. But the secondary structure is formed prior to and independent of the tertiary 3D structure [Washeitl 2005], [Washietl 2012]. Formation of tertiary structure does not alter the secondary structure. Also, the secondary structure is made up of sub-structural elements, which are responsible for most of the overall folding energy and can be seen as a coarse-grained approximation of the tertiary structure. Thus the secondary structure obviously is the first step in understanding the far more complicated three-dimensional tertiary structure and thereby the function of the ncRNA sequence. The secondary structure that is “optimum” is the minimum free energy (MFE) structure and MFE is the factor which decides this optimal structure [Pederson 2000], [Washietl 2012].

There are many computational approaches to predict the secondary structure of RNA sequences. A few examples are discussed here briefly. An improvised version of hidden Markov model (HMM), namely the context sensitive HMM (csHMM) has been used for prediction of secondary structure of noncoding RNA sequences [Yoon 2004]. This approach predicts secondary structures, taking into account the formation of pseudo-knots also. Secondary structure prediction based on a Boltzmann-weighted ensemble is presented by Ding et. al. [Ding 2005]. A centroid structure is thought to be the representative of a set of structures and a method is developed for the identification of this centroid structure. The authors claim this method make lesser errors when compared to energy based structure prediction algorithms.

Energy based algorithms are another popular approach for secondary structure prediction in RNAs. The most popular among them is the minimum free energy (MFE) based secondary structure prediction [Mathews 2010] because of the fact that in the natural environment of a biomolecule, the minimization of free energy is the most decisive factor of structure formation [Pedersen 2000]. Hajiaghayi et. al. present an analysis of energy-based algorithms for the prediction of RNA secondary structure [Hajiaghayi 2012]. The authors conclude the study with one of their findings being that MFE based structure prediction algorithms represent a reliable estimate within 2% accuracy with high confidence.

2.6.2. Novel model for MFE

The parameters, sequence length and MFE have been used in analyzing RNA from a very early time [Grüner 1996], [Galzitskaya 1998]. There have been studies which explore the influence of length and MFE on sequence stability [Pervouchine 2003], [Trotta 2014]. MFE has also been used as an index to study the relationship between entropy and structural properties of RNA sequences [Wolfsheimer 2010]. Washeitl describes a noncoding RNA gene finder which makes use of MFE *z* score computations, together with comparative genomic techniques. The mean and standard deviation of MFE of sequences are made use of here [Washietl 2005]. Clote et.al. describes a method of ‘asymptotic z score’ that sets asymptotic limits for mean and standard deviations of MFE per nucleotide of random RNA. They perform certain pre-computations that speed up z score computations for the entire genome using a sliding window scan. This method provides a filter, which can be used together with MFE computations and pattern matching to identify functional RNA genes in expressed sequence tags and genomic data. RNAs for which native state (the free energy structure) is functionally important were found to have lower folding energy, when compared to random RNAs having the same length and dinucleotide frequency [Clote 2005]. As MFE is a discerning factor, knowing its value would be useful in situations where it is needed to know quickly whether a given sequence is a functional or a random RNA sequence.

MFE is a vital tool in identifying noncoding RNA genes. Lim et.al describes a technique for identifying miRNA genes where a moving window scan searches for stem-loop structures having at least 25 base-pairs and has a predicted MFE of -25 kcal/mol or less. A window which accommodates 21 nucleotides is passed over each conserved stem-loop structure and a log-likelihood score is assigned to each window to determine how well its attributes resemble those of experimentally verified miRNA [Lim 2003]. Warris et.al. describe yet another method of prediction of small regulatory RNAs in genomes using MFE distribution of sequences as the discerning factor. The underlying principle is that the secondary structures of small regulatory RNAs have lower free energies than random RNA or other ncRNA sequences of the same length and dinucleotide composition [Warris 2014].

As is evident from the above, both MFE and sequence length are important parameters to be analyzed in the study of ncRNA. Computational methods have been widely employed to study noncoding RNA. Even though DSP methods have become as popular as computational methods in the analysis of genomic data right from the turn of this century [Anastassiou 2001(1)], [Cristea 2002], [Vaidyanathan 2004], [Yon 2007], [George 2010], little work has been done which makes use of Digital Signal Processing techniques to analyze the noncoding genome. Sequence length and MFE have been used extensively in analysing RNA, but a mathematical relationship linking MFE to the length or any other signal property of the sequence has not been reported in literature till date. Here in this work we have introduced a novel approach which links MFE, a thermodynamic property of ncRNA sequences to their signal properties. Making use of this relationship, a novel mathematical model has been arrived at for finding MFE from the signal properties of the sequence, without using any folding algorithm [George 2016].

2.7. Analysis of long noncoding RNA

Long noncoding RNA refers to those ncRNA molecules that are more than 200 nucleotides in length [Mercer 2009], [Ponting 2009]. Defining lncRNAs by what they are not is deemed rather inapt [Ponting 2009] but the current level of knowledge we have about these sequences makes this classification convenient. These molecules could be categorized based on their empirical features like genomic context, origin of transcription, tissue specificity, molecular function or mechanism of action. Long ncRNAs transcribed from intergenic regions are called long intervening ncRNAs and those transcribed from within introns are called intronic lncRNAs [Kung 2013], [Ma 2013]. Introns are regions within the gene that is not used in protein coding (explained in detail in Chapter 3). Nevertheless, their classification is not standardized and we find that very often human genes possess both coding and noncoding transcripts which are difficult to distinguish without detailed experimental studies [Ponting 2009].

Though thought to be "dark matter', "transcriptional noise" etc. initially, long noncoding RNAs are now recognized as crucial elements in biological regulation. There are diverse classes of lncRNA which control numerous processes across almost every realm of life [Ponting 2010], [Kung 2013], [Quinn 2016]. Long noncoding RNA sequences have been much studied soon after they were discovered. They are implicated in many diseases [Wapinski 2011], [Harries 2012], [Chen 2016], [Fang 2016] and various stages of development in organism [Smola 2016], [Perry 2016], [Brazao 2016], [Mercer 2009], [Calabrese 2013], [Chen 2014]. The possibility of using them in drug development too have been discovered [Li 2015], [Ling 2015], [Matsui 2017], [Boon 2016].

There are quite a few number of computational methods in literature which analyse long noncoding RNA sequences. Signal et.al. [Signal 2016] describe a computational method for functional prediction and characterisation of long noncoding RNA. Core features of functional lncRNAs are probed via an array of computational methods. Long noncoding RNA function is also predicted by using tissue specific evolutionary conserved expression as done by Perron et. al. [Perron 2017]. These authors make use of the 'guilt-by-association' principle which is explained as follows. If an lncRNA gene shows an expression profile that correlates with the expression profiles of a set of coding genes involved in a known function, then the lncRNA gene analysed probably is involved in the same function. Zhao et.al. describe prediction of lncRNA function using a co-expression network which is found to be useful in large-scale annotation of long ncRNA. The nodes in the network correspond to protein-coding gene or lncRNA and the edges connecting the nodes denote whether they are co-expressed [Zhao 2014]. Functions of lncRNAs across multiple cancers are explored through co-expression networks by Li et.al. Weighted correlation network analysis is made use of to express the functions of lncRNAs altered in more than two cancer types. The authors conclude that the lncRNAs expressed in cancers show high tissue-specificity and are weakly expressed than protein-coding genes [Li 2017].

Though there are many computational methods to analyse the long ncRNA, DSP based methods which analyse lncRNA were not found in literature. Here, we analyse lncRNA sequences (taken from NCBI GenBank) in order to locate exons present in them. Some lncRNA transcripts have been found to contain exons within them [Ponting 2009], [Niazi, 2012], and some noncoding RNAs have been found to encode peptides [Dinger 2008]. We detect exons in long noncoding RNA making use of digital filters. This method of detecting exons has been successfully carried out on DNA sequences by the authors [George 2010]. The property used here in this work [George 2017] in locating exons in long noncoding RNA sequences is the period 3 property which is an established feature in locating exons in coding regions [Tiwari 1997], [Trifonov 1980], [Li 1997], [Vaidyanathan 2002], [Anastassiou 2002].

Chapter 3

Molecular Biology background of the study

This Chapter contains a brief overview of molecular biology as we see it today which is the background for this work. The primary concepts of DNA, genes and RNA are explained. The importance of RNA in molecular biology is detailed upon as this research focuses on the analysis of noncoding RNA sequences. Much of the studies in microbiology and genomics were concentrated in the coding region or the DNA during the initial days. The relevance of studying the noncoding portion of the genome was understood only after the screening of various genomes identified myriads of noncoding RNAs (ncRNAs). Non coding RNAs are RNA molecules that do not participate directly in the formation of proteins but have regulatory functions.

Abstract

The genetic code for every organism is stored in bio-molecules called nucleic acids. There two types of nucleic acids: the deoxyribonucleic acid (DNA) and the ribonucleic acid (RNA). In higher organisms DNA is found inside the nucleus and RNA outside the nucleus. DNA has two regions – the gene and the non-gene. Gene is the portion of the DNA that is directly responsible for coding of proteins which are needed by the organism. RNA has many support functions. RNA molecules like messenger RNA (mRNA) take part directly in the formation of protein by acting as the template for DNA. Transfer RNA (tRNA) is a type of RNA molecule that helps to decode a messenger RNA (mRNA) sequence into a protein. tRNA and rRNA are the noncoding RNA sequences that were well studied since the early days of genomics and molecular biology. There are a whole other set of noncoding RNA sequences which are encoded by the noncoding region of the genome. They play vital roles in the formation of protein and expression /suppression of genes. Some examples are snRNA, snoRNA, miRNA, siRNA etc., each of which have specific functions in various stages of protein formation and gene expression. There are yet another set of noncoding RNAs which have been called long ncRNA (lncRNA). Although two thirds of the human genome gets transcribed, only 2% of the transcribed genome encodes proteins. It has been found that the remaining gets converted into long ncRNA molecules and other ncRNAs. Long ncRNAs were thought of as “transcriptional noise” even in the genomic era. But they have assumed prime importance in molecular biology in the current decade after their varied functions have been unveiled. Epigenetic regulation, chromatin modelling, gene transcription, protein transport, protein trafficking, cell differentiation, organ or tissue development, cellular transport, metabolic processes and chromosome dynamics are just a few examples of long ncRNA functions.

3.1. Introduction

This chapter introduces the basic concepts of molecular biology which are relevant to this study. It could be said that Molecular Biology as we know it today had its inception back in 1856 when Gregor Mendel conducted his famous experiments with the pea and concluded that ‘certain factors’ called ‘genes’ are passed on from the parent to the offspring [Watson 2007]. Nearly half a century later it became clear due to the work of Walter Sutton (medical student, Columbia University) and T. H. Morgan (also at Columbia), that these “factors” were located within chromosomes which were known to contain proteins and DNA molecules. In 1930 the DNA was shown to be a long molecule made of the nitrogenous bases A, T, C and *G*. In those days proteins were considered to be the “genes” that carried hereditary information. In 1944, the experiments of O. T. Avery (Rockfeller Inst., NY) showed that DNA, rather than protein, carried genetic traits. Alfred Hershey and Martha Chase verified this experimentally (1952, Cold Spring Harbor).

It was accepted that genes were contained in the DNA but nothing was known about their nature or how they worked. The helical structure of DNA as we know it today, was revealed by the work of Watson & Crick in 1953. Watson, a young scientist from Chicago, who worked at the Cavendish Laboratories, Cambridge, England, together with Maurice Wilkins of London, studied the X-ray diffraction pattern of DNA. They soon realized that finding the structure of DNA would be the only way to understand genes. Watson later worked with Francis Crick at the Cavendish Laboratories, and their studies lead them to conclude that DNA had a helical structure [Pray 2008].

After the historical announcement of the double helix structure of the DNA molecule in 1953 by Watson and Crick (for which they were awarded the Nobel Prize), there has been phenomenal progress in genomics in the last five and a half decades. It is known that all characteristics of living beings are determined by the gene sequences. At present, there are quite a few number of public databases which provide genomic and proteomic data which can be put to use so that it benefits humanity.

Genomic/proteomic information available in public data resources are in the form of character strings rather than numerical sequences. However, if we properly map a character string into a numerical sequence, then it can be processed with digital signal processing techniques. Digital signal processing has the potential to provide a set of novel and useful tools for solving highly relevant problems [Anastassiou 2001(1)], [Anastassiou 2001(2)], [Vaidyanathan 2004]. For example, colour spectrograms provide significant visual information about bio-molecular sequences which facilitates understanding of local nature, structure, and function. Also, the magnitude and the phase of properly defined Fourier transforms [Anastassiou 2001(1)], [Vaidyanathan 2004] can be used to predict important features like the location and properties of protein coding regions in DNA, which are indicative of their functions.

Genomic information is digital in a very real sense. It is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences namely, DNA and proteins, have been mathematically represented by numerical sequences, in which each character is a mapped to a numeric [Nair 2006], [Cristea 2002(2)], [Anastassiou 2001(1)].

3.2. DNA, Genes, Formation of protein

Nucleic acids which are found in living organisms are polymers specialized for storage, transmission and use of information [Watson 2007], [Alberts 2007]. There are two types of such nucleic acids: DNA (de-oxyribose nucleic acid) and RNA (ribose nucleic acid). Single-celled organisms like bacteria do not have a nucleus and the DNA just resides in the cell. Such cells are called prokaryotes. Higher organisms (worms, insects, plants, mammals etc.) have cells with nucleus and are called eukaryotes. In the case of eukaryotic cells, DNA resides within the nucleus of the cell and RNA outside the nucleus. In eukaryotic cells, DNA is found in combination with proteins within the chromosome inside the nucleus. An exception is the red blood cell which has no nucleus. Cells also have a small quantity of DNA in the mitochondria [NLM Website]. It is not relevant to this work and we shall not discuss this here.

3.2.1. DNA

Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms including some viruses. The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints or a recipe, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules [Watson 2007], [Alberts 2007].

A schematic diagram for the DNA molecule is shown Figures 3.1. The DNA molecule has the structure of a double helix as shown.

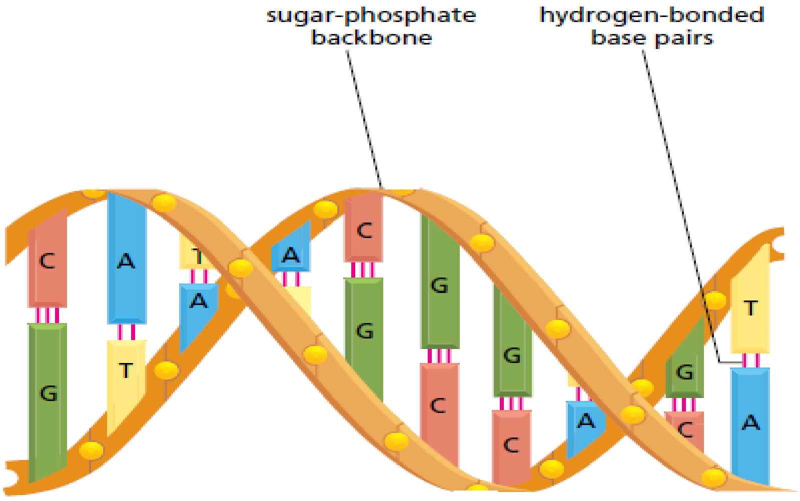


Figure 3.1. The double-helical structure of DNA

Between the two strands of the backbone which is outside, there are pairs of bases like the rungs of a ladder. The backbone is a very regular structure made from sugar-phosphate. There are four types of bases (or nucleotides), denoted with the letters *A*, *C*, *G,* and *T* (respectively, adenine, cytosine, guanine, and thymine). A and G are called the purines, while T and C are called the pyrimidines.

In Figure 3.2 (A) the double helix is shown straightened out for simplicity. The genome sequence corresponding to the top strand of the DNA molecule in this example is *ACTGGCAATG*. Note that the ordering is from the so-called 5' to the 3'end (left to right). DNA sequences are typically listed from the 5' to the 3' end because they are scanned in that direction when bases are used by the cell machinery to signal the production of amino acids. The reason for directed flow arises from the way the sugar and phosphate are glued together. The sugar-phosphate back-bone together with the nitrogenous base is called the nucleotide, shown in Figure 3.2(B).

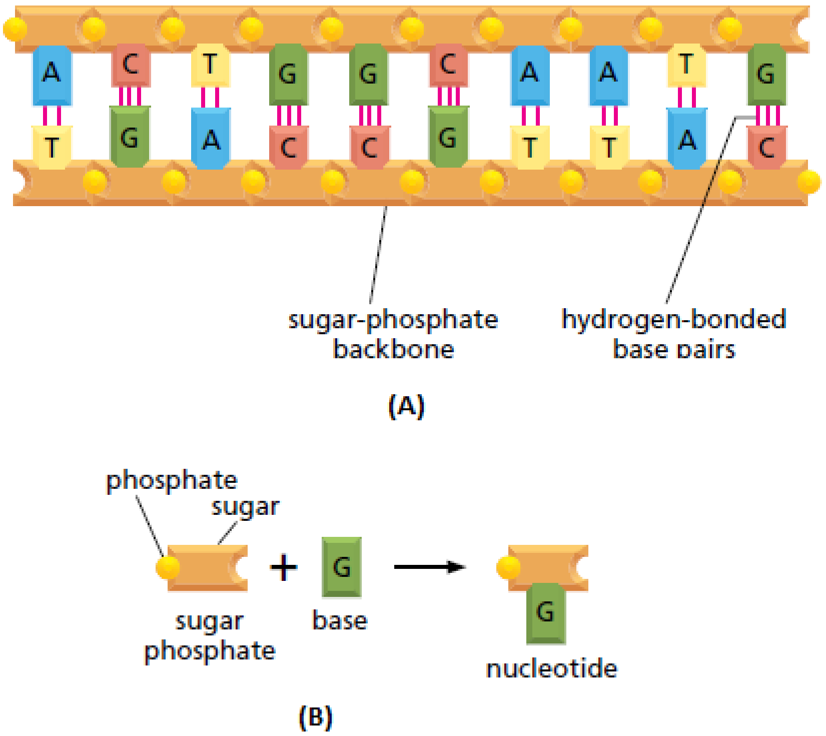


Figure 3.2. (A) Linearised schematic of the DNA double helix (B) Building block of DNA, the sugar phosphate backbone, the nitrogenous base and the nucleotide.

In the double stranded DNA, the base *A* always pairs with *T*, and *C* pairs with *G*. Thus the bottom strand *TGACCGTTAC* is the complement of the top strand. This is called the Watson-Crick base-pairing or canonical base-pairing; it occurs through a weak bond called the hydrogen bond [Watson 2007], [Alberts 2007]. Nevertheless as there are several million base pairs, the two strands are held together strongly. Typically in any given region of the DNA molecule, only one of the two strands is active in gene expression [Watson 2007], [Alberts 2007].

The internal atomic details of the molecules *A*, *T*, *C*, and *G* are shown in Figure 3.3. These molecules are made from carbon, nitrogen, hydrogen and oxygen atoms. There are about three billion of these bases in the DNA of a single human cell.

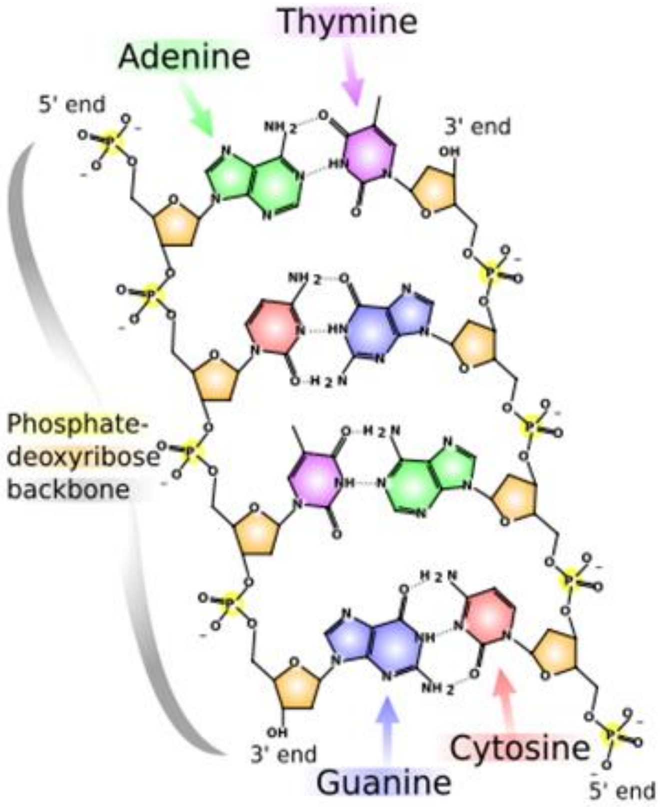


Figure 3.3.The chemical structure of DNA. Hydrogen bonds are shown as dotted lines. Bold lines indicate covalent bonds

3.2.2. Genes and formation of protein

A DNA sequence has two regions as shown in Figure 3.4: genes (marked blue) and intergenic spaces (marked yellow). Genes contain the information for generation of proteins. Each gene is responsible for the production of a different protein. Gene has two sub-regions called the exons (marked red in Figure 3.4) and introns (marked green in Figure 3.4). Exons are the regions which are directly take part in the formation of protein. Introns do not take part directly in the coding of proteins [Watson 2007], [Alberts 2007]. Procaryotes like bacteria do not have introns [Alberts 1998].

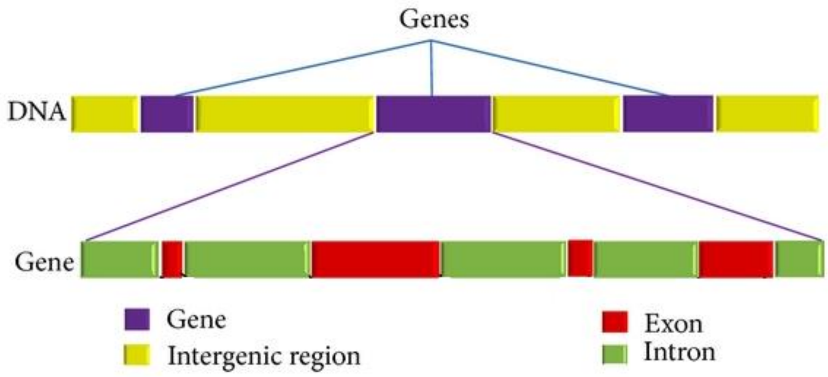


Figure 3.4. DNA sequence, introns and exons

Even though all the cells in an organism have identical genes, only a selected subset is active in any particular family of cells. For example the set of genes

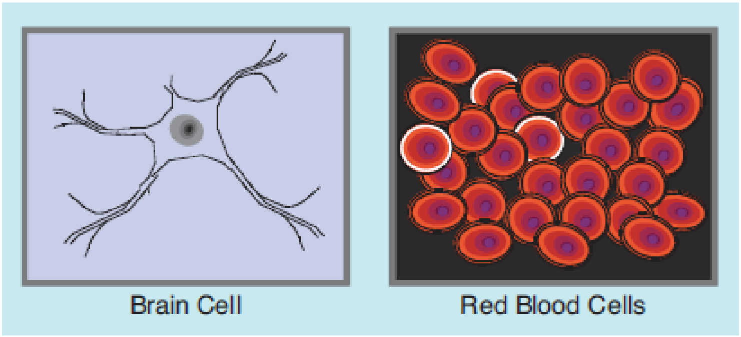


Figure 3.5. Representation of a) Brain cells and b) blood cells

that are active in blood cells are different from those that are active in nerve cells, which explains why these cells look so different. An illustration is given in Figure 3.5.

The central dogma of molecular biology states that genetic information flows from DNA – RNA – protein. The central dogma is represented in Figure 3.6.

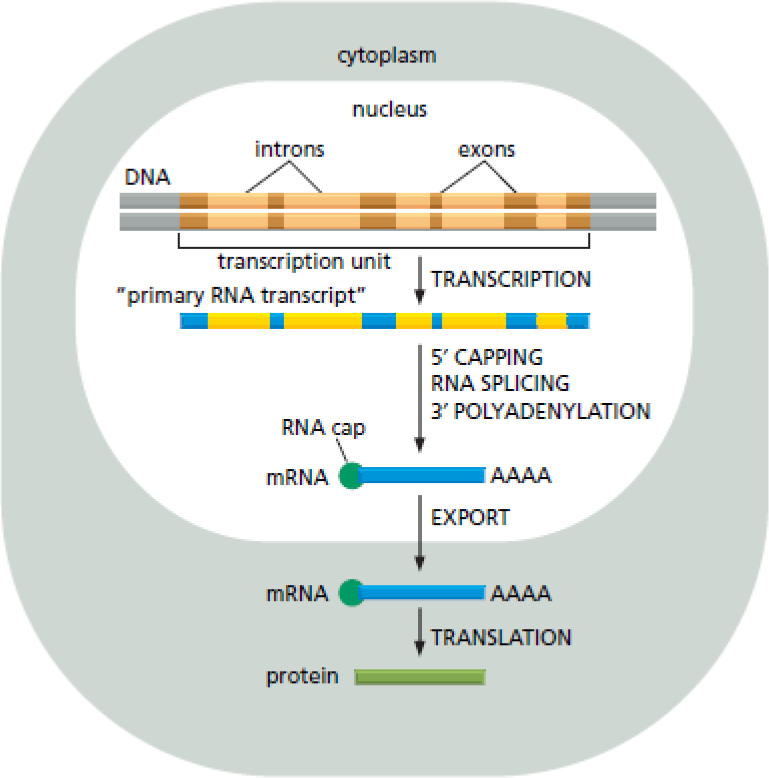


Figure 3.6. The central dogma of molecular biology. Formation of protein from DNA.

Figure 3.7 shows a simplified representation of the key steps involved in the production of protein from a gene. The gene is first copied into a single stranded chain called the messenger RNA or the mRNA molecule. This process is called transcription. The introns are then removed from the mRNA by a process called splicing. The spliced mRNA is then used by a large molecule called the ribosome to produce the appropriate protein. The translation from mRNA to protein is aided by adaptor molecules called the transfer RNA or tRNA. It could be said that the tRNA molecules also store the genetic code [Alberts 1998] as we shall see in the next section.

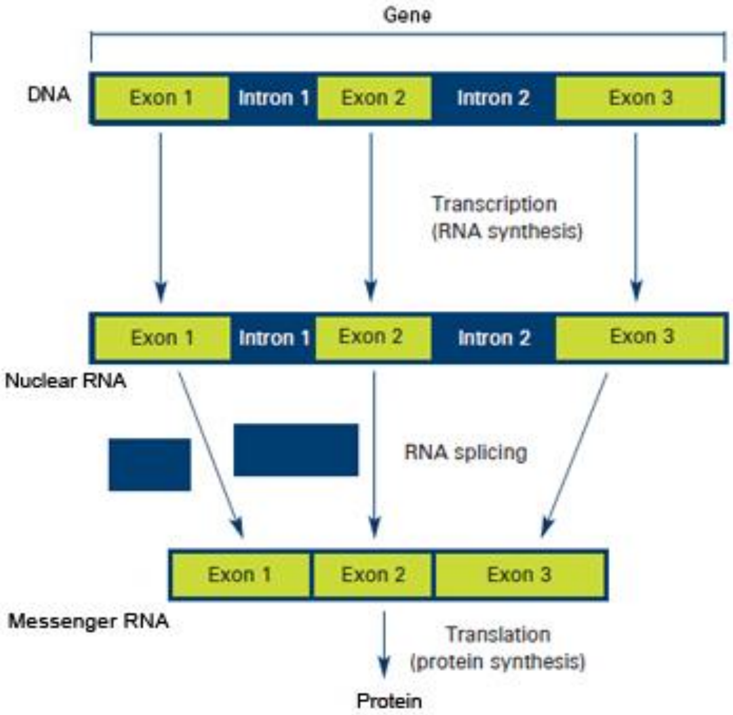


Figure 3.7. DNA to protein

3. 3. RNA

The RNA (ribonucleic acid) molecule is closely related to the DNA. It is also made of four bases but instead of thymine, a molecule called uracil (denoted as *U*) is found in RNA. Figure 3.8 shows a comparative representation of DNA and RNA. Figure 3.9 shows the primary chemical structure of an RNA sequence and Figure 3.10 shows the chemical structure of the sugar and phosphate backbone and also that of the nitrogenous bases.

Unlike DNA, RNA is single-stranded. The single stranded RNA molecule folds onto itself to form what is called the secondary structure of the RNA. While doing so, hydrogen bonds are formed between the bases. Base-pairing occurs as explained in the case of DNA. *U* pairs with *A* by hydrogen bonding just like *T* pairs with *A* as in DNA. The sugar in the sugar-phosphate backbone is also slightly different from the DNA molecule. DNA contains the sugar, deoxyribose, while RNA contains the sugar ribose. The only difference between ribose and deoxyribose is that ribose has one more -OH group than deoxyribose, which has -H attached to the second (2') carbon in the ring as shown in Figure 3.9. DNA is stable under alkaline conditions while RNA is not stable [Watson 2007], [Alberts 2007].

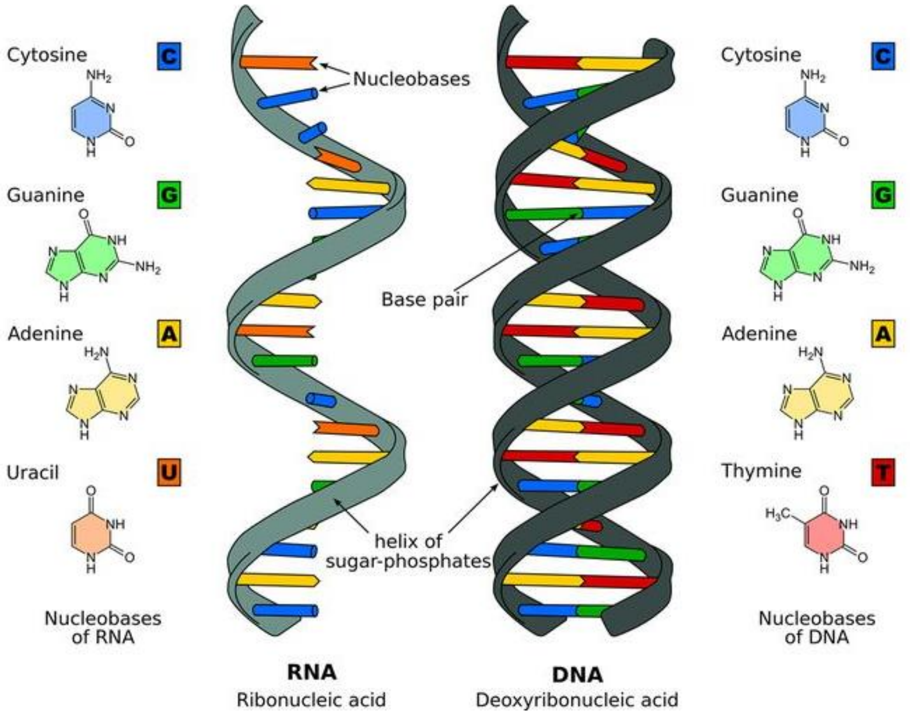


Figure 3.8. Comparison of DNA and RNA

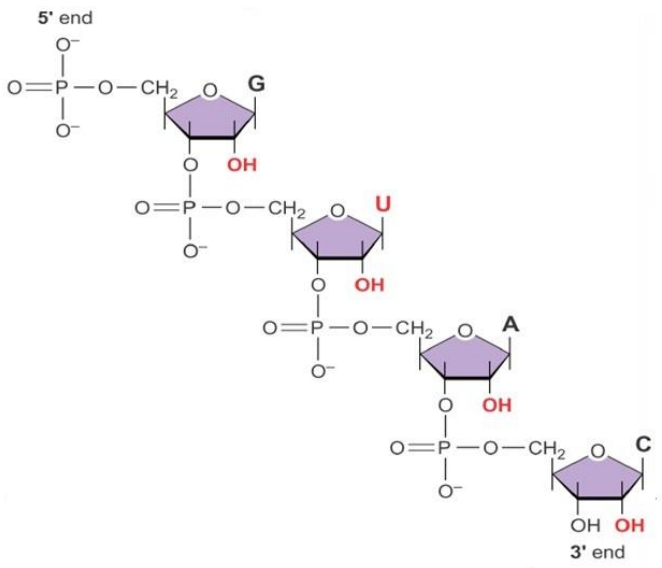
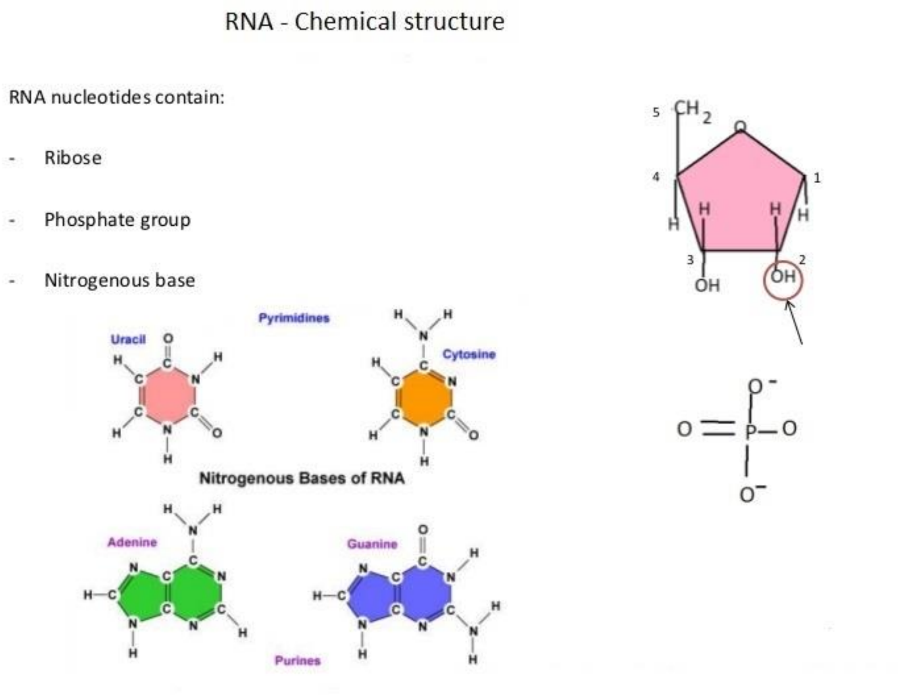


Figure 3.9. Primary, chemical structure of RNA



RNA - Chemical structure

RNA nucleotides contain:

- Ribose
- Phosphate group
- Nitrogenous base

Figure 3.10. Chemical structure of the sugar-phosphate backbone and the nitrogenous bases in RNA

3.4. mRNA, rRNA, tRNA and the formation of protein

The classical knowledge of the RNA molecules was that they are short, typically short-lived and are used by the cell as temporary copies of portions of DNA [Watson 2007], [Alberts 2007]. A typical example is the messenger RNA (mRNA). Messenger RNA is a large family of RNA molecules that convey genetic information from DNA to the ribosome, where they specify the amino acid sequence of the protein products of gene expression. Messenger RNA molecules have short life span beginning with transcription. A very simple description of transcription and translation which happens in eukaryotes is given below [Lodish 2000], [Alberts 2007], [Watson 2007], [Nature Website].

A copy of the gene from the DNA is written on to the mRNA by molecules called RNA polymerase which associates with mRNA-processing enzymes during transcription so that processing can start immediately after transcription. The short-lived, unprocessed/partially processed molecule is termed precursor mRNA or pre-mRNA and when completely processed it is termed mature mRNA. The pre-mRNA/pre-cursor mRNA has both introns and exons of the DNA template strand.

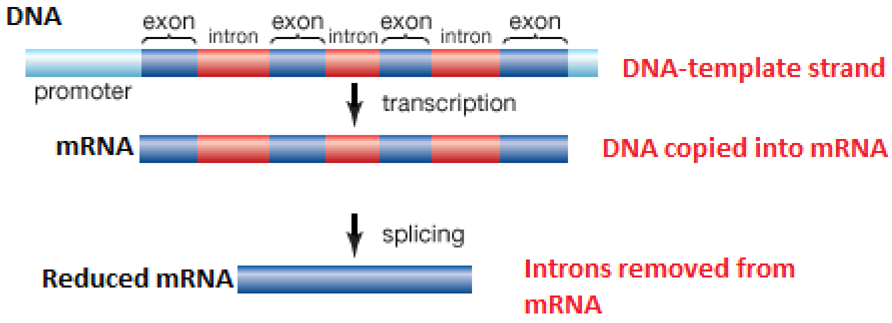


Figure 3.11. Transcription and splicing

After the DNA has been copied into the mRNA, the introns are removed by splicing and only the exons of the DNA strand are retained on the mRNA. This mRNA is termed ‘reduced mRNA’. A process called 5' capping occurs immediately after transcription commences. A modified guanine nucleotide is added to the front end or the 5' end of the eukaryotic messenger RNA shortly after the start of transcription. The 5' consists of a terminal 7 methylguanosine residue that is linked through a 5'-5' tri-phosphate bond to the first transcribed nucleotide. 5' cap ensures recognition by the ribosome and protection of the RNA molecule from ribonucleases (RNases). RNase is a type of nuclease which catalyses the degradation of RNA into smaller components.

After transcription, polyadenylation occurs. Polyadenylation involves linking of the polyadenylyl moiety to the messenger RNA molecule. The polydenylyl moiety is attached to the 3' end of the mRNA molecule. As nucleic acid molecules are read from the 5' to the 3' end, polyadenylation is also termed 'tailing'. Polyadenylation is important for the termination of transcription, export of the mRNA from the nucleus and its translation to protein. Once transcription is terminated, mRNA chain is cleaved through the action of an endonuclease complex associated with the RNA polymerase. A simple representation of transcription is given in Figure 3.11.

The next step is transportation of the mature mRNA from the nucleus to the cytoplasm. Transportation is controlled by different signaling pathways. Once in the cytoplasm, the mature mRNA is translated into protein by the ribosome. Translation is the process by which a protein is synthesized from the information contained in a molecule of messenger RNA (mRNA). During translation, an mRNA sequence is read using the genetic code, which is a set of rules that defines how an mRNA sequence is to be translated into the 20-letter code of amino acids, which are the building blocks of proteins.

Translation is represented in Figure 3.12. Translation takes place in specialized cellular structures called ribosomes. This means that ribosomes are the sites at which the genetic code is actually read by a cell. The ribosome is a complex molecule made of ribosomal RNA (rRNA) molecules and proteins that form a factory for protein synthesis in cells. The ribosome translates each codon, or set of three nucleotides, of the mRNA template and matches it with the appropriate amino acid. The amino acid is provided by the transfer RNA (tRNA) molecule. Transfer ribonucleic acid (tRNA) is a type of RNA molecule that helps decode a messenger RNA (mRNA) sequence into a protein. tRNAs function at specific sites in the ribosome during translation. Proteins are built from smaller units called amino acids, which are specified by three-nucleotide mRNA sequences called codons. Each codon represents a particular amino acid, and each codon is recognized by a specific tRNA. The tRNA molecule has a distinctive folded structure with three hairpin loops that form the shape of a three-leafed clover. One of these hairpin loops contains a sequence called the anticodon, which can recognize and decode an mRNA codon. Each tRNA has its corresponding amino acid attached to its end. When a tRNA recognizes and binds to its corresponding codon in the ribosome, it transfers the appropriate amino acid to the end of the growing amino acid chain. Then the tRNAs and ribosome continue to decode the mRNA molecule until the entire sequence is translated into a protein. Translation of an mRNA molecule by the ribosome occurs in three stages: initiation, elongation, and termination. During initiation, the small ribosomal subunit binds to the start of the mRNA sequence. Then a transfer RNA (tRNA) molecule carrying the amino acid methionine binds to what is called the start codon of the mRNA sequence. The start codon in all mRNA molecules has the sequence AUG and codes for methionine.

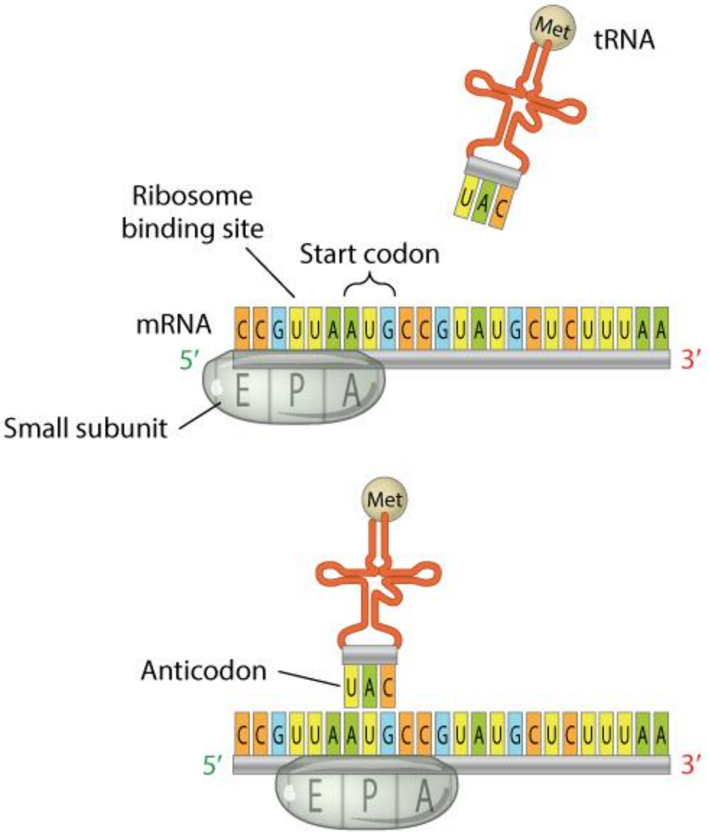


Figure 3.12. Translation – tRNA with the START anticodon binding onto the mRNA to initiate translation

Next, the large ribosomal subunit binds to form the complete initiation complex. During the elongation stage, the ribosome continues to translate each codon in turn. Each corresponding amino acid is added to the growing chain and linked via a bond called a peptide bond. Elongation continues until all of the codons are read. Lastly, termination occurs when the ribosome reaches a stop codon (UAA, UAG, and UGA). Since there are no tRNA molecules that can recognize these codons, the ribosome recognizes that translation is complete. The new protein is then released, and the translation complex comes apart.

3.5. Noncoding RNA

The central dogma of molecular biology has exerted a substantial influence on our understanding of the genetic activities in the cells. Based on the central dogma, the prevailing assumption in the past was that genes are basically repositories for protein coding information and that proteins are responsible for most of the important biological functions in all cells [Watson 2007], [Alberts 2007]. Thus RNA was seen as a passive intermediary that bridges the gap between DNA and protein. Examples of RNAs that do not directly participate in protein formation are tRNA and rRNA; they have other functions in the formation of protein. These two molecules could be called the “classical functional ncRNA” molecules [Washietl 2005 (Dissertation)]. These are the most ubiquitous noncoding RNA species in the genome and these “structural” RNAs are highly evolutionarily conserved, and occur in all known forms of life [Dinger 2008], [Morris 2015].

Little was known about other functional RNAs. Besides tRNA and rRNA, functional RNAs were considered to be very rare. This view underwent a drastic change in the last decade of the 20th century when screening of various genomes identified a wide variety of noncoding RNAs (ncRNAs) [Yoon 2007]. There are many functional RNA molecules that do not directly take part in protein coding, but have other regulatory functions [Eddy 2001]. These facts were not known and a majority of the genome was regarded as “junk” mainly because it was not well understood. It has come to light that these “junk” portions of the genome holds the keys to the functions that are vital to life including alternative splicing, control of epigenetic variations etc. [Yoon 2007].

Francis Crick proposed the existence of adaptor RNA molecules that were able to bind to the nucleotide code of mRNA, thereby facilitating the transfer of amino acids to growing polypeptide chains [Nature Website]. The work of Hoagland et al. (1958) confirmed that a specific fraction of cellular

RNA was covalently bound to amino acids. Later, the fact that rRNA was found to be a structural component of ribosomes suggested that, like tRNA, rRNA was also noncoding. In addition to rRNA and tRNA, a number of other noncoding RNAs exist in eukaryotic cells. These molecules assist in many essential functions, which are still being enumerated and defined. As a group, these RNAs are frequently referred to as small regulatory RNAs (sRNAs). These regulatory RNAs exert their effects through a combination of complementary base pairing, complexing with proteins, and their own enzymatic activities. RNAs can interact with other RNAs and DNAs in a sequence-specific manner and they are very relevant in tasks that require highly specific nucleotide recognition [Eddy 2001], [Eddy 2002]. Micro RNAs (miRNAs) that regulate gene expression, small interfering RNAs (siRNA) that take part in RNA interference (RNAi) pathways for gene silencing are just two examples [Bartel 2004], [Hannon 2004], [McManus 2002], [Novina 2004]. Micro RNA (miRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA) are other examples of this category of small ncRNA. Functions of ncRNA include transcription, control of translation, translocation, RNA processing and modification, chromosome replication, to name a few [Garst 2011], [Cech 2014].

The noncoding RNA sequences which have been called “long noncoding RNA” are also analysed in this work. These are also functional molecules like the small noncoding RNA [Brosnan 2009]. Long ncRNA sequences are transcribed from the non gene region of the DNA and they have many implications in cellular development and in diseases. Long ncRNA are discussed in section 3.6. Regulatory role of many classes of ncRNAs is broadly recognized; however, long intronic ncRNAs have received little attention. In the past few years, it has come to light that intronic regions are key sources of regulatory ncRNAs [Cech 2014], [Derrien 2012], [Kung 2013]. Most of the eukaryotic genome is transcribed, yielding a complex network of transcripts that includes tens of thousands of long noncoding RNAs with little or no protein-coding capacity. Initially these were thought to be transcriptional “noise” but it is now clear that a significant number of these long noncoding transcripts have cell type-specific expression, localization to subcellular compartments, and are associated with human diseases [Mercer 2009], [Kapusta 2014].

A large number of noncoding RNA molecules have been identified in organisms and the list is growing constantly. Many of the newly discovered ncRNAs could not be assigned a function. In the rare cases when the function is known, the underlying molecular mechanisms are often poorly understood. In this chapter, an overview of the current knowledge on ncRNAs, relevant to this study is presented. It is also to be noted that studies have shown that it is difficult to unequivocally classify RNAs as protein coding or noncoding [Dinger 2008]. Protein coding and noncoding transcripts may overlap, certain transcripts can function intrinsically at the RNA level and also code for proteins. Such facts lead us to conclude that the functionality of any transcript should not be discounted at the RNA level. The Figure 3.16 shows the split-up of the human genome as per values in Gencode version 27 [Gencode v27].

3.6. RNA analyzed in this work

Figure 3.13 gives a brief look at the various types of RNA molecules that are involved in transcription and translation. During transcription, DNA is used as a template to produce an RNA transcript as shown.

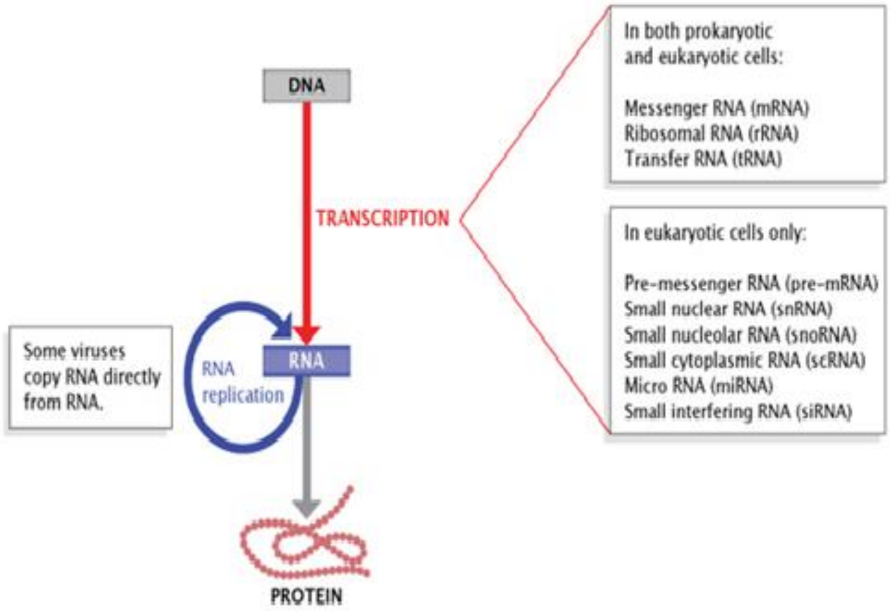

DNA
TRANSCRIPTION
In both prokaryotic and eukaryotic cells:
Messenger RNA (mRNA)
Ribosomal RNA (rRNA)
Transfer RNA (tRNA)
In eukaryotic cells only:
Pre-messenger RNA (pre-mRNA)
Small nuclear RNA (snRNA)
Small nucleolar RNA (snoRNA)
Small cytoplasmic RNA (scRNA)
Micro RNA (miRNA)
Small interfering RNA (siRNA)
Some viruses copy RNA directly from RNA.
RNA replication
RNA
PROTEIN

Figure 3.13. Several forms of RNA are involved in gene expression/suppression

RNA is translated to build the protein molecule or the polypeptide molecule encoded by the original gene. mRNA, rRNA and tRNA are present in both prokaryotic and eukaryotic molecules. Pre-messenger RNA (pre-mRNA), snRNA, snoRNA, small cytoplasmic RNA (scRNA), miRNA and siRNA are found exclusively in eukaryotic cells. Four classes of small ncRNA are studied here viz. micro RNA (miRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA) and snoRNA (small nucleolar RNA). We will see a brief overview of these four classes of molecules.

3.6.1. rRNA

Ribosomal RNA (rRNA) is a noncoding RNA molecule which could be called the ‘classical’ ncRNA [Kung 2013]. It was discovered during 1951 – 1965 right during the inception of Molecular Biology and hence is a well studied one. Though length of rRNA molecules are in the region 600 – 900 nucleotides, they still are studied along with small noncoding RNA [Edddy 2001]. The function and properties of rRNA have already been discussed in section 3.3 of this Chapter. The functions of rRNA are largely dependant on the tertiary 3D structure which in turn is derived from the secondary structure [Brimacombe 1985], [Garst 2011]. Though they are longer than 200 nucleotides, rRNA have been included along with the other ncRNA in this work for the analysis of secondary structure.

3.6.2. miRNA

Micro RNA (miRNA) is a small noncoding RNA molecule which has about twenty-two nucleotides which are found in animals, plants and certain viruses [Cia 2009], [Hannon 2004]. These molecules negatively regulate gene expression post-transcriptionally [Esau 2010]. Many genes in eukaryotic cells are silenced by not being transcribed into the mRNA. But in some other cases even transcribed genes are silenced by post-transcriptional mechanisms which prevents them from being translated. Micro RNA (miRNA) are one set of molecules that control translation. In simple words, miRNA are produced by cleavage of double-stranded RNA arising from small hairpins within RNA which is mostly single stranded. miRNAs combine with proteins to form a complex that binds rather imperfectly to mRNA molecules and inhibits translation. miRNAs function by base-pairing with complimentary sequences within mRNA molecules. Once this happens, the mRNA strand splits into two pieces or gets de-stabilized because its poladenylyl tail (at the 3’ end) gets shortened. In some other situations silencing of the mRNA occurs by its lesser efficient translation into proteins by the ribosomes [Bartel 2004], [Hannon 2004], [Cai 2009], [Wahid 2010]. Besides post-transcriptional control of gene expression, micro RNAs have been found to play crucial roles in cancer, metabolic diseases, viral infections and so on [Esau 2007], [Jansson 2012]. This means that miRNAs represent a class of molecules which have the potential to be used as drug targets for these diseases by therapeautic modulation of their activity. Different miRNAs have been found to be deregulated in kidney and bladder cancers [Gottardo 2007].

3.6.3. snRNA

These are a class of small RNA molecules which are found within the splicing speckles and cajal bodies of the cell nucleus in eukaryotic cells. snRNA is also referred to as U-RNA (U for Uridine rich). Uridine is a glycosylated pyrimidine-analog containing uracil attached to a ribose ring. The length of snRNA molecules is around 80 to 350 nucleotides [Padgett 2015] in higher eukaryotes. They are transcribed by either RNA polymerase II or RNA polymerase III, and studies have shown that their primary function is in the processing of pre- messenger RNA in the nucleus. They have also been shown to aid in the regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA), and maintaining the telomeres [Matera 2007].

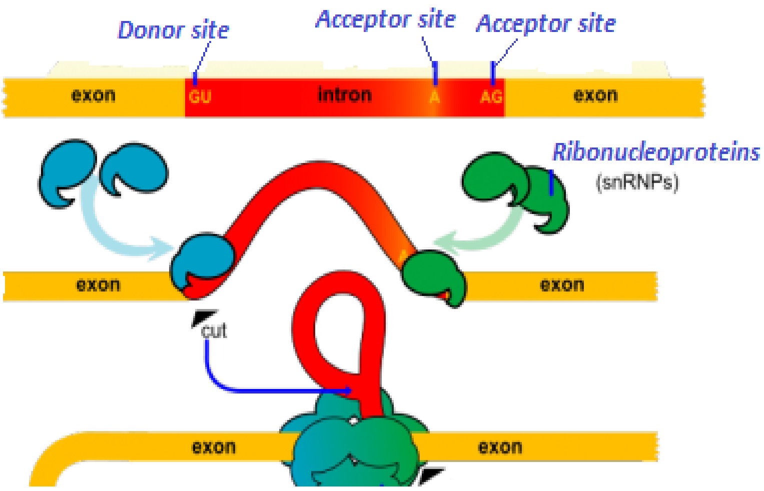


Figure 3.14. Simplified representation of snRNA initiated splicing

Biochemical, genetic and structural evidences suggest that snRNAs are the key components of the catalytic centre of the spliceosomes [Padgett 2015]. snRNAs form complexes with proteins to form snRNPs (small nuclear ribonucleoproteins) which are made use of in the unspliced primary RNA (the pre-mRNA) transcript to form spliceosomes. Thus small nuclear RNAs are essential parts of spliceosomes.

3.6.4. snoRNA

Small nucleolar RNAs (snoRNAs) are a class of small ncRNAs that carry out a fundamental role in modification and processing of ribosomal RNA. They guide site-specific rRNA modification [Deici 2009] and their function is similar in all types of organisms they are found in. These molecules have been found to have extensive similarities with other types of small non-coding RNA, in particular miRNA. snoRNAs can be roughly divided into two classes depending on their content of one of the two conserved structural elements known as the C/D and H/ACA boxes; the two classes are the box C/D and box H/ACA snoRNAs, that function differently in rRNA maturation. Generally, C/D box snoRNAs are ∼70–120 nucleotides (nt) and guide the methylation of target RNAs, while H/ACA box snoRNAs are ∼100–200 nt and guide pseudouridulation [Matera 2007]. Besides site-specific rRNA modification, snoRNAs also target spliceomal rRNA [Scott 2011].

As already mentioned, it is now been accepted that the noncoding genome is a region which contains many functional surprises. But the intronic area within genes too have their own surprises. snoRNAs have been found to be coded from intronic regions of eukaryotes [Brown 2008]. The sequences encoding H/ACA and C/D box snoRNAs are generally located in introns of their host gene, in the same orientation. One intron usually carries one snoRNA gene, but a host gene can carry several snoRNA genes in different introns. Intronic snoRNAs are produced by exonucleolytic degradation of the de-branched lariat after splicing. The stable (which goes on to form protein) part is protected by the binding of snoRNP core proteins, and/or of ancillary proteins, to the pre-mRNA [Brown 2008]. snoRNAs are further processed into smaller molecules similar to miRNA some of which display functionality. Small nucleolar RNAs (snoRNAs) guide RNA modification and are localized in nucleoli and Cajal bodies in eukaryotic cells. Components of the RNA silencing pathway associate with these structures, and studies reveal that a human and a protozoan snoRNA can be processed into miRNA-like RNAs [Taft 2009].

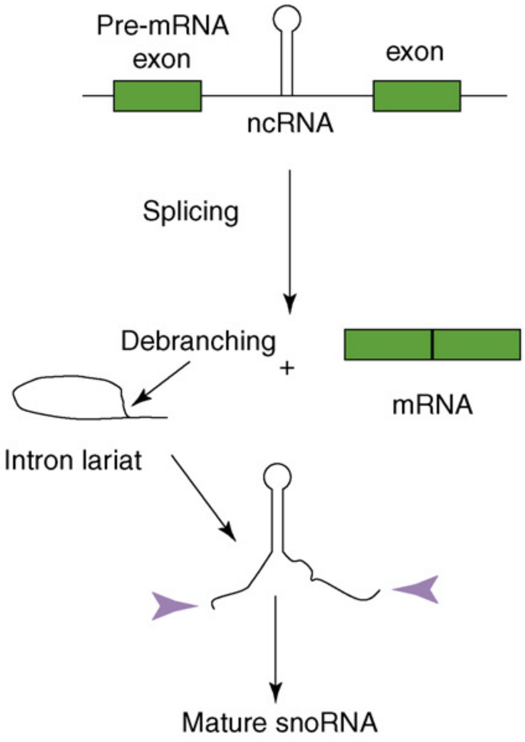


Figure 3.15. Simplified representation of snoRNA formed from intronic region of pre-mRNA.

3.7. Long noncoding RNA

The possibility of the genome having a noncoding component was not even thought of in the early days of genome studies. But with the completion of the human genome project in 2003, the number of protein coding genes has come down to around 20,000 from the estimated 60,000 in the mid 1990s. With screening of genomes of different organisms, it is becoming more clear that noncoding transcripts are vital to almost all stages of biogenesis of the cell.

Long non-coding RNA molecules, in simple terms, (long ncRNAs, lncRNA) are non-protein coding transcripts longer than 200 nucleotides [Kung 2013], [Perkel 3013]. This is more or less an arbitrary feature to distinguish long ncRNAs from small regulatory RNAs such as microRNAs (miRNAs), short interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs), and other short RNAs. They are an abundant component in the human transcriptome and have been implicated in several cellular functions, including the regulation of gene transcription [Perkel 2013] and development of tissues [Smola 2016], [Brazao 2016], [Perry 2016], [Calabrese 2013] they have also been implied in many diseases [Harries 2011], [Chen 2014], [Fang 2016], [Chen 2016].

These molecules have gained much attention in the recent years as a possible new layer of biological regulation and their possible roles in drug discovery are also explored [Ling 2016], [Matsui 2017], [Li 2015]. It could be said that a new lncRNA is found to be up or down regulated in a particular disease almost on a weekly basis. Though a wide range of lncRNAs have been implicated in a range of developmental processes and diseases, much is to be learnt yet about their mechanism of action. They have been found to be a diverse class of RNAs that engage in numerous biological processes across every branch of life [Quinn 2016], [Mercer 2013]. A total of 20 lncRNA sequences taken from the NCBI GenBank [NCBI record] were used in this work. In this section we discuss an overview of the salient features, and functions of lncRNA which are known.

3.7.1. Discovery of long ncRNA

Even in the genomics era lncRNAs continue to remain more or less in the dark due to their low expression levels, their presence in specific cell types only or their existence during narrow time frames [Mercer 2008], [Cabili 2011, [Gloss 2015]. These molecules were identified as a class of RNA molecules in 2002 [Okazaki 2002] even though some lncRNA such as H19 and Xist were known since the early 1990s [Brannan 1990, Brockdorff 1992]. Long ncRNAs are usually transcribed by RNA polymerase II (Pol II), spliced, and mostly polyadenylated [Carninci 2005, Bertone 2004]. They are thought to be comparable to protein coding genes but with seemingly low coding potential [Niazi 2012], [Kung 2013]. The development of novel DNA sequencing technologies [Wheeler 2013], [Offit 2014] have revolutionized our understanding of the human genome and its transcriptome complexity. Only 1 – 2% of the whole genome codes for proteins and it is understood that 80% of the remaining is actively transcribed [Carninci 2005], [Clark 2011]. These noncoding portions of the genome produce a variety of regulatory RNAs which have been found to range from miRNA to the heterogenous category of long noncoding RNA. Long noncoding RNAs have been found to differ in their biogenesis, properties and functions [Engstrom 2006], [Kapranov 2007].

3.7.2. The human genome and lncRNA

In the study of vertebrate genome, thousands of genes that code for lncRNAs have been identified. Eukaryotic genomes transcribe [Ponting 2009] a wide spectrum of RNA molecules which include long protein-coding mRNAs to short noncoding transcripts. One of the striking observations made from transcriptome studies is that a much larger fraction of the genome is represented as exons in mature RNAs than what would be predicted from the amount of DNA covered by exons of protein-coding genes. Long ncRNAs are the major component of this all-encompassing transcription [Ponting 2009]. Early studies revealed that only around 5% - 10% of the human genome is accounted for, by mRNA sequences and spliced noncoding RNAs that are transcribed in cell lines. It means that only around 1% of the human genome encodes proteins, leaving around 4% – 9% that is transcribed but whose functions are largely unknown [Ponting 2009]. Recent studies suggest that out of the human genome transcribed, only 2% accounts for protein-coding exons [Boon 2016]. The exonic portion of human lncRNAs accounts for 1% of the genome which is about the same amount of DNA as protein-coding exons [Kapusta 2014].

Evolutionary studies prove that there is a large amount of apparently functional, yet non-coding DNA contained in the human genome, the volume of which was estimated to be four times the amount of protein coding sequences [Ponting 2010]. Long noncoding RNAs (lncRNAs) are a part of these ‘functional yet non-coding’ sequences. Mammalian genomes have been found to contain thousands of loci that transcribe long ncRNAs [Joung 2017]. Although there have also been claims that almost the entire mammalian genome is transcribed into functional noncoding transcripts, such claims still remain contentious [Kapustha 2014], [Kung 2013].

Long ncRNAs are implicated as gene regulators and maybe they are more numerous than protein coding genes in the human genome. However they have lower and tighter tissue-specific expression compared to mRNAs and hence their reference annotations are incomplete [Kornieko 2016]. lncRNAs are found abundantly in the human genome [Cabili 2011] and in other vertebrates and plants. The latest version (2017 version) of the GENCODE project, release 27, has annotated 15,778 long noncoding RNA genes [GENCODE v27] and 7569 small noncoding RNA genes out of a total of 58,288. Protein coding genes account for only 34.031% (19,836) of the total number of genes. A pie-chart which shows coding and non-coding components of the human genome based on statistics available in GENCODE version 27 is shown in Figure 3.16.

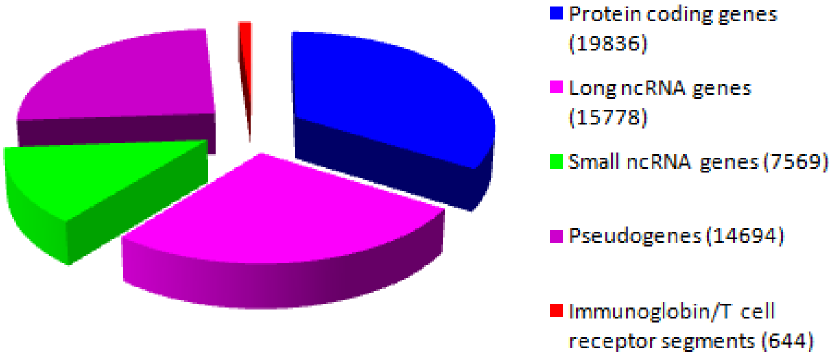


Figure 3.16. A pie chart comparing the number of coding and non-coding components of the human genome, based on values in GENCODE v27

3.7.3. Why study lncRNAs?

lncRNAs have attracted much attention with the availability of increasing evidences that these molecules play critical roles in multiple processes. Epigenetic regulation, chromatin modelling, gene transcription, protein transport, protein trafficking, cell differentiation, organ or tissue development, cellular transport, metabolic processes and chromosome dynamics are just a few examples [Chen 2016], [Cao 2014], [Rin 2012], [Wapinski 2011].

A brief discussion of the various functions of lncRNA in disease and development is presented in this section. Long ncRNAs are functionally heterogeneous. They interact with DNA, proteins and other RNAs to take part in all processes from transcription, intra-cellular trafficking to chromosome remodelling [Quinn 2016], [Rinn 2012], [Chen 2016]. It has been observed that lncRNAs control complex cellular behaviours like growth, differentiation and establishment of cell identity which are often deregulated in cancers [Hu 2012], [Flynn 2014], [Rossi 2014].

3.7.3.1. Involvement of lncRNAs in transcriptional and post-transcriptional modification, and other stages of cell biogenesis

Many lncRNA act as key regulators of transcription and translation and thus influence cell identity and function to a great extent [Mercer 2010], [Chen 2010], [Dinger 2008], [Loewer 2010]. lncRNA targeting mechanisms are diverse. Based on these mechanisms, lncRNAs may play critical regulatory roles in diverse cellular processes such as chromatin remodelling, transcription, post-transcriptional processing and intracellular trafficking [Wilusz 2008], [Chen 2010], [Hung 2010], [Pauli 2011]. Few examples of possible lncRNA targeting mechanisms are represented in Figure 3.17.

lncRNAs have been known to be involved in post-transcriptional regulation. That is, regulation of gene expression after transcription of the DNA into the mRNA molecules has occurred. As already seen in section 3.4.of this chapter, miRNA are molecules which are involved in gene post transcriptional expression. miRNAs combine with proteins to form a complex that binds rather imperfectly to mRNA molecules and inhibits translation. Certain long ncRNAs have been found to interfere with the microRNA pathways involving in different cellular processes [Cao 2014].

Epigenetics is the study of potentially heritable changes in gene expression (active versus inactive genes) that does not involve changes to the underlying DNA sequence — a change in phenotype without a change in genotype — which in turn affects how cells read the genes. Epigenetic change is a regular and natural occurrence but can also be influenced by several factors including age, the environment/lifestyle, and disease state [Weinhold 2006]. In general the term refers to modifications within a cell due to variation in gene expression focusing on genome or proteome of one cell. Epigenetic modifications [Russell 2010]. Epigenetic control is thought to occur at the chromatin-level [Koike 2013, Chang 2012, Wutz 2013]. Chromatin is the combination of DNA and proteins which together make up the cell nucleus [Saffhill 1975, Bustin 1973]. Chromatin is in charge of DNA packaging, gene expression and DNA replication [Prioleau 1994], [Voss 2006]. lncRNAs have been found to interfere with acetylation, methylation and SUMOylation of histones (Histones are highly alkaline proteins found in eukaryotic cell nuclei that package and order the DNA into structural units called nucleosomes.

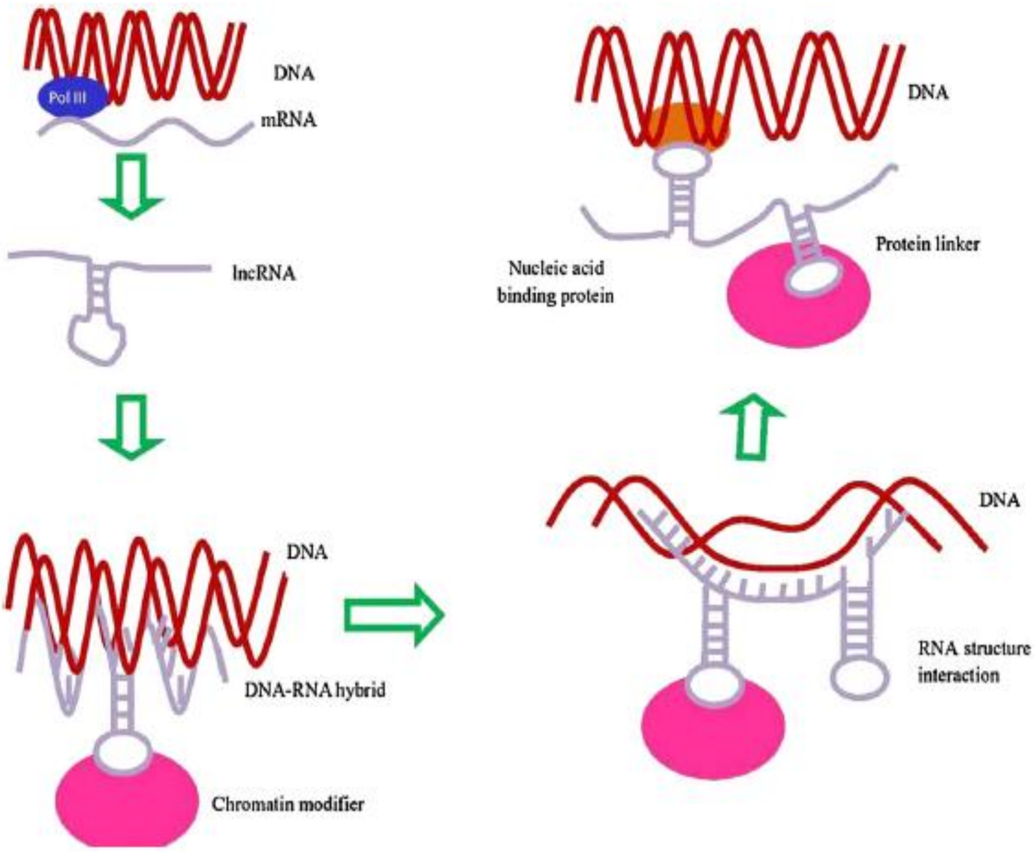


Figure 3.17. Possible lncRNA targeting mechanisms

Through their interactions with DNA, RNA and protein molecules or their combinations, lncRNA act as essential regulator in chromatin organization, transcriptional and post-transcriptional regulation. And aberrations in their expressions have been seen to confer pro-cancer features to the cell [Gibb 2011]. Aberrations in their expressions have been seen to confer pro-cancer features to the cell [Gibb 2011].

3.7.3.2. lncRNAs and cancers in humans

A prominent role of lncRNAs is in the development and progress of cancers which makes the study of these sequences quite relevant [Bartonicek 2016], [Yan 2015(2)]. A wide variety of lncRNAs have been implied in cancers. Figure 3.18 shows examples of various lncRNAs which are implicated in different types of cancer. lncRNAs have different expression levels in cancerous tissues as compared to normal tissues. lncRNAs MALAT1 (Metastasis associated lung adenocarcinoma transcript 1), PCA3 (Homo sapiens prostate cancer associated 3 (non-protein coding) RNA) are examples of some of the lncRNAs which were identified to be associated with cancer in the initial days [Brannan 1990], [Brockdorff 1992].

Six properties required for cell transformation have been termed "***the six hallmarks of cancer***" by Hanahan and Weiberg in 2000 [Hanahan 2000]. These properties are: 1) self-sustained growth signalling, 2) insensitivity to growth inhibition, 3) avoidance of apoptosis (the death of cells which occurs as a normal and controlled part of an organism's growth or development), 4) uncontrolled proliferation, 5) angiogenesis (the development of new blood vessels) and 6) metastasis (the spread of a cancer from one organ/part of the body to another organ/part without being directly connected with it) [Hanahan 2000], [Hanahan 2011]. lncRNA molecules have been found to play regulatory roles in majority of these functions [Gutschner 2012]. A simplified, brief discussion of the involvement of lncRNAs in the "hallmarks of cancer" is given below.

1. Growth signalling is vital to cell growth. Signalling happens through signalling pathways to the signal receptors which are present on the exterior of the cell. lncRNAs promote self-sufficiency in growth by acting on the signal receptors. lncRNAs have been observed to specifically bind nuclear receptors [Cathcart 2015] so that there is no need for the exterior cell receptors to receive growth signals through the external pathways. The cells becomes self-sufficient as far as growth signalling is concerned. lncRNAs have been observed to bind nuclear receptors, either alone or by being a part of ribonucleoprotein complexes. Examples are SRA1 (steroid receptor RNA activator protein1) [NCBI website] which stabilizes the estrogen receptor and signals the growth of breast cancer cells [Lanz 1999], [Shi 2001]. lncRNAs like PVT1 (Plasmacytoma Variant translocation1, PVT1 oncogene- long noncoding RNA) [NCBI website] do not affect the receptor instead regulates receptor abundance such that proliferation is ensured [Zhou 2016].

2. Growth inhibition in cells is naturally inhibited by a variety of processes. Evasion of growth inhibition is found to be achieved by lncRNA molecules by influencing tumour suppressor proteins like CDKs (cyclin dependent kinase) [Kitagawa 2013]. Some lncRNAs counter

growth inhibition by regulating the expression of tumour suppressors by influencing various stages of their transcription and translation. [Dimitrova 2014].

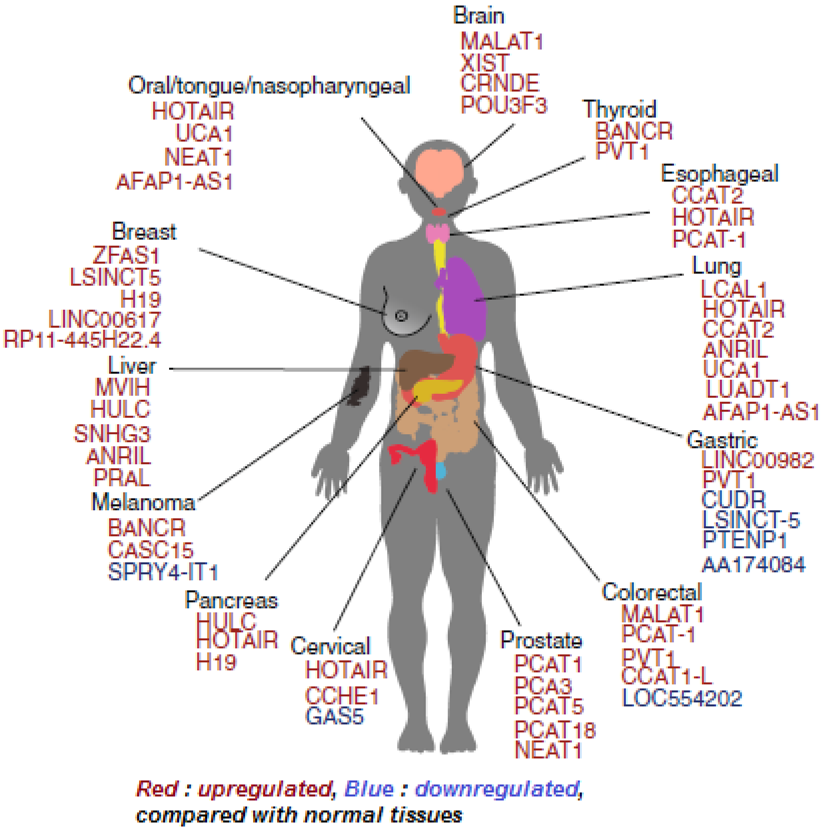


Figure 3.18. Some examples of lncRNAs implicated in various cancers

lncRNAs like gadd7 and cdk6 modifies transcription elongation by destabilization of mRNA transcripts [Liu X 2012]. Transcript stability and translation are also found to be influenced by lncRNA molecules such that repression caused by miRNA is inhibited and a tumour promoter protein is formed [Poliseno 2010].

3. Apoptosis is controlled cell death maintained in all healthy tissues for the control of carcinogenesis (initiation of cancer formation). lncRNAs

have been found to inhibit apoptosis, aiding carcinogenesis and in some cases aiding apoptosis of tumour suppressor proteins which again helps carcinogenesis [DeOcesano-Pereira 2014].

4. Cancer cells have limitless replication potential. This is achieved in cancer cells by maintaining long telomeres (a region of repetitive nucleotide sequences at each end of a chromosome) as nucleoprotein structures that stabilizes ends of chromosomes. In the natural process, telomeres shorten in dividing cells. Hence there is a need for a ribonucleoprotein complex telomerase in order to elongate telomeric repeats through reverse transcription of an internal template RNA. Shortening of telomerases induce production of lncRNA molecules called TERRA (telomeric repeat-containing RNA) [Cusanelli 2015]. When TERRA is activated they recruit protein complexes which bring about homology-directed repair of the shortened /damaged telomeric sequences [Redon 2010].

5. lncRNAs have been found to regulate nutrient supply to tumours by regulating transcription of the protein VEGF (vascular endothelial growth factor) that is essential for the formation of blood vessels. lncRNAs HOTAIR (HOX transcript anti-sense RNA) and MIAT (myocardial infarction-associated transcript) have been found to regulate transcription of VEGF [Fu 2015, Yan 2015(1)]. LncRNA MALAT1 (metastasis associated lung adenocarcinoma transcript 1) when expressed in endothelial cells has been found to promote angiogenic sprouting and migration [Michalik 2014] in cancer of the lung.

6. Many lncRNAs have been implied in the invasive nature of cancer cells that promotes metastasis. MALAT1 in colorectal and nasopharyngeal carcinoma [Yang 2015], CCAT2 (colon cancer associated transcript 2) in lung cancer, are just a couple of examples of lncRNA involvement in cancer metastasis.

3.7.3.3. Involvement of lncRNAs in other diseases in humans

Besides cancers, lncRNA involvements have been associated with the development and progression of a variety of diseases. Some of the diseases besides cancer with which lncRNAs have been found to be associated with include cardiovascular diseases, neurological disorders, diabetes, AIDS, Alzheimer's diseases (AD), cardiovascular diseases and neurodegenerative diseases. The comprehensive lncRNA-Disease database (http://www.cuilab.cn/lncrnadisease), gives more than 200 lncRNA–disease associations. There are more than 200 diseases associated with various lncRNAs and more than 300 lncRNAs playing critical roles in various complex human diseases [Chen 2013]. One lncRNA is found to be involved in many types of diseases due to the functional diversity of these molecules. For example, the lncRNA MALAT1 is associated with cancer of the lung, cancer of the colon and nasopharyngeal cancers. However, the general features of most lncRNAs, such as structure, transcriptional regulation, functions and molecular mechanisms in the biological processes still remain largely unknown and so, annotations of their disease associations too are not complete [Lu 2013], [Li 2013], [Chen 2016]. A brief overview of a couple of known diseases which have high social relevance in the current times and their lncRNA associations is presented here.

1.Alzheimer's disease

Alzheimer's disease (AD) causes dementia or short-term memory loss which is rapidly increasing among all populations throughout the world. AD is a chronic, progressive neurodegenerative disorder, which is caused by the loss of synapses between neurons in specific brain regions (such as the CA1 region of the hippocampus [Palop 2006, Tan 2013]. Researches have shown that the lncRNA BACE1-AS (beta-secretase 1 anti-sense) is involved in AD [Faghihi 2008]. Dementia and AD progress with age and it was found that the RNA BC200 (RNA Brain Cytoplasmic 200) was found significantly up-regulated in brain areas that have developed AD as against brains in a lower age group [Ng 2013, Mus 2007].

2. Heart failure (HF)

Heart failure is a clinical situation with very high rate of mortality [Li 2013], [Barsheshet 2012]. Several lncRNAs have been found to be associated with it [Chen 2016]. lncRNA Fendrr (FOXF1 adjacent non-coding developmental regulatory RNA, coded by the FOXF1 gene) [Ren 2014], [Xu 2014] has been found to play crucial roles in the development of embryonic vasculature. Mutations which inactivate the FOXF1 (Forkhead Box transcription factor 1) gene and affect Fendrr have been observed in patients with acute cardiovascular problems [Ren 2014]. Trpm 3 (TRPM is a family of transient receptor potential ion channels (M standing for melastatin)) and Scarb2 (scavenger receptor class B member 2) [Chen 2016], [Li 2013]. These lncRNAs have been found to have critical functions in heart development and in heart failure too. It is also thought that these lncRNAs could have key role in the developing therapies for heart failure [Papait 2013]. Long ncRNA Nkx2-5 (NK2 homeobox 5) genetically modifies myotonic muscular dystrophy RNA toxicity which has a vital role in heart dysfunction [Schonrock 2012]. Long ncRNA LIPCAR/MT-LIPCAR (mitochondrially encoded long non-coding cardiac associated RNA) is found associated with myocardial infarction. It is seen downregulated early after myocardial infarction but found upregulated in later stages. LIPCAR is a novel biomarker of cardiac remodeling and predicts future death in patients with heart failure [Kumarswamy 2014].

3. Other diseases

The lncRNA PVT1 (plasmacytoma variant translocation 1) has been found to be linked with the development and progress of Diabetic nephropathy [Alvarez 2011], besides its active involvement in breast and ovarian cancers [Guan 2007]. Pvt1 is the oncogene which codes for lncRNA of the same name. (Homo sapiens Pvt1 oncogene (non-protein coding), long non-coding RNA. NCBI accession number of the lncRNA is NR_003367.3) lncRNAs have been found to have roles in neurodegenerative disorders [Salta 2012, Qureshi 2010] and brain development [Qureshi 2012]. Studies of patients with alcohol addiction reveal upregulated MALAT1 in the cerebellum, hippocampus, and brain stem [Kryger 2012], which suggests that the lncRNA network may have key roles in neurodegenerative processes in Huntington’s disease [Johnson 2012]. Long ncRNAs have been found to be involved in many other diseases as well. Discussing each case is outside the area of interest of this Thesis. What is intended in this sub-section is to convey to the reader why the study of lncRNA in all its aspects is beneficial to humanity.

3.8. Conclusion

The non-coding gene which was largely ignored in the initial days of molecular biology have come to the centre space after the prime role it occupies in the various stages of biogenesis of organisms have come to light. The noncoding RNA molecules which were known from early days of molecular biology are molecules like tRNA and rRNA. The central roles of these molecules in the formation of proteins is well understood. The noncoding portion of the eukaryotic genome has been found to be responsible for the formation of a large variety noncoding RNA molecules. Small noncoding RNA sequences like miRNA, siRNA, snRNA, snoRNA, which were discovered later have been found to play pivotal regulatory roles in protein formation as well. In-depth study of these molecules is vital in understanding the gene expression and suppression which controls the occurrences of genetic diseases. Aberrations or mutations in these molecules which inhibit their proper functioning could also lead to pathologic situations which are not genetic but confined to that individual in whom this aberration occurs.

Genomic studies have demonstrated that although less than 2% of the mammalian genome encodes proteins, at least two thirds is transcribed. The noncoding portion of the genome, especially the human genome encodes another wide range of noncoding RNA molecules which are called long ncRNA. These were dismissed as “transcriptional noise” even in the genomic era. But they have been found to play critical roles in various biological processes. These molecules have been found to act as regulators at different levels of gene expression including chromatin organization, transcriptional regulation and post-transcriptional control. This means that long ncRNAs control all stages of cell biogenesis and had critical roles in development and diseases. As much as they are vital to development, evidences from researches prove that mutations and dysregulations of these long ncNA molecules are linked to diverse human diseases ranging from neuro-degeneration to cancers.

From these facts it is evident why the study of such molecules is important. Study of noncoding RNA molecules is central in molecular biology today and they are immensely researched in drug discovery too. Computational methods have been widely used to analyze the genome and is available in literature. In this work, novel approaches based on digital signal processing methods are made use of to analyze the noncoding genome.

Chapter 4

MFE based Prediction of ncRNA Secondary Structure

The purpose of this chapter is to introduce the reader to the concepts of MFE and secondary structure of RNA and their importance. The secondary structures of four classes of non coding RNA sequences are found out making use of the established folding algorithm based on the thermodynamic nearest neighbour model and the MFEs recorded.

Abstract

After the study of different genomes which unveiled enormous information over the past few decades, noncoding RNA has gained prime importance in genome studies. It has been found that the function of ncRNA is decided by its secondary structure to a great extent. Secondary structure of RNA molecules is conserved across organisms rather than the sequence itself. In this chapter, the optimal secondary structure or the minimum free energy (MFE) structure of more than 200 non-coding RNA belonging to four different classes is found out based on the thermodynamic nearest neighbour model. The MFEs are also recorded. The thermodynamic nearest neighbour algorithm makes use of the established principle of free energy minimization. Free energy minimization has been a very popular method for RNA secondary structure prediction for almost three decades.

4.1. Introduction.

One of the most important recent advancements in molecular biology has perhaps been the discovery that noncoding region of the genome can regulate transcription, translation and gene expression. The past three decades have witnessed steep rise in the study of the non-coding RNA. Systematic screening of various genomes has brought to light a completely new knowledge database of the noncoding RNA [Eddy 2001], [Gisela 2002], [Mattick 2006]. Functions of ncRNA include translocation, RNA processing and modification, chromosome replication, to name a few [Garst 2011], [Cech 2014].

As already seen, the structure of bio-molecules governs their function [Tinoco 1999], [Pederson 2000], [Washietl 2012] and many functional RNAs have well conserved structures across species [Eddy 2014]. RNA is a single stranded molecule, which folds onto itself due to complementary base-pairing via hydrogen bonds. RNA involves in complementary base-pairing via hydrogen bonds (A-U, C-G, Watson-Crick/canonical base-pairing) in the same strand [Eddy 2001], [Gisela 2002]. The folded structure thus obtained is the secondary structure of the RNA molecule. RNA secondary structure is seen to influence every step in gene expression [Wan 2011].

Today, there are many computational approaches to predict secondary structure of noncoding RNA sequences. Dynamic programming with the thermodynamic nearest neighbour approach is a popular method of minimum free energy (MFE) secondary structure prediction of RNA. This folding algorithm uses a nearest neighbour energy model. A secondary structure is uniquely decomposed into sub-structural elements (stacked bases, hairpin-loops, bulges, interior-loops and multi-way-junctions) which are assigned energies. The free energy of the secondary structure is computed as the sum of energy contributions of the individual substructures that make up the secondary structure. MFE based secondary structure prediction of four types of noncoding RNA molecule sequences is presented in this chapter. The importance of minimum free energy in the structure of the molecule and its function is also highlighted.

First we will see a brief re-cap of the basic ideas that have been discussed in detail in Chapter 3. The complete genetic information or the genetic code pertaining to an organism is stored in its DNA [Watson 2007], [Alberts 2007]. Nucleic acids – DNA and RNA are both involved in the storage and transmittance of this genetic information. DNA is found in combination with proteins within the chromosome inside the nucleus and RNA outside the nucleus in the case of eukaryotes [Alberts 2007], [Lodish 2000]. DNA has two strands which entwine with each other because of complimentary base-pairing via hydrogen bonds to form a double helical structure. Though genetic studies were DNA centric initially, the rapid advances made in microbiology research has shifted the attention towards the study of the RNA [Eddy 2001], [Mattick 2006], [Gisela 2002], [Matera 2007]. There are quite a few studies which analyse small noncoding RNA using computational methods [Yoon 2007(1)], [Eddy 2002], [Pederson 2000], [Washietl 2012]. But Digital Signal Processing (DSP) based methods that study noncoding RNA are not found in the existing literature.

4.2. Secondary structure of RNA and its relevance

Unlike the DNA, RNA is a single stranded molecule, read from the 5' end to the 3' end, which folds onto itself due to nucleotide pairing via hydrogen bonds between the bases. The reason behind base-pairing is the fact that isolated bases are unstable. RNA is made up of the four nucleotide bases, A (adenine), U (uracil), C (cytosine), G (guanine). RNA involves in complementary base-pairing via hydrogen bonds (A-U, C-G, Watson-Crick/canonical base-pairing) in the same strand [Eddy 2001], [Gisela 2002]. The folded structure thus obtained is the secondary structure of the RNA molecule. Many functional RNAs have secondary structures that are well conserved across different species. In fact, it is often said that a guiding rule of molecular biology is that the structure is more conserved than the sequence itself [Eddy 2013]. While pairing, both the canonical i.e. Watson-Crick pairs (A-U, C-G) and the non-canonical (G-U, A-A) pairs can be formed (the non-canonical pairs are also called wobble pairs). The primary structure of a DNA/RNA molecule, is the sequence expressed from the 5' end to the 3' end. RNA molecules that have the same primary structure may have different secondary structures. Also, molecules that have different primary structure may fold into the same secondary structure [Yoon 2007]. An example of the latter is shown below.

Figure 4.1 shows two sequences, A-A-A-A-C-C-U-U-U and C-U-A-A-C-C-U-A-G. Obviously the sequences though of the same length have different primary structure. But they can have the same secondary folded structure as shown. For an RNA molecule the secondary structure is a set of base pairs. Biomolecules have a tertiary structure which is formed by the 3D folding of the secondary structure. The *quaternary structure* describes how several biomolecules come together and interact to form larger aggregated structures.

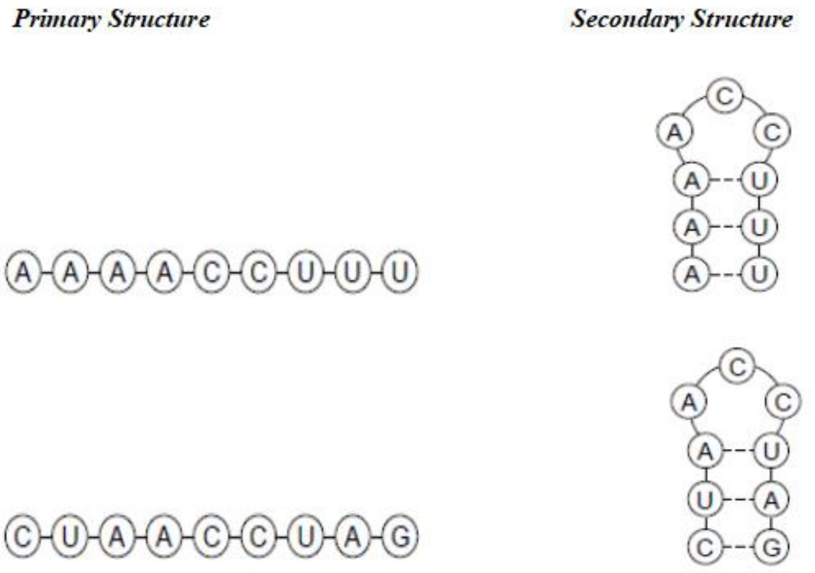


Figure.4.1. Sequences having different primary structure but the same secondary structure.

The *secondary structure* of a biomolecule describes the structural elements that are important in the formation of its three-dimensional tertiary structure. The 3D structure is thought to store all the genetic information about the molecule [Pedersen 2000] and it decides the function [Tinoco 1999]. But formation of tertiary structure does not alter the secondary structure and the secondary structure is made up of substructural elements, which are responsible for most of the overall folding energy and can be seen as a coarse-grained approximation of the tertiary structure. Besides, the secondary structure is formed prior to and independent of the tertiary 3D structure [Washeitl 2005], [Wan 2011], [Washietl 2102]. Thus the secondary structure obviously is the first step in understanding the far more complicated three-dimensional tertiary structure and thereby the function of the ncRNA sequence. A schematic diagram of the primary, secondary and tertiary structures of an RNA sequence is shown in Figure 4.2.

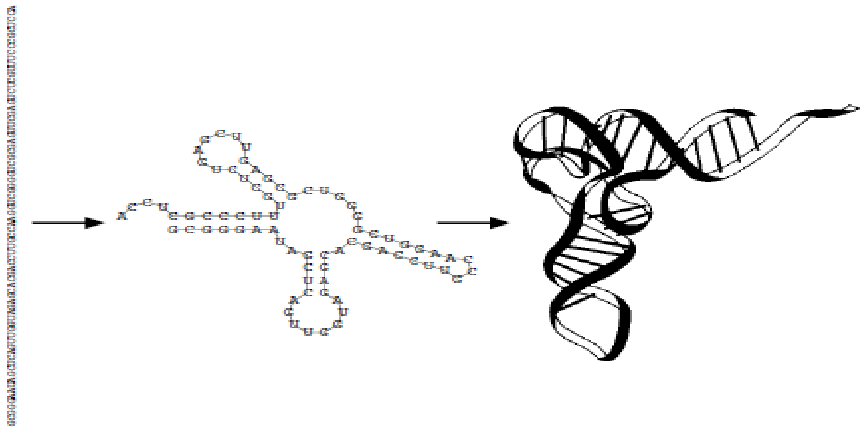


Figure.4.2. Representation of primary, secondary and tertiary structure of an RNA sequence.

Structure of biomolecules governs their function [Tinoco 1999]. Though the tertiary 3D structure is the one which carries the genetic information of the biomolecule, it is not static. It vibrates around an equilibrium called the ’native state’. Structure prediction problem of the biomolecule thus reduces to the one of predicting the native conformation in a model of structure formation. [Pedersen 2000], [Yoon 2007]. RNA secondary structure is seen to influence every step in gene expression [Wan 2011]. RNA secondary structure has been observed to be as important as the genetic code for protein synthesis [Wan 2014].The importance of RNA secondary structure has lead to the development of various approaches in secondary structure prediction [Gardener 2004], [Ding 2004]. Many computational approaches to predict the secondary structure exists today. Broadly, they could be listed as probabilistic, thermodynamic, and phylogenetic predictions and predictions with pseudoknots [Washietl 2012]. But the most popular method is the minimum free energy secondary structure prediction [Washietl 2005], [Mathews 2010], [Hajiaghayi 2012] because of the fact that in the natural environment of a biomolecule, the minimization of free energy is the most decisive factor of structure formation [Pedersen 2000].

In this chapter, the minimum free energy (MFE) secondary structure of the sequences analysed is found out and the MFEs are also recorded. The dynamic programming algorithm based on the thermodynamic nearest neighbour model is made use of in finding the optimal secondary structure of specimen. It makes use of the established principle of free energy minimization. Free energy minimization has been the most popular method for RNA secondary structure prediction for decades. This method of secondary structure prediction is based on an empirical method to find change of free energy denoted as ΔG. It is derived from experiments using the nearest-neighbour model [Zuker 1999], [Zuker 2000], [Mathews 2010]. While predicting the ncRNA secondary structure, a vital point is that many plausible secondary structures can be drawn from a sequence and the number of secondary structures increases exponentially with the length of the sequence. Hence the biologically correct structures have to be distinguished from the incorrect ones. The number of possible secondary structures has been found to be approximately equal to 1.8^{L} where L is the length of the sequence [Mathews 2010], [Wolfsheimer 2010]. So the method resorted to is finding the secondary structure which has the least value for thermodynamic free energy, ie the Gibbs free energy. But it would be too naive to find the free energy of every secondary structure as the computational time would run into impossible values. The dynamic programming algorithm is used to circumvent this problem. The algorithm was formulated by Richard Bellman in 1954 [Dreyfus 2002]. It can be applied to the situations where cost/score is built progressively from smaller solutions.

4.3. The specimen used

As already stated in Chapter 3, the non-coding RNA are that portion of the genome that do not go into protein coding and are never translated [Eddy 2001]. This portion of the genome which constituted the non-gene, goes into forming the ncRNA, and was thought to be “junk-DNA” for quite a long time. The structure, transcription and processing of ncRNA genes are basically different from that of protein-coding genes and this partially explains why ncRNA genes have been largely over-looked. But these ncRNAs have been found to have several other functions. They have been found to be actively involved in core functions in the cell including metabolism, gene expression or suppression, regulation of translation etc. A number of abundant ncRNA gene families have been well studied and research advancements in the last seventeen years have made it clear that the number and diversity of ncRNAs have been largely under-estimated [Eddy 2001], [Wan 2014].

Specimen used in this work were selected from organisms which are quite relevant in biological and medical research viz. Mus musculus (house mouse) and Sus scrofa (pig). The non-coding RNA studied belong to the classes as follows. miRNA (micro RNA), rRNA (ribosomal RNA), snRNA (small nuclear RNA), siRNA (small interfering RNA), snoRNA (small nucleolar RNA). A total of over 200 ncRNA sequences of the above mentioned four classes of ncRNA from Mus Musculus (house mouse), and Sus Scrofa (pig) are analysed in this chapter. These were downloaded from public databases viz., the nucleotide database of National Centre for Biotechnology Information (NCBI).

4.3.1. Organisms

<u>Mus musculus.</u>

Figure 4.3 Mus musculus or house-mouse.

The laboratory mouse is a major model organism for basic mammalian biology, human disease, and genome evolution, and its genome has been sequenced [NCBI Genome Resource].

<u>Sus scrofa.</u>

Sus scrofa a member of the artiodactyls, or cloven-hoofed mammals, is an important model organism for health research due to the parallels with humans. Swine are omnivores and their digestive physiology is similar to humans. Similarities between humans and pigs also exist in renal function, vascular structure, and respiratory rates. Pigs are used as model organism in many areas of medical research including obesity, cardiovascular disease, endocrinology, alcoholism, diabetes, nephropathy, and organ transplantation [NCBI Genome Resource].

Figure 4.4 Sus scrofa or pig.

4.3.2. Sequences

Micro RNA (miRNA).

Micro RNA, abbreviated miRNA, is a small non coding RNA molecule, usually of the length 20 – 24 nucleotides, identified in some viruses and in eukaryotes, nematode to human. miRNAs are well conserved in both plants and animals, and are thought to be a vital and evolutionarily ancient component of genetic regulation. The first miRNA was discovered in the early 1990s. Aberrant expression of miRNAs has been implicated in numerous disease states, and miRNA-based therapies are under investigation. microRNAs (miRNAs) play important roles in gene-silencing and post-transcriptional gene regulation. In animal cells, miRNAs regulate their targets by translational inhibition and mRNA destabilization [Bushanthi 2007], [Cai 2009], [Scott 2011].

Ribosomal RNA (rRNA)

rRNA or the ribosomal RNA is the large molecule which are grouped along with the ncRNA or the non-coding RNA. It is called the cell’s protein factory, but strictly speaking rRNAs do not make proteins, they make polypeptides that assemble to make up proteins. The large rRNA molecules have well defined secondary structures that have been strongly conserved across the evolutionary spectrum, and there is an increasing body of evidence that the rRNA plays key roles in both assembly and function of the ribosomal particles. In every ribosome the bulk of the ribosomal RNA consists of two large molecules, one in each ribosomal subunit. Sequencing of this is done from the RNA or from the DNA. In the case of the latter, the rRNA sequence is found to contain introns. The functions of rRNA are largely dependent on the tertiary 3D structure which in turn is derived from the secondary structure [Brimacombe 1985], [Garst 2011]. Though they are longer than 200 nucleotides, rRNAs have been included along with the other ncRNA in this work for the analysis of MFE of secondary structure.

Small interfering RNA (siRNA)

Small interfering RNA (siRNA), also known as short interfering RNA or silencing RNA, is a class of double-stranded RNA molecules, 20-24 bases in length, falling into the broad class of small non coding RNA. Formation of siRNA is brought about by the Dicer enzyme which catalyzes its production from long double-stranded RNAs and small hairpin RNAs. This is represented in Figure 4.5. siRNAs can also be introduced into cells by transfection.
Since in principle any gene can be knocked down by a synthetic siRNA with a complementary sequence, siRNAs are an important tool for validating gene function and drug targeting in the post-genomic era. RNA interference or RNAi is a prominent area of function of the siRNA. It is an endogenous mechanism of gene expression via siRNA or miRNA, in order to promote messenger RNA degradation, and this is done in a highly sequence-specific manner. The RNAi mechanism was first discovered in transgenic plants in 1990 [Napoli 1990] and later in animal cells (C elegans) [Fire 1998]. Due to this high sequence-specific gene silencing property exhibited siRNA is ideally suited for and is much researched in genomic medicine. Synthetic siRNAs are used in trials, as the inherent physio-chemical properties of naturally occurring siRNA, viz. low charge density, high structural stiffness and rapid enzymatic degradation severely hampers its medical use. The use of small and compact siRNA polyplexes is vital to ensure efficient, systemic siRNA delivery [Lee 2013]. The gene silencing efficiency of siRNA is said to be strongly dependant on the local RNA secondary structure of the targeted region.

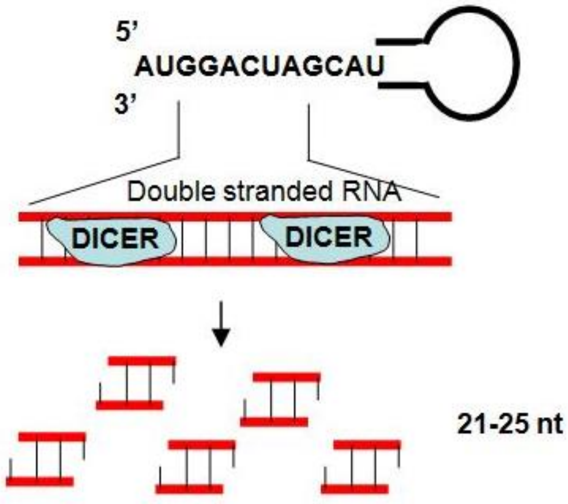


Figure 4.5 Dicer enzyme catalyzes formation of siRNA from double sided RNA(dsRNA)

Small nuclear RNA (snRNA)

A small nuclear RNA (snRNA) is one of many small RNA species confined to the nucleus. Several of the snRNAs are involved in splicing or other RNA processing reactions. Eukaryotic cells contain snRNA, designated U-snRNAs. Their nomenclature derives from their high uridine content [Reddy 1998], [Padgett 2015]. The length of an average snRNA is approximately 150 nucleotides. Studies have shown that their primary function is the processing of pre-messenger RNA in the nucleus [Scott 2011]. snRNAs are transcribed by either RNA polymerase II or RNA polymerase III. rRNA, tRNA, mRNA which comprise 99% of the cellular RNA are largely part of the cell's protein processing machinery. But later studies have revealed the presence of small nuclear RNA which comprises 0.1% – 1% of the total cellular RNA. There has been reported to be evidence for atleast fifteen distinct snRNAs in humans and mice (Mus musculus), out of which six (named U1 to U6) have been found to be capped metabolically stable, synthesized by polymerase II, which are present as ribonucleoprotein particles, and are present in concentrations comparable to that of ribosomes. The metabolically stable U1 to U6 have been found in other organisms including yeast. snRNA molecules, like other ncRNA, play fundamental regulatory roles in gene expression. The U 1 to U6 snRNAs are involved in the regulation of transcriptional elongation. U1 snRNAs are seen to be involved in transcriptional initiation as well [O Gorman 2006].

Small nucleolar RNA (snoRNA).

Small nucleolar RNAs are a class of small RNA molecules that are untranslated, but, guide the chemical reactions in other RNA molecules, like rRNA, tRNA and snRNA. They represent an abundant, evolutionarily ancient group of non-coding RNAs which have diverse functions which include 2'- O - methylation and pseudouridylation of other classes of RNA, nucleolytic processing of rRNAs, synthesis of telomeric DNA, to put down a few [Kiss 2002]. The eukaryotic cell has been found to contain the extremely complex populations of snoRNAs. The snoRNAs can be classified into different groups which are functionally and structurally different [Tollervey 1997]. snoRNAs have lengths varying from 80 to 200 nucleotides and some in yeast are found to be 1000 nucleotides long. By large, the most commonly seen classification of snoRNAs are the two classes : 1) box C/D that guide by base-pairing 2'-O-ribose methylation and 2) box H/ACA which guide by base pairing pseudouridylation of specific rRNA nucleotides. [Dieci 2009].

4.4. Secondary structure prediction using the thermodynamic nearest neighbour algorithm

4.4.1 A brief note on the Nussinov Algorithmn

The dynamic algorithm for prediction of RNA secondary structure was developed by Nussinov, as early as 1978 [Nussinov 1978], [Nussinov 1980], [MIT lecture notes]. Nussinov algorithm predicts the secondary folding pattern of RNA by discovering parts of the given sequences that are complementary, ie with an intent to maximising the base pairs. Unlike similarity search algorithms this algorithm works with a single sequence. It computes the sequence against itself using the dynamic programming table. The letters (A, U, C, G) are treated as ‘matching’ if the corresponding nucleotides would pair. ‘A’ matches ‘U’ and ‘G’ matches ‘C’, but none of the four nucleotides are considered to match with itself as such pairs are not formed, canonically

Dynamic programming table is an approach to implementing the dynamic programming algorithm. Needleman & Wunsch, Smith & Waterman, Four Russians are some of the other popular approaches.

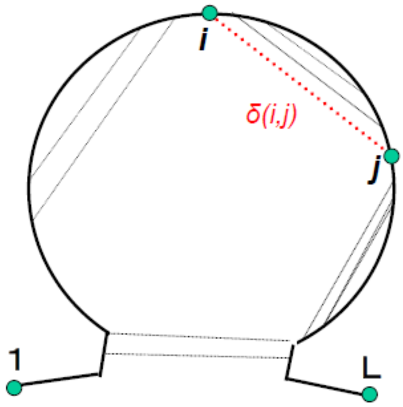


Figure 4.6 Maximising the base-pairs.

The dynamic programming approach makes use of a rectangular table whose horizontal edge is formed with one sequence and vertical edge formed by the other, generally. In the Nussinov algorithm, however only one sequence is used, as the sequence is computed against itself. Cells in the table contains edit distances of the substrings that start from position zero (cell at the top left corner) and end at the current column or row in the table. The alignment between the two sequences is found as the route that follows the lowest value entries i.e. shortest edit distances in the cells from position zero to position 'n' (cell at the bottom right corner).

The Nussinov algorithm works on the idea of maximising base-pairs. In the simplest way, it can be briefed up as follows [MIT Lecture notes]. Figure 4.6 shows a representation of an RNA sequence of length L, the base at position 'i' pairs with the one in position 'j'. 1 is the first base, and L the last base; the sequence has length L.

δ is defined so that it indicates the probability of base-pairing. $\delta(i,j) = 1$; if base i pairs with base j, and $\delta(i,j) = 0$, if they do not pair.

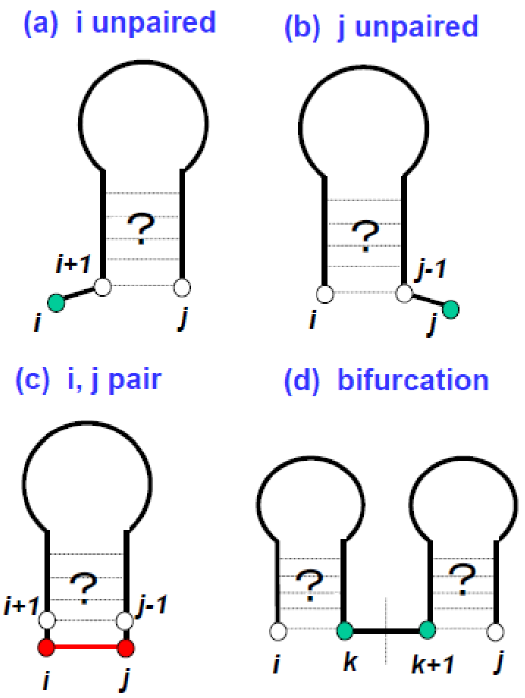


Figure 4.7. Possibilities of Watson-Crick bonding status of the nucleotides

Figure 4.7 shows the position of the possible Watson-Crick bonding status' of the nucleotides to find out matrix E of the algorithm. Figure 4.7 indicates two situations for *(i,j)* unpaired Figure 4.7 (a) and (b), *(i,j)* paired (Figure 4.7(c)) and bifurcation (Figure 4.7(d)).

E is defined such that

$$E(1,L) = \sum_{1\leq i\leq j\leq L} \delta(i,j) \tag{4.1}$$

The Nusssinov algorithm works on the idea of maximising base-pairs. That is,

$$\text{Maximise,}\ \ E(1,L) = \sum_{1\leq i\leq j\leq L} \delta(i,j) \tag{4.2}$$

The basic recursion formula for the Nussinov algorithm can be formally written as;

$$E(i,j) = max \begin{Bmatrix} E(i+1,j) \\ E(i,j-1) \\ E(i+1,j-1) + \delta(i,j) \\ max_{i<k<j} E(i,k) + E(k+1,j) \end{Bmatrix} \qquad (4.3)$$

Here *δ(i,j) = 1*, if characters at positions *i* and *j* are complementary otherwise it is 0 (ie if they are complementary, they pair up). Maximising the base-pairing alone cannot lead to accurate structure prediction. Better structure prediction can be attained by minimizing the free energy of the secondary structure formed [MIT Lecture notes].

4.4.2 Free Energy Minimization Approach

The Nussinov algorithm has certain drawbacks. The one-point focus of the algorithm is maximization of base-pairs and this does not yield biologically relevant structures [Washietl 2012]. Besides, stacking of base-pairs, size of internal loops etc. are not considered. Also, only one structure is predicted, and there is no room for possible sub-optimal solutions. However this has been overcome by Wuchty et.al, which can be considered as an add-on to the Nussinov algorithm [Wuchty 1999]. Nevertheless, the Nussinov algorithm is a classical example of dynamic programming algorithm and all modern variants of folding algorithms in computational biology make use of the same principle [Washietl 2012].

Prediction of RNA secondary structure based on free energy minimization has been a standard for more than three decades now. The basic dynamic programming algorithm for the thermodynamic nearest neighbour model was originally proposed by Zuker and Stiegler [Zuker 1981] and refined, [Mathews 1999], [Zuker 2003], [Markham 2008] later on. Optimal MFE secondary structure is predicted for the sequences analyzed starting from the primary sequence. In this computation, canonical base-pairs (A-U, C-G) and wobble pairs (G-U) are considered, other non-canonical pair formations are ignored. The energy contribution of coaxially stacked helices is not accounted for, and the formation of pseudoknots is forbidden.

The RNA structure can be uniquely decomposed into sub-structural elements (stacked bases, hairpin loops, bulges, interior loops, and multi-way junctions) and energies are assigned to these substructures. An up-to-date set of energy parameters is maintained by the Turner's Laboratories [Mathews 1999], [Xia 1998]. MFE is estimated in kilocalorie per mole by summing individual energy contributions from the secondary substructures, viz., base pair stacks, hairpins, bulges, internal loops, and multibranch loops. Figure 4.8 shows a sample illustration for the contributing energies of the different substructures and the net energy Δ*G* expressed in kilocalorie per mole.

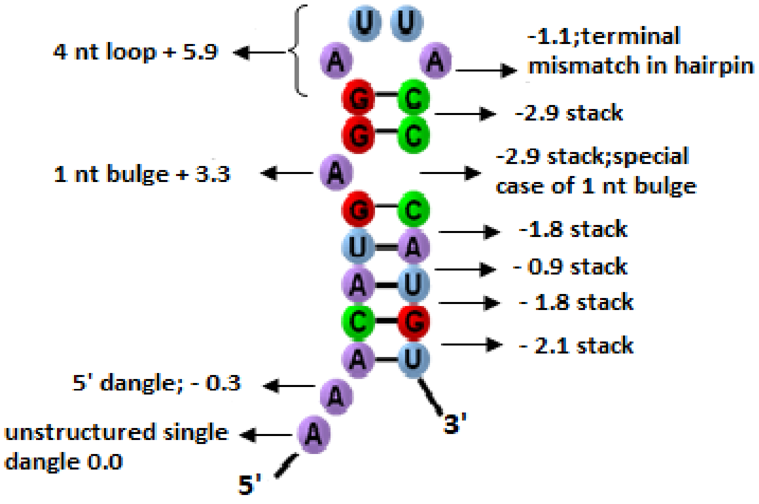


**Figure 4.8. The contributing energies of sub-structures.
Overall *ΔG* = -4.6 kcal/mol.**

A brief explanation of the figure starting from the 3' end follows. The energy contribution of the C-G pair stacked atop the A-U pair at the 3' end is -2.1kcal/mol. The A-U pair stacked atop C-G pair the contributes an energy of -1.8 kcal/mol and the contribution of the U-A base-pair stacked atop the A-U pair is -0.9 kcal/mol. The G-C pair stacked on the U-A pair contributes an energy of -1.8kcal/mol. G-C pair stacked on top of another G-C pair contributes an energy of -2.9kcal/mol. The energy contribution of the mismatch at the termination of the hairpin contributes energy of -1.1 kcal/mol. A four nucleotide bulge has energy of +59 kcal/mol while a single nucleotide bulge has an energy contribution of +3.3 kcal/mol. These two are de-stabilizing energies. The unpaired nucleotide (A) just after the last base-pair is called the dangle which contributes -0.3 kcal/mol and the last nucleotide at the 5' end is called an unstructured dangle which has no energy.

The secondary substructures have energy contributions that are sequence and length dependent [Mathews 2010] and are experimentally determined. Douglas Turner's lab maintains an up-to-date database of energy parameters [Xia 1998], [Mathews 1999]. The algorithm implemented uses dynamic programming to compute the energy contributions of all possible elementary substructures and then predicts the secondary structure by considering the combination of elementary substructures whose total free energy is minimum [Zuker 1999], [Mathews 1999], [Mathews 2010].

We will see a brief over view of the dynamic programming algorithm for the thermodynamic nearest neighbour model which is made use of here in predicting the optimal/MFE secondary structure of ncRNA sequences analysed. In classical thermodynamics [Trout and Tester], [MIT Open course ware] the Gibbs free energy, denoted by *G* describes the energetics of a system of gas molecules in equilibrium or molecules in some aqueous solution. It can be stated as

$$G = H - TS \qquad (4.4)$$

Where, *H* is the enthalpy (potential to perform work), *T* the absolute temperature (in kelvin) and *S* the entropy (measure of disorder).

In the case of a nucleotide sequence, the enthalpy is contributed by base-pairs and entropy by the disorder of being unpaired i.e. "disorder in unpaired regions", and the difference in free energy can be notated as ΔG. The change ΔG of the free energy in a chemical process, such as nucleic acid folding, determines the direction of the process:

- $\Delta G = 0$ indicates equilibrium,
- $\Delta G > 0$ indicates an unfavourable process and
- $\Delta G < 0$ indicates a favourable process.

$$\Delta G = \Delta H - T.\ \Delta S \qquad (4.5)$$

Hence, bio-molecules in solution arrange themselves so as to minimize the free energy of the entire system (bio-molecules + solvent). The net energy of a nucleotide sequence, RNA sequence, is the sum of the energy contributions of the individual sub-structures in its secondary pattern. While measuring the free energy, flexible rules are applicable to the different energy sub-structure viz. loops, stacks etc. Each secondary structure element is defined by its closing base-pair. The complete free energy is found as a summation of energies of these individual sub-structures. Secondary structure is mapped on the basis of the free energy. The one with the minimum value of Gibbs free energy is the most stable structure. By making use of the Bioinformatics toolbox of the platform, MATLAB 2015B, the sub-optimal structures have been avoided. Only the optimal secondary structure is mapped out. Next section gives a brief view on the different energy sub-structures in an RNA secondary structure.

4.4.2.1. Secondary sub-structures

The different secondary sub-structures are shown in Figure 4.9. Given S is the fixed RNA sequence and P a possible RNA secondary structure for S, i and j two locations of nucleotides in the sequence. ΔG, the change in Gibbs free energy due to formation of the structure *'P'* is computed as the sum of contributions from loops, base-pairs and other secondary structure elements. This method does not handle pseudo-knots. Also, hairpin loops with less than 3 nucleotides are not considered. Energies of stems are calculated by adding stacking contributions for the interface between neighboring base pairs.

The ΔG for the RNA molecules folding into a certain structure P is calculated from the logic, $P_{unfolded}$ = { }. Let S = $(x_1, x_2 \ldots\ldots\ x_n)$ be a string over the alphabet $\sum = \{A,\ G, C, U\}$. P is an RNA structure of S having 'n' bases. The base 'i' pairs with the base 'j' if they can form any of the canonical pairs or the wobble pair.

The base 'i', $1 \leq i \leq n$ in P is unpaired iff there is no 'j' such that $(i, j) \in P$ or $(j, i) \in P$

The secondary sub-structure elements are defined as follows.

A. ***Hairpin loop***: The unpaired nucleotide bases enclosed by a base-pair *(i, j)* can be called a hairpin loop, shown in Figure 4.9 (a),

$$\text{if } (i, j) \in P \text{ and if } i < i^{'} \leq j^{'} < j : (i^{'}, j^{'}) \notin P\ ;$$
$$\text{provided } \{i', j'\} \in \{n\};\ i < i' < j' < j$$

Here, hairpin loops have a constraint that they have at least 3 unpaired nucleotides enclosed in them. So every hairpin loop $(i, j) \epsilon P$ adheres to the constraint: $i < j - 3$

B. ***Stacking***: The base pairs stacked one after the other, shown in Figure 4.9 (b), (base-pair i, j stacked over base-pair i+1, j-1).
if, $(i, j)\ \epsilon\ P$ such that $(i$+i, j-$1)\ \epsilon\ P$

C. Internal loop: Two base pairs, *(i, j)* and *(i', j')* enclose an internal loop of unpaired bases, as shown in Figure 4.9 (c) if, the following holds good.

$i < i' < j' < j$

$(i' – i) + (j' – j) >2$ i.e. there is no stacking in between them.

There is no base pair *(k, l)* between base-pair *(i, j)* and *(i', j')*

A bulge is a special case of an internal loop. An internal loop is called a left bulge, if $j = j' + 1$ and called a right bulge if, $i' = i + 1$

D. Multiloop.

A multiloop encloses multiple loops, and multiple stacking pairs, as shown in Figure 4.9 (d). A k-multiloop consists of multiple base-pairs,

(i, j) .... $(i_k, j_k) \epsilon \ P$ with a closing base-pair $(j_0, i_{k+1}) \epsilon \ P$ with the property that,

$\forall \ 0 \leq 1 \ \leq k \quad : \ (j_0, i_{l+1})$

$\forall \ 0 \ \leq \ l, \ l' \ \leq \ k$ is true that there is no base pair (i', j') $\epsilon \ P$ with

$i' \epsilon \ [j_1 \dots i_{l+1}]$ and $j' \in [j_{l'} \dots \dots i_{l'+1}]$

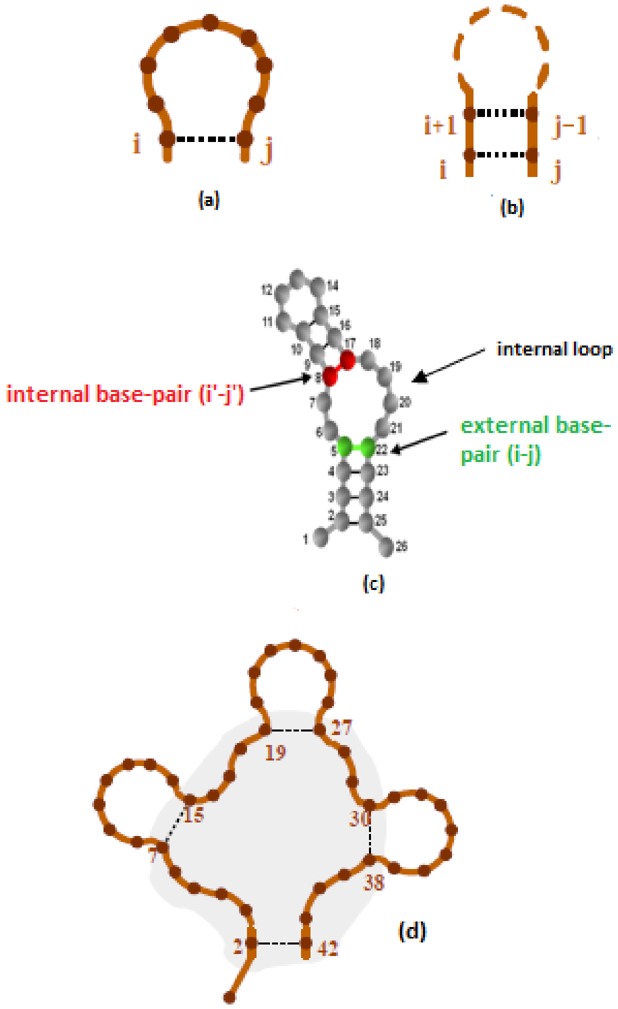


Figure 4.9. The RNA secondary sub-structures. (a) Loop (hair-pin loop), (b) Stacking, (c) Internal loop (*showing internal and external base pairs*), **(d) Multiloop**

$(i_1, j_1) \dots .(i_k, j_k)$ closes the helices of the multiloop

The energy of the various sub-structures are represented here as follows.

Hairpin loop (i, j) : $eH(i,j)$

Stacking base-pairs (i, j) : $eS(i,j,I + 1,j - 1$

Internal loop (i, j, i, j') : $eL(i,j,i',j')$

Multiloop : $eM(j_0, i_1, j_1 \dots \dots. i_k, j_k, i_{k+1})$

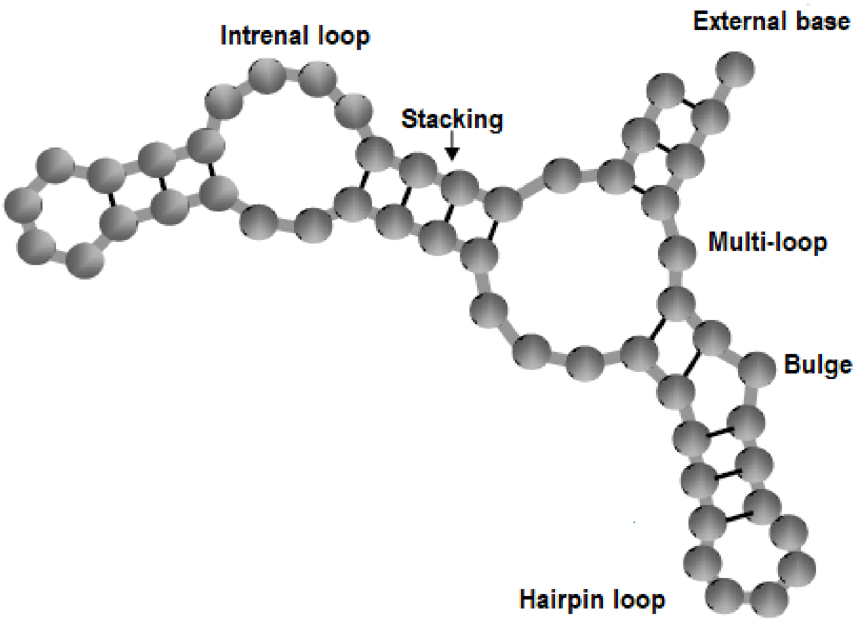


Figure 4.10. A sample RNA secondary structure having different sub-structures

A sample RNA secondary structure having all the above sub-structural elements is shown in Figure 4.10.

4.4.2.2. The algorithm

First matrices *W, V* and *WM* are to be defined.

For a sequence *S* of length *n* with a structure *P*, the *Zuker matrix W* is defined as a matrix of entries W_{ij} for $1 \leq i \leq j \leq n$. *W(i, j)* is the minimum folding energy of all non-empty foldings of the sub-sequence *1* through n. The entries W_{ij} to the matrix are as,

W_{ij} := min { E(P) | P non-crossing RNA i - j substructure of S }, for *$1 \leq i \leq j \leq n$.*

Where *E* indicates energy (Gibbs free energy), and *P* a sub-structure of the total sequence *S. E(P)* can be used to evaluate an *i – j* substructure *P*, since *P* is still an RNA structure. Tacitly, we assume that sequence outside of base pairs does not contribute to the energy.

Initialization: for *(j – i) ≤ m; $W_{ij} = 0$*

Recursion: for *i < (j – 1)*

$$W_{ij} = \begin{cases} minW_{ij-1} & \text{- } j \text{ unpaired} \\ min_{i\leq j\leq k-1} W_{i\;k-1} + W_{k+1\;j-1}E(?\,?) & \text{- } j \text{ paired} \end{cases} \qquad (4.6)$$

The term $W_{k+1\;j-1}E(?\,?)$ is replaced by V_{kj}

Now, we define Zuker matrix *V* as a matrix of entries $V_{\ ij}$ for $1 \leq i \leq j \leq n$ as

$$V_{ij} \coloneqq \text{ min } E(P) \left| \begin{array}{l} \text{P non-crossing RNA } i\text{-}j \text{ substructure of } S \\ (i,j) \in P \end{array} \right. \qquad (4.7)$$

i.e. initialization; (for $j - i \leq m$) $V_{ij} = \infty$

Recursion; for *i < (j -1)*

$$V_{ij} = \begin{cases} min\ eH(i,j) & \textit{hairpin loop} \\ V_{i+1,j-1,j-1} + eS(i,j) & \textit{stacking loop} \\ min_{i<i'<j'j} V_{i'j'} + eL(i,j,i',j') & \textit{internal loop} \\ min_{k,i<j_1<\cdots i_k<j} e_M(i,j,i_1,j_1,\ldots\ldots i_k,j_k) + & \\ \qquad + \sum_{1\leq k'\leq k} V_{i_k,j_k} & \textit{multi-loop} \end{cases} \qquad (4.8)$$

If *V(i, j)* denote the minimum folding energy of all non-empty foldings of the subsequence $x_i, \ldots \ldots\ x_j$, containing the base-pair *(i, j)*, which is non-crossing. The energies of the different secondary sub-structures have already been mentioned earlier. These four arrays hold the minimum free energy of specific substructure of the sub-sequence *P*. Their computations are done inter-dependently and they are calculated recursively using pre-specified free energy functions for each type of loop.

Little is known about all the effects of multi-branch loops on RNA stability.
The total free energy of a multi-loop is given as,

$$E(P) = \sum_{(i,j)\in P} \quad E_{ij}^{P} \tag{4.9}$$

Where E_{ij}^{P} is the energy of individual structural element *S(i, j)* in the multi-loop. RNA molecules fold by intra-molecular base pairing and are stabilized by hydrogen bonds that result from the base pairing. In addition, the stacking of base pairs in a helix also stabilizes the molecule i.e. decreases the free energy of the folded RNA. Loops and bulges destabilize the structure. As the energy contribution of the multi-loop *M* is very high, a simplified form that is practically feasible is used here, viz.

$$M = a + bk + ck' \tag{4.10}$$

where, *a* : energy contribution for the closing loop (it is the constant energy term associated with the multi-loop)

b : number of inner base-pairs

c : number of unpaired bases within the multi-loop

k : number of external base-pairs

k' : number of external unpaired bases

The last row in the energy matrix *V* in equation 4.8, $min_{k,i<i_1<j_1<\cdots..<i_k<j_k<j} eM(i,j,i_1,j_1,\ldots,i_k,j_k) + \sum_{1\le k'\le k} V_{i_k,j_k,}$

Can be replaced by $min_{i<k<j} WM_{i+1k} + WM_{k+1j-1} + a$

Now we define the Zuker matrix *WM*. For an RNA sequence S of length '*n*', the Zuker matrix WM has entries WM_{ij} for $1 \le i \le j \le n$

$$WM_{ij} := \quad minE_{ij}^{m}(P) \left| \begin{array}{l} \text{P non-crossing RNA sub-structure of S} \\ \text{P not empty} \end{array} \right. \qquad (4.11)$$

Where E_{ij}^{m} evaluates P as a part of a multi-loop.

Initialisation for $-i \leq m$; $WM_{ij} = \infty$ (*ij* sub-structure; *P* non-empty)

Recursion; for $i < j - m$

$$V_{ij} \;=\; \begin{cases} min\ eH(i,j) & \textit{hairpin loop} \\ V_{i+1,j-1,j-1} + eS(i,j) & \textit{stacking loop} \\ min_{i<i'<j'j} V_{i'j'} + eL(i,j,i',j') & \textit{internal loop} \\ \quad min_{k,i<j_1<\cdots i_k<j}\, e_M(i,j,i_1,j_1,\ldots..i_k,j_k) + & \\ \qquad\qquad + \sum_{1\leq k'\leq k} V_{i_{k'}j_k} & \textit{multi-loop} \end{cases} \qquad (4.12)$$

The computational time is in the range of around 142 – 148 seconds, varying with the length of the sequence selected. The most practical way to reduce the run time is to limit the size of a bulge or interior loop to some fixed number *d*, usually about 30. An up-to-date set of energy parameters is maintained by Douglas Turner’s Laboratories [Xia 1998], [Mathews 1999]. MFE based secondary structure on small ncRNA sequences were found out making use of the Bioinformatics toolbox of the platform MATLAB (R 2015 B). The secondary structures maps and tabulated MFE values are given in the following section.

4.5. MFE and optimal secondary structures of ncRNA sequences analysed

The dynamic programming approach using the thermodynamic nearest neighbour model was used to predict the MFE based secondary structure of a small sample of over 200 non coding RNA sequences downloaded from the NCBI GenBank. Tabulated MFE values of the sample followed by the secondary structures maps for a random set of 8 specimen, one from each type of snRNA studied are included here. Only 8 secondary structure maps have been included to conserve space. Table 4.1 shows the tabulated MFE Values of the ncRNA sequences. Column 2 of Table 4.1 shows the type of ncRNA sequence along with its GenBank accession, column 3 shows the length of the sequence, column 4 shows the MFE values. The ncRNA sequences are of the two organisms Mus musculus and Sus scrofa.

4.5.1. Tabulated MFE values of ncRNA sequences

Table 4.1. Lengths and MFE values of non coding RNA sequences analysed in this chapter

MUS MUSCULUS							
Sl. No.	**Specimen**	**NT length**	**Calculated MFE (kcal/mol)**	**Sl. No.**	**Specimen**	**NT length**	**Calculated MFE (kcal/mol)**
1	miRNA FV523919.1	24	-3.2	103	miRNA JC031756.1	83	-39
2	miRNA FV523920.1	22	0	104	miRNA JC031760.1	24	0
3	miRNA FV523921.1	24	-3.3	105	miRNA JC031761.1	32	-0.7
4	miRNA FV523922.1	22	-4.4	106	miRNA JC031762.1	31	-6.1
5	miRNA FV523923.1	22	-4.2	107	miRNA JC105046.1	73	-44
6	miRNA FW342845.1	24	-3.2	108	miRNA JC258548.1	21	-6
7	miRNA FW342846.1	22	0	109	miRNA JC258549.1	21	0
8	miRNA FW342847.1	24	-3.3	110	miRNA JC258561.1	73	-33.1
9	miRNA FW342848.1	22	-4.4	111	miRNA JC428339.1	21	0
10	miRNA FW342849.1	22	-4.2	112	miRNA JC428351.1	73	-33.1
11	miRNA FV524066.1	23	-2.5	113	miRNA NR_039546	75	-16.5
12	miRNA FV524067.1	21	-3.9	114	rRNA M27441.1	50	-18.3
13	miRNA FV524068.1	20	0	115	rRNA NR_046153.1	121	-54.3
14	miRNA FV524069.1	22	0	116	rRNA M27443.1	50	-14

15	miRNA FV524070.1	23	-1.4	117	rRNA NR_046118.1	121	-54.3
16	miRNA FW393855.1	23	-2.5	118	rRNA NR_046119.1	121	-54.3
17	miRNA FW393856.1	21	-3.9	119	rRNA NR_046153.1	121	-54.3
18	miRNA FW393857.1	20	0	120	siRNA HW523334.1	20	-2
19	miRNA FW393858.1	22	0	121	siRNA HW523335.1	21	-4.2
20	miRNA FW393859.1	23	-1.4	122	siRNA HW523336.1	21	-0.6
21	miRNA HD065369.1	23	-1	123	siRNA HW523337.1	21	0
22	miRNA HD065370.1	22	-0.5	124	siRNA HW523338.1	24	-0.1
23	miRNA JA368631.1	60	-21.5	125	siRNA HW523339.1	21	-0.9
24	miRNA JC031747.1	61	-5.4	126	siRNA HW523340.1	21	-0.9
25	miRNA JC031755.1	83	-39	127	siRNA HW523341.1	23	-3
26	siRNA HW504921.1	21	-0.9	128	snoRNA NR_028434.1	30	0
27	siRNA HW504922.1	23	-3	129	snoRNA NR_046302.1	69	-10.6
28	siRNA HW504923.1	21	-0.6	130	snoRNA NR_046303.1	67	-13.1
29	siRNA HW504924.1	21	0	131	snoRNA NR_046304.1	71	-17.6
30	siRNA DD346880.1	23	-2.7	132	snoRNA NR_046305.1	72	-23.8
31	siRNA DL076424.1	21	0	133	snoRNA NR_046306.1	71	-24.1
32	siRNA DL076425.1	21	-0.2	134	snRNA M34036.1	54	-14.8
33	siRNA HM596744.1	24	-2.3	135	snRNA X94291.1	200	-77.9
34	siRNA HW040442.1	23	-0.8	136	snRNA X07183.1	63	-18.1

35	siRNA DL076423.1	21	0	137	snRNA X04239.2	138	-49.9
36	siRNA DL076422.1	21	-3.6	138	snRNA NR_028276.1	87	-21.9
37	siRNA DL076421.1	21	-0.3	139	snRNA NR_024201.3	62	-18.6
38	siRNA DL076420.1	21	0	140	snRNA NR_024200.3	165	-68.4
39	siRNA DL076419.1	21	0	141	snRNA NR_004432.2	150	-58.8
40	siRNA DL076418.1	21	-1.2	142	snRNA NR_004414.1	187	-66.8
41	siRNA DL076417.1	21	-0.2	143	snRNA NR_004413.2	166	-70.8
42	siRNA DL076416.1	21	-2.7	144	snRNA NR_004411.3	164	-66.6
43	siRNA DL076415.1	21	-0.7	145	snRNA M34036.1	54	-14.8
44	siRNA DL076414.1	21	-4.1	146	snRNA HQ148158.1	88	-18.2
45	siRNA DL076413.1	21	-7.3	147	snRNA FM991919.1	132	-50.1
46	siRNA DL076429.1	33	-1.2	148	snRNA FM991918.1	97	-34.8
47	siRNA DL076428.1	35	-7.1	149	snRNA FM991916.1	186	-66.2
48	snoRNA DQ267101.1	72	-23.8	150	snRNA FM991912.1	169	-75.8
49	snoRNA AF357362.1	98	-31.6	151	snRNA FM991908.1	214	-89.4
50	snoRNA AF357368.1	48	-12.2	152	snRNA FM991907.1	115	-25.5
51	snoRNA AF357369.1	61	-9.6	153	snRNA BK005202.1	134	-50.9
52	snoRNA AF357371.1	65	-4.2	154	snRNA AB021173.1	29	-4.7
53	snoRNA AF357371.1	65	-4.2	155	snoRNA AF357376.1	58	-8.2
54	snoRNA AF357372.1	62	-13.4	156	snoRNA AF357377.1	68	-14.3

55	snoRNA AF357373.1	63	-9.8	157	snoRNA AF357378.1	63	-9
56	snoRNA AF357374.1	61	-6.8	158	snoRNA AJ278763.1	66	-20.2
57	snoRNA AF357375.1	59	-13.4	159	snoRNA DQ267100.1	71	-17.6
MUS MUSCULUS				**SUS SCROFA**			
Sl. No.	**Specimen**	**NT length**	**Calculated MFE (kcal/mol)**	**Sl. No.**	**Specimen**	**NT length**	**Calculated MFE (kcal/mol)**
58	snoRNA DQ267101.1	72	-23.8	160	snoRNA NR_028129.1	94	-23.6
59	snoRNA DQ267102.1	71	-24.1	161	snoRNA NR_028433.2	71	-17.5
SUS SCROFA							
Sl. No.	**Specimen**	**NT length**	**Calculated MFE (kcal/mol)**	**Sl. No.**	**Specimen**	**NT length**	**Calculated MFE (kcal/mol)**
60	miRNA AM777934.1	23	0	162	miRNA AM777927.1	20	-0.8
61	miRNA JN646111.1	54	-13.5	163	miRNA AM777928.1	21	-1.7
62	miRNA JN646112.1	20	-5.6	164	miRNA AM777929.1	21	-3.4
63	miRNA JN646113.1	17	-2	165	miRNA AM777930.1	20	-5.6
64	miRNA JX185552.1	20	0	166	miRNA AM777931.1	21	-0.4
65	miRNA JX185553.1	20	-2.7	167	miRNA AM777932.1	21	-0.4
66	miRNA JX185554.1	17	-1.50	168	miRNA AM777933.1	23	0
67	miRNA JX185555.1	20	-0.7	169	miRNA JX185562.1	21	0
68	miRNA JX185556.1	22	0	170	miRNA JX185563.1	20	-5.6

69	miRNA JX185557.1	12	-0.7	171	miRNA NR_031532.1	85	-40.2
70	miRNA JX185558.1	21	-0.7	172	miRNA NR_038548.1	80	-32.2
71	miRNA JX185559.1	18	-0.2	173	rRNA AB117609.1.	565	-226.1
72	miRNA JX185560.1	20	-0.7	174	rRNA AB117610.1	380	-138
73	miRNA JX185561.1	21	0	175	rRNA AF080393.1	218	-73.2
74	rRNA KM520146.1	881	-215.60	176	snoRNA JN899136.1	75	-26.9
75	snRNA JN617885.1	17	-1.1	177	snoRNA JN899138.1	69	-16.1
76	snRNA JN617886.1	63	-9	178	snoRNA JN899139.1	141	-33.2
77	snRNA JN617884.1	41	-5.6	179	snRNA JN617883.1	55	-21.2
78	rRNA AF329851.1	83	-30.7	180	rRNA KM520147.1	967	-257.80
79	rRNA AJ583551.1	404	-85.3	181	rRNA KM520148.1	920	-225.40
80	rRNA AJ849443.2	440	-99.7	182	rRNA KM520149.1	980	-251.2
81	rRNA AM158315.1	715	-171.4	183	snoRNA AJ240060.1	73	-21.7
82	rRNA GQ926971.1	440	-101.9	184	snoRNA AJ543323.1	68	-18.1
83	rRNA KC984217.1	207	-46.6	185	snoRNA JN831366.1	132	-48.8
84	rRNA AF080393.1	218	-73.2	186	snoRNA JN899116.1	70	-11.2
85	rRNA KF908861.1	324	-62.7	187	snoRNA JN899117.1	75	-13.9
86	rRNA KJ192659.1	392	-91.1	188	snoRNA JN899118.1	77	-20.4
87	rRNA KJ193217.1	523	-141.60	189	snoRNA JN899119.1	64	-16.5
88	rRNA KJ361825.1	532	-108.00	190	snoRNA JN899120.1	76	-18.9

89	rRNA KM520132.1	979	-262.20	191	snoRNA JN899121.1	71	-11.8
90	rRNA KM520133.1	974	-250.00	192	snoRNA JN899122.1	65	-10.5
91	rRNA KM520134.1	983	-244.80	193	snoRNA JN899123.1	66	-16
92	rRNA KM520135.1	983	-244.40	194	snoRNA JN899124.1	75	-13.7
93	rRNA KM520136.1	949	-244.80	195	snoRNA JN899125.1	85	-18.7
94	rRNA KM520137.1	968	-241.80	196	snoRNA JN899126.1	75	-14.20
95	rRNA KM520138.1	986	-250.60	197	snoRNA JN899127.1	77	-17.2
96	rRNA KM520139.1	981	-247.40	198	snoRNA JN899128.1	72	-12.5
97	rRNA KM520140.1	983	-243.40	199	snoRNA JN899129.1	86	-28.7
98	rRNA KM520141.1	966	-241.80	200	snoRNA JN899130.1	70	-7.9
99	rRNA KM520142.1	939	-228.90	201	snoRNA JN899131.1	80	-17
100	rRNA KM520143.1	981	-246.80	202	snoRNA JN899133.1	82	-12.4
101	rRNA KM520144.1	920	-228.60	203	snoRNA JN899134.1	79	-16.2
102	rRNA KM520145.1	919	-227.50	204	snoRNA JN899135.1	112	-38

Source of sequences : NCBI GenBank

4.5.2. Optimal secondary structure plots.

1. HD065369.1

Secondary structure plot of Musmusculus miRNA sequencewith NCBI accession number HD065369.1. The length of this sequence is 23 and the MFE is -1.0kcal/mol.

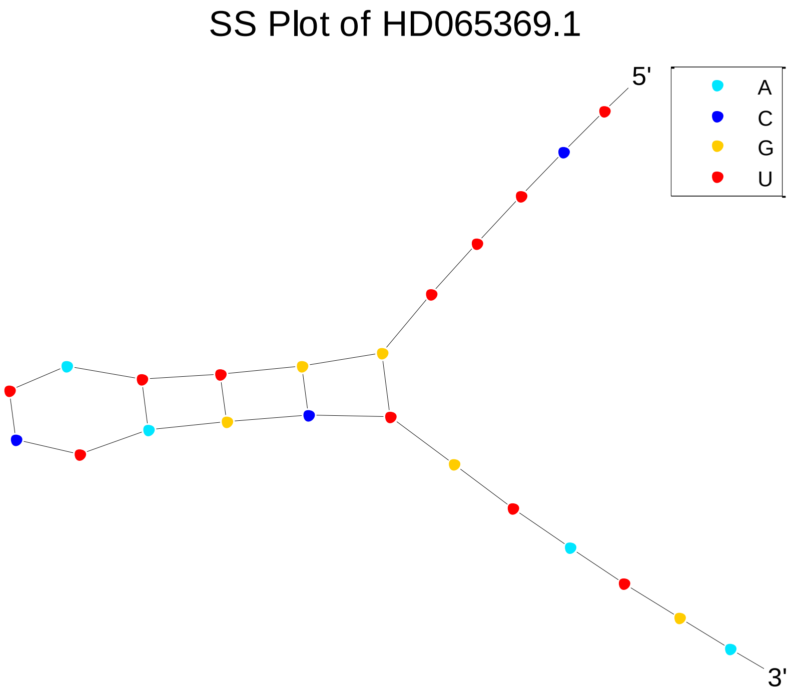


Figure 4.11. Secondary Structure plot of HD065369.1.
Mus musculus micro RNA, miR9.
Sequence 49 from Patent WO2010066384

2. NR_046118.1

Secondary structure plot of Mus musculus rRNA sequence with NCBI accession number NR_046118.1. Length of the sequence is 121 and MFE is -54.3 kcal/mol

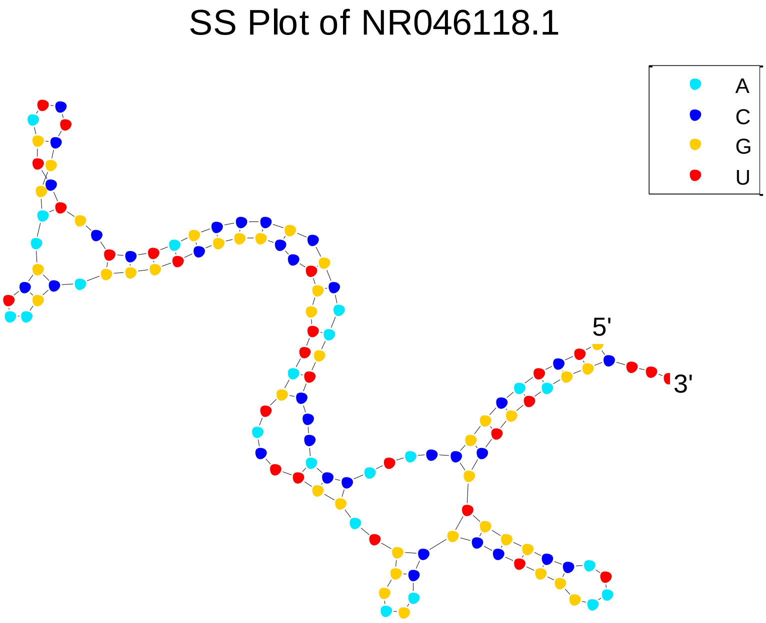


Figure 4.12. Secondary Structure plot of NR_046118.1. Mus musculus ribosomal RNA(rRNA).

3. DL076418.1

Secondary structure plot of Mus musculus siRNA sequence with NCBI accession number DL076418.1. Length of the sequence is 21 and MFE is -1.2 kcal/mol

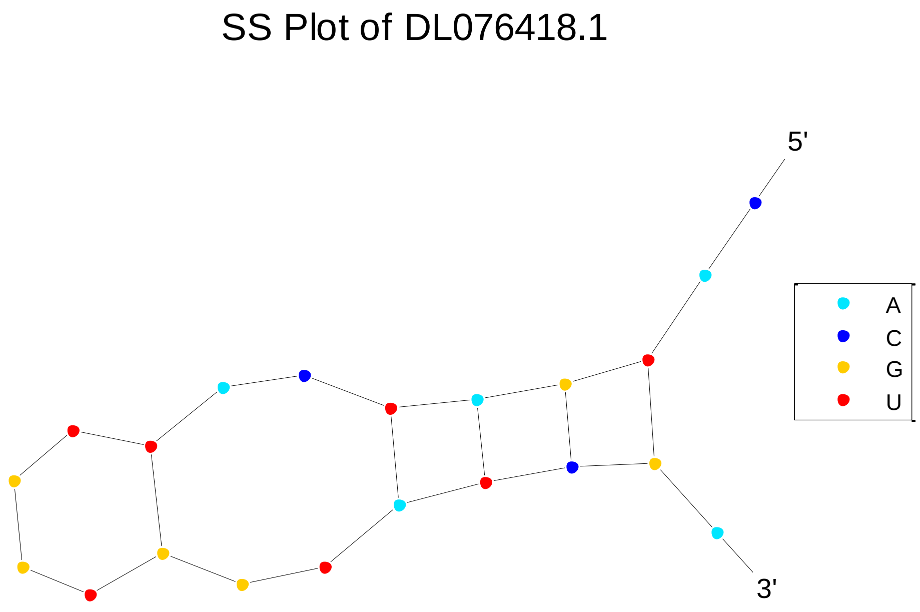


Figure 4.13. Secondary Structure plot of DL076418. Mus musculus small interfering RNA(siRNA)

4. M34036.1

Secondary structure plot of Mus musculus snRNA sequence with NCBI accession number M34036.1. Length of the sequence is 54 and MFE is -14.8 kcal/mol

SS Plot of M34036.1

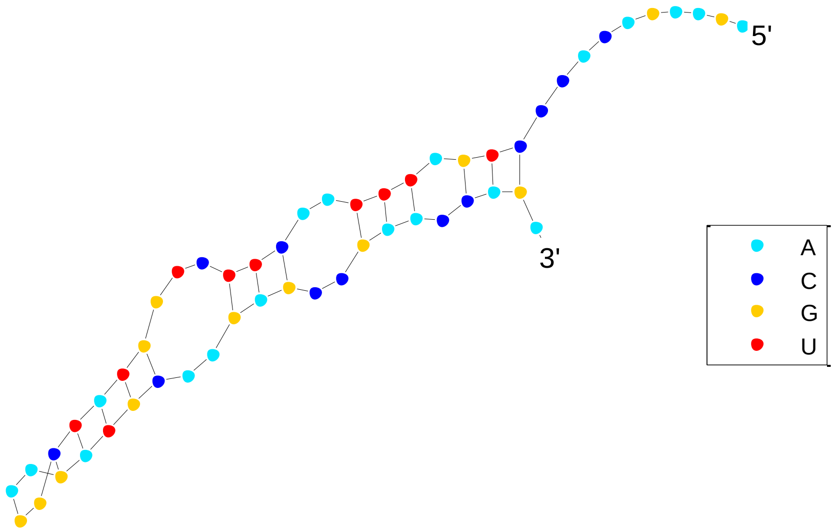


Figure 4.14. Secondary Structure plot of M34036.1. Mus musculus small nuclear (snRNA)

5. AF357371.1

Secondary structure plot of Mus musculus snoRNA sequence with NCBI accession number AF357371.1. Length of the sequence is 65 and MFE is -4.2 kcal/mol

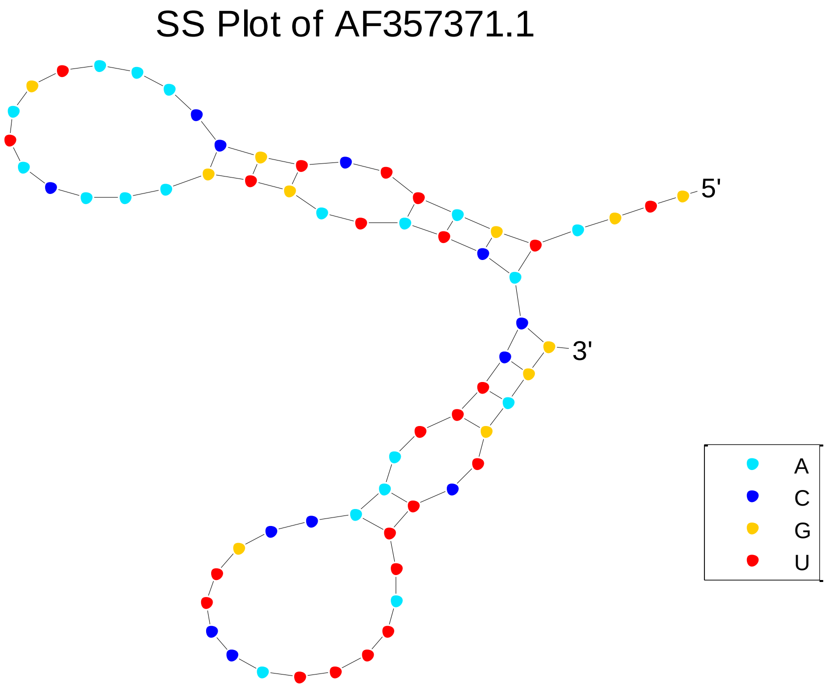


Figure 4.15. Secondary Structure plot of AF357371.1. Mus Musculus small nucleolar RNA (snoRNA)

6. AM777931.1

Secondary structure plot of Sus scrofa miRNA sequence with NCBI accession number AM777931.1. The sequence has length 21 and MFE is -0.4kcal/mol.

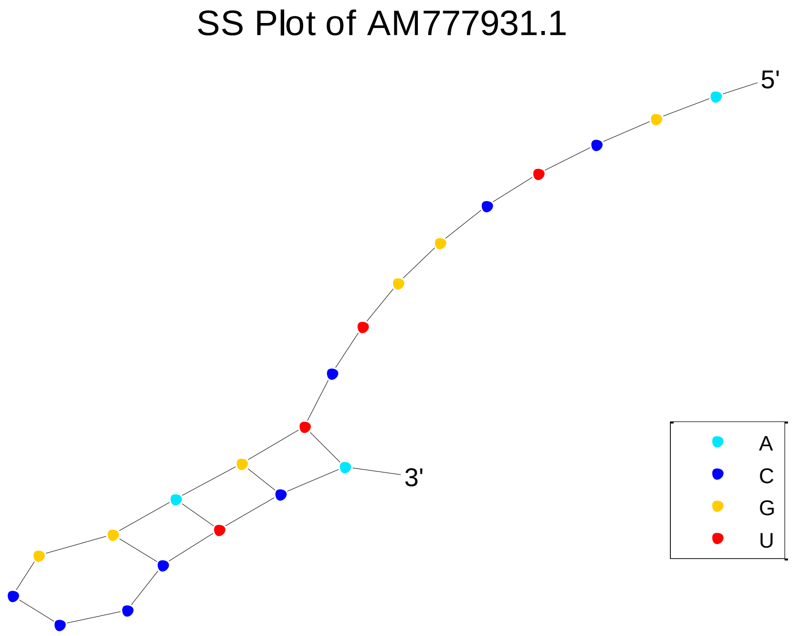


Figure 4.16. Secondary Structure plot of AM777931.1. Sus scrofa microRNA(miRNA), miR-423-clone pMPa3

7. JN899117.1

Secondary structure plot of Sus scrofa snoRNA with NCBI accession number JN899117.1. The sequence is 75 nucleotides long and the MFE of the sequence is –13.9 kcal/mol.

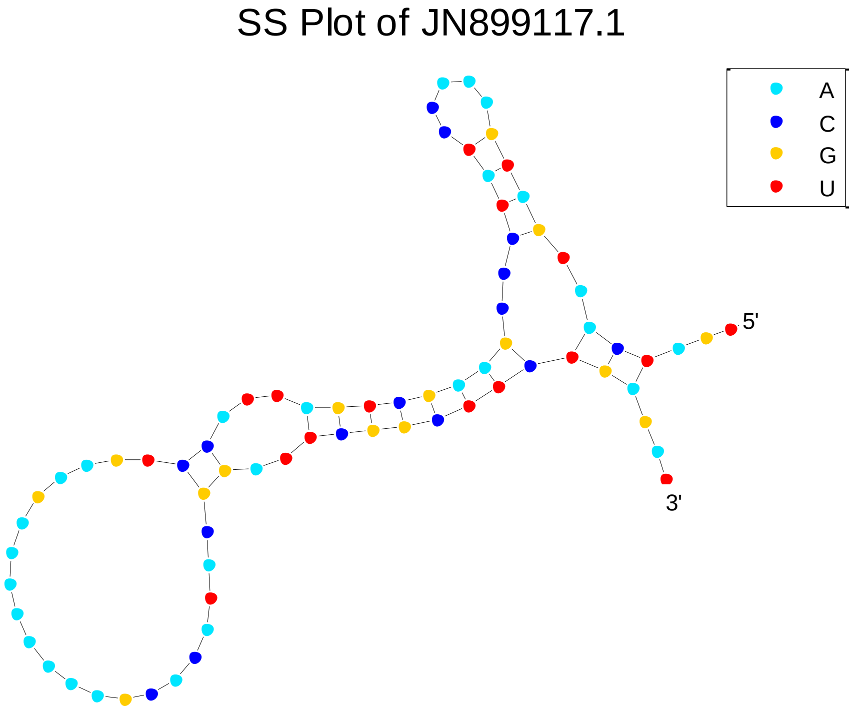


Figure 4.17. Secondary Structure plot of JN899117.1
Sus scrofa pSNORD50B snoRNA complete sequence

8. AF080393.1

Secondary structure plot of Sus scrofa rRNA with NCBI accession number AF080393.1. The sequence is 218 nucleotides long and the MFE of the sequence is -73.2 kcal/mol.

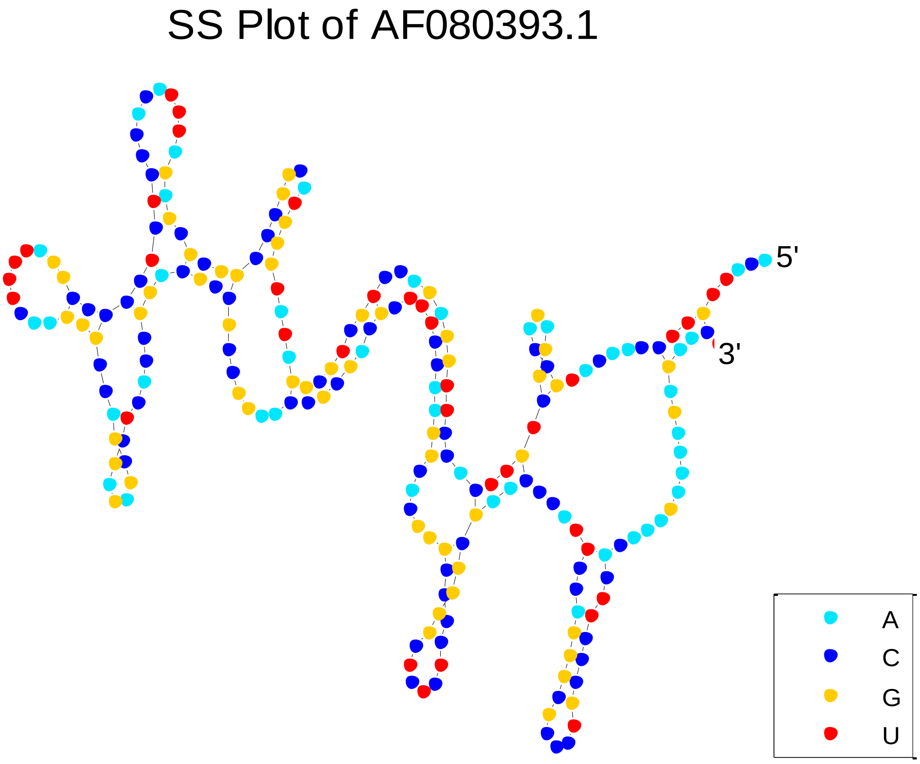


Figure 4.18. Secondary Structure plot of AF080393.1.
Sus scrofa 28S ribosomal RNA(rRNA) gene, partial sequence

4.6. Discussion

This chapter intends to introduce the reader to the concepts of MFE and secondary structure of RNA sequences. The MFE was computed using the folding algorithm based on the thermodynamic nearest neighbour model making use of the platform MATLAB 2015B. There are many energy based algorithms for the prediction of secondary structure of RNA sequences. The most popular among them is MFE based secondary structure prediction [Mathews 2010], [Washietl 2012] because of the fact that in the natural environment of a biomolecule, the minimization of free energy is the most decisive factor of structure formation [Pedersen 2000]. MFE based structure prediction algorithms are found to give results within 2% accuracy with high confidence [Hajiaghayi 2012].

Although the secondary structures arrived at using the free-energy minimization approach are reported to be much more reliable than the ones arrived at by using the base-pair maximization approach, it should be remembered that there are more than one structure with the same value of MFE [Zuker 1989]. Many secondary structures are possible for the same MFE and the number of such possible sub-optimal structures increases exponentially with the length of the sequence. Quality of the mapped secondary structure depends on the accuracy of the thermodynamic parameters and the ability of the NNTM (Nearest neighbour thermodynamic model) to represent the secondary structure. But the optimization step used in the Zuker algorithm rules out the possibility of sub-optimal structures [Zuker 1999]. The Zuker approach used here facilitates determination of a structure which has complementary regions that is stable energy-wise and the optimal one for that sequence length and energy. Here the formation of non-canonical base-pairs has not been considered. Also, this approach does not take pseudo-knots into consideration. The structures mapped here are therefore only of canonical, A-U/T, C-G base pairs and wobble-pairs, without pseudo-knots.

4.7. Conclusion

Recent advancements in molecular biology have brought to the forefront the importance of ncRNA in regulating numerous functions of the cell. Understanding the structure of RNA is one of the keys to understanding its function. The optimal secondary structure or the minimum free energy secondary structure is the fundamental state of a nucleic acid molecule which governs its higher structures and hence the function. The most popular approach of secondary structure prediction is the minimum free energy method. The thermodynamic nearest neighbour approach to finding the MFE based secondary structure was made use of in this chapter.

MFE of sequences is a common index used to study RNA. Structural RNA were found to have more folding energy than random RNA of the same dinucleotide frequency [Clote 2005]. MFE is a vital tool in identifying ncRNA genes [Lim 2003]. Warris et.al. describes a method of prediction of small regulatory RNAs in genomes using MFE distribution of sequences as the discerning factor [Warris 2014].

The importance of MFE in understanding various functions of RNA sequences is evident. Chapters 5 and 6 analyse MFE from a DSP point of view.

Chapter 5

Novel relationship between MFE and Signal Parameters of the ncRNA Sequences

In this Chapter, the relationship between the Minimum Free Energy (MFE) of the RNA secondary structure and the signal parameters of the RNA sequence viz., the length of the sequence and its spectral coefficients is explored. The Minimum Free Energy is computed based on the thermodynamic nearest neighbour model as seen in Chapter 4. The spectral coefficients of nucleotide sequences are obtained by making use of the Discrete Fourier Transform via the FFT algorithm. A novel linear relationship is noticed, between the values of MFE and nucleotide length, MFE and the standard deviation of spectral coefficient matrix of the sequences making use of simple linear regression.

Abstract

Minimum Free Energy (MFE) is a decisive parameter in determining the optimum secondary structure of RNA sequences. For most RNA molecules, there is a good relation between the structure and function and it could be said that the tertiary structure of a bio-molecule decides its function. However, the secondary structure is formed prior to and is independent of the tertiary structure. Also, the secondary structure decides the tertiary structure. Hence it is quite logical to reason that MFE of RNA sequences which decides the secondary structure, controls their function. In this chapter the MFE of noncoding RNA sequences is found out and the possibility of a mathematical relationship between MFE and signal parameters of the non coding RNA sequences is explored. The signal parameters of the sequence are the length of the sequence and the standard deviation of the spectral coefficient matrix. MFE was found to vary linearly as the length of the sequence and also the standard deviation of the spectral coefficient matrix of the sequence.

5.1. Introduction

Digital Signal Processing (DSP) methods have by now become an accepted way of analysing genomic sequences. Digital filtering methods, transform domain analyses, statistical tools etc. play important roles in analysing genomic data [Anastassiou 2001], [Vaidyanathan 2004]. Our understanding of microbiology and the genetic activities at the cellular level was based on the central dogma of molecular biology, in which the flow of genetic information happens from DNA to RNA to protein. Much of the work in molecular biology has been DNA- centric, and so have the DSP methods to analyse the genome. The non-coding region was ignored mostly. However, in the present and the last two decades, with systematic screening of different genomes, a variety of non-coding RNAs have been identified. These have other functions and are not converted into proteins nor do they participate in protein coding directly, unlike tRNA (transfer RNA) or mRNA (messenger RNA) [Watson 2007], [Alberts 2007]. Many researchers have turned to the non-coding region of the gene, arriving at interesting results. It has been shown that non-coding RNAs play vital roles in the biological processes of the cell. Small ncRNAs (non coding RNAs) like miRNA (micro RNA), siRNA (small interfering RNA) are involved in various gene-regulatory functions [Bartel 2004], [Malone 2009], [Jansson 2012], [Ling 2013].

RNA, unlike the DNA is a single stranded molecule, which folds upon itself due to nucleotide pairing via hydrogen bonds between the bases [Eddy 2001] forming what is termed the secondary structure. While pairing, both the canonical Watson-Crick pairs (A-U, C-G) and the non-canonical/wobble (G-U, A-A) pairs can be formed. Some RNAs are found to conserve their secondary structures across different organisms. It is also noticed that RNAs conserve their base-paired structures even when they have primary structures which can hardly be correlated [Yoon 2004].

For many RNAs there is a close relation between structure and function [Washietl 2012]. Also, RNA structure is seen to influence every step in gene expression [Wan 2011], [Wan 2014]. Knowledge of secondary structure aids in understanding the function of the RNA, though the function of the RNA molecule depends ultimately on the tertiary structure [Washeitl 2005], [Washietl 2012] which is formed by the arrangement of the secondary structure elements in space. However, secondary structure is formed prior to, and is independent of tertiary structure, and it decides the tertiary structure. From the secondary structure, the tertiary structure can be found out, by following simple tertiary folding principles [Tinoco 1999], [Washietl 2012]. RNA sequences fold into the secondary structure such that the thermodynamic free energy is minimum, which is the most sTable structure. Thus, we could say that the Minimum Free Energy (MFE) also called the Gibbs free energy, is a thermodynamic entity relating to the secondary structure and hence function of an RNA sequence.

The work presented in this chapter aims to find an association between MFE, which is a thermodynamic property of ncRNA sequences, and the signal properties of the sequence. The ncRNA sequences were converted to digital signals, using appropriate mathematical mapping and the spectrum found out, using a popular DSP tool, the Discrete Fourier Transform (DFT). The standard deviation of the spectral coefficient matrix was computed, for each of the specimen. It was found that the MFE of the ncRNA sequences studied are linearly related, to the length of the sequence in terms of number of nucleotide bases, and also to the standard deviation of spectral coefficients.

5.2. The specimen studied

The specimen studied in this work belong to the following classes; miRNA (micro RNA), rRNA (ribosomal RNA), snRNA (small nuclear RNA), siRNA (small interfering RNA), snoRNA (small nucleolar RNA). The sequences were downloaded from public database, NCBI GenBank. A small sample of around 194 noncoding RNA sequences in all, belonging to the organisms Mus musculus and Sus scrofa are analysed. This number 194 is not a pre-defined number. The total number of sequences used summed up to 194.An overview of the four classes of ncRNAs studied in this work have already been explained in chapters 3 and 4.

5.3. The parameters explored

In this chapter, the relationship between the Minimum Free Energy (MFE) and the signal parameters of four classes of ncRNA sequences analysed is explored. The signal parameters being, length of the sequence in terms of the number of nucleotides in its primary structure, the standard deviation of the spectral coefficients of the nucleotide sequence.

5.3.1. Relevance of MFE and length of ncRNA sequences

The parameters sequence length, and MFE have been used in analyzing RNA from a very early time [Grüner 1996], [Galzitskaya 1998]. Sequence stability of RNA is found to be influenced by length and MFE [Pervouchine 2003], [Trotta 2014]. MFE has also been used as an index to study the relationship between entropy and structural properties of RNA sequences [Wolfsheimer 2010]. Washeitl et.al. describes a noncoding RNA gene finder which makes use of mean and standard deviation of MFE of sequences [Washietl 2005]. Clote et.al. makes use of a method which applies the mean and standard deviations of MFE values in order to differentiate between functional and random RNA sequences [Clote 2005]. Lim et.al describes a technique for identifying miRNA in which MFE values are used as a threshold [Lim 2003]. Warris et.al. describe yet another method of prediction of small regulatory RNAs in genomes using MFE distribution of sequences as the discerning factor [Warris 2014].

As is evident from the above, MFE and sequence length are important parameters to be analyzed in the study of RNA. Computational methods have been widely employed to study noncoding RNA. Even though DSP methods have become as popular as computational methods in the analysis of the coding region of the genome, little work has been done which makes use of Digital Signal Processing techniques to analyze the noncoding genome.

5.3.2. Minimum Free Energy (MFE) – the thermodynamic background

Chapter 4 dealt with the calculation of MFE using a popular algorithm, the Zuker algorithm which makes use of the nearest neighbour thermodynamic model. The free energy principle tries to explain how (in our context biological) systems maintain their order by restricting themselves to a limited number of states. It states that biological systems minimise a free energy function of their internal states [Zuker 1999], [Karl 2012]. The minimum free energy is also termed Helmholtz free energy. The principle of minimum free energy is said to be a re-statement of the second law of thermodynamics which can be stated as follows. In a closed system with constant external parameters and entropy, the internal energy will decrease and approach a minimum value at equilibrium. External parameters could be anything from the volume, to a constant magnetic field. Estimation of MFE is done, classically making use of chemical methods, though computational methods are rampant in use these days.

Minimum Free Energy or MFE is a thermodynamic entity relating to the secondary structure of an RNA sequence. RNA in fact folds into a tertiary structure too, which depends on the secondary structure. Chemically, the two are distinguished from each other by the presence or absence of Mg^{++} ion [Tinoco 1999]. Proteins too fold into secondary and tertiary structures. But in proteins, folding is much more complex owing to the fact that there are twenty residues or amino acids in proteins whereas in RNA there are only four nucleotides or bases. This implies that no direct rules exists by which the tertiary structure of protein can be predicted from the secondary structure [Dorn 2014].

In comparison, RNA structure prediction is much simpler. The four nucleotide bases in RNA are very similar, two of these are purines and the other two are pyrimidines. They differ only in the placement of carbonyl and amino groups and their interactions are either through hydrogen bonding or base stacking. A useful algorithm to predict RNA folding need not tell us anything at all about the folding pathway of the RNA molecule. Such an algorithm will depend on the stabilities of the sub-structures involved, which directly points to the thermodynamic parameters of the structures, and not on the kinetics of folding [Mathews 1999], [Zuker 1999].

5.3.2.1. The Melting Temperature T_m

Thermodynamic parameters of RNA are measured chemically at controlled conditions of temperature (37^0C) and ionic concentrations (Na^+, 1 M NaCl solution) [Zuker 1999]. The formation of base-pairs is temperature dependant and also dependant on the concentration of the Na^+ ions. At high temperatures and low ionic concentrations, base pair formation is not possible, which means that RNA would then exist as the single stranded molecule which does not fold. As the temperature is lowered, or the ionic strength is raised slightly, so that base-pair formation is barely possible. Loops and bulges decrease the entropy of the single strand, so loops and bulges are formed only if the lowering of free energy due to base pair formation more than balances the cost of loop closure. As the temperature is further lowered, or the ionic strength is raised further, more secondary substructures are formed, including the entire range of loops, bulges and junctions.

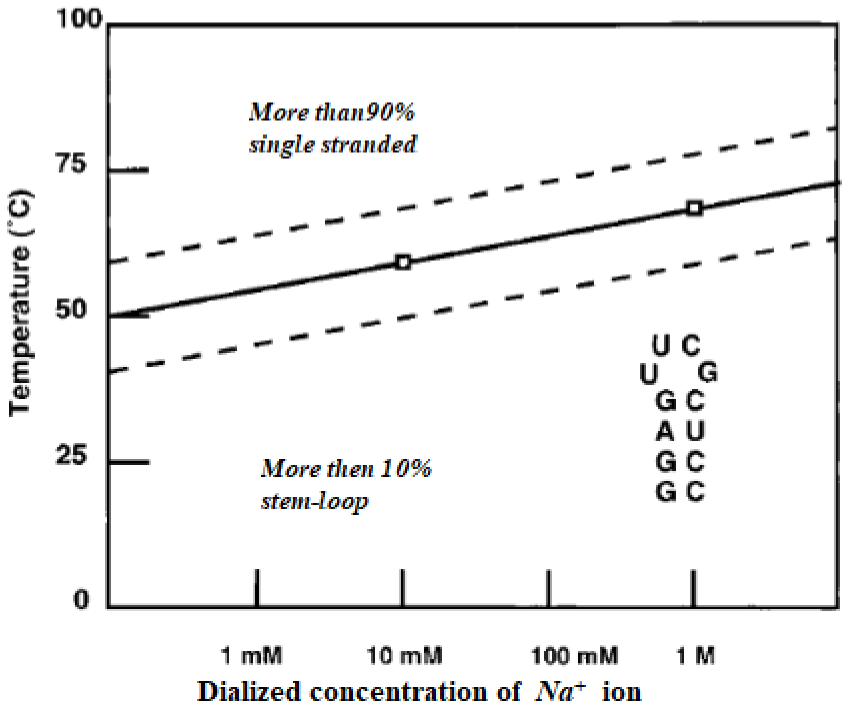


Figure 5.1 Dependence of secondary structure element formation on temperature and Na^+ concentration

The temperature dependence of sub-structure formation is represented in Figure 5.1.

The term nucleic acid thermodynamics refers to the study of the dependence of nucleic acid structure on temperature. As mentioned earlier, the formation of base-pairs occurs at lower temperatures. As temperatures are increased, the RNA or DNA exists in solution as single stranded, as the hydrogen bonds are broken. The T_m is defined as the temperature in degree Celsius, at which 50% of all molecules of a given DNA/RNA sequence are hybridized into a double/coiled strand, and 50% are present as single/uncoiled strands. T_m depends on the length of the DNA/RNA molecule and its specific nucleotide sequence. The nucleic acid molecule is said to have been de-natured by the high temperature. It can be thought that the opposite process of denaturisation is hybridization. At conformable temperatures and ionic concentrations, the single nucleic acid strands (DNA/RNA/Oligonucleotide) form double strands through complimentary base-pairing. Thus, hybridization is the process of establishing a non-covalent, sequence-specific interaction between two or more complementary strands of nucleic acids into a single complex, which in the case of two strands is referred to as a duplex. Nucleic acid strands will bind to their complement under normal conditions [Wu, 2002], [Dirks 2007].

In order to obtain the most energetically preferred complexes, a technique called annealing is used in laboratory practice. However, due to the different molecular geometries of the nucleotides, a single inconsistency between the two strands will make binding between them less energetically favourable. Measuring the effects of base incompatibility by quantifying the temperature at which two strands anneal can provide information as to the similarity in base sequence between the two strands being annealed. The complexes may be dissociated by thermal denaturation, also referred to as melting. Melting point of nucleic acid sequences are determined chemically using UV spectrophotometer method or calorimetric method [Privalov 2015]

5.3.2.2. Thermodynamics of the two-state model

DNA/RNA denaturation i.e. breaking up of the hydrogen bond and unwinding of the double/folded strand to form the single stranded nucleic acid molecule and, renaturation which is joining together of the single strands/folding of the single strand onto itself are both are temperature dependant. [Frazen 2011]. For the double stranded DNA sequence, it has been established chemically that, in a reversible manner,

$$[AB] \leftrightarrow [A] + [B] \qquad (5.1)$$

Where $[AB]$ indicate the concentration of the double stranded nucleic acid molecule and $[A]$ and $[B]$ indicate the concentrations of the single strands. At T_m, half of the total number of nucleic acid molecules are double stranded and half of them are single stranded. The equilibrium constant of the reaction, K can be expressed as,

$$K = \frac{[A] \times [B]}{[AB]} \qquad (5.2)$$

According to the Van´t Hoff equation, [ISU Lecture Notes], [MIT Open Course Ware] the relation between free energy, ΔG, and K is

$$\Delta G° = -RT \ln K \qquad (5.3)$$

Where,

- R is the universal gas constant (R = 1.98×10-3 kcal/mol-deg, or R = 8.3 ×10-3 kJ/mol-deg).
- T is the temperature in Kelvin.

- $\Delta G°$, measured in kcal/mol, represents the change in Gibbs free energy. (Gibbs free energy is referred to as Helmholtz free energy too, with regard to nucleic acids, though they have different implications in the thermodynamics of gases).

  - $\Delta G = 0$ indicates equilibrium,
  - $\Delta G > 0$ indicates an unfavourable process and
  - $\Delta G < 0$ indicates a favourable process

$\Delta G°$ is called the standard Gibbs free energy, where the naught specifies a standard set of reaction conditions that include **constant pressure** (almost always 1 atm for biochemical reactions), a given **temperature**, and a set of **standard-state concentrations**. The temperature used in calculating $\Delta G°$ is that for which K for the reaction was measured.

The standard-state concentrations of reactants and products are assumed to be 1 M unless different values are explicitly specified.

At $T = T_m$, [A], [B] and [AB] are equal, and will have a value $\frac{[AB_{initial}]}{2}$, where $[AB_{initial}]$ represents the initial concentration of the double stranded nucleic acid molecule.

The free energy minimization methods aims at keeping change in Gibbs free energy minimum. This minimum value of $\Delta G°$ is termed MFE or minimum free energy.

The expression for $\Delta G°$, in thermodynamics is as follows,

$$\Delta G°_{total} = \Delta H°_{total} - T.\Delta S°_{total} \qquad (5.4)$$

Where $\Delta H°$ *and* $\Delta S°$ are the enthalpy and entropy components of the system considered. It follows directly from the equation in thermodynamics,

$$G = H - TS \qquad (5.5)$$

Where, G is Gibbs Free Energy, H is the enthalpy (potential to perform work), T the absolute temperature (in Kelvin) and S the entropy (measure of disorder). In the case of a nucleotide sequence, the enthalpy is contributed by base-pairs and entropy by the disorder of being unpaired ie “disorder in unpaired regions”, the difference in free energy, which can be notated as ΔG. The change ΔG of the free energy in a chemical process, such as nucleic acid folding, determines the direction of the process.

5.3.3. Length of the nucleotide sequence

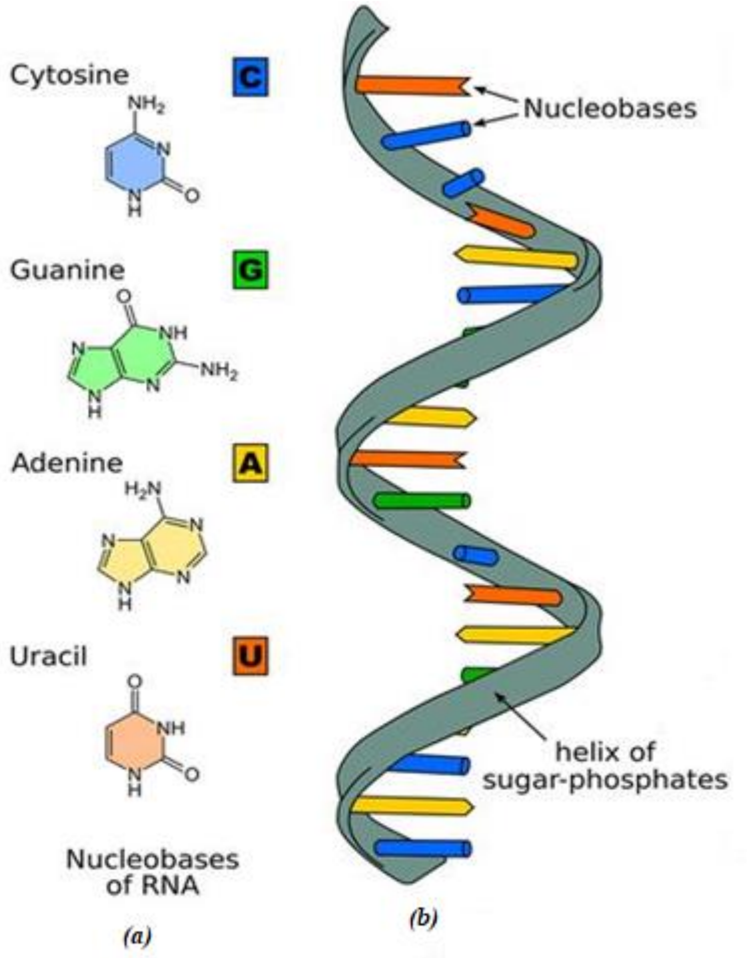


Figure 5.2. (a)Chemical and structure of RNA bases and the (b)RNA single strand

The difference in nucleotides between the DNA and the RNA is that, RNA has Uracil – U, instead of thymine. RNA sequences are made up of four nucleotide bases viz. Adenine, Uracil, Cytosine, Guanine attached to the sugar-phosphate back bone. The four letters representing the RNA sequence will be A, U, C, and G. A Figure depicting the molecular structure of RNA is shown in Figure 5.2 (a). As already seen, the RNA are single stranded molecules unlike DNA, see Figure 5.2. (b). However, they too form hydrogen bonds between the complimentary base pairs, A-U, C-G by folding over which results in their secondary structures [Watson 2007].

Every sequence has two distinct ends, the 5' end and the 3′ end. The RNA molecules too are read from the 5' end to the 3' end. 3' and 5' indicate the carbon atom number in the sugar back-bone of the nucleotide sequence. The 3' end has the hydroxyl group attached to it whereas the 5' end has the phosphate group attached to it. Therefore, each RNA single strand is mathematically represented by a character string, which, by convention specifies the 5′ to 3′ direction when read from left to right. Length of the RNA sequence obviously means the number of nucleotide bases in the sequence, when it is in the uncoiled form. That is the number of bases in the sequence in its primary structure when read from the 5' end to the 3' end. These strings, comprising of the four letters of the alphabet however need to be represented as numerical sequences in order to perform digital signal processing on them.

5.3.4. Spectral coefficient matrix of the nucleotide sequence

To apply signal processing techniques to nucleotide sequences, these sequences comprising of letters of the alphabet have to be converted into sequences of numbers. Once that is done, we can perform signal processing operations such as Fourier transformation [Tiwari 1997], [Anastassiou 2001], digital filtering [Vaidyanathan 2002.], wavelet transformations [George 2010], and Markov modelling [Durbin 1998] etc.

5.3.4.1. Mathematical representation of biomolecular sequences.

A number of techniques have evolved in the last decade [Akhtar 2008], [Kwan 2009], for the mapping of nucleotide sequences, besides the classical ones [Voss 1992]. A summary is given below.

a. Using complex conjugates.

Use of complex conjugates for representing the nucleotide string has been used widely [Anastassiou 2001,], [Cristea, 2002(1)], [Cristea 2002(2)], [Cristea 2005]. In a DNA sequence of length *N*, assume that we assign the numbers *a*, *t*, *c*, *g* to the characters *A*, *T*, *C*, *G*, respectively. A proper choice of the numbers *a*, *t*, *c* and *g* can provide potentially useful properties to the numerical sequence *x*[*n*]. One such example is, we could choose complex conjugate pairs $t=a^*$ and $g=c^*$, then the complementary DNA strand is represented by

$$\bar{x}(n) = x^*[-n + N - 1], \qquad n = 0, 1, 2......., N-1 \qquad (5.6)$$

and, in this case, all palindromes will yield conjugate, symmetric numerical sequences which have interesting mathematical properties, including generalized linear phase. One such assignment (the simplest out of many possible ones) is the following:

$a = 1 + j, \quad t = 1 - j, \quad c = -1 - j, \quad g = -1 + j$ (5.7)

b. Using binary sequences.

Another popular method of numerical representation of DNA sequences is the use of binary indicator sequences to represent the nucleotide sequences. This is called the Voss mapping technique. This technique maps the nucleotides A, C, G, and T into four binary indicator sequences $x_A(n)$, $x_T(n)$, $x_C(n)$, and $x_G(n)$. Consequently it is a four dimensional mapping [Anastassiou 2001(1)], [George 2010], because each base in the DNA sequence is represented by a four dimensional vector composed of either '0' or '1'. For example in the indicator sequence $x_A(n)$ '1' indicates the presence of base A and '0' indicates its absence as shown in the following example.

Let the nucleotide sequence read from the 5' end to the 3' end be GACTGTTACG. Here, the length of the sequence is 10. The indicator binary sequences for A, T, C, G would be as shown below.

$x_A(n)$= [0 1 0 0 0 0 0 1 0 0] (5.8)
$x_T(n)$= [0 0 0 1 0 1 1 0 0 0] (5.9)
$x_C(n)$= [0 1 0 0 0 0 0 0 1 0] (5.10)
$x_G(n)$= [1 0 0 0 1 0 0 0 0 1] (5.11)

so that the sum of the individual indicator sequences yield 1. $x_A(n) + x_T(n) + x_C(n) + x_G(n) = 1$. If we assign the number *a* to the character 'A' the number *t* to the character 'T', the number *c* to the character 'C', and the number *g* to the character 'G', In general, *a, t, c* and *g* can be complex numbers. The numerical sequence resulting from a character string of length N can then be written as:

$x(n) = a.x_A(n) + t.x_T(n) + c.x_C(n) + g.x_G(n)$ (5.12)
where $n = 0,1,2,3 \ldots\ldots\ldots. N - 1$
in which $x_A(n)$, $x_T(n)$, $x_C(n)$, and $x_G(n)$ are the *binary indicator sequences*, which take the value of either 1 or 0 at location *n*, depending on whether the corresponding character exists or not, respectively, at location *n*. The four binary indicator sequences uniquely determine the character string corresponding to a

DNA segment. For each n, three of the four sequences take the value of 0 and one takes the value of 1. They are a redundant, "linearly dependent" set of sequences;

$$x_A(n) + x_T(n) + x_C(n) + x_G(n) = 1 \qquad (5.13)$$

Therefore three of these four binary sequences would be enough to uniquely determine the DNA character string (in fact, the minimum number of binary sequences of length N required to uniquely determine the character string is 2, because each of the four possible characters can be encoded using two bits). The proper choice of the numbers a, t, c and g for a DNA segment can provide potentially useful properties to the numerical sequence $x[n]$. Voss mapping method does not predefine any mathematical relationship among the bases, but only indicates the frequencies of the bases. This method is an efficient representation among fixed mapping methods for spectral analysis of DNA sequences.

c. Using Electron Ion Interaction Pseudo-potentials (EIIP).

This technique is detailed by Cosic [Cosic 1994] and also by A S Nair [Nair 2006]. The four binary indicator sequences are replaced by just one sequence which we call as 'EIIP indicator sequence'. The energy of delocalized electrons in amino acids and nucleotides has been calculated as the Electron-ion interaction pseudopotential (EIIP). The EIIP values of amino acids have been used in Resonant Recognition Models (RRM) to substitute for the corresponding amino acids in protein sequences, whose Discrete Fourier Transforms are taken to extract the information contents. If we substitute the EIIP values for A, G, C & T in a DNA string x[n], we get a numerical sequence which represents the distribution of the free electrons' energies along the DNA sequence. This sequence is named as the EIIP indicator sequence, $x(n)$. EIIP values are 0.1260, 0.1335, 0.0806, and 0.1340 for A, T, G and C respectively.

d. Using integers.

Nucleotide sequences are to be converted into numerical sequences. The RNA sequence is to be read from the 5' end to the 3' end, as per conventional norms. MATLAB (R2016a) release notes describe a method of representing nucleotides with integers. It makes use of the integers 1 to 4. Mapping is done to 3 for Guanine and to 4 for Thymine/Uracil, Adenine 1 and Cytosine 2. A similar method is used here.

In order to make it conducive for digital signal processing, the sequences of letters from the four-character alphabet were first converted into numerical sequences. The binary indicator sequence representation was used here $u_a[n]$, $u_u[n]$, $u_c[n]$, $u_g[n]$ are the binary indicator sequences corresponding to A, U, C G which take on a value of 0 or 1 at location n, depending on whether the corresponding character exists or not at n.

$$u_a[n] + u_u[n] + u_c[n] + u_g[n] = 1 \qquad (5.13)$$

N is the sequence length, NTL.

The numerical sequence resulting from a character string of length N can be written as:

$$x[\mathrm{n}] = \mathrm{A}au_a[n] + \mathrm{U}uu_u[\mathrm{n}] + \mathrm{C}cu_c[\mathrm{n}] + \mathrm{G}gu_g[\mathrm{n}] \qquad (5.14)$$

n = 0, 1, 2, 3....... (*N-1*) and $a = 1 + j,\ u = 1 - j,\ c = -1 - j,\ g = -1 + j,$ following the convention of complex representation of bases [Cristea 2002] where purines and pyrimidines are represented by numbers that are complex conjugates. The multipliers A, U, C, G are taken as 1, 2, 3 and 4 respectively.

Frequency domain analysis of nucleotide sequences, has already been recognized as an important tool in bioinformatics by many authors including ones who could be called authorities in the field. [Vaidyanathan 2004], [Anastassiou 2001]. Some of the popular DSP tools used in frequency domain analysis are Discrete Fourier Transform (DFT), Short time Fourier Transform (STFT), Discrete Wavelet Transform (DWT) etc. They are used in spectral component sorting, comparison/correlation, removal or enhancement of certain regions of the spectrum by filtering. The technique used here is the Discrete Fourier Transform, the DFT.

5.3.4.2. Spectrum of the nucleotide sequence

Discrete Fourier Transform (DFT) is a very effective and simple tool which can be very conveniently used to analyse the frequency domain properties of signals [Proakis 2006], [Oppenheim 2009]. In digital signal processing, the function is any quantity or signal that varies over time, such as the pressure of a sound wave, a radio signal, or daily temperature readings, sampled over a finite time interval (often defined by a window function). The DFT is also used to efficiently solve partial differential equations, and to perform other operations such as convolutions or multiplying large integers. The DFT differs from the discrete-time Fourier transform (DTFT) in that its input and output sequences are both finite. It is therefore said to be the Fourier analysis of finite-domain (or periodic) discrete-time functions.

Mathematically, the DFT converts a finite list of equally spaced samples of a function into the list of coefficients of a finite combination of complex sinusoids, ordered by their frequencies, that has those same sample values. It can be said to convert the sampled function from its original domain (often time or position along a line) to the frequency domain. Being expansion of sinusoids, DFT provides the spectrum within the finite frequency bounds, **-π** to +**π** radians/sec or 0 to 2**π** radians/sec, ie. $-\frac{1}{2}\ to + \frac{1}{2}$ hertz, frequencies being commonly expressed in radians/second.

The input samples can be complex numbers in which case, the output coefficients are complex as well. The frequencies of the output sinusoids are integer multiples of a fundamental frequency, whose corresponding period is the length of the sampling interval. The combination of sinusoids obtained through the DFT is therefore periodic with that same period. The Discrete Fourier Transform (DFT) of a sequence $x[n]$, of length *N*, is itself another sequence $X[k]$, of the same length *N* [Proakis 2007] , [Oppenheim 2009],

$$X\ (k) = \sum_{n=0}^{N-1} x(n) e^{-(jk2n\pi)/N} \tag{5.15}$$

Where *k* = 0,1,2,3.... *N-1*

The DFT represented mathematically in equation 5.15 is a ‘k point’ DFT.

The sequence *X*[*k*] provides a measure of the frequency content at “frequency” *k*, which corresponds to an underlying “period” of *N/k* samples.

The popular algorithm for finding the Discrete Fourier Transform the Fast Fourier Transform, abbreviated FFT, was used here. FFT algorithm was designed on the logic of quickness of automation rather than simplicity and was formulated by Cooley and Tukey, 1964, it is also called the ‘prime-factor algorithm’. It works on the concept of decimating the computation of an ‘N’ point DFT into two N/2 point *N* DFTs in the case of radix 2 computation, or into three N/3 point DFTs in the case of radix 3 computation. The two N/2 point FFTs are further divided into four N/4 point DFTs in the former or into nine N/9 point DFTs in the latter. Decimation in this manner reduces the number of complex multiplications and additions required in the computation. The time taken to evaluate a DFT on a computer depends principally on the number of multiplications involved. DFT needs N^2 multiplications. Radix 2 FFT only needs $N\log_2(N)$

Here, $x(n)$ is the nucleotide sequence considered. As mentioned earlier, short ncRNA are in the range of 20 – 24 nucleotides commonly, though some of the sequences considered here are slightly longer. The expression for DFT is as already seen in equation 5.15. The evaluated DFT has two parts, the real part or the magnitude part, and the imaginary part or the phase or the argument part. $|X(k)|$ and $Arg(X(k))$ respectively. But MATLAB, which is the platform used, plots only the real part or the magnitude of the transform against the position in the nucleotide sequence. That is, from the DFT computed, only the magnitude of spectral coefficients are plotted. A sample spectrum is shown in the Figure 5.3 below. The spectrum is of miRNA of Mus musculus with the GenBank id HW523334.1.

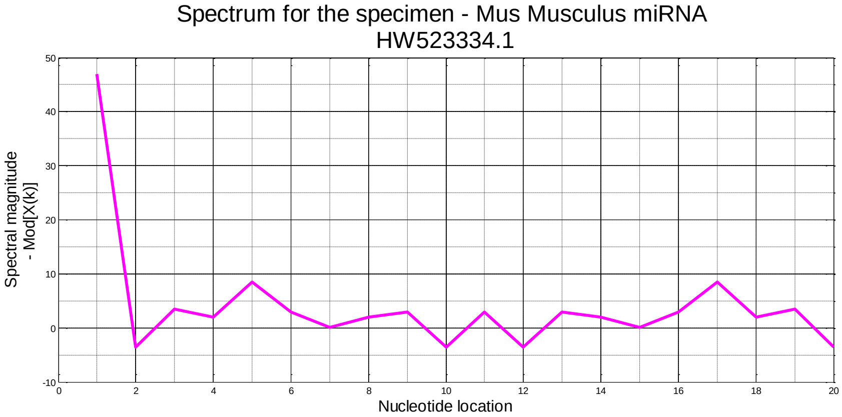


Figure 5.3. Plot of DFT spectrum of Mus Musculus miRNA – HW523334.1

5.4. Exploring the relationship between MFE and signal parameters of the ncRNA sequences.

In this chapter, it is tried to relate the Minimum Free Energy of the sequences analysed, with the signal properties of the sequence. The sequences are grouped organism-wise and also based on the class of the ncRNA for analysis.

A linear relationship is found to exist between MFE and the two signal parameters considered here, viz 1) length of the sequence and 2) the spectral coefficients of the sequence with the help of simple linear regression analysis.

A small sample 194 non coding RNA sequences were downloaded from the public database viz., the NCBI GenBank. The MFEs of these specimens were computed based on the thermodynamic nearest neighbour approach, the unit being kcal/mol. It was observed that the ***MFE varies linearly as the length of the sequence; MFE varies linearly as the standard deviation of the spectral coefficient matrix of the sequences.*** The relationship was found to be true for all the sequences studied. Simple linear regression analysis was done taking the length of the specimen and the MFE as the independent and the dependent variable respectively and taking MFE as the dependent variable and the standard deviation of the spectral coefficient matrix as the independent variable in the second case. Each one of the specimen were taken individually to study the regression.

The following Table, Table 5.1 shows a tabulation of the values of the parameters studied here, for the same sample space of 194 sequences. The spectrum of the nucleotide sequence was found out using DFT as explained in section 5.2.3. The magnitude of spectral coefficients were picked out and their standard deviation (SD) was evaluated and this parameter is called SD_DFT henceforth in this Chapter.

Table 5.1. MFE, sequence length and standard deviation of spectral coefficients of the noncoding RNA sequences

MUS MUSCULUS miRNA				
Sl. No.	**Specimen**	**NT length**	**Std. Dev. Of spectral coefficients**	**Calculated MFE (kcal/mol)**
1	FV523919.1	24	12.8807	-3.2
2	FV523920.1	22	12.4097	0
3	FV523921.1	24	14.6614	-3.3
4	FV523922.1	22	11.6701	-4.4
5	FV523923.1	22	13.9215	-4.2
6	FW342845.1	24	12.8807	-3.2
7	FW342846.1	22	12.4097	0
8	FW342847.1	24	14.6614	-3.3
9	FW342848.1	22	11.6701	-4.4
10	FW342849.1	22	13.9215	-4.2

11	FV524066.1	23	14.0938	-2.5
12	FV524067.1	21	14.1986	-3.9
13	FV524068.1	20	11.8766	0
14	FV524069.1	22	10.6369	0
15	FV524070.1	23	11.4773	-1.4
16	FW393855.1	23	14.0938	-2.5
17	FW393856.1	21	14.1986	-3.9
18	FW393857.1	20	11.8766	0
19	FW393858.1	22	10.6369	0
20	FW393859.1	23	11.4773	-1.4
21	HD065369.1	23	13.9447	-1
22	HD065370.1	22	11.3053	-0.5
23	JA368631.1	60	21.0327	-21.5
24	JC031747.1	61	20.2666	-5.4
25	JC031755.1	83	25.8662	-39
26	JC031756.1	83	25.8662	-39
27	JC031760.1	24	11.6917	0
28	JC031761.1	32	13.4404	-0.7
29	JC031762.1	31	13.8263	-6.1
30	JC105046.1	73	24.9098	-44
31	JC258548.1	21	13.0019	-6
32	JC258549.1	21	12.4238	0
33	JC258561.1-	73	23.0275	-33.1
34	JC428339.1	21	12.4238	0
35	JC428351.1	73	23.0275	-33.1
36	NR_039546	75	21.98863343	-16.5
MUS MUSCULUS siRNA				
37	HW523334.1	20	11.2858	-2
38	HW523335.1	21	13.7859	-4.2
39	HW523336.1	21	11.6383	-0.6
40	HW523337.1	21	12.339	0
41	HW523338.1	24	15.1514	-0.1
42	HW523339.1	21	11.3644	-0.9
43	HW523340.1	21	11.3644	-0.9
44	HW523341.1	23	11.8801	-3
45	HW504921.1	21	11.3644	-0.9
46	HW504922.1	23	11.8801	-3
47	HW504923.1	21	11.6383	-0.6

48	HW504924.1	21	12.339	0
49	DD346880.1	23	13.014	-2.7
50	DL076424.1	21	11.1781	0
51	DL076425.1	21	12.7574	-0.2
52	HM596744.1	24	12.7177	-2.3
53	HW040442.1	23	14.0567	-0.8
54	DL076423.1	21	9.4472	0
55	DL076422.1	21	13.9374	-3.6
56	DL076421.1	21	12.339	-0.3
57	DL076420.1	21	10.5995	0
58	DL076419.1	21	11.0838	0
59	DL076418.1	21	13.4387	-1.2
60	DL076417	21	12.1244	-0.2
61	DL076416.1	21	13.4778	-2.7
62	DL076415.1	21	11.9059	-0.7
63	DL076414.1	21	13.7859	-4.1
64	DL076413.1	21	13.9374	-7.3
65	DL076429.1	33	13.6622	-1.2
66	DL076428.1	35	16.3914	-7.1
MUS MUSCULUS snoRNA				
67	DQ267101.1	72	23.7665	-23.8
68	AF357362.1	98	27.2873	-31.6
69	AF357368.1	48	19.8548	-12.2
70	AF357369.1	61	19.9887	-9.6
71	AF357370.1	45	17.0427	-4.9
72	AF357371.1	65	20.9948	-4.2
73	AF357372.1	62	21.7866	-13.4
74	AF357373.1	63	20.0089	-9.8
75	AF357374.1	61	20.0649	-6.8
76	AF357375.1	59	20.222	-13.4
77	AF357376.1	58	20.1242	-8.2
78	AF357377.1	68	23.302	-14.3
79	AF357378.1	63	20.436	-9
80	AJ278763.1	66	22.8231	-20.2
81	DQ267100.1	71	25.1981	-17.6
82	DQ267101.1	72	23.7665	-23.8
83	DQ267102.1	71	24.996	-24.1
84	NR_028129.1	94	24.3787	-23.6

85	NR_028433.2	71	24.9554	-17.5
86	NR_028434.1	30	13.4933	0
87	NR_046302.1	69	23.3213	-10.6
88	NR_046303.1	67	21.8199	-13.1
89	NR_046304.1	71	25.1981	-17.6
90	NR_046305.1	72	23.7665	-23.8
91	NR_046306.1	71	24.996	-24.1
MUS MUSCULUS snRNA				
92	M34036.1	54	19.2579	-14.8
93	X94291.1	200	41.3224	-77.9
94	X07183.1	63	22.3592	-18.1
95	X04239.2	138	32.273	-49.9
96	NR_028276.1	87	26.401	-21.9
97	NR_024201.3	62	21.8797	-18.6
98	NR_024200.3	165	37.0448	-68.4
99	NR_004432.2	150	35.0311	-58.8
100	NR_004414.1	187	37.5973	-66.8
101	NR_004413.2	166	37.2877	-70.8
102	NR_004411.3	164	36.5533	-66.6
103	M34036.1	54	19.2579	-14.8
104	HQ148158.1	88	25.4233	-18.2
105	FM991919.1	132	30.0976	-50.1
106	FM991918.1	97	29.2545	-34.8
107	FM991916.1	186	37.3699	-66.2
108	FM991912.1	169	36.7422	-75.8
109	FM991908.1	214	41.9191	-89.4
110	FM991907.1	115	28.3547	-25.5
111	BK005202.1	134	33.6968	-50.9
112	AB021173.1	29	15.0606	-4.7
SUS SCROFA miRNA				
113	AM777927.1	20	12.5237	-0.8
114	AM777928.1	21	12.9615	-1.7
115	AM777929.1	21	13.8996	-3.4
116	AM777930.1	20	12.3969	-5.6
117	AM777931.1	21	13.2023	-0.4
118	AM777932.1	21	13.2023	-0.4
119	AM777933.1	23	12.3546	0
120	AM777934.1	23	12.3546	0

121	JN646111.1	54	20.2634	-13.5
122	JN646112.1	20	11.4248	-5.6
123	JN646113.1	17	12.4122	-2
124	JX185552.1	20	11.4248	0
125	JX185553.1	20	13.2982	-2.7
126	JX185554.1	17	11.5244	-1.50
127	JX185555.1	20	11.7429	-0.7
128	JX185556.1	22	12.619	0
129	JX185557.1	12	9.686	-0.7
130	JX185558.1	21	10.3995	-0.7
131	JX185559.1	18	12.044	-0.2
132	JX185560.1	20	11.1921	-0.7
133	JX185561.1	21	12.4238	0
134	JX185562.1	21	10.7471	0
135	JX185563.1	20	12.6491	-5.6
136	NR_031532.1	85	24.8855	-40.2
137	NR_038548.1	80	26.6054	-32.2
SUS SCROFA rRNA				
138	AB117609.1	565	66.3535	-226.1
139	AB117610.1	380	55.5348	-138
140	AF080393.1	218	40.3792	-73.2
141	AF329851.1	83	26.8644	-30.7
142	AJ583551.1	404	49.2748	-85.3
143	AJ849443.2	440	51.3614	-99.7
144	AM158315.1	715	65.0763	-171.4
145	GQ926971.1	440	51.6533	-101.9
146	KC984217.1	207	34.6236	-46.6
147	KF908860.1	556	60.6681	-150.4
148	KF908861.1	324	43.1828	-62.7
149	KJ192659.1	392	48.8989	-91.1
150	KJ193217.1	523	58.3825	-141.60
151	KJ361825.1	532	56.3646	-108.00
152	KM520132.1	979	77.6412	-262.20
153	KM520133.1	974	77.3508	-250.00
154	KM520134.1	983	77.2468	-244.80
155	KM520135.1	983	77.1106	-244.40
156	KM520136.1	949	75.5779	-244.80
157	KM520137.1	968	76.6686	-241.80

158	KM520138.1	986	77.3438	-250.60
159	KM520139.1	981	77.4281	-247.40
160	KM520140.1	983	77.0197	-243.40
161	KM520141.1	966	76.6621	-241.80
162	KM520142.1	939	75.2664	-228.90
163	KM520143.1	981	77.4475	-246.80
164	KM520144.1	920	74.6395	-228.60
165	KM520145.1	919	74.5255	-227.50
166	KM520146.1	881	73.5877	-215.60
167	KM520147.1	967	77.1502	-257.80
168	KM520148.1	920	74.9407	-225.40
169	KM520149.1	980	77.4152	-251.2
SUS SCROFA snoRNA				
170	AJ240060.1	73	23.2029	-21.7
171	AJ543323.1	68	24.1784	-18.1
172	JN831366.1	132	34.7005	-48.8
173	JN899116.1	70	22.3639	-11.2
174	JN899117.1	75	21.8719	-13.9
175	JN899118.1	77	25.2243	-20.4
176	JN899119.1	64	22.873	-16.5
177	JN899120.1	76	24.2432	-18.9
178	JN899121.1	71	23.619	-11.8
179	JN899122.1	65	22.0794	-10.5
180	JN899123.1	66	20.6509	-16
181	JN899124.1	75	23.351	-13.7
182	JN899125.1	85	24.9666	-18.7
183	JN899126.1	75	22.8686	-14.20
184	JN899127.1	77	24.8193	-17.2
185	JN899128.1	72	22.4047	-12.5
186	JN899129.1	86	25.5854	-28.7
187	JN899130.1	70	22.6569	-7.9
188	JN899131.1	80	23.3845	-17
189	JN899133.1	82	23.0978	-12.4
190	JN899134.1	79	23.0593	-16.2
191	JN899135.1	112	30.2184	-38
192	JN899136.1	75	23.8449	-26.9
193	JN899138.1	69	23.5594	-16.1
194	JN899139.1	141	32.6891	-33.2

Next, we will see the background of analysis done on the data. The data was subjected to regression analysis as the parameters studied were found to have a prominent linear relationship.

5.4.1.Regression analysis

As already mentioned, the linear relationship observed between the MFE values of the RNA sequences and the lengths of the sequences themselves, was analysed with the help of regression. Regression is a generic term for all methods attempting to fit a model to observed data in order to *quantify the relationship* between two groups of variables. The fitted model may then be used either to merely *describe* the relationship between the two groups of variables, or to *predict* new values [Sykes 1993].

In statistics, regression analysis is a statistical process for estimating the relationships among variables [Chatterjee C 2012], [Rohatgi 2000]. It includes many techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. In all cases, the target variable(s) (dependent variable) is a function of the independent variable(s) called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.

In general terms, if the two data matrices involved in regression are usually denoted as *X* and *Y*, where *X* represents the independent variable and *Y*, the dependent variable. The purpose of regression is to build a model *Y = f(X)*. Such a model tries to explain, or predict, the variations in the Y-variable(s) from the variations in the X-variable(s). The link between X and Y is achieved through a common set of samples for which both X- and Y-values have been collected. The literature on regression analysis present different types of regression.

Authorities classify regression under different heads. Broadly we have non-linear regression and linear regression. Linear regression has been used here as a linear relationship was noticed between the parameters analysed. Here, in this chapter, we have used simple linear regression.

Simple Linear regression

This examines the linear relationship between a single predictor variable or dependant variable and one dependant or response variable. Simple linear regression is the most commonly used technique for determining how one variable of interest (the response variable) is affected by changes in another variable (the explanatory variable). The terms "response" and "explanatory" mean the same thing as "dependent" and "independent", but the former terminology is preferred because the "independent" variable may actually be interdependent with many other variables as well. [Douglas C Montgomery et.al. “Introduction to Linear Regression Analysis”, Wiley Student Edition, December 2006].

The predictor variable and the response variable are mathematically represented by the equation,

$$y = a + bx \qquad (5.16)$$

Where the analysis examines the relationship between response variable y and predictor variable x.

Simple linear regression can be used for three main purposes:

1. To describe the linear dependence of one variable on another
2. To predict values of one variable from values of another, for which more data are available
3. To correct for the linear dependence of one variable on another, in order to clarify other features of its variability.

In our case, the first purpose was made use of.

5.4.1.a. Error and minimisation of mean squared error (MMSE).

Here, the mathematical problem is a straightforward one: given a set of n points (x,y) on a scatter-plot, find the best-fit line, $\hat{y} = a + bx$ such that the sum of squared errors in Y, E, is given as,

$$E = (\, y - \,\hat{y})\,^2 \qquad (5.17)$$

is minimized.

Equation 4.16 an be re-written as

$$E = \sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(5.18)

Any line fitted through a cloud of data will deviate from each data point to greater or lesser degree. The vertical distance between a data point and the fitted line is termed a "residual". This distance is a measure of prediction error, in the sense that it is the discrepancy between the actual value of the response variable and the value predicted by the line.

Linear regression determines the best-fit line through a scatter plot of data, such that the sum of squared residuals is minimized; equivalently, it minimizes the error variance. The fit is "best" in precisely that sense: the sum of squared errors is as small as possible. That is why it is also termed "Ordinary Least Squares" regression [Kirchner 2001].

5.4.1.b. Intercept 'a' and Regression slope 'b'.

Minimisation of squared error means, error 'e ' has to be minimized with respect to 'a' and 'b', where, $\frac{\partial E}{\partial a}$=0, $\frac{\partial E}{\partial b}$ = 0 respectively.

$E = \sum[y - (a + bx)]^2$ (5.19)

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{n} -2\,(y_i - a - bx_i) = 2\left(na + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right) = 0$$

(5.20)

Equating , $\frac{\partial E}{\partial a}$ to 0 yields,

$$a = \bar{y} - b.\bar{x} \qquad (5.21)$$

Equation 4.18, says that the ***intercept a*** is such that the line should pass through the mean of x and y. That is, given the 'data cloud' of the scatter plot, the regression curve, the straight line here, would go through the centre of it.

$\bar{y}, \bar{x}$ represents the mean of *y* and *x* respectively.

To find ***b,*** the ***regression slope*** we equate the partial derivative $\frac{\partial E}{\partial b}$ to 0.

Equating , $\frac{\partial E}{\partial b}$ to 0 yields,

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{n} -2x_i(y_i - a - bx_i) = \sum_{i=1}^{n} -2\left(x_i y_i - ax_i - bx_i^2\right) = 0$$

(5.22)

Upon re-arranging and after making appropriate substitutions, we have,

$$b = \frac{\sum_{i=1}^{n}(x_i y_i - x_i\bar{y}) + \sum_{i=1}^{n}(\bar{x}\bar{y} - y_i\bar{x})}{\sum_{i=1}^{n}\left(x_i^2 - x_i\bar{x}\right) + \sum_{i=1}^{n}(\bar{x}^2 - x_i\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(5.23)

Which can be re-stated as, $b = \frac{cov(x,y)}{var(x)}$ (5.24)

The quantities that result from regression analysis are mathematically equivalent, although they appear different. Besides regression intercept a and slope b, a third parameter that is of primary importance is, the correlation coefficient *‘r’*, or the coefficient of determination $r^2 . r^2$ is ratio between the variance in *y* as indicated by the regression, and the total variance in *y*. As in the case with ‘b’, *r* can also expressed in different yet equivalent mathematical terms.

$$r^2 = \frac{var(\hat{y})}{var(y)} = \frac{b^2 var(x)}{var(y)} = \frac{\left(cov(x,y)\right)^2}{var(x)var(y)} = \frac{var(y) - var(y - \hat{y})}{var(y)}$$

$$= \frac{S_{xy}^2}{SS_x SS_y}$$

(5.25)

S_x and S_y represent standard deviations of x and y respectively.

$$r = b\frac{S_x}{S_y}$$

(5.26)

In regression estimation, the response variable $\bar{y}_i$ is estimated from the observed variable x_i. That is, $\widehat{y_i}$ is the variable that closely fits into the regression curve. Here, we have tried to relate the predictor variable, namely MFE with nucleotide length and standard deviation of spectral coefficients, individually, via simple linear regression and, together via multiple linear regression. The actual response variable y_i does, however, deviate from the ideal or the predicted value $\hat{y}_i$. The deviations of the response variable y_i from its estimate $\widehat{y_i}$, is called the ‘residue’ or 'residual'. The estimation of $\widehat{y_l}$, is done such that the mean squared error is minimum, i.e. the MMSE estimation.

In our case, two regression relationships were explored, as already mentioned,

- MFE vs NTL
- MFE vs SD_DFT

The response variable being MFE in both cases and the predictor variable being NTL in the first case and SD_DFT in the second case. The simple linear regression analysis was done for each separately. In each case, a, b, and R^2 were found out. A simple linear regression equation of the form, $y = bx + a$ was found, on groups of data, of the two organisms studied. The results of the analysis are given in the section 5.5.

5.4.1.c. Coefficient of determination - R^2.

The coefficient of determination denoted often as R^2 or r^2, indicates how well minimum mean squared error based equation $y = b_0 + b_1 x$ describes the data set (*x*,*y*). In this work, it is an index of how well the set of data values relating to the parameters studied, is described by the linear relationship. In our case, how well MFE is regressed to NTL and SD_DFT individually. R^2 is computed as,

$$R^2 = \frac{SS_{yy} - E}{SS_{yy}} = \frac{SS_{yy}}{SS_{yy}} - \frac{E}{SS_{yy}} = 1 - \frac{E}{SS_{yy}}$$

(5.27)

SS_{yy} measures the deviations of the observations from their mean, and can be expressed mathematically as,

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

(5.28)

E is the sum of squared errors given by equation 4.17,

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(5.29)

Like the standard error, the coefficient of determination gives an indication of how well a linear-regression model serves as an estimator of values for the dependent variable. It works by measuring the fraction of total variation in the dependent variable that can be explained by variation in the independent variable.

For a simple linear regression with one independent variable, the simple method for computing the coefficient of determination is squaring the correlation coefficient between the dependent and independent variables. Since the correlation coefficient is given by r, the coefficient of determination is popularly known as "R^2, or R-squared". For example, if the correlation coefficient is 0.76, the R-squared is $(0.76)^2 = 0.578$. R-squared terms are usually expressed as percentages; thus 0.578 would be 57.8%. R^2 takes on a value between 0 and 1. The rule of thumb is, the higher the value of R^2, more useful the model would be. In our context, it means that, higher the value of the R^2, the more linear are the relationships explored here, between NTL and MFE, and between SD_DFT and MFE . The values of the coefficient of determination, computed in regression analysis for several sets of the ncRNA analysed and for the entire sample space is given in the results sub-section of this chapter.

The very expression of the equation for simple linear regression, $\hat{y} = a + bx$ implies that an estimate of y, $\hat{y}$, is made from the knowledge of x, which implies that, better the estimate, better the linear relationship, as then the error would be smaller, and as is obvious, smaller value of E would increase the value of R^2. The simple linear regression plots for all the specimen given in the results sub-section of this chapter. This relationship was found to be true for all the specimen analysed.

The ncRNA from mus musculus and sus scrofa were grouped into their respective classes, viz. miRNA, snRNA, snoRNA, siRNA, rRNA and equations relating the MFE with sequence length and spectral coefficient matrices were found out for each class. The regression analysis which clearly points to a linear relationship between MFE and spectral coefficients was further ratified through polynomial curve fitting using MATLAB. Polynomial curve fitting can simply be stated as the process of formulating a mathematical equation or constructing a curve that subject to certain constraints, fits best into a set of data points. Here, curve fitting was carried out just to ratify the relationship between MFE and spectral coefficients that was arrived at via regression analysis. The complete result of the simple linear regression analysis for both the two sets, including the values of the regression coefficients are given in the next section.

5.5. Representation of the relationship between MFE and signal parameters

In this work, as already mentioned, MFE was calculated for the ncRNA sequences with the thermodynamic nearest neighbour approach, making use of the Bioinformatics toolbox of MATLAB, and it has been related linearly to two signal parameters of the ncRNA sequence, viz. length of the nucleotide sequence and standard deviation of the DFT coefficients of the sequence. The analysis was done on a total of about 200 specimen of ncRNA sequences miRNA, rRNA, siRNA, snoRNA, snRNA belonging to the organisms, C elegans, and Sus scrofa. The specimen have been taken from the NCBI database, GenBank.

These results of simple linear regression have been tabulated for the groups of ncRNA considered individually. The value of the coefficient of determination, R^2 has also been computed in every case. The mathematical expressions and the values of regression coefficients and that of R^2 are given in tabular columns. All through, in the results, a point to be noted is that the MFE obtained making use of the thermodynamic nearest neighbour algorithm is negative, hence the slope of the regression curve is negative, and also the value of the coefficient of determination R^2 is negative.

The results are organized into two sections,

- *1. Linear relationship between NTL and MFE.*

- *2. Linear relationship between SD_DFT and MFE.*

The first section (section 5.5.1) presents, graphically and algebraically the linear relationship between the minimum free energy (MFE) and the length of the nucleotide sequences (NTL). The second section (section 5.5.2) gives the results of the regression analysis which shows the linear relationship between minimum free energy (MFE) and the standard deviation of the spectral coefficients (SD_DFT).

5.5.1.Linear relationship between MFE and Nucleotide length.

Graphical representation of the linear relationship between MFE and NTL is depicted in Figures 5.4 to 5.10. Figures 5.4 to 5.7 shows the MFE vs. NTL plot for Mus musculus miRNA, siRNA, snoRNA, snRNA respectively. Figures 5.8 to 5.10 show the MFE vs NTL relationship for sus scrofa miRNA, rRNA, snoRNA, respectively.

The MFE evaluated is negative, so, the regression lines have a negative slope, and also, the value of the coefficient of determination R^2 is negative. Table 5.2 gives the mathematical relationship depicted graphically in the Figures mentioned above. Column 2 of Tale 5.2 gives the name of the organism and the class of ncRNA which was analysed. Columns 3 and 4 give the values of the regression coefficients ‘b’ and ‘a’ respectively. Column 5gives the values of coefficient of determination R^2 and column 6 gives the regression equation i.e. the equation linking MFE with NTL for each f te class of ncRNA analysed.

5.5.1.1. Graphical representation of MFE-NTL relationship.

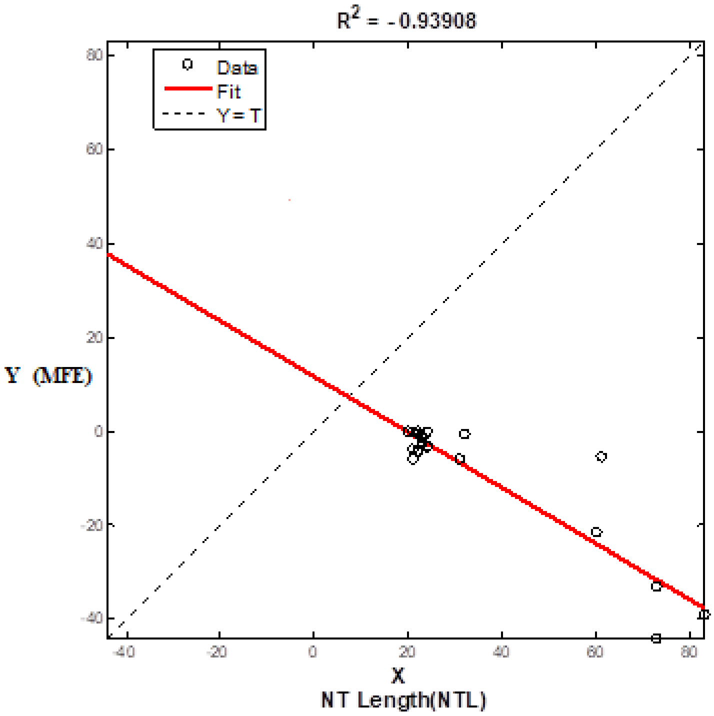


Figure 5.4. Single Linear Regression plot of Mus Musculus miRNA

The NTL-MFE relationship of for miRNA sequences of Mus musculus is shown in Figure 5.4. The equation linking MFE and NTL is $y = -0.59x$+38 and the value of the coefficient of determination, R^2 for this regression curve is -0.93908. The red line in Figure 5.4 indicates the line of fit for the given sample. As can be seen from this scatter plot, there are very few outliers from the line of fit. The value of coefficient of determination is indicative of this. Fewer the outliers, better the fit and higher the magnitude of R^2. Similar logic is to be applied while interpreting the other regression plots.

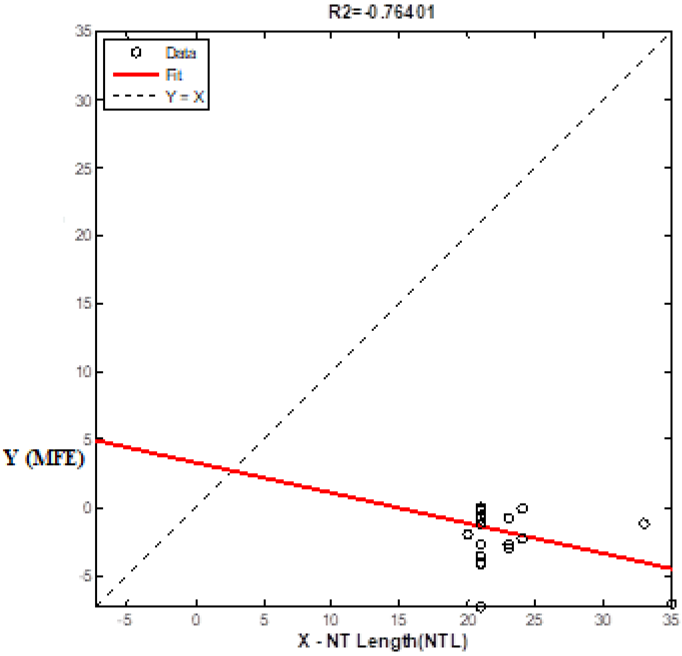


Figure 5.5. Single Linear Regression plot of Mus Musculus siRNA

In this regression plot, given in Figure 5.5 above, there are more outliers relative to the other MFE-NTL regression graphs given in this section And hence the coefficient of determination R^2 has a lower value relative (- 0.76401) to the R^2 values of the other regression graphs.

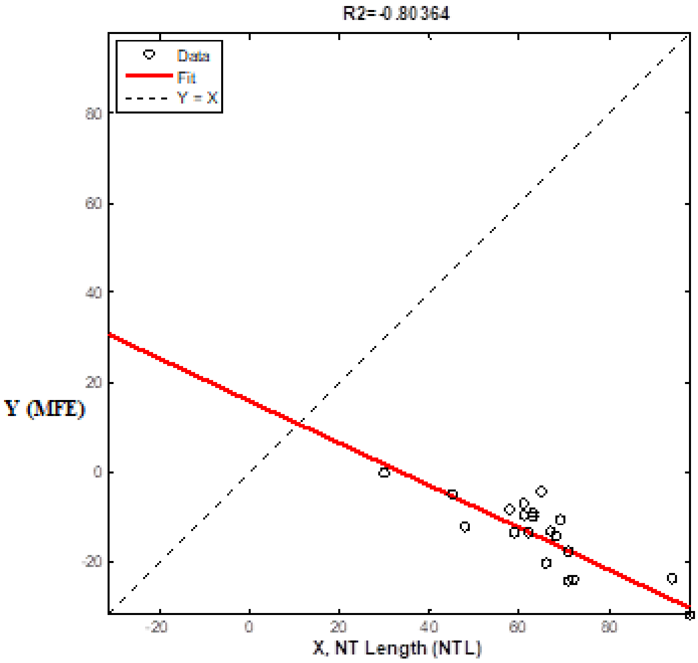


Figure 5.6. Single Linear Regression plot of Mus Musculus snoRNA

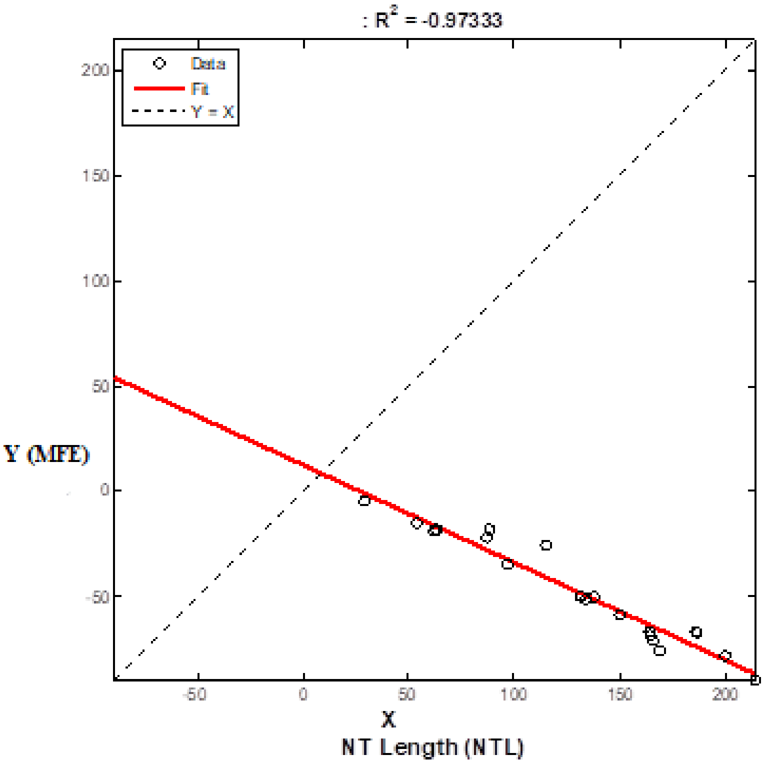


Figure 5.7. Single Linear Regression plot of Mus musculus snRNA

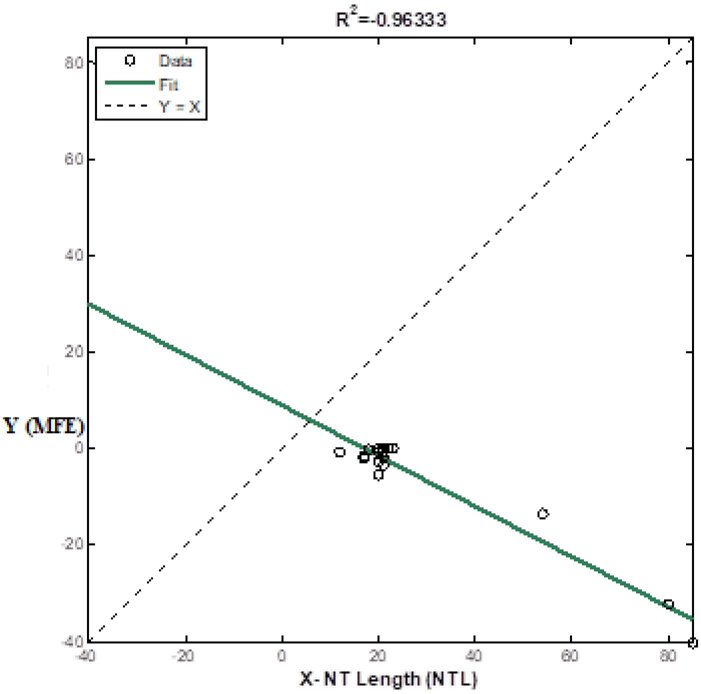


Figure 5.8. Single Linear Regression plot of Sus scrofa miRNA

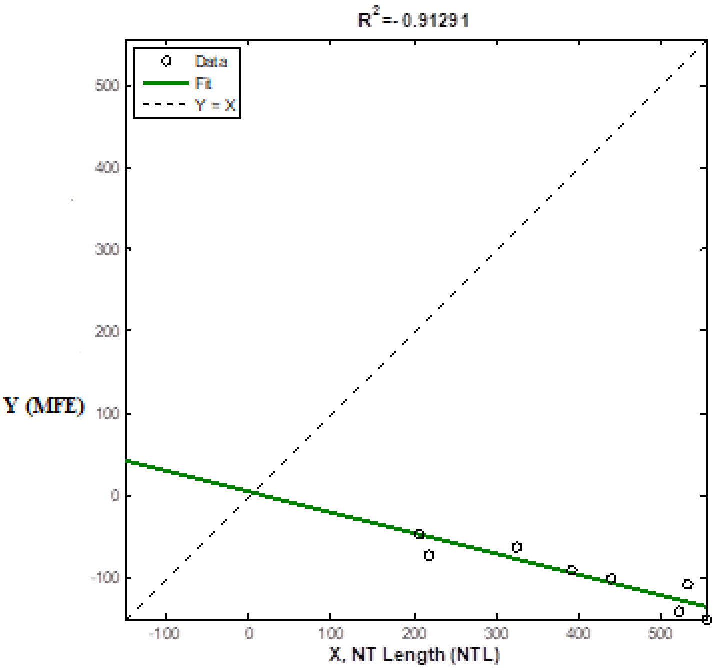


Figure 5.9. Single Linear Regression plot of Sus scrofa rRNA.

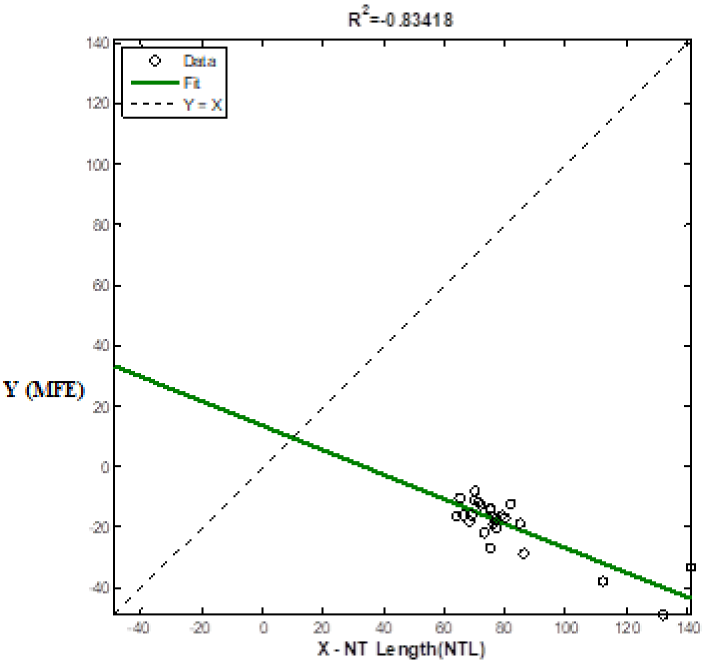


Figure 5.10. Single Linear Regression plot of Sus scrofa snoRNA

5.5.1.2. Mathematical expressions relating MFE and nucleotide length

Table 5.2. Regression coefficients, coefficient of determination and regression equations for MFE vs. NTL

Sl. No.	**Specimen**	**Regression coefficients**			**Mathematical relationship**
		b	a	$\boldsymbol{R^2}$	
1	Mus musculus miRNA	-0.59	12	- 0.93908	$y = -0.59x + 38$
2	Mus musculus siRNA	-0.22	3.3	- 0.76401	$y = -0.22x + 3.3$
3	Mus musculus snoRNA	-0.47	16	- 0.80364	$y = -0.47x + 30$
4	Mus musculus snRNA	-0.46	13	- 0.97333	$\mathrm{y} = -0.46x + 53$
5	Sus scrofa miRNA	-0.52	9.1	- 0.96333	$y = -0.52x + 31$
6	Sus scrofa rRNA	-0.25	-6.4	- 0.91291	$y = -0.25x + 40.6$
7	Sus scrofa snoRNA	-0.41	14	- 0.83418	$y = -0.41x + 34$

5.5.2.Linear relationship between MFE and spectral coefficients

The section gives the results of the regression analysis which shows the linear relationship between minimum free energy (MFE) and the standard deviation of the spectral coefficients (SD_DFT). As already mentioned, the analysis was done after separating the organisms and taking individual classes of ncRNA for each. Figures 5.11 to 5.14 given in the pages that follow show the graphical expression of the linear MFE vs SD_DFT relationship for miRNA, siRNA, snoRNA, snRNA sequences of mus musculus respectively. And Figures 5.15, 5.16, 5.17 is the graphical plot of the above mentioned linear relationship for miRNA, rRNA, snoRNA sequences of Sus scrofa studied here.

The specimen were grouped into their respective classes and equations linking MFE with the SD of the spectral coefficient matrices were found for each case. Table 5.3 gives the mathematical relationship between MFE and SD_DFT of specimen, grouped as shown. The value of the regression coefficients a, b and the coefficient of determination R^2 obtained, in each case is also given.

As mentioned earlier, the MFE evaluated is negative, so, the regression lines have a negative slope, and also, the value of the coefficient of determination R^2 is negative. Table 5.3 gives the mathematical relationship depicted graphically in the Figures mentioned above. Column 2 of Tale 5.3 gives the name of the organism and the class of ncRNA which was analysed. Columns 3 and 4 give the values of the regression coefficients ‘b’ and ‘a’ respectively. Column 5gives the values of coefficient of determination R^2 and column 6 gives the regression equation i.e. the equation linking MFE with SD_DFT for each of the classes of ncRNA analysed.

5.5.2.1. Graphical representation of MFE-SD_DFT relationship.

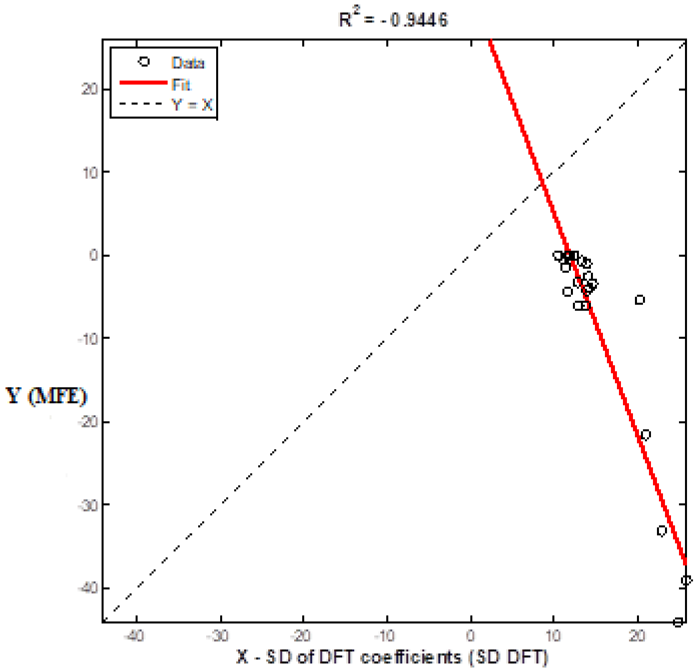


Figure 5.11. Single Linear Regression plot of Mus musculus miRNA

The NTL-SD_DFT relationship of for miRNA sequences of Mus musculus is shown in Figure 5.11. The equation linking MFE and NTL is $y = -2.7x + 72$ and the value of the coefficient of determination, R^2 for this regression curve is -0.9446. The red line in Figure 5.11 indicates the line of fit for the given sample. The value of coefficient of determination is indicative of the closeness to which the regression line represents the given sample of points . As can be seen from the scatter plot, there are very few outliers from the line of fit. Fewer the outliers, better the fit and higher the magnitude of R^2. Similar logic is to be applied while interpreting the other regression plots.

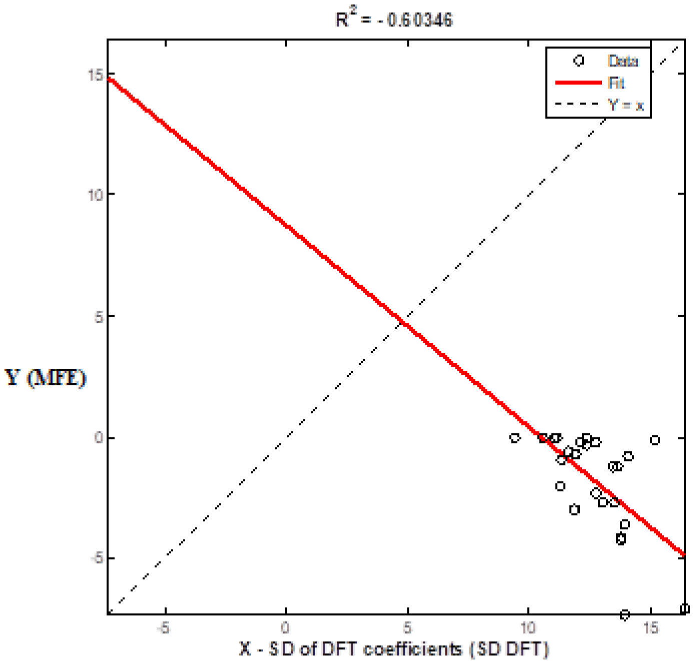


Figure 5.12. Single Linear Regression plot of Mus musculus siRNA

In this regression plot, given in Figure 5.12 above, relative to the other ones there are more outliers. And hence the coefficient of determination R^2 has a lower value (-0.60346) relative to the other R^2 values seen in this section.

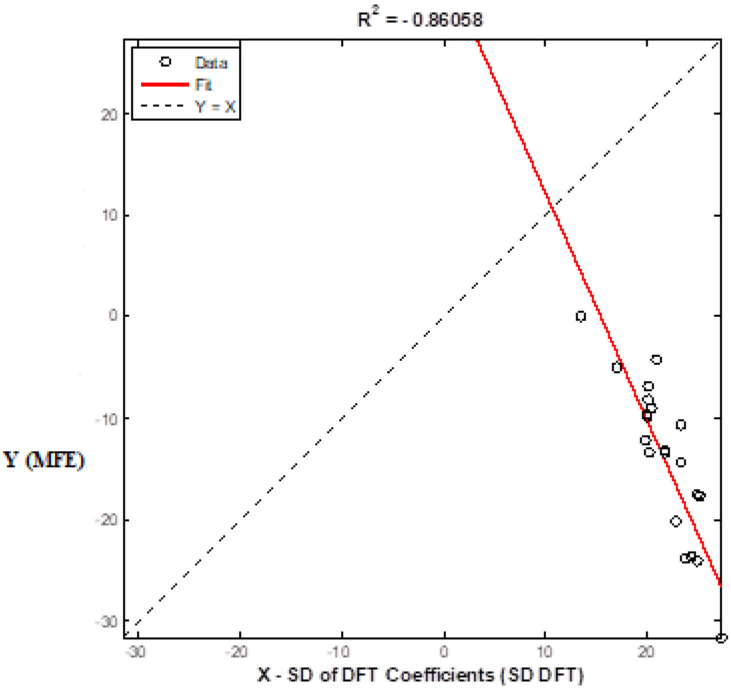


Figure 5.13. Single Linear Regression plot of Mus musculus snoRNA

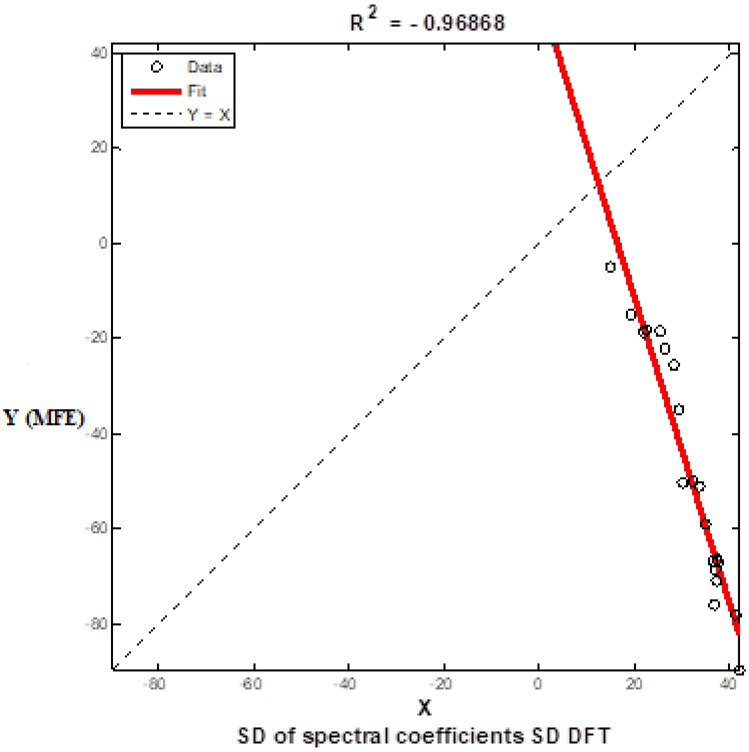


Figure 5.14. Single Linear Regression plot of Mus musculus snRNA

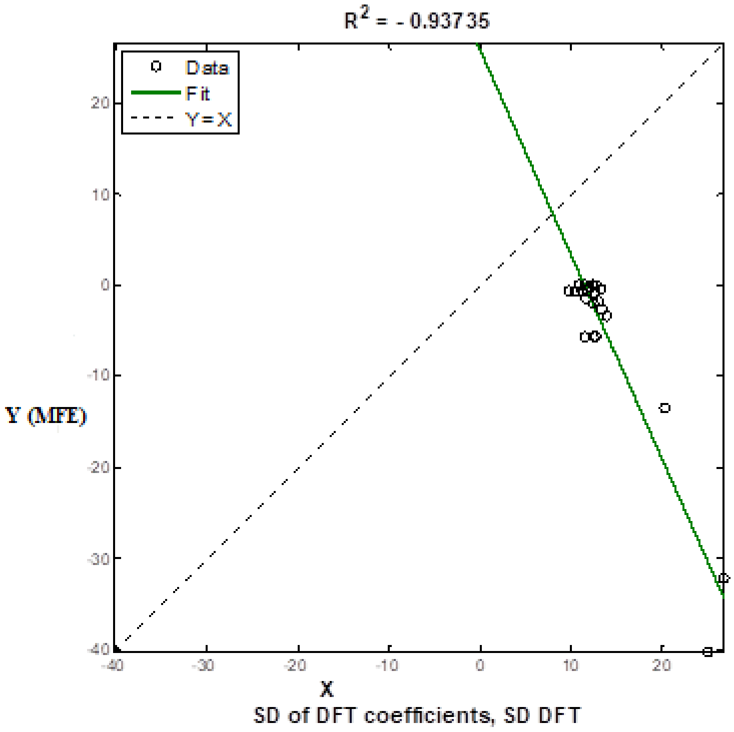


Figure 5.15. Single Linear Regression plot of Sus scrofa miRNA

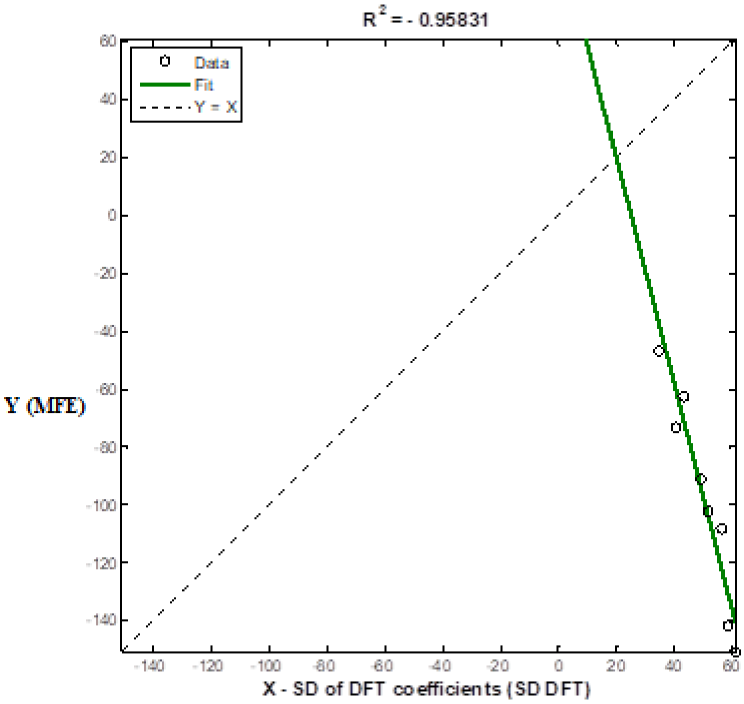


Figure 5.16. Single Linear Regression plot of Sus scrofa rRNA

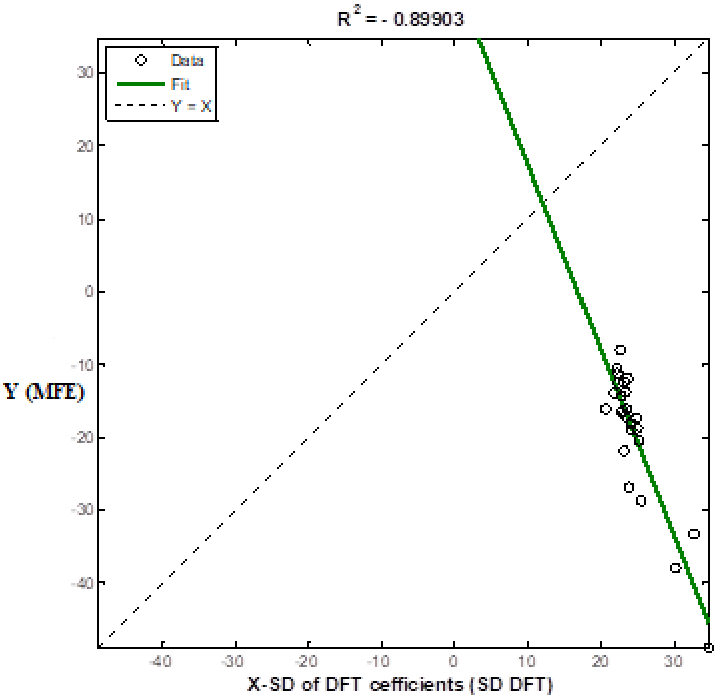


Figure 5.17. Single Linear Regression plot of Sus scrofa snoRNA

5.5.2.2. Regression equations R^2, *a, b for MFE vs. SD of the spectral coefficient matrix*

Table 5.3. Regression coefficients, coefficient of determination and regression equations for MFE vs. SD_DFT

Sl. No.	Specimen	Regression coefficients		R^2	Mathematical relationship (Simple linear regression Equation : $y = bx + a$)
		b	a		
1	Mus musculus miRNA	-2.7	32	-0.9446	$y = -2.7x + 72$
2	Mus musculus siRNA	-0.83	8.7	-0.60346	$y = -0.83x + 14$
3	Mus musculus snoRNA	-2.2	35	-0.86058	$\mathrm{y} = -2.2x + 85$
4	Mus musculus snRNA	-3.2	53	$-$ 0.96868	$\mathrm{y} = -3.2x + 93$
5	Sus scrofa miRNA	-2.3	26	-0.93735	$y = -2.3x + 66$
6	Sus scrofa rRNA	-5	142	-0.98161	$y = -5x + 192$
7	Sus scrofa snoRNA	-2.6	43	-0.89903	$y = -2.6x + 93$

5.6. Discussion

Although MFE and sequence length have been widely used in the study of functional RNA, a relationship between MFE and length or any other signal property of RNA sequences have not been found in literature. In the work presented in this chapter, the possibility of a relationship between MFE and signal parameters of ncRNA sequences was explored. The signal parameters studied were sequence length and the standard deviation of the spectral coefficient matrix of noncoding RNA sequences. It was observed with the help of simple linear regression analysis that a linear relationship exists between MFE and sequence length and MFE and the standard deviation of spectral coefficients of noncoding RNA sequences studied.

It is to be noted here, that all the calculated values of MFE are negative, as they are expected to be, hence the regression curve has a negative slope. Hence, all through, in the regression analysis, the value of the coefficient of determination, R^2, was obtained as negative, for the very reason that MFE value is negative

The magnitude of R^2 in simple linear regression takes on a value between 0 and +1, the higher the magnitude of R^2, ie the closer magnitude is to 1, the more useful the model would be. In our context, it means that, higher the magnitude of the R^2, the more linear are the relationships explored here, and better the model for representing the relationship between variables explored. Here, between NTL and MFE, and between SD_DFT and MFE . The values of R^2 are in the range 0.80364 to 0.98263 in the case on MFE vs NTL, except for Mus musculus siRNA which gave a magnitude of R^2 as 0.76401. The R^2 values indicate the correctness of the mathematical expressions. The details given in Table 5.3 are the results of regression analysis of MFE vs SD_DFT for the different classes on ncRNA taken from sus scrofa and mus musculus. The range of magnitudes of R^2 are from 0.98161 to 0.86058, except in the case of Mus musculus siRNA which gave a lower magnitude for R^2, 0.60346. The equations linking MFE with the spectral coefficients are given in the Table for the individual classes of ncRNA studied.

5.7. Conclusion

This part of the work aims to relate the thermodynamic entity, minimum free energy of noncoding RNA sequences, which decides their secondary structure and hence the function, with signal parameters of the ncRNA sequences. And as seen, a linear relationship was observed, between MFE and the sequence length in terms of nucleotides and also between MFE and the SD of spectral coefficients of the ncRNA sequence. Linear equations of the form $y = mx + c$ are found to link the above parameters. The high magnitudes of coefficients of determination R^2, obtained in the linear regression analysis show that the linear equations obtained amply represent the relationships between and MFE and spectral coefficient matrix.

MFE is an entity which needless to say, is a very relevant one as far as functional RNA sequences are concerned, as it decides the secondary structure and thereby its function. Structural RNA have been found to have lower folding than random RNA of the same dinucloetide frequency. Various instances where MFE values are used in identification of non coding RNA genes, used as a filter in picking functional RNA from ESTs and other genomic data [Washeitl 2005], [Clote 2003], [Warris 2014] have already been discussed. The relevance of the knowledge of MFE in RNA biology is very evident from this. This chapter brings to light the possibility of expressing MFE in terms of signal parameters of the sequence. *Transcriptome* refers to the set of all RNA molecules from protein coding (mRNA) to noncoding RNA, including rRNA, tRNA, lncRNA, pri-miRNA, and others. Transcriptome may apply to an entire organism or a specific cell type [Wang et.al. 2009]. The transcriptome structure does affect its function [Piao 2017] and therefore MFE does have a role in the functions of the transcriptome. The functions of noncoding RNA sequences have not been studied here, the aim of the study was to establish the relationship the signal properties of ncRNA sequences have to their minimum free energy.

Coding DNA has been studied widely with both computational and DSP methods. The noncoding region of the genome has been explored with computational and statistical tools. But DSP tools have not yet been made use of in the analysis of noncoding genomic sequences. DSP techniques are inherently simple and in most cases require lesser computational time. It is hoped that this relationship linking a bio-chemical property (MFE) to the signal parameters of ncRNA sequences could be an initial step to exploring non coding genomic sequences using DSP techniques.

From the results of the study which is recorded in this chapter, it is evident that MFE is linearly related to length as well as the standard deviation of the spectral coefficient matrix of noncoding RNA sequences analysed. This finding is made use of to arrive at a mathematical model for MFE from the signal properties of noncoding RNA sequences. This is discussed in chapter 6.

Chapter 6

Novel mathematical model for MFE

In this chapter, a novel mathematical model for MFE of noncoding RNA sequences is arrived from their signal properties. In Chapter5, the relationship between MFE and sequence length, MFE and standard deviation of spectral coefficients of ncRNA sequences of Mus musculus and Sus scrofa was presented. The relationships being linear in nature, the combined relationship between them needs to be analysed too. In this Chapter, multiple regression analysis is done with the response variable being MFE and the predictor variables being length of the nucleotide sequence and standard deviation of the spectral coefficients and the results are presented. The mathematical models are developed for MFE from sequence length and SD of the spectral coefficients. These models are made use of in evaluating MFE. This method proved to be one saving time and computational complexity as against the traditional computational algorithms for evaluating MFE. The correctness of the models developed is checked with standard webservers, RNAfold and RNAstructure.

Abstract

Noncoding RNA studies occupy the prime slot in genome research now after their functional potential was revealed. The function of ncRNA is decided by its secondary structure, and across organisms, the secondary structure is more conserved than the sequence itself. In this chapter, the optimal secondary structure or the minimum free energy (MFE) structure of non-coding RNA is found out based on the thermodynamic nearest neighbour model. MFE of over 2600 noncoding RNA sequences were analyzed with view of its signal properties. Mathematical models linking MFE to the signal properties were found out for each of the four classes of ncRNA analyzed. MFE values computed with the proposed models were in concordance with those obtained with the standard web servers. 95% of the sequences analyzed had deviation of MFE values within ± 15% relative to those obtained from standard web servers. It is hoped that this relationship between MFE and the signal properties of the sequence leads to new discoveries between the biological features of genomic sequences and their signal parameters. This in turn may bring up efficient Digital Signal Processing approach to study the noncoding genome, enabling better understanding of the ncRNA world.

6.1. Introduction

Computational methods are quite popular and rampantly used in molecular biology. However, over the past two decades the theory and methods of digital signal processing too have gained attention in molecular biology. Good amount of digital signal processing methods (DSP) has been employed to analyze DNA and proteins after the initial work in the turn of this century [Anastassiou 2001 (2)], [Anastassiou 2001 (2)], [Cristea 2002]. Nevertheless, there has not been much published work on DSP methods to analyze the non-coding region of the genome.

In this chapter, we develop a model for MFE (minimum free energy) of the secondary structure of noncoding RNA sequences from their signal parameters viz. length and the spectral coefficient matrix making use of multiple linear regression analysis [George 2016]. This model is made use of to evaluate MFE without employing the folding algorithm. The correctness of the model is checked using standard webservers (RNAfold and RNAstructure). To begin with, MFE of noncoding RNA sequences is found out using the thermodynamic nearest neighbour algorithm. Multiple linear regression analysis is done by considering MFE as the response variable sequence length and the standard deviation of the spectral coefficient matrix as the predictor variables [Chatterjee 2012], [Montgomery 2006] to arrive at the model.

The parameters, sequence length and MFE are key entities in ncRNA studies and have been used in their analysis from a very early time [Grüner 1996], [Galzitskaya 1998]. There have been studies which explore the influence of length and MFE on sequence stability [Pervouchine 2003], [Trotta 2014]. MFE has also been used as an index to study the relationship between entropy and structural properties of RNA sequences [Wolfsheimer 2010]. Washeitl et.al. describes a noncoding RNA gene finder which makes use of MFE *z* score computations, together with comparative genomic techniques. The mean and standard deviation of MFE of sequences are made use of here [Washietl 2005]. Clote et.al. describes a method of 'asymptotic z score' that sets asymptotic limits for mean and standard deviations of MFE per nucleotide of random RNA. They perform certain pre-computations that speed up z score computations for the entire genome using a sliding window scan. This method provides a filter, which can be used together with MFE computations and pattern matching to identify functional RNA genes in expressed sequence tags and genomic data. RNAs for which native state (the free energy structure) is functionally important were found to have lower folding energy, when compared to random RNAs having the same length and dinucleotide frequency [Clote 2005]. As MFE is a discerning factor, knowing its value would be useful in situations where it is needed to know quickly whether a given sequence is functional or a random RNA sequence.

MFE is a vital tool in identifying noncoding RNA genes. Lim et.al describe a technique for identifying miRNA genes where a moving window scan searches for stem-loop structures having at least 25 base-pairs and has a predicted MFE of -25 kcal/mol or less. A window which accommodates 21 nucleotides is passed over each conserved stem-loop structure and a log-likelihood score is assigned to each window to determine how well its attributes resemble those of experimentally verified miRNA [Lim 2003]. Warris et.al. describe yet another method of prediction of small regulatory RNAs in genomes using MFE distribution of sequences as the discerning factor [Warris 2014]. The underlying principle is that the secondary structures of small regulatory RNAs have lower free energies than random RNA or other ncRNA sequences of the same length and di-nucleotide composition. The importance of the length and MFE in RNA studies sequences is obvious from this discussion.

Both computational [Yoon 2007], [Washietl 2005], [Tran 2009] and signal processing based approaches [Anastassiou 2000], [Vaidyanathan 2002], [Yoon 2004], [George 2010] are popular in the analysis of the coding region of the genome. Although computational methods have been widely employed to study noncoding RNA, little work has been done which makes use of Digital Signal Processing techniques to analyze the noncoding genome. As seen, MFE and sequence length are important parameters to be analyzed in the study of RNA, however a mathematical relationship linking MFE to length or any other signal parameter of the sequence has not been reported in literature until date. Here in this work we have introduced a novel approach, which links MFE, a thermodynamic property of ncRNA sequences to their signal properties.

6.2. Non coding RNA sequences used

Over 2600 ncRNA sequences downloaded from the benchmarked database, Rfam [Rfam/Pfam database] were used in this work. The classes of ncRNA whose MFE were analyzed are snRNA (902), snoRNA (573), miRNA (376), rRNA (805) taken from across bacteria, archaea, fungi and eukaryotes. A model for MFE (minimum free energy) of the secondary structure of these noncoding RNA sequences is developed from their signal parameters viz. length and the spectral coefficient matrix making use of multiple linear regression analysis [George 2016]. The function and properties of the four classes of noncoding RNA sequences have already been discussed in chapters 3 and 4.

6.2.1. snRNA

The snRNA sequences used in this work were taken from the Rfam database. 902 snRNA sequences were used in this study. snRNA belonging to different Rfam families were used. Each Rfam family contains multiple number of sequences. The family names and the sequence identifiers of these sequences are given in Tables 6.4 and 6.5 of the section 6.4.

6.2.2. snoRNA

In this study, more than 500 snRNA sequences were analysed which were taken from the Rfam database [Rfam/Pfam database]. The Rfam families used in this study is given in the Table 6.1 given below.

Table 6.1. Rfam snoRNA families used in this study

RF00012	RF00147	RF00049	RF00157
RF00045	RF00152	RF00054	RF00160
RF00090	RF00205	RF00093	RF00221
RF00091	RF00218	RF00188	RF00056
RF00181	RF00190	RF00055	RF00067
RF00134			

6.2.3. miRNA

The micro RNA sequences used in this study were taken from the Rfam database [Rfam database]. For example the Rfam miRNA family RF00694 contains 28 sequences whereas Rfam family RF00813 contains 15 sequences. A total of around 380 sequences from different Rfam families belonging to different organisms were used. Table 6.2 below gives the names of the Rfam miRNA families used in this work.

Table 6.2. Rfam miRNA families used in this study

RF00706	RF00754
RF00694	RF00795
RF00813	RF00948
RF00824	RF00645
RF00728	RF00747
RF00641	RF00782

6.2.4. rRNA

The rRNA sequences used in this work were taken from the Rfam database. The Rfam rRNA families used in this study are: RF00001, RF00002, RF01118. 805 rRNA sequences from these families across different organisms was used in this study.

6.3. Developing a novel model for MFE

The optimal two-dimensional MFE structures of a sample of over 2600 non-coding RNA sequences were found out with the thermodynamic nearest neighbour algorithm using MATLAB R2015b and the free energies were recorded. The algorithm is as detailed in Chapter 4. A novel mathematical model for MFE was developed in terms of signal parameters of ncRNA sequences using multiple linear regression analysis. This model was used to compute MFE of ncRNA sequences directly from the signal parameters, without using any folding algorithm. MFE values so obtained were compared and ratified with those obtained using standard web servers, RNAfold and RNAstructure [George 2016].

6.3.1. Secondary Structure prediction and evaluation of MFE

The basic dynamic programming algorithm for the thermodynamic nearest neighbour model was proposed by Zuker and Steigler in 1981 [Zuker 1981]. Optimal minimum free energy secondary structure was predicted for the sequences analyzed starting from the primary sequence [Mathews 1999], [Mathews 2010], [Markham 2008]. As already explained in Section 4.4.2 of Chapter 4, RNA secondary structure can be uniquely decomposed into stacked bases, hairpin-loops, bulges, interior-loops, and multi-way-junctions and energies are assigned to these substructures. MFE is estimated in kcal/mol by summing individual energy contributions from the secondary substructures, viz. base pair stacks, hairpins, bulges, internal loops and multi-branch loops. An up-to-date set of energy parameters is maintained by the Turner's Lab [Mathews 1999], [Xia 1998]. In this computation, canonical and non-canonical base pairings are considered, the energy contribution of coaxially stacked helices is not accounted for, and the formation of pseudoknots is forbidden. The secondary substructures have energy contributions that are sequence and length-dependent. The algorithm implemented uses dynamic programming to compute the energy contributions of all possible elementary substructures and then predicts the secondary structure by considering the combination of elementary substructures whose total free energy is minimum.

A sample secondary structure plot of rRNA sequence of mus musculus with GenBank accession number NR_046118.1 is shown in Figure 6.1. The signal properties considered here for developing the mathematical model are 1) the length of the ncRNA sequences in terms of the number of nucleotides (mentioned as NTL) and 2) standard deviation of the spectral coefficient matrix of the sequences (mentioned as SD_SCM).

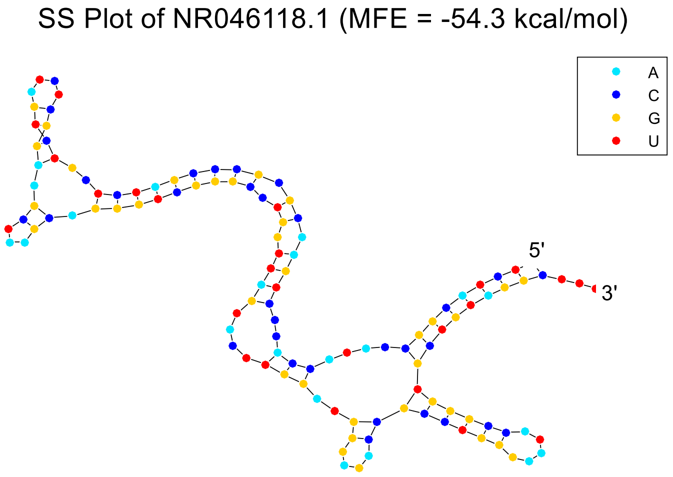


Figure 6.1. Secondary Structure plot of NR_046118.1. Mus musculus ribosomal RNA (rRNA)

6.3.2 Novel model for MFE

The signal properties considered here for developing the mathematical model are 1) the length of the ncRNA sequences in terms of the number of nucleotides (mentioned as NTL) and 2) standard deviation of the spectral coefficient matrix of the sequences (mentioned as SD_SCM).

6.3.2.1. Signal length, coefficient matrix of the signal spectrum

In order to make it conducive for digital signal processing, the sequences of letters from the four-character alphabet were first converted into numerical sequences. The binary indicator sequence representation was used here [Anastassiou 2001 (1)]. $u_a[n]$, $u_u[n]$, $u_c[n]$, $u_g[n]$ are the binary indicator sequences corresponding to A, U, C G which take on a value of 0 or 1 at location n, depending on whether the corresponding character exists or not at n.

$u_a[n]$ +$u_u[n]$ +$u_c[n]$ +$u_g[n]$ = 1 (6.1)

N is the sequence length, NTL.

The numerical sequence resulting from a character string of length N can be written as:

x[n] = a$u_a[n]$ + uu_u[n] + cu_c[n] + gu_g[n] (6.2)

n = 0, 1, 2, 3....... (N-1) and $a = 1 + j,\ u = 1 - j,\ c = -1 - j,\ g = -1 + j$, following the convention of complex representation of bases [Cristea 2002] where purines and pyrimidines are represented by numbers that are complex conjugates. The multipliers a, u, c, g are taken as 1, 2, 3 and 4 respectively. The length of the sequence is the number of nucleotide bases in it, indicated as NTL (nucleotide length).

To obtain the spectral coefficients, the Digital Fourier Transform (DFT) of the sequence was found out using the FFT (Fast Fourier Transform) algorithm. DFT of a sequence x[n], of length N, is itself another sequence X[k], of the same length N [Proakis 2006], [Oppenheim 2009] can be expressed mathematically as,

$$X\ (k) = \sum_{n=0}^{N-1} x(n) e^{-(jk2n\pi)/N}$$

(6.3)

Magnitudes of spectral coefficients were separated from the spectrum and their standard deviation computed. This is SD_SCM.

6.3.2.2. Multiple Linear Regression analysis

The mathematical models linking MFE, NTL and SD_SCM were arrived at, making use of regression analysis. Regression is a generic term for all methods that attempt to fit a model to observed data in order to *quantify the relationship* between two groups of variables. The fitted model may then be used either to merely *describe* the relationship between the two groups of variables, namely the predictor or the independent variable(s) and the dependent or the target or the response variable(s). In all cases, the target (dependent variable) is a function of the independent variables called the regression function. In general terms, if the two data matrices involved in regression are usually denoted as *X* and *Y*, where *X* represents the independent variable and *Y*, the dependent variable. The purpose of regression is to build a model *Y* = *f(X)*. Such a model tries to explain, or predict, the variations in the Y-variable(s) from the variations in the X-variable(s). The link between *X* and *Y* is achieved through a common set of samples for which both *X*- and *Y*-values have been collected.

As there are more than one predictor variables (NTL, SD_SCM multiple linear regression (MLR) was used for developing the mathematical models for MFE in this work. The iteration done here is based on the minimum squared errors approach. MLR examines the linear relationships between one continuous response and two or more predictors. If the number of predictors is large, then before fitting a regression model with all the predictors, you should use stepwise or best subsets model-selection techniques to screen out predictors not associated with the responses [Montgomery 2006], [Chatterjee 2012], [Sanford 2005]. The general format for the multiple linear regression relationship can be written as,

$$y|x_1, x_2. \ldots . x_n = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad (6.4)$$

This is referred to as the regression equation where, *y* is the response variable or the dependent variable and $x_1, x_2 \ldots . x_n$ are the predictor variables or the independent variables. b_0 is the intercept, b_1, b_2, ... b_n are the slopes or the coefficients.

While performing modelling making use of the multiple linear regression, the following assumptions are made. Let x represent $\{x\}$.

- The observations 'y' are assumed to be statistically independent.

- The standard deviation of 'y' within a particular set of values of 'x' is constant for all range of values of 'x'.

- The distribution of 'y' within 'x' is normal.

We saw the regression analysis MFE vs NTL and MFE vs SD_SCM individually, in the earlier chapter, which was simple linear regression where, the relationship between the predictor and the response variables, x and y, were modelled to fit into a straight line, the regression line. [Montgomery 2006]. Here in our case, we have three variables, MFE, NTL, SD_SCM, which are represented by y, x_1, x_2 respectively, and the resultant regression curve would be a plane called the regression plane, not a line. Thus the regression model describes a plane, in the three variable space of y, x_1, x_2. Multiple linear regression with three variables, y, x_1, x_2, which are representative of MFE, NTL and SD_SCM.

The regression equation in this case, reduces to

$$y = b_0 + b_1x_1{+}b_2x_2 \qquad (6.5)$$

Here again, the minimum mean squared error algorithm is used to evaluate the coefficients b_0, b_1 and b_2 and the dispersion parameter R^2. The underlying principle of the minimum mean squared error method is, the residual sum of estimates be minimum. Suppose that k observations are available. Let y_j denote the j^{th} value of the response variable and let $x_{11}, x_{12}, x_{13 \ldots\ldots} x_{1k}$, $x_{21}, x_{22}, x_{23} \ldots. x_{2k}, \ldots.$ $x_{n1}, x_{n2}, x_{n3} \ldots.. x_{nk}$ represent the k values for each of the predictor variables $x_1, x_2, \ldots.. x_n$. It is depicted as shown below in Table 6.3.

Table 6.3. Representation of data for multiple linear regression

Observation j	**Response** y_j	**Regressors**				
		x_1	x_2	x_3	**.......**	x_n
1	y_1	x_{11}	x_{21}	x_{31}	......	x_{n1}

2	y_2	x_{12}	x_{22}	x_{32}	……	x_{n2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	y_k	x_{1k}	x_{2k}	x_{3k}	……	x_{nk}

Given the set of points $Y = [y_1, y_2, \ldots . y_k]$ (6.6)

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{bmatrix} \qquad (6.7)$$

And

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{1k} & \cdots & b_{nk} \end{bmatrix} \qquad (6.8)$$

In the above, $j = 1, 2.\ldots . k$, the number of observations and $i = 1,2,3. \ldots n$, the number of predictor/independent variables. In our case, $n = 2$ (2 independent variables) and k is the number of specimen taken. Here, it is total number of the different kinds on ncRNAs taken across organisms. That is, in this study, k is 902 for snRNA, 572 for snoRNA, 376 for miRNA and 805 for rRNA.

An assumption regarding the variables in *X* is that they are mathematical, that is to say, that they are non-random, and measured without error, in a designed experiment and are random variables in an observational study. When *X* s are random variables, the observations x_{ij} are independent and their distribution does not depend on regression coefficients. While testing a hypothesis or constructing confidence intervals, we assume the conditional distribution of *y* given x_1, x_2, x_3 ..... x_k is normal with mean of $b_0 + b_1x_1 + b_2x_2 + \ldots . . + b_kx_k$.

Let y_k and $\hat{y}_k$ represent the k thvalue of observed and predicted values of the response variable respectively. Then,

$$y_k = b_{0k} + b_{1k}x_{1k} + b_{2k}x_{2k} + \cdots + b_{nk}x_{nk} \qquad (6.9)$$

$$\hat{y}_k = \hat{b}_{0k} + \hat{b}_{1k}x_{1k} + \hat{b}_{2k}x_{2k} + \cdots + \hat{b}_{nk}x_{nk} + \ \varepsilon_i \qquad (6.10)$$

$$y_j = \sum_{i=1}^{n} b_0 + b_{ij}$$

(6.11)

$$\hat{y}_j = \sum_{i=1}^{n} \hat{b}_0 + \hat{b}_{ij}$$

(6.12)

$j \ = \ 1,2,3.\ldots n$, where n represents the number of predictor variables and k, the number of observations that can be made of the response variable. ε_i represents the error between the predicted and observed values of *y*.

The error squared function is

$$S(\{b_1, b_2 \ldots . b_k\}) = \sum_{j=1}^{k}\left(y_j - \hat{y}_j\right)^2 = \sum_{j=1}^{k} \varepsilon_j^2$$

(6.13)

The algorithm for regression here works on the MMSE principle. So the function *S* is to be minimised with respect to $b_0, b_1 \ldots . b_k$. It mathematically means that,

$$\frac{\partial S}{\partial b} = 0 \qquad (6.14)$$

For each value of S and b

$$S = \sum_{j=1}^{k}\left(y_j - \hat{y}_j\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(b_{0j} + b_{ij}x_{ij} \ - \left(\hat{b}_{0j} + \hat{b}_{ij}x_{ij}\right)\right)^2$$

(6.15)

$$\frac{\partial S}{\partial B} = -2\sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_i - \hat{b}_0 - b_0 x_{ij}\right) x_{ij} = 0$$

(6.16)

Simplifying the above equation, we get the least squares normal equations,

$$n.\hat{b}_0 + \hat{b}_1\sum_{i=1}^{n} x_{i1} + \hat{b}_2\sum_{i=1}^{n} x_{i2} + \ldots\ldots + \hat{b}_k\sum_{i=1}^{n} x_{ik} = \sum_{i-1}^{n} y_i$$

$$\hat{b}_0\sum_{i=1}^{n} x_{i1} + \hat{b}_1\sum_{i=1}^{n} x_{i1}^2 + \hat{b}_2\sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{b}_k\sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

:

:

:

$$\hat{b}_0\sum_{i=1}^{n} x_{ik} + \hat{b}_1\sum_{i=1}^{n} x_{ik}\, x_{i1} + \hat{b}_2\sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{b}_k\sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i$$

(6.17)

The solution to the normal equations will be, least squares estimators, $\hat{b}_0, \hat{b}_1, \hat{b}_2 \ldots\ldots \hat{b}_k$.

There are *p = k+1* normal equations, one for each of the unknown regression coefficients. For easiness of representation, we can use the matrix notation. So then the equations become,

$$Y = XB + \varepsilon \qquad (6.18)$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_k \end{bmatrix} \quad (6.19) \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} (6.20)$$

$$B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \quad (6.21) \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_K \end{bmatrix} \qquad (6.22)$$

Here Y, is an $k \times 1$ (column) matrix of observations, X is an k×n matrix of the levels of the regressor variables, B is a n×1 matrix of the regression coefficients, and ε is a k×1 matrix of random errors. We have to find out the vector of least square estimators, $\hat{B}$ such that S(B) is made minimum.

$$S(B) = \sum_{j=1}^{k} \varepsilon_j^2 = \varepsilon'\varepsilon = (Y - XB)'(Y - XB)$$

(6.23)

$S(B)$ can be expressed as,

$$S(B) = Y'Y - B'X'Y - Y'XB + B'X'XB = Y'Y - 2B'X'Y + B'X'XB \qquad (6.24)$$

Since $B'X'y$ is a 1 × 1 matrix, it's transpose $(B'X'Y)' = Y'XB$ is the same scalar. The LS estimators should satisfy,

$$\frac{\partial S}{\partial B} = -2X'Y + 2X'XB' = 0 \qquad (6.25)$$

This simplifies to, $X'X\hat{B} = X'Y$ (6.26)

Equations 6.25 and 6.26 are the LS normal equations, and are the matrix analogue of the scalar representation. Upon multiplying 6.26 throughout with $(X'X)^{-1}$, we've,

$$\hat{B} = (X'X)^{-1}X'Y \qquad (6.27)$$

Provided that the inverse $(X'X)^{-1}$ exists. $(X'X)^{-1}$ will exist if the regressors are linearly independent ie if no column of the X matrix is a linear combination of the other.

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}{}^{2} & \sum_{i=1}^{n} x_{i1}\, x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}{}^{2} \end{bmatrix} \cdot \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \vdots \\ \hat{b}_k \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}\, y_i \\ \vdots \\ \vdots \\ \sum_{i=1}^{n} x_{ik} y_i \end{bmatrix}$$

(6.28)

The fitted regression model corresponding to the levels of the regressor variables X'=[1, $x_1, x_2, \ldots.., x_k$] is,

$$\hat{Y} = X'\hat{B} = \hat{b}_0 + \sum_{J=1}^{K} \hat{b}_j\, x_j$$

(6.29)

The vector of fitted values $\hat{Y}$, corresponding to the observed values Y, is given as,

$$\hat{Y} = X'\hat{B} = X(X'X)^{-1}X'Y = HY \qquad (6.30)$$

Where, the matrix, H is called the hat matrix. The hat matrix and its properties play a central role in regression analysis. The difference between the observed value y_j and the corresponding fitted value $\hat{y}_j$ is the residual $e_j = y_j - \hat{y}_j$. This can be expressed in matrix form as $E = Y - \hat{Y}$.

In the work presented here, multiple linear regression analysis was done taking MFE as the response variable and NTL, SD_SCM as the predictor variables. The statistical toolbox of MATLAB R2015b was used to perform regression analysis. Linear equations were arrived at linking the three, for the four classes of ncRNA analyzed. The equations are explained in the section 6.5 and have been tabulated in Table 6.4. MFE was computed from sequence length and standard deviation of the spectral coefficient matrix using these mathematical models, for each class of the ncRNA analyzed. The accuracy of the model developed was probed by comparing the MFE values obtained by using the model (named MFE_C) with MFE values obtained via MATLAB (named as MFE_M) and relative deviations were found. The performance of the model was evaluated by comparing MFE values computed using it with the ones obtained from standard web servers RNAfold (MFE_F) and RNAstructure (MFE_S). Deviations in MFEs computed with the models developed were found out relative to the MFE values obtained using these two web servers. Results are given in section 6.4.

6.3.2.3. Mathematical models

6.3.2.3.1. 3D scatter plots of multiple linear regression analysis

MFEs computed with the thermodynamic nearest neighbour algorithm via MATLAB were related to the signal properties of the sequences namely the length and the standard deviation of the spectral coefficient matrices of the sequences. Figures 6.2 to 6.5 show the plots of MFE scattered against NTL (length of the sequence) and SD_SCM (standard deviation of the spectral coefficients) in 3D space of snRNA, snoRNA, miRNA and rRNA sequences respectively. The plots have MFE marked along the z axis, NTL along the x axis and SD_SCM along the y axis. The rainbow grid in the graphs indicates the ideal fit plane. In regression, perfect fit is said to occur when the iterations of the predictor variables are perfect and there is zero error. The scatter plots taken for the different classes of noncoding RNA sequences show that most of the points fall on one plane, the number of outliers are very few. This indicates the correctness of the analysis.

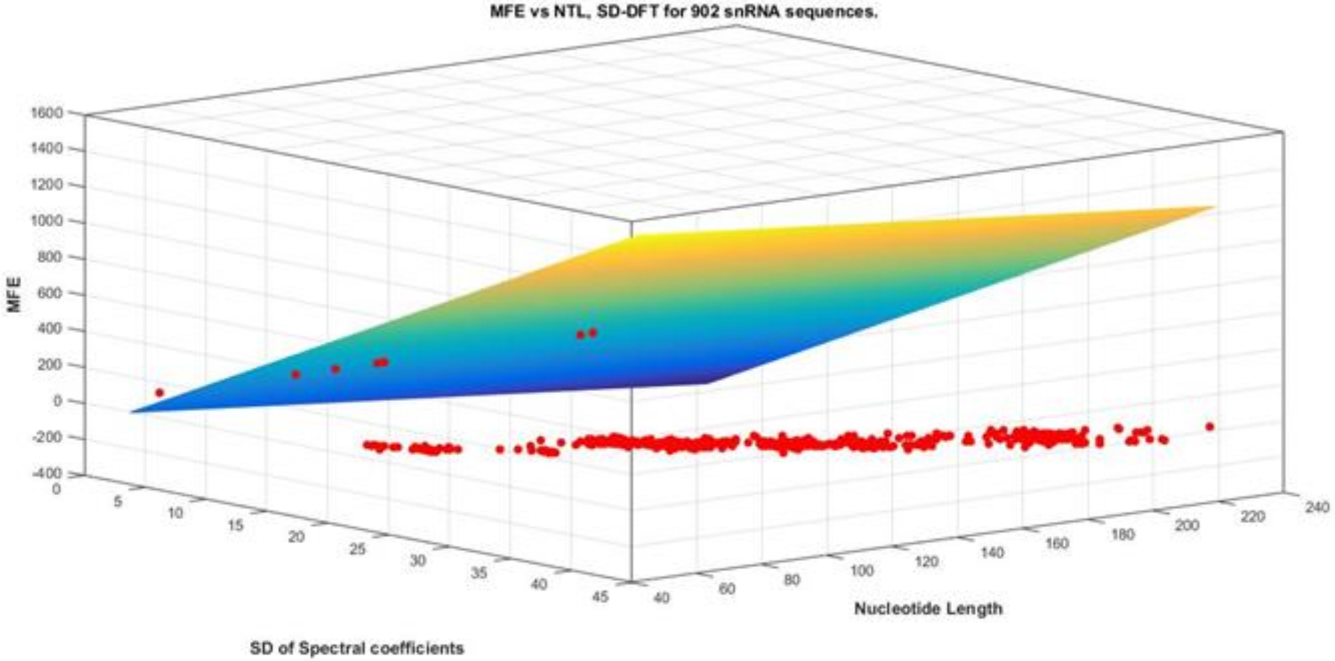


Figure 6.2. Plot of MFE vs NTL, SD_SCM for snRNA (902) sequences

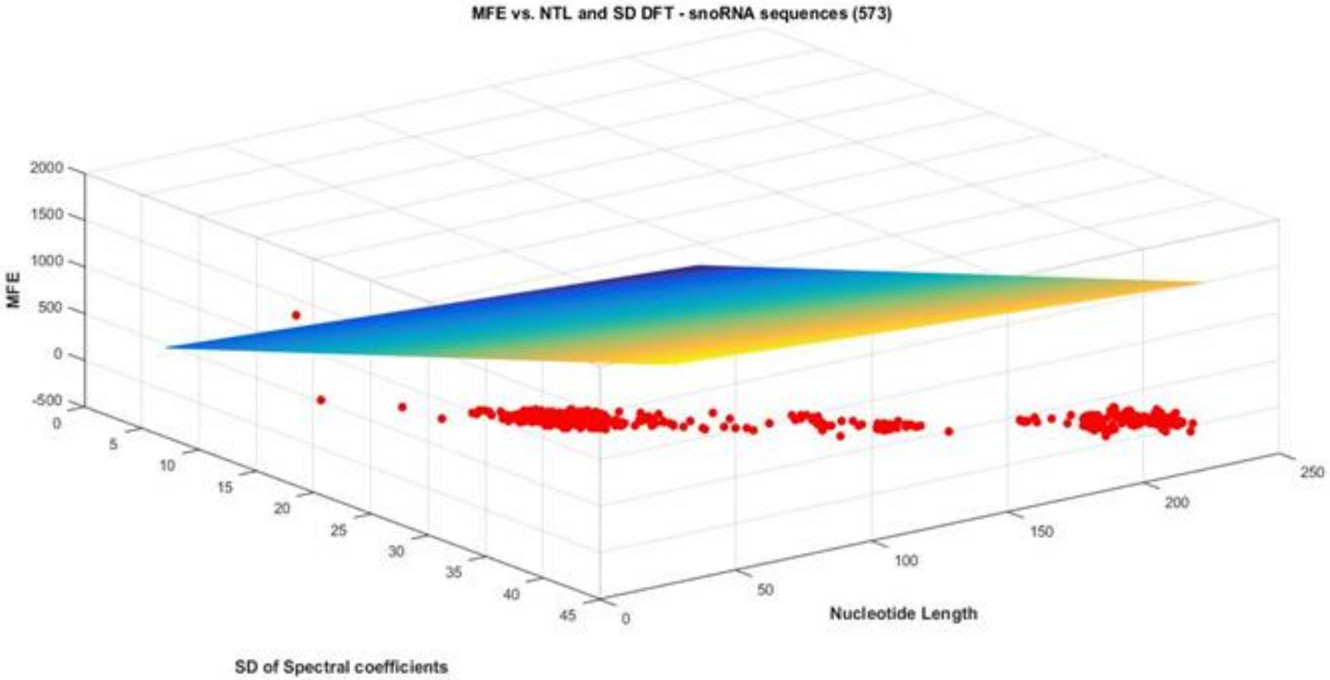


Figure 6.3. Plot of MFE vs NTL, SD_SCM for snoRNA (573) sequences

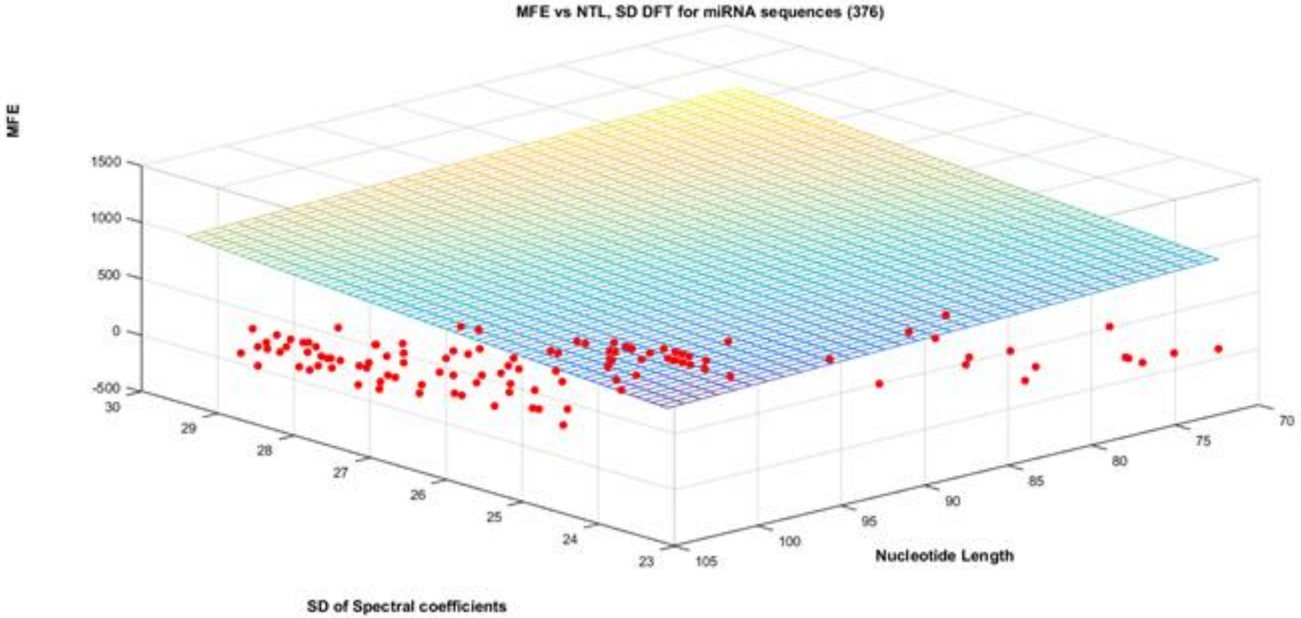


Figure 6.4. Plot of MFE vs NTL, SD_SCM for miRNA (376) sequences

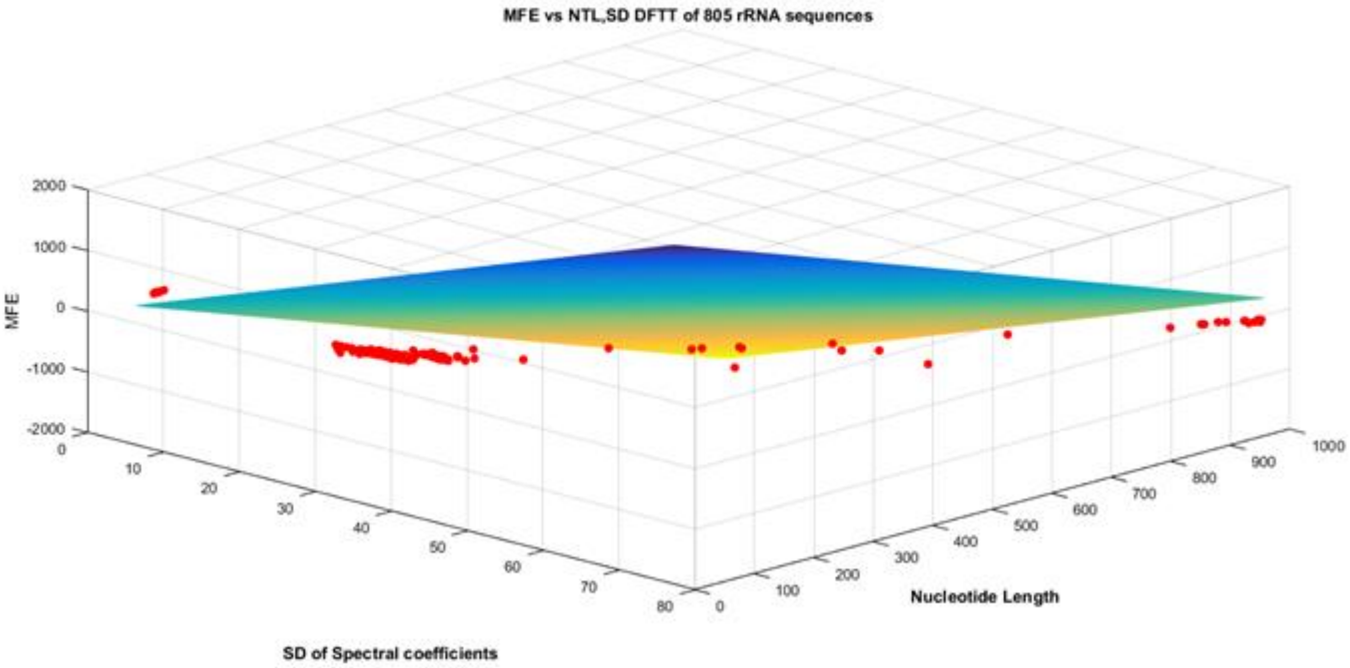


Figure 6.5. Plot of MFE vs NTL, SD_SCM for rRNA (805) sequences

6.3.2.3.2. Mathematical models developed for MFE.

Table 6.4 The equations developed relating MFE with sequence length and SD of spectral coefficient matrices for the four classes of ncRNA

Sl. No	**Class of ncRNA**	**Regression coefficients**			**Equation linking MFE (y) with NTL (x_1) and SD_SCM (x_2)**
		$\boldsymbol{b_0}$	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	
1	miRNA	45.5857	0.3455	-4.2116	$y = 0.3455x_1 - 4.2116x_2 + 45.5857$
2	rRNA	21.2110	-0.1485	-1.5454	$y = -0.1485x_1 - 1.5454x_2 + 21.2110$
3	snoRNA	41.7028	-0.2219	-1.7731	$y = -0.2219x_1 - 1.7731x_2 + 41.7028$
4	snRNA	36.2222	-0.1996	-1.5913	$y = -0.1996x_1 - 1.5913x_2 + 36.2222$

The general form of the MLR equation for one response variable (y) and two predictor variables (x_1, x_2) is $y = b_1x_1 + b_2x_2 + b_0$. The significance of the regression parameters b_0, b_1, b_2 has been already explained in the previous section. The mathematical models developed relating MFE to the length and standard deviations of spectral coefficient matrix of sequences for the four classes of ncRNA analyzed are given in Table 6.4. The values of regression coefficients b_1 and b_2 and the intercepts b_0 obtained for each class are also shown. The mathematical model developed for each class was used in computing MFE from NTL and SD_SCM for the corresponding class of ncRNA analyzed.

The equation linking MFE (y) with NTL (x_1) and SD_SCM (x_2) shown in Table 6.4 are given below.

$y = 0.3455x_1 - 4.2116x_2 + 45.5857$ (6.31)
$y = -0.1485x_1 - 1.5454x_2 + 21.2110$ (6.32
$y = -0.2219x_1 - 1.7731x_2 + 41.7028$ (6.33)
$y = -0.1996x_1 - 1.5913x_2 + 36.2222$ (6.34)
for miRNA, rRNA, snRNA and snoRNA sequences respectively.

6.3.3. Computation of MFE from mathematical models

Mathematical models for MFE (the dependent/target variable) were developed via multiple regression analysis with the sequence length and standard deviation of the spectral coefficient matrix as the independent variables. The models were developed by grouping the ncRNAs into four classes as already mentioned. The mathematical model developed for each class was used in computing MFE from NTL and SD_SCM for the corresponding class of ncRNA analyzed. The equations linking MFE (y) with NTL (x_1) and SD_SCM (x_2) are as given in equations 6.31 to 6.34.

These equations were used to compute MFE (y) from sequence length (x_1) and SD_SCM (x_2). The value of MFEs so obtained, named MFE_C were compared with those obtained using the thermodynamic nearest neighbour algorithm making use of the Bioinformatics toolbox of MATLAB. This value is denoted by MFE_M. Deviation in the computation of MFE_C was found out relative to MFE_M (termed RD_1). Relative deviation is found out as given below.

$$RD\_1 = \frac{(MFE\_M - MFE\_C)}{MFE\_M} \times 100 \quad (6.35)$$

6.3.3.1 Checking accuracy of mathematical models using webservers

The mathematical models developed which are given in equations 6.31 to 6.34 were used to compute MFE of the all four classes of ncRNA. 376 miRNA, 805 rRNA, 902 snRNA, and 573 snoRNA sequences. The accuracy of the models was checked making use of standard webservers, RNAfold and RNAstructure for each of the sequences analyzed.

MFE of all the 2656 ncRNA sequences was computed using the webservers. The MFEs obtained from webservers RNAfold and RNAserver are denoted in this work as MFE_F and MFE_S respectively. These values of MFE thus obtained was compared with the values of MFE obtained using the mathematical models developed. The percentage of relative deviation was calculated for each of the sequences analysed. The percentage of deviation in MFE_C (MFE obtained from the model developed) relative to MFE_F (MFE obtained from RNAfold) is termed RD_2 and the percentage of deviation in MFE_C relative to MFE_S (MFE obtained from RNAserver) is termed RD_3. They are found out as given below.

$$RD\_2 = \frac{(MFE\_F - MFE\_C)}{MFE\_F} \times 100 \quad (6.36)$$

$$RD\_3 = \frac{(MFE\_S - MFE\_C)}{MFE\_S} \times 100 \quad (6.37)$$

Results are discussed in Section 6.4 and are included as Appendix (pages 305 to 422). Only one set of results are included in this thesis to conserve space.

6.4. Applying the mathematical models to compute MFE

A total of 2656 noncoding RNA sequences belonging to four classes (miRNA, rRNA, snRNA, snoRNA) were downloaded from the benchmarked database Rfam. Optimal MFE secondary structures of the sequences were found out with the thermodynamic nearest neighbour approach using MATLAB R2015b. Mathematical models for MFE for these four classes of ncRNA were developed from the signal parameters of the sequences viz. length and SD of spectral coefficient matrices of the sequences. Using the novel mathematical models developed, MFE was computed from nucleotide length (NTL) and the standard deviation of the spectral coefficient matrix (SD_SCM) for each of the four classes of ncRNA analyzed. The accuracy of the model was also checked making use of MFE values computed with standard webservers. It was found that the novel model yielded MFE values which were within ± 15% deviation in comparison to the MFE values obtained with standard webservers for all the four classes of ncRNA analysed.

6.4.1. MFE from the novel model

The MFE computed from the models developed is termed MFE_C in this work. The MFE so computed using the models was compared with the MFE values obtained with the thermodynamic nearest neighbour algorithm using the bioinformatics toolbox of MATLAB (MFE_M). The deviation of MFE_C from MFE_M, is termed RD_1, calculated as given in equation 6.35.

A sample of results of computation of MFE from the mathematical models developed for a set of 10 sequences from each class of ncRNA analysed is given in Table 6.5. Table 6.5 has the identities of sequences in column 2. The third column shows the length of the sequence (NTL) and the fourth column has the standard deviation of the spectral coefficient matrix of the sequence (SD_SCM). From these two signal parameters MFE is computed as per the mathematical model developed. The mathematical models are as in equations 6.31 to 6.34. The values of RD_1 are seen to be within ±15%. The Table 6.6 in Appendix shows the same for the entire set of 902 snRNA sequences analysed. The results of all the 2656 sequences could not be included in this thesis due to the restriction on space.

Results of computation of MFE_C is given in Table 6.6 in the Appendix I are for the set of 902 snRNA sequences analysed. The 902 snRNA sequences analysed belong to different Rfam families viz. RF00004, RF00007, RF00026, RF00283, RF00492, RF01458, RF01475, RF01490, RF00618.The mathematical model used in this computation is given in equation 6.31. Table 6.5 has the identities of sequences in column 2. The third column shows the length of the sequence (NTL) and the fourth column has the standard deviation of the spectral coefficient matrix of the sequence (SD_SCM). From these two signal parameters MFE is computed as per the mathematical model developed for snRNA is : $y = -0.1996x_1 - 1.5913x_2 + 36.2222$; where y is MFE, x_1 NTL (nucleotide length) and x_2 is SD_SCM (standard deviation of the spectral coefficient matrix of snRNA sequences). This equation was arrived at upon regression analysis of the 902 snRNA sequences studied. This MFE computed from the model is indicated as MFE_C, whereas MFE computed with the Bioinformatics toolbox of MATLAB is indicated by MFE_M. Deviation in the computation of MFE_C was found out relative to MFE_M (shown as RD1) and the percentage of relative deviation is shown in column 8 of Table 6.6. 1.55210 % (14 out of 902 sequences) of the sequences had values of RD_1 beyond ± 15%. These have been highlighted in red. One outlier was found which had a value of relative deviation 41.98703% (sequence identifier: AAFD02000024.1/69022-69131). These results indicate that the sample at hand was conducive to regression analysis. The time of computation of MFE using the method developed was noted for each of the sequences studied and compared with the time taken for computation of MFE of sequences while making use of the RNAFold webserver and the RNAStructure webserver. It was found that the time of computation of MFE using the proposed algorithm was comparable to that of RNAFold webserver.

Table 6.5. Sample computation of MFE using the model developed for 10 samples each of the four classes of ncRNA analysed. Source of sequences : Rfam database

rRNA - Rfam family RF00001							
Sl.No.	**Specimen ID**	**NTL**	**SD_DFT**	**MFE_M**	**MFE_C**	**RD_1**	**%RD1**
1	X01556.1/3-118	116	26.962295	-37.2	-37.682531	-0.0129713	-1.297126
2	X55260.1/3-119	117	30.345851	-46	-43.059979	0.0639135	6.391351
3	M16174.1/3-119	117	30.954639	-49.2	-44.000799	0.1056748	10.567483
4	X55267.1/3-119	117	30.758514	-49.2	-43.697708	0.1118352	11.18352
5	M16172.1/3-119	117	30.856732	-49.9	-43.849494	0.1212526	12.125262
6	AF001265.1/6033-6149	117	30.561132	-48.8	-43.392673	0.1108059	11.080589
7	X05057.1/3-119	117	30.561132	-49.6	-43.392673	0.1251477	12.514773
8	X15126.1/3-120	118	33.460848	-52.2	-48.022394	0.0800308	8.0030763
9	Z50737.1/3-119	117	34.214221	-52.9	-49.038157	0.0730027	7.3002704
10	X55261.1/3-119	117	31.084701	-50.4	-44.201797	0.1229802	12.298023
snoRNA - Rfam RF00012 family							
1	ABGA01262676.1/1340-1125	216	41.906208	-85.9	-79.841099	0.0705343	7.0534349
2	AANU01105435.1/2236-	217	41.326388	-83.1	-78.978957	0.0495914	4.9591377

	2452						
3	AAGV020566442.1/1857-1640	218	42.747895	-86	-81.827471	0.0485178	4.8517776
4	ABVD01644074.1/1753-1970	218	42.653788	-81.2	-81.652988	-0.0055787	-0.557867
5	CABF01044723.1/2170-2355	186	38.260558	-66.4	-66.6947	-0.0044383	-0.443825
6	AAGD02005452.1/2742-2555	188	38.547422	-70.6	-67.652375	0.0417511	4.175106
7	AY948622.1/2-186	185	37.917961	-66.8	-65.846591	0.0142726	1.4272592
8	CAAC02000606.1/1368922-1368738	185	37.79845	-69.3	-65.625006	0.0530302	5.3030213
9	ABKE01003568.1/5923-6108	186	38.062963	-67	-66.328341	0.0100248	1.0024767
10	AAQA01000362.1/29408-29193	216	42.181005	-92.5	-80.350602	0.1313448	13.134484
miRNA Rfam family RF00706							
1	AALT01209640.1/567-377	90	26.605736	-35.4	-35.356053	0.0012414	0.1241448
2	AAFR03033875.1/20528-20718	91	26.925205	-39.7	-36.355838	0.0842358	8.4235808
3	AAIY01044029.1/787-597	90	25.07667	-30.9	-28.917158	0.0641696	6.4169641
4	AAZO01007389.1/15370-15178	93	26.505742	-30.5	-33.898479	-0.1114255	-11.14255
5	AAYZ01695118.1/310-500	90	25.893038	-33.7	-32.354885	0.0399144	3.9914399

6	AAHX01044404.1/26102-26292	93	25.205573	-29.7	-28.423468	0.0429809	4.298088
7	AACN010750078.1/657-848	91	26.849995	-35	-36.039128	-0.0296894	-2.968937
8	ABAV01019481.1/5988-6180	90	26.870476	-33	-36.470874	-0.105178	-10.51779
9	AAZX01018356.1/721-913	89	27.493138	-36.5	-39.438404	-0.0805042	-8.050422
10	AY765362.1/650-458	90	26.376701	-35.9	-34.391586	0.0420171	4.2017091
Rfam snRNA family RF00004							
1	AALT01209640.1/567-377	191	37.914724	-63.1	-62.235101	0.0137068	1.3706803
2	AAFR03033875.1/20528-20718	191	37.675344	-65.8	-61.854176	0.0599669	5.9966936
3	AAIY01044029.1/787-597	191	37.80852	-63.7	-62.066098	0.0256499	2.5649946
4	AAZO01007389.1/15370-15178	193	38.609895	-70.3	-63.740525	0.0933069	9.3306896
5	AAYZ01695118.1/310-500	191	38.257851	-63.4	-62.781118	0.0097615	0.9761543
6	AAHX01044404.1/26102-26292	191	37.781923	-65.2	-62.023774	0.0487151	4.8715129
7	AACN010750078.1/657-848	192	38.309842	-66.9	-63.063451	0.0573475	5.7347515
8	ABAV01019481.1/5988-6180	193	38.570822	-64	-63.67835	0.0050258	0.5025789
9	AAZX01018356.1/721-913	193	45.570822	-79.3	-74.81745	0.0565265	5.6526488
10	AY765362.1/650-458	193	38.830561	-70	-64.091672	0.0844047	8.440468

6.4.2 Accuracy of the novel model

Accuracy of the models developed was checked by computing the relative deviations of MFE values obtained using the model (MFE_C) with those obtained using the web servers RNAfold (MFE_F) and RNAstructure (MFE_S). These are represented as RD_2 and RD_3 respectively, calculated as given on equations 6.36 and 6.37 respectively. Out of the total 2656 sequences analyzed around 95% were found to have relative deviations (both RD_2 and RD_3) within ± 15%. The deviation values were around than ± 5% for 45% and were between ± 5 to ± 10% for 35% of the sequences. 15% of the sequences had deviation values between ± 10% to ± 15%. Only around 5% of the sequences had deviation values above ± 15% (marked in red). It is also to be noted that the correlation between the MFE values obtained via RNAfold and RNAstructure was not found to be one (1) always. This is true for all the 4 classes of ncRNA sequences analysed.

Only sample results are included in this thesis due to the constraint on space. Table 6.7 given below shows the results of calculation of RD_2 and RD_3 for a sample of 10 sequences each from the four classes of ncRNA analysed. Column 2 of Table 6.7 shows the sequence identifier, columns 3, 4 show the nucleotide length and the SD_SCM respectively. Column 5 gives the MFE_C i.e. MFE value of the sequence computed using the model developed. Column 6 contains the MFE for the same sequence obtained from RNAfold, column 7 gives the relative deviation of MFE_C with respect to MFE_F and column 8 gives the same in percentage. Column 9 contains MFE_S, column 10 shows the deviation of MFE_C with respect to MFE_S and column 10 shows the same deviation in percentage. As observed in the entries for RD_2 and RD_3 for the sample of forty ncRNAs, selected 10 each from the four classes analysed, the values of RD_2 and RD_3 are within ±15%.

Table 6.7. Checking the accuracy of the model developed, making use of MFE values from web-servers

rRNA - Rfam family RF00001											
Sl.No.	**Specimen ID**	**NTL**	**SD_DFT**	**MFE_M**	**MFE_C**	**MFE_F**	**RD_2**	**%RD2**	**MFE_S**	**RD_3**	**%RD3**
1	X01556.1/3-118	116	26.962295	-37.2	-37.682531	-33.5	-0.1248517	-12.485167	-33.8	-0.1148678	-11.486778
2	X55260.1/3-119	117	30.345851	-46	-43.059979	-42.4	-0.0155655	-1.5565532	-42.9	-0.0037291	-0.3729104
3	M16174.1/3-119	117	30.954639	-49.2	-44.000799	-47	0.0638128	6.3812796	-45.6	0.0350702	3.5070206
4	X55267.1/3-119	117	30.758514	-49.2	-43.697708	-43.2	-0.011521	-1.1521021	-42.2	-0.0354907	-3.5490714
5	M16172.1/3-119	117	30.856732	-49.9	-43.849494	-42.3	-0.0366311	-3.6631068	-42.3	-0.0366311	-3.6631068
6	AF001265.1/6033-6149	117	30.561132	-48.8	-43.392673	-41.6	-0.0430931	-4.3093092	-41.6	-0.0430931	-4.3093092
7	X05057.1/3-119	117	30.561132	-49.6	-43.392673	-42	-0.0331589	-3.3158872	-42	-0.0331589	-3.3158872
8	X15126.1/3-120	118	33.460848	-52.2	-48.022394	-42.8	-0.1220186	-12.201856	-42.7	-0.1246462	-12.464623
9	Z50737.1/3-119	117	34.214221	-52.9	-49.038157	-44.5	-0.1019811	-10.198106	-44.5	-0.1019811	-10.198106
10	X55261.1/3-119	117	31.084701	-50.4	-44.201797	-44.9	0.0155502	1.5550188	-44.1	-0.0023083	-0.2308312
snoRNA - Rfam RF00012 family											
1	ABGA01262676.1/1340-1125	216	41.906208	-85.9	-79.841099	-83.1	0.0392166	3.9216613	-81.3	0.0179447	1.7944656
2	AANU01105435.1/2236-2452	217	41.326388	-83.1	-78.978957	-83.4	0.0530101	5.3010113	-80.4	0.0176747	1.767467
3	AAGV020566442.1/1857-1640	218	42.747895	-86	-81.827471	-84.2	0.0281773	2.8177301	-80.5	-0.0164903	-1.6490327
4	ABVD01644074.1/1753-1970	218	42.653788	-81.2	-81.652988	-76.6	-0.0659659	-6.5965903	-72.7	-0.1231498	-12.314977
5	CABF01044723.1/2170-2355	186	38.260558	-66.4	-66.6947	-65.1	-0.0244962	-2.4496158	-69.4	0.0389813	3.8981269
6	AAGD02005452.1/2742-2555	188	38.547422	-70.6	-67.652375	-65.1	-0.039207	-3.9206991	-67.1	-0.0082321	-0.8232118
7	AY948622.1/2-186	185	37.917961	-66.8	-65.846591	-60	-0.0974432	-9.744318	-60.2	-0.0937972	-9.379719
8	CAAC02000606.1/1368922-1368738	185	37.79845	-69.3	-65.625006	-61.2	-0.072304	-7.2304023	-63.6	-0.0318397	-3.1839721
9	ABKE01003568.1/5923-6108	186	38.062963	-67	-66.328341	-66.3	-0.0004275	-0.042746	-66.3	-0.0004275	-0.042746
10	AAQA01000362.1/29408-29193	216	42.181005	-92.5	-80.350602	-91.8	0.1247211	12.472111	-93.4	0.1397152	13.971518
miRNA Rfam family RF00706											
1	AALT01209640.1/567-377	90	26.605736	-35.4	-35.356053	-37.4	0.054651	5.4650996	-33.6	-0.0522635	-5.2263475

2	AAFR03033875.1/20528-20718	91	26.925205	-39.7	-36.355838	-34.8	-0.044708	-4.4708001	-37.7	0.0356542	3.5654153
3	AAIY01044029.1/787-597	90	25.07667	-30.9	-28.917158	-33.5	0.1368013	13.680125	-29.7	0.0263583	2.6358314
4	AAZO01007389.1/15370-15178	93	26.505742	-30.5	-33.898479	-32.9	-0.0303489	-3.0348896	-33.6	-0.0088833	-0.8883294
5	AAYZ01695118.1/310-500	90	25.893038	-33.7	-32.354885	-35.7	0.0937007	9.3700707	-35.7	0.0937007	9.3700707
6	AAHX01044404.1/26102-26292	93	25.205573	-29.7	-28.423468	-35.1	0.1902146	19.021459	-31.3	0.091902	9.1901985
7	AACN010750078.1/657-848	91	26.849995	-35	-36.039128	-36	-0.0010869	-0.1086895	-36.4	0.0099141	0.991406
8	ABAV01019481.1/5988-6180	90	26.870476	-33	-36.470874	-39.8	0.0836464	8.3646388	-36	-0.0130798	-1.3079827
9	AAZX01018356.1/721-913	89	27.493138	-36.5	-39.438404	-43.1	0.0849558	8.4955816	-43.3	0.0891823	8.9182348
10	AY765362.1/650-458	90	26.376701	-35.9	-34.391586	-38	0.0949583	9.4958252	-38.4	0.1043858	10.438577
Rfam snRNA family RF00004 (208)											
1	AALT01209640.1/567-377	191	37.914724	-63.1	-62.235101	-56.2	-0.1073861	-10.738613	-58.1	-0.0711721	-7.1172129
2	AAFR03033875.1/20528-20718	191	37.675344	-65.8	-61.854176	-60.1	-0.0291876	-2.9187614	-61.8	-0.0008766	-0.0876628
3	AAIY01044029.1/787-597	191	37.80852	-63.7	-62.066098	-60.3	-0.0292885	-2.9288531	-61.3	-0.0124975	-1.2497527
4	AAZO01007389.1/15370-15178	193	38.609895	-70.3	-63.740525	-71.2	0.1047679	10.47679	-72.6	0.1220313	12.203133
5	AAYZ01695118.1/310-500	191	38.257851	-63.4	-62.781118	-61.3	-0.0241618	-2.4161798	-61.6	-0.019174	-1.9173997
6	AAHX01044404.1/26102-26292	191	37.781923	-65.2	-62.023774	-61.4	-0.0101592	-1.0159179	-62.8	0.0123603	1.2360293
7	AACN010750078.1/657-848	192	38.309842	-66.9	-63.063451	-62.2	-0.0138819	-1.3881853	-62.7	-0.0057967	-0.5796671
8	ABAV01019481.1/5988-6180	193	38.570822	-64	-63.67835	-68.9	0.0757859	7.5785928	-70.5	0.096761	9.6761
9	AAZX01018356.1/721-913	193	45.570822	-79.3	-74.81745	-78.1	0.0420301	4.2030096	-78.9	0.0517434	5.1743352
10	AY765362.1/650-458	193	38.830561	-70	-64.091672	-67.3	0.047672	4.767203	-67.3	0.047672	4.767203

Web-servers used : RNAfold and RNAstructure. Source of sequences : Rfam

Table 6.8 given in the Appendix I of the thesis shows the above shown results from the calculations for 902 snRNA sequences analysed in this work. The maximum relative discrepancy in the values of MFE found out using RNAfold and RNAstructure webservers was found to be 25.36% for snRNA sequence with sequence identifier X69327.1/1-196 belonging to Rfam family RF00004.

Deviation in the value of MFE_C found out in relation to MFE_F and MFE_S, for 902 snRNA sequences is shown in Table 6.7 as 'RD_2' and 'RD_3' respectively. The percentage deviations are also given, indicated by RD_2 and RD_3 in columns 6 and 9 respectively. The values of relative deviations of MFE_C computed with the novel model in comparison to MFE obtained from standard webservers RNAfold and RNAstructure, MFE_F and MFE_S respectively, are given in Table 6.9.

The details shown in Table 6.9 can be summed up as follows:

- 41.695% and 43.692% of the sequences showed a deviation of 0 to ±5% when the MFE values obtained with the proposed model are compared with those obtained with RNAfold and RNAserver respectively.
- Similarly, the proposed model showed a relative deviation of ±5% to ±10% for 33.51% and 34.61% of the sequences in the three comparisons in the order mentioned above.
- 18.58%, 15.044% of the sequences had ±10% to ±15% deviation when the MFE values from the model were compared with the ones obtained using RNAfold and RNAserver respectively.
- Deviations above ±15% for MFE values were shown only by about 6.21% and 6.65% of the sequences in the comparisons with RNAfold and RNAstructure.

Table 6.9. Percentage deviations of MFE values computed with the novel model for 902 snRNA sequences, relative to MFE values computed with RNAfold and RNAstructure

Sl. No.	Percentage Deviation	Percentage of sequences for which the proposed model gives relative deviation:			
		from 0 to ± 5%	from ± 5% to ± 10%	from ± 10% to ± 15%	above ± 15%
1	Relative deviation 2 (RD_2) (comparison with MFE from RNAfold)	41.69422986	33.51327434	18.5840708	6.208425
2	Relative deviation 3 (RD_3) (comparison with MFE from RNAstructure)	43.69211	34.6121107	15.04424779	6.6518847

6.5. Discussion

Recent advancements in molecular biology have brought to the forefront the importance of ncRNA in regulating numerous functions of the cell. Understanding the structure of RNA is one of the keys to understanding its function. Length and minimum free energy of sequences are also common indices used to study RNA. In this work, the minimum free energy of ncRNA sequences, which decides the optimal secondary structure, was analyzed with respect to its relationship to the sequence length and the standard deviation of spectral coefficients.

As already seen in the introduction of this chapter, MFE and sequence length are vital parameters to be analyzed in the study of RNA. Computational methods have been widely employed to study noncoding RNA. Even though DSP methods have become as popular as computational methods in the analysis of genomic data, little work has been done which makes use of Digital Signal Processing techniques to analyze the noncoding genome. Though sequence length and MFE have been used extensively in analysing RNA, a mathematical relationship linking MFE to the length or any other signal property of the sequence has not been reported in literature till date. Here in this work we have introduced a novel approach, which links MFE, a thermodynamic property of ncRNA sequences to their signal properties.

The sequences studied in this chapter were taken from the Rfam database. More than 2600 noncoding RNA sequences belonging to four classes viz snRNA, snoRNA, rRNA and miRNA across different organisms were analyzed. Sequences having zero value for MFE, even though were considered in the analysis, they contribute to gross outliers and do not alter the results of the regression analysis. Only a total number of seven (five in snRNA, one in rRNA and one in snoRNA) sequences were found, having zero as MFE out of the database of over 2600 ncRNA sequences studied. The results of computation of MFE using the algorithm for the five snRNA sequences with 0 MFE are shown in Table 6.9 on page 384 in Appendix I. The value of RD_1, RD_2 and RD_3 cannot be calculated as the reference values, MFE_M, MFE_F, and MFE_S come in the denominator in calculations.

A novel mathematical model linking MFE, sequence length and standard deviation of spectral coefficient matrix was developed for all the classes of noncoding RNA analyzed and MFE was computed using this model. The performance of the models developed here for the four classes of ncRNA analyzed was checked for accuracy with standard web servers, RNAfold and RNA structure.

The main findings of this study presented in this chapter can be summarized as follows.

- It was found that the MFE values computed with the proposed model was in concordance with those obtained from the web servers.
- The time of computation was comparable with that of RNAfold webserver, which is faster than RNAstructure webserver.
- Upon comparing the MFE values obtained using the model with that of webservers, the relative deviations of MFE values obtained with proposed models for all four category of ncRNA sequences analysed were found to be ***within 0 to ± 5% for about 45%*** of the sequences; ***within ± 5% to ± 10% for about 35%*** of the sequences; ***between ± 10% to ± 15% for 15% of the sequences***. Only around *5%* of the sequences gave relative deviation percentages *above +/-15%* in the comparisons. This shows the accuracy of the model. In this context, it is to be noted that a maximum discrepancy of 25.3666% was observed between MFE values calculated using RNAfold and RNAstructure webservers for Rfam snRNA sequence with id X69327.1/1-196.

At this point, certain facts regarding MFE and secondary structure is to be mentioned. At room temperature, RNAs exist in an ensemble of structures and the MFE structure is not always the biologically relevant one [Washietl 2012], [Hofacker 2002]. There are several algorithms to predict these sub-optimal secondary structures [Wuchty 1999], [Zuker 1989], [McCaskill 1990]. Most of the common secondary structure prediction methods assume that the functional RNA structure depends solely on the thermodynamic equilibrium and does not consider the kinetics of folding. The impact of the kinetics of folding on the functional structure of RNA is not fully known [Washietl 2005]. However, in examples like RNA switches, kinetics of folding is significant and there are studies which analyze this aspect [Chen 2008], [Wolfinger 2004]. A sequence may fold into reliable structures other than the MFE structure or switch between structures as a consequence of energy fluctuations in the range of a few kT , where k is the Boltzmann constant and T the absolute tempetature [Fontana 2002]. This energy range is around 3 kcal/mol at 37°C. Secondary structure is also predicted based on the ensemble, making use of McCaskill's algorithm [McCaskill 1990]. The probability of a particular base pair in the thermodynamic ensemble is found out using a partition function over all possible structures, computed with the algorithm [Ding 2006]. Secondary structure prediction has also been performed by identifying a ‘centroid structure’ which is thought to represent the ensemble [Ding 2005]. In this work, we have considered only one structure from the ensemble, viz. the MFE secondary structure. The accuracy of the model examined here pertains only to the MFE structure from the ensemble of structures.

The accuracy of MFE based secondary structure prediction depends on the type of RNA. Generally, it can be assumed that only two-thirds of the actual base-pairs are predicted correctly while one-third of the true base pairs are missed [McCaskill 1990], even with the best of currently available prediction methods. In addition, all MFE based structure prediction approaches give only a rough model of the RNA structure. Base pairing possibilities are described by the Shannon entropy introduced by Huynen et al. (1997) [Huynen 1997]. Shannon entropy is a measure of how well defined the RNA structure for a given sequence is.

Mathematically, the average S value for a sequence is given by

$$S = -\sum_{i,j} P_{i,j} \log\left(P_{i,j}\right)/N \tag{6.35}$$

for all $1 \leq i \ \leq j \leq N$.[43] Where, N is the length of the sequence and $P_{i,j}$ is the probability of base *i* pairing with base *j*. Well defined structures are said to have lower Shannon entropy than those which have many alternate structures (alternate/competing base pairs) [Mathews 2004]. Hence Shannon entropy has been used to pick the most probable structure form the Botlzmann ensemble [Ding 2001], [Ding 2003]. The value of *S* is directly linked to *N* as shown in the above equation. Shannon entropy increases with the logarithm of the length *N* of the sequence and starts to saturate at a sequence length of 500 [Mathews 2004]. The mathematical models developed here link MFE linearly to the length of the sequence as well as to the standard deviation of spectral coefficients. The spectral coefficients are computed after performing mathematical mapping of the sequence string as already explained, the value of which depends only on the bases in the sequence and base-pairing is not considered. Shannon entropy is not the sole indicator to the correctness of base-pairs predicted in the MFE structure [Huynen 1997]. As Shannon entropy is not directly linked mathematically to MFE, a direct mathematical relationship between Shannon entropy and spectral coefficient matrix cannot be made within the confines of this study. However, shorter sequences have lower values for S [Huynen 1997] and have stable structures. It was found in this work that shorter sequences have lower values of SD of spectral coefficient matrix. So we could say that shorter sequences have lower Shannon entropy, lower values of SD_SCM, lower MFE and form the more stable structures in the ensemble.

As already mentioned, no MFE based secondary structure prediction algorithm ensures fool proof structures as base pairs may be missed or wrongly predicted. The authors do not claim that this is the perfect method for computing MFE. Nevertheless, the technique presented here is computationally simple and it is the first of its kind that links a thermodynamic quantity with the signal properties of the sequence. Signal processing techniques have the inherent property of computational simplicity and easiness of implementation. Genomic sequences possess more signal properties and there are varieties of DSP tools that can be put to use to analyze them. Researchers should explore noncoding RNA using DSP techniques and this work should be considered as an initial step in the direction.

6.6. Conclusion

Over 2600 noncoding RNA sequences belonging to four classes viz. snRNA, snoRNA, rRNA, miRNA were analyzed in this work as regards the relationship between their MFE and signal parameters. Synthetically generated oligonucleotide sequences too are a part of the database used in the study.

Novel mathematical models linking MFE with the signal properties of ncRNA sequences of these four classes was arrived at. Only about 5% of the sequences showed relative deviations above +/15% when MFE values obtained with the model were compared with those obtained using conventionally accepted methods. This shows the accuracy of the models developed. Thus the mathematical models are specific to the ncRNA classes studied and represent them aptly. It is not claimed that the model developed here is the perfect method to compute MFE. At the same time the easiness with which one can compute MFE just by knowing the sequence length and the sequence spectrum cannot be overlooked. This work brings to light the relationship between the thermodynamic entity MFE and the signal properties of the sequence. This shows that the noncoding genome too is conducive to analysis with DSP techniques. Digital signal processing methods have the unique convenience of ease of implementation and lesser computational complexity. It is hoped that this novel relationship linking MFE with signal properties of the sequences can be taken forward so that more signal processing approaches to study noncoding RNA evolve.

Chapter 7

Exon Mapping in lncRNA Using Digital Filters

Long noncoding RNAs (lncRNA) which were initially dismissed as "transcriptional noise" have become a vital area of study from the beginning of this decade after their roles in biological regulation were discovered. Long ncRNAs have been implicated in various developmental processes and diseases. Findings of recent studies emphasize the need for in-depth study of sequence, structural features, and genomic architecture of lncRNA. In this work, we perform exon mapping of human lncRNA sequences (taken from NCBI GenBank) using digital filters. The exon locations obtained here conform to the ranges specified in GenBank.

Abstract

Long noncoding RNAs (lncRNA) which were initially dismissed as “transcriptional noise” have become a vital area of study after their roles in biological regulation were discovered. Long ncRNAs have been implicated in various developmental processes and diseases. Findings of recent studies emphasize the need for in-depth study of the sequence, structural features and genomic architecture of lncRNA.

In this work, we perform exon mapping of human lncRNA sequences (taken from NCBI GenBank) using digital filters. Digital anti-notch filters are used to map out the exons of the lncRNA sequences analysed. The period 3 property which is an established indicator for locating exons in genes is used here. Discrete wavelet transform filter bank is used to fine-tune the exon maps by selectively removing the spectral noise. The exon locations conform to the ranges specified in GenBank. In an earlier work, a quadratic filter was successfully used by the authors to bring down the spectral noise while mapping exons of coding regions. However, it is found that this quadratic function introduces additional spectral noise when used with lncRNA sequences. This indicates that the sequence spectrum of lncRNAs cannot be amply represented by the A-T spectra alone as in protein coding genes. The spectral noise in the exon map of lncRNA occupies the same frequency ranges as that of coding regions and hence the de-noising techniques used for exon prediction in genes can be extended to lncRNAs too. As reported in literature, G-C concentration in lncRNA sequences is seen to be less than 50%, which is much lower than that found in coding regions. It is seen that none of the sequences analysed have STOP codons although different START codon patterns are found in them. This leads to the logic that the exons present in lncRNA sequences do not have coding potential. The function of these regions is yet to be analysed.

7.1 Introduction

Long noncoding RNAs (lncRNAs) constitute a heterogenic class of RNAs that include intergenic lncRNAs, antisense transcripts, and enhancer RNAs etc. Long ncRNAs generally refer to those sequences that are more than 200 nucleotides in length [Ponting 2009], [Mercer 2009], [Brosnan 2009]. Defining lncRNAs by the virtue of what they are not, viz. neither short nor protein coding is rather inapt [Ponting 2009]. Nevertheless the current imperfect level of understanding of their functions makes such a categorization practical. Long noncoding RNAs were called so primarily to distinguish them from small non coding RNAs. They could be categorized based on their diverse empirical features, viz. genomic context [Kung 2013], origin of transcription, tissue specificity, molecular function, or mechanism of action. For example, based on the genomic context, we could categorize lncRNAs as “stand-alone” sequences these are lincRNAs (ling intergenic/intervening lncRNAs) which are transcription units that do not overlap protein-coding genes [Cabili 2011], [Ulitsky 2011]. There are natural anti-sense transcripts [Kanduri 2006], pseudo-genes [Pink 2011], long intronic RNAs [Louro 2009], divergent transcripts, promoter-associated transcript [Kanhere 2010] and enhancer RNAs [Kim 2010]. Based on the origin of transcription we could have the following different categories: lncRNAs transcribed from intergenic regions are called long intervening noncoding RNAs, those transcribed from within introns of protein-coding genes are called intronic lncRNAs, those transcribed from the antisense strand of a given gene are called natural antisense transcripts and so on [Ma 2013]. However, their classification is not standardized. For example defining a transcript, or its locus, as being coding or non-coding is unsatisfactory simply because of the inherent contrariness. Very often human genes possess both coding and non-coding transcripts which are difficult to distinguish without detailed experimental studies. It is equally difficult to label a transcript as being “intergenic” [Ponting 2010].

In this context it also needs to be mentioned that many methods attempting to classify RNAs into protein coding and noncoding have come up. Some ncRNA sequences could actually code for peptides and some which are thought to be coding RNAs might not be so. Besides, protein-coding and noncoding transcripts often overlap as already mentioned. Such factors make it practically impossible to classify RNAs under this feature. RNAs cannot be unequivocally classified as being protein coding or non-protein coding [Dinger 2008]. The functionality of any transcript at the RNA level should not be discounted. Hence the very name ‘long noncoding RNA’ is not always truly descriptive of the function of a sequence.

Long noncoding RNAs of all kinds have been implicated in a range of developmental processes and diseases [Wapinski 2011], [Harries 2012], [Chen 2014], [Chen 2016], [Fang 2016], [Smola 2016] but knowledge of the mechanisms by which they act is still surprisingly limited. At the same time, there are a small number of lncRNAs which have been well-studied from which we have been able to deduce important clues about the biology of these molecules. For example, metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) and myocardial infarction-associated transcript (MIAT) were shown to affect endothelial cell functions, whereas lincRNA-p21 controls neointima formation [Boon 2016]. The Xist long noncoding RNA has been found to be essential in X-chromosome inactivation during female eutherian mammalian development [Calabrese 2013], [Smola 2016]. However it is to be noted that the functions/involvement of lncRNAs are not limited to the ones mentioned here. The same lncRNA can be implicated in more than one disease/function.

In the vertebrate genomes studied so far, thousands of genes encoding long noncoding RNAs (lncRNAs) have been identified [Kapusta 2014]. The human genome consists of many thousands of lncRNA. Analyses show that human lncRNAs are generated through pathways similar to that of protein-coding genes, with similar histone-modification profiles, splicing signals and intron/exon lengths [Derrien 2012]. It has also been seen that lncRNAs exhibit a striking bias towards two-exon transcripts unlike protein-coding genes [Derrien 2012]. Expression analyses have shown that lncRNAs are generally lower expressed than protein-coding genes. They show positive correlation with the expression of anti-sense strand of coding genes [Derrien 2012], [Harrow 2012].

Recent studies suggest the need for in-depth study of the sequence, structural features and genomic architecture of lncRNA [Niazi 2012]. There are quite a few number of computational methods in literature which analyse long noncoding RNA sequences. Signal et.al. [Signal 2016] describe a computational method for functional prediction and characterisation of long noncoding RNA. Core features of functional lncRNAs are probed via an array of computational methods. Long noncoding RNA function is also predicted by using tissue specific evolutionary conserved expression as done by Perron et. al. [Perron 2017]. These authors make use of the ‘guilt-by-association’ principle which is explained as follows. If a long ncRNA gene shows an expression profile that correlates with the expression profiles of a set of coding genes involved in a known function, then the lncRNA gene analysed probably is involved in the same function. Zhao et.al. [Zhao 2014] describe prediction of lncRNA function using a co-expression network which is described to be useful in large-scale annotation of long ncRNA function and is based on a non-coding gene co-expression network. The nodes in the network correspond to protein-coding gene or lncRNA and the edges connecting the nodes denote whether they are co-expressed. Functions of lncRNAs across multiple cancers are explored through co-expression networks by Li et.al. [Li 2017]. Weighted correlation network analysis is made use of to express the functions of lncRNAs altered in more than two cancer types. The authors conclude that the lncRNAs expressed in cancers show high tissue-specificity and are weakly expressed than protein-coding genes.

Though there are many computational methods to analyse the long ncRNA, DSP based methods which analyse lncRNA were not found in literature. Digital signal processing methods inherently have simplicity of implementation and ease of use. In the work presented in this chapter, we focus on IncRNA, typically said to be those with more than 200 nucleotides [Mercer 2013], [Kung 2013], a heterogeneous group of sequences which are implied in diseases and cell development. The aim of this work is to study lncRNA sequences using digital signal processing techniques and search for similarity/differences they have with coding genes as regards the signal/spectral properties of their sequences. Here, we apply digital filtering technique to map out the exons in lncRNA sequences taken from the benchmarked, public database (NCBI Genbank). Period 3 property which is an established feature [Tiwari 1997], [Trifonov 1980], [Li 1997] in locating exons in coding regions is made use of here. Long ncRNAs have been found to have low values of GC concentration [Niazi 2012] which is considered to be one of the reasons for their lack of protein coding capability. In this work, G-C content of the sequences in percentage relative to the net nucleotide content is computed. In this study, the sequences are also searched for START codon (both AUG and the alternate START codons), and the STOP codon patterns.

7.2 Locating exons within lncRNA sequences

First we will have a brief look at the specimen used in the study reported in this chapter.

7.2.1 Specimen used in the study

In this work we make use of human lncRNA sequences which are available in the NCBI GenBank. Only sequences of length more than 200 nucleotides are considered. The lncRNAs analyzed in this work is random, assorted list. It includes stand alone lncRNAs (e.g. MALAT1), natural anti-sense transcripts (BACE-AS1, FOXC2-AS1), lncRNAs implicated in diseases (CAHM, CCEPR) and so on. The list of sequences used in this work in Table 7.1. Column 2 of Table 7.1 gives the name of the lncRNA, column 3 has the NCBI GenBank accession number, and column 4 gives the location of the sequence in the UCSC (University of California Santa Cruz) Genome Browser [Genome Browser] and column 5 gives a brief description of the sequence as found in the corresponding NCBI record. More details about the sequences can be had from the NCBI website [NCBI Website]. This part of the work has been reported by the authors [George 2017].

7.2.2. Brief overview of long noncoding RNA sequences used

Table 7.1. lncRNA sequence list

Sl. No.	Name of the lncRNA	GenBank Accession No.	Location in Genome Browser	GenBank information of the sequence
1	CCEPR	NR_131782.1	chr6:163413065 - 163413950	Homo sapiens cervical carcinoma expressed PCNA regulatory lncRNA (CCEPR)
2	BACE1-AS	NR_037803.2	chr11:117,291,346-117,292,170	Homo sapiens BACE1 antisense RNA (BACE1-AS), antisense RNA
3	CAHM	NR_037593.1	chr6:163,413,065-163,413,950	Homo sapiens colon adenocarcinoma hypermethylated (non-protein coding) (CAHM)
4	BGLT3	NR_121648.1	chr11:5,244,554-5,245,546	Homo sapiens beta globin locus transcript 3 (non-protein coding) (BGLT3), long non-coding RNA
5	ABALON	NR_131907.1	chr20:31,721,507-31,723,409	Homo sapiens apoptotic BCL2L1-antisense long non-coding RNA (ABALON)
6	DISC2	NR_002227.2	chr1:231,814,626-231,818,517	Homo sapiens disrupted in schizophrenia 2 (non-protein coding) (DISC2), long non-coding RNA
7	GHET1	NR_130107.1	chr7:148,987,527-148,989,429	Homo sapiens gastric carcinoma proliferation enhancing transcript 1 (GHET1), long non-coding RNA

8	HEIH	NR_045680.1	chr5:180,829,954-180,831,618	Homo sapiens hepatocellular carcinoma up-regulated EZH2-associated long non-coding RNA (HEIH)
9	NEAT1	NR_028272	chr11:65,422,798-65,426,532	Homo sapiens nuclear paraspeckle assembly transcript 1 (NEAT1), transcript variant MENepsilon, long non-coding RNA
10	MALAT1(TV1*)	NR_002819.4	chr11:65,497,679-65,504,494	Homo sapiens metastasis associated lung adenocarcinoma transcript 1 (MALAT1), transcript variant 1, long non-coding RNA
11	NKILA	NR_131157.1	chr20:57,710,183-57,712,780	Homo sapiens NF-kappaB interacting lncRNA (NKILA), long non-coding RNA
12	FOXC2-AS1	NR_125795.1	chr16:86,565,145-86,567,761	Homo sapiens FOXC2 antisense RNA1 long noncoding RNA
13	DLEU1(TV2**)	NR_002605.2	chr13:50082169-50107218	Homo sapiens deleted in lymphocytic leukemia 1 (DLEU1), transcript variant 2, long non-coding RNA
14	HULC	NR_004855.2	chr6:8,652,209-8,653,846	Homo sapiens hepatocellular carcinoma up-regulated long non-coding RNA (HULC)
15	KIAA0087	NR_022006.1	chr7:26,533,121-26,538,825	Homo sapiens KIAA0087 lncRNA (KIAA0087), long non-coding RNA
16	MHENCR	NR_132417.1	chr20:63,627,235-63,628,824	Homo sapiens melanoma highly expressed competing endogenous lncRNA for miR-425 and miR-489 (MHENCR), transcript variant 1, long non-coding RNA

17	FALEC	NR_051960.1	chr1:150,515,757-150,518,032	Homo sapiens antisense of IGF2R non-protein coding RNA (AIRN), transcript variant 1
18	PRNT	NR_024267.1	chr1:150,515,757-150,518,032	Homo sapiens prion protein (testis specific) (PRNT), transcript variant 1, long non-coding RNA
19	HOTAIRM1	NR_038366.1	chr7:27,096,094-27,100,258	Homo sapiens HOXA transcript antisense RNA, myeloid-specific 1 (HOTAIRM1), transcript variant 1, long non-coding RNA
20	CISTR(TV1*)	NR_104332.1	chr12:53,750,447-53,757,034	Homo sapiens chondrogenesis-associated transcript (CISTR), transcript variant 1, long non-coding RNA

TV1* : Transcript variant 1; TV2* : Transcript variant 2

7.2.3 Exon prediction

The period 3 property which is an established digital signal processing (DSP) method to detect protein coding regions in genes [Vaidyanathan 2002], [George 2010] and in gene detection [Anastassiou 2002], [Tiwari 1997], [Kakumani 2008] is used here. The base sequences in the coding regions (exons) of genes exhibit a strong period 3 component. This was observed by Trifonov and Sussman [Trifonov 1980] as early as 1980. They maintain that this is due to the non-uniform codon usage in the formation of amino acids. Even though there are several codons that could possibly code a given amino acid, they are not used with uniform probability and this creates a codon bias. There is an excess Guanine in position 1, which leads to a strong period 3 oscillation [Herzela 1998]. There are other authors [Tiwari 1997] who think this explanation is rather incomplete. But all authors do agree to the fact that the spectrum of protein coding DNA has a peak at every third component (ie. at frequency k = N/3, in a sequence of length N) and this property still remains widely accepted in predicting exons in eukaryotic coding regions.

It is to be noted that such periodicity was observed two decades ago in noncoding regions for procaryotes, and some viral and mitochondrial base sequences [Li 1997]. In this work we map out the exons in lncRNA sequences using the period 3 property. Algorithms which exploit the period 3 property proceed by computing the discrete Fourier transform (DFT) [Proakis 2006], [Oppenheim 2009] which is expected to exhibit a peak at frequency $2\pi/3$ in the spectrum. From the spectrum the component at frequency $\omega = 2\pi/3$ can be located by using a sharp single frequency peaking filter.

7.2.3.1. Spectrum of the lncRNA sequences

The mathematical mapping of the sequence string $x[n]$ is done making use of binary indicator sequences [Anastassiou 2001]. $u_a[n]$, $u_u[n]$, $u_c[n]$, $u_g[n]$ are the binary indicator sequences corresponding to A, U, C G which take on a value of 0 or 1 at location n, depending on whether the corresponding character exists or not at n such that,

$$u_a[n] + u_u[n] + u_c[n] + u_g[n] = 1 \qquad (7.1)$$

DFT of a sequence y[n], of length N, is itself another sequence $Y[k]$, of the same length N, expressed mathematically as,

$$Y(k) = \sum_{n=0}^{N-1} y(n) e^{-(jk2n\pi)/N}$$

(7.2)

For k = *0,1,2,3..... (N-1)*. DFTs of individual indicator sequences $u_a[n]$, $u_u[n]$, $u_c[n]$, $u_g[n]$, $(U_a(k), U_u(k), U_c(k), U_g(k)$, respectively) are computed as per the equation (6.2) and the power spectrum is obtained as follows.

$S(k) = |U_a(k)^2| + |U_u(k)^2| + |U_c(k)^2| + \ U_g(k)^2$ (7.3)

We make use of sliding overlapping windows for better time resolution and compute the STFT (short time Fourier transform). The length of the window has to be a multiple of 3. In a former work, we have found that the window length should be selected based on the length of the sequence used for optimum results [George 2010].

Due to the period 3 property we expect a peak in the spectrum at frequency 2π/3 as seen in Figure 7.1.

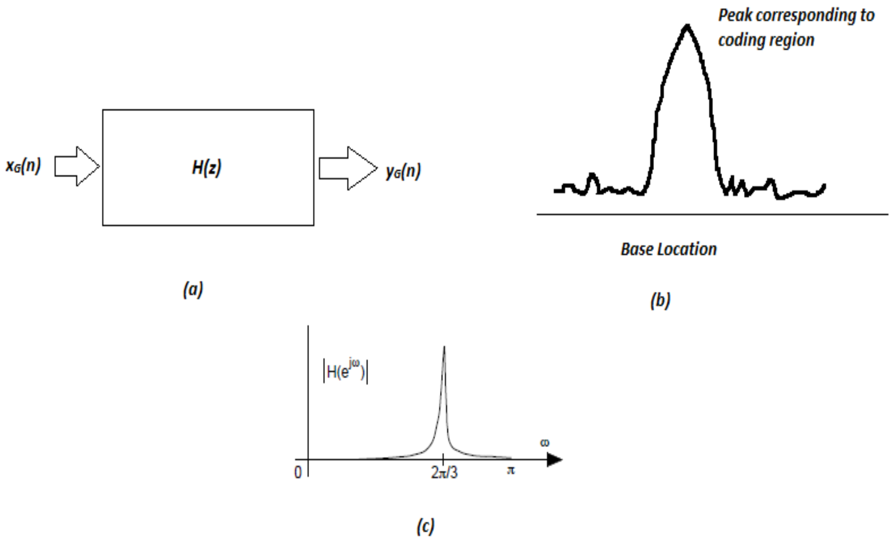


Figure 7.1. Expected O/P of the single peaking/anti-notch filter

Expected output of anti-notch filter show in Figure 7.3(c). x_G(n) – indicator sequence, $H(z)$ –anti-notch filter with pass band centred at 2π/3, y_G(n) - output of the filter. This peak was and detected using a lattice implementation of a digital IIR anti-notch filter by Vaidyanathan and Yoon [Vaidyanathan 2002].

7.2.3.2. The IIR single peaking filter

In this work, we have used a single peaking IIR filter designed using the in-built filter design utility of the platform MATLAB 2016a.

It is a direct form II, transposed, stable filter of order 2 with very high Q factor [Proakis 2006], [Oppenheim 2009]. The general form of the direct form II filter is given in Figure 7.2 and the general transfer function for IIR Direct form II implementation is

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 + \sum_{k=1}^{N} a_k z^{-k}} \tag{7.4}$$

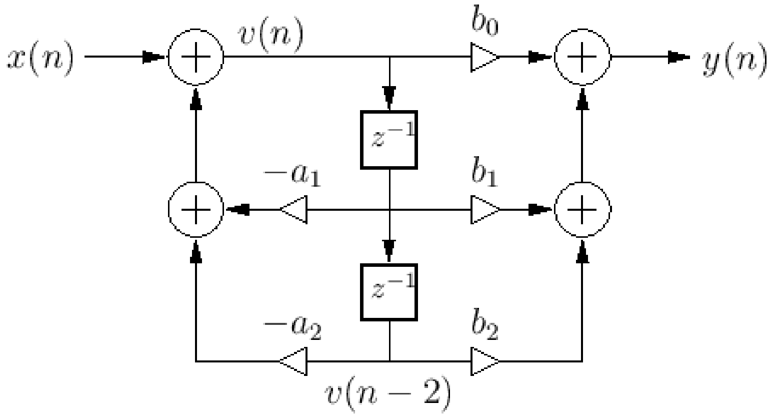


Figure 7.2. General form of the Direct Form II filter

For order 2, in equation 7.4, M=N=2.

The magnitude and phase responses of the filter are given in Figure 7.3 and the pole-zero plots given in Figure 7.4.

The filter coefficients of the IIR single peaking design:

Numerator: [2.05798×10^{-10}, 0, 2.05798×10^{-10}]

Denominator: [1, 0.99999, 0.999999]

This IIR single peaking filter [George 2010] gives a far better result than the IIR anti-notch filter. The subsequent filter bank [Vaidyanathan 2002] is not needed for removal of noise. This can be attributed to the high attenuation in the stop band of the peaking filter. Next, we see how the accuracy of an exon detection algorithm can be improved using the Discrete Wavelet Transform (DWT).

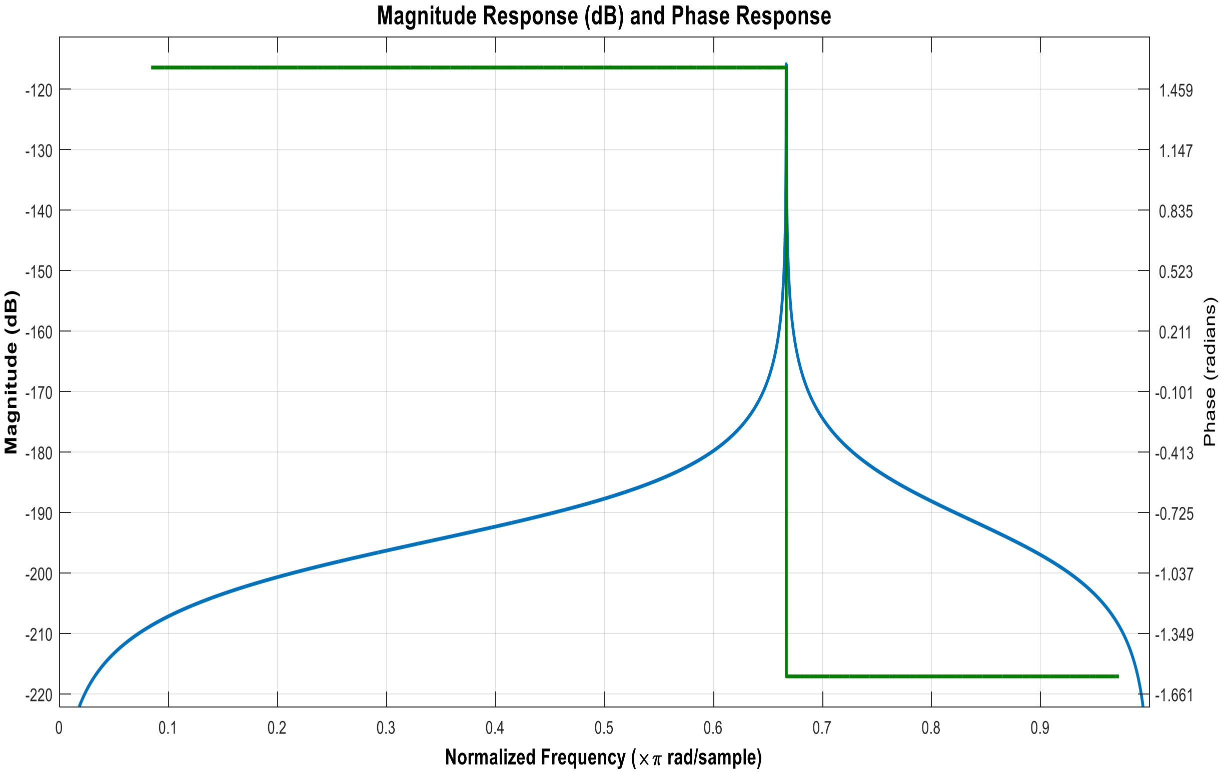


Figure 7.3. Magnitude and phase response of IIR single peaking filter at 2π/3.

The magnitude response is shown in blue and phase response is in green

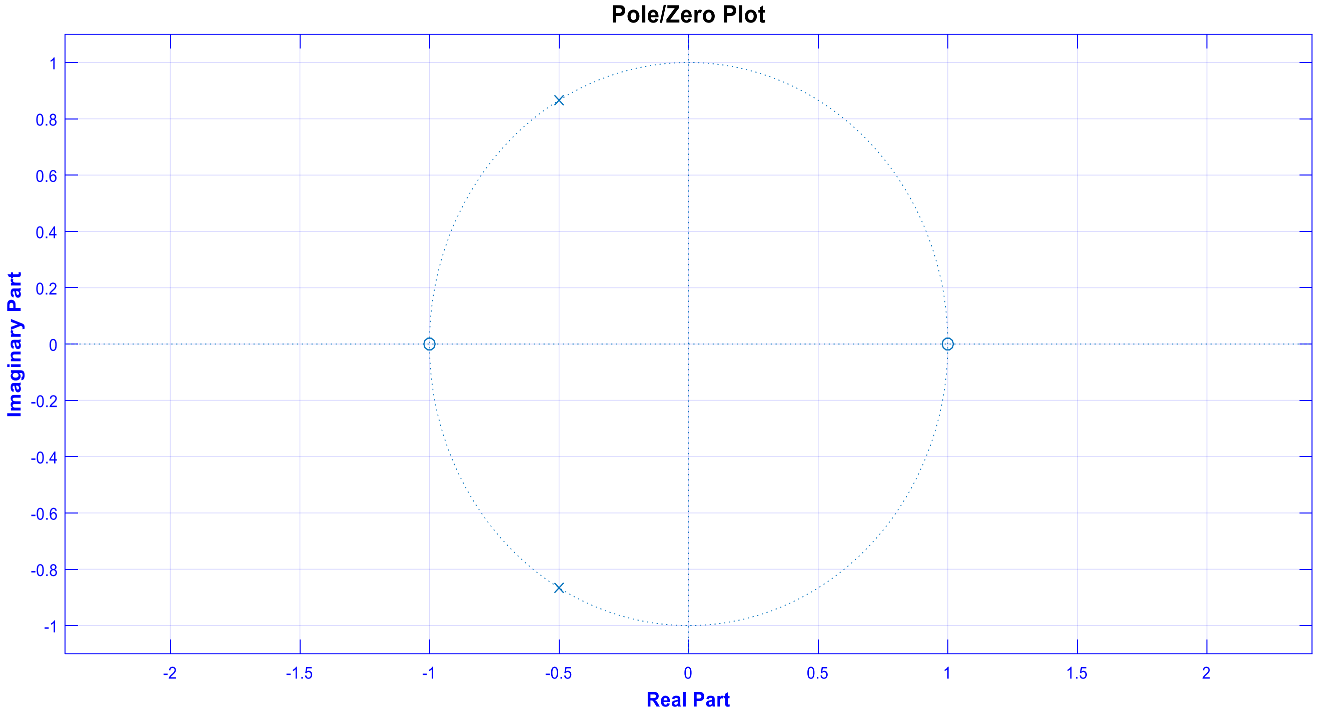


Figure 7.4. Pole-zero plot of the IIR single peaking filter

7.2.3.3. Noise removal using the DWT filter bank

The exon map is improved by de-noising the exon plot with the DWT (discrete wavelet transform) [Soman 2004], [Mallat 2009]. The discrete wavelet transform is a digital filter bank which performs sub-band coding. A simple schematic representation is shown in Figure 4. The signal is passed through a filter bank consisting of low and high pass filters followed by scaling. The scale is altered by upsampling and downsampling or subsampling operations. Subsampling reduces sampling rate while upsampling increases it. The incoming signal is split into two frequency-specific halves. The low frequency half (LF) g(n) and the high frequency half (HF) h(n).

7.2.3.3.1. Basics of wavelets

Wavelets are functions that satisfy certain mathematical requirements and are used in representing data or other functions. Approximation using superposition of functions has existed since the early 18OOs, when Joseph Fourier discovered that he could superpose sines and cosines to represent other functions. However, in wavelet analysis, the scale through data is viewed makes the analysis different. The fundamental idea behind wavelets is to analyze according to scale. Wavelet algorithms process data at different scales or resolutions. If we look at a signal (or a function) through a large “window,” we would notice gross features. Similarly, if we look at a signal through a small “window,” we would notice small features. The intention in wavelet analysis is to see both the forest and the trees, so to speak. Sines and cosines have been used in signal analysis as basis functions for a long time (Fourier analysis). Both sines and cosines are nonlocal (stretch out to infinity) hence, they are rather inadequate to approximate sharp spikes. But with wavelet analysis, we can use approximating functions that are contained neatly in finite domains. Wavelets are well-suited for approximating data with sharp discontinuities.

The wavelet analysis procedure is to adopt a wavelet prototype function, called an *analyzing wavelet* or *mother wavelet.* Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same wavelet. Because the original signal or function can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions) data operations can be performed using just the corresponding wavelet coefficients.

The fast Fourier transform (FFT) and the discrete wavelet transform (DWT) are both linear operations that generate a data structure that contains $\log_2 n$ segments of various lengths, usually filling and transforming it into a different data vector of length 2^n. The mathematical properties of the matrices involved in the transforms are also similar. The inverse transform matrix for both the FFT and the DWT is the transpose of the original. As a result, both transforms can be viewed as a rotation in function space to a different domain. For the FFT, this new domain contains basis functions that are sines and cosines. For the wavelet transform, this domain contains more complicated basis functions called wavelets, mother wavelets, or analyzing wavelets. Both transforms have another similarity. The basis functions are localized in frequency, making mathematical tools like power spectra (how much power is contained in a frequency interval) useful at picking out frequencies.

The most striking dissimilarity between these two kinds of transforms is that individual wavelet functions are *localized in space*. Fourier sine and cosine functions are not. This localization feature, along with localization of frequency provided by wavelets, makes many functions and operators using wavelets “sparse" when transformed into the wavelet domain [Mallet 2009]. This in turn, results in a number of useful applications such as data compression, detecting features in images, and removing noise from time series.

One way to see the time-frequency resolution differences between the Fourier transform and the wavelet transform is to look at the basis function coverage of the time-frequency plane. Figure 7.5 shows a windowed Fourier transform, where the window is simply a square wave. The square wave window truncates the sine or cosine function to fit a window of a particular width. Because a single window is used for all frequencies in the windowed Fourier transform, the resolution of the analysis is the same at all locations in the time-frequency plane. An advantage of wavelet transforms is that the windows vary. In order to isolate signal discontinuities, it is desirable to have very short basis functions. At the same time, in order to obtain detailed frequency analysis, it ie desirable to have very long basis functions. A way to achieve this is to have short high-frequency basis functions and long low-frequency ones. This is achieved using wavelet transforms. Figure 7.6 shows the coverage in the time-frequency plane with a wavelet function.

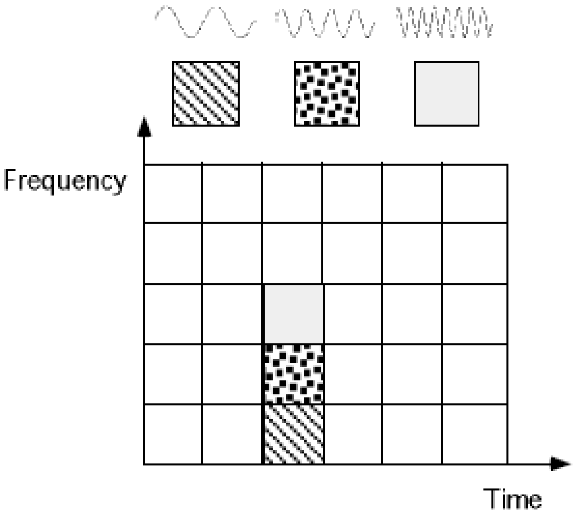


Figure 7.5. Fourier basis functions, time-frequency tiles, and coverage of the time-frequency plane

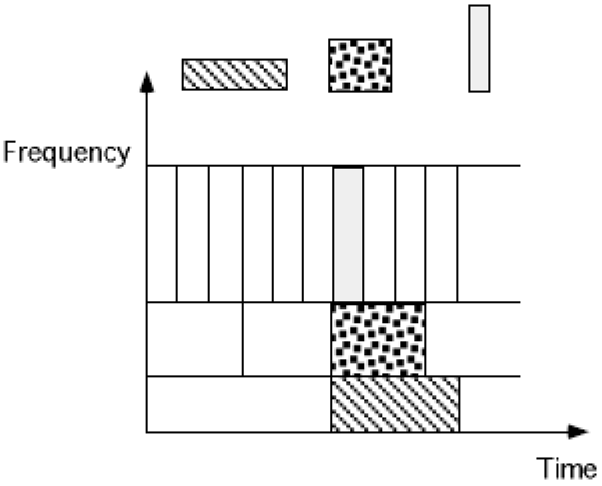


Fig. 7.6. The time-frequency tiles, and coverage of the time-frequency plane with Daubchies basis function

Wavelet transforms do not have a single set of basis functions. Instead, wavelet transforms have an infinite set of possible basis functions. Thus wavelet analysis provides immediate access to information that can be obscured by other time-frequency methods such as Fourier analysis. The wavelet mother function used in this work is the Haar wavelet.

7.2.3.3.2. The Haar wavelet transform – a brief overview

Wavelet transforms comprise an infinite set. The different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are. In general, Wavelets could be thought of as building blocks that can quickly de-correlate data. Just as signals are represented in terms of sines and cosines in Fourier analysis, we use basis functions to represent signals in wavelet analysis too. The wavelet basis is defined from the dilatations and translations of the 'Mother wavelet'. Within each family of wavelets (such as the Haar family) are wavelet subclasses distinguished by the number of coefficients and by the level of iteration. Wavelets are classified within a family most often by the *number of vanishing moments*. This is an extra set of mathematical relationships for the coefficients that must be satisfied, and is directly related to the number of coefficients. Figure 7.7 shows the basis function of the Haar wavelet transform.

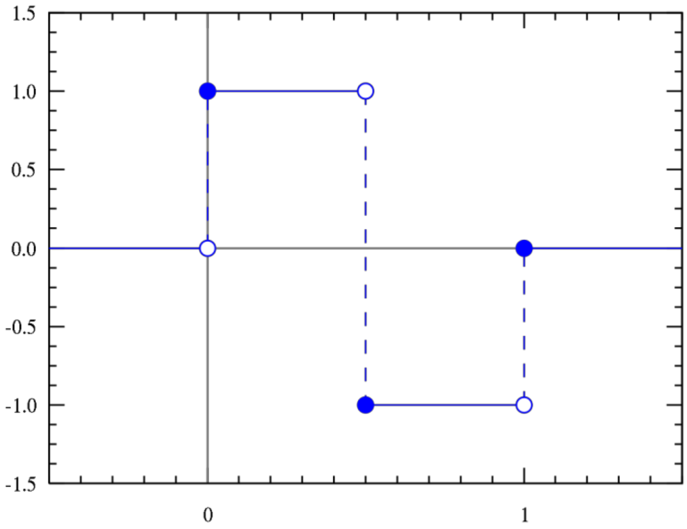


Figure 7.7. The Haar wavelet basis function

The Haar wavelet transform are defined by computing running averages and differences via scalar products with scaling signals and wavelets. These transforms are very powerful tools for performing noise removal. A Haar wavelet is the simplest type of wavelet. In discrete form, Haar wavelets are related to a mathematical operation called the *Haar transform – only the* *discrete Haar wavelet transform is discussed here.* The Haar transform serves as a prototype for all other wavelet transforms.

Let the discrete signal be expressed as ***f*** = $(f_1, f_2, f_3 \ldots . f_N)$, Where N is a positive integer denoting the length of ***f***. The *values* of ***f*** are the *N* real numbers $f_1, f_2, \ldots, f_N$.

These values are typically measured values of an analog signal g, measured at the time values $t = t_1, t_2, \ldots, t_N$ i.e. the values of ***f*** are

$$f_1 = g(t_1),\ f_2 = g(t_2), \ldots \ldots\ f_N = g(t_N)$$

Like all wavelet transforms, the Haar transform decomposes a discrete signal into two subsignals of half its length. One subsignal is a running average or *trend;* the other subsignal is a running difference or *fluctuation.*

Let us examine the trend signal first. The first trend subsignal $\mathbf{a}^1$ = (a_1, a_2, $a_3 \ldots a_{N/2}$) for the signal ***f*** is computed by taking a running average in the following way. The first value a_1 is computed by taking the average of the first pair of values of ***f*** : *(f_1 + f_2)/2* and then multiplying it with $\sqrt{2}$. That is,

$$a_1 = (f_1 + f_2)/\sqrt{2} \qquad (7.5)$$

Similarly, the next value a_2 is computed by taking the average of the next pair of values of ***f*** :

$(f_3 + f_4)/2$ and then multiplying it with $\sqrt{2}$. ie

$$a_2 = (f_3 + f_4)/\sqrt{2} \qquad (7.6)$$

Continuing in this manner, all the values of a^1 are computed by taking averages of successive pairs of values of ***f*** , and then multiplying these by $\sqrt{2}$. The precise formula for the values of $\mathbf{a}^1$ would be,

$$a_m = \frac{f_{2m-1} + f_{2m}}{\sqrt{2}} \qquad (7.7)$$

For m = 1, 2, 3 ..... N/2

For example, let ***f*** be defined by the values ***f*** = (4, 6, 10, 12, 8, 6, 5, 5), then it's first trend subsignal is a_1 = ($5\sqrt{2}$, $11\sqrt{2}$, $7\sqrt{2}$, $5\sqrt{2}$) calculated as per formula given in equation 7.7. Multiplication by $\sqrt{2}$ is done in order to ensure that the Haar transform preserves energy of the signal.

The other subsignal is called the first fluctuation. The first fluctuation of the signal ***f*** which is denoted by $\mathbf{d}^1$ = (d_1, d_2, $d_3 \ldots . . d_{N/2}$) is computed by taking a running difference as explained below.

$(f_1 - f_2)/2$ is computed first and it is multiplied with $\sqrt{2}$. ie

$d_1 = (f_1 - f_2)/\sqrt{2}$ (7.8)

The next value *d₂* is calculated by taking half the difference of the next pair of values ***f***
$(f_3 - f_4)/2$ and then multiplying it with $\sqrt{2}$

i.e. $d_2 = (f_3 - f_4)/\sqrt{2}$ (7.9)

Proceeding in this manner, all the values of $\mathbf{d}^1$ are obtained according to the formula,

$$d_m = \frac{f_{2m-1} - f_{2m}}{\sqrt{2}}$$

(7.10)

For *m = 1, 2, 3 .... N/2.*
For example, for the the signal ***f*** = (4, 6, 10, 12, 8, 6, 5, 5) mentioned above, the first fluctuation $\mathbf{d}^1$ will be obtained as $(-\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0)$ making use of the formula in equation 7.10

Haar Transform level 1

The Haar transform is performed in several levels. The first level is the mapping H_1 defined by

$\mathbf{f} \longrightarrow (\mathbf{a}_1 \mid \mathbf{d}_1)$ (7.11)

H1 involves mapping of the discrete signal **f** into it's first trend $\mathbf{a}^1$ and first fluctuation $\mathbf{d}^1$.

ie. as shown above, $\mathbf{a}^1 = (a_1, a_2, a_3....a_m)$, $\mathbf{d}^1 = (d_1, d_2, d_3 .....d_m)$

$a_m = \frac{f_{2m-1} + f_{2m}}{\sqrt{2}}$ and $d_m = \frac{f_{2m-1} - f_{2m}}{\sqrt{2}}$; *m = 1, 2, 3, 4...... N/2*

This mapping has an inverse, ie getting back **f** from the values of a_m and d_m

$\boldsymbol{f} = (f_1, f_2, f_3 ....f_m)$; *m =1, 2, 3 ... N/2* (7.12)

$$f_1 = (a_1 + d_1)/\sqrt{2} \qquad (7.13)$$

$$f_2 = (a_1 - d_1)/\sqrt{2} \qquad (7.14)$$

$$f_3 = (a_2 + d_2)/\sqrt{2} \qquad (7.15)$$

$$f_4 = (a_2 - d_2)/\sqrt{2} \qquad (7.16)$$

$$f = \left(\frac{(a_1+d_1)}{\sqrt{2}}, \frac{(a_1-d_1)}{\sqrt{2}}, \frac{(a_2+d_2)}{\sqrt{2}}, \frac{(a_2-d_2)}{\sqrt{2}} \ldots., \frac{(a_{N/2}+d_{N/2})}{\sqrt{2}}, \frac{(a_{N/2}-d_{N/2})}{\sqrt{2}}\right) \qquad (7.17)$$

The ‘small fluctuation feature’ as it is called, is the prime advantage with using the Haar transform. The fluctuation subsignal has values of magnitude which are very much smaller than the magnitudes of values of the original signal. Conservation of energy and compaction of energy are the two other significant features of the Haar transform.

The 1-D discrete Haar wavelet transform

The DWT can be interpreted as spectral analysis using a set of basis functions those are localized in both time and frequency, in contrast to the infinite-extent sinusoids used in Fourier analysis. Haar basis functions are the oldest and the simplest of all the wavelet basis functions that are used practically. Hence it was selected for this work.

1 – level Haar wavelets are defined as

$$W_1^1 = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0 \ldots..0\right) \qquad (7.18)$$

$$W_2^1 = \left(0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0 \ldots..0\right) \qquad (7.19)$$

:

:

$$W_{N/2}^1 = \left(0, 0 \ldots..0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right) \qquad (7.20)$$

These 1 – level Haar wavelets each have energy of 1. Each consists of a rapid fluctuation between $\pm\frac{1}{\sqrt{2}}$ with an average value of 0, and hence the name 'wavelets'. Each is a translation forward in time by an even number of time-units of the first Haar wavelet W_1^1 . W_2^1, W_3^1 W_4^1 .... are forward translations in time by 2, 4, 6.... units.

Once W_1^1 . W_2^1, W_3^1 W_4^1 .... are defined in this manner, **d¹** can be defined as follows.

$d_1 = \frac{f_1 - f_2}{\sqrt{2}}$, $d_2 = \frac{f_3 - f_4}{\sqrt{2}}$, ... , $d_m = \frac{f_{2m-1} - f_{2m}}{\sqrt{2}}$ become

$$d_1 = \mathbf{f}.W_1^1\ ,\ \ d_2 = \mathbf{f}.W_2^1 \ ....\quad d_m = \mathbf{f}.W_m^1 \qquad (7.21)$$

For $m = 1, 2, 3 \ldots.., \frac{N}{2}$

We can also express the 1-level trend values in a similar fashion. The elementary signals used are called the 1-level Haar scaling signals.

$$V_1^1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0,0, \ldots\ldots, 0) \qquad (7.22)$$

$$V_2^1 = (0,0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0,0, \ldots\ldots, 0) \qquad (7.23)$$

:

:

$$V_{N/2}^1 = (0,0, \ldots\ldots, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \qquad (7.24)$$

Using these Haar scaling signals, the values **a¹** = (a_1, a_2, ... , a_m) for $= 1, 2, 3 \ldots.., \frac{N}{2}$

$$a_m = \boldsymbol{f}.V_m^1 \qquad (7.25)$$

The Haar scaling signals are very similar to the Haar wavelets. They have energy 1 and have a support of just two consecutive time indices. Similar to the

1 – level Haar scaling signals, we have the 2 – level Haar scaling signals also. These are defined as follows.

$$V_1^2 = \left(\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},0,0,0\dots.,0\right) \qquad (7.26)$$

$$V_2^2 = \left(0,0,0\,0\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},0,0,0\dots.,0\right) \qquad (7.27)$$

:

:

$$V_{N/4}^2 = \left(0,0,0,\dots.0,\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2}\right) \qquad (7.28)$$

These scaling signals are all translations by multiples of four time-units of the first scaling signal V_1^2, and they all have energy 1 and average value ½. Thevalues of the 2-level trend $\mathbf{a}^2$ are scalar productsof these scaling signals with signal $\boldsymbol{f}$. $\mathbf{a}^2$ satisfies

$$\mathbf{a^2} = \left(\mathbf{f.V_1^2},\mathbf{fV_1^2},\dots\dots.,\mathbf{f.V_{N/4}^2}\right)$$

Likewise, the 2-level Haar wavelets are defined by

$$W_1^2 = \left(\frac{1}{2},\frac{1}{2},-\frac{1}{2},-\frac{1}{2},0,0,\dots..,0\right) \qquad (7.29)$$

$$W_2^2 = \left(0,0,0,0,\frac{1}{2},\frac{1}{2},-\frac{1}{2},-\frac{1}{2},0,0,\dots..,0\right) \qquad (7.30)$$

:

:

$$W_{N/4}^2 = \left(0,0,\dots,0,\frac{1}{2},\frac{1}{2},-\frac{1}{2},-\frac{1}{2}\right) \qquad (7.31)$$

These wavelets all have supports of length 4, since they are all translations by multiples of four time-units of the first wavelet $\mathbf{W_1^2}$**.** They also all have energy 1 and average value 0. Using scalar products, the 2 – level fluctuation $\mathbf{d^2}$ satisfies

$$\mathbf{d^2} = \left(\mathbf{f.W_1^2},\ \mathbf{f.W_1^2}, \dots \dots, \mathbf{f.W_{N/4}^2}\right) \qquad (7.32)$$

The Haar multiresolution Analysis.

As seen, the Haar transform can be described using scalar products with scaling signals and wavelets. The inverse Haar transform can also be described in terms of these same elemntary signals. Multiresolution analysis (MRA), the heart of wavelet analysis, is the means by which discret signals are synthesized by begining with a very low resolution signal and successively adding on details to create higher resolution versions, ending up with a complete synthesis of the signal at the finest resolution.

We have seen how the 1 –level Haar transform is expressed in terms of wavelets and scaling signals. The inverse of the 1-level Haar transform can also be expressed in terms of the same elementary signals. This leads to the first level of the Haar MRA.

Given two signals, f and g such that, $f = (f_1, f_2, f_3, \dots, f_N)$ and $g = (g_1, g_2, g_3, \dots, g_N)$

Recall equation 7.16, $f = \left(\frac{(a_1+d_1)}{\sqrt{2}}, \frac{(a_1-d_1)}{\sqrt{2}}, \frac{(a_2+d_2)}{\sqrt{2}}, \frac{(a_2-d_2)}{\sqrt{2}} \dots, \frac{(a_{N/2}+d_{N/2})}{\sqrt{2}}, \frac{(a_{N/2}-d_{N/2})}{\sqrt{2}}\right)$

(7.33)

So, f can be expressed as, $f = \left(\frac{a_1}{\sqrt{2}}, \frac{a_1}{\sqrt{2}}, \frac{a_2}{\sqrt{2}}, \frac{a_2}{\sqrt{2}}, \dots \dots, \frac{a_{N/2}}{\sqrt{2}}, \frac{a_{N/2}}{\sqrt{2}}\right) + \left(\frac{d_1}{\sqrt{2}}, \frac{d_1}{\sqrt{2}}, \frac{d_2}{\sqrt{2}}, \frac{d_2}{\sqrt{2}}, \dots \dots, \frac{d_{N/2}}{\sqrt{2}}, \frac{d_{N/2}}{\sqrt{2}}\right)$ (7.34)

This is, $\boldsymbol{f} = \mathbf{A^1} + \mathbf{D^1}$ (7.35)

$$\mathbf{A^1} = \left(\frac{a_1}{\sqrt{2}}, \frac{a_1}{\sqrt{2}}, \frac{a_2}{\sqrt{2}}, \frac{a_2}{\sqrt{2}} \dots \dots, \frac{a_{N/2}}{\sqrt{2}}, \frac{a_{N/2}}{\sqrt{2}}\right) \qquad (7.36)$$

$$\mathbf{D^1} = \left(\frac{d_1}{\sqrt{2}}, \frac{-d_1}{\sqrt{2}}, \frac{d_2}{\sqrt{2}}, \frac{-d_2}{\sqrt{2}} \dots \dots, \frac{d_{N/2}}{\sqrt{2}}, \frac{-d_{N/2}}{\sqrt{2}}\right) \qquad (7.37)$$

Using Haar scaling signals and wavelets, and using basic elementary algebraic operations with signals, the averaged and detail signals can be expressed as,

$$\mathbf{A^1} = \mathrm{a}_1\mathbf{V_1^1} + \mathrm{a}_2\mathbf{V_2^1} + \ \ldots.\ldots + \mathrm{a}_{\mathrm{N}/2}\mathbf{V_{N/2}^1} \qquad (7.38)$$

$$\mathbf{D^1} = \mathrm{d}_1\mathbf{W_1^1} + \mathrm{d}_2\mathbf{W_2^1} + \ \ldots.\ldots + \mathrm{d}_{\mathrm{N}/2}\mathbf{W_{N/2}^1} \qquad (7.39)$$

Now, recalling previous equations $d_m = \mathbf{f}.W_m^1$ and $a_m = \boldsymbol{f}.V_m^1$ for *m= 1, 2, 3, ... , N/2,* we can re-write the above as,

$$\mathbf{A^1} = (\mathbf{f.V_1^1})\mathbf{V_1^1} + (\mathbf{f.V_2^1})\mathbf{V_2^1} + \cdots + \ (\mathbf{f.V_{N/2}^1}) \qquad (7.40)$$

$$\mathbf{D^1} = (\mathbf{f.W_1^1})\mathbf{W_1^1} + (\mathbf{f.W_2^1})\mathbf{W_2^1} + \cdots + \left(\mathbf{f.W_{\frac{N}{2}}^1}\right)\mathbf{W_{N/2}^1} \qquad (7.41)$$

The above equations show that the averaged signal is a combination of Haar scaling signals, with the values of the first trend subsignal as coefficients and that the detail signal is a combination of Haar wavelets, with the values of the first fluctuation subsignal as coefficients.

The idea behind the first level of the Haar MRA of a signal can be extended to further levels, as many levels as the number of times the signal can be divided by 2. In the second level of MRA of the signal, the signal is expressed as,

$$\mathbf{f} = \mathbf{A^2} + \mathbf{D^2} + \mathbf{D^1} \qquad (7.42)$$

Since $\mathbf{A^2}$and $\mathbf{D^2}$ are the first average and detail signals and comparing the equations, $\mathbf{f} = \mathbf{A^1} + \mathbf{D^1}$ and $\mathbf{f} = \mathbf{A^2} + \mathbf{D^2} + \mathbf{D^1}$, we’ve,

$$\mathbf{A^1} = \mathbf{A^2} + \mathbf{D^2} \qquad (7.43)$$

From the above expression we see that, computing the second averaged signal $\mathbf{A^2}$ and the second detail signal $\mathbf{D^2}$ consists of performing the first level MRA of the signal $\mathbf{A^1}$**.** The second level averaged signal $\mathbf{A^2}$ satisfies

$$\mathbf{A^2} = (\mathbf{f.V_1^2})\mathbf{V_1^2} + (\mathbf{f.V_2^2})\mathbf{V_2^2} + \cdots + (\mathbf{f.V_{N/4}^2})\mathbf{V_{N/4}^2} \qquad (7.44)$$

And the second level detail signal satsfies,

$$\mathbf{D^2} = (\mathbf{f.W_1^2})\mathbf{W_1^2} + (\mathbf{f.W_2^2})\mathbf{W_2^2} + \cdots + (\mathbf{f.W_{N/4}^2})\mathbf{W_{N/4}^2} \qquad (7.45)$$

In general, if the number N of signal values is divisible k times by 2, then, a k-level MRA that can be performed on the signal is given by,

$$\mathbf{f} = \mathbf{A^k} + \mathbf{D^k} + \cdots + \mathbf{D^2} + \mathbf{D^1} \tag{7.46}$$

7.2.3.3.3. Removal of noise

Discrete wavelet transform employs filters which work at different frequencies (f_c - cutoff) ranges such that the signal is analysed at different scales. There are two filter banks – one which work at low frequencies and the other which works at high frequencies such that the incoming signal gets split into low and high frequency ranges.

When passed through the DWT filter-bank, the resolution of the signal changes. Resolution could be thought of as the amount of fine information content the signal possesses. The scale of the signal gets altered by the sampling operations of the DWT. Both up-sampling and down-sampling are performed. Up-sampling involves in increasing the number of samples the signal has in a given duration of time. This is achieved by interpolation or by introducing zeros. When we say a signal has been up-sampled by a factor 2, it means that a new value (either a zero or an interpolated value) has been added in between every two samples of the original signal. Down-sampling or sub-sampling involves reducing the number of samples of the signal, by removing certain samples. Or we could simple say, sampling the given signal.

Let the sequence to be treated be denoted by x[n], where n represents the number of samples (an integer). Treating the signal with the DWT involves passing the signal through a series of low pass and high pass filters. Filtering can be mathematically represented by convolution in time

$$x[n] * h[n] = \sum_{n=-\infty}^{\infty} x[k].h[n-k] \tag{7.47}$$

where x[n] is the signal to be filtered and h[n] represents the impulse response of the filter [Proakis 2006].

Let h[n] and g[n] represent the impulse responses of the low pass and the high pass half band filters respectively in the DWT filter bank. We consider passing x[n] through the low pass half-band filter, h[n], first. The cut off frequency, ω_c of the half-band low-pass filter is so selected as to be half of the maximum frequency content in the incoming signal x[n]. Thus if ω_{max} represents the maximum frequency contained in the signal, then ω_c is $\frac{\omega_{max}}{2}$. So, the low-pass filter h[n] removes all the frequencies that are above $\frac{\omega_{max}}{2}$. Thus after the low-pass filtering operation, we have the lower half-band of the signal between the extremes of 0 to $\frac{\omega_{max}}{2}$ radians. And the number of samples in the signal reduces by half. But lowering the number of samples by half does not distort the signal as per Nyquist rule [Proakis 2006].

Now consider passing x[n] through the high-pass filter g[n]. The cut-off frequency of the high-pass filter is also set at $\frac{\omega_{max}}{2}$ such that only the frequencies above $\frac{\omega_{max}}{2}$ are allowed to pass through it. Thus when x[n] is passed through the half-band high-pass filter with cut-off at $\frac{\omega_{max}}{2}$, the spectrum of the filtered output extends from $\frac{\omega_{max}}{2}$ to ω_{max}. Filtering x[n] with the high pass filter g[n] is represented mathematically as shown below.

$$x[n] * g[n] = \sum_{k=-\infty}^{\infty} x[k].\, g[n-k] \qquad (7.48)$$

We know that with digital signals, the unit of frequency is radians. A continuous signal (analogue signal) X(t) with maximum frequency F_m can be converted into an equivalent discrete x[n] signal without distortion or data loss if it is sampled at the Nyquist rate of $F_s = 2F_m$ [Proakis 2006]. The frequency spectrum (the Discrete Time Fourier Transform, the DTFT) of the discrete time signal x[n], X(ω) is a periodic, continuous signal which is symmetric about the Y axis and has a periodicity of $2\boldsymbol{\pi}$. One full period of X(ω) can extend to the maximum limits of $-\boldsymbol{\pi}^c$ to $+\boldsymbol{\pi}^c$ [Proakis 2006]. Thus the frequency of digital signals have a maximum limit; from $-\boldsymbol{\pi}$ to $+\boldsymbol{\pi}$ radians. Ignoring the negative half of the spectrum, it being symmetric, we can say that the maximum frequency of a discrete signal is $\boldsymbol{\pi}$ radians. Thus for a discrete signal which has a spectrum which stretches over the entire range of $-\boldsymbol{\pi}$ to $+\boldsymbol{\pi}$ radians, filtering with the half-band low pass signal reduces the frequency content from + $\frac{\pi}{2}$ to

$-\frac{\pi}{2}$ or ignoring the negative frequencies, we could say, the spectrum after filtering with the low-pass half band filter extends up to only $\frac{\pi}{2}$ from 0. Thus we could easily say that the lower half-band signal occupies the lower frequencies from 0 to $\frac{\pi}{2}$ radians and the upper half-band of the signal occupies the higher frequencies from $\frac{\pi}{2}$ to π ^{c}.

The half-band high pass or the half-band low pass filter would remove either the lower half or the upper half of the spectrum. This would be reflected in the number of samples of the signal in the DFT too. Let there be 1024 samples in the DFT X[k], of the discrete signal x[n]. spanning 0 to $\boldsymbol{\pi}^{c}$ in the spectrum. When x[n] is filtered with either h[n] (the low-pass half-band filter) or g[n] (the high-pass half-band filter), one half of the spectrum is removed. h[n] would remove the spectrum from $\frac{\pi}{2}$ to **π**, whereas g[n] would remove that part of the spectrum which is there between 0 to $\frac{\pi}{2}$ The number of samples in the DFT in either case would be half the number of samples in the original DFT of x[n]. Thus in filtered signal, be it the high frequency or the low frequency one, the number of samples in the DFT would be only 512. But it is to be noticed that the scale of the signal remains unchanged in either case. That is in the spectral plot, the spectrum of the filtered signal occupies the same spread in the X-axis as occupied by the spectrum of the original signal x[n]. Thus the frequency resolution has doubled, by either of the half-band filtering operations. The operation explained above represents one level of decomposition, either using the upped half-band or the lower half-band filter. This can be represented mathematically as seen in equations 7.7 and 7.8 below.

$$y_{high}[k] = x[n] * g[n] = \sum_{n} x[n] \,.\, g[2k-1] \tag{7.49}$$

$$y_{low}[k] = x[n] * h[n] = \sum_{n} x[n] \,.\, h[2k-1] \tag{7.50}$$

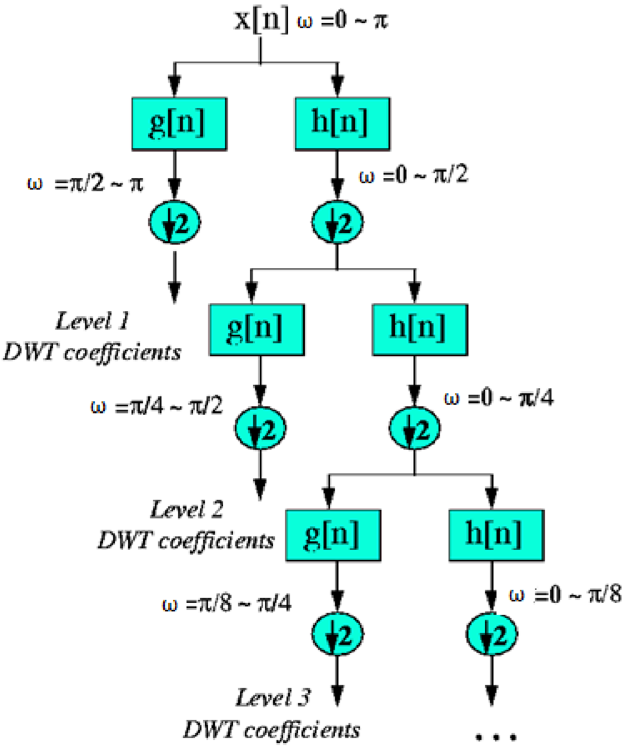


Figure 7.8. The DWT filter bank.

It can be seen that frequency resolution doubles with decomposition as half the number of samples are stretched over the entire existing scale. At the same time, it can be seen that time resolution reduces by half the previous value as only half the number of samples characterizes the entire signal.

This procedure is called sub-band coding and is repeated with every stage of decomposition. In noise removal using the DWT, the upper half-band or the lower half-band filters are used depending on the frequency ranges occupied by the noise in the signal. If the noise occupies higher frequency ranges, the signal is passed through the low pass half-band filters so that the noise gets removed. On the other hand if the noise occupies lower ranges of frequency, the signal is subjected to half-band high pass filtering. Every stage of decomposition reduces the number of samples in the spectrum by half, keeping the scale the same which doubles frequency resolution. As half the number of samples represents the resultant signal, each stage of DWT decomposition halves the time resolution. Figure 7.8 represents the procedure explained above. Where x[n] is the signal to be treated and h[n] and g[n] represent the low pass and the high filters respectively. Decimation thus splits the frequency contents in the spectrum into low and high halves. In this work, the noise in exon plots is found to occupy the higher frequency ranges. Hence while reconstructing the decimated signal, only the coefficients of approximation (the lf range of the spectrum) were used. Two levels of decimation and reconstruction using *Haar* wavelets is performed here for noise removal. Noise is found to occupy the higher frequencies and hence higher frequencies are not used in reconstruction. The advantage of using DWT is that good time resolution is obtained at high frequencies, and good frequency resolution at low frequencies with effective removal of noise.

7.2.3.4. G-C content, START and STOP codons

Long ncRNA is reported to have lower G-C content when compared to coding regions [Niazi 2012]. The G-C content of the sequences is found, relative to the total number of nucleotides. Sequence matching is done to locate START and STOP codon patterns. The sequences were checked for ATG and the alternative START codons too viz. ATG, CTG and GTC and also for the STOP codons, TAA, TAG, and TGA. The results of the study are detailed n the next section.

7.3. Exons in long noncoding RNA sequences studied

Here, we present the exon maps of lncRNA sequences obtained using the period 3 property making use of digital filters. STFT is used to obtain the spectrum of the sequences. While computing the spectrum using STFT, optimum window size is mandatory for locating the exons. Window sizes depend on the length of the sequence analysed [George 2010]. De-noising of exon plots is done with the help of the DWT filter bank which filters out HF noise, and only the low frequency components of decimation are used in reconstruction. It is found that the reduced computation technique [George 2010] which applied a quadratic window and reduced noise in the case of exon prediction of coding DNA sequences is not found to be of use here. Applying the quadratic window is seen to introduce additional spectral noise and is not used in the algorithm here.

7.3.1. Exon maps of lncRNA sequences

Figures 7.9 to 7.28 show the exon plots obtained using the algorithm described in section 5 of this chapter. 20 lncRNA sequences were used in this study and exon plots of all the sequences are given here. We will see the detailed explanation of the first two plots; that of lncRNA CCEPR and FALEC. Similar logic is to be applied while interpreting the other exon plots. CCEPR has one exon and FALEC has two exons.

Figure 7.9 shows the exon plot of lncRNA CCEPR (Homo sapiens cervical carcinoma expressed PCNA regulatory lncRNA) and its GenBank accession number is NR_131782.1. It is 2502 bases long and contains a single exon as per the NCBI record (https://www.ncbi.nlm.nih.gov/nuccore/NR_131782.1), from 1 to 2502 ie. spanning the entire length of the sequence taken. The exon plots has nucleotide location along the X axis and the power spectral density (PSD) along the Y axis. As per the period 3 property, the energy peaks (peaks in the PSD) should correspond to exons. The peak power in this plot is between $\mathbf{6 \times 10^{-17}}$and $\mathbf{7 \times 10^{-17}}$ the half-power value is between 3**×** $\mathbf{10^{-17}}$ and $\mathbf{3.5 \times 10^{-17}}$. Though there are dips in the plot, on an average, the plot retains the half power throughout and does not touch the 0 PSD value at any point. Hence we count only one peak in this plot. Thus, there is a single exon extending from 1 to around 2450. This range conforms to the value given in the NCBI database.

The next plot given in Figure 7.10 is that of lncRNA FALEC (focally amplified long non-coding RNA in epithelial cancer) with NCBI accession number NR_051960.1. As per the NCBI record (https://www.ncbi.nlm.nih.gov/nuccore/NR_051960.1) FALEC has two exons; 1 - 306 and 307 - 566. The exon plot given in Figure 7 shows two energy peaks corresponding to two exons. Peak power value is between 0.6×10^{-17} and 0.8×10^{-17} and half power values between 0.3×10^{-17} and 0.4×10^{-17}. Based on the very definition of half power, PSD values less than the half-power are not considered as peaks. The first exon in the plot spans from 1 to around 190 and the second from around 220 to 560. The net length of the sequences in terms of nucleotides is 566.

Figures 7.11 to 7.28 show the exon plots of lncRNAs BACE-AS, CAHM, BGLT3, ABALON, DISC2, GHET1, HEIH, NEAT1, MALAT1, NKILA, FOXC2-AS1, DLEU1, HULC, KIAA0087, MHENCR, PRNT, HOTAIRM1 and CISTR respectively.

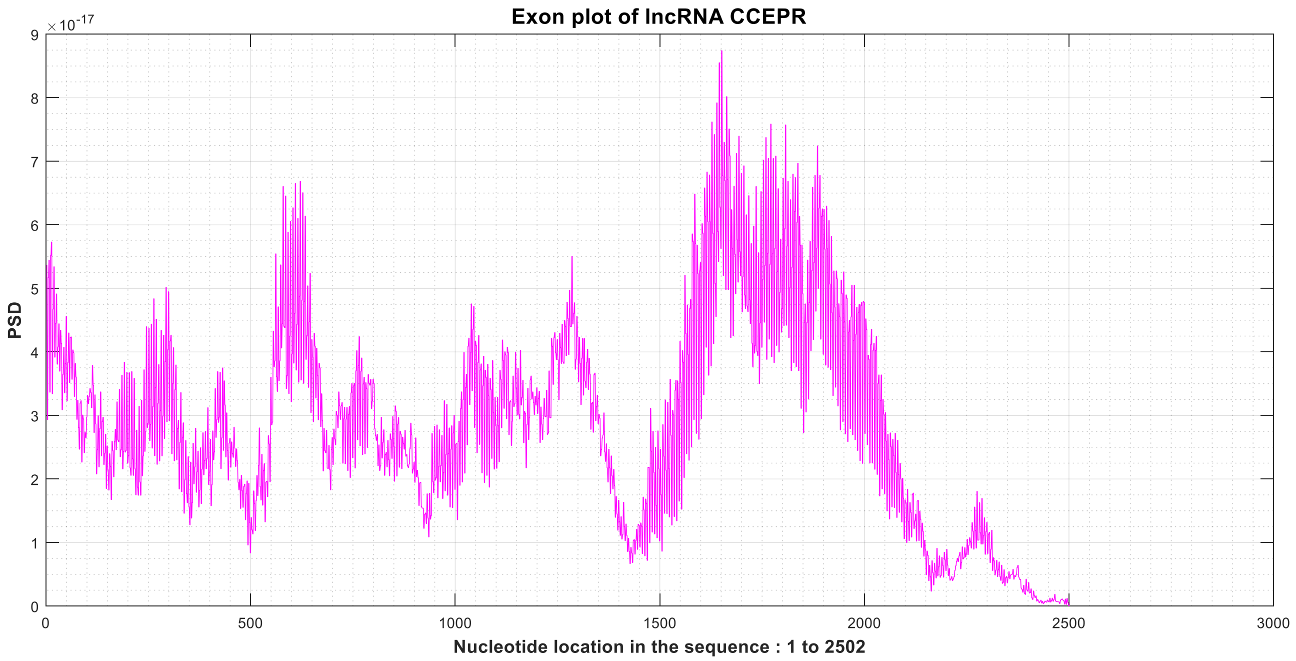


Figure 7.9. Exon plot of lncRNA CCEPR. Exon plot of lncRNA CCEPR with GenBank accession no. NR_131782.1

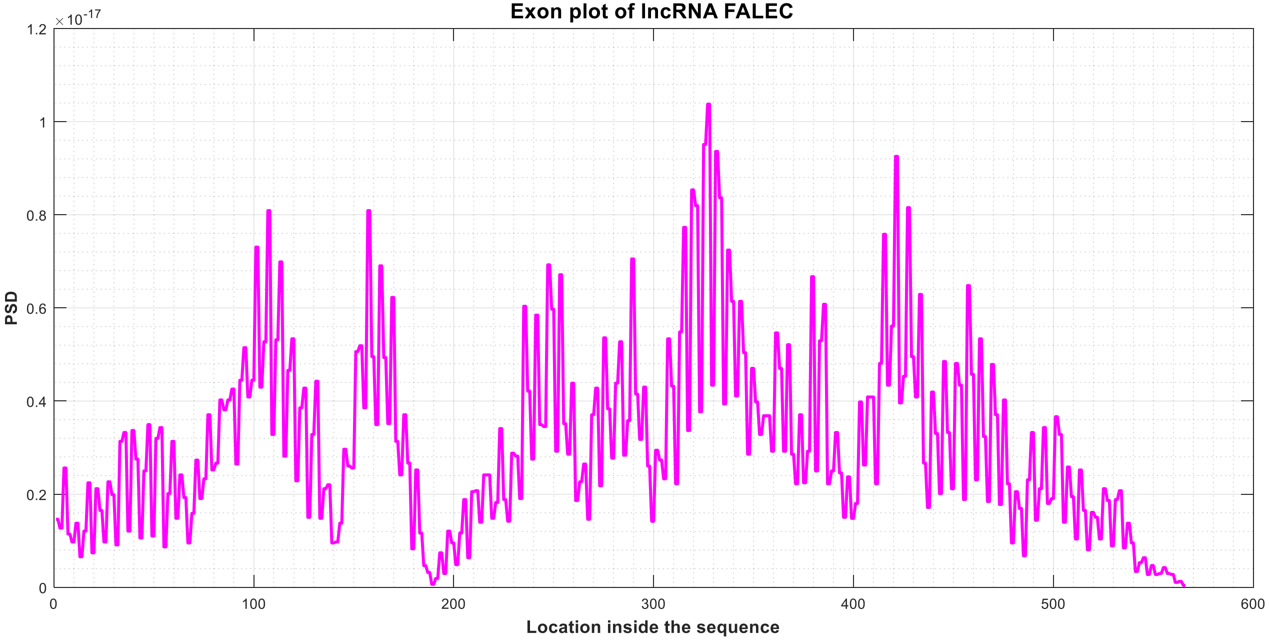


Figure 7.10. Exon plot of lncRNA FALEC. Exon plot of lncRNA FALEC with GenBank accession no. NR_051960.1

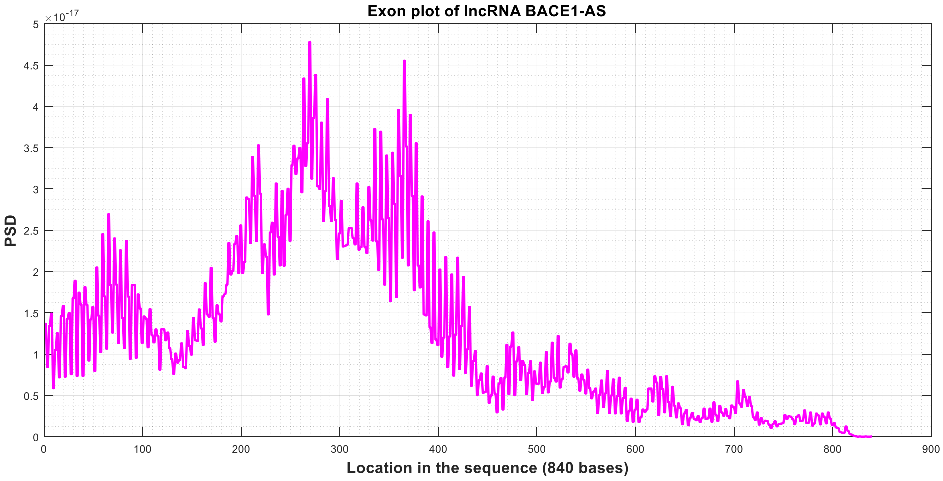


Figure 7.11 Exon plot of lncRNA BACE-AS (has 840 bases)

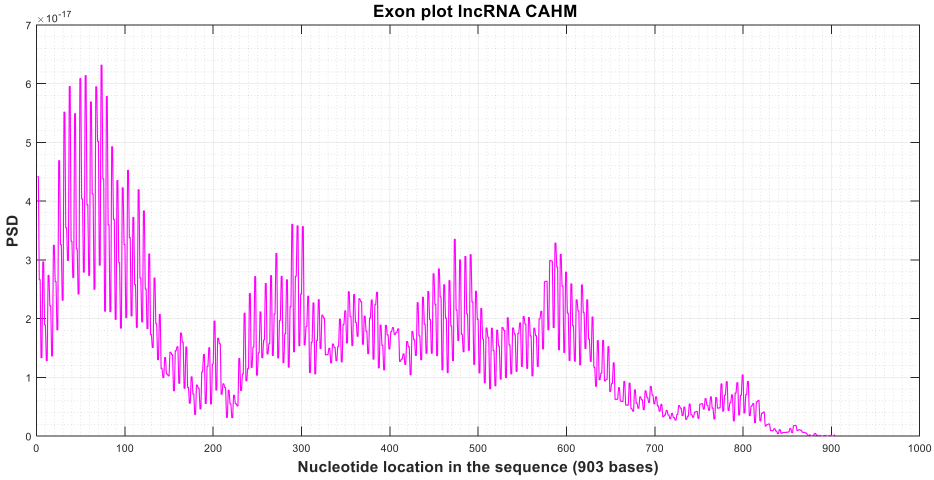


Figure 7.12. Exon plot of lncRNA CAHM

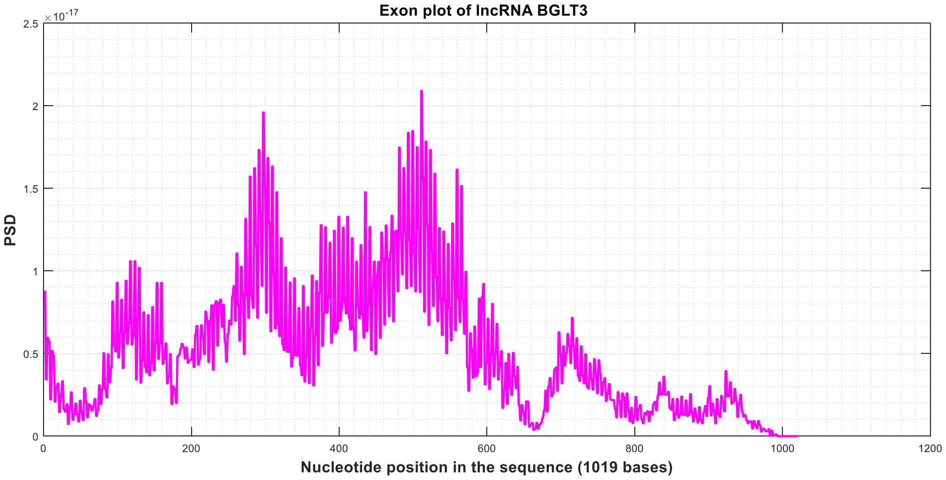


Figure 7.13. Exon plot of BGLT3. 1019 bases; 1 exon

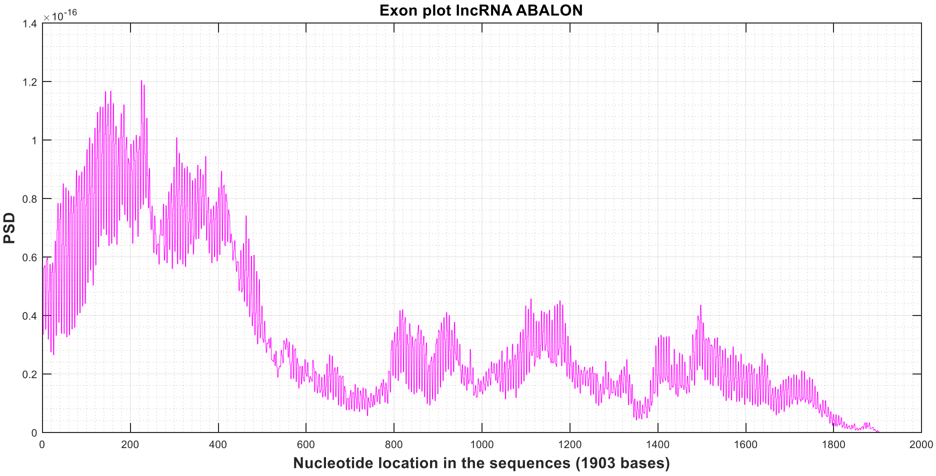


Figure 7.14 Exon plot of lncRNA ABALON

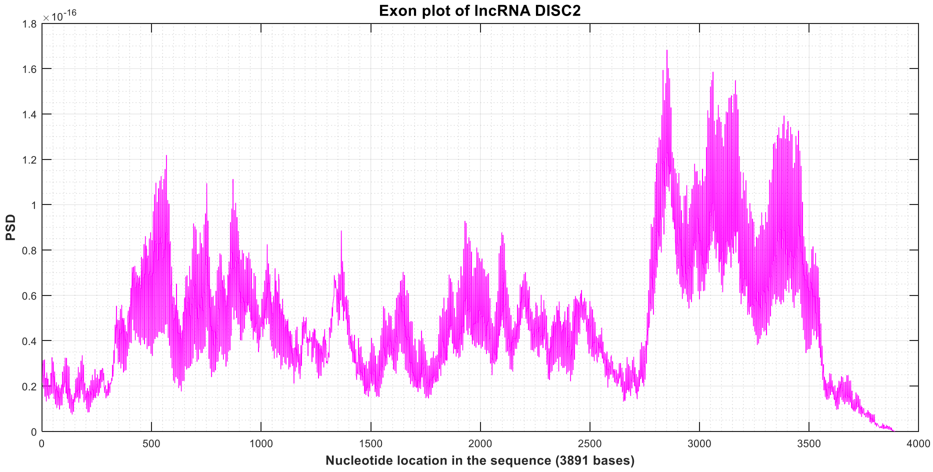


Figure 7.15. Exon plot of lncRNA DISC2

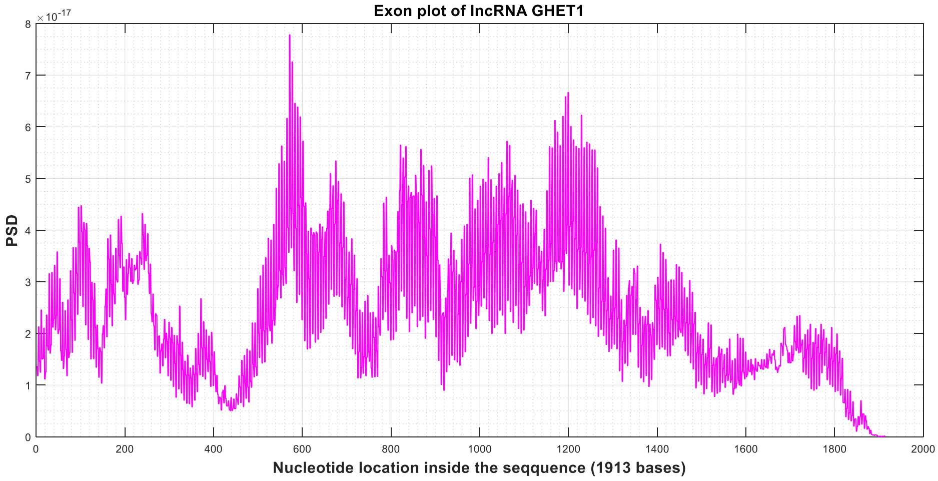


Figure 7.16. Exon plot of lncRNA GHET1

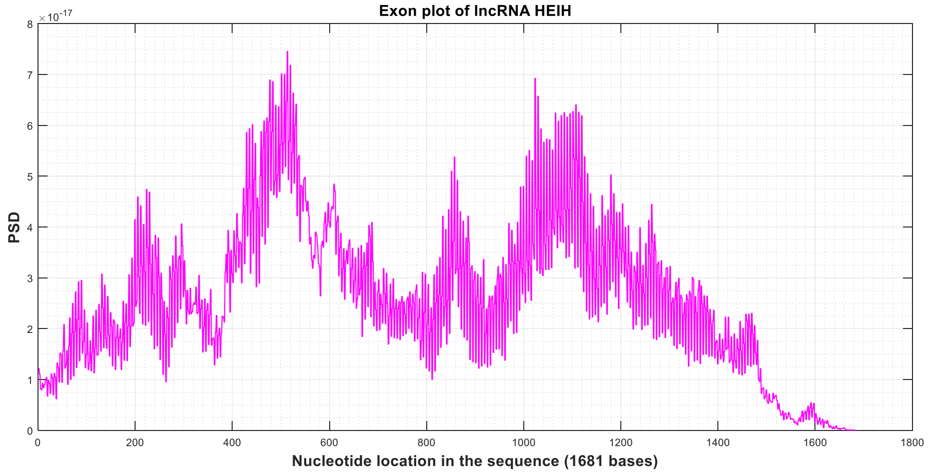


Figure 7.17. Exon plot of lncRNA HEIH

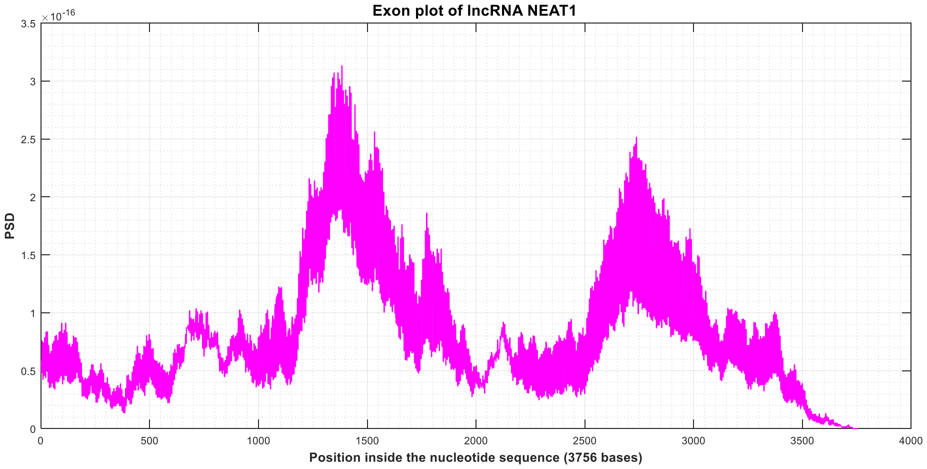


Figure 7.18. Exon plot of lncRNA NEAT1

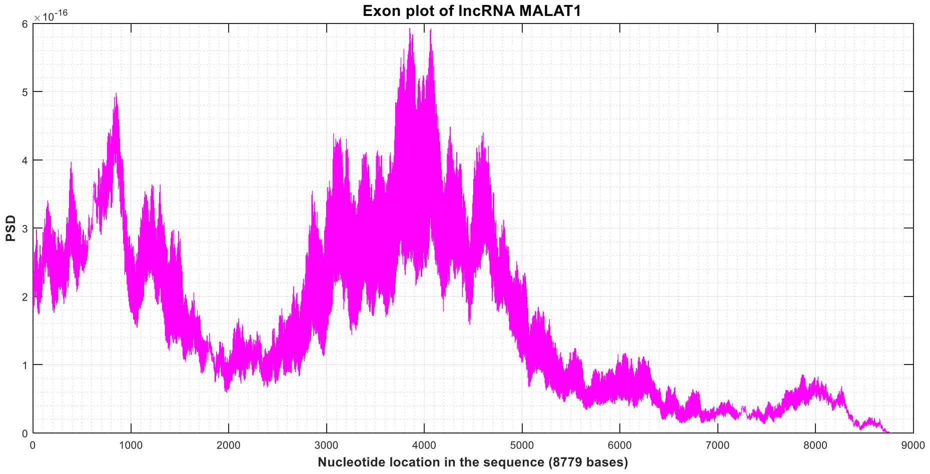


Figure 7.19. Exon plot of lncRNA MALAT1

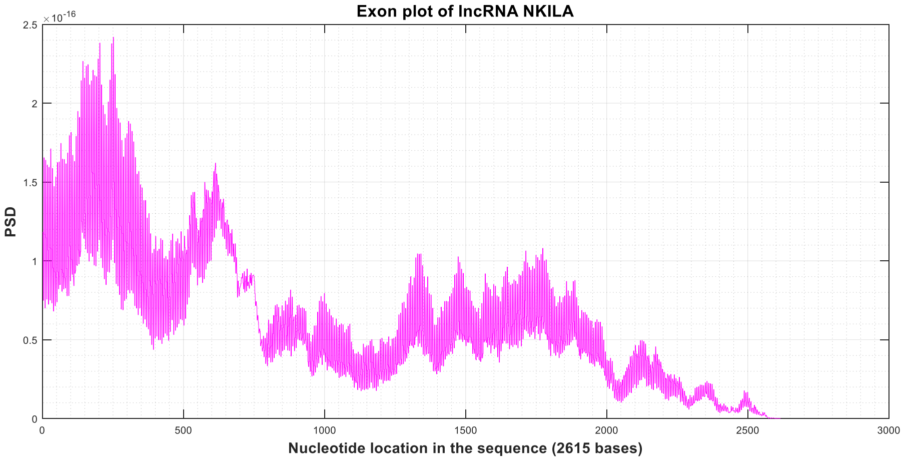


Figure 7.20. Exon plot of lncRNA NKILA

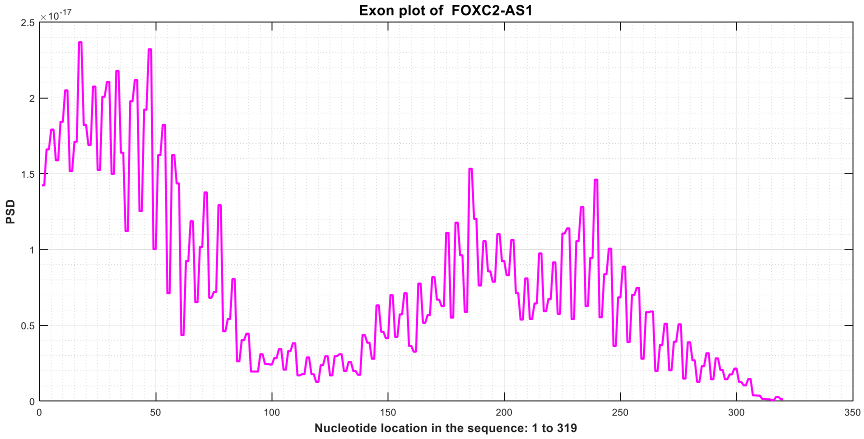


Figure 7.21. Exon plot of lncRNA FOXC2-AS1

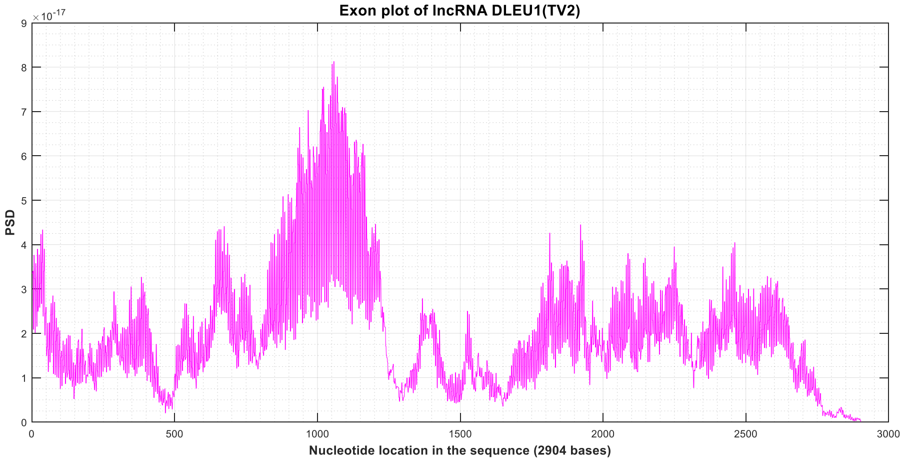


Figure 7.22. Exon plot of lncRNA DLEU1 (transcript variant 2)

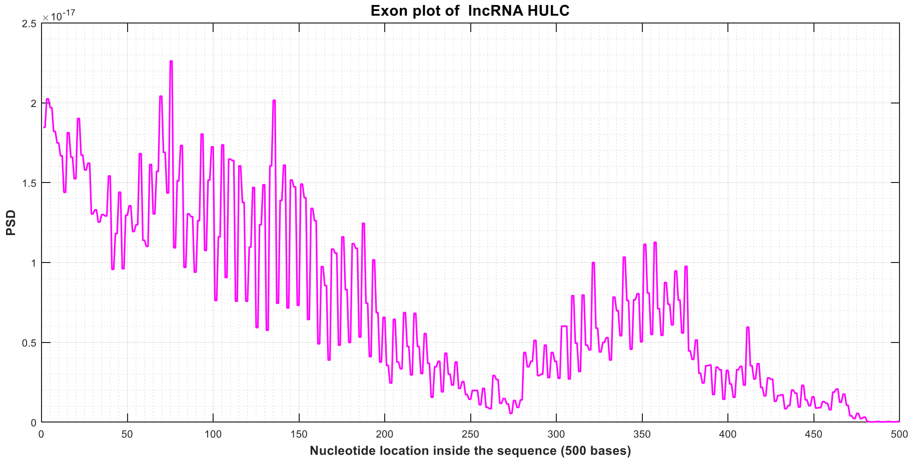


Figure 7.23. Exon plot of lncRNA HULC

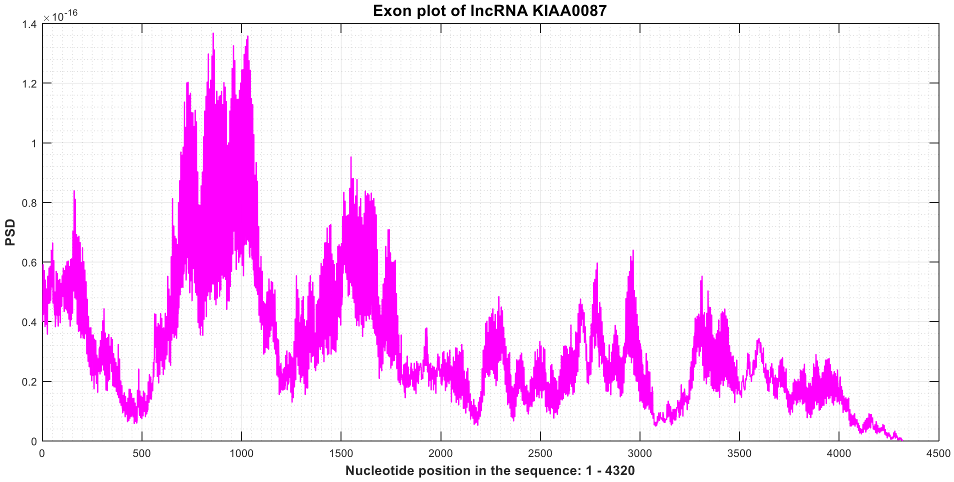


Figure 7.24. Exon plot of lncRNA KIAA0087

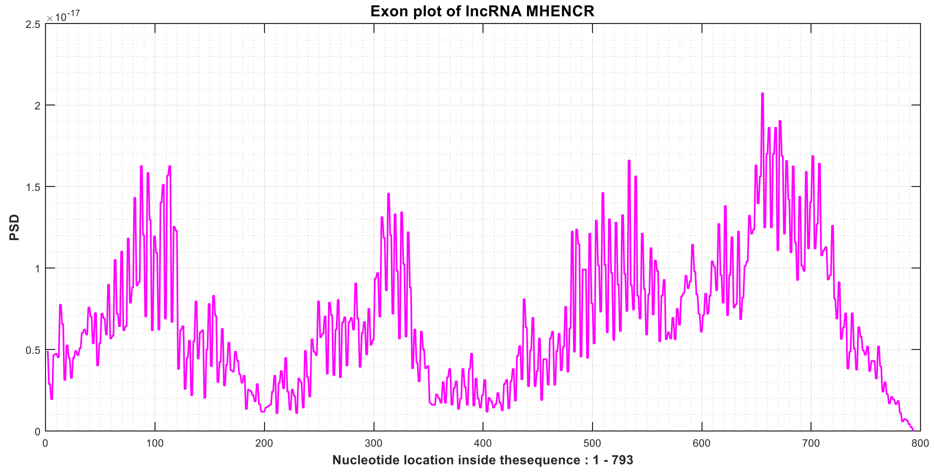


Figure 7.25. Exon plot of lncRNA MHENCR

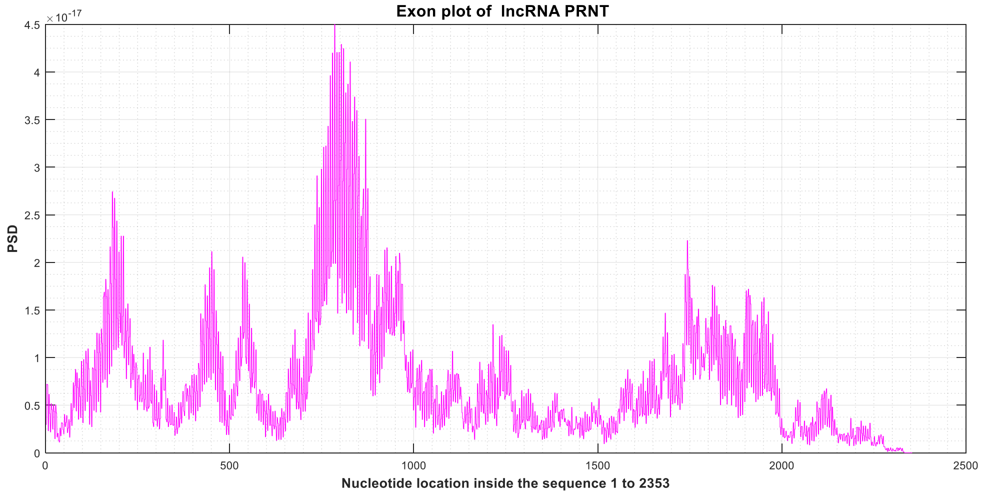


Figure 7.26. Exon plot of lncRNA PRNT

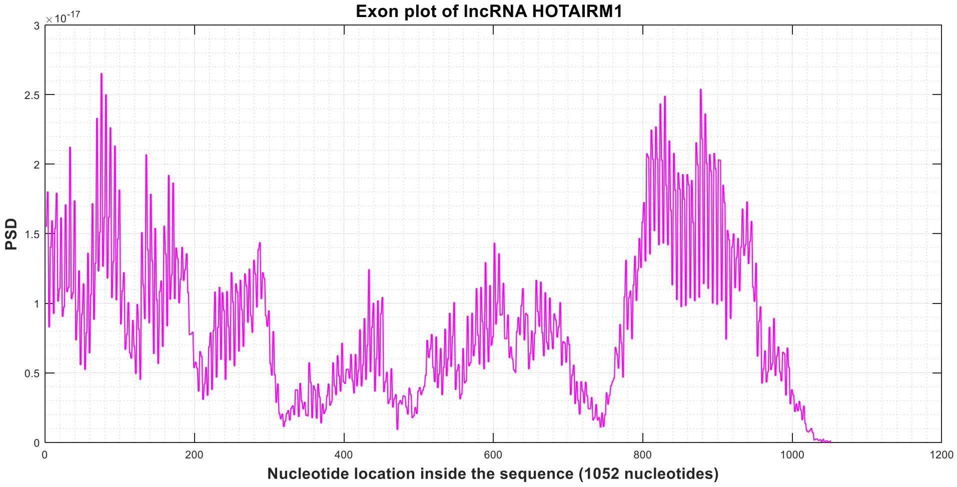


Figure 7.27. Exon plot of lncRNA HOTAIRM1

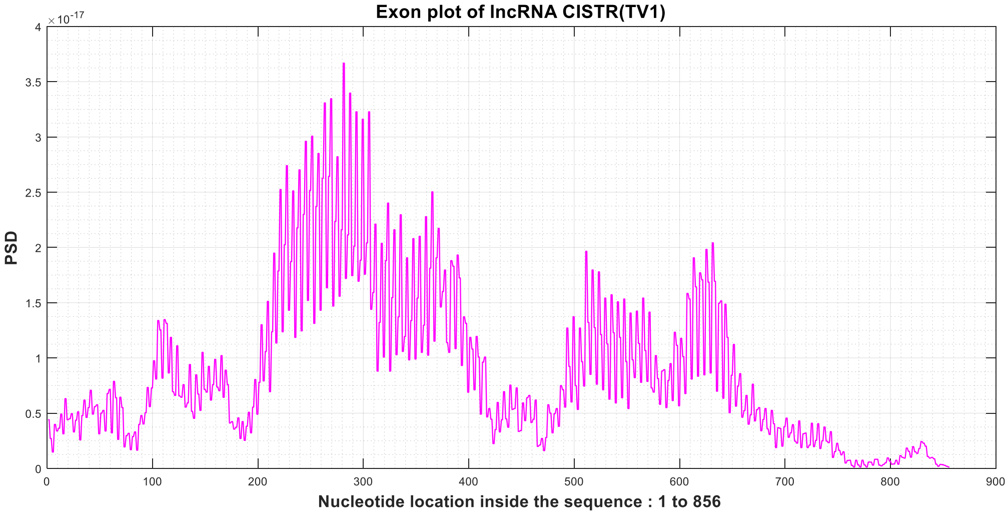


Figure 7.28. Exon plot of lncRNA CISTR (transcript variant 1)

7.3.2. Comparison of exon locations obtained with the values in NCBI records

A summary of the exon locations obtained in this work has been compared with wet-lab results which are found in the NCBI records and is given in Table 7.2. Column 2 of the Table shows the name of the lncRNA sequence, column 3 gives the GenBank accession number and column 4 gives the length of the sequence. Columns 5 and 6 show the start and end positions of exons as per the NCBI records which are results of wet-lab methods, while columns 7 and 8 show the same obtained in this work. Columns 9 and 10 display the deviation in exon locations observed at the start and the end with reference to the NCBI records. Each of the NCBI records for these sequences site literature which ascertains that wet-lab techniques have been used in the analysis of the long noncoding RNAs. Sample references [Peng 2016, Yang 2015, Choy 2006] are included in this Chapter for the sequence CCEPR (https://www.ncbi.nlm.nih.gov/nuccore/NR_131782.1). Records corresponding to the NCBI accession numbers can be found for each of the sequences presented here.

Among sequences analyzed, CCEPR, BACE-AS1, CAHM, BGLT3, ABALON, DISC2, GHET1, NEAT1, HEIH, NEAT1, MALAT1, NKILA have one exon each. Long ncRNAs FOXC2-AS1, DLEU1, HULC, KIAA0087, MHENCR, FALEC, PRNT have 2 exons each. CISTR and HOTAIRM1 have 3 exons.

The range of deviation of exon locations obtained in this work is around 36 to 100 nucleotides with respect to the exon ranges given in their NCBI records except for lncRNAs PRNT and HOTAIRM1. These two sequences show deviations of 180 and 175 respectively.

Table 7.2 Comparison of exon locations of lncRNA sequences obtained with the exon ranges in the NCBI records.

Sl.No.	Name of the lncRNA	GenBank Accession number	Length	Exon location				Deviation in location	
				Reference (GenBank)		Observed			
				Start	End	Start	End	Start	End
1	CCEPR	NR_131782.1	2502	1	2502	1	2500	0	2
2	BACE-AS1	NR_037803.2	840	1	825	1	800	0	25
3	CAHM	NR_037593.1	903	1	886	1	850	0	36
4	BGLT3	NR_121648.1	1019	1	993	1	980	0	13
5	ABALON	KC505631.1	1903	1	1903	1	1900	0	3
6	DISC2	NR_002227.2	3892	1	3892	1	3880	0	12
7	GHET1	NR_130107.1	1903	1	1903	1	1890	0	13
8	HEIH	NR_045680.1	1681	1	1681	1	1680	0	1
9	NEAT1	NR_028272	3756	1	3735	1	3700	0	35
10	MALAT1 (TV1)*	NR_002819.4.	8779	1	8779	1	8750	0	29
11	NKILA	NR_131157.1	2615	1	2598	1	2500	0	98
12	FOXC2-AS1	NR_125795.1	319	1	145	1	140	0	5
				146	319	140	319	6	0
13	DLEU1 (TV2)**	NR_002605.2	2904	1	389	1	450	0	-61
				390	2904	500	2900	-110	4
14	HULC	NR_004855.2	500	1	182	1	230	0	-48
				183	484	250	480	-67	4
15	KIAA0087	NR_022006.1	4320	1	420	1	500	0	-80
				421	4320	500	4300	-79	20
16	MHENCR	NR_132417.1	793	1	158	1	200	0	-42

				159	793	200	793	-41	0
17	FALEC	NR_051960.1	566	1	306	1	200	0	106
				307	566	220	560	87	6
18	PRNT	NR_024267.1	2353	1	529	1	350	0	179
				530	2333	350	2300	180	33
19	HOTAIRM1	NR_038366.1	1502	1	295	1	300	0	-5
				296	564	300	740	-4	-176
				565	1044	740	1000	-175	44
20	CISTR (TV1)*	NR_104332.1	856	1	221	1	200	0	21
				222	337	200	420	22	-83
				338	856	450	800	-112	56

7.3.3. G-C content of lncRNAs

Table 7.3. G-C Concentration

Sl No	lncRNA	NCBI accession number	Sequence Length	% GC Concentration
1	CCEPR	NR_131782.1	2502	41.9265
2	BACE1-AS	NR_037803.2	840	47.1429
3	CAHM	NR_037593.1	903	59.4684
4	BGLT3	NR_121648.1	1019	38.5672
5	ABALON	NR_131907.1	1903	56.5423
6	DISC2	NR_002227.2	3892	37.7698
7	GHET1	NR_130107.1	1913	44.5896
8	HEIH	NR_045680.1	1681	58.5366
9	NEAT1	NR_028272.1	3756	47.9499
10	MALAT1 (TV1*)	NR_002819.4	8779	40.3463
11	NKILA	NR_131157.1	2615	53.3461
12	FOXC2-AS1	NR_125795.1	319	54.8589
13	DLEU1 (TV2**)	NR_002605.2	2904	38.6708
14	HULC	NR_004855.2	500	36
15	KIAA0087	NR_022006.1	4320	41.4583
16	MHENCR	NR_132417.1	793	55.9899
17	FALEC	NR_051960.1	566	56.8905
18	PRNT	NR_024267.1	2353	47.0463
19	HOTAIRM1	NR_038366.1	1052	50.7605
20	CISTR (TV1*)	NR_104332.1	856	51.8692

***TV1 – Transcript variant 1, *TV2 – Transcript variant 2**

The G-C content of the sequences is computed relative to the total number of nucleotides and is shown in Table 7.3. It is found that out of 20 sequences, 9 sequences have relative GC concentration more than 50%. The sequence with the highest GC content is lncRNA CAHM (GenBank accession number NR_037593.1), 59.4684%. The average value of GC concentration is found to be 47.9865%.

7.3.4. START and STOP codons in lncRNA sequences

Sequence matching is done to locate START and STOP codon patterns. The sequences are checked for ATG and the alternative START codons too viz. ATG, CTG and GTC and also for the STOP codons, TAA, TAG, and TGA. It is found that all the sequences have START codon patterns but none of them have STOP codons.

7.4. Discussion

Period 3 property is a feature which has been combined with a proven signal processing based automated method of detecting exons in coding DNA sequences [Vaidyanathan 2002], [George 2010], [Anastassiou 2002]. It was observed in noncoding regions for procaryotes and some viral and mitochondrial base sequences two decades ago [Li 1997] and this concept is explored here. There are other signal processing based automated methods to detect exons in coding regions. One such method [Song 2010] detects short exons in DNA sequences by analyzing their structural properties viz. DNA bending stiffness, disrupt energy, free energy, and propeller twist making use of the autoregressive model to arrive at linear prediction matrices for these features. The linear prediction matrices for the four features are combined to find the linear prediction coefficients from which the spectrum of the DNA sequence is estimated and exons detected based on the $\frac{1}{3}$ rd frequency component. Short exons have also been detected by evaluating the complex wavelet transform of the structural features of DNA sequences [Provazník 2012]. In this work, we opted for period 3 property because of its proven robustness and relative simplicity [Vaidyanathan 2002], [Anastassiou 2002], [George 2010].

In my former work [George 2010] the exons for the sequence AF099922 (former GenBank accession number) has been mapped out. The nucleotide sequence was taken from the gene SL1 trans-splice acceptor F56F11.4, which is a part of the F56F11 DNA sequence. Exons are located making use of the period 3 property and the best method of locating was found to be the one using IIR peaking filters followed by DWT de-noising using the Haar wavelet. This GenBank record for AF099922 is obsolete now, but it is mentioned here, as the plot [George 2010] is easy to relate to in the context of exon locations.

Figure 7.29 shows a sample exon plot of a coding region which is obtained by making use of period 3 property along with digital filtering. The sequence is that of homo sapiens gene for Osteomodulin with GenBank accession number AB009589.1. The region of the sequence considered is 8000 to 11,000. As per the GenBank record [https://www.ncbi.nlm.nih.gov/nuccore/AB009589.1], this region has two exons: 8524 – 9479 and 10624 – 11846. In the PSD plot (Figure 8), we find two energy peaks in the regions around 500 – 1350 and 2600 – 3500. As the segment considered here is 8000 to 12000, the energy peaks are from around 8500 to 9350 and from around 10600 to 11500. These two energy peaks correspond to the exons in this particular segment (8000 to 11,000) of AB009589.1. The peak power is seen to be between 0.8 X 10^{-16} and 1 X 10^{-16} and the half power values between 0.4 X 10^{-16} and 0.5 X 10^{-16} .The same technique has been adopted here with minor variations in plotting the exon locations of lncRNAs.

While interpreting the exon plots of lncRNAs seen in this work, the two points to be noted are:

- Even the most accurate of exon prediction algorithms do not pin point the exon locations to the precision of a nucleotide.
- While interpreting graphs of power spectrum, generally, the values below half power are not considered as signals.

As seen from the sample exon plots in Figures 7.9 to 7.28, period three property in conjunction with digital filtering techniques can be used to locate the exons in lncRNA sequences.

While computing the STFT for obtaining the spectrum, optimum window lengths are mandatory in locating the exons from the sequences as they are in the case of coding DNA sequences.

A reduced computation technique which makes use of a quadratic function (using only T and G sequences) was used to compute the spectrum in our former work [George 2010]. This effectively reduced spectral noise with coding DNA sequences. When the same approach is used here with lncRNA sequences, it is found to insert spectral noise. The exon plot of CCEPR making use of this reduced computation technique is shown in Figure 7.30. The noise in the spectral plot is evident and the exon (1 to 2502, refer Figure 7.8) is not discernible from the noise. This means that T and G sequences are insufficient to represent the signal spectrum unlike the case of coding DNA sequences. This indicates the difference in spectral properties of coding and non coding sequences.

Certain lncRNA sequences contain exons but their coding ability is still not been confirmed yet due to a variety of reasons. Most of the lncRNAs were found to have low GC concentration when compared to coding sequences [Derrien 2012]. This also suggests poor coding capacity. But the lncRNAs viz. CAHM, ABALON, HEIH, NKILA, FOXC2-AS1, MHENCR, FALEC, HOTAIRM1 and CISTR have G-C concentrations above 50%. This could imply protein coding capacity. But the lack of introns and the lack of STOP codons suggest otherwise. Computational analysis of functional lncRNA has been reported to reveal lack of protein coding capacity and also was found to have similarities with 3’UTRs. Long ncRNA sequences have been found to possess low G-C content and scantiness of introns. In previous studies opening reading

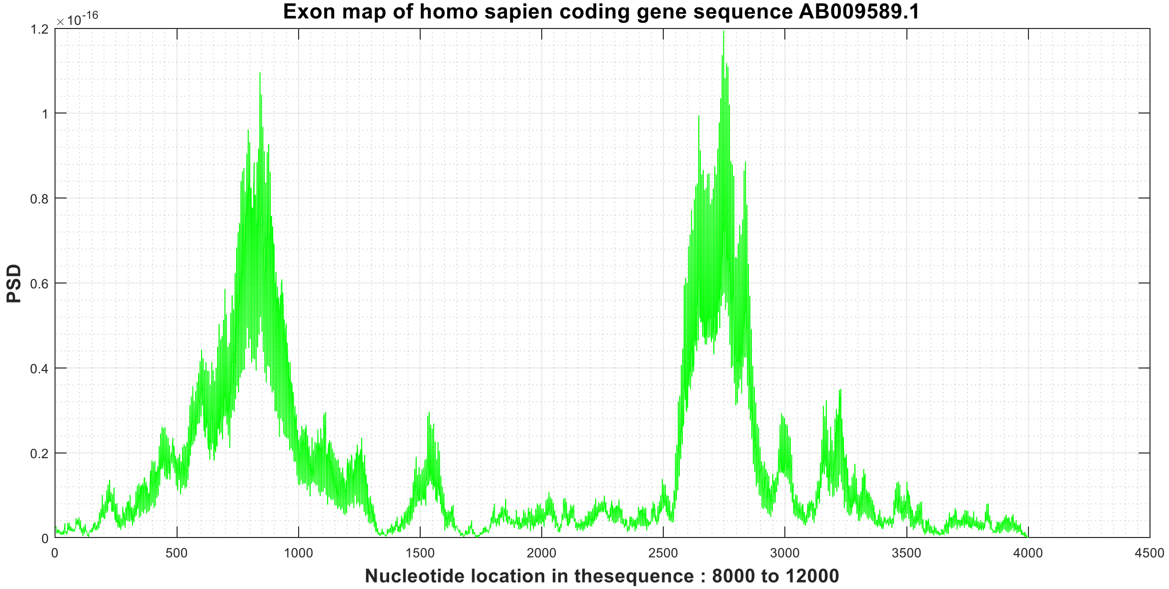


Figure 7. 29. Sample exon plot of a coding sequence. The exon plot of locations8000 to 11000 of Homo sapien gene for osteomodulin. GenBank accession number AB009589.1

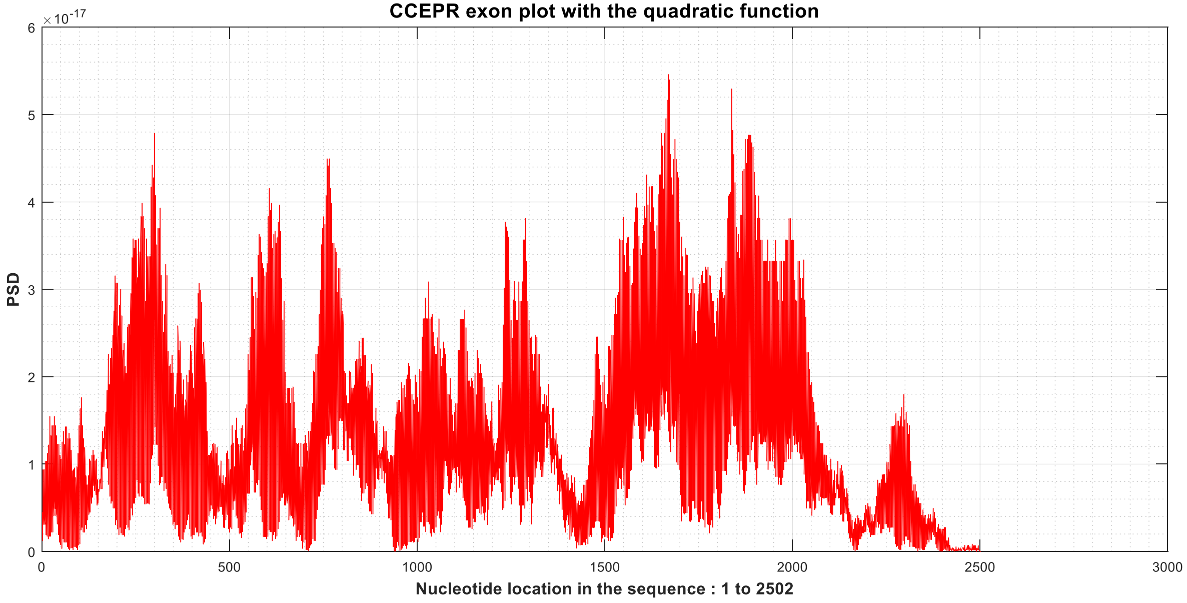


Figure 7.30. Sample exon plot of CCEPR with reduced computation technique. Exon plot of lncRNA CCEPR with GenBank accession no. NR_131782.1 obtained when the reduced computation technique is used

detected in some lncRNA sequences, but they have a poor start codon and ORF contexts which would make it unlikely for these lncRNAs to be protein coding [Niazi 2012]. The lncRNAs analysed in this work have very short or practically non-existent introns. These sequences have START codon corresponding to the exon locations mentioned in the reference database, they do not have any of the STOP codon patterns within them (UAA, UAG or UGA). Though such a stretch after the START codon might appear to be an ORF, there are no STOP codons, which would make it unlikely for the lncRNA to code for peptides. Most of long noncoding RNAs have been found to be spliced (98%) and they exhibit a striking bias towards two exon transcripts. 42% of lncRNAs have only two exons as against 6% of the protein coding genes [Derrien 2012]. But here, lncRNAs with more than 2 exons are also included in the study to establish the robustness of the algorithm in locating exons.

7.5. Conclusion

Exons of human long noncoding RNA sequences are predicted in this work by making use of the period 3 property which is a widely accepted approach for predicting exons in coding DNA sequences. The IIR anti-notch filter picks the spectral component at 2**π**/3, and de-noising of the exon map with DWT filter bank refines it as the noise is seen to occupy the HF part of the spectrum. For obtaining the spectrum the choice of the window used for computing the STFT is found to be crucial just as the case with coding DNA sequences. Window is to be selected based on the length of the sequence used. The reduced computation technique which makes use of the T and G binary sequences alone to compute the spectrum was found to suppress spectral noise with coding DNA sequences. But this is not so with lncRNA sequences. In this case the quadratic function introduces spectral noise. Thus it is clear that T and G binary sequences alone cannot represent the spectrum amply as in the case of coding DNA sequences. This indicates that the spectral properties of lncRNA sequences are different from those of coding DNA sequences. Long ncRNA sequences may contain information in their spectrum which could be made use of in further studies. Comparing the exon plots for lncRNA sequences (Figures 7.9 to 7.28) with that of the exon plot of a coding DNA sequence (Figure 7.28) it is clear that the algorithm based on period 3 property followed by digital filtering techniques can effectively be extended to locate exons in lncRNA sequences. NCBI records are the widely accepted reference data for many features of genomic sequences including exon locations in DNA/lncRNA sequences. The exons found in long ncRNA sequences analysed have lengths which are in the range 566 to 8779 nucleotides. Hence a deviation of 36 – 100 nucleotides can be taken as quite acceptable. This proves the correctness of the algorithm developed for exon prediction

Period 3 property which picks exons from coding DNA sequences has been used successfully in identifying genes [Tiwari 1997] from DNA sequences. On parallel logic, it is to be investigated whether the technique used in locating exons within long ncRNAs can be adapted to identify long ncRNAs themselves. This could be yet another area in which this work could be taken forward. However, this has multiple constraints as the functional implications of the exons present in long ncRNAs have not been fully unveiled yet.

There are many computational methods to predict functional features of long noncoding RNA that are listed in literature [Zhao 2014], [Signal 2016], [Perron 2017]. The former [Zhao 2014] details a method to predict long noncoding RNA functions based on a coding-noncoding gene co-expression network. Several in-silico methods for the prediction of function and characterisation of long ncRNAs are outlined by Signal et.al [Signal 2016]. Computational prediction of lncRNA function using tissue specific co-expression and from the genes in different species is detailed by Perron et.al. [Perron 2017]. The works mentioned above are just examples of methods to predict functions of long ncRNAs, it not an all-inclusive reference list.

The study presented here is not a method to identify lncRNA nor is it sufficient to predict regulatory properties/functions of long ncRNA. The signal processing technique that is widely used to locate exons in coding genes is used here to detect exons in lncRNA. The similarity/differences of lncRNA sequences with sequences of coding genes in terms of their spectral properties have been highlighted. Such a study which predicts exons in lncRNAs using signal processing principles was not found in literature. The novelty of the study is this very fact and hence a comparative study of this work with existing techniques is not presented. Signal processing methods are inherently easy to implement and robust. The authors expect that this novel approach to analyse long ncRNA would be helpful in bringing to light many of their sequence and spectral properties. The future direction of this work would be to explore the possibility of predicting the regulatory functions of long ncRNA from their sequence properties or by frequency domain analysis of sequences.

Chapter 8

Conclusion and the future scope of this work

8.1. Conclusion

The discovery of the double structure of DNA is the milestone in the history of science and gave rise to modern molecular biology. There has been dramatic progress in genomics in the last seven decades. We are now in the genomic era, with the human genome project completed in 2003. Today large amounts of genomic and proteomic data are available in the public domain and it is to be processed in ways which are beneficial to mankind. Genomic signal processing is primarily the processing of DNA sequences, RNA sequences, and proteins and other forms of genomic data. This is an area where traditional as well as modern signal processing methods can find wide application.

Noncoding RNA molecules were ignored for a long time in genome studies. But they have come to be of vital importance in both molecular biology as well as in genome studies as it has become evident that they play vital roles in many biological processes. The work presented in this Thesis analyses noncoding RNA sequences. Though computational techniques have been used for this purpose, analysis of noncoding region of the genome using DSP methods has not been reported in literature till date. In this work, we have analysed two types of noncoding RNA sequences using DSP methods.

In the first part of the work, a mathematical model was developed for MFE (minimum free energy) of the secondary structure of noncoding RNA sequences from their signal parameters viz. length and the spectral coefficient matrix making use of multiple linear regression analysis. This model was made use of to evaluate MFE without employing the folding algorithm. The correctness of the model was checked using standard webservers (RNAfold and RNAstructure). To begin with, MFE of noncoding RNA sequences was found out using the thermodynamic nearest neighbour algorithm. Multiple linear regression analysis was performed by considering MFE as the response variable sequence length and the standard deviation of the spectral coefficient matrix as the predictor variables. The method developed in computing MFE, making use of signal properties of the sequence is simpler and does not involve the folding algorithm followed in traditional computing methods.

The second part of this work analyses long noncoding RNA sequences. DSP methods which are used to locate exons in coding regions of the genome have been employed to identify exons in long ncRNAs. Period 3 property which picks exons from coding DNA sequences has been used successfully in identifying genes [Tiwari 1997] from DNA sequences. On parallel logic, it is to be investigated whether the technique used in locating exons within long ncRNAs can be adapted to identify long ncRNAs themselves. This could be yet another area in which this work could be taken forward. However, this has multiple constraints as the functional implications of the exons present in long ncRNAs have not been fully unveiled yet.

There are many computational methods to predict functional features of long noncoding RNA that are listed in literature. In this work, the signal processing technique that is widely used to locate exons in coding genes was used to detect exons in lncRNA. The similarity/differences of lncRNA sequences with sequences of coding genes in terms of their spectral properties have been highlighted. Such a study which predicts exons in lncRNAs using signal processing principles was not found in literature. This is the novelty of the study presented here

8.2. Future scope of this work

The first part of this work has brought to light the relationship between the thermodynamic entity MFE and the signal properties of ncRNA sequences. This shows that the noncoding genome too is conducive to analysis with DSP techniques. The easiness with which the algorithm developed here computes MFE unlike the traditional folding algorithms cannot be overlooked. Digital signal processing methods have the unique convenience of ease of implementation and lesser computational complexity. It is hoped that this novel relationship linking MFE with signal properties of the sequences can be taken forward so that more signal processing approaches to study noncoding RNA evolve.

In the second part of this work which analyses long noncoding RNA sequences, DSP methods which are used to locate exons in coding regions of the genome have been employed to identify exons in long ncRNAs. Period 3 property which picks exons from coding DNA sequences has been used successfully in identifying genes from lncRNA sequences. On parallel logic, it is to be investigated whether the technique used in locating exons within long ncRNAs can be adapted to identify long ncRNAs themselves.

REFERENCES

Akhtar 2008; Akhtar M, Julien Epps J, Ambikairajah E. Signal Processing in Sequence Analysis, Advances in Eukaryotic Gene Prediction . IEEE Journal of selected topics in signal processing. 2008;2(3):310-321

Alberts 1998; Alberts B, Bray D, Johnson A, Lewis J, Raff M, Roberts K, and Walter P, Essential Cell Biology, Garland Publishing Inc., New York, 1998.

Alberts 2007; Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of The Cell. Student Edition 2007. Garland Science, Taylor & Francis Group. NY, USA.

Alvarez 2011; Alvarez ML, Di Stefano JK. Functional characterization of the plasmacytoma variant translocation 1 gene (PVT1) in diabetic nephropathy. PLoS One 2011;6:e18671.

Anastassiou 2000; Anastassiou D. Frequency-domain analysis of bio-molecular sequences. Bioinformatics, Oxford Academic. 2000;16(12):1073-1081

Anastassiou 2001(1) (DSP); Anastassiou D. DSP in genomics : processing and frequency domain analysis of character strings. Proceedings, IEEE International Conference on Acoustics Speech and Signal Processing 2001, 2:1053 – 1056.

Anastassiou 2001(2) (GSP); Anastassiou D. Genomic Signal Processing. IEEE Signal Processing magazine 2001, 14 (4):8 – 20.

Barsheshet 2012; Barsheshet A, Brenyo A, Goldenberg I, et al. Sex-related differences in patients' responses to heart failure therapy. Nat Rev Cardiol 2012;9:234–42.

Bartel 2004; Bartel D P. MicroRNAs : Genomics, biogenesis, mechanism and function. Cell 2004; 116(2): 281 – 297

Bertone 2004; Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X et.al.. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004;306:2242–6.

Bhushanthi 2007; N Bushanthi, S M Cohen, micro RNA functions. Annual Review of Cell Devolopment Biology 2007; DOI: 10.1146/annurev.cellbio.23.090506.123406

Boon 2016; Boon RA, Jae N, Holdt L , Dimmeler S. Long Noncoding RNAs From clinical Genetics to Therapeutic Targets? Journal of the American college of Cardiology (Elsevier). 2016; 67(10): 1214-1226.

Brannan 1990; Brannan CI, Dees EC, Ingram RS, Tilghman SM. The product of the H19 gene may function as an RNA. Molecular Cell Biology. 1990;10:28–36.

Brazao 2016; Brazao TF, Johnson JS, Muller J, Heger A, Ponting CP, Victor LJ et.al. Long noncoding RNAs in B-cell development and activation. doi: 10.1182/blood-2015-11-680843

Brimacombe 1985; Brimacombe R and Stiege W. Structure and function of ribosomal RNA. Biochemical Journal. 1985(229):1-17

Brockdorff 1992; Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ et.al. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell. 1992; 71:515–26.

Brosnan 2009; Brosnan CA, Vionnet O. The long and short of noncoding RNAs. Curent Opinion in Cell Biology, Elsevier. 2009;21(3):416-425

Brown 2008; Brown JWS, Marshall D F, Echeverrria. Intronic noncoding RNAs and splicing. Cell Review 2008; doi:10.1016/j.tplants.2008.04.010

Bustin 1973; Bustin M: Arrangement of histones in chromatin. Nature New Biology. 1973;245:207–209.doi:10.1038/newbio245207a0

Cabili 2011; Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25:1915–27.

Cai 2009; Cai Y, Yu X, Hu S. A Brief Review on the Mechanisms of miRNA Regulation. Genomics, Proteomics & Bioinformatics. Elsevier 2009;7(4):147-154

Calabrese 2013; Calabrese JM, Magnuon T. Roles of Long Non-coding RNAs in X-Chromosome Inactivation. Molecular Biology of Long Non-coding RNAs. 2013. DOI: 10.1007/978-1-4614-8621-3_3.

Cao 2014; Cao J. The functional role of long non-coding RNAs and epigenetics. BMC Biological Procedures Online. 2014;16:11.

Carninci 2005; Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science. 2005;309:1559–63.

Carninci 2009; Carninci P. Molecular biology: the long and short of RNAs. Nature. 2009;457(7232):974-975

Cathcart 2015; Cathcart P, Lucchesi W, Ottaviani S, De Giorgio A, Krell J, Stebbing J, Castellano L. Noncoding RNAs and the control of signalling via nuclear receptor regulation in health and disease. Best Practice & Research Clinical Endocrinology & Metabolism. Elsevier 2015;29:529–43.

Cech 2014; Cech TR, Steitz JA. The non-coding RNA revolution – Trashing old rules to forge new ones. Cell 2014, 157(1):77 – 94.

Chang 2012; Chang CP, Bruneau BG: Epigenetics and cardiovascular development. Annu Rev Physiol 2012, 74:41–68.

Chatterjee 2012; Chatterjee S, Hadi AS. Regression analysis by example. Wiley Series in Probability and Statistics. Fifth Edition. 2012

Chen 2010; Chen LL, Carmichael CG. Decoding the function of nuclear long non-coding RNAs. Current Opinion Cell Biology. 2010;22:357-364.

Chen 2008; Chen SJ. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu Rev Biophys 2008, 37:197–214.

Chen 2013; Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X et.al.. LncRNA disease : a database for long nono-coding RNA-associated diseases. Nucleic Acids Research 2013;41:D983-6

Chen 2014; Chen LL, Zhao JC. Functional analysis of Long Noncoding RNAs in Development and Diseae. In : Yeo G (eds) Systems Biology of RNA Binding Proteins. Advances in Experimental Medicine and Biology. 2014; 825. Springer. New York.

Chen 2016; Chen X, Yan CC, Zhang X, You ZH. Long non-coding RNAs and complex diseases: from experimental results to computational models. Oxford. Briefings in Bioinformatics. 2017;18(4):558–576

Chen 2017; Chen X, Yan CC, Zhang X, You Z. Long non-coding RNAs and complex diseases: from experimental results to computational models. Briefings in Bioinformatics. Oxford Academic 2017;18(4):558–576

Choy 2006; Choy KW, Wang CC, Ogura A, Lau TK, Rogers MS, Ikeo K et.al. Genomic annotation of 15,809 ESTs identified from pooled early gestation human eyes. Physiol Genomics. 2006; 25:9 –15

Clark 2011; Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL et.al.. The reality of pervasive transcription. PLoS Biology. 2011; https://doi.org/10.1371/journal.pbio.1000625

Clote 2005; Clote P, Ferré F, Kranakis E et.al. Structural RNA have more folding energy than random RNA of the same dinucleotide frequency. RNA 2005, 11:578–591

Cosic 1994; Cosic I. Macromolecular Bioactivity: Is it resonant interaction between macromolecules? Theory and Applications. IEEE Transactions on Biomedical Engineering. 1994; 41:1101-1114

Cristea 2002(1); Cristea PD. Genetic signal representation and analysis. Proc. SPIE Inter. Conf. on Biomedical Optics 2002;4623:77–84.

Cristea 2002(2); Cristea PD. Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine. 2002;6(2):279-303.

Cristea 2005; Cristea PD. Representation and analysis of DNA sequences in Genomic signal processing and statistics. EURASIP Book Series in Signal Processing and Communications. 2005;2:15-66

Cucanelli 2015; Cusanelli E, Chartrand P. Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity. Front. Genet., 2015. https://doi.org/10.3389/fgene.2015.00143

DeOcesano-Pereira 2014; DeOcesano-Pereira C, Amaral MS, Parreira KS, Ayupe AC, Jacysyn JF, Amarante-Mendes GP. Long non-coding RNA INXS is a critical mediator of BCL-XS induced apoptosis. Nucleic Acids Res. 2014;42:8343–55.

Derrien 2012; Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H et.al. The GENCODE v7 catalog of human long noncoding RNAs : Analysis of their gene structure, evolution and expression. Cold Spring Harbor Laboratory Press. 2012;22:1775–1789.

Dieci 2009, G Dieci, Preti M, Montanini B. Eukaryotic snoRNAs: A paradigm for gene expression flexibility. ScienceDirect 2009; 94(2):83–88

Dimitrova 2014; Dimitrova N, Zamudio JR, Jong RM, Soukup D, Resnick R, Sarma K, et.al.. LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. Mol Cell. 2014;54:777–90.

Ding 2006; Ding Y. Statistical and Bayesian approaches to RNA secondary structure prediction. RNA 2006, 12:323–331.

Ding 2001; Ding Y, Lawrence C. Statistical prediction of single stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. Nucleic Acids Research 2001. 29: 1034–1046.

Ding 2003; Ding Y, Lawrence C. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Research 2003. 31: 7280–7301.

Ding 2005; Ding y, Chan CY, Lawerence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. CSHL –RNA. 2005; 11:1157-1166

Dinger 2008; Dinger ME, Pang KC, Mercer TR, Mattick J. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. PLOS Computational Biology. 2008 http://dx.doi.org/10.1371/journal.pcbi.1000176

Dirks 2007; Dirks RM, Bois JS, Schaeffer JM, Winfree R, Pierce NA. Thermodynamic Analysis of Interacting Nucleic Acid Strands. Society for Industrial and Applied Mathematics Review . 2007;49(1):65–88.

Dorn 2014; Dorn M, e Silva MB, Buriol LS, Lamb LC. Three-dimensional protein structure prediction: Methods and computational strategies. ScienceDirect 2014;53(B):251-276

Dougherty 2005; Edited by Dougherty ER, Smulevich I, Chen J, Wang ZJ. Genomics Signal Processing and Statistics. EURASIP Book Series on Signal Processing and Communications. 2005, Volume 2.

Dreyfus 2002; Dreyfus S. Richard Bellman on the birth of dynamic programming. INFORMS. 2002;50(1);48-51.

Eddy 2001; Eddy SR 2001. Non–coding RNA genes and the modern RNA world. Nature Reviews Genetics. 2001;2(12):919-929

Eddy 2002; S.R. Eddy. Computational genomics of noncoding RNA genes. *Cell* 2002; vol. 109(2):137-140.

Eddy 2014; Eddy SR. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annual Review Biophysics. 2014;43:433-456

Engstrom 2006; Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G et.al.. Complex Loci in human and mouse genomes. PLoS Genetics. 2006;2:e47.

Erdmann 2001; Erdmann VA, Barciszewska MZ, Szymanski M, Hochberg A, Groot N, Barciszewski J. The non-coding RNAs as riboregulators. Nucleic Acids Research. 2001;29:189-193

Esau 2007; Esau CC, Monia B P. Therapeautic potential for microRNAs. Advanced Drug Delivery Reviews 59. Elsevier 2007; 101–114

Esteller 2011; Manel Esteller. Non-coding RNAs in human disease. Nature Reviews. Genetics. 2011;12:861-874

Faghihi 2008; Faghihi MA, Modarresi F, Khalil A et al.. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. Nature Medicine. 2008;14:723 – 730

Fang 2016; Fang W, Fullwood MJ. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. Elsevier. Genomics Proteomics and Bioinformatics. 2016;14(1):42-54

Fire et.al. 1998; Fire A, Xu S, Montgomery MK, Kostas SA, Samuel E. Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans.* Letters to Nature. 1998;391

Flynn 2014; Flynn RA, Chang HY. Long noncoding RNAs in cell-fate programming and reprogramming. Cell Stem Cell. 2014;14:752–61.

Fontana 2002; Fontana W. Modelling 'evo-devo' with RNA. Wiley Periodicals BioEssays 2002, 24:1164–1177.

Fox 2004; Fox TW, Carreira A. A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression. EURASIP Journal on Applied Signal Processing 2004:1, 108–114

Frazen 2011; Franzen S. Chapter 15; Thermodynamics of Nucleic Acid Structural Modifications for Biotechnology Applications. BIOTECHNOLOGY OF BIOPOLYMERS. 2011:299.

Fu 2015; Fu W-M, Lu Y-F, Hu B-G, Liang W-C, Zhu X, Yang H-D et.al.. Long noncoding RNA Hotair mediated angiogenesis in nasopharyngeal carcinoma by direct and indirect signalling pathways. Oncotarget. 2015;7:4712.

Galzitskaya 1998; Galzitskaya OV and Finkelstein AV. Folding rate dependence on the chain length for RNA-like heteropolymers. Folding & Design 1998, 3:69–78.

Gardner 2004; Gardner PP, Giegerich R. A comprehensive comparison of RNA structure prediction approaches. BMC Bioinformatics. 2004;5:140

Garst 2011; Garst AD, Edwards AL, and Batey RT. Riboswitches : Structures and Mechanism. Cold Spring Harbor. Perspectives in Biology. 2011; doi: 10.1101/cshperspect.a003533

GENCODE v27; https://www.gencodegenes.org/releases/current.html

Genome Browser; https://genome.ucsc.edu/

George 2010; George TP, Thomas T. Discrete wavelet transform de-noising in eukaryotic gene splicing. BMC Bioinformatics 2010. 11(Suppl 1):S50 doi: 10.1186/1471-2105-11-S1-S50

George 2016; George TP, Thomas T. Novel Approach to analyzing MFE of Noncoding RNA Sequences. Genomics Insights. 2016;9:41–49 doi:10.4137/GEI.S39995

George 2017; George TP, Thomas T. Exon mapping in long noncoding RNAs using digital filters. Genomic Insights. 2017;10:1-12

Gibb 2011; Gibb EA, Brown CJ, Lam WL: The functional role of long non-coding RNAin human carcinomas. Mol Cancer 2011, 10:38–55.

Gisela 2002; Gisela S. An expanding universe on noncoding RNAS. Science 2002, 296(5571):1260 -1263.

Gottardo 2007; Gottardo F, Lui C G, Ferracin M, Calin G A, Fassan M, Bassi P et.al. Micro-RNA profiling in kidney and bladder cancers. Urologic Oncology: Seminars and Original Investigations. 2007;25(5):387-392.

Gottesman 2002; Gottesman S. Stealth regulation: biological circuit switch small RNA switches. Genes & Development. 2002;16:2829-2842
.

Grüner 1996; Grüner W, Giegerich R, Strothmann D et.al. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration I. Neutral Networks. Chemical Monthly 1996. 127; 355-374.

Guan 2007; Guan Y, Kuo WL, Stilwell JL, Takano H, Lapuk AV, Fridlyand J et al. Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. Clinical Cancer Research. 2007 Oct 1;13(19):5745-5755. Available from, DOI: 10.1158/1078-0432.CCR-06-2882

Gutschner 2012; Gutschner T, Diederichs S. The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol. 2012;9:703–19.

Hajiaghayi 2012; Hajiaghayi M, Condon A, Hoos HH. Analysis of energy-based algorithms for RNA secondary structure prediction. BMC Bioinformatics. 2012;13:22 http://www.biomedcentral.com/1471-2105/13/22

Han 2014; Han Y, Ye J, Wu D, Wu P, Chen Z, Chen J et.al.. LEIGC long noncoding RNA acts as a tumor suppressor in gastric carcinoma by inhibiting the epithelial-to-mesenchymal transition. BMC Cancer. 2014;14:932.

Hanahan 2000; Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57–70.

Hanahan 2011; Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.

Hannon 2004; He L, Hannon J. MicroRNAs : Small RNAs with a big role in gene regulation. Nature Rev. Genetics 2004, 5(7):522 – 531.

Hannon 2006; Hannon GJ, Rivas FV, Murchison EP, Steitz JA. The expanding universe of noncoding RNAs. Cold Spring Harbour Symposium on Quantitative Biology. 2006;71:551-564

Harries 2012; Harries LW. Long noncoding RNAs and human disease. Biochemical Society Transactions. 2012;40(4).

Harrow 2012; Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski et.al. GENCODE: The reference human genome annotation for The ENCODE Project. Cold Spring Harbor Laboratory Press. 2012; 22:1760–1774.

He 2004; He L and Hannon J. MicroRNAs: Small RNAs with a big role in gene regulation. Nature Review Genetics. 2004;5(7):522-31.

Herzela 1998; Herzel H, Trifonov EN, Weiss O, Gro*β*e I. Interpreting correlations in biosequences. Physica A, Statistical Mechanics and it's Applications. 1998;249:449–459

Hofacker 2002; Hofacker IL and Stadler PF. RNA Secondary Structures. Journal of Molecular Biology 2002, 319(5):1059–1066.

Hu 2012; Hu W, Alvarez-Dominguez JR, Lodish HF. Regulation of mammalian cell differentiation by long non-coding RNAs. EMBO Rep. 2012;13:971–83.

Hung 2010; Hung T, Chang HY. Long noncoding RNA in genome regulation : prospects and mechanisms. RNA Biology. 2010;7:582-585

Huynen 1997; Huynen M, Gutellz R, Koningsy D. Assessing the reliability of RNA folding using statistical mechanics. Journal of Molecular Biology 1997, 267(5):1104-12.

Iyer 2015; Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et.al.. The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics. 2015;47:199–208.

Jansson 2012; Jansson M D, Lund A H. MicroRNA and cancer. Molecular Oncology. Elsevier. 2012; http://dx.doi.org/10.1016/j.molonc.2012.09.006

Johnson 2012; Johnson R. Long non-coding RNAs in Huntington's disease neurodegeneration. Neurobiology of Disease. 2012;46(2):245–254.

Joung 2017; J. Joung J, Engreitz JM, Konermann S, Abudayyeh OO, Verdine VK, Aguet F et.al.. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. Nature 2017;548(7667):343-346

ISU Lecture Notes. Illinois State University. Available at : http://chemistry.illinoisstate.edu/standard/che360/handouts/360vanthoff.pdf

Kakumani 2008; Kakumani R. Prediction of Protein-Coding Regions in DNA Sequences Using a Model-Based Approach. International Symposium on Circuits and Systems IEEE. 2008;1918 - 1921, DOI: 10.110/ ISCAS. 2008. 4541818

Kanduri 2006; Kanduri C, Thakur N, Pandey RR. The length of the transcript encoded from the *Kcnq1ot1* antisense promoter determines the degree of silencing. The EMBL Journal. 10.1038/sj.emboj.7601090

Kanhere 2010; Kanhere A, Viiri K, Araújo CC, Rasaiyaah J, Bouwman RD, Whyte WA et.al. Short RNAs Are Transcribed from Repressed Polycomb Target Genes and Interact with Polycomb Repressive Complex-2. 2010;38(5):675-688

Kapranov 2007; Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT et.al.. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007;316:1484–8.

Kapusta 2014; Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends in Genetics. 2014;30(10):439-452.

Karl 2012; Karl F. A Free Energy Principle for Biological Systems. Entropy (Basel). 2012;14(11):2100-2121.

Kim 2010; Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–187.

Kirchner 2001; Kirchner Data Analysis Toolkit #10: Simple linear regression. , Prof. James Kirchner. University of Berkeley. 2001. Available at: http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Kiss 2002; T Kiss. Small nucleolar RNAs: an abundant group of non-coding RNAs with diverse cellular functions. Cell. 2002; 109(2) : 145-148.

Kitagawa 2013; Kitagawa M, Kitagawa K, Kotake Y, Niida H, Ohhata T. Cell cycle regulation by long non-coding RNAs. Cell Mol Life Sci. 2013;70:4785–94.

Koike 2013; Koike K, Kasamatsu A, Iyoda M, Saito Y, Kouzu Y, Koike H et.al. High prevalence of epigenetic inactivation of the human four and a half LIM domains 1 gene in human oral cancer. Int J Oncol 2013, 42:141–150.

Kornieko 2016; Kornieko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C et.al. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. BMC Genome Biology. 2016;17:4, DOI 10.1186/s13059-016-0873-8

Kryger 2012; Kryger R, L. Fan L, P. A. Wilce PA, and V. Jaquet V. MALAT-1, a non protein-coding RNA is upregulated in the cerebellum, hippocampus and brain stem of human alcoholics. Alcohol. 2012;46(7):629–63

Kumarswamy 2014; Kumarswamy R, Bauters C, Volkmann I, et al. Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure. Circ Res 2014;114:1569–75.

Kung 2013; Kung JTY, Colognori D, Lee JT. Long Noncoding RNAs: Past, Present and Future. Genetics 2013;193:651–669.

Kwan 2009; Kwan HK, Arniker SB. Numerical representation of DNA sequences. Proceedings of IEEE Inter, Conf. on Electro/Information Technology, EIT. 2009:307-310

Lanz 1999; Lanz RB, McKenna NJ, Onate SA, Albrecht U, Wong J, Tsai SY et.al.. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell. 1999;97:17–27.

Lee 2013; Lee SJ, Son S, Yhee JY, Choi K, Kwon IC, Kim SH, Kim K. Structural modification of siRNA for efficient gene silencing. Biotechnology Advances. 2013;13(5):491-503

Li 1997; Li W. The study of correlation structures of DNA sequences: a critical review. Computers and Chemistry 1997;21(4):257-271

Li 2013; Li J, Xuan Z, Liu C. Long non-coding RNAs and complex human diseases. International Journal of Molecular Sciences. 2013 Sep 12;14(9):18790-808. doi: 10.3390/ijms140918790.

Li 2015; Li CH, Chen Y. Small and Long Non-Coding RNAs: Novel Targets in Perspective Cancer Therapy. Current Genomics. 2015;16(5):319-326

Li 2017; Li S, Li B, Zheng Y. Exploring functions of long noncoding RNAs across multiple cancers through co-expression network. Nature Scientific Reports. 2017;754. Doi : 10.1038/s41598-017-00856-8

Lim 2003; Lim LP, Glasner ME, Yekta S. Vertebrate micro RNA genes. Science. 2003; 299(5612):1540.

Ling 2013; Ling H, Fabbri M, Calin GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. Nature Reviews Drug Discovery. 2013;12:847-865

Liu X 2012; Liu X, Li D, Zhang W, Guo M, Zhan Q. Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. EMBO J. 2012;31:4415–27

Lodish 2000; Lodish H, Berk A, Zipursky S L, Matsudaira P, Baltimore D, Darnell J. Molecular Cell Biology. W H Freeman, New York. 2000.

Loewer 2010; Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH et.al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Mature Genetics. 2010;42:1113-1117.

Louro 2009; Louro R, Smirnova AS, Verjovski-Almeida S. Long intronic noncoding RNA transcription: Expression noise or expression choice? Elsevier Genomics Review. 2009;93(4):291-298.

Lu 2013; Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X et.al.. Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics. 2013 Sep 24;14:651. doi: 10.1186/1471-2164-14-651.

Ma 2013; Ma L, Bajic VB, Zhang Z. On the classification of long non-coding RNAs. RNA Biology. 2013;10: 925–33.

Mallat 2009; Mallat S. A Wavelet Tour of Signal Processing: The Sparse Way. Elsevier 2009. ISBN 13:978-0-12-37437--1

Malone 2009; Malone CD, Hannon GJ. Small RNAs as Guardians of the Genome. Elsevier. Cell. 2009;136:656-668

Markham 2008; Markham NR, Zuker M. UNAFold: Software for nucleic acid folding and hybridization. Methods Mol Biol. 2008;453:3–31.

Matera 2007; Matera A G, Terns RM, Terns MP. Non-coding RNAs : lessons from the small nuclear and small nucleolar RNAs. Nature Reviews, Molecular Biology 2007;8:209 – 220.

Mathews 1999; Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology 1999;288(5): 911-40.

Mathews 2004; Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization in RNA 2004, 10:1178–1190.

Mathews 2010; Mathews DH, Moss WN, Turner DH. Folding and Finding RNA Secondary Structure. Cold Spring Harb Perspect Biol 2010;2:a003665

Matsui 2017; Matsui M, Corey DR. Non-coding RNA as drug targets. Nature Reviews drug Discovery. 2016;(16):167-197

Mattick 2006; Mattick JS and Makunin IV. Non-coding RNA. Human Molecular Genetics 2006;Vol. 15, Review Issue 1 R17–R29doi:10.1093/hmg/ddl046.

Matzke 2005; Matzke MA and Birchler JA. RNAi-mediated pathways in the nucleus. Nature Review. Genetics. 2005;6(1):24-25

McCaskill 1990; McCaskill J. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 1990, 29:1105–1119.

McManus 2002; McManus MT, Sharp PA. Gene silencing in mammals by small interfering RNAs. Nature Review Genetics 2002, 3(10).

Mercer 2009; Mercer TR, Dinger ME, Mattick JS. Long noncoding RNAs: insights into functions. Nature Reviews Genetics 2009;10:155–9.

Mercer 2013; Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nature Structural and Molecular Biology 2013;20:300-307

Michalik 2014; Michalik KM, You X, Manavski Y, Doddaballapur A, Zornig M, Braun T et.al.. Long noncoding RNA MALAT1 regulates endothelial cell function and vessel growth. Circ Res. 2014;114:1389–97

MIT Lecture notes; http://math.mit.edu/classes/18.417/Slides/rna-prediction-nussinov.pdf

MIT Open Course Ware. Available at: http://ocw.mit.edu/courses/biology/7-51-graduate-biochemistry-fall-2001/lecture-notes/fa01lec06.pdf

Montgomery 2006; Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. Wiley Student Edition 2006.

Morris 2015; Morris KV, J.S. Mattick JS. The rise of regulatory RNA. Nat. Rev. Genet. 2014;15: 423–437.

Muhammad 2004; Sohail Muhammad, ed. Gene silencing by RNA interference: technology and application. crc press, 2004

Mus 2007; Mus E, Hof PR, Tiedge H. Dendritic BC200 RNA in aging and in Alzheimer’s disease. Proc Natl Acad Sci USA 2007;104: 10679–84.

Nair 2006; Nair AS, Pillai SS. A coding measure scheme employing electron-ion interaction pseudo potential (EIIP)’. Bio-information. 2006; 1: 197-202.

Napoli 1990; Napoli C, Lemieux C, Jorgensen R. lntroduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes. Nature Rev. Genetics. 2004;5(7):522-531

Nature Website; https://www.nature.com/scitable

NCBI Genome Resource; https://www.ncbi.nlm.nih.gov/genome

NCBI Website; NCBI GenBank: https://www.ncbi.nlm.nih.gov/nucleotide

Ng 2013; Ng S-Y, Lin L, Soh BS, et al. Long noncoding RNAs in development and disease of the central nervous system. Trends Genet 2013;29:461–8.

NHGRI Website (HGP); https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/

Niazi 2012; Niazi F, Vlaladkhan S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3’ UTRs. RNA CSHL Press. 2012; 8:825–843.

NIH Website; https://www.genome.gov/18016863/a-brief-guide-to-genomics/

NLM Website#hgp : https://ghr.nlm.nih.gov/primer#hgp

Novina 2004; C D Novina, Sharp PA. The RNAi revolution. Nature 2004, 30:161 – 164.

Nussinov 1978, Nussinov R, Piecznik G, Grigg JR and Kleitman DJ : Algorithms for loop matchings. SIAM Journal on Applied Mathematics 1978

Nussinov 1980, R Nussinov and A B Jacobson Fast algorithm for predicting the secondary structure of single-stranded RNA. PNAS, Proc Natl Acad Sci U S A. Nov 1980; 77(11): 6309–6313

O’Gorman 2006; O’Gorman W, Kwek KY, Thomas B, Akoulitchev A. Non-coding RNA in transcription initiation. Biochemical Society Symposia 2006; DOI: 10.1042/bss0730131

Offit 2014; Offit K. Decade in review–genomics: a decade of discovery in cancer genomics. Nature Review Clinical Oncology. 2014;11:632–4.

Okazaki 2002; Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature Genetics. 2002;420:563–73

Oppenheim 2009; Oppenheim AV, Schafer R W. Discrete-Time Signal Processing. 3rd edition. Prentice Hall Signal Processing Series 2009.

Padgett 2015; Padgett R A. mRNA splicing : Role of snRNAs. John Wiley & Sons Limited. 2005; http://www.els.net [doi: 10.1002/9780470015902.a0000879.pub3]

Palop 2006; Palop JJ, Chin J, Mucke L. A network dysfunction perspective on neurodegenerative diseases. Nature 2006;443:768–73.

Papait 2013; Papait R, Kunderfranco P, Stirparo GG, et al. Long noncoding RNA: a new player of heart failure? J Cardiovasc Transl Res 2013;6:876–83.

Pauli 2011; Pauli A, Rinn JL, Schier AF. Noncoding RNAs as regulators of embryogenesis. Nature Review Genetics. 2011;12:136-149

Pedersen 2000; Pedersen CNS. Algorithmns in Computational Biology. Doctoral Thesis. Available at: http://www.brics.dk/DS/00/4/BRICS-DS-00-4.pdf

Peng 2016; Peng W, Fang H. Long noncoding RNA CCHE1 indicates a poor prognosis of hepatocellular carcinoma and promotes carcinogenesis via activation of the ERK/MAPK pathway. Biomedicine and Pharmacotherapy. Elsevier 2016;83:450-45

Perkel 2013; Perkel J M. Visiting "noncodarnia". Biotechniques. 2013; doi: 10.2144/000114037.

Perron 2017; Perron U, Provero P, Molineris I. In silico prediction of lncRNA function using tissue specific and evolutionary conserved expression. BMC Bioinformatics. 2017; DOI: 10.1186/s12859-017-1

Perry 2016; Perry RB-T, Ulitsky I. The functions of long noncoding RNAs in development and stem cells. Development. 2016;143:3882-3894; doi: 10.1242/dev.140962

Pervouchine 2003; Pervouchine DD, Graber JH and Kasif S. On the normalization of RNA equilibrium free energy to the length of the sequence. Nucleic Acids Research 2003, 31(9): e49. doi : 10.1093/nar/gng049.

Pink 2011; Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DRF. Pseudogenes: Pseudo-functional or key regulators in health and disease? RNA. CSHL Press. 2011; 17(5):792-8.

Piao 2017; Piao M, Sun L, Zhang QC. RNA Regulations and Functions Decoded by Transcriptome-wide Structure Probing. Elsevier Genomics Proteomics Bioinformatics 2017:15 ;267–278

Poliseno 2010; Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465:1033–8.

Ponting 2009; Ponting CP, Oliver PL, Reik W. Evolution and Functions of Long Noncoding RNAs. Cell 2009;136,629-641.

Ponting 2010; Ponting CP, Belgard TG. Transcribed dark matter: meaning or myth? Human Molecular Genetics. 2010; doi:10.1093/hmg/ddq362

Pray 2008; Pray L. Discovery of DNA structure and function: Watson andCrick. Nature Education 2008;1(1):100

Prioleau 1994; Prioleau MN, Huet J, Sentenac A, Méchali M: Competition between chromatin and transcription complex assembly regulates gene expression during early development. Cell 1994, 77:439–449.

Privalov 2015; Privalov PL. Microcalorimetry of Macromolecules: The Physical Basis of Biological Structures. Journal of Solution Chemistry. 2015;44: 1141. doi:10.1007/s10953-015-0337-x

Proakis 2006; Proakis JG, Manolakis DK. Digital Signal Processing. 4th edition. Pearson New Delhi. 2006.

Provazník 2012; Provazník I, Kubicová V, Škutková H, Nedvěd J, Tkacz E, Babula P et.al.. Detection of short exons in DNA sequence using complex wavelet transform of structural features. Proceedings IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS). 2012; DOI: 10.1109/GENSIPS.2012.6507740

Quinn 2016; Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nature Reviews Genetics 2016;17:47-62.

Qureshi 2010; Qureshi IA, Mattick, Mehler MF. Long non-coding RNAs in nervous system function and disease. Brain Research. 2010;1338,20–35

Qureshi 2012; Qureshi IA, Mehler MF. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nature Reviews Neuroscience 2012;13(8):528–541.

Rao 2008; Rao KD, Swamy MNS. Analysis of Genomics and Proteomics Using DSP Techniques. IEEE Transactions on Circuits And syatems-I:Regular Papers. 2008;55(1):370378.

Reddy 1998; Reddy R, Busch H. Small Nuclear RNAs : RNA sequences, Structure, and Modifications. In: Birnstiel ML. (eds) Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles. Springer 1998, Berlin, Heidenberg.

Redon 2010; Redon S, Reichenbach P, Lingner J. The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. Nucleic Acids Res. 2010;38:5797–806.

Ren 2014; Ren X, Ustiyan V, Pradhan A, Cai Y, Havrilak JA, Bolte CS et. al. FOXF1 Transcription Factor Is Required for Formation of Embryonic Vasculature by Regulating VEGF Signaling in Endothelial Cells. Circulation Research. 2014;115:790-720

Rfam/Pfam database; http://rfam.xfam.org/

Rinn 2012; Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annual Review Biochem. 2012; 81:145–66.

Rohatgi 2000; Rohatgi VK, A K Md. E. Saleh. An Introduction to Probability and Statistics. Wiley Series in Probability and Statistics. 2nd Edition 2000

Rossi 2014; Rossi MN, Antonangeli F. LncRNAs: new players in apoptosis control. International Journal of Cell Biology . 2014;2014:473857.

Russell 2010; Russell PJ: iGenetics. 3rd ed. San Francisco: Pearson Benjamin Cummings. Expert Opin Biol. 2010;2:217–230.

Saffhill 1975; Saffhill R, Itzhaki RF: Accessibility of chromatin to DNA polymerase I and location of the F1 histone. Nucleic Acids Res 1975, 2:113–119.

Salta 2012; Salta E, De Strooper B. Non-coding RNAs with essential roles in neurodegenerative disorders. The Lancet Neurology. 2012;11(2):189–200

Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: a laboratory manual. Cold spring harbor laboratory press. Edition 2. 1989.

Sanford 2005; Sanford W. Applied Linear Regression. Wiley Series in Probability and Statistics. 3rd Edition 2005.

Schonrock 2012; Schonrock N, Harvey RP, Mattick JS. Long noncoding RNAs in cardiac development and pathophysiology. Circ Res 2012;111:1349–62.

Scott 2011; Scott MS, Ono M. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. Biochime. 2011; 93(11):1987-1992

Shi 2001; Shi Y, Downes M, Xie W, Kao HY, Ordentlich P, Tsai CC et.al.. Sharp, an inducible cofactor that integrates nuclear receptor repression and activation. Genes Dev. 2001;15:1140–51

Signal 2016; Signal B, Gloss BS, Dinger ME. Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. Trends in Genetics. 2016;32(10):620-637

Smola 2016; Smola M, Christy TW, Inoue K, Nicholson CO, Friedersdorf M, Keene JD et.al. SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. PNAS 2016;113(37):10322-103227

Soman 2004; Soman KP and Ramachandran KI: Insights into Wavelets – From Theory to Practice. Prentice - Hall of India Pvt. Ltd; Second Edition 2004

Song 2010; Song NY, Yan H. Short Exon Detection in DNA Sequences Based on Multifeature Spectral Analysis. EURASIP Journal on Advances in Signal Processing. 2010;2011:780794

Sykes 1993; Sykes AO. An Introduction to regression analysis. The Inaugural Coase Lecture, University of Chicago. 1993. Available at: http://chicagounbound.uchicago.edu/law_and_economics/51/

Taft 2009; Taft RJ, Glazov EG, Lassman T, Hayashizaki Y, Carminci P, Mattick JS. Small RNAs derived from snoRNAs. RNA 2009;15(7):1233-1240.

Tan 2013; Tan L, Yu J-T, Hu N, et al. Non-coding RNAs in Alzheimer's disease. Mol Neurobiol 2013;47:382–93.

Tinoco 1999; Tinoco I Jr and Bustamante C. How RNA folds. Journal of Molecular Biology (JOMB) 1999; 293(2), 271

Tiwari 1997; Tiwari S, Ramachandran S, Bhattachrya A, Bhattacharya S, Krishnaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS*. 1997;13(3):263-270

Tollervey 1997; Tollervey D, Kiss T. Current Opinion in Cell Biology. ScienceDirect 1997; 9(3): 337–342 (page 11)

Tran 2009; Tran TTT, Zhou F, Marshburn S, Stead M, Kushner SR, Xu Y. *De novo* computational prediction of non-coding RNA genes in prokaryotic genomes. Bioinformatics. Oxford Journals. 2009;25(22): 2897–2905.

Trifonov 1980; Trifonov EN, Sussman JL. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proceedings of the National Academy of Science USA. 1980;77:3816–3820

Trotta 2014; Trotta E. On the Normalization of the Minimum Free Energy of RNAs by Sequence Length. PLoS ONE 2014, 9(11): e113380. doi:10.1371/journal.pone.0113380

Trout and Tester; Trout B, and Jefferson Tester J. *10.40 Chemical Engineering Thermodynamics.* Fall 2003. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Vaidyanathan 2002; Vaidyanathan PP, Yoon BJ. Gene and exon prediction using allpass-based filters. Proceedings IEEE Workshop on Genomics Signal Processing and Statistics. 2002.

Vaidyanathan 2004; Vaidyanathan PP, Yoon B-J. The role of signal-processing concepts in genomics and proteomics. Journal of the Franklin Institute 2004; 341(1):111-135.

Valadkhan 2010; Valadkhan S, Nilden TW. Reprogramming of the non-coding transcriptome during brain development. Journal Biology. 2010;9:5

Valadkhan 2013; Roles of small nuclear RNAs in eukaryotic gene expression. Essays Biochem. 2013;54:79-90. doi: 10.1042/BSE0540079

Voss 1992; R. F. Voss. Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. Physical Review Letters. 1992;68(25):3805–3808

Voss 2006; Voss TC, John S, Hager GL: Single-cell analysis of glucocorticoid receptor action reveals that stochastic post-chromatin association mechanisms regulate ligand-specific transcription. Mol Endocrinol 2006, 20:2641–2655.

Wahid 2010; Wahid F, Shehzad A, Khan T, Kim YY. MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. Biocima et Biophysica Acta (BBA)-Molecular Research, 2010;1803(11):1231-1243

Wan 2011; Wan Y, Kertesz M, Spitale RC, Segal E, Chang, H Y. Understanding the transcriptome through RNA structure. Nature Rev. Genet. 2011;12:641–655

Wan 2014; Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z et.al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature. 2014;505(7485):706-9. doi: 10.1038/nature12946

Wapinski 2011; Wapinski O, Howard Y. Long noncoding RNAs and human disease. Trends in Cell Biology June. Cell Press. 2011;21(6):354-361

Warris 2014; Warris et. al. Fast selection of miRNA candidates based on large-scale pre-computed MFE sets of randomized sequences. BMC Research notes 2014.7:34.

Washeitl 2005; Washeitl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences USA(PNAS) 2005;102(7):2454–2459.doi: 10.1073/pnas.0409169102

Washietl 2005 (Dissertation); Stefan Washietl. Prediction of structural non-coding RNAs by comparative sequence analysis. Dissertation. Stefan Washeitl 2005. Available at: https://www.tbi.univie.ac.at/papers/Abstracts/wash_diss.pdf

Washietl 2012; Washietl S, Will S, Hendrix DA, Goff LA, Rinn JL, Berger B and Kellis M. Computational analysis of noncoding RNAs. Advanced Review. WIREs RNA 2012. doi: 10.1002/wrna.1134

Watson 2007; Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R. Molecular Biology of the Gene. 5th Edition 2007. Pearson Education Inc. USA

Weinhold 2006; Weinhold B. Epigenetics: The Science of Change. Environmental Health Perspectives. 2006;114(3): A160–A167

Wheeler 2013; Wheeler DA, Wang L. From human genome to cancer genome: the first decade. Genome Research. 2013;23:1054–62.

Wilusz 2008; Wilusz JE, Freier SM, Spector DL. 3’ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell 2008;135:919-932.

Wolfinger 2004; Wolfinger MT, Svrcek-SeilerWA, Flamm C, Hofacker IL, Stadler PF. Efficient computation of RNA folding dynamics. Journal of Physics A: Mathematical and General 2004, 37:4731–4741. [http://stacks.iop.org/0305-4470/37/4731].

Wolfsheimer 2010; Wolfsheimer S and Hartmann AK. Minimum-Free-Energy Distribution of RNA Secondary Structures: Entropic and Thermodynamic Properties of Rare Events. Physical Review E 2010. 82(2 Pt 1):021902.

Wu 2002; Wu Peng, Nakano S, Sugimoto N. Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. European Journal of Biochemistry. 2002;269;2821-2830.

Wuchty 1999; Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 1999, 49:145–165.

Wutz 2013; Wutz A: Epigenetic regulation of stem cells: the role of chromatin in cell differentiation. Adv Exp Med Biol 2013, 786:307–328.

Xia 1998; Xia T, SantaLucia J Jr, Burkard ME et.al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 1998, 37:14719–14735.

Xu 2014; Xu TP, Huang MD, Xia R, Liu XX, Sun M, Yin L et.al. Decreased expression of the long non-coding RNA FENDRR is associated with poor prognosis in gastric cancer and FENDRR regulates gastric cancer cell metastasis by affecting fibronectin1 expression. Journal of Hematol Oncol. 2014 Aug 29;7:63. doi: 10.1186/s13045-014-0063-7.

Yan 2015(1); Yan B, Yao J, Liu J-Y, Li X-M, Wang X-Q, Li Y-J et.al.. lncRNA-MIAT regulates microvascular dysfunction by functioning as a competing endogenous RNA. Circ Res. 2015;116:1143–56.

Yan 2015(2); Yan X, Hu Z, Feng Y, Hu X, Yuan J, Zhao SD et.al.. Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. Cancer Cell 2015;28, 529–540. http://dx.doi.org/10.1016/j.ccell.2015.09.006

Yang 2015; Yang M, Zhai X, Xia B, Wang Y, Lou G. Long noncoding RNA CCHE1 promotes cervical cancer cell proliferation via upregulating PCNA. Tumor Biology. 2015;36(10):7615-7622

Yang 2015; Yang M-H, Hu Z-Y, Xu C, Xie L-Y, Wang X-Y et.al.. MALAT1 promotes colorectal cancer cell proliferation/migration/invasion via PRKA kinase anchor protein 9. Biochim Biophys Acta. 2015;1852:166–74.

Yoon 2004; Yoon BJ, Vaidyanathan PP. RNA secondary structure prediction using context-sensitive hidden Markov models. IEEE Explore 2005. DOI: 10.1109/BIOCAS.2004.1454177

Yoon 2007; Yoon BJ, Vaidyanathan PP. Computational Identification and Analysis of Noncoding RNAs. IEEE Signal Processing Magazine; DOI: 10.1109/MSP.2007.273058

Zhao 2014; Zhao Y, Luo H, Chen X, Xiao Y, Chen R. Computational Methods to Predict Long Noncoding RNA Functions Based on Co-expression Network. Methods in molecular Biology. 2014;1182:209-218

Zhou 2016; Zhou Q, Chen J, Feng J, Wang J. Long noncoding RNA PVT1 modulates thyroid cancer cell proliferation by recruiting EZH2 and regulating thyroid stimulating hormone receptor (TSHR). Tumour Biol. 2016;37(3):3105-3113

Zuker 1981; Zuker M, Stiegler P. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9:133–148.

Zuker 1989; Zuker M. On finding all suboptimal foldings of an RNA molecule. Science. 1989;244:48–52.

Zuker 1999; Zuker M, Mathews DH, Turner DH. Algorithms and thermodynamics for RNA secondary structure prediction : A practical guide. NA Biochemistry and Biotechnology. 1999;70:11-43. DOI : 10.1007/978-94-011-4485-8_2

Zuker 2003; Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research. 2003;31(13):3406-3451

APPENDIX –I

Appendix I - Results for 902 snRNA sequences from Chapter 6

Table 6.6. Sample computation of MFE using the model developed for 902 snRNA sequences taken from the Rfam database

Rfam snRNA family RF00004 (208)							
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE _M**	**MFE_C**	**RD1**	**% RD1**
1	AALT01209640.1/567-377	191	37.91472	-63.1	-62.2351007	0.013706803	1.370680322
2	AAFR03033875.1/20528-20718	191	37.67534	-65.8	-61.8541756	0.059966936	5.996693597
3	AAIY01044029.1/787-597	191	37.80852	-63.7	-62.0660984	0.025649946	2.564994628
4	AAZO01007389.1/15370-15178	193	38.60989	-70.3	-63.7405252	0.093306896	9.330689585
5	AAYZ01695118.1/310-500	191	38.25785	-63.4	-62.7811182	0.009761543	0.976154279
6	AAHX01044404.1/26102-26292	191	37.78192	-65.2	-62.0237736	0.048715129	4.871512853
7	AACN010750078.1/657-848	192	38.30984	-66.9	-63.0634513	0.057347515	5.734751475
8	ABAV01019481.1/5988-6180	193	38.57082	-64	-63.6783495	0.005025789	0.502578852
9	AAZX01018356.1/721-913	193	45.57082	-79.3	-74.8174495	0.056526488	5.652648758
10	AY765362.1/650-458	193	38.83056	-70	-64.0916724	0.08440468	8.440468023
11	BX927129.10/97355-97165	191	43.04706	-72.6	-70.4021902	0.030272862	3.02728624
12	AAFC03011281.1/26348-26158	191	38.44134	-66.2	-63.0731089	0.047234004	4.723400425
13	AAVX01416582.1/429-619	191	43.34971	-73.8	-70.8837882	0.039515065	3.951506519
14	X00093.1/360-550	191	38.07347	-67.4	-62.4877202	0.072882489	7.288248891

15	CAAE01009132.1/1078-888	191	38.09987	-66.6	-62.5297212	0.061115298	6.111529778
16	AF095839.1/1586-1389	198	45.21437	-76.8	-75.2482205	0.020205462	2.020546161
17	AANH01015084.1/457-647	191	37.80852	-62.2	-62.0660984	0.002152758	0.215275849
18	AC004138.3/33098-33293	196	45.56928	-75.3	-75.4138032	-0.00151133	-0.151133045
19	AAJJ01003841.1/8097-7907	191	45.28412	-72.7	-73.9620168	-0.017359241	-1.735924115
20	AACY020405974.1/944-1135	192	38.36229	-68.7	-63.1469043	0.080831087	8.083108661
21	AM465080.2/15550-15355	196	38.76424	-75.1	-64.5849426	0.14001408	14.00140801
22	M72891.1/1-196	196	37.93868	-77.6	-63.2712246	0.184649167	18.46491672
23	BAAB01070452.1/1509-1701	193	38.89523	-67.3	-64.1945718	0.046143064	4.61430635
24	X04243.1/69-264	196	38.79017	-72	-64.62619	0.102414027	10.24140275
25	AAPY01817437.1/27757-27567	191	38.89629	-60.6	-63.7970597	-0.05275676	-5.275676018
26	AANG01476605.1/2311-2501	191	34.3366	-58.7	-56.5412282	0.036776351	3.677635089
27	AANN01265286.1/964-773	192	38.89575	-70.5	-63.9958113	0.092257995	9.225799526
28	AASG02001826.1/33924-33729	196	38.85489	-72.9	-64.7291882	0.112082466	11.20824658
29	AAEU02000279.1/6988-6794	195	39.26713	-68.5	-65.1855783	0.048385719	4.838571889
30	AC189506.1/7126-6931	196	38.00486	-72.3	-63.3765305	0.123422814	12.34228144
31	X69327.1/1-196	196	41.72469	-68.4	-69.2959049	-0.013098025	-1.309802536
32	AAAA02007579.1/14294-14490	197	47.12697	-82.1	-78.0921424	0.04881678	4.881677997
33	AC149482.1/90998-90803	196	43.91951	-73.9	-72.7885152	0.01504039	1.504039013
34	AAPU01010615.1/193731-193537	195	38.92002	-67.2	-64.6332296	0.038195989	3.819598881

35	AADA01287270.1/15999-15809	191	38.06027	-68.6	-62.4667089	0.089406576	8.940657632
36	AAPT01020503.1/50330-50135	196	38.45183	-68.9	-64.0878045	0.069843186	6.984318555
37	AAAB01008933.1/737524-737331	194	43.46592	-71.7	-71.6675173	0.000453036	0.045303644
38	ABDC01189504.1/6312-6122	191	35.1921	-58.4	-57.9025958	0.008517195	0.851719509
39	AAWU01010867.1/21259-21065	195	38.19004	-67.1	-63.4716154	0.054074287	5.407428665
40	AANI01016115.1/56636-56831	196	38.66039	-66.2	-64.4196763	0.026893108	2.689310782
41	AC157776.1/115175-114979	197	39.6482	-72	-66.1911772	0.080678094	8.067809416
42	AAPP01015704.1/576899-577092	194	38.15105	-65.6	-63.2099653	0.036433455	3.643345526
43	AC151964.12/72254-72449	196	38.64739	-63.5	-64.3989867	-0.014157272	-1.415727158
44	BAAE01249332.1/245-435	191	38.25785	-68.5	-62.7811182	0.083487326	8.348732573
45	AAGE02006086.1/44202-44008	195	38.37384	-69.5	-63.7640906	0.082531071	8.253107113
46	AP009284.1/2157-1962	196	38.95823	-74.6	-64.8936295	0.130112206	13.01122056
47	AAQB01006449.1/663346-663151	196	35.00979	-60.5	-58.6104867	0.031231625	3.123162455
48	AB202073.1/636-444	193	34.79171	-58.4	-57.6646504	0.012591603	1.2591603
49	AACT01003467.1/10053-10247	195	39.12608	-64.5	-64.9611388	-0.007149439	-0.714943903
50	AF106845.1/1870-2065	196	39.61067	-67.1	-65.9318582	0.017408969	1.740896909
51	ABBA01028418.1/3128-3319	192	27.99366	-44.8	-46.6473138	-0.041234682	-4.123468222
52	X15930.1/1-195	195	38.69988	-65.9	-64.2829129	0.024538499	2.453849949
53	AASR01035668.1/1184-988	197	39.06078	-67.7	-65.2564266	0.036094141	3.609414146
54	D25323.1/7242-7047	196	39.96434	-75.7	-66.4946591	0.121602917	12.16029175

55	AADE01000447.1/58660-58464	197	39.31726	-69	-65.6645559	0.048339769	4.833976904
56	X55772.1/223-413	191	39.03817	-59.4	-64.0228446	-0.077825667	-7.782566696
57	AP007151.1/1974779-1974972	194	38.15105	-66.7	-63.2099653	0.052324358	5.23243578
58	X54113.1/230-419	190	38.97428	-69.6	-63.721576	0.084460115	8.446011495
59	AATM01000136.1/58626-58436	191	39.1282	-71.1	-64.1660997	0.097523211	9.75232115
60	AAHF01000004.1/976953-976761	193	37.8739	-64.4	-62.5693369	0.028426446	2.842644613
61	CAAA01181619.1/3682-3492	191	31.07347	-52.7	-51.3486202	0.02564288	2.564287956
62	X05084.1/1-193	193	38.54475	-65.8	-63.6368641	0.032874406	3.287440635
63	ABAR01000024.1/421358-421550	193	38.09883	-60.9	-62.9272679	-0.033288471	-3.328847103
64	AAQA01000616.1/21523-21715	193	38.53171	-65.4	-63.6161108	0.027276593	2.727659326
65	AY661656.1/2159-2358	200	39.07215	-61.7	-65.8733176	-0.067638859	-6.763885889
66	ABAS01000032.1/96356-96164	193	37.94019	-63	-62.674831	0.005161413	0.516141308
67	AAJN01000116.1/21716-21908	193	36.20422	-59.4	-59.9123768	-0.008625873	-0.862587258
68	ABDB01000030.1/249236-249428	193	37.86063	-66.2	-62.5482159	0.055162902	5.516290209
69	CAAL01000198.1/59590-59777	188	38.04866	-66.1	-61.8494328	0.064305101	6.430510131
70	AACD01000084.1/492354-492546	193	38.20422	-65.5	-63.0949768	0.036717911	3.671791098
71	AACM02000140.1/27654-27846	193	38.30932	-61	-63.2622244	-0.037085646	-3.708564592
72	AAKD03000004.1/610702-610510	193	37.75428	-60.6	-62.3789807	-0.029356117	-2.935611669
73	AANU01167734.1/2229-2419	191	37.11079	-57.4	-60.9557957	-0.06194766	-6.194765967

74	AAEE01000010.1/89978-89784	195	38.42619	-62.8	-63.8473979	-0.016678311	-1.667831129
75	AARE01006627.1/693-886	194	37.54017	-54.9	-62.2378701	-0.133658836	-13.36588358
76	AACW02000228.1/581455-581260	196	38.15004	-59.7	-63.6075605	-0.065453275	-6.545327485
77	AM270020.1/5920-6112	193	37.83407	-59	-62.5059517	-0.05942291	-5.942290979
78	AAIM02000113.1/450420-450612	193	38.38796	-57.2	-63.3873596	-0.108170622	-10.8170622
79	ABBB01000093.1/1472-1279	194	37.36572	-63.4	-61.9602673	0.022708718	2.270871835
80	AAFU01000671.1/40760-40955	196	34.07092	-61	-57.1164534	0.063664699	6.366469887
81	CR382129.1/446178-446370	193	38.59687	-62.8	-63.719807	-0.014646608	-1.464660806
82	AAPN01113121.1/7070-7251	182	37.21032	-62.7	-59.3177784	0.053942929	5.394292865
83	AAQQ01759780.1/669-855	187	33.08578	-56.4	-53.7523966	0.046943323	4.694332313
84	AAQX01002532.1/7801-7990	190	43.32402	-71.3	-70.6433058	0.009210298	0.921029797
85	X63786.1/549-739	191	33.36281	-54.9	-54.991641	-0.001669236	-0.166923584
86	AAIW01000278.1/13786-13979	194	30.54017	-52.8	-51.0987701	0.032220264	3.222026354
87	AAIW01000278.1/13786-13979	194	30.67381	-50.2	-51.311436	-0.02214016	-2.214015962
88	AASM01001106.1/7439-7242	198	30.9377	-59.5	-52.5297611	0.117146873	11.71468731
89	AAWC01001022.1/39371-39181	191	32.44786	-58	-53.5356772	0.076971083	7.697108258
90	AC187487.2/140157-140345	189	26.64721	-45.9	-43.9059013	0.043444415	4.344441541
91	CAAI01005114.1/671-870	200	33.16126	-56.1	-56.4673056	-0.006547337	-0.654733652
92	AAGK01000002.1/903654-903460	195	32.99742	-56.8	-55.2085986	0.028017631	2.801763071
93	AAXI01000029.1/108904-109093	190	31.82234	-53.4	-52.3406833	0.019837392	1.983739181

94	DQ114948.1/136-329	194	38.08512	-60.3	-63.1050581	-0.046518377	-4.651837662
95	AAKM01000005.1/1407395-1407197	199	38.56779	-67.9	-64.8711309	0.044607793	4.460779279
96	AAXT01000001.1/1056302-1056495	194	38.83004	-72.6	-64.2904429	0.114456709	11.44567087
97	AAJI01001427.1/51501-51693	193	34.03281	-59.5	-56.4570138	0.051142625	5.114262522
98	AABS01001062.1/345-536	192	34.84403	-62.7	-57.5483044	0.082164205	8.216420466
99	X71483.1/1-192	192	38.79224	-70.8	-63.8310877	0.098430965	9.843096528
100	AAXJ01017415.1/259-455	197	39.33004	-67.6	-65.6848925	0.028329992	2.832999203
101	AF325695.1/199-9	191	38.88336	-68.9	-63.7764929	0.074361496	7.436149625
102	AAFT01000058.1/3256-3060	197	30.05722	-50.1	-50.9290552	-0.016548007	-1.654800723
103	AP004918.1/55019-55207	189	38.54688	-65.1	-62.8418449	0.034687483	3.468748283
104	AAID01003241.1/2743-2932	190	37.82234	-54.5	-61.8884833	-0.135568501	-13.55685005
105	AATT01000021.1/233305-233109	197	39.6482	-67.4	-66.1911772	0.017935056	1.793505608
106	CP000498.1/1665432-1665627	196	34.28155	-54.8	-57.4516275	-0.048387362	-4.838736228
107	AAFM01000022.1/42807-42996	190	37.70254	-58.8	-61.697851	-0.049283181	-4.928318091
108	AAGT01000476.1/108951-108760	192	38.15208	-61.3	-62.8124037	-0.024672165	-2.467216518
109	DQ235686.1/7795-7608	188	38.48216	-62.5	-62.5392688	-0.000628301	-0.062830076
110	AATU01001299.1/7335-7151	185	45.5882	-77	-73.2482972	0.048723413	4.872341336
111	AAGI01000215.1/38061-38253	193	28.13839	-48.5	-47.0772131	0.029335813	2.93358134
112	AAGD02001363.1/26266-26450	185	37.74521	-62.5	-60.7677571	0.027715887	2.771588674

113	AADS01000047.1/234093-233905	189	38.63805	-62.7	-62.9869295	-0.004576229	-0.457622885
114	AANW02001910.1/4438-4250	189	37.90252	-69	-61.8164809	0.104108972	10.41089718
115	DQ158857.1/121770-121585	186	30.37979	-46.5	-49.2467641	-0.059070195	-5.907019535
116	CP000599.1/149439-149632	194	33.75377	-52.8	-56.2125742	-0.064632088	-6.463208756
117	AAWT01070971.1/3028-3217	190	33.32738	-56.3	-54.7356606	0.027785781	2.778578056
118	AACP01000091.1/4899-4707	193	38.44029	-65.5	-63.4706408	0.030982583	3.098258292
119	AAFP01000557.1/11449-11264	186	38.48327	-65.6	-62.1418295	0.052716014	5.271601435
120	AAFI02000140.1/17830-17618	213	33.58744	-59.3	-59.7402941	-0.007424858	-0.742485839
121	AAFB02000004.1/138313-138494	182	32.89825	-55.2	-52.4559811	0.049710487	4.971048684
122	AAPO01000010.1/72363-72562	200	32.37137	-56.3	-55.2103564	0.019354238	1.935423834
123	AAFX01115267.1/519-717	199	40.82124	-79.1	-68.4570412	0.13455068	13.45506801
124	AANV02000585.1/5693-5873	181	36.34969	-61	-57.7486625	0.053300615	5.330061468
125	DQ012953.1/31-235	205	33.0138	-57.4	-57.2306606	0.002950165	0.295016463
126	AAFO01000053.1/189105-188894	212	30.30711	-57	-54.3207119	0.047005055	4.700505522
127	AABY01000227.1/3654-3483	172	30.21662	-46.5	-46.192712	0.006608345	0.660834469
128	Z36100.1/1808-1619	190	30.44159	-47.2	-50.1435008	-0.062362305	-6.236230456
129	AC167922.2/17996-18182	187	37.58396	-57.1	-60.9103495	-0.066731165	-6.673116457
130	AAZN01000309.1/86876-87072	197	27.76374	-46.1	-47.2794397	-0.025584376	-2.558437556
131	AADM01000307.1/26094-25895	200	33.58616	-60.1	-57.1434524	0.049193803	4.919380307
132	AL590446.1/168546-168725	180	38.60414	-71.4	-61.1365619	0.143745632	14.37456318

133	AF053589.1/90-279	190	47.07732	-89	-76.6159413	0.139146727	13.91467267
134	Z50072.1/229-412	184	24.41713	-38.5	-39.3591854	-0.022316503	-2.231650304
135	AAHC01001365.1/13705-13888	184	37.18217	-60.9	-59.6721849	0.020161168	2.016116781
136	AF287991.1/4898-5088	191	37.90147	-68.6	-62.2140013	0.09309036	9.309035966
137	M33777.1/191-381	191	37.90147	-68.6	-62.2140013	0.09309036	9.309035966
138	S64581.1/735-926	192	38.90867	-67.3	-64.016371	0.048790922	4.879092155
139	AAKO01002676.1/16743-16938	196	38.94533	-71	-64.8730982	0.086294392	8.629439199
140	AAPQ01007349.1/370124-370319	196	38.94533	-71	-64.8730982	0.086294392	8.629439199
141	AAIZ01004041.1/16829-17024	196	38.91951	-68	-64.8320152	0.046588012	4.658801222
142	AAYL01000061.1/287788-287596	193	38.99846	-68.3	-64.358856	0.057703425	5.770342535
143	AL683874.1/16263-16071	193	37.8739	-64.4	-62.5693369	0.028426446	2.842644613
144	AAIH02000488.1/9382-9189	194	38.15105	-66.7	-63.2099653	0.052324358	5.23243578
145	AAKE03000002.1/1730972-1731164	193	37.83407	-61.7	-62.5059517	-0.013062426	-1.306242589
146	AAEL01000160.1/10170-10364	195	32.62188	-58	-54.610995	0.058431121	5.843112135
147	AATX01000107.1/79359-79166	194	37.16341	-60.2	-61.6383371	-0.023892642	-2.389264247
148	AANS01001054.1/6664-6467	198	32.9377	-59.5	-55.7123611	0.063657797	6.365779747
149	AABL01000318.1/11806-12005	200	31.16126	-55.1	-53.2847056	0.032945452	3.294545229
150	CAAJ01003844.1/3754-3953	200	31.16126	-55.1	-53.2847056	0.032945452	3.294545229
151	EF140768.1/2-192	191	30.51491	-54.8	-50.4597752	0.079201183	7.920118256
152	AB179181.1/1-163	163	28.95694	-44	-42.3917856	0.036550326	3.655032615

153	AC198944.2/184805-184614	192	25.45438	-43.8	-42.6065471	0.027247782	2.724778217
154	AACQ01000018.1/49571-49783	213	30.41868	-54	-54.6978478	-0.012923108	-1.292310767
155	AE017345.1/926592-926777	186	38.48327	-65.6	-62.1418295	0.052716014	5.271601435
156	AAEY01000026.1/44265-44450	186	38.48327	-65.6	-62.1418295	0.052716014	5.271601435
157	AACO02000044.1/38107-37922	186	38.48327	-65.6	-62.1418295	0.052716014	5.271601435
158	AACI02000565.1/65-236	172	28.85777	-45.7	-44.0303739	0.036534487	3.653448746
159	AACF01000175.1/12372-12544	173	27.45797	-41.7	-42.0024633	-0.007253316	-0.725331638
160	AAFW02000011.1/661534-661345	190	28.44159	-47.2	-46.9609008	0.005065662	0.506566154
161	AAEG01000106.1/129944-130133	190	29.44159	-47.2	-48.5522008	-0.028648322	-2.864832151
162	AY007788.1/537-683	147	32.60725	-50	-45.00692	0.099861601	9.986160091
163	M58665.1/571-739	169	32.81627	-52.5	-49.7307296	0.052748008	5.27480078
164	U23406.1/206-352	147	33.06718	-45.6	-45.7388026	-0.003043918	-0.304391763
165	EF052257.1/89-253	165	21.11825	-29.5	-30.3172688	-0.027704026	-2.770402631
166	AACA01000784.1/509-337	173	26.51465	-41.1	-40.50137	0.014565207	1.456520665
167	AABZ01000169.1/11839-11668	172	22.64068	-35.1	-34.1371117	0.027432714	2.743271437
168	X56454.1/125-277	153	32.60627	-48.4	-46.2029621	0.045393345	4.53933454
169	AC008368.21/83891-84039	149	34.15891	-50	-47.8752723	0.042494554	4.249455424
170	M58666.1/571-718	148	37.248	-57.2	-52.5913346	0.080571073	8.057107269
171	AAHK01000589.1/17217-17069	149	33.54928	-44	-46.9051661	-0.066026502	-6.602650171
172	AF326335.1/1-142	142	32.20722	-45.9	-43.3723478	0.055068675	5.50686751

173	CAAC02000548.1/434719-434909	191	38.88336	-67.5	-63.7764929	0.055163068	5.516306803
174	AACE03000009.1/583940-584128	189	32.8327	-58.8	-53.7488783	0.085903431	8.590343099
175	AABX02000002.1/241678-241484	195	38.72584	-67.9	-64.32423	0.052662298	5.266229791
176	AASC02023314.1/7214-7405	192	38.64945	-70	-63.6038722	0.091373254	9.137325367
177	AAFD02000010.1/1244775-1244583	193	32.04199	-50.2	-53.2890262	-0.061534387	-6.153438734
178	AAZY02000001.1/986053-985858	196	37.64613	-61	-62.8056796	-0.029601305	-2.960130482
179	ABFM01000169.1/24531-24724	194	37.36572	-63.4	-61.9602673	0.022708718	2.270871835
180	AAQM02000124.1/160019-159827	193	39.12713	-65.7	-64.5636018	0.017296777	1.729677689
181	CR382136.2/900319-900516	198	28.60736	-49.4	-48.8214852	0.011710825	1.171082549
182	AAGV020390824.1/478-668	191	37.94123	-69.3	-62.2772774	0.101337989	10.13379888
183	X58842.1/1-191	191	38.74091	-65.3	-63.5498057	0.026802363	2.680236278
184	AAKN02019678.1/9356-9546	191	38.00741	-66.9	-62.3825903	0.067524808	6.752480808
185	AAWR02015112.1/62289-62483	195	37.33831	-61.1	-62.1162546	-0.016632645	-1.663264549
186	ABDF02000003.1/1932068-1932260	193	38.80467	-60.7	-64.0504646	-0.05519711	-5.51971103
187	M12856.1/361-551	191	39.2052	-72.9	-64.2886279	0.118125818	11.81258177
188	AACS02000012.1/1565410-1565598	189	38.58598	-62.7	-62.904066	-0.003254641	-0.325464106
189	ABEG02004067.1/53561-53371	191	38.23157	-66.4	-62.7392908	0.055131162	5.513116227
190	AAIL02000026.1/644385-644193	193	38.53171	-65.2	-63.6161108	0.024292779	2.429277913

191	AAQY02000293.1/11238-11050	189	39.07787	-66.6	-63.6868118	0.043741565	4.374156529
192	AADG06003467.1/1334-1527	194	38.51815	-66.8	-63.7941277	0.044998088	4.499808779
193	AAGJ04020931.1/13670-13861	192	38.58437	-65.1	-63.5003145	0.024572741	2.457274122
194	AAGU03035529.1/35295-35485	191	38.29724	-69.3	-62.8438054	0.093162981	9.31629812
195	AFQF01002518.1/111488-111680	193	38.37486	-59.3	-63.3665215	-0.068575405	-6.857540515
196	AAGW02065159.1/39337-39527	191	37.95447	-68.1	-62.2983547	0.08519303	8.519302999
197	AAWZ02022241.1/1219-1409	191	38.41518	-67.9	-63.0314813	0.071701306	7.170130564
198	CAAB02025078.1/1453-1643	191	37.91472	-71.3	-62.2351007	0.127137437	12.71374374
199	AAQR03042593.1/1155-1347	193	38.76579	-63.6	-63.9886013	-0.006110083	-0.611008331
200	AACU03000093.1/631552-631747	196	38.4649	-65	-64.1085992	0.013713858	1.371385821
201	AE014186.2/1461495-1461298	198	30.9377	-56.5	-52.5297611	0.070269716	7.026971592
202	FR799006.1/251703-251558	146	38.62345	-58.5	-54.3809	0.070411966	7.041196554
203	AAFN02000024.1/475809-475596	214	30.48033	-57.6	-54.9955523	0.045216106	4.521610633
204	K00034.1/420-610	191	37.71535	-62.5	-61.917831	0.009314704	0.931470354
205	ABDG02000029.1/618164-617972	193	38.41414	-64	-63.4290144	0.00892165	0.892165027
206	AP004871.3/124344-124540	197	51.9339	-89.1	-85.7414113	0.037694598	3.7694598
II. Rfam snRNA family RF00007 (62)							
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_M**	**MFE_C**	**RD1**	**% RD1**

207	AANN01056468.1/521-372	150	39.56829	-58.4	-56.6828146	0.02940386	2.94038601
208	AADA01322814.1/1127-978	150	39.56829	-58.4	-56.6828146	0.02940386	2.94038601
209	AANU01293824.1/906-757	150	39.56829	-58.4	-56.6828146	0.02940386	2.94038601
210	ABBA01017933.1/18965-18816	150	40.56829	-61	-58.2741146	0.044686646	4.468664639
211	ABDC01356688.1/591-740	150	39.56829	-58.4	-56.6828146	0.02940386	2.94038601
212	AAFC03093377.1/31487-31636	150	34.64102	-56.3	-48.842049	0.132468046	13.24680462
213	AAQQ01306058.1/1332-1183	150	34.4516	-56.3	-48.5406313	0.137821824	13.78218243
214	AANG01542153.1/730-879	150	34.56829	-56.3	-48.7263146	0.13452372	13.45237199
215	AAIY01326656.1/671-820	150	34.94483	-57.2	-49.3255033	0.137666026	13.76660257
216	AAPN01231707.1/282-431	150	34.46621	-56.2	-48.5638761	0.135874091	13.58740911
217	AAHX01055169.1/34088-34238	151	34.55295	-57.4	-48.9015165	0.148057203	14.80572031
218	AALT01414211.1/1221-1073	149	39.32063	-56.8	-56.089113	0.012515616	1.251561625
219	AAHY01168842.1/1247-1097	151	37.15945	-59	-53.049227	0.10086056	10.08605601
220	AAPY01023785.1/1285-1135	151	35.15945	-54.7	-49.866627	0.088361482	8.836148163
221	AAVX01293999.1/160-11	150	35.14591	-51.9	-49.6454845	0.043439604	4.343960431
222	CAAE01014653.1/353935-353782	154	39.76852	-60.7	-57.7998388	0.047778603	4.777860318
223	BAAE01110703.1/791-944	154	35.64054	-59	-51.2309974	0.13167801	13.16780097
224	AANH01004214.1/17675-17828	154	34.55071	-55.5	-49.4967465	0.10816673	10.81667299
225	ABAV01000136.1/26200-26351	152	33.96453	-55.7	-48.1647591	0.135282602	13.5282602
226	AAZO01006159.1/30950-30796	155	34.41864	-49.2	-49.4861836	-0.00581674	-0.581674041

227	AAAB01008960.1/5726267-5726086	182	47.24875	-77.8	-75.29193	0.032237403	3.223740317
228	AAJJ01000520.1/29099-28948	152	34.08288	-50.1	-48.3530794	0.034868676	3.48686755
229	AAZX01007808.1/26508-26658	151	34.45083	-45.4	-48.739013	-0.073546542	-7.354654167
230	AABS01000019.1/293615-293466	150	34.42237	-47	-48.4941121	-0.03178962	-3.17896196
231	AAYZ01032557.1/2146-2294	149	26.97419	-36.9	-36.4422307	0.012405671	1.240567083
232	AACT01038531.1/95659-95808	150	35.11725	-50.8	-49.5998851	0.023624309	2.362430881
233	AASG02002046.1/26669-26822	154	36.05682	-57.3	-51.8934123	0.094355806	9.435580625
234	AADK01040274.1/1841-1691	151	34.94405	-55.5	-49.5238676	0.107678061	10.76780614
235	AC198009.4/84558-84712	155	34.95544	-47.8	-50.3403992	-0.053146427	-5.314642705
236	AAGE02014219.1/46421-46245	177	37.58967	-68.9	-58.9234348	0.144797754	14.47977539
237	AARH01003540.1/648176-648022	155	35.68213	-53.7	-51.4967709	0.041028475	4.102847503
238	AC004255.1/89334-89170	165	35.42038	-58.4	-53.0762583	0.09115996	9.115996039
239	AP005874.3/27602-27759	158	46.08659	-70.1	-68.652192	0.020653466	2.06534665
240	AAAA02006813.1/31506-31663	158	46.08659	-70.1	-68.652192	0.020653466	2.06534665
241	AAWT01090545.1/3123-3274	152	33.69674	-48.5	-47.7386203	0.015698551	1.569855082
242	AAWU01013761.1/27501-27308	194	38.72635	-71.7	-64.1254487	0.105642278	10.56422779
243	AATU01009112.1/74318-74156	163	35.74688	-62.1	-53.1966118	0.14337179	14.33717899
244	AAQX01001042.1/31224-31386	163	41.92937	-65.8	-63.0348092	0.042024176	4.202417572
245	DQ888370.1/1-162	162	45.29229	-68.9	-68.1866237	0.010353792	1.035379191
246	AAEU02001091.1/68476-68268	209	49.98827	-87.9	-85.0405376	0.032530858	3.253085755

247	AAPQ01006579.1/170624-170417	208	51.82909	-90	-87.770229	0.024775233	2.477523314
248	AF459090.1/1-212	212	41.4379	-82.4	-72.0331297	0.125811533	12.58115329
249	AAKO01000167.1/274946-275154	209	53.03727	-88.5	-89.8924109	-0.015733457	-1.573345668
250	AASV01051056.1/691-484	208	49.95196	-83.9	-84.7831496	-0.010526216	-1.052621643
251	AAQB01007708.1/1369-1503	135	44.04067	-58.2	-60.8057168	-0.044771766	-4.47717658
252	AAIZ01008995.1/21704-21918	215	41.65417	-87	-72.9760848	0.161194427	16.11944273
253	AAFS01000475.1/55736-55522	215	43.83468	-86.6	-76.4459213	0.117252641	11.72526413
254	AAPU01011411.1/133226-133037	190	39.33374	-69.7	-64.2935755	0.077567066	7.756706598
255	AM487500.2/4877-4718	160	35.38112	-60.9	-52.0157818	0.145882072	14.58820721
256	AASC02028416.1/49450-49599	150	40.877	-58.6	-58.7653672	-0.002821966	-0.282196635
257	AAGV020551495.1/1052-903	150	34.71359	-56.7	-48.957541	0.136551306	13.65513059
258	AAWR02025158.1/532-681	150	34.93042	-56.5	-49.302577	0.127388017	12.73880169
259	AAWZ02031009.1/19848-19997	150	40.74258	-60.5	-58.5514702	0.032207104	3.220710441
260	EU240273.1/1-149	149	41.74336	-61.5	-59.9444153	0.025294059	2.529405941
261	AAEX03007273.1/32318-32169	150	40.64102	-59.4	-58.389849	0.017005909	1.700590906
262	AAGJ04047220.1/14069-13916	154	34.92738	-57.1	-50.0961328	0.12265967	12.265967
263	AADG06004603.1/1703-1555	149	33.29323	-46.2	-46.4977175	-0.006444102	-0.644410184
264	AAGW02065961.1/10800-10651	150	39.53915	-57.3	-56.6364526	0.011580235	1.158023455
265	AAQY02000248.1/644530-644694	165	40.19308	-68.5	-60.6710479	0.114291272	11.42912721
266	AAQR03135034.1/15819-15671	149	29.26298	-40.8	-40.0843762	0.0175398	1.753979952

267	AAGU03077213.1/2720-2571	150	39.78602	-58.7	-57.0292919	0.028461807	2.846180667
268	CAAB02002948.1/53519-53671	153	40.35776	-60.5	-58.5379003	0.032431399	3.243139922

III. Rfam snRNA family RF00015 (170)

Sl. No.	Sequence ID	NTL	SD_DFT	MFE_M	MFE_C	RD1	% RD1
269	AAIY01144063.1/2359-2499	141	28.71992	-37.9	-37.6234012	0.007298121	0.729812099
270	AC193264.3/174224-174084	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
271	AADD01128634.1/1009-869	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
272	AAFR03070450.1/4019-4159	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
273	AAPN01043183.1/815-675	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
274	AAHY01048392.1/24763-24623	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
275	ABDC01319198.1/612-472	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
276	AANG01100342.1/1096-956	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
277	AALT01138445.1/833-693	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
278	AANU01246434.1/3940-4080	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
279	AANN01193588.1/6106-6246	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
280	AACN010181221.1/232-92	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
281	AAFC03029538.1/15458-15318	141	30.41065	-35.4	-40.3138616	-0.13880965	-13.88096501
282	AAPY01772888.1/1357-1497	141	28.71992	-35.9	-37.6234012	-0.048005605	-4.800560486
283	AAYZ01170597.1/2807-2947	141	27.84281	-36.4	-36.2276595	0.00473463	0.473463005

284	AB168678.1/1-141	141	26.76605	-35.9	-34.5142215	0.038601072	3.860107211
285	K00476.1/2-142	141	30.364	-37.1	-40.2396351	-0.084626282	-8.462628179
286	BAAE01269333.1/2018-1878	141	32.56565	-35.1	-43.7431156	-0.246242611	-24.62426107
287	AAZO01005801.1/5213-5353	141	27.03558	-35.9	-34.9431231	0.026653954	2.665395373
288	BC124577.1/1-141	141	32.53471	-38.4	-43.6938788	-0.137861428	-13.78614281
289	AANH01001405.1/82396-82256	141	27.90408	-36.6	-36.3251647	0.007509161	0.750916122
290	ABAV01003454.1/17582-17444	139	25.98221	-34.3	-32.8676886	0.041758349	4.175834947
291	AAZX01000257.1/602-462	141	25.12976	-33.2	-31.9103859	0.038843799	3.884379902
292	AAGE02020535.1/44175-44035	141	32.41065	-47.1	-43.4964616	0.076508246	7.650824597
293	AAVX01087583.1/989-849	141	33.10244	-40.1	-44.5973102	-0.112152375	-11.2152375
294	K03095.1/1-139	139	28.80056	-36.6	-37.3525248	-0.020560786	-2.056078615
295	AACT01063233.1/31971-31831	141	28.91938	-38.8	-37.9408126	0.022144004	2.214400387
296	AAWU01039949.1/7621-7481	141	32.16109	-44.7	-43.0993426	0.03580889	3.580888986
297	AAAB01008933.1/1598961-1598821	141	32.28611	-47.8	-43.2982858	0.094178121	9.417812106
298	X15933.1/1-149	149	34.33529	-54.8	-48.1559475	0.121241834	12.12418339
299	AC084591.1/18183-18321	139	33.63239	-39.8	-45.0414228	-0.13169404	-13.16940396
300	CU302335.1/69797-69946	150	34.12866	-55.3	-48.0267289	0.13152389	13.15238897
301	AABS01000042.1/351938-351798	141	33.42037	-41.8	-45.1032419	-0.079024925	-7.902492505
302	AADK01043685.1/1410-1272	139	32.06865	-41.7	-42.5530486	-0.020456802	-2.04568021
303	AACG02000302.1/224-363	140	28.05761	-36.4	-36.369879	0.000827499	0.082749888

304	AACA01000783.1/223-362	140	28.05761	-36.4	-36.369879	0.000827499	0.082749888
305	AAPU01010717.1/67287-67148	140	30.00495	-35.9	-39.46867	-0.09940585	-9.940585043
306	AAPQ01007039.1/986580-986441	140	32.16191	-38.3	-42.9010482	-0.120131808	-12.01318077
307	AANI01016129.1/182015-181878	138	31.46705	-42.1	-41.3961153	0.016719351	1.671935109
308	AAJJ01000336.1/45867-45727	141	33.072	-43.7	-44.5488726	-0.019425003	-1.942500252
309	AAAA02007064.1/44912-44768	145	33.55241	-54	-46.1117423	0.146078847	14.60788471
310	AAKO01002834.1/25643-25504	140	32.25572	-42	-43.0503317	-0.025007898	-2.50078982
311	AAEU02000313.1/218060-217921	140	32.25572	-42	-43.0503317	-0.025007898	-2.50078982
312	AASS01015485.1/140-1	140	32.25572	-42	-43.0503317	-0.025007898	-2.50078982
313	X07113.1/1-150	150	34.50999	-53.6	-48.6335514	0.092657623	9.265762294
314	AAPP01015712.1/40466-40605	140	32.33369	-41.4	-43.1744038	-0.042859995	-4.285999452
315	AM479189.1/4511-4661	151	39.58208	-59.6	-56.904358	0.045228893	4.522889282
316	AASG02000802.1/49546-49696	151	42.20237	-62.9	-61.0740272	0.029029775	2.902977495
317	AAPT01020986.1/3963-4102	140	32.55101	-39.6	-43.5202245	-0.098995568	-9.899556802
318	AARH01003623.1/42069-42219	151	40.05909	-59.9	-57.6634346	0.03733832	3.733831977
319	X67145.1/194-344	151	34.53838	-54.9	-48.8783312	0.109684314	10.96843139
320	AP004858.3/49137-48993	145	33.62735	-51.4	-46.2310007	0.100564189	10.05641886
321	AAQA01000004.1/379392-379252	141	33.5407	-45.7	-45.2947163	0.008868353	0.886835308
322	AP006099.1/69439-69589	151	34.91523	-54.5	-49.4780066	0.092146668	9.214666828
323	AAWT01050999.1/11219-11362	144	26.65372	-35.7	-34.9342659	0.021449133	2.144913316

324	AATT01000006.1/24344-24478	135	31.75642	-39.6	-41.2577896	-0.041863374	-4.186337357
325	AAQB01006409.1/413450-413312	139	31.76887	-41.1	-42.075997	-0.023746886	-2.374688644
326	AP007155.1/2359336-2359191	146	30.70133	-43.4	-41.7744186	0.037455794	3.745579361
327	AAIH02000036.1/105327-105472	146	30.70133	-43.4	-41.7744186	0.037455794	3.745579361
328	AACW02000210.1/206231-206366	136	33.122	-46.6	-43.6304334	0.063724606	6.37246056
329	AC192395.1/10098-9958	141	23.53471	-30.4	-29.3721788	0.033809907	3.380990657
330	CAAJ01010632.1/7788-7919	132	25.15046	-32	-30.1469267	0.057908542	5.790854183
331	CAAI01006665.1/11117-11248	132	25.15046	-32	-30.1469267	0.057908542	5.790854183
332	AAKM01000017.1/282531-282664	134	26.13576	-35.8	-32.1140274	0.102960129	10.29601295
333	L22250.1/171-310	140	33.76602	-43.1	-45.4536624	-0.054609336	-5.460933645
334	AAFU01001086.1/4264-4401	138	33.19617	-40.9	-44.1476581	-0.079404844	-7.940484356
335	AAYL01000045.1/85592-85722	131	42.09625	-61.8	-56.9131572	0.079075126	7.907512619
336	AAXJ01014857.1/1040-1174	135	32.894	-46.2	-43.0680147	0.06779189	6.77918899
337	AANS01000355.1/46407-46542	136	26.81893	-36	-33.6003668	0.066656479	6.665647883
338	AATM01000105.1/197284-197424	141	28.76605	-34.7	-37.6968215	-0.086363732	-8.636373231
339	AAGT01000338.1/12653-12519	135	32.32242	-41.5	-42.1584662	-0.015866656	-1.586665638
340	AAXI01000285.1/23727-23857	131	31.96561	-41.3	-40.7922697	0.012293713	1.229371306
341	AAID01000631.1/14812-14946	135	29.36914	-40	-37.4589117	0.063527207	6.352720658
342	CU329671.1/467615-467481	135	24.78813	-31.2	-30.1691479	0.033040131	3.304013115

343	AAXT01000001.1/1022888-1022758	131	26.07575	-32	-31.4197437	0.018133011	1.813301074
344	AAFT01000065.1/55802-55646	157	30.50325	-40.2	-43.65482	-0.085940795	-8.594079497
345	ABAS01000010.1/147598-147449	150	30.64102	-44.5	-42.476849	0.045464067	4.546406738
346	AACM02000196.1/54942-55125	184	30.39285	-51.4	-48.8683422	0.049254044	4.925404358
347	AAVQ01000002.1/194294-194149	146	23.98397	-33.4	-31.0850945	0.069308549	6.930854926
348	ABAR01000001.1/643558-643722	165	30.91402	-46.1	-45.9052854	0.004223745	0.422374498
349	CAAL01000681.1/399056-398923	134	31.18101	-39.2	-40.1425477	-0.024044584	-2.404458435
350	BX842620.1/46696-46842	147	34.46857	-48.1	-47.9688325	0.002726974	0.272697441
351	AAFO01000045.1/228200-228380	181	23.62528	-38.5	-37.5003075	0.025966038	2.596603784
352	AACQ01000084.1/105435-105615	181	23.62528	-38.5	-37.5003075	0.025966038	2.596603784
353	AAJN01000077.1/141935-142086	152	34.77001	-51.5	-49.446518	0.039873437	3.987343712
354	AACD01000007.1/178412-178544	133	32.47971	-38.4	-42.0095688	-0.093999188	-9.399918767
355	CR382130.1/2966271-2966119	153	30.856	-45.7	-43.417754	0.049939737	4.993973697
356	AM269959.1/11788-11936	149	30.18837	-45.5	-41.556952	0.086660395	8.666039454
357	AAFM01000021.1/757822-757672	151	26.74181	-39.6	-36.4716416	0.07899895	7.899894963
358	AACY020397167.1/942-809	134	33.18461	-42.6	-43.3308628	-0.017156404	-1.715640381
359	AAKE03000003.1/920249-920379	131	32.48159	-43.7	-41.613358	0.047749244	4.774924443
360	AB189720.1/1-134	134	27.50248	-35.2	-34.2888906	0.02588379	2.588379015

361	AAGK01000002.1/935734-935865	132	32.05911	-42	-41.1406549	0.020460597	2.046059687
362	AAHF01000007.1/1275121-1274962	160	36.18263	-53	-53.291212	-0.005494566	-0.549456602
363	AC198144.2/107120-106981	140	20.006	-22.2	-23.5573459	-0.061141709	-6.114170922
364	ABDB01000004.1/78964-78804	161	36.19582	-52.7	-53.5118109	-0.015404381	-1.540438109
365	AAEE01000007.1/708153-708014	140	23.81557	-33.8	-29.6195096	0.123683149	12.36831487
366	AAEL01000435.1/3438-3299	140	23.81557	-33.8	-29.6195096	0.123683149	12.36831487
367	AAPO01000090.1/100324-100179	146	23.57147	-34.3	-30.4286726	0.112866688	11.28666883
368	AAQQ01631221.1/1703-1837	135	22.53748	-28	-26.5876909	0.050439611	5.043961088
369	CR382124.1/1170861-1171036	176	26.17229	-40.7	-40.555359	0.003553834	0.355383393
370	AAKD03000006.1/347523-347653	131	32.68262	-38.9	-41.933258	-0.077975784	-7.797578354
371	AAWC01000056.1/46279-46148	132	32.61995	-36	-42.0331247	-0.167586798	-16.75867985
372	ABCN01001426.1/26711-26844	134	33.38137	-44.7	-43.6439723	0.02362478	2.362478016
373	AANV02000065.1/936-807	130	30.46538	-38.8	-38.2053633	0.015325689	1.532568874
374	AAIM02000062.1/67925-68097	173	28.21567	-48	-43.2081897	0.099829381	9.982938139
375	AATU01001408.1/348748-348878	131	23.44115	-30.7	-27.2273081	0.113117001	11.31170006
376	AF270843.1/940-1107	168	28.70136	-47.7	-42.9830679	0.098887465	9.888746459
377	AAQX01001192.1/50846-50711	136	28.69341	-35.8	-36.5832282	-0.021877884	-2.187788394
378	AARE01001618.1/3212-3345	134	32.89491	-44.1	-42.8698723	0.027894052	2.789405179
379	AADM01000245.1/23026-22856	171	25.65719	-41.4	-38.7376886	0.064307039	6.430703922

380	ABBC01001255.1/19365-19501	137	28.38293	-37.9	-36.2887628	0.042512855	4.251285466
381	AATX01000107.1/49292-49428	137	28.38293	-37.9	-36.2887628	0.042512855	4.251285466
382	AASO01000240.1/58-194	137	28.38293	-37.9	-36.2887628	0.042512855	4.251285466
383	ABBB01000091.1/60646-60782	137	27.61541	-41.2	-35.0673955	0.148849624	14.88496237
384	CP000582.1/123212-123076	137	29.22702	-40.1	-37.6319585	0.061547169	6.154716926
385	AAIW01000278.1/37304-37168	137	28.5381	-37.7	-36.5356783	0.030883865	3.088386539
386	AACI02000576.1/2356-2546	191	25.71535	-43.1	-42.822231	0.006444756	0.644475571
387	AACF01000119.1/10356-10566	211	27.22105	-53.3	-49.210251	0.07673075	7.673074963
388	AAJI01000076.1/250-381	132	28.72737	-37.8	-35.8388626	0.051881943	5.188194302
389	AABZ01000001.1/29367-29202	166	26.81518	-41.1	-39.5823994	0.03692459	3.692458974
390	U18778.1/15676-15446	231	31.28412	-56.2	-59.6678143	-0.061704881	-6.170488097
391	AAEG01000006.1/103924-103694	231	31.28412	-56.2	-59.6678143	-0.061704881	-6.170488097
392	AABY01000063.1/21244-21029	216	30.20552	-58.1	-54.9574416	0.054088784	5.408878399
393	AADS01000270.1/32451-32584	134	31.88395	-40.9	-41.2611346	-0.008829697	-0.882969658
394	AE017356.1/390372-390503	132	30.97204	-35.1	-39.410805	-0.122814958	-12.28149583
395	AAEY01000066.1/357595-357726	132	30.97204	-35.1	-39.410805	-0.122814958	-12.28149583
396	AACO02000129.1/36387-36520	134	28.21331	-35.1	-35.4200391	-0.009117922	-0.911792221
397	AAZN01000370.1/73687-73536	152	23.71146	-35.2	-31.8490492	0.095197466	9.519746559
398	AAFP01000576.1/27349-27217	133	31.03615	-35.1	-39.7124208	-0.131408	-13.14080004
399	AANW02001764.1/480-354	127	22.21904	-27.7	-24.4841609	0.116095274	11.60952735

400	AACP01000036.1/91007-91135	129	28.79365	-36.9	-35.3455292	0.042126581	4.212658081
401	X13840.1/1-118	118	30.2782	-37.4	-35.5122953	0.050473387	5.047338668
402	AC149882.2/166659-166543	117	25.53738	-28.4	-27.7686349	0.022231164	2.223116444
403	AC190402.1/26804-26930	127	23.10208	-24.6	-25.889332	-0.052411869	-5.241186921
404	DQ451048.1/205-341	137	25.72333	-34.6	-32.0565314	0.073510654	7.351065358
405	AL590450.1/23479-23642	164	35.52032	-52.5	-53.0356879	-0.010203579	-1.020357854
406	ABIT01000802.1/5235-5099	137	29.61541	-41.2	-38.2499955	0.071602051	7.16020509
407	AAZY02000012.1/107951-108083	133	33.71269	-44.3	-43.9716033	0.007413018	0.741301801
408	AAKN02007150.1/17728-17588	141	27.94996	-36.3	-36.3981746	-0.002704535	-0.270453532
409	M15957.1/271-411	141	28.48824	-36.1	-37.2547357	-0.031987139	-3.198713917
410	AL844502.1/365248-365383	136	27.70793	-35.2	-35.0150232	0.005255023	0.525502329
411	AASC02060889.1/1492-1632	141	33.64563	-44.4	-45.4616946	-0.023912042	-2.391204155
412	AM055942.3/1615661-1615791	131	40.09625	-61.8	-53.7305572	0.130573508	13.05735081
413	AC189493.2/83218-83068	151	34.78525	-53.6	-49.2711618	0.080761906	8.076190627
414	AACE03000002.2/368765-368919	155	27.66631	-39.8	-38.7412036	0.026602924	2.660292358
415	AAGF03001264.1/335958-336087	130	31.36185	-36.7	-39.6319184	-0.079888785	-7.988878477
416	ABRQ01104616.1/4922-4782	141	27.71992	-36.9	-36.0321012	0.023520292	2.352029229
417	CR382136.2/1404151-1404006	146	24.95433	-34.4	-32.6292259	0.05147599	5.147599018
418	AAWR02035467.1/41938-42078	141	28.64287	-37.8	-37.5008034	0.007915253	0.791525282
419	AAGD02002631.1/646-784	139	33.51238	-41.3	-44.8504526	-0.085967375	-8.596737516

420	AAGJ04107959.1/2135-2275	141	28.2549	-35	-36.8834223	-0.053812066	-5.381206589
421	ABEG02000949.1/11902-11764	139	33.48231	-41.9	-44.8026031	-0.069274537	-6.927453695
422	AACS02000001.1/566465-566333	133	31.77406	-40.5	-40.8866622	-0.009547214	-0.954721442
423	AE016817.6/721964-721809	156	33.64904	-45.5	-48.4611214	-0.065079591	-6.507959142
424	AAPE02005503.1/37803-37943	141	27.33287	-33.4	-35.4161912	-0.060365006	-6.036500569
425	AAEC03000003.1/2600987-2601123	137	32.27387	-39.1	-42.4804157	-0.086455643	-8.645564327
426	ADTU01005844.1/39857-39997	141	28.31729	-38.6	-36.9827014	0.041898929	4.189892852
427	AAIL02000028.1/1015369-1015502	134	32.04156	-41.9	-41.5119366	0.009261657	0.926165666
428	AAWZ02025418.1/39904-40044	141	32.51923	-38.1	-43.6692429	-0.146174354	-14.61743537
429	AAQY02000456.1/44132-43994	139	32.8265	-39.4	-43.7590043	-0.110634627	-11.06346274
430	AACU03000146.1/753634-753781	148	26.92308	-38.4	-36.1613008	0.058299458	5.829945821
431	Z74042.2/30433-30295	139	30.81161	-39.2	-40.5527072	-0.034507836	-3.450783557
432	AFQF01001265.1/35199-35370	172	30.10803	-51.9	-46.0199072	0.113296585	11.32965852
IV. Rfam snRNA family RF00020 (180)							
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_M (kcal/mol)**	**MFE_C**	**RD1**	**% RD1**
433	AAFC03028536.1/14883-14999	117	29.50321	-33.1	-34.0794649	-0.029591086	-2.959108599
434	AAFR03051183.1/13754-13869	116	29.97912	-35.5	-34.6371787	0.024304825	2.430482531

435	AALT01479958.1/1552-1436	117	29.84312	-33.4	-34.6203637	-0.036537836	-3.653783607
436	AC083892.19/144818-144933	116	29.29846	-33.4	-33.554046	-0.004612154	-0.461215449
437	AANG01141002.1/575-691	117	29.29738	-34.8	-33.7519135	0.030117428	3.011742768
438	ABDC01046604.1/3549-3664	116	29.21227	-31.3	-33.4168799	-0.067631945	-6.763194538
439	AAQQ01236639.1/2054-2171	118	29.58748	-33.3	-34.4131524	-0.033427999	-3.342799911
440	AAIY01547840.1/3111-2996	116	28.98695	-35.4	-33.0583392	0.066148609	6.614860895
441	AAYZ01307320.1/25593-25710	118	29.51922	-34.9	-34.304542	0.017061835	1.70618351
442	AC068213.7/527-412	116	29.29846	-34.5	-33.554046	0.027418958	2.741895769
443	AANN01833323.1/1168-1053	116	29.17772	-31.3	-33.3619	-0.065875399	-6.587539941
444	AANN01833323.1/1168-1053	116	29.7257	-32.3	-34.2339104	-0.059873388	-5.987338844
445	AAPN01296676.1/948-833	116	29.84423	-36.1	-34.4225284	0.046467356	4.646735624
446	CAAE01011816.1/72155-72041	115	29.4541	-36.6	-33.6021054	0.08190969	8.190968976
447	K03164.1/1-115	115	35.09779	-45.9	-42.5829193	0.072267553	7.226755316
448	AAVX01303608.1/479-595	117	28.53738	-33.1	-32.5425349	0.016841845	1.684184502
449	BAAF04017217.1/172-285	114	28.63904	-37.8	-32.1055065	0.150647976	15.06479756
450	X63789.1/2235-2348	114	28.28458	-33.7	-31.5414587	0.06405167	6.405167009
451	X06020.1/401-515	115	28.91837	-32.6	-32.7495969	-0.004588861	-0.458886054
452	K03096.1/1-119	119	29.07063	-34	-33.7902981	0.006167704	0.616770379
453	AC174762.1/131410-131525	116	29.17772	-34	-33.3619	0.018767647	1.876764702
454	AASG02001471.1/28708-28826	119	35.14358	-44.7	-43.4541836	0.027870614	2.787061404

455	AM454615.1/12361-12244	118	34.77437	-43.6	-42.6670473	0.021397998	2.139799829
456	AP004339.3/129793-129675	119	36.00946	-46.4	-44.8320565	0.033791885	3.379188496
457	AAAA02022467.1/54175-54057	119	36.00946	-46.4	-44.8320565	0.033791885	3.379188496
458	ABAV01004466.1/10927-10807	121	29.75427	-38.9	-35.2773721	0.093126682	9.312668169
459	CAAJ01000127.1/3421-3534	114	28.48008	-29.7	-31.8525537	-0.072476557	-7.247655713
460	CAAI01002944.1/2056-1943	114	28.19527	-28.9	-31.3993388	-0.086482312	-8.648231217
461	AAJJ01001278.1/15201-15319	119	30.89161	-37.4	-36.6880224	0.019036836	1.903683551
462	X15935.1/3-121	119	28.79177	-39.4	-33.3465459	0.153640968	15.36409679
463	AC158186.2/4528-4644	117	33.86383	-42	-41.0185129	0.023368741	2.336874089
464	AC007202.3/14338-14454	117	28.82886	-34.2	-33.0063726	0.034901385	3.490138513
465	AATU01003637.1/103086-103202	117	28.95106	-36.2	-33.2008195	0.082850289	8.28502894
466	Z14994.1/1543-1661	119	33.16031	-44.8	-40.2981952	0.100486714	10.04867144
467	AADK01028234.1/592-479	114	28.70941	-33.3	-32.2174804	0.032508096	3.250809556
468	AASM01000961.1/2142-2261	120	29.12159	-38.9	-34.0709923	0.124139017	12.41390166
469	AARH01001853.1/639606-639489	118	34.70654	-42.3	-42.5591201	-0.006125772	-0.612577181
470	AC146755.23/40564-40682	119	34.79023	-43.7	-42.8918883	0.018492259	1.849225882
471	AAZO01006170.1/89711-89592	120	29.50004	-41.9	-34.6732067	0.172477168	17.2477168
472	AAZX01000523.1/17560-17680	121	33.63542	-42.9	-41.4534496	0.033719123	3.371912327
473	X74440.1/1-120	120	31.0164	-42	-37.0861906	0.116995462	11.69954621
474	EF647601.1/94040-94158	119	32.6885	-39.6	-39.5474033	0.0013282	0.132819998

475	AABL01000519.1/10755-10640	116	28.77741	-29.1	-32.7248871	-0.124566568	-12.45665676
476	AAKM01000004.1/1465347-1465233	115	28.63794	-37.9	-32.3033531	0.147668783	14.76687828
477	AAWU01000380.1/47408-47286	123	25.95488	-31.3	-29.6306071	0.053335237	5.333523699
478	AAPT01020503.1/50561-50682	122	22.83785	-24.6	-24.4708768	0.005248911	0.524891115
479	CR855038.1/57351-57233	119	33.67151	-40.5	-41.1116685	-0.015102926	-1.510292553
480	AAGE02022633.1/4541-4666	126	29.83394	-34.6	-36.4021494	-0.052085242	-5.208524174
481	AAQX01002881.1/912-1027	116	28.81244	-30.8	-32.7806309	-0.064306198	-6.430619753
482	AACY020397167.1/1194-1084	111	28.3058	-34.3	-30.9764158	0.096897498	9.689749753
483	AACT01019277.1/3233-3346	114	29.40381	-33.4	-33.3224836	0.00232085	0.232084975
484	AASV01046355.1/666-787	122	22.05026	-25.2	-23.2175719	0.078667783	7.866778287
485	AAEU02000660.1/100051-99931	121	25.02416	-25.4	-27.7503409	-0.092533108	-9.253310786
486	AAPQ01007319.1/728227-728345	119	29.26081	-31.9	-34.0929246	-0.068743718	-6.874371781
487	AAPP01015704.1/527093-527216	124	30.73464	-40.2	-37.4362272	0.068750567	6.875056711
488	AACW02000228.1/444426-444540	115	23.58505	-23.9	-24.2626951	-0.015175526	-1.517552576
489	AAAB01008944.1/3738569-3738447	123	30.05569	-37.9	-36.1562132	0.046010207	4.601020665
490	AAKO01001397.1/16278-16396	119	23.17452	-25.7	-24.4078116	0.050279703	5.027970294
491	AAPU01010615.1/222531-222648	118	23.26154	-24.7	-24.3466855	0.014304232	1.430423229
492	X01693.1/1-112	112	28.34027	-28.3	-31.2308792	-0.103564635	-10.35646347
493	AAYL01000045.1/86345-86236	110	31.93012	-36	-36.5442039	-0.015116776	-1.511677615

494	AANI01017247.1/37618-37740	123	29.92121	-36.5	-35.9422183	0.01528169	1.528169042
495	AY462110.1/1391-1506	116	30.67206	-42.1	-35.7398472	0.151072513	15.10725134
496	AY462110.1/1391-1506	118	35.36508	-46.6	-43.6070455	0.064226491	6.422649051
497	AF271469.1/310-194	117	29.89378	-34.2	-34.7009681	-0.014648189	-1.46481894
498	AAQB01006740.1/366264-366386	123	27.37269	-32.3	-31.8867538	0.012794	1.279400016
499	AADE01000447.1/59086-59201	116	23.21227	-25.1	-23.8690799	0.049040642	4.904064182
500	AB202073.1/931-819	113	30.25059	-32	-34.4703644	-0.077198886	-7.719888641
501	AABS01000112.1/71173-71286	114	29.38665	-32.2	-33.2951768	-0.034011701	-3.401170088
502	AAIZ01003066.1/82049-82166	118	25.33071	-30.5	-27.6393602	0.09379147	9.379147025
503	AAQA01000616.1/21218-21104	115	29.69287	-39.6	-33.9820677	0.141866977	14.18669767
504	AY705674.1/1315-1203	113	28.67537	-34.8	-31.9637146	0.081502453	8.15024527
505	Z69659.1/6667-6789	123	29.25678	-36	-34.8849097	0.030974731	3.097473127
506	CAAL01001847.1/78328-78213	116	28.33587	-33.4	-32.0222758	0.041249226	4.124922632
507	L22251.1/140-257	118	29.82513	-40.2	-34.7913316	0.134543991	13.45439911
508	AAXJ01002433.1/2315-2426	112	28.07198	-32.8	-30.8039411	0.060855455	6.08554554
509	AAXT01000002.1/226294-226405	112	33.30301	-39.1	-39.1280872	-0.000718342	-0.07183424
510	AAFP01000428.1/22975-22864	112	27.90976	-27.1	-30.5458083	-0.127151596	-12.71515964
511	AE017344.1/587581-587470	112	27.90976	-27.1	-30.5458083	-0.127151596	-12.71515964
512	AAEY01000021.1/77428-77539	112	27.90976	-27.1	-30.5458083	-0.127151596	-12.71515964
513	AAFI02000148.1/105951-105837	115	27.76152	-31.6	-30.9086999	0.021876584	2.1876584

514	DQ001173.1/3-117	115	27.76152	-31.6	-30.9086999	0.021876584	2.1876584
515	AAGK01000001.1/548506-548395	112	28.62368	-32.9	-31.6818692	0.037025253	3.702525315
516	AACO02000021.1/6655-6544	112	25.14378	-25.8	-26.1442892	-0.013344542	-1.334454189
517	AATT01000070.1/145415-145303	113	28.78073	-33.3	-32.1313739	0.035093876	3.509387598
518	X00386.1/1-104	104	27.14935	-25.5	-27.7389657	-0.087802578	-8.78025777
519	AAWT01083394.1/3003-2887	117	25.63197	-27.8	-27.9191566	-0.00428621	-0.428620953
520	AF095839.1/891-776	116	28.8649	-39.5	-32.8641197	0.16799697	16.79969699
521	X16573.1/258-372	115	24.74342	-27.3	-26.1060047	0.043736091	4.373609073
522	AAFU01001022.1/48358-48247	112	28.46461	-31.2	-31.4287398	-0.007331404	-0.733140426
523	ABCN01002017.1/34559-34665	107	29.0671	-33.4	-31.3894741	0.060195385	6.019538474
524	AAFB02000352.1/4039-4148	110	27.45639	-25.6	-29.4251493	-0.149419896	-14.94198956
525	AACM02000265.1/29897-29783	115	25.37629	-26.1	-27.1130957	-0.038815926	-3.881592565
526	CR382132.1/1370879-1370995	117	30.17921	-39.1	-35.1551707	0.100890774	10.08907745
527	AANV02000637.1/5704-5595	110	27.91205	-26.5	-30.1502461	-0.137745137	-13.77451367
528	AATM01000006.1/4956-4838	119	29.51816	-32	-34.5024551	-0.078201723	-7.820172272
529	AAIM02000161.1/229078-228967	112	28.35807	-28.1	-31.2591981	-0.112426978	-11.2426978
530	AARE01001511.1/1260-1149	112	28.054	-30.1	-30.7753333	-0.022436324	-2.243632388
531	AAJI01001476.1/13492-13374	119	29.07063	-34.2	-33.7902981	0.011979589	1.197958857
532	AF529186.1/1009-894	116	29.50431	-36	-33.8816096	0.058844179	5.884417885
533	AAEE01000007.1/707570-707686	117	29.22844	-38	-33.6422171	0.114678497	11.46784974

534	AAEL01000435.1/2854-2970	117	29.17663	-39.4	-33.5597747	0.148229069	14.82290686
535	AAWC01002764.1/14312-14199	114	33.3523	-40	-39.6057152	0.009857121	0.985712062
536	AAFT01000039.1/59328-59208	121	25.97374	-29.2	-29.2614101	-0.002103085	-0.210308507
537	Z11883.1/951-834	118	29.92641	-34.4	-34.952489	-0.016060727	-1.606072686
538	AAKE03000008.1/776195-776083	113	28.35694	-32.5	-31.4569993	0.032092328	3.209232828
539	AAHF01000006.1/1883955-1883843	113	28.35694	-32.5	-31.4569993	0.032092328	3.209232828
540	ABDB01000059.1/258270-258158	113	28.35694	-32.5	-31.4569993	0.032092328	3.209232828
541	AATX01000063.1/152569-152455	115	28.9358	-32.6	-32.7773435	-0.005439984	-0.543998444
542	ABBB01000033.1/420348-420462	115	28.9358	-32.6	-32.7773435	-0.005439984	-0.543998444
543	AASO01001658.1/2059-1945	115	28.9358	-32.6	-32.7773435	-0.005439984	-0.543998444
544	ABBC01000392.1/6892-7006	115	28.9358	-32.6	-32.7773435	-0.005439984	-0.543998444
545	AAXI01000301.1/8428-8544	117	28.93363	-38.3	-33.1730917	0.133861836	13.38618362
546	AAVQ01000002.1/3936-3813	124	22.52299	-24.5	-24.3690345	0.005345532	0.534553212
547	AAGI01000327.1/72210-72324	115	28.54974	-31.6	-32.1630031	-0.017816553	-1.781655316
548	AAKD03000017.1/303987-303868	120	29.277	-38.7	-34.3182936	0.113222388	11.3222388
549	AM270115.1/128938-128823	116	28.24674	-28.8	-31.8804357	-0.106959572	-10.69595722
550	AANW02001233.1/8188-8079	110	24.83965	-26	-25.2611282	0.028418147	2.841814732
551	AP007155.1/1026766-1026646	121	29.80506	-37	-35.3581937	0.044373143	4.437314284
552	AAIW01000368.1/11893-11779	115	25.79602	-29.1	-27.781	0.045326462	4.532646211

553	AAJN01000094.1/53887-53768	120	29.5342	-39.9	-34.7275708	0.129634817	12.96348174
554	AANS01001320.1/16875-16995	121	26.75427	-30.6	-30.5034721	0.003154507	0.315450712
555	AAFM01000003.1/226344-226452	109	25.55664	-25.1	-26.2024755	-0.043923326	-4.392332572
556	AL590450.1/114088-114197	110	28.46695	-33.3	-31.03325	0.068070571	6.807057077
557	AC004395.1/1702-1592	111	28.5012	-27.9	-31.2873534	-0.121410517	-12.14105172
558	AACD01000010.1/103824-103714	111	28.5012	-27.9	-31.2873534	-0.121410517	-12.14105172
559	AAPO01000006.1/285980-286104	125	25.53336	-28.2	-29.3590327	-0.041100449	-4.110044884
560	AAIH02000216.1/25877-25996	120	29.70443	-38.7	-34.9984536	0.095647195	9.564719498
561	AAGT01000497.1/29058-28944	115	23.07031	-24	-23.4435902	0.023183742	2.318374178
562	X87329.1/2199-2085	115	33.4258	-38.4	-39.9222814	-0.039642745	-3.964274537
563	AAFO01000011.1/73327-73206	122	23.01553	-25.9	-24.7536086	0.044262214	4.426221444
564	AACQ01000039.1/46591-46712	122	22.41243	-23.4	-23.7938964	-0.016833178	-1.68331777
565	CR382125.1/1422292-1422448	157	34.49024	-48.8	-49.9993251	-0.024576334	-2.457633382
566	AAID01003647.1/370-250	121	26.03389	-30	-29.3571264	0.021429121	2.142912115
567	AC091619.3/67508-67375	134	20.02881	-24.1	-22.3960425	0.07070363	7.070362964
568	AACA01000117.1/2098-2219	122	23.8885	-25.6	-26.1427669	-0.021201832	-2.120183217
569	AACG02000018.1/45887-45766	122	23.8885	-25.6	-26.1427669	-0.021201832	-2.120183217
570	AC189540.1/19288-19170	119	30.32703	-37.4	-35.7896045	0.043058702	4.305870207
571	AACI02000988.1/1421-1535	115	22.77849	-23.7	-22.9792185	0.03041272	3.04127199
572	AACH01000658.1/4206-4082	125	24.03694	-29.5	-26.977782	0.085498914	8.549891449

573	DQ028748.1/5155-5269	115	20.06868	-18.6	-18.6670928	-0.003607138	-0.360713757
574	AADM01000037.1/36531-36659	129	25.09337	-30.9	-29.4572791	0.046689996	4.668999638
575	AABY01000025.1/44800-44680	121	24.41343	-26.3	-26.7784948	-0.01819372	-1.81937197
576	CR380948.1/197408-197529	122	22.62166	-24.5	-24.1268447	0.015230829	1.523082937
577	AAFD02000027.1/316326-316217	110	32.62603	-36.1	-37.6516001	-0.042980612	-4.29806118
578	AAKN02051087.1/45335-45220	116	28.72478	-30.3	-32.641144	-0.077265479	-7.72654787
579	AAGV020174570.1/3903-4018	116	29.33287	-31.9	-33.6087995	-0.053567383	-5.356738328
580	AASC02039457.1/2620-2506	115	28.54974	-33.7	-32.1630031	0.045608217	4.560821721
581	ABIS01000186.1/30072-30186	115	28.9358	-32.6	-32.7773435	-0.005439984	-0.543998444
582	AABX02000002.1/78945-78831	115	33.37252	-38.7	-39.8374941	-0.029392613	-2.939261314
583	AP009663.1/4379-4498	120	25.91308	-30.2	-28.9652895	0.040884455	4.088445454
584	AAWR02001149.1/42020-41905	116	28.81244	-32.4	-32.7806309	-0.011747867	-1.174786678
585	BX890568.8/147618-147503	116	33.17772	-38.9	-39.7271	-0.021262211	-2.126221083
586	AC110235.13/80163-80048	116	33.33287	-38.1	-39.9739995	-0.049186339	-4.918633928
587	AAGD02000134.1/1661-1782	122	29.1369	-34.4	-34.4945418	-0.002748307	-0.274830706
588	AM055942.3/1616414-1616305	110	31.93012	-36	-36.5442039	-0.015116776	-1.511677615
589	CAAC02000457.1/1551549-1551428	122	29.41243	-33.3	-34.9329964	-0.04903893	-4.903892967
590	CR382134.2/219036-219157	122	24.01553	-26	-26.3449086	-0.013265717	-1.326571715
591	AAFN02000018.1/172560-172433	128	21.91642	-23.6	-24.2021976	-0.025516846	-2.551684576
592	AE016817.6/459194-459336	143	32.2689	-44.9	-43.6700952	0.027392089	2.739208869

593	ABDG02000015.1/1095640-1095526	115	26.28235	-29.4	-28.5549068	0.028744667	2.874466722
594	AADG06003819.1/2345-2227	119	30.00946	-34.1	-35.2842565	-0.034728931	-3.472893074
595	AFQF01001308.1/86869-86981	113	28.25	-28.9	-31.286825	-0.0825891	-8.258910035
596	AACU03000132.1/1014141-1014032	110	28.14608	-25.6	-30.5226575	-0.19229131	-19.22913102
597	CAAB02011239.1/13744-13857	114	29.07604	-35.3	-32.8008977	0.070796099	7.079609855
598	AACN010031598.1/2460-2578	119	29.70548	-36.1	-34.8005226	0.035996603	3.599660345
599	AAWZ02006407.1/13458-13573	116	29.65776	-36.9	-34.1257903	0.075181835	7.518183519
600	AACS02000007.1/597289-597173	117	28.89875	-39.7	-33.1175858	0.165803884	16.58038837
601	ABDF02000090.1/29527-29640	114	28.93692	-27.9	-32.5795151	-0.167724556	-16.77245564
602	AAGW02002965.1/6736-6851	116	28.98695	-33.2	-33.0583392	0.00426689	0.426689027
603	AAGJ04035404.1/487-369	119	29.00117	-36.6	-33.67976	0.079787977	7.978797693
604	AAIL02000016.1/31650-31761	112	28.37586	-27.4	-31.2874992	-0.141879534	-14.18795345
605	AAQR03026016.1/3436-3551	116	29.45298	-33.2	-33.7999333	-0.01807028	-1.807027973
606	AE014187.2/1889353-1889471	119	28.96637	-37.7	-33.6243916	0.108106323	10.81063229
607	AAPE02065766.1/2803-2918	116	29.6067	-36.2	-34.0445373	0.059543168	5.95431678
608	AACZ03099104.1/4171-4286	116	29.38441	-33.5	-33.6908097	-0.005695811	-0.569581081
609	AAQY02000250.1/229279-229396	118	28.75776	-32.2	-33.0928176	-0.027727254	-2.772725432
610	ABEG02003930.1/15093-15214	122	29.20602	-33.5	-34.6045432	-0.032971439	-3.297143889
V. Rfam snRNA family RF00026							

Sl. No.	Sequence ID	NTL	SD_DFT	MFE_M	MFE_C	RD1	% RD1
611	AB010698.1/46416-46518	103	22.99796	-22.2	-20.9332543	0.057060618	5.706061829
612	AARH01001853.1/272694-272592	103	22.4727	-22.1	-20.0974098	0.090614941	9.061494106
613	X60506.1/390-492	103	24.56789	-25.5	-23.4314882	0.081118109	8.111810932
614	AC146705.11/15272-15374	103	24.56789	-25.5	-23.4314882	0.081118109	8.111810932
615	AASG02002949.1/2307-2409	103	24.56789	-25.5	-23.4314882	0.081118109	8.111810932
616	AACV01009611.1/38269-38167	103	24.56789	-25.5	-23.4314882	0.081118109	8.111810932
617	AAAA02013555.1/2292-2394	103	24.56789	-25.5	-23.4314882	0.081118109	8.111810932
618	CR855100.1/43897-43999	103	24.56789	-25.5	-23.4314882	0.081118109	8.111810932
619	X52315.1/1-103	103	22.97717	-23.4	-20.9001647	0.106830567	10.6830567
620	X51447.1/262-364	103	23.41542	-25.4	-21.5975607	0.149702333	14.97023331
621	AAXJ01018701.1/864-762	103	25.02014	-25.9	-24.1511567	0.067522906	6.752290645
622	AAQA01000086.1/111608-111711	104	23.08723	-21.1	-21.2749128	-0.008289705	-0.828970457
623	L26849.1/150-253	104	22.08723	-21.1	-19.6836128	0.067127357	6.712735704
624	X51387.1/158-259	102	21.89545	-20	-18.9792273	0.051038633	5.103863332
625	AANU01133867.1/371-269	103	23.92655	-24	-22.4109226	0.066211559	6.6211559
626	X63066.1/363-465	103	21.4727	-21	-18.5061098	0.118756676	11.87566761
627	AAAB01008807.1/5121648-5121542	107	27.294	-26.6	-28.5679372	-0.073982601	-7.398260114

628	AAWU01008690.1/11499-11605	107	24.10845	-25.7	-23.4987725	0.085650877	8.565087674
629	AAGE02013372.1/83708-83814	107	24.10845	-25.7	-23.4987725	0.085650877	8.565087674
630	AABU01002774.1/16804311-16804417	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
631	AAGH01007200.1/1053-947	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
632	AADE01001799.1/12821-12927	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
633	AAEU02000254.1/81013-80907	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
634	AAPQ01001284.1/654-760	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
635	AAPP01016905.1/3608-3714	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
636	AAQB01008633.1/461913-461807	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
637	AANI01001778.1/524-630	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
638	AAIZ01003051.1/11867-11761	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
639	AAPT01019380.1/23511-23405	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
640	AAPU01011102.1/63061-63167	107	22.10845	-21	-20.3161725	0.032563216	3.256321583
641	AABS01000051.1/274751-274857	107	22.90845	-23.9	-21.5892125	0.096685671	9.668567081
642	AAHY01132583.1/7739-7845	107	27.294	-25.9	-28.5679372	-0.103009158	-10.30091579
643	AAFC03115769.1/4330-4436	107	23.62485	-22.2	-22.7292163	-0.02383857	-2.383857015
644	AAPY01153714.1/3278-3172	107	23.62485	-22.2	-22.7292163	-0.02383857	-2.383857015
645	AALT01529386.1/701-595	107	25.53334	-24.4	-25.7662067	-0.055992076	-5.599207629
646	AAYZ01436191.1/1883-1777	107	25.53334	-24.4	-25.7662067	-0.055992076	-5.599207629
647	BAAB01141103.1/711-817	107	24.10845	-24.3	-23.4987725	0.032972326	3.297232643

648	AAZX01000860.1/13921-13815	107	24.10845	-24.3	-23.4987725	0.032972326	3.297232643
649	AACY022149992.1/529-635	107	24.10845	-24.3	-23.4987725	0.032972326	3.297232643
650	AATU01006594.1/28185-28081	105	27.18526	-29.8	-27.9957072	0.060546739	6.054673917
651	ABAV01046662.1/5606-5712	107	27.294	-27.3	-28.5679372	-0.046444586	-4.644458573
652	AAFR03037834.1/2280-2178	103	26.35802	-28.4	-26.2801136	0.074643887	7.464388749
653	ABDC01230141.1/8073-7967	107	26.53334	-24.4	-27.3575067	-0.121209289	-12.12092894
654	AANN01390182.1/1050-1156	107	27.53334	-26.1	-28.9488067	-0.109149681	-10.91496805
655	M31687.1/705-811	107	27.53334	-26.1	-28.9488067	-0.109149681	-10.91496805
656	AACT01041609.1/35558-35664	107	23.34942	-23.6	-22.290926	0.055469238	5.546923787
657	AANG01770494.1/575-681	107	24.07118	-24.7	-23.4394767	0.051033333	5.10333332
658	AAQQ01629113.1/2249-2143	107	25.53334	-23.9	-25.7662067	-0.078083961	-7.808396073
659	ABBA01062195.1/38876-38770	107	27.44153	-26.1	-28.8027115	-0.103552166	-10.35521661
660	AAIY01587713.1/2016-2122	107	27.44153	-26.1	-28.8027115	-0.103552166	-10.35521661
661	AAPN01022574.1/939-833	107	27.44153	-26.1	-28.8027115	-0.103552166	-10.35521661
662	CR956385.13/177771-177665	107	27.44153	-26.1	-28.8027115	-0.103552166	-10.35521661
663	U43841.1/336-439	104	20.35678	-20.6	-16.9299368	0.17815841	17.81584097
664	AANV02000039.1/44432-44535	104	20.35678	-20.6	-16.9299368	0.17815841	17.81584097
665	AAFB02000174.1/3561-3664	104	20.35678	-20.6	-16.9299368	0.17815841	17.81584097
666	CAAI01006173.1/984-1090	107	24.21993	-23.8	-23.6761728	0.005202822	0.520282232
667	BAAF04101838.1/670-565	106	27.10965	-27.6	-28.0749922	-0.017209863	-1.720986285

668	AAVX01043085.1/4211-4317	107	27.56998	-26.1	-29.0071085	-0.111383467	-11.13834671
669	AC148181.3/26639-26533	107	27.34942	-26.4	-28.656126	-0.085459318	-8.545931766
670	CT573239.9/51807-51913	107	27.44153	-26.2	-28.8027115	-0.099340135	-9.934013489
671	AAYL01000007.1/235430-235323	108	27.47709	-25.6	-29.0588985	-0.135113223	-13.5113223
672	CAAE01010022.1/48689-48584	106	24.42435	-26.7	-23.8018758	0.108543976	10.85439757
673	AF529186.1/459-564	106	26.94156	-27.8	-27.8075034	-0.000269905	-0.026990501
674	AAXT01000001.1/1039074-1039179	106	24.979	-22	-24.6844894	-0.122022243	-12.20222434
675	DQ666642.1/4-106	103	22.68167	-23.7	-20.4299489	0.137976838	13.79768377
676	AAWT01067003.1/344-451	108	27.87824	-28.7	-29.6972432	-0.03474715	-3.474714993
677	AANH01010141.1/91130-91025	106	24.05374	-23.2	-23.2121139	-0.000522152	-0.052215217
678	AB220565.1/723-829	107	27.18282	-26.6	-28.3910202	-0.067331587	-6.733158699
679	AAGK01000002.1/918046-917941	106	22.43086	-20.3	-20.629631	-0.016237981	-1.623798056
680	AC136964.2/84202-84308	107	23.21993	-21.3	-22.0848728	-0.03684849	-3.684848961
681	X71486.1/1-101	101	23.03286	-22.3	-20.5895962	0.07669972	7.66997203
682	AACM02000382.1/262414-262518	105	19.09179	-17.6	-15.1165601	0.141104541	14.11045409
683	AC146661.3/155499-155393	107	20.08368	-15.9	-17.0941639	-0.075104648	-7.51046475
684	AC087806.3/115795-115689	107	26.84652	-26.1	-27.8558736	-0.06727485	-6.727485008
685	L25920.1/3-109	107	27.3863	-26.5	-28.7148194	-0.08357809	-8.357808952
686	CU326409.1/117472-117365	108	22.80773	-23.2	-21.6285353	0.067735546	6.773554556

687	AC188110.1/71045-71151	107	25.64311	-22.8	-25.9408803	-0.13775791	-13.77579095
688	CR382132.1/1089192-1089093	100	21.99534	-20.6	-18.7389807	0.090340743	9.034074336
689	AF095841.1/1-107	107	23.89756	-21.1	-23.1631796	-0.097781022	-9.778102155
690	AASM01002098.1/3099-2992	108	23.69889	-23.9	-23.0466404	0.035705424	3.570542437
691	X58843.1/3-106	104	23.05622	-23.9	-21.2255568	0.111901388	11.19013876
692	CAAJ01009065.1/446-339	108	23.2737	-24.5	-22.370032	0.08693747	8.693747
693	AAFU01001153.1/10970-11070	101	22.47527	-20	-19.7022995	0.014885025	1.48850246
694	AABL01000365.1/9537-9644	108	23.32719	-22.9	-22.4551634	0.019425177	1.942517744
695	AADS01000210.1/17114-17228	115	28.9358	-30.2	-32.7773435	-0.0853425	-8.534249976
696	AAPO01000024.1/134030-133927	104	20.00761	-17.1	-16.3743159	0.042437666	4.243766623
697	AAWC01002368.1/53508-53403	106	23.84447	-24.6	-22.8790987	0.069955337	6.995533683
698	CT990557.10/71286-71393	108	22.16306	-21.7	-20.6026758	0.050567934	5.056793412
699	AAFM01000021.1/681699-681594	106	20.48809	-19.4	-17.5381017	0.095974139	9.597413942
700	AAFT01000065.1/268653-268548	106	20.04611	-18.5	-16.8347795	0.090011917	9.001191658
701	AAEY01000056.1/129419-129306	114	22.88948	-26.6	-22.9562271	0.136983945	13.69839445
702	AACO02000104.1/123353-123240	114	22.88948	-26.6	-22.9562271	0.136983945	13.69839445
703	AE017348.1/872264-872377	114	22.88948	-26.6	-22.9562271	0.136983945	13.69839445
704	AAFP01000223.1/16410-16523	114	22.88948	-26.6	-22.9562271	0.136983945	13.69839445
705	CP000496.1/1065610-1065505	106	20.62115	-19.5	-17.7498386	0.089751867	8.975186697

706	AAID01000554.1/5500-5599	100	18.34609	-13	-12.9319319	0.005236008	0.523600837
707	X14196.1/133-285	153	22.47914	-30.2	-30.0876604	0.003719854	0.371985354
708	AAZN01000268.1/119544-119646	103	20.43453	-16.1	-16.8540657	-0.046836379	-4.683637919
709	AACW02000046.1/25302-25425	124	24.87427	-28.4	-28.1106289	0.010189125	1.018912464
710	CU104654.2/178130-178232	103	17.73774	-13.1	-12.5626691	0.041017624	4.101762353
711	EF419774.1/1-102	102	22.61295	-21.2	-20.1209942	0.050896498	5.089649786
712	AACI02000106.1/4087-4199	113	28.46348	-33.3	-31.6265343	0.050254225	5.025422515
713	CR382126.1/1858703-1858820	118	28.66995	-30.4	-32.9530853	-0.083983069	-8.39830691
714	AACH01000157.1/7933-7821	113	28.14265	-30.9	-31.116004	-0.006990421	-0.69904214
715	AATM01000137.1/47790-47943	154	17.71006	-26.3	-22.6982109	0.136950157	13.6950157
716	Z73279.1/2843-2955	113	28.28569	-30.9	-31.3436213	-0.014356675	-1.435667521
717	AACA01000433.1/2210-2322	113	28.28569	-30.9	-31.3436213	-0.014356675	-1.435667521
718	AAEG01000112.1/87347-87459	113	28.28569	-30.9	-31.3436213	-0.014356675	-1.435667521
719	AABY01000279.1/8828-8716	113	28.28569	-30.9	-31.3436213	-0.014356675	-1.435667521
720	AACG02000194.1/7974-8086	113	28.28569	-30.9	-31.3436213	-0.014356675	-1.435667521
721	AACF01000007.1/86925-87036	112	28	-26.4	-30.6894	-0.162477273	-16.24772727
722	AADM01000279.1/397-288	110	24.37818	-25.7	-24.5267979	0.045649889	4.564988885
723	AF083031.2/127905-127809	97	22.50899	-19.3	-18.9575484	0.017743607	1.774360728
724	AC144401.2/82868-82974	107	20.39147	-18.8	-17.5839385	0.064684121	6.468412078
725	AANW02001116.1/942-1048	107	23.8841	-24.3	-23.1417649	0.047663996	4.76639958

726	AY953942.1/4-110	107	24.7146	-23.3	-24.4633417	-0.04992883	-4.992882978
727	AJ416571.1/12089-11984	106	18.21995	-15.6	-13.9288045	0.107127917	10.71279166
728	AATT01000229.1/70559-70673	115	23.90092	-28.6	-24.7653335	0.134079249	13.40792492
729	AL590448.1/66612-66503	110	24.02862	-27.5	-23.9705367	0.128344119	12.83441186
730	AY136823.1/430-532	103	20.44656	-18.2	-16.8732129	0.072900388	7.290038751
731	AF305715.1/117-214	98	26.03685	-28.4	-24.7710384	0.127780337	12.77803372
732	X82228.1/412-509	98	26.03685	-28.4	-24.7710384	0.127780337	12.77803372
733	AAFI02000006.1/10321-10427	107	24.48691	-23.3	-24.101026	-0.034378797	-3.437879673
734	AF053588.1/116-223	108	29.92139	-37.4	-32.9485116	0.119023753	11.90237531
735	AAJI01001561.1/684-795	112	19.00273	-19.2	-16.3720454	0.147289304	14.72893036
736	AACQ01000098.1/66592-66693	102	21.28255	-19.8	-18.0039171	0.09071126	9.071125991
737	AAFO01000026.1/267271-267372	102	21.28255	-19.8	-18.0039171	0.09071126	9.071125991
738	AAIM02000091.1/125555-125392	164	28.84169	-44.6	-42.4079834	0.049148355	4.914835521
739	X78552.1/318-415	98	26.1916	-28.6	-25.0172958	0.125269378	12.52693777
740	X79014.1/475-572	98	26.153	-28.4	-24.9558681	0.121272251	12.12722514
741	X78551.1/329-426	98	26.2494	-28.4	-25.109268	0.115870844	11.58708443
742	AC149301.1/92055-91961	95	25.65006	-30.4	-23.556739	0.225107268	22.51072681
743	AAEE01000007.1/553473-553580	108	27.44033	-28.8	-29.0004044	-0.006958487	-0.695848747
744	AAEL01000070.1/7754-7647	108	27.44033	-28.8	-29.0004044	-0.006958487	-0.695848747
745	X82229.1/194-291	98	25.92018	-28	-24.5853806	0.121950694	12.19506944

746	AAHK01000939.1/1885-1993	109	27.82269	-33.8	-29.8084401	0.118093488	11.8093488
747	CP000581.1/336841-336949	109	30.29033	-37.6	-33.7351971	0.102787311	10.27873112
748	AC152105.2/17973-17865	109	30.29033	-37.6	-33.7351971	0.102787311	10.27873112
749	AC092562.4/129660-129562	99	22.55414	-20.7	-19.4286059	0.061420004	6.142000417
750	AAXI01000109.1/30199-30115	85	24.61973	-21.1	-19.9211714	0.055868654	5.586865365
751	DQ103593.1/29389-29298	92	21.2173	-16.7	-15.9040853	0.04765956	4.765956019
752	AP004520.1/55775-55881	103	18.66274	-15.3	-14.0346258	0.082704195	8.270419506
753	AAHF01000007.1/1651650-1651551	100	22.1889	-21	-19.0469997	0.093000015	9.300001483
754	U58510.1/7462-7568	107	21.515	-21.3	-19.3718266	0.090524571	9.052457119
755	ABAR01000008.1/592107-592006	102	21.8151	-22	-18.8513662	0.14311972	14.31197198
756	AP007171.1/1600650-1600749	100	24.55279	-24.8	-22.8086502	0.080296364	8.02963637
757	AARE01000569.1/1814-1714	101	22.64639	-20.9	-19.9745974	0.044277635	4.42776354
758	AASO01000114.1/1873-1980	108	24.67839	-24.4	-24.6053273	-0.008415054	-0.841505425
759	AAIW01000495.1/17855-17954	100	22.49566	-21.7	-19.5351499	0.099762679	9.976267928
760	X04788.1/1-98	98	21.90068	-20.1	-18.1891563	0.095066852	9.506685209
761	AAGI01000262.1/1755-1855	101	19.34482	-15.2	-14.7208152	0.031525317	3.152531716
762	AY102720.1/4016-3910	107	17.5559	-14.2	-13.0716958	0.079458039	7.945803884
763	X57046.1/441-540	100	26.51472	-26.2	-25.9306717	0.010279705	1.027970512
764	AC148038.3/33473-33580	108	20.78758	-21	-18.4138741	0.123148854	12.31488542
765	AAZY02000001.1/305276-	108	27.42194	-30.2	-28.971128	0.040691126	4.069112584

	305169						
766	CR380959.2/1159577-1159689	113	28.0349	-28	-30.944529	-0.105161749	-10.51617488
767	CAAC02000605.1/61448-61549	102	18.73898	-15.6	-13.9563376	0.105362976	10.53629758
768	M10329.1/1-108	108	27.60536	-25.9	-29.2630144	-0.129846115	-12.98461152
769	CR382136.2/1319844-1319739	106	20.20069	-19.5	-17.080759	0.12406364	12.40636395
770	X12565.1/540-652	113	28.28569	-30.9	-31.3436213	-0.014356675	-1.435667521
771	AASC02049566.1/23588-23693	106	22.56421	-21	-20.8418241	0.007532187	0.753218684
772	AC121317.7/68985-69092	108	27.45872	-26.1	-29.0296613	-0.112247558	-11.2247558
773	AAPN01095965.1/240-347	108	27.73304	-28.8	-29.4661862	-0.023131465	-2.313146523
774	AAFD02000024.1/69022-69131	110	27.3459	-20.6	-29.2493283	-0.419870306	-41.98703055
775	ABPA01000003.1/180549-180442	108	27.47709	-25.6	-29.0588985	-0.135113223	-13.5113223
776	AACE03000008.2/905684-905797	114	28.26674	-32.7	-31.5130707	0.036297533	3.629753328
777	FR796420.1/621391-621294	98	26.09499	-28.4	-24.8635561	0.124522673	12.45226733
778	AAFN02000036.1/11182-11077	106	18.3352	-15.2	-14.1122067	0.071565347	7.156534711
779	AAGV020896327.1/1288-1181	108	27.55046	-26.7	-29.1756525	-0.092721068	-9.272106813
780	AATU01006589.1/7799-7695	105	25.18526	-26.1	-24.8131072	0.049306239	4.930623859
781	AFQF01002057.1/12805-12639	167	28.93962	-45.5	-43.1626123	0.051371159	5.137115919
782	CAAA01180605.1/4-106	103	22.73838	-22.7	-20.5201904	0.096026855	9.602685524
783	ABDF02000090.1/1201765-1201660	106	21.58951	-20	-19.2907841	0.035460795	3.546079497
784	AACU03000132.1/2986497-2986398	100	21.05356	-19	-17.2403267	0.092614386	9.261438567

785	AAIL02000075.1/131155-131066	90	18.09252	-11.5	-10.5324244	0.084137011	8.413701076
786	X76546.1/281-387	107	24.20138	-24.2	-23.6466566	0.02286543	2.286543009
787	CAAB02029384.1/647-542	106	25.33217	-26.7	-25.2464851	0.054438761	5.443876051
788	AAEX03005034.1/14917-14815	103	23.83263	-24.2	-22.261469	0.080104588	8.01045883
789	AC242743.1/25579-25465	115	26.9358	-30.2	-29.5947435	0.020041606	2.00416062
790	BAAB01206473.1/1073-967	107	23.10845	-24.3	-21.9074725	0.098457923	9.845792314
791	M24606.1/401-507	107	24.10845	-24	-23.4987725	0.020884481	2.088448051
792	AL844509.2/1632740-1632633	108	23.69889	-23.9	-23.0466404	0.035705424	3.570542437
793	AC188639.6/7118-7225	108	23.44033	-22.7	-22.6352044	0.00285443	0.285443

VI. Rfam snRNA family RF00283

Sl. No.	Sequence ID	NTL	SD_DFT	MFE_M	MFE_C	RD1	% RD1
794	AF357342.1/1-62	62	24.90107	-14.5	-15.778073	-0.088142966	-8.814296637
795	AY077741.1/1-83	83	26.25462	-22.5	-22.1235709	0.01673018	1.673018008
796	AC090227.10/15359-15277	83	26.25462	-22.5	-22.1235709	0.01673018	1.673018008
797	AL592064.14/82927-83010	84	26.90008	-26	-23.3503019	0.101911467	10.19114672
798	AANU01101212.1/1480-1562	83	26.46581	-21.7	-22.4596412	-0.035006507	-3.500650743
799	ABDC01297764.1/1645-1727	83	26.93963	-21.4	-23.2136402	-0.084749544	-8.474954417
800	AAYZ01094280.1/2675-2757	83	29.27389	-27.1	-26.9281344	0.006341903	0.634190342
801	AAFC03053505.1/66323-66241	83	26.78892	-24.8	-22.9738116	0.07363663	7.363663044

802	AAPY01347300.1/1993-1911	83	27.19672	-26.1	-23.6227452	0.094913977	9.491397659
803	AAIY01249536.1/1104-1186	83	26.63736	-25.3	-22.732626	0.101477234	10.14772343
804	AANN01109527.1/812-730	83	26.46581	-24.3	-22.4596412	0.07573493	7.573492958
805	AANG01025936.1/3125-3205	81	25.79171	-22.7	-20.9877541	0.075429333	7.542933273
806	AAHX01095967.1/17609-17687	79	25.22006	-21.8	-19.6788769	0.097299224	9.729922367
807	AAKN02020097.1/24241-24159	83	28.02227	-26.5	-24.9364383	0.059002329	5.900232942
808	AAGV020429307.1/439-358	82	26.58227	-21.9	-22.4453724	-0.02490285	-2.490284954
809	AAGU03013210.1/58113-58031	83	28.59933	-26.1	-25.8547148	0.009397899	0.939789885
810	AAGW02032389.1/3932-4014	83	26.98334	-26.8	-23.2831947	0.131224078	13.12240775
811	AAEX03017760.1/16604-16686	83	26.63736	-25.7	-22.732626	0.115462024	11.54620245
812	AAPE02017035.1/34119-34194	76	24.5134	-20.1	-17.9555698	0.106688072	10.66880719
813	AAQR03077543.1/26098-26180	83	29.59933	-27.6	-27.4460148	0.005579172	0.557917247

VI. Rfam snRNA family RF00492

Sl. No.	Sequence ID	NTL	SD_DFT	MFE_M	MFE_C	RD1	% RD1
814	AC090227.10/15587-15444	144	39.9353	-64.6	-56.0692506	0.132054944	13.20549443
815	AC129097.27/183442-183585	144	32.68669	-45.1	-44.5345268	0.012538208	1.253820822
816	AC018751.30/166114-165972	143	37.2071	-59.1	-51.5282529	0.128117548	12.81175484
817	AC125020.7/181527-181661	135	34.25922	-45.8	-45.2404939	0.012216291	1.221629085
818	AC127289.4/19294-19156	139	34.72282	-51	-46.7766162	0.082811446	8.281144644

819	AC023490.5/121842-121984	143	35.38856	-55.8	-48.6344116	0.128415562	12.84155625
VII. Rfam snRNA family RF01458							
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_M**	**MFE_C**	**RD1**	**% RD1**
820	AB261975.1/7738-7641	98	22.2688	-19.6	-18.7749466	0.042094563	4.209456334
821	CP000255.1/68682-68585	98	21.36127	-19.3	-17.3307913	0.102031541	10.20315412
822	CP000703.1/63982-64079	98	22.36127	-18.6	-18.9220913	-0.017316734	-1.731673416
823	AM263198.1/671271-671368	98	21.82052	-18.6	-18.0615909	0.028946726	2.894672589
824	AL591976.1/215482-215385	98	28.12601	-31.4	-28.095512	0.105238472	10.52384722
825	AL591973.1/172171-172268	98	28.12601	-31.4	-28.095512	0.105238472	10.52384722
826	AL591974.1/157606-157509	98	28.12601	-31.4	-28.095512	0.105238472	10.52384722
VIII. Rfam snRNA family RF01475							
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_M**	**MFE_C**	**RD1**	**% RD1**
827	AADR01000003.1/142171-142094	78	21.34496	-14.4	-13.3128299	0.075497922	7.54979218
828	AM263198.1/2127523-2127600	78	21.36664	-14.2	-13.3473388	0.060046561	6.004656108
829	AL596171.1/97397-97474	78	21.53941	-14	-13.6222656	0.026981029	2.698102865
830	AL591982.1/53775-53852	78	21.77387	-16.1	-13.9953662	0.130722598	13.07225979
831	AADQ01000011.1/3685-3762	78	21.77387	-16.1	-13.9953662	0.130722598	13.07225979

832	AARL02000916.1/597-673	77	21.25997	-14.4	-12.9779941	0.098750412	9.875041172

IX. Rfam snRNA family RF01490

Sl. No.	Sequence ID	NTL	SD_DFT	MFE_M	MFE_C	RD1	% RD1
833	AY168080.1/216-96	121	27.79179	-34	-32.154468	0.054280352	5.42803518
834	AY510072.1/4036-4154	119	20.77558	-23.8	-20.5903864	0.134857716	13.48577156
835	AY510073.1/4042-4162	121	25.31298	-29	-28.2099497	0.027243115	2.724311485
836	AY512490.1/3938-4058	121	27.70093	-33.2	-32.009894	0.035846565	3.584656508
837	AY512446.2/3933-4053	121	27.24212	-32.9	-31.2797923	0.049246434	4.924643436
838	EU372052.1/3947-4067	121	27.79179	-34	-32.154468	0.054280352	5.42803518
839	EU372053.1/3947-4067	121	27.79179	-34	-32.154468	0.054280352	5.42803518
840	FJ041145.1/3922-4042	121	27.70093	-32.2	-32.009894	0.005903912	0.590391182
841	EU372028.1/3954-4074	121	28.20594	-33.4	-32.8135109	0.017559555	1.755955463

X. Rfam snRNA family RF00618

Sl. No.	Sequence ID	NTL	SD_DFT	MFE_M	MFE_C	RD1	% RD1
842	AAPP01019634.1/122532-122382	151	39.30442	-60	-56.4625259	0.058957902	5.895790155

843	AAPQ01006438.1/743481-743332	150	34.37847	-54.6	-48.4242592	0.113108806	11.31088059
844	AE014297.2/1020885-1020736	150	34.33452	-51	-48.354317	0.051876138	5.187613791
845	AAST01029695.1/1294-1145	150	34.15814	-54.3	-48.0736481	0.114665781	11.46657814
846	AAKO01001557.1/51647-51498	150	34.15814	-54.3	-48.0736481	0.114665781	11.46657814
847	AAEU02010290.1/66-215	150	34.52458	-49.8	-48.6567569	0.022956689	2.295668907
848	AADA01241850.1/15965-15840	126	26.10302	-32.8	-30.4651407	0.071184736	7.118473581
849	AL389925.10/20736-20611	126	28.01919	-31.5	-33.5143432	-0.063947403	-6.394740284
850	AADA01047294.1/887-1012	126	30.92882	-36.5	-38.1444337	-0.045052977	-4.505297723
851	AL135914.25/92223-92098	126	30.92882	-36.5	-38.1444337	-0.045052977	-4.505297723
852	AL161445.10/77816-77941	126	30.13649	-34.1	-36.8835958	-0.081630374	-8.163037412
853	AADA01054074.1/15964-15839	126	30.13649	-34.2	-36.8835958	-0.078467712	-7.846771221
854	AC136636.6/172138-172014	125	31.02743	-35.6	-38.1017545	-0.070274003	-7.027400282
855	AAXN01018884.1/118-242	125	30.73364	-35.6	-37.6342368	-0.057141482	-5.714148243
856	AACN010332835.1/830-706	125	30.91351	-37.9	-37.92047	-0.000540106	-0.054010626
857	AAFC03121196.1/26240-26365	126	30.41947	-42.3	-37.3338986	0.117401925	11.74019254
858	AAFR03008173.1/71752-71627	126	30.01919	-33.1	-36.6969432	-0.108668979	-10.86689785
859	AAPN01427475.1/38-163	126	35.41388	-48.2	-45.2815109	0.060549567	6.054956692
860	BAAF04053164.1/10519-10646	128	30.28474	-38	-37.5187035	0.012665696	1.266569625
861	AANH01011402.1/5162-5288	127	38.6367	-50.2	-50.6095849	-0.008159061	-0.81590612
862	ABAV01030669.1/9778-9657	122	26.32387	-31.6	-30.0181769	0.050057693	5.005769271

863	AACT01014164.1/14797-14663	135	32.92461	-41.7	-43.1167296	-0.033974332	-3.397433208
864	AABS01000098.1/66915-66782	134	31.19717	-37.7	-40.16825	-0.065470823	-6.547082329
865	AAZX01007551.1/46173-46008	166	34.27942	-47	-51.4602446	-0.094898821	-9.489882089
866	BAAB01203970.1/2470-2315	156	34.72361	-50.8	-50.1710792	0.01238033	1.238032998
867	AAAB01008986.1/3372038-3372230	193	39.15281	-58.5	-64.6044702	-0.104349917	-10.43499172
868	AAFS01000016.1/19569-19724	156	34.82491	-57.5	-50.3322756	0.124656077	12.46560772
869	AAIZ01001811.1/683-838	156	34.99788	-54.4	-50.6075266	0.069714584	6.971458391
870	AANI01017162.1/86143-85990	154	21.59058	-29.1	-28.8732836	0.007790941	0.779094085
871	AANI01014648.1/138479-138633	155	34.03636	-55.9	-48.8778649	0.12561959	12.56195903
872	AAPU01011105.1/262422-262573	152	33.77134	-53.9	-47.8573311	0.112108885	11.21088845
873	AAPT01020183.1/127226-127384	159	34.77941	-58.9	-50.8586805	0.136524949	13.65249489
874	ABDC01347327.1/313-438	126	29.74935	-39.2	-36.2675453	0.074807517	7.480751724
875	ABDC01347327.1/313-438	126	30.26999	-38.6	-37.0960273	0.038963022	3.896302219
876	AANU01295318.1/748-623	126	30.01919	-41.7	-36.6969432	0.119977382	11.99773816
877	AANN01562320.1/870-745	126	30.26999	-37.8	-37.0960273	0.018623615	1.862361525
878	AAIY01042223.1/294-419	126	30.60118	-38.5	-37.6230521	0.022777868	2.277786801
879	AAPY01611414.1/511-386	126	30.60118	-40	-37.6230521	0.059423698	5.942369796
880	CAAE01014614.1/129416-129543	128	41.53103	-56.5	-55.4149232	0.019204898	1.920489846
881	AAVX01595596.1/659-533	127	31.42893	-44.1	-39.1398598	0.112474835	11.24748351
882	AANG01209374.1/1-113	113	28.37472	-34.9	-31.4852994	0.097842425	9.784242508

883	AAWU01017057.1/6040-5883	158	34.86681	-54.3	-50.7981508	0.064490777	6.449077667
884	AAJJ01000001.1/47747-47867	121	29.99055	-33.5	-35.6533687	-0.064279662	-6.427966236
885	DQ682679.1/205-355	151	34.64025	-52.9	-49.040424	0.072959849	7.295984868
886	AAGE02008333.1/19499-19656	158	35.22575	-52.6	-51.3693369	0.023396637	2.339663724
887	AAZO01007334.1/46791-46906	116	29.43586	-30.3	-33.7726762	-0.114609776	-11.4609776
888	AAGV020469602.1/2323-2199	125	30.20428	-35.1	-36.791867	-0.048201339	-4.820133878
889	AASC02027737.1/1238-1108	131	38.06579	-55.8	-50.4994841	0.094991325	9.499132509
890	AAKN02006802.1/64876-65001	126	29.81704	-36	-36.3752591	-0.010423864	-1.042386402
891	CAAK05033158.1/3332-3460	129	31.427	-40.2	-39.5359831	0.016517833	1.651783334
892	AAWR02006087.1/50015-49891	125	29.90239	-39.9	-36.3114811	0.089937816	8.993781607
893	AAWZ02013490.1/81111-81235	125	30.76642	-43.8	-37.6864037	0.139579825	13.95798252
894	AAGW02073287.1/24794-24920	127	31.17131	-39.2	-38.7299114	0.011992056	1.19920565
895	AADG06006595.1/14072-13949	124	24.94711	-28.8	-28.2265351	0.019911975	1.991197529
896	CAAB02003742.1/18728-18856	129	30.66413	-44.1	-38.3220351	0.131019612	13.10196123
897	EU240318.1/2-119	118	38.50254	-49.6	-48.5996977	0.020167385	2.016738488
898	AAQR03093718.1/17057-17182	126	30.46913	-39.1	-37.4129296	0.04314758	4.314757991
899	AAPE02048822.1/2601-2476	126	31.30137	-39.6	-38.7372761	0.021785957	2.17859575
900	AAGJ04111208.1/10003-10134	132	28.66141	-39.7	-35.7339037	0.099901671	9.990167106
901	FJ916040.1/3-128	126	30.01919	-41.7	-36.6969432	0.119977382	11.99773816
902	U62822.1/2-128	127	30.78084	-43.6	-38.108556	0.12595055	12.59505499

Table 6.8. Comparison of MFE computed with the model with MFE from webservers, RNAfold and RNAstructure for 902 snRNA sequences

I. Rfam snRNA family RF00004 (208)										
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_C**	**MFE_F**	**RD_2**	**%RD_2**	**MFE_S**	**RD_3**	**%RD_3**
1	AALT01209640.1/567-377	191	37.91472	-62.235101	-56.2	-0.10739	-10.7386134	-58.1	-0.07117	-7.117212937
2	AAFR03033875.1/20528-20718	191	37.67534	-61.854176	-60.1	-0.02919	-2.91876142	-61.8	-0.00088	-0.087662804
3	AAIY01044029.1/787-597	191	37.80852	-62.066098	-60.3	-0.02929	-2.9288531	-61.3	-0.0125	-1.249752727
4	AAZO01007389.1/15370-15178	193	38.60989	-63.740525	-71.2	0.104768	10.47679042	-72.6	0.122031	12.2031333
5	AAYZ01695118.1/310-500	191	38.25785	-62.781118	-61.3	-0.02416	-2.41617975	-61.6	-0.01917	-1.917399655
6	AAHX01044404.1/26102-26292	191	37.78192	-62.023774	-61.4	-0.01016	-1.01591795	-62.8	0.01236	1.236029268
7	AACN010750078.1/657-848	192	38.30984	-63.063451	-62.2	-0.01388	-1.38818531	-62.7	-0.0058	-0.579667086
8	ABAV01019481.1/5988-6180	193	38.57082	-63.67835	-68.9	0.075786	7.578592837	-70.5	0.096761	9.67609995
9	AAZX01018356.1/721-913	193	45.57082	-74.81745	-78.1	0.04203	4.203009558	-78.9	0.051743	5.17433519
10	AY765362.1/650-458	193	38.83056	-64.091672	-67.3	0.047672	4.767202996	-67.3	0.047672	4.767202996
11	BX927129.10/97355-97165	191	43.04706	-70.40219	-72.1	0.023548	2.354798627	-72.7	0.031607	3.160673741
12	AAFC03011281.1/26348-26158	191	38.44134	-63.073109	-65.3	0.034102	3.410246679	-66.9	0.057203	5.720315518

13	AAVX01416582.1/429-619	191	43.34971	-70.883788	-71.5	0.008618	0.8618347	-72.1	0.016868	1.686840237
14	X00093.1/360-550	191	38.07347	-62.48772	-66.6	0.061746	6.174594223	-67.2	0.070123	7.012321061
15	CAAE01009132.1/1078-888	191	38.09987	-62.529721	-58.1	-0.07624	-7.62430494	-62.3	-0.00369	-0.368733816
16	AF095839.1/1586-1389	198	45.21437	-75.248221	-76	0.009892	0.989183489	-76.8	0.020205	2.020546161
17	AANH01015084.1/457-647	191	37.80852	-62.066098	-57.2	-0.08507	-8.50716507	-58.4	-0.06278	-6.277565791
18	AC004138.3/33098-33293	196	45.56928	-75.413803	-75.4	-0.00018	-0.01830661	-76.4	0.012908	1.29083353
19	AAJJ01003841.1/8097-7907	191	45.28412	-73.962017	-73.9	-0.00084	-0.08391993	-75.1	0.015153	1.515290504
20	AACY020405974.1/944-1135	192	38.36229	-63.146904	-74.7	0.15466	15.4659915	-75.5	0.163617	16.36171609
21	AM465080.2/15550-15355	196	38.76424	-64.584943	-72.3	0.106709	10.67089546	-73.7	0.123678	12.36778483
22	M72891.1/1-196	196	37.93868	-63.271225	-74.2	0.147288	14.72880779	-74.3	0.148436	14.84357386
23	BAAB01070452.1/1509-1701	193	38.89523	-64.194572	-72.1	0.109645	10.96453283	-73	0.120622	12.06223038
24	X04243.1/69-264	196	38.79017	-64.62619	-68.1	0.05101	5.101042551	-69.2	0.066096	6.60955199
25	AAPY01817437.1/27757-27567	191	38.89629	-63.79706	-57.1	-0.11729	-11.7286509	-58.5	-0.09055	-9.05480285
26	AANG01476605.1/2311-2501	191	34.3366	-56.541228	-55.1	-0.02616	-2.61565917	-56	-0.00966	-0.966478934
27	AANN01265286.1/964-773	192	38.89575	-63.995811	-66.6	0.039102	3.910193192	-67.2	0.047681	4.768137896
28	AASG02001826.1/33924-33729	196	38.85489	-64.729188	-68.3	0.052281	5.228128488	-70.7	0.084453	8.445278299

29	AAEU02000279.1/6988-6794	195	39.26713	-65.185578	-63.8	-0.02172	-2.17175275	-65.2	0.000221	0.022119239
30	AC189506.1/7126-6931	196	38.00486	-63.376531	-68.2	0.070725	7.072535897	-68.6	0.076144	7.614387
31	X69327.1/1-196	196	41.72469	-69.295905	-68.2	-0.01607	-1.60689873	-85.5	0.189522	18.95215797
32	AAAA02007579.1/14294-14490	197	47.12697	-78.092142	-79.8	0.021402	2.140172476	-79.8	0.021402	2.140172476
33	AC149482.1/90998-90803	196	43.91951	-72.788515	-70.4	-0.03393	-3.39277723	-72.1	-0.00955	-0.954944756
34	AAPU01010615.1/193731-193537	195	38.92002	-64.63323	-62.2	-0.03912	-3.91194462	-64	-0.00989	-0.989421175
35	AADA01287270.1/15999-15809	191	38.06027	-62.466709	-66.7	0.063468	6.346763321	-67.5	0.074567	7.456727608
36	AAPT01020503.1/50330-50135	196	38.45183	-64.087805	-64.5	0.006391	0.639062766	-64.6	0.007929	0.792872267
37	AAAB01008933.1/737524-737331	194	43.46592	-71.667517	-71	-0.0094	-0.94016519	-70.4	-0.018	-1.800450692
38	ABDC01189504.1/6312-6122	191	35.1921	-57.902596	-56.6	-0.02301	-2.30140602	-57.1	-0.01406	-1.40559686
39	AAWU01010867.1/21259-21065	195	38.19004	-63.471615	-66.4	0.044102	4.410217823	-68.4	0.072052	7.205240694
40	AANI01016115.1/56636-56831	196	38.66039	-64.419676	-62.9	-0.02416	-2.41601949	-63.4	-0.01608	-1.60832218
41	AC157776.1/115175-114979	197	39.6482	-66.191177	-67.7	0.022287	2.228689483	-68.1	0.02803	2.802970308
42	AAPP01015704.1/576899-577092	194	38.15105	-63.209965	-63.4	0.002997	0.29973922	-65.4	0.033487	3.348676858
43	AC151964.12/72254-72449	196	38.64739	-64.398987	-62.2	-0.03535	-3.53534847	-63.3	-0.01736	-1.736155996

44	BAAE01249332.1/245-435	191	38.25785	-62.781118	-67.5	0.069909	6.990936019	-67.8	0.074025	7.40248055
45	AAGE02006086.1/44202-44008	195	38.37384	-63.764091	-67.5	0.055347	5.534680657	-68.1	0.06367	6.36697422
46	AP009284.1/2157-1962	196	38.95823	-64.893629	-71.4	0.091126	9.112563783	-71.7	0.094928	9.492845943
47	AAQB01006449.1/663346-663151	196	35.00979	-58.610487	-57.3	-0.02287	-2.28706233	-57.7	-0.01578	-1.577966576
48	AB202073.1/636-444	193	34.79171	-57.66465	-55.4	-0.04088	-4.08781658	-58.5	0.014279	1.42794806
49	AACT01003467.1/10053-10247	195	39.12608	-64.961139	-62.4	-0.04104	-4.10438913	-63.2	-0.02787	-2.786612053
50	AF106845.1/1870-2065	196	39.61067	-65.931858	-61.2	-0.07732	-7.7317944	-62.3	-0.0583	-5.829627887
51	ABBA01028418.1/3128-3319	192	27.99366	-46.647314	-44.1	-0.05776	-5.77622169	-44.9	-0.03892	-3.891567402
52	X15930.1/1-195	195	38.69988	-64.282913	-59	-0.08954	-8.95408963	-59.1	-0.0877	-8.769734151
53	AASR01035668.1/1184-988	197	39.06078	-65.256427	-61.2	-0.06628	-6.62814808	-61.8	-0.05593	-5.592923339
54	D25323.1/7242-7047	196	39.96434	-66.494659	-72.6	0.084096	8.409560405	-72.3	0.080295	8.029517087
55	AADE01000447.1/58660-58464	197	39.31726	-65.664556	-62.3	-0.05401	-5.40057133	-63.7	-0.03084	-3.084075254
56	X55772.1/223-413	191	39.03817	-64.022845	-57.9	-0.10575	-10.5748612	-58.6	-0.09254	-9.254001053
57	AP007151.1/1974779-1974972	194	38.15105	-63.209965	-67.7	0.066323	6.632252091	-66.2	0.045167	4.516668679
58	X54113.1/230-419	190	38.97428	-63.721576	-67.5	0.055977	5.597665186	-67.8	0.060154	6.015374632
59	AATM01000136.1/58626-58436	191	39.1282	-64.1661	-67.5	0.049391	4.939111611	-66.4	0.033643	3.364307737

60	AAHF01000004.1/976953-976761	193	37.8739	-62.569337	-58.7	-0.06592	-6.59171528	-60.5	-0.0342	-3.42039152
61	CAAA01181619.1/3682-3492	191	31.07347	-51.34862	-49.9	-0.02903	-2.90304659	-51.4	0.001	0.099960608
62	X05084.1/1-193	193	38.54475	-63.636864	-64.8	0.01795	1.794962867	-65.4	0.026959	2.69592651
63	ABAR01000024.1/421358-421550	193	38.09883	-62.927268	-57.7	-0.09059	-9.05938975	-58.7	-0.07201	-7.201478511
64	AAQA01000616.1/21523-21715	193	38.53171	-63.616111	-62.9	-0.01138	-1.13849094	-65	0.021291	2.129060307
65	AY661656.1/2159-2358	200	39.07215	-65.873318	-55.3	-0.1912	-19.1199233	-56.6	-0.16384	-16.38395335
66	ABAS01000032.1/96356-96164	193	37.94019	-62.674831	-61.7	-0.0158	-1.57995296	-61	-0.02746	-2.745624551
67	AAJN01000116.1/21716-21908	193	36.20422	-59.912377	-59.4	-0.00863	-0.86258726	-56.6	-0.05852	-5.852255885
68	ABDB01000030.1/249236-249428	193	37.86063	-62.548216	-63.3	0.011877	1.187652636	-62.2	-0.0056	-0.559832607
69	CAAL01000198.1/59590-59777	188	38.04866	-61.849433	-64.7	0.044058	4.40582256	-64.4	0.039605	3.960508069
70	AACD01000084.1/492354-492546	193	38.20422	-63.094977	-64.7	0.024807	2.480715872	-64.4	0.020264	2.026433492
71	AACM02000140.1/27654-27846	193	38.30932	-63.262224	-57.9	-0.09261	-9.26118204	-58.4	-0.08326	-8.325726714
72	AAKD03000004.1/610702-610510	193	37.75428	-62.378981	-57.7	-0.08109	-8.10915194	-60.2	-0.0362	-3.619569221
73	AANU01167734.1/2229-2419	191	37.11079	-60.955796	-57.9	-0.05278	-5.27771272	-47.7	-0.2779	-27.78992802
74	AAEE01000010.1/89978-89784	195	38.42619	-63.847398	-53.5	-0.19341	-19.3409307	-54.3	-0.17583	-17.58268499

75	AARE01006627.1/693-886	194	37.54017	-62.23787	-55.5	-0.1214	-12.1403065	-53.8	-0.15684	-15.68377339
76	AACW02000228.1/581455-581260	196	38.15004	-63.607561	-58.9	-0.07992	-7.99246266	-60.2	-0.0566	-5.660399516
77	AM270020.1/5920-6112	193	37.83407	-62.505952	-57.7	-0.08329	-8.32920568	-56.2	-0.11221	-11.22055459
78	AAIM02000113.1/450420-450612	193	38.38796	-63.38736	-53.3	-0.18926	-18.9256277	-52.7	-0.2028	-20.2796197
79	ABBB01000093.1/1472-1279	194	37.36572	-61.960267	-56.8	-0.09085	-9.08497756	-56.7	-0.09277	-9.277367295
80	AAFU01000671.1/40760-40955	196	34.07092	-57.116453	-54.5	-0.04801	-4.80083187	-55.1	-0.0366	-3.659624989
81	CR382129.1/446178-446370	193	38.59687	-63.719807	-59.5	-0.07092	-7.09211258	-59.8	-0.06555	-6.554861181
82	AAPN01113121.1/7070-7251	182	37.21032	-59.317778	-57.2	-0.03702	-3.70240974	-57.2	-0.03702	-3.702409744
83	AAQQ01759780.1/669-855	187	33.08578	-53.752397	-50.2	-0.07076	-7.0764872	-51	-0.05397	-5.396856031
84	AAQX01002532.1/7801-7990	190	43.32402	-70.643306	-70.7	0.000802	0.08018988	-71.6	0.013362	1.336165147
85	X63786.1/549-739	191	33.36281	-54.991641	-52	-0.05753	-5.75315586	-59.8	0.080407	8.040734034
86	AAIW01000278.1/13786-13979	194	30.54017	-51.09877	-51.4	0.005861	0.586050418	-51.7	0.011629	1.162920531
87	AAIW01000278.1/13786-13979	194	30.67381	-51.311436	-46.2	-0.11064	-11.0637143	-47.8	-0.07346	-7.346100445
88	AASM01001106.1/7439-7242	198	30.9377	-52.529761	-50.5	-0.04019	-4.01932881	-54.2	0.030816	3.081621678
89	AAWC01001022.1/39371-39181	191	32.44786	-53.535677	-49.4	-0.08372	-8.37181622	-50.2	-0.06645	-6.644775319

90	AC187487.2/140157-140345	189	26.64721	-43.905901	-41.4	-0.06053	-6.05290177	-43	-0.02107	-2.106747286
91	CAAI01005114.1/671-870	200	33.16126	-56.467306	-52.5	-0.07557	-7.55677253	-54.3	-0.03991	-3.991354657
92	AAGK01000002.1/903654-903460	195	32.99742	-55.208599	-57	0.031428	3.142809516	-56.5	0.022857	2.285666237
93	AAXI01000029.1/108904-109093	190	31.82234	-52.340683	-51.6	-0.01435	-1.43543271	-53.5	0.021669	2.166947145
94	DQ114948.1/136-329	194	38.08512	-63.105058	-59.4	-0.06237	-6.23747157	-59.5	-0.06059	-6.058921194
95	AAKM01000005.1/1407395-1407197	199	38.56779	-64.871131	-57.1	-0.1361	-13.6096863	-58.7	-0.10513	-10.51299978
96	AAXT01000001.1/1056302-1056495	194	38.83004	-64.290443	-66.2	0.028845	2.884527273	-66.3	0.03031	3.031006115
97	AAJI01001427.1/51501-51693	193	34.03281	-56.457014	-53	-0.06523	-6.52266755	-53.6	-0.0533	-5.330249625
98	AABS01001062.1/345-536	192	34.84403	-57.548304	-54.1	-0.06374	-6.37394523	-54.5	-0.05593	-5.593219023
99	X71483.1/1-192	192	38.79224	-63.831088	-70.7	0.097156	9.715576155	-70.8	0.098431	9.843096528
100	AAXJ01017415.1/259-455	197	39.33004	-65.684893	-72.2	0.090237	9.023694544	-72.5	0.094001	9.400148222
101	AF325695.1/199-9	191	38.88336	-63.776493	-65	0.018823	1.882318603	-66.3	0.038062	3.806194709
102	AAFT01000058.1/3256-3060	197	30.05722	-50.929055	-48.5	-0.05008	-5.00836116	-48.4	-0.05225	-5.225320583
103	AP004918.1/55019-55207	189	38.54688	-62.841845	-60.9	-0.03189	-3.18857942	-61.6	-0.02016	-2.015981929
104	AAID01003241.1/2743-2932	190	37.82234	-61.888483	-53.7	-0.15249	-15.2485722	-52.2	-0.1856	-18.56031279
105	AATT01000021.1/233305-233109	197	39.6482	-66.191177	-66	-0.0029	-0.28966246	-67.4	0.017935	1.793505608

106	CP000498.1/1665432-1665627	196	34.28155	-57.451627	-49.3	-0.16535	-16.5347413	-51	-0.1265	-12.65024991
107	AAFM01000022.1/42807-42996	190	37.70254	-61.697851	-53.1	-0.16192	-16.1918099	-54	-0.14255	-14.2552797
108	AAGT01000476.1/108951-108760	192	38.15208	-62.812404	-57.5	-0.09239	-9.238963	-57	-0.10197	-10.19719952
109	DQ235686.1/7795-7608	188	38.48216	-62.539269	-57.4	-0.08953	-8.95342996	-56.9	-0.09911	-9.910841472
110	AATU01001299.1/7335-7151	185	45.5882	-73.248297	-77.1	0.049957	4.995723513	-78.8	0.070453	7.045308158
111	AAGI01000215.1/38061-38253	193	28.13839	-47.077213	-45.5	-0.03466	-3.46640231	-46.9	-0.00378	-0.377852986
112	AAGD02001363.1/26266-26450	185	37.74521	-60.767757	-59.2	-0.02648	-2.64823831	-61.3	0.008683	0.868259251
113	AADS01000047.1/234093-233905	189	38.63805	-62.98693	-62.1	-0.01428	-1.42822794	-65.9	0.044204	4.420440745
114	AANW02001910.1/4438-4250	189	37.90252	-61.816481	-60.9	-0.01505	-1.50489482	-61.9	0.001349	0.134925773
115	DQ158857.1/121770-121585	186	30.37979	-49.246764	-44.5	-0.10667	-10.6668856	-45.5	-0.08235	-8.234646338
116	CP000599.1/149439-149632	194	33.75377	-56.212574	-53.3	-0.05464	-5.46449198	-55.9	-0.00559	-0.559166768
117	AAWT01070971.1/3028-3217	190	33.32738	-54.735661	-49.3	-0.11026	-11.0256806	-51.1	-0.07115	-7.114795605
118	AACP01000091.1/4899-4707	193	38.44029	-63.470641	-60.7	-0.04564	-4.5644824	-60.5	-0.0491	-4.910150114
119	AAFP01000557.1/11449-11264	186	38.48327	-62.141829	-57.7	-0.07698	-7.69814464	-57.6	-0.07885	-7.885120588
120	AAFI02000140.1/17830-17618	213	33.58744	-59.740294	-53.1	-0.12505	-12.505262	-54.4	-0.09817	-9.8167171

121	AAFB02000004.1/138313-138494	182	32.89825	-52.455981	-46.3	-0.13296	-13.2958556	-48	-0.09283	-9.283294013
122	AAPO01000010.1/72363-72562	200	32.37137	-55.210356	-48.5	-0.13836	-13.8357864	-49.4	-0.11762	-11.76185502
123	AAFX01115267.1/519-717	199	40.82124	-68.457041	-70.5	0.028978	2.897813894	-71.5	0.042559	4.255886427
124	AANV02000585.1/5693-5873	181	36.34969	-57.748663	-50.5	-0.14354	-14.3537871	-50.7	-0.13903	-13.90268739
125	DQ012953.1/31-235	205	33.0138	-57.230661	-50.3	-0.13779	-13.7786492	-52	-0.10059	-10.0589626
126	AAFO01000053.1/189105-188894	212	30.30711	-54.320712	-51.8	-0.04866	-4.8662391	-53.5	-0.01534	-1.534040846
127	AABY01000227.1/3654-3483	172	30.21662	-46.192712	-41.2	-0.12118	-12.1182329	-45.6	-0.013	-1.299806956
128	Z36100.1/1808-1619	190	30.44159	-50.143501	-45.7	-0.09723	-9.72319644	-45.7	-0.09723	-9.723196445
129	AC167922.2/17996-18182	187	37.58396	-60.910349	-55.9	-0.08963	-8.96305813	-55.1	-0.10545	-10.54509891
130	AAZN01000309.1/86876-87072	197	27.76374	-47.27944	-44.5	-0.06246	-6.24593194	-45.5	-0.03911	-3.910856512
131	AADM01000307.1/26094-25895	200	33.58616	-57.143452	-55.2	-0.03521	-3.52074717	-56.2	-0.01679	-1.678740988
132	AL590446.1/168546-168725	180	38.60414	-61.136562	-61	-0.00224	-0.22387195	-61.5	0.00591	0.590956275
133	AF053589.1/90-279	190	47.07732	-76.615941	-82.2	0.067933	6.793258729	-81.6	0.061079	6.107915043
134	Z50072.1/229-412	184	24.41713	-39.359185	-39	-0.00921	-0.92098812	-39.8	0.011076	1.107574455
135	AAHC01001365.1/13705-13888	184	37.18217	-59.672185	-62.4	0.043715	4.371498589	-63.1	0.054324	5.432353597
136	AF287991.1/4898-5088	191	37.90147	-62.214001	-63.5	0.020252	2.02519476	-64.2	0.030935	3.093455876

137	M33777.1/191-381	191	37.90147	-62.214001	-63.5	0.020252	2.02519476	-64.2	0.030935	3.093455876
138	S64581.1/735-926	192	38.90867	-64.016371	-65.8	0.027107	2.710682402	-66.2	0.032985	3.298533264
139	AAKO01002676.1/16743-16938	196	38.94533	-64.873098	-64.4	-0.00735	-0.73462449	-65	0.001952	0.195233587
140	AAPQ01007349.1/370124-370319	196	38.94533	-64.873098	-64.4	-0.00735	-0.73462449	-65	0.001952	0.195233587
141	AAIZ01004041.1/16829-17024	196	38.91951	-64.832015	-60.7	-0.06807	-6.80727375	-61.9	-0.04737	-4.736696557
142	AAYL01000061.1/287788-287596	193	38.99846	-64.358856	-67.8	0.050754	5.075433557	-68.9	0.065909	6.590920104
143	AL683874.1/16263-16071	193	37.8739	-62.569337	-58.7	-0.06592	-6.59171528	-60.5	-0.0342	-3.42039152
144	AAIH02000488.1/9382-9189	194	38.15105	-63.209965	-67.7	0.066323	6.632252091	-66.2	0.045167	4.516668679
145	AAKE03000002.1/1730972-1731164	193	37.83407	-62.505952	-58.3	-0.07214	-7.21432535	-59	-0.05942	-5.942290979
146	AAEL01000160.1/10170-10364	195	32.62188	-54.610995	-52.7	-0.03626	-3.6261764	-53.5	-0.02077	-2.076626097
147	AATX01000107.1/79359-79166	194	37.16341	-61.638337	-55.1	-0.11866	-11.8663105	-55	-0.1207	-12.06970378
148	AANS01001054.1/6664-6467	198	32.9377	-55.712361	-53.1	-0.0492	-4.91970066	-54.2	-0.0279	-2.790334041
149	AABL01000318.1/11806-12005	200	31.16126	-53.284706	-52.5	-0.01495	-1.49467729	-54.3	0.018698	1.869787147
150	CAAJ01003844.1/3754-3953	200	31.16126	-53.284706	-52.5	-0.01495	-1.49467729	-54.3	0.018698	1.869787147
151	EF140768.1/2-192	191	30.51491	-50.459775	-51.4	0.018292	1.829231136	-52	0.02962	2.961970777
152	AB179181.1/1-163	163	28.95694	-42.391786	-41.8	-0.01416	-1.41575514	-43.1	0.016432	1.643188749

153	AC198944.2/184805-184614	192	25.45438	-42.606547	-40.7	-0.04684	-4.68439101	-41.1	-0.03666	-3.66556482
154	AACQ01000018.1/49571-49783	213	30.41868	-54.697848	-51.8	-0.05594	-5.5943008	-53.5	-0.02239	-2.238967877
155	AE017345.1/926592-926777	186	38.48327	-62.141829	-57.7	-0.07698	-7.69814464	-57.6	-0.07885	-7.885120588
156	AAEY01000026.1/44265-44450	186	38.48327	-62.141829	-57.7	-0.07698	-7.69814464	-57.6	-0.07885	-7.885120588
157	AACO02000044.1/38107-37922	186	38.48327	-62.141829	-57.7	-0.07698	-7.69814464	-57.6	-0.07885	-7.885120588
158	AACI02000565.1/65-236	172	28.85777	-44.030374	-42.3	-0.04091	-4.09071849	-43.6	-0.00987	-0.987096154
159	AACF01000175.1/12372-12544	173	27.45797	-42.002463	-38.6	-0.08815	-8.81467174	-40.8	-0.02947	-2.947213953
160	AAFW02000011.1/661534-661345	190	28.44159	-46.960901	-45.7	-0.02759	-2.75908266	-45.7	-0.02759	-2.759082659
161	AAEG01000106.1/129944-130133	190	29.44159	-48.552201	-45.7	-0.06241	-6.24113955	-45.7	-0.06241	-6.241139552
162	AY007788.1/537-683	147	32.60725	-45.00692	-49.6	0.092602	9.260242027	-52.2	0.137798	13.77984683
163	M58665.1/571-739	169	32.81627	-49.73073	-49.9	0.003392	0.339219258	-52.1	0.045475	4.547543973
164	U23406.1/206-352	147	33.06718	-45.738803	-47.4	0.035046	3.504635772	-47.9	0.045119	4.511894271
165	EF052257.1/89-253	165	21.11825	-30.317269	-30.3	-0.00057	-0.05699266	-30.8	0.015673	1.567309169
166	AACA01000784.1/509-337	173	26.51465	-40.50137	-39.8	-0.01762	-1.7622362	-41	0.012162	1.216170716
167	AABZ01000169.1/11839-11668	172	22.64068	-34.137112	-33.7	-0.01297	-1.29706744	-33.6	-0.01599	-1.598546803
168	X56454.1/125-277	153	32.60627	-46.202962	-45.5	-0.01545	-1.54497161	-45.5	-0.01545	-1.54497161

169	AC008368.21/83891-84039	149	34.15891	-47.875272	-48.5	0.012881	1.288098375	-52.7	0.091551	9.155081048
170	M58666.1/571-718	148	37.248	-52.591335	-57.1	0.078961	7.896086441	-57.5	0.085368	8.536809318
171	AAHK01000589.1/17217-17069	149	33.54928	-46.905166	-44	-0.06603	-6.60265017	-45.4	-0.03315	-3.315343778
172	AF326335.1/1-142	142	32.20722	-43.372348	-46.1	0.059168	5.916816024	-46.7	0.071256	7.125593549
173	CAAC02000548.1/434719-434909	191	38.88336	-63.776493	-65.6	0.027797	2.779736421	-66.4	0.039511	3.951064898
174	AACE03000009.1/583940-584128	189	32.8327	-53.748878	-52.8	-0.01797	-1.79711791	-53.3	-0.00842	-0.842173092
175	AABX02000002.1/241678-241484	195	38.72584	-64.32423	-62.7	-0.0259	-2.59047842	-63.3	-0.01618	-1.618056827
176	AASC02023314.1/7214-7405	192	38.64945	-63.603872	-68.6	0.07283	7.282985068	-69	0.078205	7.82047501
177	AAFD02000010.1/1244775-1244583	193	32.04199	-53.289026	-50.5	-0.05523	-5.52282425	-52.5	-0.01503	-1.502907132
178	AAZY02000001.1/986053-985858	196	37.64613	-62.80568	-62.1	-0.01136	-1.13636005	-62.8	-9E-05	-0.00904394
179	ABFM01000169.1/24531-24724	194	37.36572	-61.960267	-56.8	-0.09085	-9.08497756	-56.7	-0.09277	-9.277367295
180	AAQM02000124.1/160019-159827	193	39.12713	-64.563602	-66.9	0.034924	3.492374054	-68.4	0.056088	5.608769359
181	CR382136.2/900319-900516	198	28.60736	-48.821485	-45.2	-0.08012	-8.01213544	-46.7	-0.04543	-4.542794906
182	AAGV020390824.1/478-668	191	37.94123	-62.277277	-67.6	0.078739	7.873850039	-68.2	0.086843	8.684344027
183	X58842.1/1-191	191	38.74091	-63.549806	-63.3	-0.00395	-0.39463777	-63.3	-0.00395	-0.394637774
184	AAKN02019678.1/9356-9546	191	38.00741	-62.38259	-60.7	-0.02772	-2.7719775	-61.1	-0.02099	-2.099165858

185	AAWR02015112.1/62289-62483	195	37.33831	-62.116255	-59.4	-0.04573	-4.57281926	-60.2	-0.03183	-3.183147242
186	ABDF02000003.1/1932068-1932260	193	38.80467	-64.050465	-58.5	-0.09488	-9.48797367	-59.5	-0.07648	-7.647839655
187	M12856.1/361-551	191	39.2052	-64.288628	-70.2	0.084208	8.420757994	-72.3	0.110807	11.08073598
188	AACS02000012.1/1565410-1565598	189	38.58598	-62.904066	-58.1	-0.08269	-8.26861617	-57.4	-0.09589	-9.588965147
189	ABEG02004067.1/53561-53371	191	38.23157	-62.739291	-66.5	0.056552	5.655201767	-68	0.077363	7.736337022
190	AAIL02000026.1/644385-644193	193	38.53171	-63.616111	-61.4	-0.03609	-3.60930098	-62.9	-0.01138	-1.138490939
191	AAQY02000293.1/11238-11050	189	39.07787	-63.686812	-65.7	0.030642	3.064213468	-66.1	0.036508	3.650814294
192	AADG06003467.1/1334-1527	194	38.51815	-63.794128	-61.4	-0.03899	-3.89923084	-64.6	0.012475	1.247480285
193	AAGJ04020931.1/13670-13861	192	38.58437	-63.500315	-61.7	-0.02918	-2.91785178	-62.9	-0.00954	-0.954395146
194	AAGU03035529.1/35295-35485	191	38.29724	-62.843805	-61.7	-0.01854	-1.85381751	-62.1	-0.01198	-1.197754272
195	AFQF01002518.1/111488-111680	193	38.37486	-63.366522	-58.6	-0.08134	-8.13399578	-58.5	-0.08319	-8.318840214
196	AAGW02065159.1/39337-39527	191	37.95447	-62.298355	-61.8	-0.00806	-0.80639912	-64.6	0.035629	3.562918487
197	AAWZ02022241.1/1219-1409	191	38.41518	-63.031481	-64.5	0.022768	2.276773105	-66.2	0.047863	4.786281953
198	CAAB02025078.1/1453-1643	191	37.91472	-62.235101	-70.8	0.120973	12.09731537	-71.4	0.12836	12.83599339
199	AAQR03042593.1/1155-1347	193	38.76579	-63.988601	-55	-0.16343	-16.3429115	-57.2	-0.11868	-11.86818409

200	AACU03000093.1/631552-631747	196	38.4649	-64.108599	-63	-0.0176	-1.7596813	-63.7	-0.00641	-0.641443039
201	AE014186.2/1461495-1461298	198	30.9377	-52.529761	-53	0.008872	0.887243301	-55.3	0.050095	5.009473688
202	FR799006.1/251703-251558	146	38.62345	-54.3809	-53	-0.02605	-2.60547173	-55.3	0.01662	1.662025288
203	AAFN02000024.1/475809-475596	214	30.48033	-54.995552	-50.9	-0.08046	-8.04627166	-52.6	-0.04554	-4.554281893
204	K00034.1/420-610	191	37.71535	-61.917831	-59.1	-0.04768	-4.7679036	-60.7	-0.02006	-2.006311415
205	ABDG02000029.1/618164-617972	193	38.41414	-63.429014	-60.7	-0.04496	-4.49590508	-60	-0.05715	-5.715023971
206	AP004871.3/124344-124540	197	51.9339	-85.741411	-86.1	0.004165	0.416479305	-85.5	-0.00282	-0.282352419
	II. Rfam snRNA family RF00007 (62)									
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_C**	**MFE_F**	**RD_2**	**%RD_2**	**MFE_S**	**RD_3**	**%RD_3**
207	AANN01056468.1/521-372	150	39.56829	-56.682815	-57.7	0.017629	1.76288636	-58.3	0.027739	2.773902967
208	AADA01322814.1/1127-978	150	39.56829	-56.682815	-57.7	0.017629	1.76288636	-58.3	0.027739	2.773902967
209	AANU01293824.1/906-757	150	39.56829	-56.682815	-57.7	0.017629	1.76288636	-58.3	0.027739	2.773902967
210	ABBA01017933.1/18965-18816	150	40.56829	-58.274115	-58.3	0.000444	0.044400395	-58.9	0.010626	1.06262382
211	ABDC01356688.1/591-740	150	39.56829	-56.682815	-57.7	0.017629	1.76288636	-58.3	0.027739	2.773902967
212	AAFC03093377.1/31487-31636	150	34.64102	-48.842049	-55.6	0.121546	12.15458813	-56.2	0.130924	13.0924395

213	AAQQ01306058.1/1332-1183	150	34.4516	-48.540631	-55.6	0.126967	12.69670631	-56.2	0.136288	13.62876995
214	AANG01542153.1/730-879	150	34.56829	-48.726315	-55.6	0.123627	12.36274358	-56.2	0.132984	13.29837265
215	AAIY01326656.1/671-820	150	34.94483	-49.325503	-57.6	0.143654	14.36544561	-57.7	0.145139	14.51385904
216	AAPN01231707.1/282-431	150	34.46621	-48.563876	-55.4	0.123396	12.33957386	-56	0.132788	13.27879272
217	AAHX01055169.1/34088-34238	151	34.55295	-48.901517	-58.9	0.169754	16.97535392	-59.9	0.183614	18.36140811
218	AALT01414211.1/1221-1073	149	39.32063	-56.089113	-55.8	-0.00518	-0.51812365	-55.8	-0.00518	-0.518123651
219	AAHY01168842.1/1247-1097	151	37.15945	-53.049227	-59.1	0.102382	10.23819466	-59.7	0.111403	11.14032336
220	AAPY01023785.1/1285-1135	151	35.15945	-49.866627	-55.4	0.09988	9.988037988	-56.4	0.11584	11.58399476
221	AAVX01293999.1/160-11	150	35.14591	-49.645485	-48	-0.03428	-3.42809278	-47.9	-0.03644	-3.644017822
222	CAAE01014653.1/353935-353782	154	39.76852	-57.799839	-56	-0.03214	-3.21399783	-55.9	-0.03399	-3.398638259
223	BAAE01110703.1/791-944	154	35.64054	-51.230997	-57.8	0.113651	11.36505635	-57.7	0.112114	11.21144293
224	AANH01004214.1/17675-17828	154	34.55071	-49.496746	-53.6	0.076553	7.655323709	-54	0.083394	8.339358348
225	ABAV01000136.1/26200-26351	152	33.96453	-48.164759	-54.6	0.117862	11.78615555	-53.2	0.094647	9.464738593
226	AAZO01006159.1/30950-30796	155	34.41864	-49.486184	-48.4	-0.02244	-2.24418105	-47.9	-0.03311	-3.311448075
227	AAAB01008960.1/5726267-5726086	182	47.24875	-75.29193	-75.3	0.000107	0.010717087	-72	-0.04572	-4.572125046

228	AAJJ01000520.1/29099-28948	152	34.08288	-48.353079	-49.3	0.019207	1.920731527	-50.3	0.038706	3.87061758
229	AAZX01007808.1/26508-26658	151	34.45083	-48.739013	-47.9	-0.01752	-1.75159288	-47.9	-0.01752	-1.751592885
230	AABS01000019.1/293615-293466	150	34.42237	-48.494112	-50.3	0.035902	3.590234352	-52.6	0.078059	7.805870492
231	AAYZ01032557.1/2146-2294	149	26.97419	-36.442231	-34.8	-0.04719	-4.71905387	-34.5	-0.0563	-5.629654337
232	AACT01038531.1/95659-95808	150	35.11725	-49.599885	-56.6	0.123677	12.36769415	-53.2	0.067671	6.767133248
233	AASG02002046.1/26669-26822	154	36.05682	-51.893412	-59.5	0.127842	12.78418101	-59.7	0.130764	13.0763613
234	AADK01040274.1/1841-1691	151	34.94405	-49.523868	-54.6	0.092969	9.296945805	-52.2	0.051267	5.126690439
235	AC198009.4/84558-84712	155	34.95544	-50.340399	-53.9	0.066041	6.604083093	-54.2	0.07121	7.121034663
236	AAGE02014219.1/46421-46245	177	37.58967	-58.923435	-61.5	0.041895	4.189536983	-62.6	0.058731	5.873107419
237	AARH01003540.1/648176-648022	155	35.68213	-51.496771	-56.7	0.091768	9.176770916	-58.3	0.116693	11.66934667
238	AC004255.1/89334-89170	165	35.42038	-53.076258	-57.6	0.078537	7.853718206	-57.6	0.078537	7.853718206
239	AP005874.3/27602-27759	158	46.08659	-68.652192	-67.7	-0.01406	-1.40648744	-68.7	0.000696	0.069589522
240	AAAA02006813.1/31506-31663	158	46.08659	-68.652192	-67.7	-0.01406	-1.40648744	-68.7	0.000696	0.069589522
241	AAWT01090545.1/3123-3274	152	33.69674	-47.73862	-43.3	-0.10251	-10.2508552	-44.4	-0.07519	-7.519415057
242	AAWU01013761.1/27501-27308	194	38.72635	-64.125449	-66.4	0.034255	3.4255291	-66.7	0.038599	3.859897035

243	AATU01009112.1/74318-74156	163	35.74688	-53.196612	-59.8	0.110425	11.04245511	-63	0.155609	15.56093358
244	AAQX01001042.1/31224-31386	163	41.92937	-63.034809	-65	0.030234	3.023370403	-66.9	0.057776	5.777564667
245	DQ888370.1/1-162	162	45.29229	-68.186624	-90.3	0.244888	24.48878877	-92.6	0.263643	26.36433722
246	AAEU02001091.1/68476-68268	209	49.98827	-85.040538	-90.3	0.058244	5.824432313	-92.6	0.081636	8.163566283
247	AAPQ01006579.1/170624-170417	208	51.82909	-87.770229	-90.2	0.026938	2.693759404	-91.8	0.043897	4.389728739
248	AF459090.1/1-212	212	41.4379	-72.03313	-81.9	0.120475	12.0474607	-82.6	0.127928	12.7928212
249	AAKO01000167.1/274946-275154	209	53.03727	-89.892411	-90.7	0.008904	0.890395903	-95.3	0.056743	5.674280256
250	AASV01051056.1/691-484	208	49.95196	-84.78315	-84.8	0.000199	0.019870803	-90.4	0.062133	6.213330134
251	AAQB01007708.1/1369-1503	135	44.04067	-60.805717	-60.3	-0.00839	-0.83866794	-60.5	-0.00505	-0.505316974
252	AAIZ01008995.1/21704-21918	215	41.65417	-72.976085	-83	0.12077	12.07700623	-84	0.131237	13.12370854
253	AAFS01000475.1/55736-55522	215	43.83468	-76.445921	-82.9	0.077854	7.785378447	-84	0.08993	8.992950873
254	AAPU01011411.1/133226-133037	190	39.33374	-64.293576	-75.7	0.150679	15.06793197	-74.8	0.14046	14.04602206
255	AM487500.2/4877-4718	160	35.38112	-52.015782	-58.1	0.10472	10.47197623	-58.9	0.11688	11.68797656
256	AASC02028416.1/49450-49599	150	40.877	-58.765367	-57.7	-0.01846	-1.84639034	-57.5	-0.02201	-2.200638657
257	AAGV020551495.1/1052-903	150	34.71359	-48.957541	-55.6	0.119469	11.94686879	-56.6	0.135026	13.50257782
258	AAWR02025158.1/5	150	34.93042	-49.302577	-54.1	0.088677	8.867694925	-54.7	0.098673	9.867318015

	32-681									
259	AAWZ02031009.1/19848-19997	150	40.74258	-58.55147	-58.4	-0.00259	-0.25936675	-57.7	-0.01476	-1.475684893
260	EU240273.1/1-149	149	41.74336	-59.944415	-58.7	-0.0212	-2.119958	-57.6	-0.0407	-4.070165532
261	AAEX03007273.1/32318-32169	150	40.64102	-58.389849	-57.8	-0.01021	-1.0205	-58.3	-0.00154	-0.154114926
262	AAGJ04047220.1/14069-13916	154	34.92738	-50.096133	-57	0.12112	12.11204764	-56.6	0.114909	11.49093137
263	AADG06004603.1/1703-1555	149	33.29323	-46.497718	-44.9	-0.03558	-3.55839088	-45.6	-0.01969	-1.968678739
264	AAGW02065961.1/10800-10651	150	39.53915	-56.636453	-57.7	0.018432	1.843236464	-58.3	0.028534	2.85342614
265	AAQY02000248.1/644530-644694	165	40.19308	-60.671048	-59.4	-0.0214	-2.13981122	-63.4	0.043043	4.30434091
266	AAQR03135034.1/15819-15671	149	29.26298	-40.084376	-39.7	-0.00968	-0.96820196	-40.2	0.002876	0.287621443
267	AAGU03077213.1/2720-2571	150	39.78602	-57.029292	-58	0.016736	1.673634572	-58.6	0.026804	2.680389167
268	CAAB02002948.1/53519-53671	153	40.35776	-58.5379	-56.6	-0.03424	-3.4238522	-58.1	-0.00754	-0.753701113
III. Rfam snRNA family RF00015 (170)										
Sl. No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_C**	**MFE_F**	**RD_2**	**%RD_2**	**MFE_S**	**RD_3**	**%RD_3**
269	AAIY01144063.1/2359-2499	141	28.71992	-37.623401	-36.1	-0.0422	-4.21994796	-37.5	-0.00329	-0.329069905
270	AC193264.3/174224-174084	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264

271	AADD01128634.1/1009-869	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
272	AAFR03070450.1/4019-4159	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
273	AAPN01043183.1/815-675	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
274	AAHY01048392.1/24763-24623	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
275	ABDC01319198.1/612-472	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
276	AANG01100342.1/1096-956	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
277	AALT01138445.1/833-693	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
278	AANU01246434.1/3940-4080	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
279	AANN01193588.1/6106-6246	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
280	AACN010181221.1/232-92	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.702544264
281	AAFC03029538.1/15458-15318	141	30.41065	-40.313862	-35.4	-0.13881	-13.880965	-37.1	-0.08663	-8.66269977
282	AAPY01772888.1/1357-1497	141	28.71992	-37.623401	-35.8	-0.05093	-5.09329948	-37.5	-0.00329	-0.329069905
283	AAYZ01170597.1/2807-2947	141	27.84281	-36.227659	-34.9	-0.03804	-3.80418185	-36.6	0.010173	1.017323863
284	AB168678.1/1-141	141	26.76605	-34.514222	-35.8	0.035916	3.591560024	-37.5	0.079621	7.96207597
285	K00476.1/2-142	141	30.364	-40.239635	-36.2	-0.11159	-11.1592129	-37.9	-0.06173	-6.173179563
286	BAAE01269333.1/20	141	32.56565	-43.743116	-37.8	-0.15723	-15.7225281	-39	-0.12162	-12.16183496

	18-1878									
287	AAZO01005801.1/5213-5353	141	27.03558	-34.943123	-35.3	0.01011	1.01098283	-36.3	0.03738	3.737953
288	BC124577.1/1-141	141	32.53471	-43.693879	-38.5	-0.13491	-13.4905944	-39.8	-0.09784	-9.783615176
289	AANH01001405.1/82396-82256	141	27.90408	-36.325165	-37.8	0.039017	3.90168069	-39.1	0.070968	7.096765475
290	ABAV01003454.1/17582-17444	139	25.98221	-32.867689	-33.8	0.027583	2.758317712	-34.4	0.044544	4.454393566
291	AAZX01000257.1/602-462	141	25.12976	-31.910386	-31	-0.02937	-2.93672862	-34.2	0.066948	6.694778151
292	AAGE02020535.1/44175-44035	141	32.41065	-43.496462	-42.3	-0.02829	-2.82851446	-41.4	-0.05064	-5.063916944
293	AAVX01087583.1/989-849	141	33.10244	-44.59731	-39.2	-0.13769	-13.7686486	-40.5	-0.10117	-10.11681541
294	K03095.1/1-139	139	28.80056	-37.352525	-36.8	-0.01501	-1.50142601	-38.7	0.034818	3.481848132
295	AACT01063233.1/31971-31831	141	28.91938	-37.940813	-37	-0.02543	-2.54273689	-38.8	0.022144	2.214400387
296	AAWU01039949.1/7621-7481	141	32.16109	-43.099343	-39.3	-0.09668	-9.66753848	-40.6	-0.06156	-6.156016314
297	AAAB01008933.1/1598961-1598821	141	32.28611	-43.298286	-42.6	-0.01639	-1.63916858	-45.2	0.042073	4.207332272
298	X15933.1/1-149	149	34.33529	-48.155948	-56.5	0.147682	14.7682345	-57.4	0.161046	16.10462107
299	AC084591.1/18183-18321	139	33.63239	-45.041423	-40.2	-0.12043	-12.0433402	-40	-0.12604	-12.60355694
300	CU302335.1/69797-69946	150	34.12866	-48.026729	-53.5	0.102304	10.23041327	-53.8	0.10731	10.73098717
301	AABS01000042.1/351938-351798	141	33.42037	-45.103242	-41.1	-0.0974	-9.74024785	-40.6	-0.11092	-11.09172874

302	AADK01043685.1/1410-1272	139	32.06865	-42.553049	-42.6	0.001102	0.110214442	-44.8	0.050155	5.015516412
303	AACG02000302.1/224-363	140	28.05761	-36.369879	-35.9	-0.01309	-1.30885527	-35.6	-0.02163	-2.162581575
304	AACA01000783.1/223-362	140	28.05761	-36.369879	-35.9	-0.01309	-1.30885527	-35.6	-0.02163	-2.162581575
305	AAPU01010717.1/67287-67148	140	30.00495	-39.46867	-39.5	0.000793	0.079316379	-40.3	0.020629	2.062853523
306	AAPQ01007039.1/986580-986441	140	32.16191	-42.901048	-39.5	-0.0861	-8.61024869	-41.2	-0.04129	-4.128757847
307	AANI01016129.1/182015-181878	138	31.46705	-41.396115	-37.4	-0.10685	-10.6848003	-37.1	-0.1158	-11.57982566
308	AAJJ01000336.1/45867-45727	141	33.072	-44.548873	-41.2	-0.08128	-8.12833158	-42.9	-0.03844	-3.843525898
309	AAAA02007064.1/44912-44768	145	33.55241	-46.111742	-52.9	0.128322	12.83224526	-55.5	0.169158	16.91577972
310	AAKO01002834.1/25643-25504	140	32.25572	-43.050332	-42.3	-0.01774	-1.77383386	-43	-0.00117	-0.117050522
311	AAEU02000313.1/218060-217921	140	32.25572	-43.050332	-42.3	-0.01774	-1.77383386	-43	-0.00117	-0.117050522
312	AASS01015485.1/140-1	140	32.25572	-43.050332	-42.3	-0.01774	-1.77383386	-43	-0.00117	-0.117050522
313	X07113.1/1-150	150	34.50999	-48.633551	-53.7	0.094347	9.434727355	-52.7	0.077162	7.716221233
314	AAPP01015712.1/40466-40605	140	32.33369	-43.174404	-43.8	0.014283	1.428301888	-45.6	0.053193	5.319289971
315	AM479189.1/4511-4661	151	39.58208	-56.904358	-59.3	0.040399	4.039868485	-60.9	0.06561	6.560988525
316	AASG02000802.1/49546-49696	151	42.20237	-61.074027	-61.5	0.006926	0.692638772	-62.4	0.02125	2.124956482

317	AAPT01020986.1/3963-4102	140	32.55101	-43.520224	-41.1	-0.05889	-5.88862407	-41.9	-0.03867	-3.866884233
318	AARH01003623.1/42069-42219	151	40.05909	-57.663435	-58.7	0.017659	1.765869428	-60.2	0.042136	4.213563711
319	X67145.1/194-344	151	34.53838	-48.878331	-52.4	0.067207	6.72074205	-53.3	0.082958	8.295813948
320	AP004858.3/49137-48993	145	33.62735	-46.231001	-50.3	0.080895	8.089461812	-52.9	0.126068	12.60680395
321	AAQA01000004.1/379392-379252	141	33.5407	-45.294716	-45.3	0.000117	0.011663876	-46.1	0.017468	1.746819383
322	AP006099.1/69439-69589	151	34.91523	-49.478007	-52.9	0.064688	6.468796638	-53.8	0.080334	8.033445021
323	AAWT01050999.1/11219-11362	144	26.65372	-34.934266	-33.8	-0.03356	-3.35581641	-34	-0.02748	-2.747841018
324	AATT01000006.1/24344-24478	135	31.75642	-41.25779	-35.5	-0.16219	-16.2191256	-36.6	-0.12726	-12.72620108
325	AAQB01006409.1/413450-413312	139	31.76887	-42.075997	-40.6	-0.03635	-3.63546067	-41	-0.02624	-2.624383007
326	AP007155.1/2359336-2359191	146	30.70133	-41.774419	-40.4	-0.03402	-3.40202613	-40.7	-0.0264	-2.639849035
327	AAIH02000036.1/105327-105472	146	30.70133	-41.774419	-40.4	-0.03402	-3.40202613	-40.7	-0.0264	-2.639849035
328	AACW02000210.1/206231-206366	136	33.122	-43.630433	-36	-0.21196	-21.1956483	-37.2	-0.17286	-17.2861112
329	AC192395.1/10098-9958	141	23.53471	-29.372179	-28.6	-0.027	-2.69992601	-27.7	-0.06037	-6.0367466
330	CAAJ01010632.1/7788-7919	132	25.15046	-30.146927	-29	-0.03955	-3.95491952	-29.4	-0.02541	-2.54056687
331	CAAI01006665.1/11117-11248	132	25.15046	-30.146927	-29	-0.03955	-3.95491952	-29.4	-0.02541	-2.54056687

332	AAKM01000017.1/282531-282664	134	26.13576	-32.114027	-30.4	-0.05638	-5.63824791	-32.7	0.01792	1.79196524
333	L22250.1/171-310	140	33.76602	-45.453662	-40.6	-0.11955	-11.9548335	-42.4	-0.07202	-7.20203396
334	AAFU01001086.1/4264-4401	138	33.19617	-44.147658	-35.7	-0.23663	-23.6629078	-35.5	-0.2436	-24.3596002
335	AAYL01000045.1/85592-85722	131	42.09625	-56.913157	-54.3	-0.04812	-4.8124442	-53.6	-0.06181	-6.18126343
336	AAXJ01014857.1/1040-1174	135	32.894	-43.068015	-39.8	-0.08211	-8.21109218	-41.7	-0.03281	-3.28061075
337	AANS01000355.1/46407-46542	136	26.81893	-33.600367	-32.7	-0.02753	-2.75341518	-33.8	0.005906	0.59063088
338	AATM01000105.1/197284-197424	141	28.76605	-37.696822	-34.7	-0.08636	-8.63637323	-36.8	-0.02437	-2.43701497
339	AAGT01000338.1/12653-12519	135	32.32242	-42.158466	-33.8	-0.24729	-24.7291901	-34	-0.23995	-23.9954889
340	AAXI01000285.1/23727-23857	131	31.96561	-40.79227	-35.7	-0.14264	-14.2640606	-36.1	-0.12998	-12.9979768
341	AAID01000631.1/14812-14946	135	29.36914	-37.458912	-36.4	-0.02909	-2.90909818	-36.9	-0.01515	-1.51466595
342	CU329671.1/467615-467481	135	24.78813	-30.169148	-26.9	-0.12153	-12.1529662	-28	-0.07747	-7.74695681
343	AAXT01000001.1/1022888-1022758	131	26.07575	-31.419744	-31.3	-0.00383	-0.38256759	-32.4	0.030255	3.02548254
344	AAFT01000065.1/55802-55646	157	30.50325	-43.65482	-38.8	-0.12512	-12.5124226	-39.2	-0.11364	-11.3643366
345	ABAS01000010.1/147598-147449	150	30.64102	-42.476849	-41	-0.03602	-3.60207074	-42.3	-0.00418	-0.4180827
346	AACM02000196.1/54942-55125	184	30.39285	-48.868342	-49.6	0.014751	1.475116612	-50.4	0.03039	3.03900365

347	AAVQ01000002.1/194294-194149	146	23.98397	-31.085094	-31.3	0.006866	0.686599186	-30.4	-0.02254	-2.2536001
348	ABAR01000001.1/643558-643722	165	30.91402	-45.905285	-41	-0.11964	-11.9641106	-43	-0.06756	-6.7564775
349	CAAL01000681.1/399056-398923	134	31.18101	-40.142548	-35	-0.14693	-14.6929934	-38.7	-0.03728	-3.7275134
350	BX842620.1/46696-46842	147	34.46857	-47.968833	-40.4	-0.18735	-18.734734	-39.2	-0.22369	-22.369470
351	AAFO01000045.1/228200-228380	181	23.62528	-37.500308	-33.5	-0.11941	-11.9412165	-33.4	-0.12276	-12.276369
352	AACQ01000084.1/105435-105615	181	23.62528	-37.500308	-33.5	-0.11941	-11.9412165	-33.4	-0.12276	-12.276369
353	AAJN01000077.1/141935-142086	152	34.77001	-49.446518	-42.4	-0.16619	-16.6191462	-42.2	-0.17172	-17.171843
354	AACD01000007.1/178412-178544	133	32.47971	-42.009569	-35.2	-0.19345	-19.3453659	-36.4	-0.15411	-15.410903
355	CR382130.1/2966271-2966119	153	30.856	-43.417754	-41.1	-0.05639	-5.63930419	-41.2	-0.05383	-5.3828981
356	AM269959.1/11788-11936	149	30.18837	-41.556952	-38.9	-0.0683	-6.83021092	-39.2	-0.06013	-6.0126327
357	AAFM01000021.1/757822-757672	151	26.74181	-36.471642	-36.1	-0.01029	-1.0294781	-34.5	-0.05715	-5.7149031
358	AACY020397167.1/942-809	134	33.18461	-43.330863	-43.6	0.006173	0.617287151	-45.6	0.049762	4.97617806
359	AAKE03000003.1/920249-920379	131	32.48159	-41.613358	-34.9	-0.19236	-19.2359829	-36.2	-0.14954	-14.954027
360	AB189720.1/1-134	134	27.50248	-34.288891	-32.7	-0.04859	-4.85899262	-33.4	-0.02661	-2.6613490
361	AAGK01000002.1/935734-935865	132	32.05911	-41.140655	-35.2	-0.16877	-16.8768606	-37.8	-0.08838	-8.8377114

362	AAHF01000007.1/1275121-1274962	160	36.18263	-53.291212	-48.3	-0.10334	-10.3337723	-48.8	-0.09203	-9.2033032
363	AC198144.2/107120-106981	140	20.006	-23.557346	-20.8	-0.13256	-13.2564709	-20.6	-0.14356	-14.356048
364	ABDB01000004.1/78964-78804	161	36.19582	-53.511811	-47.3	-0.13133	-13.1327926	-48.1	-0.11251	-11.251166
365	AAEE01000007.1/708153-708014	140	23.81557	-29.61951	-30.3	0.022458	2.245842991	-32.8	0.096966	9.6966171
366	AAEL01000435.1/3438-3299	140	23.81557	-29.61951	-30.3	0.022458	2.245842991	-32.8	0.096966	9.69661715
367	AAPO01000090.1/100324-100179	146	23.57147	-30.428673	-30	-0.01429	-1.42890864	-30.3	-0.00425	-0.4246620
368	AAQQ01631221.1/1703-1837	135	22.53748	-26.587691	-21.5	-0.23664	-23.6636786	-22	-0.20853	-20.853140
369	CR382124.1/1170861-1171036	176	26.17229	-40.555359	-36.3	-0.11723	-11.722752	-36.8	-0.10205	-10.204779
370	AAKD03000006.1/347523-347653	131	32.68262	-41.933258	-34.7	-0.20845	-20.8451239	-35.1	-0.19468	-19.467971
371	AAWC01000056.1/46279-46148	132	32.61995	-42.033125	-39.5	-0.06413	-6.41297404	-40.7	-0.03275	-3.2754907
372	ABCN01001426.1/26711-26844	134	33.38137	-43.643972	-40	-0.0911	-9.10993082	-41.3	-0.05675	-5.6754777
373	AANV02000065.1/936-807	130	30.46538	-38.205363	-33.3	-0.14731	-14.7308207	-34.4	-0.11062	-11.062102
374	AAIM02000062.1/67925-68097	173	28.21567	-43.20819	-44.6	0.031207	3.120650912	-45.1	0.041947	4.1947013
375	AATU01001408.1/348748-348878	131	23.44115	-27.227308	-28.1	0.031057	3.10566519	-27.9	0.024111	2.41108214
376	AF270843.1/940-1107	168	28.70136	-42.983068	-44.6	0.036254	3.625408209	-44.1	0.025327	2.53272576

377	AAQX01001192.1/50846-50711	136	28.69341	-36.583228	-35.7	-0.02474	-2.4740287	-35.8	-0.02188	-2.1877883
378	AARE01001618.1/3212-3345	134	32.89491	-42.869872	-38.5	-0.1135	-11.3503177	-38.1	-0.12519	-12.519349
379	AADM01000245.1/23026-22856	171	25.65719	-38.737689	-35.8	-0.08206	-8.20583401	-37	-0.04696	-4.6964556
380	ABBC01001255.1/19365-19501	137	28.38293	-36.288763	-32.1	-0.13049	-13.0491053	-34.4	-0.05491	-5.4905895
381	AATX01000107.1/49292-49428	137	28.38293	-36.288763	-32.1	-0.13049	-13.0491053	-34.4	-0.05491	-5.4905895
382	AASO01000240.1/58-194	137	28.38293	-36.288763	-32.1	-0.13049	-13.0491053	-34.4	-0.05491	-5.4905895
383	ABBB01000091.1/60646-60782	137	27.61541	-35.067396	-31.8	-0.10275	-10.2748286	-34.4	-0.0194	-1.9401032
384	CP000582.1/123212-123076	137	29.22702	-37.631959	-37.3	-0.0089	-0.8899692	-39.5	0.047292	4.72921895
385	AAIW01000278.1/37304-37168	137	28.5381	-36.535678	-32.8	-0.11389	-11.389263	-36.4	-0.00373	-0.3727425
386	AACI02000576.1/2356-2546	191	25.71535	-42.822231	-40.4	-0.05996	-5.99562136	-41.8	-0.02446	-2.4455287
387	AACF01000119.1/10356-10566	211	27.22105	-49.210251	-46.9	-0.04926	-4.92590841	-47	-0.04703	-4.7026617
388	AAJI01000076.1/250-381	132	28.72737	-35.838863	-31.2	-0.14868	-14.8681492	-32	-0.11996	-11.996445
389	AABZ01000001.1/29367-29202	166	26.81518	-39.582399	-38.8	-0.02016	-2.0164932	-39.3	-0.00719	-0.718573
390	U18778.1/15676-15446	231	31.28412	-59.667814	-53.2	-0.12158	-12.1575457	-54.8	-0.08883	-8.8828728
391	AAEG01000006.1/103924-103694	231	31.28412	-59.667814	-53.2	-0.12158	-12.1575457	-54.8	-0.08883	-8.8828728

392	AABY01000063.1/21244-21029	216	30.20552	-54.957442	-52.1	-0.05485	-5.48453292	-53.4	-0.02917	-2.9165577
393	AADS01000270.1/32451-32584	134	31.88395	-41.261135	-37.4	-0.10324	-10.3238893	-36.7	-0.12428	-12.428159
394	AE017356.1/390372-390503	132	30.97204	-39.410805	-35.6	-0.10705	-10.7045085	-34.8	-0.13249	-13.249439
395	AAEY01000066.1/357595-357726	132	30.97204	-39.410805	-35.6	-0.10705	-10.7045085	-34.8	-0.13249	-13.249439
396	AACO02000129.1/36387-36520	134	28.21331	-35.420039	-34.4	-0.02965	-2.96522985	-33.3	-0.06366	-6.3664836
397	AAZN01000370.1/73687-73536	152	23.71146	-31.849049	-30.3	-0.05112	-5.11237363	-31	-0.02739	-2.7388684
398	AAFP01000576.1/27349-27217	133	31.03615	-39.712421	-35.6	-0.11552	-11.5517439	-34.8	-0.14116	-14.116151
399	AANW02001764.1/480-354	127	22.21904	-24.484161	-24	-0.02017	-2.01733718	-24.9	0.0167	1.6700364
400	AACP01000036.1/91007-91135	129	28.79365	-35.345529	-34	-0.03957	-3.95743873	-34.2	-0.0335	-3.3495004
401	X13840.1/1-118	118	30.2782	-35.512295	-29.4	-0.2079	-20.7901202	-29.2	-0.21617	-21.617449
402	AC149882.2/166659-166543	117	25.53738	-27.768635	-29.5	0.05869	5.869034136	-32	0.13223	13.223015
403	AC190402.1/26804-26930	127	23.10208	-25.889332	-24.8	-0.04392	-4.39246767	-26.1	0.008072	0.80715715
404	DQ451048.1/205-341	137	25.72333	-32.056531	-30.7	-0.04419	-4.41866901	-34.1	0.059926	5.99257658
405	AL590450.1/23479-23642	164	35.52032	-53.035688	-48.1	-0.10261	-10.2613053	-49.9	-0.06284	-6.2839436
406	ABIT01000802.1/5235-5099	137	29.61541	-38.249996	-31.8	-0.20283	-20.2830047	-34.4	-0.11192	-11.191847
407	AAZY02000012.1/10	133	33.71269	-43.971603	-43.4	-0.01317	-1.3170583	-44.6	0.01409	1.40896120

	7951-108083									
408	AAKN02007150.1/17728-17588	141	27.94996	-36.398175	-35.8	-0.01671	-1.67087886	-38	0.042153	4.2153299
409	M15957.1/271-411	141	28.48824	-37.254736	-36.2	-0.02914	-2.9136346	-37.9	0.017025	1.70254426
410	AL844502.1/365248-365383	136	27.70793	-35.015023	-32.7	-0.0708	-7.07958159	-33.4	-0.04835	-4.8353987
411	AASC02060889.1/1492-1632	141	33.64563	-45.461695	-42.8	-0.06219	-6.21891272	-43.5	-0.0451	-4.5096428
412	AM055942.3/1615661-1615791	131	40.09625	-53.730557	-54.3	0.010487	1.048697603	-53.6	-0.00244	-0.2435768
413	AC189493.2/83218-83068	151	34.78525	-49.271162	-54.2	0.090938	9.093797373	-55.2	0.107406	10.740648
414	AACE03000002.2/368765-368919	155	27.66631	-38.741204	-37.2	-0.04143	-4.14302054	-38.4	-0.00889	-0.8885511
415	AAGF03001264.1/335958-336087	130	31.36185	-39.631918	-37	-0.07113	-7.11329298	-37.5	-0.05685	-5.6851157
416	ABRQ01104616.1/4922-4782	141	27.71992	-36.032101	-35.8	-0.00648	-0.64832741	-38	0.051787	5.1786810
417	CR382136.2/1404151-1404006	146	24.95433	-32.629226	-30.2	-0.08044	-8.0437945	-30	-0.08764	-8.7640864
418	AAWR02035467.1/41938-42078	141	28.64287	-37.500803	-37.6	0.002638	0.26382063	-38.9	0.035969	3.5969063
419	AAGD02002631.1/646-784	139	33.51238	-44.850453	-40.5	-0.10742	-10.7418583	-40.3	-0.11291	-11.291445
420	AAGJ04107959.1/2135-2275	141	28.2549	-36.883422	-32.9	-0.12108	-12.1076666	-34.3	-0.07532	-7.5318434
421	ABEG02000949.1/11902-11764	139	33.48231	-44.802603	-39.5	-0.13424	-13.4243116	-39.3	-0.14002	-14.001534
422	AACS02000001.1/56	133	31.77406	-40.886662	-36.4	-0.12326	-12.325995	-37	-0.10504	-10.504492

	6465-566333									
423	AE016817.6/721964-721809	156	33.64904	-48.461121	-38.4	-0.26201	-26.200837	-37.1	-0.30623	-30.622968
424	AAPE02005503.1/37803-37943	141	27.33287	-35.416191	-34.4	-0.02954	-2.95404416	-35.7	0.00795	0.79498266
425	AAEC03000003.1/2600987-2601123	137	32.27387	-42.480416	-30.4	-0.39738	-39.7382094	-33.5	-0.26807	-26.80721
426	ADTU01005844.1/39857-39997	141	28.31729	-36.982701	-35.8	-0.03304	-3.30363508	-36.7	-0.0077	-0.7703034
427	AAIL02000028.1/1015369-1015502	134	32.04156	-41.511937	-36.8	-0.12804	-12.8041755	-37.9	-0.0953	-9.5301756
428	AAWZ02025418.1/39904-40044	141	32.51923	-43.669243	-38.5	-0.13427	-13.4266049	-39.2	-0.11401	-11.401129
429	AAQY02000456.1/44132-43994	139	32.8265	-43.759004	-36.4	-0.20217	-20.2170448	-38.3	-0.14253	-14.253274
430	AACU03000146.1/753634-753781	148	26.92308	-36.161301	-32.9	-0.09913	-9.9127684	-34.3	-0.05427	-5.4265329
431	Z74042.2/30433-30295	139	30.81161	-40.552707	-39.6	-0.02406	-2.40582615	-40.5	-0.0013	-0.1301411
432	AFQF01001265.1/35199-35370	172	30.10803	-46.019907	-42.9	-0.07273	-7.27251102	-44.6	-0.03184	-3.1836484

IV. Rfam snRNA family RF00020 (180)

Sl.no.	Sequence ID	NTL	SD_DFT	MFE_C	MFE_F	RD_2	%RD_2	MFE_S	RD_3	%RD_3
433	AAFC03028536.1/14883-14999	117	29.50321	-34.079465	-32.4	-0.05184	-5.18353379	-33.8	-0.00827	-0.8268193
434	AAFR03051183.1/13754-13869	116	29.97912	-34.637179	-33.2	-0.04329	-4.32885151	-32.8	-0.05601	-5.6011545

435	AALT01479958.1/1552-1436	117	29.84312	-34.620364	-32.7	-0.05873	-5.87267194	-34.1	-0.01526	-1.5259933
436	AC083892.19/144818-144933	116	29.29846	-33.554046	-32.1	-0.0453	-4.52973819	-31.6	-0.06184	-6.1836896
437	AANG01141002.1/575-691	117	29.29738	-33.751914	-34.4	0.01884	1.883972335	-34.5	0.021684	2.1683666
438	ABDC01046604.1/3549-3664	116	29.21227	-33.41688	-30.1	-0.1102	-11.0195345	-31.5	-0.06085	-6.0853329
439	AAQQ01236639.1/2054-2171	118	29.58748	-34.413152	-31	-0.1101	-11.0101689	-32.4	-0.06213	-6.2134332
440	AAIY01547840.1/3111-2996	116	28.98695	-33.058339	-34.2	0.033382	3.338189347	-34.7	0.04731	4.7310108
441	AAYZ01307320.1/25593-25710	118	29.51922	-34.304542	-34.5	0.005665	0.566545058	-35.9	0.044442	4.4441728
442	AC068213.7/527-412	116	29.29846	-33.554046	-33.1	-0.01372	-1.37174006	-34.5	0.027419	2.7418957
443	AANN01833323.1/1168-1053	116	29.17772	-33.3619	-30.8	-0.08318	-8.31785715	-30.3	-0.10105	-10.105280
444	AANN01833323.1/1168-1053	116	29.7257	-34.23391	-32	-0.06981	-6.98097015	-31.5	-0.08679	-8.6790807
445	AAPN01296676.1/948-833	116	29.84423	-34.422528	-33.7	-0.02144	-2.1440013	-33.1	-0.03996	-3.9955541
446	CAAE01011816.1/72155-72041	115	29.4541	-33.602105	-37.7	0.108697	10.86974707	-38.6	0.129479	12.947913
447	K03164.1/1-115	115	35.09779	-42.582919	-40.4	-0.05403	-5.40326562	-40.2	-0.05928	-5.9276599
448	AAVX01303608.1/479-595	117	28.53738	-32.542535	-32.6	0.001763	0.176273222	-32.1	-0.01379	-1.3786134
449	BAAF04017217.1/172-285	114	28.63904	-32.105507	-36.4	0.117981	11.798059	-36.7	0.125191	12.519055
450	X63789.1/2235-2348	114	28.28458	-31.541459	-31.5	-0.00132	-0.13161498	-31.3	-0.00771	-0.7714336

451	X06020.1/401-515	115	28.91837	-32.749597	-33.1	0.010586	1.058619778	-33.9	0.033935	3.3935196
452	K03096.1/1-119	119	29.07063	-33.790298	-32.3	-0.04614	-4.61392592	-31.4	-0.07612	-7.612414
453	AC174762.1/131410-131525	116	29.17772	-33.3619	-34.8	0.041325	4.13247126	-34.3	0.02735	2.7349854
454	AASG02001471.1/28708-28826	119	35.14358	-43.454184	-44.1	0.014644	1.464436389	-43.9	0.010155	1.0155272
455	AM454615.1/12361-12244	118	34.77437	-42.667047	-40.6	-0.05091	-5.09124944	-40.3	-0.05874	-5.8735664
456	AP004339.3/129793-129675	119	36.00946	-44.832057	-45.2	0.00814	0.814034208	-44.7	-0.00295	-0.2954284
457	AAAA02022467.1/54175-54057	119	36.00946	-44.832057	-45.2	0.00814	0.814034208	-44.7	-0.00295	-0.2954284
458	ABAV01004466.1/10927-10807	121	29.75427	-35.277372	-37.5	0.05927	5.927007781	-37.5	0.05927	5.9270077
459	CAAJ01000127.1/3421-3534	114	28.48008	-31.852554	-29	-0.09836	-9.83639223	-29.8	-0.06888	-6.8877642
460	CAAI01002944.1/2056-1943	114	28.19527	-31.399339	-29.2	-0.07532	-7.53198227	-29.1	-0.07902	-7.9015079
461	AAJJ01001278.1/15201-15319	119	30.89161	-36.688022	-36.7	0.000326	0.032636644	-37.8	0.029417	2.94173981
462	X15935.1/3-121	119	28.79177	-33.346546	-38.6	0.1361	13.6099848	-38.3	0.129333	12.9333006
463	AC158186.2/4528-4644	117	33.86383	-41.018513	-43.3	0.05269	5.269023366	-43.2	0.050497	5.04973869
464	AC007202.3/14338-14454	117	28.82886	-33.006373	-31.5	-0.04782	-4.78213533	-31.2	-0.0579	-5.7896558
465	AATU01003637.1/103086-103202	117	28.95106	-33.20082	-35.4	0.062124	6.212374226	-36	0.077755	7.7755013
466	Z14994.1/1543-1661	119	33.16031	-40.298195	-45.3	0.110415	11.04151171	-45	0.104485	10.448455

467	AADK01028234.1/592-479	114	28.70941	-32.21748	-32.4	0.005633	0.563332044	-32.4	0.005633	0.5633320
468	AASM01000961.1/2142-2261	120	29.12159	-34.070992	-37.2	0.084113	8.411311147	-36.7	0.071635	7.1635088
469	AARH01001853.1/639606-639489	118	34.70654	-42.55912	-42.9	0.007946	0.794591731	-42.6	0.00096	0.0959620
470	AC146755.23/40564-40682	119	34.79023	-42.891888	-43.1	0.004829	0.482857797	-43.2	0.007132	0.7132215
471	AAZO01006170.1/89711-89592	120	29.50004	-34.673207	-36.3	0.044815	4.481524347	-37.4	0.072909	7.29089127
472	AAZX01000523.1/17560-17680	121	33.63542	-41.45345	-40	-0.03634	-3.63362403	-40	-0.03634	-3.6336240
473	X74440.1/1-120	120	31.0164	-37.086191	-37.7	0.016281	1.628141663	-39.4	0.058726	5.8726127
474	EF647601.1/94040-94158	119	32.6885	-39.547403	-37.3	-0.06025	-6.02520987	-38.5	-0.02721	-2.7205280
475	AABL01000519.1/10755-10640	116	28.77741	-32.724887	-29	-0.12844	-12.8444383	-28.8	-0.13628	-13.628080
476	AAKM01000004.1/1465347-1465233	115	28.63794	-32.303353	-34.7	0.069068	6.906763313	-34.4	0.060949	6.0949036
477	AAWU01000380.1/47408-47286	123	25.95488	-29.630607	-30.2	0.018854	1.885407013	-29.8	0.005684	0.56843261
478	AAPT01020503.1/50561-50682	122	22.83785	-24.470877	-23	-0.06395	-6.39511646	-22.2	-0.10229	-10.229174
479	CR855038.1/57351-57233	119	33.67151	-41.111668	-39.6	-0.03817	-3.81734466	-40.4	-0.01762	-1.7615556
480	AAGE02022633.1/4541-4666	126	29.83394	-36.402149	-32.9	-0.10645	-10.6448309	-32.5	-0.12007	-12.006613
481	AAQX01002881.1/912-1027	116	28.81244	-32.780631	-32.2	-0.01803	-1.8032015	-31.7	-0.03409	-3.4089302

482	AACY020397167.1/1194-1084	111	28.3058	-30.976416	-35	0.11496	11.49595476	-35.4	0.12496	12.4960004
483	AACT01019277.1/3233-3346	114	29.40381	-33.322484	-33.4	0.002321	0.232084975	-32.9	-0.01284	-1.2841447
484	AASV01046355.1/666-787	122	22.05026	-23.217572	-23.8	0.024472	2.447177009	-23.4	0.007796	0.7796073
485	AAEU02000660.1/100051-99931	121	25.02416	-27.750341	-23.8	-0.16598	-16.5980712	-24.8	-0.11897	-11.896536
486	AAPQ01007319.1/728227-728345	119	29.26081	-34.092925	-29.7	-0.14791	-14.7909919	-30.2	-0.1289	-12.89047
487	AAPP01015704.1/527093-527216	124	30.73464	-37.436227	-37.3	-0.00365	-0.36522038	-38.2	0.019994	1.99940523
488	AACW02000228.1/444426-444540	115	23.58505	-24.262695	-23.3	-0.04132	-4.13173848	-23.9	-0.01518	-1.5175525
489	AAAB01008944.1/3738569-3738447	123	30.05569	-36.156213	-31.6	-0.14418	-14.4183961	-30.7	-0.17773	-17.772681
490	AAKO01001397.1/16278-16396	119	23.17452	-24.407812	-24.2	-0.00859	-0.85872576	-24.7	0.011829	1.18294884
491	AAPU01010615.1/222531-222648	118	23.26154	-24.346685	-24	-0.01445	-1.44452276	-24.9	0.022221	2.2221467
492	X01693.1/1-112	112	28.34027	-31.230879	-29.9	-0.04451	-4.45110087	-29.4	-0.06227	-6.2274801
493	AAYL01000045.1/86345-86236	110	31.93012	-36.544204	-36.3	-0.00673	-0.67273813	-37.6	0.02808	2.8079682
494	AANI01017247.1/37618-37740	123	29.92121	-35.942218	-33.9	-0.06024	-6.02424277	-33.5	-0.0729	-7.2902038
495	AY462110.1/1391-1506	116	30.67206	-35.739847	-40.3	0.113155	11.31551566	-40.6	0.119708	11.970819
496	AY462110.1/1391-1506	118	35.36508	-43.607046	-46.7	0.06623	6.623028817	-50.3	0.133061	13.306072

497	AF271469.1/310-194	117	29.89378	-34.700968	-30.1	-0.15286	-15.2856082	-30.5	-0.13774	-13.773665
498	AAQB01006740.1/366264-366386	123	27.37269	-31.886754	-30.1	-0.05936	-5.93605912	-30.7	-0.03866	-3.865647
499	AADE01000447.1/59086-59201	116	23.21227	-23.86908	-23.3	-0.02442	-2.44240296	-24.6	0.029712	2.9712199
500	AB202073.1/931-819	113	30.25059	-34.470364	-26.5	-0.30077	-30.0768467	-28.1	-0.2267	-22.670335
501	AABS01000112.1/71173-71286	114	29.38665	-33.295177	-34	0.02073	2.073009505	-33.5	0.006114	0.6114126
502	AAIZ01003066.1/82049-82166	118	25.33071	-27.63936	-27.5	-0.00507	-0.50676421	-27.2	-0.01615	-1.6152946
503	AAQA01000616.1/21218-21104	115	29.69287	-33.982068	-33.5	-0.01439	-1.43900813	-33.8	-0.00539	-0.5386619
504	AY705674.1/1315-1203	113	28.67537	-31.963715	-32.1	0.004246	0.424564965	-31.7	-0.00832	-0.8319074
505	Z69659.1/6667-6789	123	29.25678	-34.88491	-32.1	-0.08676	-8.67573107	-32.4	-0.07669	-7.6694743
506	CAAL01001847.1/78328-78213	116	28.33587	-32.022276	-33.1	0.03256	3.255964227	-31.6	-0.01336	-1.3363159
507	L22251.1/140-257	118	29.82513	-34.791332	-39.7	0.123644	12.36440414	-41.6	0.16367	16.366991
508	AAXJ01002433.1/2315-2426	112	28.07198	-30.803941	-32.7	0.057983	5.798345373	-32.3	0.046318	4.6317614
509	AAXT01000002.1/226294-226405	112	33.30301	-39.128087	-40.8	0.040978	4.09782552	-40.5	0.033874	3.3874390
510	AAFP01000428.1/22975-22864	112	27.90976	-30.545808	-29.3	-0.04252	-4.25190534	-29.1	-0.04968	-4.9684132
511	AE017344.1/587581-587470	112	27.90976	-30.545808	-29.3	-0.04252	-4.25190534	-29.1	-0.04968	-4.9684132
512	AAEY01000021.1/77428-77539	112	27.90976	-30.545808	-29.3	-0.04252	-4.25190534	-29.1	-0.04968	-4.9684132

513	AAFI02000148.1/105951-105837	115	27.76152	-30.9087	-30.6	-0.01009	-1.00882335	-30.6	-0.01009	-1.0088233
514	DQ001173.1/3-117	115	27.76152	-30.9087	-30.6	-0.01009	-1.00882335	-30.6	-0.01009	-1.0088233
515	AAGK01000001.1/548506-548395	112	28.62368	-31.681869	-33.5	0.054273	5.427256204	-33.4	0.051441	5.1441042
516	AACO02000021.1/6655-6544	112	25.14378	-26.144289	-25.4	-0.0293	-2.93027237	-24.9	-0.04997	-4.9971453
517	AATT01000070.1/145415-145303	113	28.78073	-32.131374	-29.2	-0.10039	-10.0389518	-28.7	-0.11956	-11.956006
518	X00386.1/1-104	104	27.14935	-27.738966	-24.1	-0.15099	-15.0994429	-24.1	-0.15099	-15.099442
519	AAWT01083394.1/3003-2887	117	25.63197	-27.919157	-24	-0.1633	-16.3298193	-23.5	-0.18805	-18.804921
520	AF095839.1/891-776	116	28.8649	-32.86412	-33.5	0.018982	1.898150182	-33.2	0.010117	1.0116876
521	X16573.1/258-372	115	24.74342	-26.106005	-25.9	-0.00795	-0.79538503	-26.1	-0.00023	-0.0230066
522	AAFU01001022.1/48358-48247	112	28.46461	-31.42874	-32.5	0.032962	3.296185191	-32.5	0.032962	3.2961851
523	ABCN01002017.1/34559-34665	107	29.0671	-31.389474	-32.3	0.02819	2.81896548	-32.9	0.045913	4.5912639
524	AAFB02000352.1/4039-4148	110	27.45639	-29.425149	-25.8	-0.14051	-14.0509664	-25.6	-0.14942	-14.941989
525	AACM02000265.1/29897-29783	115	25.37629	-27.113096	-28.1	0.035121	3.512115091	-28.3	0.04194	4.19400827
526	CR382132.1/1370879-1370995	117	30.17921	-35.155171	-36.2	0.028863	2.886268735	-37.7	0.067502	6.75021029
527	AANV02000637.1/5704-5595	110	27.91205	-30.150246	-26.2	-0.15077	-15.0772753	-26	-0.15962	-15.962485
528	AATM01000006.1/4956-4838	119	29.51816	-34.502455	-31.6	-0.09185	-9.18498458	-32.6	-0.05836	-5.8357519

529	AAIM02000161.1/229078-228967	112	28.35807	-31.259198	-31.1	-0.00512	-0.51189094	-31.3	0.001304	0.1303575
530	AARE01001511.1/1260-1149	112	28.054	-30.775333	-29.8	-0.03273	-3.2729307	-30.4	-0.01235	-1.2346491
531	AAJI01001476.1/13492-13374	119	29.07063	-33.790298	-29.6	-0.14156	-14.1564124	-29.6	-0.14156	-14.156412
532	AF529186.1/1009-894	116	29.50431	-33.88161	-34.6	0.020763	2.076272944	-35.1	0.034712	3.47119783
533	AAEE01000007.1/707570-707686	117	29.22844	-33.642217	-35	0.038794	3.879379721	-35.3	0.046963	4.69626884
534	AAEL01000435.1/2854-2970	117	29.17663	-33.559775	-36.4	0.078028	7.802816763	-36.1	0.070366	7.03663518
535	AAWC01002764.1/14312-14199	114	33.3523	-39.605715	-40.8	0.029272	2.927168688	-41.4	0.04334	4.33402131
536	AAFT01000039.1/59328-59208	121	25.97374	-29.26141	-28.8	-0.01602	-1.60211835	-29.6	0.011439	1.14388485
537	Z11883.1/951-834	118	29.92641	-34.952489	-29	-0.20526	-20.5258242	-29.8	-0.1729	-17.2902315
538	AAKE03000008.1/776195-776083	113	28.35694	-31.456999	-34.3	0.082886	8.288631688	-34.3	0.082886	8.28863168
539	AAHF01000006.1/1883955-1883843	113	28.35694	-31.456999	-34.3	0.082886	8.288631688	-34.3	0.082886	8.28863168
540	ABDB01000059.1/258270-258158	113	28.35694	-31.456999	-34.3	0.082886	8.288631688	-34.3	0.082886	8.28863168
541	AATX01000063.1/152569-152455	115	28.9358	-32.777343	-30.8	-0.0642	-6.4199464	-31.1	-0.05393	-5.3933874
542	ABBB01000033.1/420348-420462	115	28.9358	-32.777343	-26.4	-0.24157	-24.1566041	-26.9	-0.21849	-21.848860
543	AASO01001658.1/2059-1945	115	28.9358	-32.777343	-30.8	-0.0642	-6.4199464	-31.1	-0.05393	-5.3933874

544	ABBC01000392.1/6892-7006	115	28.9358	-32.777343	-30.8	-0.0642	-6.4199464	-31.1	-0.05393	-5.3933874
545	AAXI01000301.1/8428-8544	117	28.93363	-33.173092	-37.4	0.113019	11.30189392	-37.4	0.113019	11.3018939
546	AAVQ01000002.1/3936-3813	124	22.52299	-24.369034	-25.5	0.044352	4.435158968	-25.5	0.044352	4.43515896
547	AAGI01000327.1/72210-72324	115	28.54974	-32.163003	-34.3	0.062303	6.23031172	-34.3	0.062303	6.23031172
548	AAKD03000017.1/303987-303868	120	29.277	-34.318294	-39.4	0.128977	12.89773202	-40.2	0.146311	14.6311104
549	AM270115.1/128938-128823	116	28.24674	-31.880436	-31.4	-0.0153	-1.53004993	-31.4	-0.0153	-1.5300499
550	AANW02001233.1/8188-8079	110	24.83965	-25.261128	-25.5	0.009368	0.936752275	-25.5	0.009368	0.93675227
551	AP007155.1/1026766-1026646	121	29.80506	-35.358194	-36.2	0.023254	2.325431727	-36.6	0.033929	3.39291334
552	AAIW01000368.1/11893-11779	115	25.79602	-27.781	-26.8	-0.0366	-3.66044758	-27.4	-0.01391	-1.3905107
553	AAJN01000094.1/53887-53768	120	29.5342	-34.727571	-39.8	0.127448	12.74479702	-40.6	0.144641	14.464111
554	AANS01001320.1/16875-16995	121	26.75427	-30.503472	-29.7	-0.02705	-2.70529321	-30.1	-0.0134	-1.3404388
555	AAFM01000003.1/226344-226452	109	25.55664	-26.202475	-24.5	-0.06949	-6.94887949	-24.5	-0.06949	-6.9488794
556	AL590450.1/114088-114197	110	28.46695	-31.03325	-27.5	-0.12848	-12.8481818	-27.9	-0.1123	-11.230286
557	AC004395.1/1702-1592	111	28.5012	-31.287353	-29.3	-0.06783	-6.78277621	-30.6	-0.02246	-2.246253
558	AACD01000010.1/103824-103714	111	28.5012	-31.287353	-29.3	-0.06783	-6.78277621	-30.6	-0.02246	-2.2462530

559	AAPO01000006.1/285980-286104	125	25.53336	-29.359033	-29.1	-0.0089	-0.89014659	-29.5	0.004779	0.4778553
560	AAIH02000216.1/25877-25996	120	29.70443	-34.998454	-37.1	0.056645	5.664545676	-37.9	0.076558	7.6557953
561	AAGT01000497.1/29058-28944	115	23.07031	-23.44359	-22.7	-0.03276	-3.27572774	-22	-0.06562	-6.5617736
562	X87329.1/2199-2085	115	33.4258	-39.922281	-41	0.026286	2.628581897	-41.9	0.047201	4.7200920
563	AAFO01000011.1/73327-73206	122	23.01553	-24.753609	-24.3	-0.01867	-1.86670225	-25.2	0.017714	1.7713942
564	AACQ01000039.1/46591-46712	122	22.41243	-23.793896	-24.3	0.020827	2.082731036	-25.3	0.05953	5.9529788
565	CR382125.1/1422292-1422448	157	34.49024	-49.999325	-47.1	-0.06156	-6.1556796	-47.8	-0.04601	-4.6010985
566	AAID01003647.1/370-250	121	26.03389	-29.357126	-28	-0.04847	-4.84687988	-30.4	0.034305	3.4305053
567	AC091619.3/67508-67375	134	20.02881	-22.396043	-22.7	0.01339	1.339019711	-23.5	0.046977	4.69769138
568	AACA01000117.1/2098-2219	122	23.8885	-26.142767	-25.4	-0.02924	-2.92427915	-26.7	0.02087	2.08701534
569	AACG02000018.1/45887-45766	122	23.8885	-26.142767	-25.4	-0.02924	-2.92427915	-26.7	0.02087	2.08701534
570	AC189540.1/19288-19170	119	30.32703	-35.789605	-38.2	0.063099	6.309935752	-37.6	0.048149	4.81488153
571	AACI02000988.1/1421-1535	115	22.77849	-22.979219	-21.6	-0.06385	-6.38527101	-24.1	0.046505	4.65054548
572	AACH01000658.1/4206-4082	125	24.03694	-26.977782	-28.4	0.050078	5.00780978	-29.4	0.082388	8.23883665
573	DQ028748.1/5155-5269	115	20.06868	-18.667093	-19	0.017521	1.752143375	-22.4	0.166648	16.6647644

574	AADM01000037.1/36531-36659	129	25.09337	-29.457279	-28.9	-0.01928	-1.92830143	-33.4	0.118046	11.8045535
575	AABY01000025.1/44800-44680	121	24.41343	-26.778495	-26.4	-0.01434	-1.43369253	-27.1	0.011864	1.18636594
576	CR380948.1/197408-197529	122	22.62166	-24.126845	-24.4	0.011195	1.119489014	-24.6	0.019234	1.92339560
577	AAFD02000027.1/316326-316217	110	32.62603	-37.6516	-35.7	-0.05467	-5.46666691	-38.7	0.02709	2.70904370
578	AAKN02051087.1/45335-45220	116	28.72478	-32.641144	-30	-0.08804	-8.80381335	-30.5	-0.0702	-7.0201442
579	AAGV020174570.1/3903-4018	116	29.33287	-33.6088	-31	-0.08415	-8.41548234	-32.4	-0.03731	-3.7308627
580	AASC02039457.1/2620-2506	115	28.54974	-32.163003	-30.9	-0.04087	-4.08738861	-31.1	-0.03418	-3.4180163
581	ABIS01000186.1/30072-30186	115	28.9358	-32.777343	-30.8	-0.0642	-6.4199464	-31.1	-0.05393	-5.3933874
582	AABX02000002.1/78945-78831	115	33.37252	-39.837494	-37.6	-0.05951	-5.95078226	-37.6	-0.05951	-5.9507822
583	AP009663.1/4379-4498	120	25.91308	-28.965289	-27.5	-0.05328	-5.32832536	-28.2	-0.02714	-2.7137924
584	AAWR02001149.1/42020-41905	116	28.81244	-32.780631	-31.1	-0.05404	-5.40395783	-31.7	-0.03409	-3.4089302
585	BX890568.8/147618-147503	116	33.17772	-39.7271	-41.2	0.03575	3.574999997	-40.7	0.023904	2.39041768
586	AC110235.13/80163-80048	116	33.33287	-39.974	-41.2	0.029757	2.975729305	-40.7	0.017838	1.78378494
587	AAGD02000134.1/1661-1782	122	29.1369	-34.494542	-30.8	-0.11995	-11.9952655	-31.5	-0.09506	-9.5064817
588	AM055942.3/1616414-1616305	110	31.93012	-36.544204	-36.3	-0.00673	-0.67273813	-37.6	0.02808	2.8079682

589	CAAC02000457.1/1551549-1551428	122	29.41243	-34.932996	-30.8	-0.13419	-13.4188193	-32.8	-0.06503	-6.5030376
590	CR382134.2/219036-219157	122	24.01553	-26.344909	-25.2	-0.04543	-4.54328828	-26.7	0.013299	1.32993016
591	AAFN02000018.1/172560-172433	128	21.91642	-24.202198	-23.4	-0.03428	-3.4281947	-24.7	0.020154	2.01539449
592	AE016817.6/459194-459336	143	32.2689	-43.670095	-45.8	0.046504	4.650447123	-45.8	0.046504	4.65044712
593	ABDG02000015.1/1095640-1095526	115	26.28235	-28.554907	-29.8	0.041782	4.178165155	-29.8	0.041782	4.17816515
594	AADG06003819.1/2345-2227	119	30.00946	-35.284257	-35.6	0.008869	0.886919837	-35.8	0.014406	1.44062419
595	AFQF01001308.1/86869-86981	113	28.25	-31.286825	-31.9	0.019222	1.922178683	-32.1	0.025333	2.53325545
596	AACU03000132.1/1014141-1014032	110	28.14608	-30.522658	-25.5	-0.19697	-19.6966962	-27.2	-0.12216	-12.215652
597	CAAB02011239.1/13744-13857	114	29.07604	-32.800898	-37.3	0.120619	12.0619364	-38.2	0.141338	14.1337756
598	AACN010031598.1/2460-2578	119	29.70548	-34.800523	-35.5	0.019704	1.970358829	-35.7	0.025195	2.51954449
599	AAWZ02006407.1/13458-13573	116	29.65776	-34.12579	-36.6	0.067601	6.760135843	-36.9	0.075182	7.51818351
600	AACS02000007.1/597289-597173	117	28.89875	-33.117586	-39.2	0.155164	15.51636271	-38.8	0.146454	14.6453973
601	ABDF02000090.1/29527-29640	114	28.93692	-32.579515	-29.3	-0.11193	-11.1928844	-30.4	-0.07169	-7.1694576
602	AAGW02002965.1/6736-6851	116	28.98695	-33.058339	-32.7	-0.01096	-1.09583866	-34.1	0.030547	3.05472362
603	AAGJ04035404.1/487-369	119	29.00117	-33.67976	-37.6	0.104262	10.42617009	-37.4	0.099472	9.94716565

604	AAIL02000016.1/31650-31761	112	28.37586	-31.287499	-30.8	-0.01583	-1.58278976	-31.7	0.013013	1.3012642
605	AAQR03026016.1/3436-3551	116	29.45298	-33.799933	-30.5	-0.10819	-10.8194534	-35.6	0.050564	5.05636717
606	AE014187.2/1889353-1889471	119	28.96637	-33.624392	-36.1	0.068576	6.857640925	-35.6	0.055495	5.54946172
607	AAPE02065766.1/2803-2918	116	29.6067	-34.044537	-34.2	0.004546	0.454569224	-34	-0.00131	-0.1309921
608	AACZ03099104.1/4171-4286	116	29.38441	-33.69081	-33.1	-0.01785	-1.78492345	-34.5	0.023455	2.34547924
609	AAQY02000250.1/229279-229396	118	28.75776	-33.092818	-32	-0.03415	-3.41505497	-33.2	0.003228	0.32283858
610	ABEG02003930.1/15093-15214	122	29.20602	-34.604543	-30.7	-0.12718	-12.7183818	-32.2	-0.07468	-7.46752547

V. Rfam snRNA family RF00026

Sl.no.	Sequence ID	NTL	SD_DFT	MFE_C	MFE_F	RD_2	%RD_2	MFE_S	RD_3	%RD_3
611	AB010698.1/46416-46518	103	22.99796	-20.933254	-20	-0.04666	-4.66627137	-20.7	-0.01127	-1.1268322
612	AARH01001853.1/272694-272592	103	22.4727	-20.09741	-17.9	-0.12276	-12.2760324	-18.6	-0.08051	-8.0505903
613	X60506.1/390-492	103	24.56789	-23.431488	-21.3	-0.10007	-10.0069869	-22.2	-0.05547	-5.5472441
614	AC146705.11/15272-15374	103	24.56789	-23.431488	-21.3	-0.10007	-10.0069869	-22.2	-0.05547	-5.5472441
615	AASG02002949.1/2307-2409	103	24.56789	-23.431488	-21.3	-0.10007	-10.0069869	-22.2	-0.05547	-5.5472441
616	AACV01009611.1/38269-38167	103	24.56789	-23.431488	-21.3	-0.10007	-10.0069869	-22.2	-0.05547	-5.5472441

617	AAAA02013555.1/2292-2394	103	24.56789	-23.431488	-21.3	-0.10007	-10.0069869	-22.2	-0.05547	-5.5472441
618	CR855100.1/43897-43999	103	24.56789	-23.431488	-21.3	-0.10007	-10.0069869	-22.2	-0.05547	-5.5472441
619	X52315.1/1-103	103	22.97717	-20.900165	-19.2	-0.08855	-8.85502464	-20.1	-0.03981	-3.9809190
620	X51447.1/262-364	103	23.41542	-21.597561	-21.3	-0.01397	-1.39699877	-21.8	0.009286	0.9286204
621	AAXJ01018701.1/864-762	103	25.02014	-24.151157	-21.7	-0.11296	-11.2956531	-23.2	-0.041	-4.0998134
622	AAQA01000086.1/111608-111711	104	23.08723	-21.274913	-18.9	-0.12566	-12.565676	-19.2	-0.10807	-10.806837
623	L26849.1/150-253	104	22.08723	-19.683613	-18.9	-0.04146	-4.14609929	-19.2	-0.02519	-2.5188164
624	X51387.1/158-259	102	21.89545	-18.979227	-17.8	-0.06625	-6.62487266	-18.1	-0.04858	-4.8576095
625	AANU01133867.1/371-269	103	23.92655	-22.410923	-22.9	0.021357	2.13570924	-24.4	0.08152	8.1519566
626	X63066.1/363-465	103	21.4727	-18.50611	-17.9	-0.03386	-3.38608828	-20.7	0.105985	10.5985033
627	AAAB01008807.1/5121648-5121542	107	27.294	-28.567937	-25.5	-0.12031	-12.0311262	-25.9	-0.10301	-10.300915
628	AAWU01008690.1/11499-11605	107	24.10845	-23.498772	-23.1	-0.01726	-1.72628774	-23.7	0.008491	0.84906131
629	AAGE02013372.1/83708-83814	107	24.10845	-23.498772	-23.1	-0.01726	-1.72628774	-23.7	0.008491	0.84906131
630	AABU01002774.1/16804311-16804417	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
631	AAGH01007200.1/1053-947	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
632	AADE01001799.1/12821-12927	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022

633	AAEU02000254.1/81013-80907	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
634	AAPQ01001284.1/654-760	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
635	AAPP01016905.1/3608-3714	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
636	AAQB01008633.1/461913-461807	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
637	AANI01001778.1/524-630	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
638	AAIZ01003051.1/11867-11761	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
639	AAPT01019380.1/23511-23405	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
640	AAPU01011102.1/63061-63167	107	22.10845	-20.316172	-23.1	0.120512	12.05120144	-23.5	0.135482	13.5482022
641	AABS01000051.1/274751-274857	107	22.90845	-21.589212	-23.3	0.073424	7.342435761	-23.3	0.073424	7.34243576
642	AAHY01132583.1/7739-7845	107	27.294	-28.567937	-25.3	-0.12917	-12.9167478	-25.3	-0.12917	-12.916747
643	AAFC03115769.1/4330-4436	107	23.62485	-22.729216	-24.9	0.08718	8.718006998	-24.9	0.08718	8.71800699
644	AAPY01153714.1/3278-3172	107	23.62485	-22.729216	-24.9	0.08718	8.718006998	-24.9	0.08718	8.71800699
645	AALT01529386.1/701-595	107	25.53334	-25.766207	-24.9	-0.03479	-3.47874161	-24.9	-0.03479	-3.4787416
646	AAYZ01436191.1/1883-1777	107	25.53334	-25.766207	-24.9	-0.03479	-3.47874161	-24.9	-0.03479	-3.4787416
647	BAAB01141103.1/711-817	107	24.10845	-23.498772	-24.9	0.056274	5.627419809	-24.9	0.056274	5.627419809

648	AAZX01000860.1/13921-13815	107	24.10845	-23.498772	-24.9	0.056274	5.627419809	-24.9	0.056274	5.6274198
649	AACY022149992.1/529-635	107	24.10845	-23.498772	-24.9	0.056274	5.627419809	-24.9	0.056274	5.62741980
650	AATU01006594.1/28185-28081	105	27.18526	-27.995707	-26.1	-0.07263	-7.26324587	-27.5	-0.01803	-1.8025715
651	ABAV01046662.1/5606-5712	107	27.294	-28.567937	-26.9	-0.06201	-6.20051	-27.2	-0.05029	-5.0291808
652	AAFR03037834.1/2280-2178	103	26.35802	-26.280114	-27.4	0.040872	4.08717666	-27.4	0.040872	4.08717666
653	ABDC01230141.1/8073-7967	107	26.53334	-27.357507	-26	-0.05221	-5.22117947	-26.4	-0.03627	-3.6269191
654	AANN01390182.1/1050-1156	107	27.53334	-28.948807	-26.6	-0.0883	-8.83010023	-27	-0.07218	-7.2178024
655	M31687.1/705-811	107	27.53334	-28.948807	-26.6	-0.0883	-8.83010023	-27	-0.07218	-7.2178024
656	AACT01041609.1/35558-35664	107	23.34942	-22.290926	-21.9	-0.01785	-1.78505017	-22.3	0.000407	0.04069064
657	AANG01770494.1/575-681	107	24.07118	-23.439477	-21.9	-0.0703	-7.02957384	-21	-0.11617	-11.616555
658	AAQQ01629113.1/2249-2143	107	25.53334	-25.766207	-26.6	0.031346	3.134561423	-26.6	0.031346	3.1345614
659	ABBA01062195.1/38876-38770	107	27.44153	-28.802712	-26.6	-0.08281	-8.28087043	-27	-0.06677	-6.6767093
660	AAIY01587713.1/2016-2122	107	27.44153	-28.802712	-26.6	-0.08281	-8.28087043	-27	-0.06677	-6.6767093
661	AAPN01022574.1/939-833	107	27.44153	-28.802712	-26.6	-0.08281	-8.28087043	-27	-0.06677	-6.6767093
662	CR956385.13/177771-177665	107	27.44153	-28.802712	-26.6	-0.08281	-8.28087043	-27	-0.06677	-6.6767093

663	U43841.1/336-439	104	20.35678	-16.929937	-15.3	-0.10653	-10.6531814	-16.4	-0.03231	-3.2313217
664	AANV02000039.1/44432-44535	104	20.35678	-16.929937	-15.3	-0.10653	-10.6531814	-16.4	-0.03231	-3.2313217
665	AAFB02000174.1/3561-3664	104	20.35678	-16.929937	-15.3	-0.10653	-10.6531814	-16.4	-0.03231	-3.2313217
666	CAAI01006173.1/984-1090	107	24.21993	-23.676173	-23.4	-0.0118	-1.18022576	-23.4	-0.0118	-1.1802257
667	BAAF04101838.1/670-565	106	27.10965	-28.074992	-26.6	-0.05545	-5.54508351	-26.6	-0.05545	-5.545083
668	AAVX01043085.1/4211-4317	107	27.56998	-29.007108	-27.2	-0.06644	-6.64378122	-27.2	-0.06644	-6.6437812
669	AC148181.3/26639-26533	107	27.34942	-28.656126	-26.5	-0.08136	-8.13632448	-26.5	-0.08136	-8.1363244
670	CT573239.9/51807-51913	107	27.44153	-28.802712	-26.3	-0.09516	-9.51601344	-26.7	-0.07875	-7.87532409
671	AAYL01000007.1/235430-235323	108	27.47709	-29.058899	-26.6	-0.09244	-9.24397936	-26.6	-0.09244	-9.2439793
672	CAAE01010022.1/48689-48584	106	24.42435	-23.801876	-22.6	-0.05318	-5.31803473	-22.6	-0.05318	-5.3180347
673	AF529186.1/459-564	106	26.94156	-27.807503	-25.9	-0.07365	-7.36487784	-25.9	-0.07365	-7.3648778
674	AAXT01000001.1/1039074-1039179	106	24.979	-24.684489	-26.6	0.072012	7.201167837	-26.6	0.072012	7.2011678
675	DQ666642.1/4-106	103	22.68167	-20.429949	-18.3	-0.11639	-11.6390653	-18.3	-0.11639	-11.639065
676	AAWT01067003.1/344-451	108	27.87824	-29.697243	-24.1	-0.23225	-23.2250755	-24.1	-0.23225	-23.225075
677	AANH01010141.1/91130-91025	106	24.05374	-23.212114	-21.8	-0.06478	-6.47758684	-21.8	-0.06478	-6.4775868
678	AB220565.1/723-829	107	27.18282	-28.39102	-26.6	-0.06733	-6.7331587	-26.6	-0.06733	-6.7331586

679	AAGK01000002.1/918046-917941	106	22.43086	-20.629631	-20.9	0.012936	1.293631553	-20.9	0.012936	1.2936315
680	AC136964.2/84202-84308	107	23.21993	-22.084873	-20.1	-0.09875	-9.8749892	-19.4	-0.1384	-13.839550
681	X71486.1/1-101	101	23.03286	-20.589596	-16.8	-0.22557	-22.5571205	-16.8	-0.22557	-22.557120
682	AACM02000382.1/262414-262518	105	19.09179	-15.11656	-14.2	-0.06455	-6.45464845	-15.2	0.005489	0.5489468
683	AC146661.3/155499-155393	107	20.08368	-17.094164	-17.1	0.000341	0.034129267	-17.1	0.000341	0.03412926
684	AC087806.3/115795-115689	107	26.84652	-27.855874	-26	-0.07138	-7.13797533	-26	-0.07138	-7.1379753
685	L25920.1/3-109	107	27.3863	-28.714819	-26.7	-0.07546	-7.54613997	-26.7	-0.07546	-7.5461399
686	CU326409.1/117472-117365	108	22.80773	-21.628535	-20.6	-0.04993	-4.99289001	-19.1	-0.13238	-13.238404
687	AC188110.1/71045-71151	107	25.64311	-25.94088	-27.4	0.053253	5.325254243	-27.4	0.053253	5.3252542
688	CR382132.1/1089192-1089093	100	21.99534	-18.738981	-14.7	-0.27476	-27.4760591	-15.2	-0.23283	-23.282767
689	AF095841.1/1-107	107	23.89756	-23.16318	-23.6	0.018509	1.850934091	-23.7	0.022651	2.26506517
690	AASM01002098.1/3099-2992	108	23.69889	-23.04664	-22.3	-0.03348	-3.34816304	-21.5	-0.07194	-7.1936760
691	X58843.1/3-106	104	23.05622	-21.225557	-16.6	-0.27865	-27.8648002	-17.5	-0.21289	-21.2888962
692	CAAJ01009065.1/446-339	108	23.2737	-22.370032	-23.7	0.056117	5.611679388	-22.3	-0.00314	-0.3140447
693	AAFU01001153.1/10970-11070	101	22.47527	-19.7023	-19.8	0.004934	0.493436829	-19.9	0.009935	0.9934698
694	AABL01000365.1/9537-9644	108	23.32719	-22.455163	-24.1	0.06825	6.825047981	-23.7	0.052525	5.25247495

695	AADS01000210.1/17114-17228	115	28.9358	-32.777343	-27.1	-0.2095	-20.949607	-28.3	-0.15821	-15.821001
696	AAPO01000024.1/134030-133927	104	20.00761	-16.374316	-14.3	-0.14506	-14.5057056	-14	-0.16959	-16.959399
697	AAWC01002368.1/53508-53403	106	23.84447	-22.879099	-22.5	-0.01685	-1.68488317	-23.4	0.022261	2.2260738
698	CT990557.10/71286-71393	108	22.16306	-20.602676	-19	-0.08435	-8.43513595	-19.2	-0.07306	-7.3056032
699	AAFM01000021.1/681699-681594	106	20.48809	-17.538102	-15.3	-0.14628	-14.6281157	-14.4	-0.21792	-21.792372
700	AAFT01000065.1/268653-268548	106	20.04611	-16.83478	-15.4	-0.09317	-9.31675028	-15.4	-0.09317	-9.3167502
701	AAEY01000056.1/129419-129306	114	22.88948	-22.956227	-22	-0.04346	-4.34648671	-21.4	-0.07272	-7.2720891
702	AACO02000104.1/123353-123240	114	22.88948	-22.956227	-22	-0.04346	-4.34648671	-21.4	-0.07272	-7.2720891
703	AE017348.1/872264-872377	114	22.88948	-22.956227	-22	-0.04346	-4.34648671	-21.4	-0.07272	-7.2720899
704	AAFP01000223.1/16410-16523	114	22.88948	-22.956227	-22	-0.04346	-4.34648671	-21.4	-0.07272	-7.2720891
705	CP000496.1/1065610-1065505	106	20.62115	-17.749839	-16.7	-0.06286	-6.28645865	-16.5	-0.07575	-7.5747793
706	AAID01000554.1/5500-5599	100	18.34609	-12.931932	-11.9	-0.08672	-8.67169656	-11.7	-0.10529	-10.52933
707	X14196.1/133-285	153	22.47914	-30.08766	-28.4	-0.05942	-5.94246628	-28.9	-0.0411	-4.1095516
708	AAZN01000268.1/119544-119646	103	20.43453	-16.854066	-15.9	-0.06	-6.00041324	-15.9	-0.06	-6.0004132
709	AACW02000046.1/25302-25425	124	24.87427	-28.110629	-25.6	-0.09807	-9.80714398	-27.2	-0.03348	-3.3479002

710	CU104654.2/178130-178232	103	17.73774	-12.562669	-12.3	-0.02136	-2.1355214	-11.7	-0.07373	-7.3732404
711	EF419774.1/1-102	102	22.61295	-20.120994	-20.2	0.003911	0.391117597	-20.2	0.003911	0.3911175
712	AACI02000106.1/4087-4199	113	28.46348	-31.626534	-31.4	-0.00721	-0.72144682	-31.3	-0.01043	-1.0432405
713	CR382126.1/1858703-1858820	118	28.66995	-32.953085	-29.7	-0.10953	-10.9531492	-30.2	-0.09116	-9.1161764
714	AACH01000157.1/7933-7821	113	28.14265	-31.116004	-29	-0.07297	-7.29656559	-28.9	-0.07668	-7.6678339
715	AATM01000137.1/47790-47943	154	17.71006	-22.698211	-21.7	-0.046	-4.60005009	-23.1	0.017393	1.7393468
716	Z73279.1/2843-2955	113	28.28569	-31.343621	-29	-0.08081	-8.08145263	-28.9	-0.08455	-8.4554369
717	AACA01000433.1/2210-2322	113	28.28569	-31.343621	-29	-0.08081	-8.08145263	-28.9	-0.08455	-8.4554369
718	AAEG01000112.1/87347-87459	113	28.28569	-31.343621	-29	-0.08081	-8.08145263	-28.9	-0.08455	-8.4554369
719	AABY01000279.1/8828-8716	113	28.28569	-31.343621	-29	-0.08081	-8.08145263	-28.9	-0.08455	-8.4554369
720	AACG02000194.1/7974-8086	113	28.28569	-31.343621	-29	-0.08081	-8.08145263	-28.9	-0.08455	-8.4554369
721	AACF01000007.1/86925-87036	112	28	-30.6894	-30.5	-0.00621	-0.62098361	-30.4	-0.00952	-0.9519736
722	AADM01000279.1/397-288	110	24.37818	-24.526798	-22.3	-0.09986	-9.98564061	-22.6	-0.08526	-8.5256542
723	AF083031.2/127905-127809	97	22.50899	-18.957548	-17.8	-0.06503	-6.50308078	-17.8	-0.06503	-6.5030807
724	AC144401.2/82868-82974	107	20.39147	-17.583939	-16.6	-0.05927	-5.92734054	-17.1	-0.0283	-2.8300498
725	AANW02001116.1/9	107	23.8841	-23.141765	-24.1	0.039761	3.976079245	-25.5	0.09248	9.24798077

	42-1048									
726	AY953942.1/4-110	107	24.7146	-24.463342	-26	0.059102	5.9102241	-26.2	0.066285	6.6284666
727	AJ416571.1/12089-11984	106	18.21995	-13.928805	-12.2	-0.14171	-14.1705287	-12.2	-0.14171	-14.170528
728	AATT01000229.1/70559-70673	115	23.90092	-24.765333	-23.4	-0.05835	-5.83475844	-22.9	-0.08146	-8.1455610
729	AL590448.1/66612-66503	110	24.02862	-23.970537	-25.9	0.074497	7.4496651	-26.7	0.102227	10.2227088
730	AY136823.1/430-532	103	20.44656	-16.873213	-15.9	-0.06121	-6.12083615	-17.6	0.041295	4.12947189
731	AF305715.1/117-214	98	26.03685	-24.771038	-25.6	0.032381	3.238131162	-26.2	0.054541	5.45405182
732	X82228.1/412-509	98	26.03685	-24.771038	-25.6	0.032381	3.238131162	-26.2	0.054541	5.45405182
733	AAFI02000006.1/10321-10427	107	24.48691	-24.101026	-23.8	-0.01265	-1.26481497	-23.3	-0.03438	-3.4378796
734	AF053588.1/116-223	108	29.92139	-32.948512	-33.6	0.01939	1.938953475	-33.4	0.013518	1.35176158
735	AAJI01001561.1/684-795	112	19.00273	-16.372045	-16.3	-0.00442	-0.44199614	-17.2	0.048137	4.8136897
736	AACQ01000098.1/66592-66693	102	21.28255	-18.003917	-15.8	-0.13949	-13.9488421	-14.1	-0.27687	-27.687354
737	AAFO01000026.1/267271-267372	102	21.28255	-18.003917	-15.8	-0.13949	-13.9488421	-14.1	-0.27687	-27.687354
738	AAIM02000091.1/125555-125392	164	28.84169	-42.407983	-38.5	-0.10151	-10.1506061	-39.6	-0.07091	-7.0908670
739	X78552.1/318-415	98	26.1916	-25.017296	-26	0.037796	3.77963155	-26.2	0.045141	4.5141381
740	X79014.1/475-572	98	26.153	-24.955868	-25.6	0.025161	2.516140384	-26.2	0.047486	4.7485951
741	X78551.1/329-426	98	26.2494	-25.109268	-25.6	0.019169	1.91692179	-26.2	0.041631	4.1630991
742	AC149301.1/92055-	95	25.65006	-23.556739	-26.2	0.100888	10.08878226	-26.7	0.117725	11.7725129

	91961									
743	AAEE01000007.1/553473-553580	108	27.44033	-29.000404	-22.9	-0.26639	-26.6393207	-23.4	-0.23933	-23.93335
744	AAEL01000070.1/7754-7647	108	27.44033	-29.000404	-22.9	-0.26639	-26.6393207	-23.4	-0.23933	-23.933352
745	X82229.1/194-291	98	25.92018	-24.585381	-21.6	-0.13821	-13.8212063	-23.5	-0.04619	-4.6186406
746	AAHK01000939.1/1885-1993	109	27.82269	-29.80844	-30.8	0.032194	3.219350306	-31.4	0.050687	5.06866208
747	CP000581.1/336841-336949	109	30.29033	-33.735197	-34.9	0.033375	3.337544134	-34.9	0.033375	3.33754413
748	AC152105.2/17973-17865	109	30.29033	-33.735197	-34.9	0.033375	3.337544134	-34.9	0.033375	3.33754413
749	AC092562.4/129660-129562	99	22.55414	-19.428606	-20.6	0.056864	5.68637906	-20.2	0.038188	3.81878260
750	AAXI01000109.1/30199-30115	85	24.61973	-19.921171	-20.9	0.046834	4.683390393	-20.8	0.042251	4.22513746
751	DQ103593.1/29389-29298	92	21.2173	-15.904085	-15.2	-0.04632	-4.63214043	-15.6	-0.01949	-1.9492650
752	AP004520.1/55775-55881	103	18.66274	-14.034626	-13.2	-0.06323	-6.32292285	-14.4	0.025373	2.53732072
753	AAHF01000007.1/1651650-1651551	100	22.1889	-19.047	-17.1	-0.11386	-11.3859631	-17.8	-0.07006	-7.0056162
754	U58510.1/7462-7568	107	21.515	-19.371827	-19	-0.01957	-1.95698228	-19	-0.01957	-1.9569822
755	ABAR01000008.1/592107-592006	102	21.8151	-18.851366	-17.6	-0.0711	-7.11003503	-18.9	0.002573	0.2573218
756	AP007171.1/1600650-1600749	100	24.55279	-22.80865	-21.2	-0.07588	-7.58797255	-21.2	-0.07588	-7.5879725
757	AARE01000569.1/1814-1714	101	22.64639	-19.974597	-17.7	-0.12851	-12.8508329	-19.3	-0.03495	-3.4953234

758	AASO01000114.1/1873-1980	108	24.67839	-24.605327	-21.7	-0.13389	-13.3886052	-22.8	-0.07918	-7.9181022
759	AAIW01000495.1/17855-17954	100	22.49566	-19.53515	-18.4	-0.06169	-6.16929271	-18.2	-0.07336	-7.3359882
760	X04788.1/1-98	98	21.90068	-18.189156	-15.9	-0.14397	-14.3972093	-19.8	0.081356	8.1355743
761	AAGI01000262.1/1755-1855	101	19.34482	-14.720815	-15.3	0.037855	3.785521705	-16.5	0.107829	10.7829383
762	AY102720.1/4016-3910	107	17.5559	-13.071696	-11.4	-0.14664	-14.6639987	-11.7	-0.11724	-11.723896
763	X57046.1/441-540	100	26.51472	-25.930672	-22.6	-0.14737	-14.7374855	-23.8	-0.08952	-8.9524022
764	AC148038.3/33473-33580	108	20.78758	-18.413874	-18.1	-0.01734	-1.73411084	-19.2	0.040944	4.09440592
765	AAZY02000001.1/305276-305169	108	27.42194	-28.971128	-24.9	-0.1635	-16.3499116	-26.1	-0.11	-11.000490
766	CR380959.2/1159577-1159689	113	28.0349	-30.944529	-26.9	-0.15035	-15.0354237	-26.8	-0.15465	-15.464660
767	CAAC02000605.1/61448-61549	102	18.73898	-13.956338	-13.3	-0.04935	-4.93486901	-13.8	-0.01133	-1.1328808
768	M10329.1/1-108	108	27.60536	-29.263014	-26.6	-0.10011	-10.0113323	-26.6	-0.10011	-10.011332
769	CR382136.2/1319844-1319739	106	20.20069	-17.080759	-16.6	-0.02896	-2.89613873	-15.5	-0.10198	-10.198445
770	X12565.1/540-652	113	28.28569	-31.343621	-29	-0.08081	-8.08145263	-28.9	-0.08455	-8.4554369
771	AASC02049566.1/23588-23693	106	22.56421	-20.841824	-19.9	-0.04733	-4.7327843	-19.9	-0.04733	-4.7327843
772	AC121317.7/68985-69092	108	27.45872	-29.029661	-27.2	-0.06727	-6.72669582	-27.2	-0.06727	-6.7266958
773	AAPN01095965.1/240-347	108	27.73304	-29.466186	-26.5	-0.11193	-11.1931555	-26.5	-0.11193	-11.193155

774	AAFD02000024.1/69022-69131	110	27.3459	-29.249328	-25.6	-0.14255	-14.2551886	-25.9	-0.12932	-12.931769
775	ABPA01000003.1/180549-180442	108	27.47709	-29.058899	-26.6	-0.09244	-9.24397936	-26.6	-0.09244	-9.2439793
776	AACE03000008.2/905684-905797	114	28.26674	-31.513071	-30.5	-0.03322	-3.32154315	-30.4	-0.03661	-3.6614166
777	FR796420.1/621391-621294	98	26.09499	-24.863556	-25.6	0.028767	2.87673407	-26.2	0.051009	5.100931
778	AAFN02000036.1/11182-11077	106	18.3352	-14.112207	-12.4	-0.13808	-13.8081187	-12.4	-0.13808	-13.808118
779	AAGV020896327.1/1288-1181	108	27.55046	-29.175653	-27.6	-0.05709	-5.70888594	-27.6	-0.05709	-5.7088859
780	AATU01006589.1/7799-7695	105	25.18526	-24.813107	-22.2	-0.11771	-11.770753	-23.6	-0.0514	-5.1402846
781	AFQF01002057.1/12805-12639	167	28.93962	-43.162612	-40.9	-0.05532	-5.53205931	-41.8	-0.0326	-3.2598379
782	CAAA01180605.1/4-106	103	22.73838	-20.52019	-18.5	-0.1092	-10.919948	-18.7	-0.09734	-9.7336384
783	ABDF02000090.1/1201765-1201660	106	21.58951	-19.290784	-18.4	-0.04841	-4.84121794	-20.5	0.058986	5.89861414
784	AACU03000132.1/2986497-2986398	100	21.05356	-17.240327	-16.3	-0.05769	-5.76887529	-17	-0.01414	-1.4136863
785	AAIL02000075.1/131155-131066	90	18.09252	-10.532424	-10.4	-0.01273	-1.27331131	-11.6	0.092032	9.2032381
786	X76546.1/281-387	107	24.20138	-23.646657	-22.9	-0.03261	-3.26050913	-22.9	-0.03261	-3.2605091
787	CAAB02029384.1/647-542	106	25.33217	-25.246485	-22.6	-0.1171	-11.710111	-22.6	-0.1171	-11.710111
788	AAEX03005034.1/14917-14815	103	23.83263	-22.261469	-23	0.03211	3.211004508	-23	0.03211	3.2110045

789	AC242743.1/25579-25465	115	26.9358	-29.594743	-27.1	-0.09206	-9.20569554	-28.3	-0.04575	-4.5750653
790	BAAB01206473.1/1073-967	107	23.10845	-21.907472	-23.3	0.059765	5.976513014	-23.7	0.075634	7.5634073
791	M24606.1/401-507	107	24.10845	-23.498772	-23.1	-0.01726	-1.72628774	-23.5	5.22E-05	0.005223542
792	AL844509.2/1632740-1632633	108	23.69889	-23.04664	-23.3	0.010874	1.087380439	-23.7	0.027568	2.75679174
793	AC188639.6/7118-7225	108	23.44033	-22.635204	-22.3	-0.01503	-1.50315892	-21.5	-0.0528	-5.2800206
VI. Rfam snRNA family RF00283										
Sl No.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_C**	**MFE_F**	**RD_2**	**%RD_2**	**MFE_S**	**RD_3**	**%RD_3**
794	AF357342.1/1-62	62	24.90107	-15.778073	-16.7	0.055205	5.520520884	-17.1	0.077306	7.73056717
795	AY077741.1/1-83	83	26.25462	-22.123571	-21	-0.0535	-5.35033785	-21.9	-0.01021	-1.0208719
796	AC090227.10/15359-15277	83	26.25462	-22.123571	-21	-0.0535	-5.35033785	-21.9	-0.01021	-1.0208719
797	AL592064.14/82927-83010	84	26.90008	-23.350302	-27	0.135174	13.51740054	-27.3	0.144678	14.4677587
798	AANU01101212.1/1480-1562	83	26.46581	-22.459641	-22.4	-0.00266	-0.26625541	-22.8	0.014928	1.49280170
799	ABDC01297764.1/1645-1727	83	26.93963	-23.21364	-21.9	-0.05998	-5.99835728	-21.9	-0.05998	-5.9983572
800	AAYZ01094280.1/2675-2757	83	29.27389	-26.928134	-27.6	0.024343	2.43429559	-26.9	-0.00105	-0.1045889
801	AAFC03053505.1/66323-66241	83	26.78892	-22.973812	-23	0.001139	0.113862761	-22.9	-0.00322	-0.3223212
802	AAPY01347300.1/19	83	27.19672	-23.622745	-24.8	0.04747	4.746995117	-26.4	0.105199	10.5199045

	93-1911									
803	AAIY01249536.1/1104-1186	83	26.63736	-22.732626	-26	0.125668	12.56682319	-26	0.125668	12.566823
804	AANN01109527.1/812-730	83	26.46581	-22.459641	-29.9	0.248841	24.88414311	-30.2	0.256303	25.6303271
805	AANG01025936.1/3125-3205	81	25.79171	-20.987754	-20.6	-0.01882	-1.88230168	-21.4	0.019264	1.92638249
806	AAHX01095967.1/17609-17687	79	25.22006	-19.678877	-19.6	-0.00402	-0.40243329	-19.4	-0.01438	-1.4375099
807	AAKN02020097.1/24241-24159	83	28.02227	-24.936438	-25.9	0.037203	3.720315559	-26	0.040906	4.0906220
808	AAGV020429307.1/439-358	82	26.58227	-22.445372	-23.9	0.060863	6.086307929	-24	0.064776	6.47761498
809	AAGU03013210.1/58113-58031	83	28.59933	-25.854715	-26.4	0.020655	2.065474091	-28.2	0.083166	8.31661404
810	AAGW02032389.1/3932-4014	83	26.98334	-23.283195	-24.4	0.045771	4.577070808	-24	0.029867	2.98668865
811	AAEX03017760.1/16604-16686	83	26.63736	-22.732626	-24.2	0.060635	6.063529044	-25	0.090695	9.06949611
812	AAPE02017035.1/34119-34194	76	24.5134	-17.95557	-19.4	0.074455	7.445516723	-20.4	0.119825	11.9825012
813	AAQR03077543.1/26098-26180	83	29.59933	-27.446015	-27.5	0.001963	0.196309673	-28.1	0.023273	2.327349324
VI. Rfam snRNA family RF00492										
Sl.no.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_C**	**MFE_F**	**RD_2**	**%RD_2**	**MFE_S**	**RD_3**	**%RD_3**
814	AC090227.10/15587-15444	144	39.9353	-56.069251	-63.2	0.112828	11.28283133	-63.2	0.112828	11.2828313

815	AC129097.27/183442-183585	144	32.68669	-44.534527	-39.1	-0.13899	-13.8990455	-39.2	-0.13608	-13.608486
816	AC018751.30/166114-165972	143	37.2071	-51.528253	-52.5	0.018509	1.850946873	-51	-0.01036	-1.0357899
817	AC125020.7/181527-181661	135	34.25922	-45.240494	-43.3	-0.04482	-4.48151011	-44.9	-0.00758	-0.7583381
818	AC127289.4/19294-19156	139	34.72282	-46.776616	-53.2	0.12074	12.07402964	-53.3	0.12239	12.238993
819	AC023490.5/121842-121984	143	35.38856	-48.634412	-48.1	-0.01111	-1.11104286	-46.8	-0.0392	-3.9196829
VII. Rfam snRNA family RF01458										
Sl.no.	**Sequence ID**	**NTL**	**SD_DFT**	**MFE_C**	**MFE_F**	**RD_2**	**%RD_2**	**MFE_S**	**RD_3**	**%RD_3**
820	AB261975.1/7738-7641	98	22.2688	-18.774947	-18.1	-0.03729	-3.72898651	-18.8	0.001333	0.13326298
821	CP000255.1/68682-68585	98	21.36127	-17.330791	-15.9	-0.08999	-8.99868714	-16.5	-0.05035	-5.0350988
822	CP000703.1/63982-64079	98	22.36127	-18.922091	-16.7	-0.13306	-13.3059357	-17	-0.11306	-11.306419
823	AM263198.1/671271-671368	98	21.82052	-18.061591	-17.1	-0.05623	-5.62333859	-17.1	-0.05623	-5.6233385
824	AL591976.1/215482-215385	98	28.12601	-28.095512	-25.3	-0.11049	-11.0494544	-26.1	-0.07646	-7.6456397
825	AL591973.1/172171-172268	98	28.12601	-28.095512	-25.3	-0.11049	-11.0494544	-26.1	-0.07646	-7.6456397
826	AL591974.1/157606-157509	98	28.12601	-28.095512	-25.3	-0.11049	-11.0494544	-26.1	-0.07646	-7.6456397
VIII. Rfam snRNA family RF01475										

Sl.no.	Sequence ID	NTL	SD_DFT	MFE_C	MFE_F	RD_2	%RD_2	MFE_S	RD_3	%RD_3
827	AADR01000003.1/142171-142094	78	21.34496	-13.31283	-13.1	-0.01625	-1.62465592	-12.8	-0.04006	-4.00648379
828	AM263198.1/2127523-2127600	78	21.36664	-13.347339	-11.8	-0.13113	-13.113041	-13.5	0.011308	1.13082346
829	AL596171.1/97397-97474	78	21.53941	-13.622266	-12.7	-0.07262	-7.26193385	-12.6	-0.08113	-8.1132903
830	AL591982.1/53775-53852	78	21.77387	-13.995366	-11.5	-0.21699	-21.6988363	-11.5	-0.21699	-21.698836
831	AADQ01000011.1/3685-3762	78	21.77387	-13.995366	-11.5	-0.21699	-21.6988363	-11.5	-0.21699	-21.698836
832	AARL02000916.1/597-673	77	21.25997	-12.977994	-12.5	-0.03824	-3.82395257	-12.7	-0.02189	-2.1889296

IX. Rfam snRNA family RF01490

Sl.no.	Sequence ID	NTL	SD_DFT	MFE_C	MFE_F	RD_2	%RD_2	MFE_S	(RD_3	%RD_3
833	AY168080.1/216-96	121	27.79179	-32.154468	-31.7	-0.01434	-1.43365312	-31.7	-0.01434	-1.43365311
834	AY510072.1/4036-4154	119	20.77558	-20.590386	-21.1	0.024152	2.415230476	-21.3	0.033315	3.33151939
835	AY510073.1/4042-4162	121	25.31298	-28.20995	-27.4	-0.02956	-2.95602069	-28.3	0.003182	0.31819904
836	AY512490.1/3938-4058	121	27.70093	-32.009894	-31.2	-0.02596	-2.59581423	-31.2	-0.02596	-2.5958142
837	AY512446.2/3933-4053	121	27.24212	-31.279792	-28.3	-0.10529	-10.5293014	-28.2	-0.10921	-10.921249
838	EU372052.1/3947-4067	121	27.79179	-32.154468	-31.7	-0.01434	-1.43365312	-31.7	-0.01434	-1.4336531

839	EU372053.1/3947-4067	121	27.79179	-32.154468	-31.7	-0.01434	-1.43365312	-31.7	-0.01434	-1.4336531
840	FJ041145.1/3922-4042	121	27.70093	-32.009894	-30.5	-0.0495	-4.95047226	-30.5	-0.0495	-4.9504722
841	EU372028.1/3954-4074	121	28.20594	-32.813511	-31.3	-0.04835	-4.835498	-31.3	-0.04835	-4.8354980

X. Rfam snRNA family RF00618

Sl. No	Sequence ID	NTL	SD_DFT	MFE_C	MFE_F	RD_2	%RD_2	MFE_S	RD_3	%RD_3
842	AAPP01019634.1/122532-122382	151	39.30442	-56.462526	-56.9	0.007688	0.768847264	-54.7	-0.03222	-3.2221680
843	AAPQ01006438.1/743481-743332	150	34.37847	-48.424259	-52.7	0.081134	8.113360157	-48.9	0.009729	0.97288507
844	AE014297.2/1020885-1020736	150	34.33452	-48.354317	-48.6	0.005055	0.505520645	-44.9	-0.07693	-7.6933562
845	AAST01029695.1/1294-1145	150	34.15814	-48.073648	-51.4	0.064715	6.471501808	-47.6	-0.00995	-0.9950589
846	AAKO01001557.1/51647-51498	150	34.15814	-48.073648	-51.4	0.064715	6.471501808	-47.6	-0.00995	-0.9950589
847	AAEU02010290.1/66-215	150	34.52458	-48.656757	-45.5	-0.06938	-6.93792722	-42.2	-0.153	-15.300371
848	AADA01241850.1/15965-15840	126	26.10302	-30.465141	-26.3	-0.15837	-15.8370368	-26	-0.17174	-17.173617
849	AL389925.10/20736-20611	126	28.01919	-33.514343	-25.6	-0.30915	-30.9154031	-25.4	-0.31946	-31.946233

850	AADA01047294.1/887-1012	126	30.92882	-38.144434	-34.5	-0.10564	-10.5635759	-31.8	-0.19951	-19.951049
851	AL135914.25/92223-92098	126	30.92882	-38.144434	-34.5	-0.10564	-10.5635759	-31.8	-0.19951	-19.951049
852	AL161445.10/77816-77941	126	30.13649	-36.883596	-31.7	-0.16352	-16.3520371	-29.8	-0.2377	-23.770455
853	AADA01054074.1/15964-15839	126	30.13649	-36.883596	-31.1	-0.18597	-18.5967709	-29.7	-0.24187	-24.187191
854	AC136636.6/172138-172014	125	31.02743	-38.101755	-35	-0.08862	-8.86215572	-35.1	-0.08552	-8.5520071
855	AAXN01018884.1/118-242	125	30.73364	-37.634237	-35	-0.07526	-7.52639078	-35.1	-0.0722	-7.2200477
856	AACN010332835.1/830-706	125	30.91351	-37.92047	-35.3	-0.07423	-7.42342784	-35.2	-0.07729	-7.7286080
857	AAFC03121196.1/26240-26365	126	30.41947	-37.333899	-40.2	0.071296	7.129605585	-39.3	0.050028	5.00280266
858	AAFR03008173.1/71752-71627	126	30.01919	-36.696943	-32	-0.14678	-14.6779475	-31.6	-0.1613	-16.129567
859	AAPN01427475.1/38-163	126	35.41388	-45.281511	-43.5	-0.04095	-4.0954273	-42.6	-0.06295	-6.2946264
860	BAAF04053164.1/10519-10646	128	30.28474	-37.518704	-34.6	-0.08436	-8.43555937	-32.5	-0.15442	-15.442164
861	AANH01011402.1/5162-5288	127	38.6367	-50.609585	-50.2	-0.00816	-0.81590612	-49.7	-0.0183	-1.8301506
862	ABAV01030669.1/9778-9657	122	26.32387	-30.018177	-31.7	0.053054	5.305435614	-31.5	0.047042	4.70420028
863	AACT01014164.1/14797-14663	135	32.92461	-43.11673	-41.1	-0.04907	-4.90688479	-37.7	-0.14368	-14.367983
864	AABS01000098.1/66915-66782	134	31.19717	-40.16825	-35.2	-0.14114	-14.1143467	-32.8	-0.22464	-22.464176

865	AAZX01007551.1/46173-46008	166	34.27942	-51.460245	-44	-0.16955	-16.9551013	-43.1	-0.19397	-19.397319
866	BAAB01203970.1/2470-2315	156	34.72361	-50.171079	-49.3	-0.01767	-1.766895	-47.9	-0.04741	-4.7412927
867	AAAB01008986.1/3372038-3372230	193	39.15281	-64.60447	-59.3	-0.08945	-8.94514361	-58.7	-0.10059	-10.058722
868	AAFS01000016.1/19569-19724	156	34.82491	-50.332276	-55.8	0.097988	9.798789319	-53.7	0.062714	6.27136767
869	AAIZ01001811.1/683-838	156	34.99788	-50.607527	-56.2	0.09951	9.951020222	-55.9	0.094678	9.46775199
870	AANI01017162.1/86143-85990	154	21.59058	-28.873284	-28.9	0.000924	0.092444217	-27.6	-0.04613	-4.6133464
871	AANI01014648.1/138479-138633	155	34.03636	-48.877865	-55.8	0.124053	12.40526003	-50.6	0.034034	3.40342904
872	AAPU01011105.1/262422-262573	152	33.77134	-47.857331	-53.2	0.100426	10.04261067	-48.4	0.011212	1.12121668
873	AAPT01020183.1/127226-127384	159	34.77941	-50.858681	-56.6	0.101437	10.14367401	-51	0.002771	0.27709704
874	ABDC01347327.1/313-438	126	29.74935	-36.267545	-35.3	-0.02741	-2.7409216	-35.6	-0.01875	-1.8751273
875	ABDC01347327.1/313-438	126	30.26999	-37.096027	-35.9	-0.03332	-3.33155249	-35.1	-0.05687	-5.6866875
876	AANU01295318.1/748-623	126	30.01919	-36.696943	-34.6	-0.06061	-6.06052945	-34.1	-0.07616	-7.6156691
877	AANN01562320.1/870-745	126	30.26999	-37.096027	-32.8	-0.13098	-13.0976443	-32.4	-0.14494	-14.493911
878	AAIY01042223.1/294-419	126	30.60118	-37.623052	-32.5	-0.15763	-15.7632372	-32.9	-0.14356	-14.355781
879	AAPY01611414.1/511-386	126	30.60118	-37.623052	-35.7	-0.05387	-5.38670051	-34.8	-0.08112	-8.1122186

880	CAAE01014614.1/129416-129543	128	41.53103	-55.414923	-54.8	-0.01122	-1.1221227	-53.5	-0.03579	-3.5792957
881	AAVX01595596.1/659-533	127	31.42893	-39.13986	-37.3	-0.04933	-4.93259993	-36	-0.08722	-8.7218327
882	AANG01209374.1/1-113	113	28.37472	-31.485299	-34.2	0.079377	7.937721156	-33.7	0.065718	6.5718119
883	AAWU01017057.1/6040-5883	158	34.86681	-50.798151	-56.2	0.096118	9.611831269	-56.2	0.096118	9.61183126
884	AAJJ01000001.1/47747-47867	121	29.99055	-35.653369	-29.8	-0.19642	-19.6421768	-29.1	-0.2252	-22.520167
885	DQ682679.1/205-355	151	34.64025	-49.040424	-52.5	0.065897	6.589668562	-50.6	0.030822	3.08216599
886	AAGE02008333.1/19499-19656	158	35.22575	-51.369337	-56.1	0.084326	8.432554579	-55.2	0.069396	6.93960709
887	AAZO01007334.1/46791-46906	116	29.43586	-33.772676	-26.1	-0.29397	-29.3972269	-25.5	-0.32442	-32.441867
888	AAGV020469602.1/2323-2199	125	30.20428	-36.791867	-30.9	-0.19068	-19.0675307	-29.8	-0.23463	-23.462640
889	AASC02027737.1/1238-1108	131	38.06579	-50.499484	-49.7	-0.01609	-1.60861984	-47.7	-0.05869	-5.8689393
890	AAKN02006802.1/64876-65001	126	29.81704	-36.375259	-33.3	-0.09235	-9.23501233	-33.3	-0.09235	-9.2350123
891	CAAK05033158.1/3332-3460	129	31.427	-39.535983	-37.6	-0.05149	-5.14889122	-36.6	-0.08022	-8.0218117
892	AAWR02006087.1/50015-49891	125	29.90239	-36.311481	-33.9	-0.07114	-7.11351368	-32.5	-0.11728	-11.727634
893	AAWZ02013490.1/81111-81235	125	30.76642	-37.686404	-35.5	-0.06159	-6.15888354	-32.6	-0.15602	-15.60246
894	AAGW02073287.1/24794-24920	127	31.17131	-38.729911	-42.8	0.095096	9.509552838	-41.8	0.073447	7.3447096

895	AADG06006595.1/14072-13949	124	24.94711	-28.226535	-29.7	0.049612	4.96116124	-24.6	-0.14742	-14.742012
896	CAAB02003742.1/18728-18856	129	30.66413	-38.322035	-40.5	0.053777	5.377691114	-32.1	-0.19383	-19.383286
897	EU240318.1/2-119	118	38.50254	-48.599698	-44.8	-0.08481	-8.4814681	-43.4	-0.11981	-11.98087
898	AAQR03093718.1/17057-17182	126	30.46913	-37.41293	-35.9	-0.04214	-4.21428865	-34.9	-0.072	-7.2003714
899	AAPE02048822.1/2601-2476	126	31.30137	-38.737276	-39	0.006737	0.673651069	-38.8	0.001617	0.16165958
900	AAGJ04111208.1/10003-10134	132	28.66141	-35.733904	-32.8	-0.08945	-8.94482823	-29.8	-0.19912	-19.912428
901	FJ916040.1/3-128	126	30.01919	-36.696943	-34.6	-0.06061	-6.06052945	-34.1	-0.07616	-7.6156691
902	U62822.1/2-128	127	30.78084	-38.108556	-37.3	-0.02168	-2.16771053	-36.1	-0.05564	-5.5638671

Table 6.8. Comparison of MFE computed with the model with MFE from webservers, RNAfold and RNAstructure for 902 snRNA sequences

Table 6.10. Tabulation of computation of MFE using the algorithm developed, for Rfam snRNA sequences having MFE = 0 kcal/mol

Sl.No	Rfam snRNA family RF00004 (208)							
	Seq.No.	Specimen ID	NTL	SD_SCM	MFE_M	MFE_C	MFE_F	MFE_S
1	101	AALT01209640.1/567-377	192	0	0	-2.101	0	0
2	109	AALT01209640.1/567-377	196	0	0	-2.8994		
	Rfam snRNA family RF00015 (170)							
3	337	AABL01000640.1/15189-15058	132	0	0	9.875	0	0
4	340	X58844.1/1-130	130	0	0	10.2742	0	0
	Rfam snRNA family RF00020 (180)							
5	443	M10270.1/1-117	117	0	0	12.869	0	0

Table 6.10. Computation of MFE using the algorithm developed for Rfam snRNA sequences known to have 0 value of MFE

APPENDIX - II