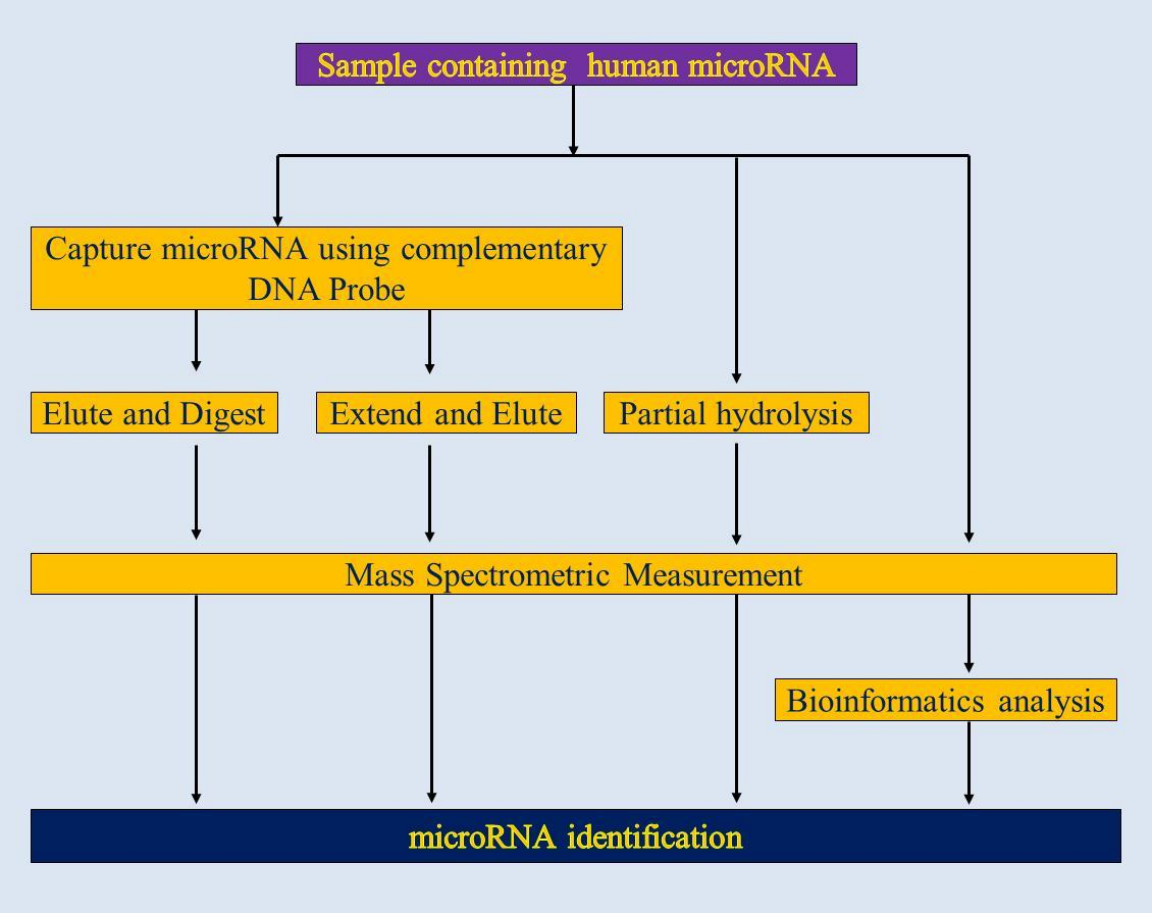
WAMBUA, DICKSON MUOKI, Ph.D. Mass Spectrometric Methods and Bioinformatics Tools for Accurate Identification of MicroRNA Biomarkers. (2012)
Directed by Dr. Norman H.L. Chiu. 172 pp.

MicroRNA (miRNA) are a class of endogenous non-protein-coding RNA of ~19-25 nucleotides long that post-transcriptionally regulate protein expression by targeting messenger RNAs for cleavage or translational repression. MiRNAs have been implicated in the initiation and progression of 160+ human diseases. Unique miRNA differential expression signatures can be used as a basis of discriminating against the presence or absence of human diseases. MiRNAs are therefore a promising and emerging class of disease biomarkers and therapeutic targets; however, the accurate detection of a specific miRNA has continued to be a challenging issue. Recently, mass spectrometry (MS) has seen remarkable technological advancements making it an attractive alternative to the conventional molecular biology miRNA characterization techniques. This study consistently documents the development of various analytical techniques aimed at characterization of miRNAs. The current literature in the field of miRNA is covered in chapter one. In chapter two, two new MS based concepts for detection of miRNA are introduced; a) the miRNA is captured using a specific complementary DNA probe, eluted and digested with specific endonuclease. The digested miRNA fragments are measured by MS resulting in a peak pattern that is dependent on the miRNA sequence i.e. an intrinsic mass signature and b) a unique mass signature is created by incorporating extra nucleotide(s) to the 3' end of miRNA and the extended miRNA is measured by using MS. The molecular mass of the extended miRNA, which is defined as extended mass signature, is expected to be different from the other miRNA within the same sample.

These two approaches can improve the accuracy on qualitative MS identification of specific miRNA. To better understand miRNA function however, it is important to elucidate the nucleotide sequence of the miRNA. Chapter three of this study introduces a novel MS based assay for the sequencing of miRNA through chemical hydrolysis. In this study, by taking advantage of the mixing between a miRNA sample and an acidic MALDI matrix prior to the MALDI-TOF MS measurements, a unique yet simple and relatively cost-effective approach to generate miRNA sequencing ladders was developed. By using this method, 100% sequence coverage and accuracy in the sequencing of selected miRNAs were achieved. When many samples are involved, the data generated from miRNA measurements can be complex and manual data processing is tedious and challenging, as such, the spectral interpretation of mass spectrometric data can quickly turn out to be the bottleneck in miRNA analysis. The success of MS as a tool for analysis of miRNA will therefore strongly depend on the development of relevant computational software with the ability to properly interpret and analyze the large data. To meet this need, chapter four of this work explains the development of MicroRNA MultiTool, a computational software for the rapid interpretation of MS data containing human miRNA. Users can directly enter data obtained from mass spectrometric measurement in order to obtain the identify of miRNA, highly reducing the time needed to process data. The development of such analytical and bioinformatics tools will provide scientists with the opportunity to better understand miRNA functions and will be influential in propelling the breakthroughs of miRNA in clinical diagnostics and therapeutic fields.



Schematic abstract depicting analytical assays and bioinformatics solutions developed in this study.

MASS SPECTROMETRIC METHODS AND BIOINFORMATICS TOOLS FOR ACCURATE IDENTIFICATION OF MicroRNA BIOMARKERS

by

Dickson Muoki Wambua

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2012

Approved by

_____
Committee Chair

To my wife and friend Sarah and my daughter Elizabeth, your support and understanding, encouragement and love gave me the energy to face every day with renewed vigor, a fire that kept burning throughout my Master's and Ph.D. studies – thank you. To Shushu Julie Brown, you are awesome; your support, advice and care throughout my Ph.D. scholarship are highly appreciated. To my mother Elizabeth, my dad David, brothers and sisters, I offer my thanks to you for letting me be away from home, your encouragement and prayers can never be overstated. My mother's words "*îkîa vitii … nîtîkûûvoyea"* have always propelled me beyond my imaginations and in ways untold.

APPROVAL PAGE

This dissertation written by DICKSON MUOKI WAMBUA has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Norman H.L. Chiu

Committee Members _____
Gregory Raner

_____
Wei Jia

_____
Will Taylor

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

My sincere thanks goes to Dr. Norman H.L Chiu, as my Ph.D. advisor, Dr. Chiu was very inspirational and a central figure throughout the entire period of my studies. He not only offered me precise guidance throughout my studies, but most importantly spend a lot of time advising me on the tricks and tactics on how to be a successful scientist. For my Ph.D. and for my life thereafter, I will forever remain very grateful to Dr. Chiu. His words *"..... get it right, get it done, make it happen ....*" will always reverberate in my life well beyond the boundaries of UNCG.

To my Ph.D. committee members, Dr. Gregory Raner, Dr. Wei Jia and Dr. Will Taylor, the time you took to review my proposal and dissertation, and the helpful hints that you gave me during my scholarship were greatly helpful in bringing this study to a successful ending. I truly appreciate your help.

I would like to acknowledge our collaborative team at JEOL USA, Inc. Peabody MA; Dr. Robert B. Cody, Dr. Masaaki Ubukata and Dr. John Dane were very supportive in the acquisition of high resolution mass spectrometry data for the microRNA sequencing project. I would also like to acknowledge the help of Zhonghao Chewy and Paolo Valerio for their contributions towards the development of the microRNA MultiTool computer software. I would also like to thank the Oracle Corporation for the Java Standard Edition Development Kit freeware that was used for development of MicroRNA MultiTool software. Thanks to the Chiu Research group members for all the ideas and support that we shared. Guidance on the choices of suitable microRNAs for the

studies in this dissertation from Dr. Tannous Bakhos of Harvard Medical School/ Massachusetts General Hospital is highly acknowledged. I also wish to thank the Department of Chemistry and Biochemistry at The University of North Carolina at Greensboro as well as the National Science Foundation GK-12 Program for supporting me financially throughout my Master and Ph.D. studies. The MALDI-TOF MS instrument used in these studies was purchased with a NSF Major Research Instrumentation Grant (Award #0420292).

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AB	Applied Biosystems
cDNA	Copy DNA
CID	Collision-induced dissociation
Da	Dalton
DBE	Double bond equivalent
dCTP	Deoxycytidine triphosphate
ddC	Deoxycytidine
ddG	Dideoxyguanosine
ddNTP	Dideoxyribo-nucleotide triphosphate
ddTTP	Dideoxythymidine-5'-Triphosphate
DGCR8	Di-George syndrome critical region gene
DHB	Dihydroxybenzoic acid
dNTP	Deoxyribo-nucleotide triphosphate
ds-RNA	Double stranded RNA
ECD	Electron capture dissociation
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FWHM	Full width half maximum
HPA	Hydroxypicolinic acid
JEOL	Japan Electron Optics Laboratory

LNA	Locked nucleic acids
m/z	Mass to charge ratio
MALDI	Matrix Assisted Laser Desorption/Ionization
miRNA	MicroRNA
mmu	Milli mass unit/ millidalton
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
ncRNA	Non-protein-coding RNA
Nd:YAG	Neodymium-doped yttrium aluminum garnet; Nd:Y3Al5O12
Nd-YLF	Neodymium-doped yttrium lithium fluoride
Oncomirs	Oncogenic miRNAs
ORF	Open reading frames
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
pKa	Acid dissociation constant
qRT-PCR	Quantitative real time polymerase chain reaction
RAKE	RNA primed-array based Klenow enzyme
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNase	Ribonuclease
RT-PCR	Reverse transcriptase polymerase chain reaction
SA	Sinapinic acid

SOS	Simple Oligonucleotide Sequencer
TOF	Time-of-flight
TRBP	Transactivating response ds-RNA binding protein
UTR	Untranslated regions
UV	Ultra Violet

CHAPTER I

INTRODUCTION

1.10 Review of the Literature

MicroRNA (miRNA) are a class of endogenous non-protein-coding RNA of ~19-25 nucleotides long that post-transcriptionally regulate protein expression by targeting messenger RNAs for cleavage or translational repression (Iorio et al. 2005). MiRNA were discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum when studying the development of the nematode *Caenorhabditis elegans* and have since revolutionized the understanding of post-transcriptional gene regulation (Omahen 2011). The genes that encode miRNA are evolutionarily conserved and are located within exons or introns of protein-coding genes or in intergenic areas (Fiorucci et al. 2012). MiRNAs have been shown to regulate biological processes that are central to cancer biology such as cell growth, angiogenesis, invasion, metastasis (Lee and Dutta 2009), as well as other key cellular functions like tissue differentiation, cell proliferation, apoptosis and embryonic growth (Lu et al. 2008). It is therefore not surprising that miRNAs are implicated in the initiation and progression of more than 160 human diseases including numerous cancer types, heart disease and diabetes (Jiang et al. 2009). It has been shown that unique miRNA differential expression signatures can be used as a basis of discriminating against the presence or absence of human diseases (Calin et al. 2005). These properties make miRNAs valuable diagnostic biomarkers (Park et al. 2009),

(Ferdin et al. 2010), classification and prognostic tools for human diseases as well as potential therapeutic targets (Calin and Croce 2006).

According to the miRBase database, 2100+ human miRNAs have already been reported and the number is constantly increasing (Griffiths-Jones et al. 2006), (Griffiths-Jones et al. 2008). A single messenger RNA can be targeted by more than one miRNA and each miRNA has the ability to regulate several different messenger RNA, this immensely increases the ability of miRNAs to regulate gene expression as well as the complexity of miRNA analysis (van Rooij and Olson 2007). MiRNAs bind to partially complimentary sequences on 3' untranslated regions (UTR) or open reading frames of their target messenger RNA and lead to either translation repression or degradation of the messenger RNA (Thomson et al. 2004). Through computational techniques, Friedman et al estimated that more than 60% of human messenger RNAs are indeed targets for miRNA (Friedman et al. 2009). It is therefore no wonder that miRNA research continues to generate increased interest as evidenced by a plot of miRNA discovered and deposited in miRBase versus the number or publications reported in SciFinder as shown in Figure 1.1.

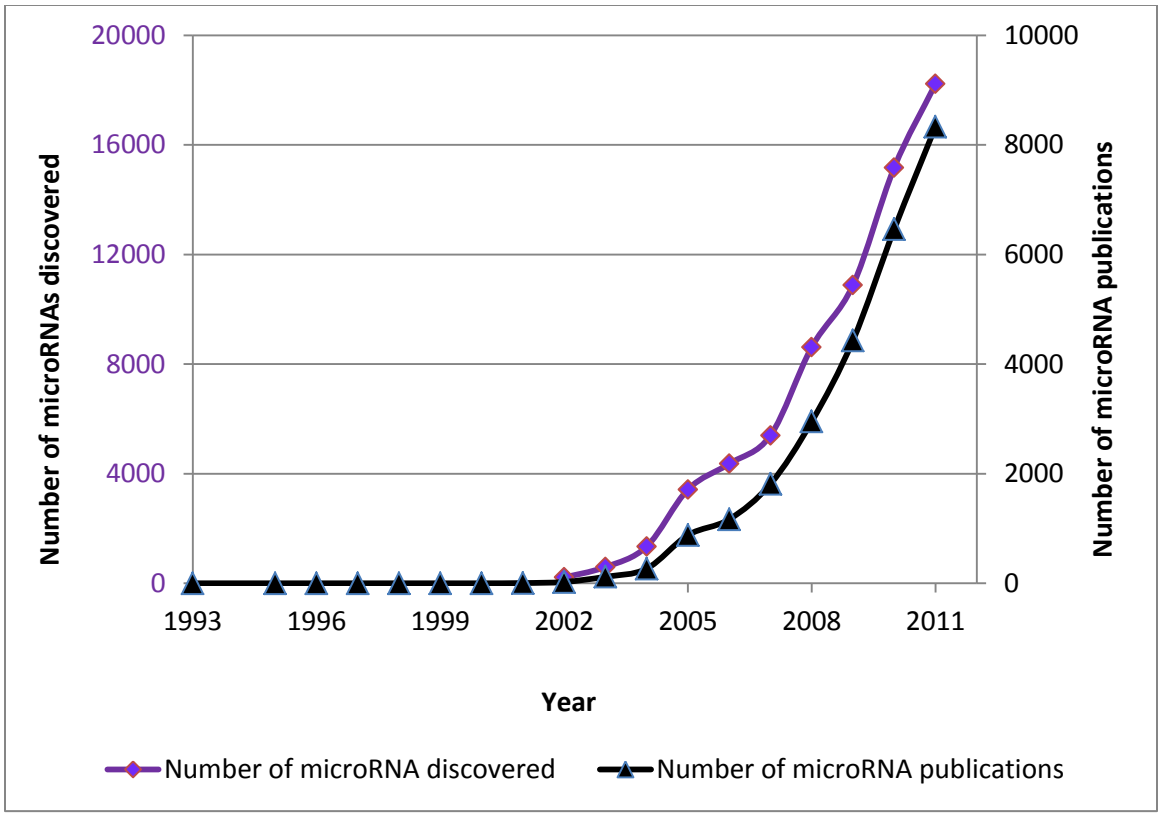


Figure 1.1. Number of miRNAs in miRBase versus publications in SciFinder.

1.11 MiRNA biogenesis

The schematic overview of miRNA biogenesis is shown in Figure 1.2. miRNA are transcribed by RNA polymerase II or RNA polymerase III as an independent gene or as part of a mRNA intron that can range from hundreds to thousands of nucleotides long. The transcript is cleaved by a double stranded RNA (ds-RNA) specific ribonuclease (RNase) called Drosha in conjunction with a ds-RNA binding protein known as Di-George syndrome critical region gene (DGCR8) into a hairpin structure of about 70 nucleotides long called pre-miRNA. The pre-miRNA has a two nucleotide 3' overhang and is exported from the nucleus to the cytoplasm by the nuclear factor exportin-5 with the help of a cofactor Ran-GTP. In the cytoplasm, the pre-miRNA is recognized and cleaved by a type III RNase endonuclease referred to as Dicer together with its corresponding immunodeficiency virus transactivating response ds-RNA binding protein (TRBP) into an imperfect miRNA duplex of ~19-25 nucleotides long. Only one of the two strands is retained while the other is removed and destroyed. TRBP then recruits a human argonaute protein hAgo2 onto the Dicer complex, forming a RNA-induced silencing complex (RISC). The Ago2 has two domains; a RNA binding domain and an endonucleolytic domain capable of cleaving target mRNA (Wiemer 2007). miRNA are thought to regulate gene expression by binding to complimentary sequences on 3' untranslated regions (UTR) or open reading frames (ORF) of target mRNA(Thomson et al. 2004). The critical base pairing process occurs through the 5' seed region (base pairs 2–8). Perfect or near perfect complementarily of miRNA to target mRNA is thought to cause mRNA degradation through RISC endonucleolytic activity while imperfect

complementarity stops the translation of miRNA by blocking the ribosome during translation (Hutvágner and Zamore 2002). The RISC complex can also effect post-transcriptional gene silencing by inducing the deadenylation of mRNA through mechanisms that are not yet fully understood (Creighton et al. 2009). The origins and mechanisms of action of miRNA were thoroughly reviewed by Carthew et al (Carthew and Sontheimer 2009).

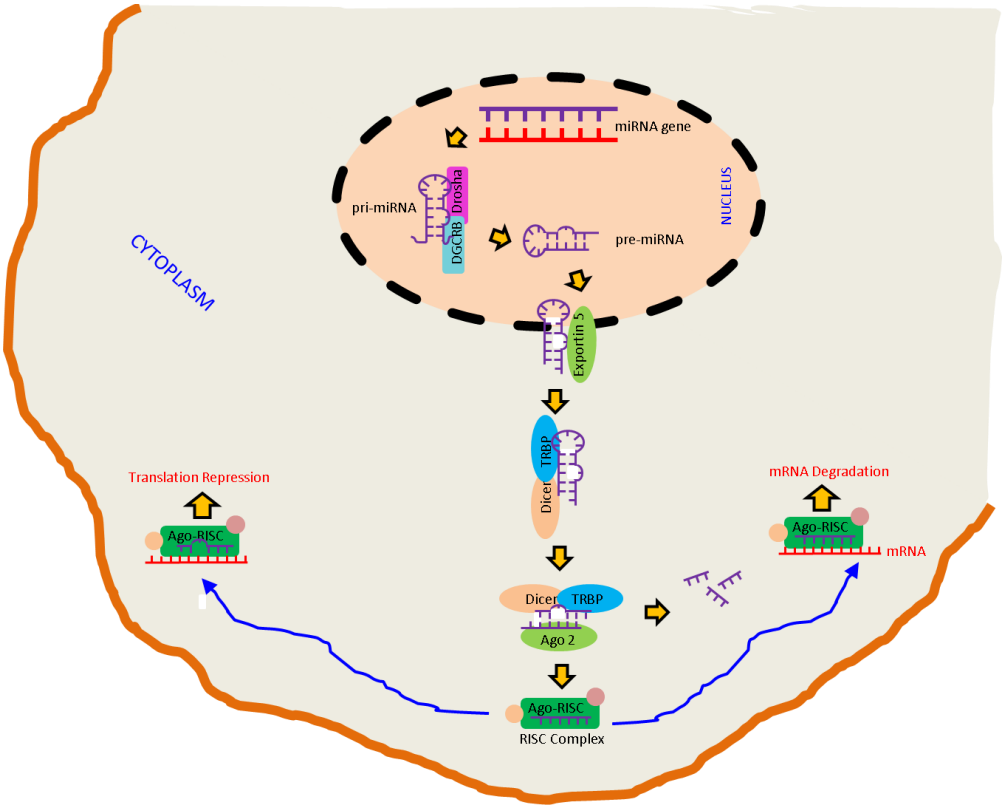


Figure 1.2. Schematic diagram of miRNA biogenesis in a mammalian cell.

1.12 MiRNA and diseases

Recent years have seen a tremendous interests in miRNA analysis especially after Calin and co-workers revealed that miRNA play a crucial role in cancer pathogenesis either as tumour suppressor genes or oncogenes (Calin et al. 2002), (Backes et al. 2010) this is evidenced by the exponential increase in miRNA publications as shown in Figure 1.1.

Cancer is a genetic disorder characterized by accumulation of mutations and deregulation of gene expression leading to uncontrolled cell proliferation, it has been one of the leading causes of death in the United States and around the world; this trend is expected to continue as the current population continues to age. The main stages in cancer include; initiation, promotion, malignancy, progression and metastasis. The regulation of these cellular activities especially gene expression have been the target by scientists to elucidate the underlying molecular mechanisms leading to cancer development with an aim of providing effective therapies or disease prevention (Negrini et al. 2009). Since the discovery of miRNA oncogenic and tumor suppression characteristics by Calin et al in 2002, more studies have accumulated evidence that deregulation of miRNA expression plays a central role in cancer pathogenesis (Bandrés et al. 2006). The expression of miRNAs has been found to vary dramatically across different tumor types, these miRNA expression patterns can be used for differentiating cancers according to their types and developmental stages. Such strategies can be used for prognosis of cancers as well as determination of better treatment choices. Many different miRNAs have been identified as diagnostic biomarkers or as possible new

therapeutic agents (Park et al. 2009). The potential for miRNA dependent diagnostics is boosted by the ease with which small RNAs can be routinely measured under normal hospital laboratories (Meltzer 2005). All research work that focuses on miRNA and/or their relationships with cancer do require high specificity and accuracy on measuring miRNA.

Numerous methods for identifying miRNA have been developed and tested. Among these methods, the traditional Northern blotting is still being used, which is mainly due to its availability in many research laboratories (Várallyay et al. 2008). The less labor intensive and more sensitive real-time polymerase chain reaction (PCR) method for measuring specific miRNA has been commercialized (Sharbati-Tehrani et al. 2008). However, due to the small sizes of miRNA, the primer design for PCR has been challenging. For achieving higher throughput, microarray technology and the next generation sequencing method are available for the analysis of multiple miRNAs (Davison et al. 2006), (Motameny et al. 2010), (Lu et al. 2005).

1.13 MiRNA as drugs and therapeutic targets

With a lot of research having been done on miRNA, it is evident that this class of small non-coding RNA can control vital biological processes, current research is therefore moving towards translating the information gathered into the development of novel therapeutic drugs (Nana-Sinkam and Croce 2011) that exploit the catalytic process of natural RNA as a new mechanism of action. Furthermore, miRNA have an added

advantage that they can regulate a wide range of oncogenes with high specificity, giving miRNA an advantage over cancer drugs that target a single genetic pathway of interest. There are currently two major strategies of developing miRNA based treatments, either as a) miRNA knockdown through miRNA Antisense Oligonucleotides b) restoration of miRNA tumour suppressors through miRNA mimics (Bader et al. 2011).

(a) MiRNA knockdown through miRNA Antisense Oligonucleotides

Oncogenic miRNAs (oncomirs) are miRNAs that promote proliferation when overexpressed in the cell. To counter the presence of oncomirs, miRNA Antisense Oligonucleotides can be used. In this case, the therapeutic strategy is to reduce the expressed miRNA. This is done through the use of anti-miRNA oligonucleotides. As shown on Figure 1.3, the anti-miRNAs consist of short nucleic acids that are complementary to the miRNA to be countered. The antisense miRNA bind onto the miRNA of interest, preventing them from reaching the target mRNA and thus effectively silencing the oncogenic miRNA. This suppresses the effect of the oncomir, and decreases disease progression (Feliciano et al. 2011).

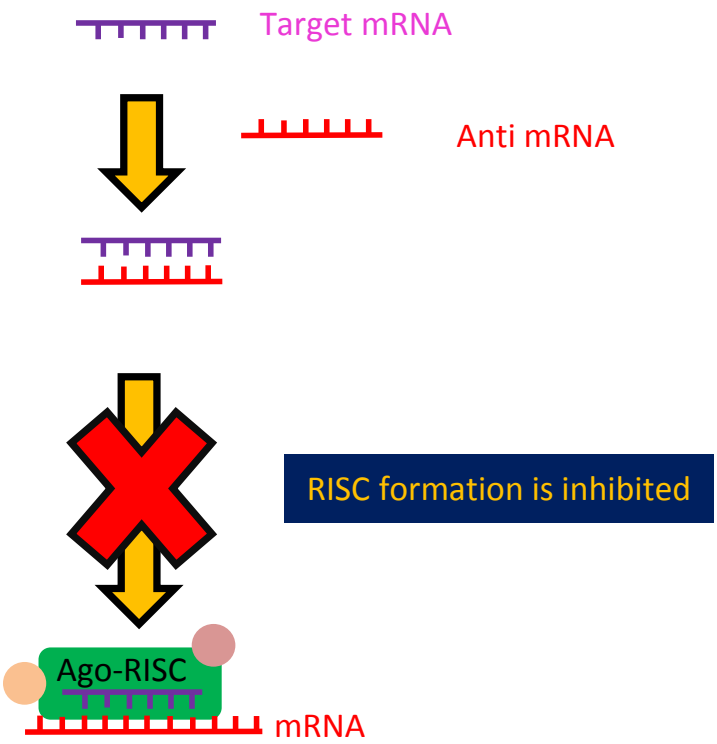


Figure 1.3. miRNA knockdown through the use of miRNA antisense oligonucleotides.

(b) Restoration of miRNA tumor suppressors through miRNA mimics

Tumor suppressor miRNAs are miRNAs whose downregulation promotes the growth of cancer. In such cases, the strategic approach is a replacement therapy that aims at restoring the miRNAs functions. This is done through exogenous expression of the miRNA of interest. The expression of miRNA mimics blocks proliferation, inducing cell death and therefore reduces cancer growth (Feliciano et al. 2011).

Other strategies of miRNA therapy including the use of small molecules that bind onto miRNA and prevent its function or that interfere with the maturation of miRNA by binding onto the pre-miRNA such as depicted in Figure 1.4 have been proposed as

suitable approaches to alternative miRNA therapies (Maiti et al. 2012). It is important to note that miRNA therapies are still in their initial stages and that factors such as cellular delivery, suitable biological distribution and general stability will be key to success of miRNA therapies (Xu and Anchordoquy 2011).

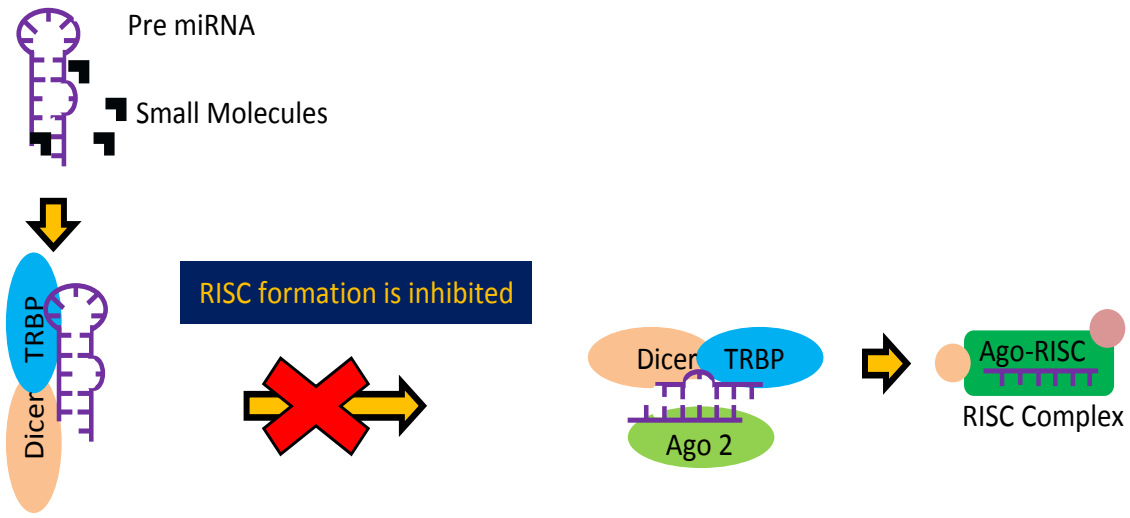


Figure 1.4. Inhibition of Dicer during miRNA biogenesis by use of small molecules.

1.14 Current miRNA detection assays

(a) Northern Blotting

To understand the functions of miRNAs, numerous methods have been developed and tested. Northern blotting is still widely used for miRNA analysis because it is readily accessible to many laboratories and does not necessitate the use of special or expensive apparatus and technical knowhow. High quality total RNA which includes small RNA is first isolated from the samples of interest. The total RNA is separated using denaturing polyacrylamide gel electrophoresis, followed by transfer of the RNA onto a membrane and fixing it using ultraviolet crosslinking. Radiolabelled probes that are complementary

to the miRNA are hybridized onto the miRNA and the membrane washed to remove non-specific hybridization before the final measurements are done. The main disadvantage of Northern blots is the poor sensitivity of the method especially when measuring microRNA with low expression levels. This necessitates the use of large sample sizes, an aspect that is trivial especially when sample sources are limited. An improvement to conventional Northern blotting that includes use of locked nucleic acids (LNA) on the oligonucleotide probes has greatly enhanced the specificity and sensitivity levels of Northern blots. LNAs are bicyclic analogues of nucleotides which exhibit high affinity towards their complementary sequences. In LNA, the O2´ and the C4´ atoms are linked by a methylene group thereby chemically locking the nucleotide in an N-type (C3'-endo) conformation. This results in enhancement of base stacking properties and pre-organization of the phosphate backbone leading to improved thermal stability of the LNAs after they hybridize onto complimentary targets (Várallyay et al. 2007), (Várallyay et al. 2008). Nonetheless, Northern blotting is considered a low throughput labor intensive method (Sipova et al. 2010), (Qavi et al. 2010). Due to the time and sensitivity constraints of Northern blotting, high throughput and more specific methods have been adopted for miRNA analysis (Várallyay et al. 2008).

(b) Microarrays

A great advantage of using microarrays is the parallel screening of multiple miRNAs and comparative miRNA expression analysis on the same platform (Davison et al. 2006). Initial steps of miRNA microarray analysis involves extraction of total RNA followed by enrichment of the RNA sample for small RNA. Alternatively, small RNA

isolation can be done directly from the samples. The miRNA are then labeled. Microarrays rely exclusively on the use of nucleic acid probes or their modified formats. For example, the use of molecular beacons, locked nucleic acids and peptide nucleic acids have already been reported (Hunt et al. 2009). To achieve a specific recognition of a known miRNA, the nucleic acid probe whose sequence is complementary to the miRNA forms a hybrid with the targeted miRNA through Watson-Crick base pairing. In other words, all the existing microarray assays for miRNA analysis rely on the precise annealing of miRNA onto complementary nucleic acid probes (Chen et al. 2008; Li and Ruan 2009). By carrying out the annealing process between the probe and targeted miRNA on a solid surface, e.g. nitrocellulose membrane, or polystyrene microtiter plate, any unwanted non-specific binding of the probe can be easily removed by washing the surface with a suitable buffer (Cissell et al. 2007). To further improve the specificity as well as achieve a high sensitivity for the detection of the miRNA-probe hybrids, various reporting labels have been explored. These include both radioactive and non-radioactive labels or fluorescent moieties (Chen et al. 2005). The major challenge of designing miRNA arrays is the capability to construct probes that can discriminate against many highly similar miRNA sequences whose difference can be as few as a single nucleotide. Despite their widespread use and acceptance, microarrays lack the ability to identify base mismatches between miRNA-probe duplexes which normally leads to non-specific hybridization and false signals (Lee et al. 2008). To harmonize the miRNA-probe duplex melting temperatures and enhance binding affinity in microarrays, the use LNAs has been reported. Regardless of the great progress made in microarray development, results so

obtained must be verified using an independent analytical method (Git et al. 2010), (Cho 2007). Other limitations of microarrays include the requirement for rather large sample volumes or alternative sample amplification step as well as the prolonged incubation time periods during sample preparation. The labeling of miRNA prior to analysis can be quite laborious. These limitations of microarrays contribute towards an increase in analysis cost and delay time for obtaining results, factors which have limited microarray techniques from point of care applications (Chen and Li 2007).

(c) Quantitative real time polymerase chain reaction

Quantitative real time polymerase chain reaction (qRT-PCR) is a powerful tool for the comparative analysis of miRNA due to its high level of accuracy, sensitivity and practical ease as well as its relatively short analysis time when compared to other methods. The most commonly used qRT-PCR method for analysis of mature miRNA is the stem-loop method which is based on TaqMan assay. The method consists of two main steps. In the first step, miRNAs to be analyzed are hybridized onto stem-loop primers with the 3' end of the miRNA next to the 5' end of the stem loop primer. The miRNA are reverse transcribed using reverse transcriptase into copy DNA (cDNA). In the second step, a miRNA specific forward primer, a reverse primer and dye labeled TaqMan probes are used to quantify the reverse transcribed miRNA product in a conventional TaqMan PCR method (Chen et al. 2005). Nevertheless, designing primers for miRNA is challenging due to their short sequences which leads to low melting temperatures that interfere with the PCR efficiency. Other variations of qRT-PCR such as the invader assay

(Allawi et al. 2004) and miR-Q have also been reported (Sharbati-Tehrani et al. 2008). Despite its sensitivity, PCR cannot amplify RNA modifications (Bartels and Tsongalis 2009).

(d) Bead-based technology

Carboxylated polystyrene microspheres are coupled to miRNA capture probes that consist of sequences complementary to those of miRNA to be analyzed. Each microsphere is designated specific miRNA and contains a unique combination of two fluorescent dyes. The miRNAs are first biotinylated before incubating them with the beads. After hybridization, the bead-miRNA complex is isolated though streptavidin phycoerythrin-biotin interaction. The complexes are then analyzed by flow cytometry, identifying each bead by its color combination. Fluorescence from the phycoerythrin is correlated with specific miRNA concentration. Bead arrays, just like microarrays are subject to non-specific hybridization and false signals (Bartels and Tsongalis 2009).

(e) Sequencing

The main advantage of sequencing is that it allows for the discovery of new miRNAs. The RNA extracts are first separated using polyacrylamide gel electrophoresis (PAGE) and the fraction containing small RNAs is obtained. 3’ and 5’ adapters are ligated to the small RNA to create sites for the annealing of primers used for reverse transcription of the miRNA. The miRNA are amplified by reverse transcriptase polymerase chain reaction (RT-PCR) and the resulting products are cloned into vectors to

form a cDNA library (Berezikov et al. 2006). The clones can be sequenced by Sanger sequencing method in which 2'-deoxynucleotides and 2',3' dideoxynucleotides are used in a polymerase dependent synthesis of miRNA cDNA. The cDNA is terminated randomly when a dNTP is incorporated into a growing chain. PAGE is used to separate the products by size in order to deduce the sequence of the miRNAs. Bioinformatics are then used to identify the location of the miRNA in the genome. Next generation sequencing technologies are also widely used for the discovery and sequencing of new miRNA (Morozova and Marra 2008). Sequencing is however, labor intensive, lengthy and cannot provide the information on modified nucleotide in the RNA molecule (Durairaj and Limbach 2008).

Other methods reported for miRNA analysis include rolling circle amplification (Jonstrup et al. 2006), and circularization based platforms of miRNA such as miR-ID which uses a ligase followed by reverse transcription of the circularized miRNA introduced by Kumar et al have also been explored (Kumar et al. 2011). Next generation sequencing technologies (Motameny et al. 2010) and bead based techniques that use a unique combination of fluorescent dyes have been developed (Lu et al. 2005). RNA primed-array based Klenow enzyme (RAKE) method (Nelson et al. 2004) and nanotechnology have also been described by Kong et al (Kong et al. 2009).

1.15 Mass spectrometric sequencing of nucleic acids

There are two main approaches for the mass spectrometric sequencing of nucleic acids 1) the top-down approach and 2) bottom-up approach. In top-down approach, the

entire nucleic acid molecule is introduced into the mass spectrometer and ionized. Precursor ions of the desired RNA are selected for fragmentation in which cleavage of the nucleic acid backbone is induced through gas phase fragmentation. Collision-induced dissociation (CID) is the most commonly used fragmentation technique. The characteristic spectra obtained from the fragmentation process are then used to piece together the nucleic acid sequence (Keough et al. 1993). Other gas-phase fragmentation techniques such as electron transfer dissociation (ETD) or electron capture dissociation (ECD) have been investigated (Taucher and Breuker 2010). The main advantage of top-down approach to sequencing is its inherent speed at which data can be generated. Fundamentally all structurally conceivable fragmentations of the molecular ion can occur, leading to numerous CID fragments. The number of fragment ions increases with the increasing length of the precursor ion. Even though the numerous fragmentation pathways can increase the confidence of sequence identification, the phenomenon makes mass spectra very complex and is therefore a great drawback of the top-down approach (Oberacher and Pitterl 2009). A recent attempt by Oberacher et al to sequence of 22 mers by the top-down approach yielded the correct sequences in only 90% of cases (Oberacher and Pitterl 2011).

Mass spectrometric sequencing of RNA by the bottom-up approach uses the principle of first fragmenting the RNA to create a "mass ladder" whose mass difference between adjacent mass spectrometric peaks can be matched to known single nucleotide masses in the sequence. The fragments are then measured by a mass spectrometer and the information obtained is used to reconstruct the sequence. The most commonly used

methods for fragmenting RNA include enzymatic digestion or chemical hydrolysis using alkaline or acidic methods. For enzymatic digestion, a combination of 5’ and 3’ exonucleases have to be used to degrade RNA from 3’ and 5’ end for a complete mass ladder, however, due to variation in enzyme kinetics, multiple MS measurements have to be carried out at different stages during the enzymatic process (Hahner et al. 1997a). Modifications on RNA such as 2' O-Methyl and phosphothionates make the nucleic acids resistant to nucleases digestion and can therefore not be sequenced through enzymatic digestion. Enzymatic assays are also time consuming and costly for large-scale applications (Schuette et al. 1995). Recently Bahr et al used trifluoroacetic acid to hydrolyze and obtain more than 90% sequence of a 22 mer RNA. The phosphodiester backbone of the RNA was randomly cleaved through acidic hydrolysis to generate fragments that represent a sequencing ladder. Cleavage of the terminal dinucleotides was however not observed, and the sequencing was completed by gas phase fragmentation using high resolution MS (Bahr et al. 2009).

1.16 Principles of MALDI mass spectrometry

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) is a mass spectrometric technique that produces protonated gas phase ions and is used for obtaining the molecular weights of analytes. MALDI is considered a soft ionization technique i.e. allows the analysis of intact molecules with minimal fragmentation and has therefore become a common method for the analysis of a wide range of samples

including biomolecules such as nucleic acids and proteins. The combination of MALDI with time-of-flight (TOF) mass analyzer to form (MALDI-TOF MS) resulted in a simple and sensitive and high speed tool for high throughput genomic research (Garcia et al. 2002).

In MALDI experiments, the sample is first mixed with a molar excess of a matrix and the liquid matrix-sample mixture is spotted onto steel MALDI plate such as that shown on Figure 1.5 and air dried. Most MALDI matrices are conjugated weak organic acids that exhibit optimal light absorption of the laser used (Aitken 2005). Common MALDI-MS lasers are in the ultra violet region with various wavelengths such as the frequency-quadrupled Nd:YAG lasers at 266 nm, nitrogen lasers at 337 nm or frequency-tripled Nd:YAG lasers at 355 nm, however infra-red lasers with a wavelength of ~3 μM have also been used (Niu et al. 1998). As shown in figure 1.5 the sample is then irradiated using the laser a few nanoseconds at a time. The matrix absorbs the laser light through electronic transitions which results in rapid heating, sufficient enough to cause the matrix-sample mixture to volatilize into gas phase (Beavis and Chait 1989). The matrix serves two main purposes, it absorbs the laser energy to enable desorption, and is acts as a source of H^+ ions that are transferred to the sample for ionization purposes. The ions enter a time of flight mass analyzer which measures the time it takes ions of different masses to travel from the ion source to the detector. Usually the ions are accelerated using an equal potential difference between the ion source and the detector. Since the ions are given the same push by the accelerating potential difference, similar ions gain similar kinetic energy and the velocity of the ions is therefore determined by the mass to

charge ratio (m/z). A schematic diagram of the Applied Biosystems 4700 MALDI-TOF-TOF –MS instrument used for some of the studies in this work is shown in Figure 1.6.

Lighter ions travel faster and reach the detector earlier than heavier ions. The time taken by an ion of a particular m/z to reach the detector is then converted into a mass and plotted against the count/intensity of the ions. However, in some cases, ions of the same m/z generated in the ion source and accelerated into the time-of-flight mass analyzer may not have exactly the same initial time spread or precisely identical kinetic energies. This results in the ions reaching the detector at different times and leads to peak broadening and loss of resolution.

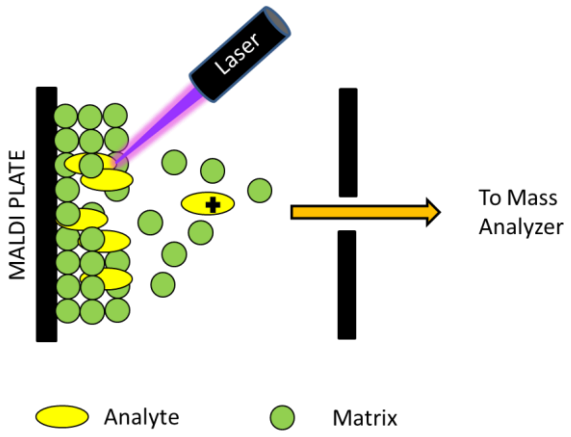


Figure 1.5. MALDI laser desorption process.

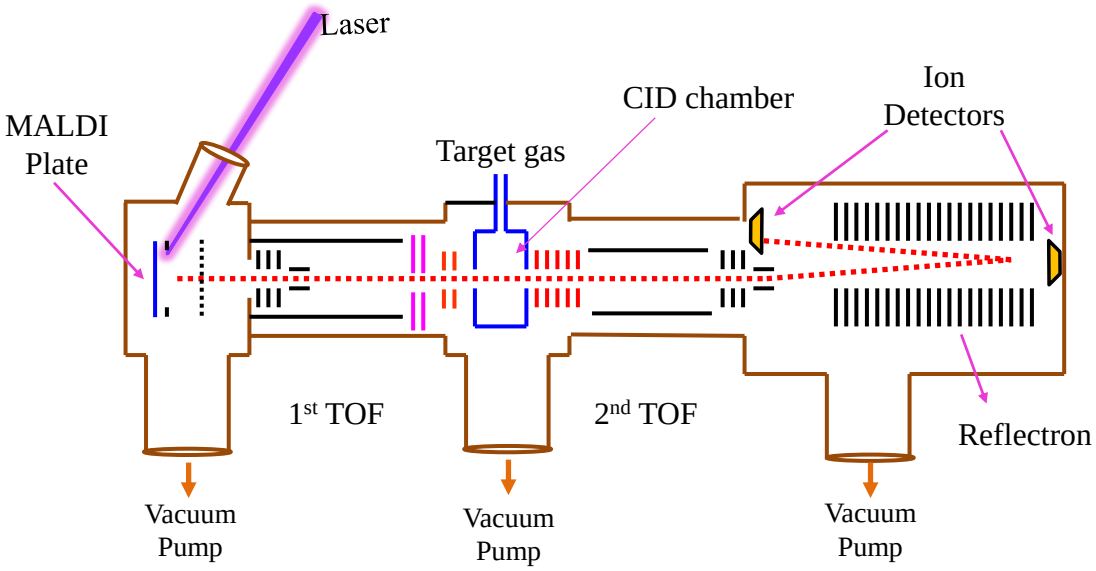


Figure 1.6. Schematic diagram of ABI 4700 MALDI TOF Mass Spectrometer.

Several solutions such as reflectrons have been developed to compensate for these differences. The reflectron is an optic ion "mirror" at the end of the time of flight tube consisting of a series of plates with increasing electric field that reverses the ions flight

path back in the direction of the flight tube. Ions of the same m/z with slightly different kinetic energies penetrate the reflectron to different depths, with the higher kinetic energy ions arriving at the reflectron first and penetrating deeper than ions will lesser kinetic energy. The increased flight path of the high kinetic energy ions allows ions of the same m/z and different initial kinetic energies to travel and reach the detector at the same time, effectively increasing mass resolution (Aitken 2005).

The development of new easy-to-use methods for the characterization of miRNA will play an important role in understand gene regulation by miRNA in both diseased and normal conditions. The subsequent chapters describe the development of various tools for the identification and characterization of miRNA.

1.20 Purpose of Study

The purpose of this study was to develop highly specific and selective assays for the characterization of miRNA as well as designing bioinformatics solutions that assist in the interpretation of data from mass spectrometric measurements of miRNA.

1.30 Hypothesis and Specific Aims

Hypothesis 1

Specific miRNA can be captured by using a complementary DNA probe. The miRNA can then be eluted and digested with specific endonuclease. The digested miRNA fragments are directly measured by using MS giving rise to an intrinsic miRNA

signature, alternatively, a unique extended mass signature can be created by incorporating one or more than one extra nucleotides to the 3' end of miRNA and the extended miRNA measured by using MS. The two different approaches to create a unique mass signature which can improve the accuracy on qualitative MS detection of specific miRNA.

Specific aim 1

To design a highly specific assays for the intrinsic and extended mass spectrometric signatures for detecting miRNA.

Hypothesis 2

By taking advantage of the mixing between a miRNA sample and MALDI matrix prior to the MALDI-TOF MS measurements and the acidity of MALDI matrix, a unique yet simple and relatively cost-effective approach to generate miRNA sequencing ladders can be developed.

Specific aim 2

To utilize acidic MALDI matrix to hydrolyze the phosphodiester bonds in each miRNA molecule, to produce sequencing ladders of the miRNA molecule prior to MALDI-TOF MS measurements, followed by elucidation of the miRNA sequence.

Hypothesis 3

The constantly increasing number of human miRNA and the recent discoveries of miRNA containing modified nucleotides has continued to make miRNA mass spectrometric data analysis a great challenge, however, the development of relevant computational software can enable the direct identification of miRNA from MS data facilitating the rapid interpretation of MS data containing modified or unmodified human miRNA.

Specific aim 3

To build a software that contains databases of all known human miRNA sequences as well as RNA modifications where users can directly enter data obtained from MS measurement in order to obtain the identify of both modified or unmodified miRNA.

1.40 Significance of the Study

According to recent studies in mir2disease website (http://www.mir2disease.org/), a publicly accessible database linking different miRNAs to their corresponding related diseases, miRNAs are implicated in the initiation and progression of more than 160 human diseases, including different cancer types, heart diseases and diabetes (Jiang et al. 2009) and have been shown to regulate biological processes that are central to cancer biology. The regulation of these processes especially gene expression have been

the target by scientists to elucidate the underlying molecular mechanisms leading to cancer development with an aim of providing effective therapies or disease prevention (Negrini et al. 2009). Since the discovery of miRNA oncogenic and tumor suppression characteristics by Calin et al in 2002, more studies have accumulated evidence that deregulation of miRNA expression plays a central role in cancer pathogenesis (Bandrés et al. 2006). The expression of miRNAs has been found to vary dramatically across different tumor types, these miRNA expression patterns can be used for differentiating cancers according to their types and developmental stages. Such strategies can be used for prognosis of cancers as well as determination of better treatment choices. Many different miRNAs have been identified as diagnostic biomarkers as well as in the development of new therapeutic treatments (Park et al. 2009). The assays and bioinformatics tools developed in this work serve to fill the void in the increased demand for the development of more efficient, accurate and rapid methods for identification, sequencing or sequence verification of miRNAs.

CHAPTER II

CREATING MASS SIGNATURES FOR THE DETECTION OF microRNA

Parts of the work presented in chapter II has been published in the Royal Society of Chemistry Analytical Methods Journal and is referenced as Wambua DM, Tannous BA, Chiu NHL. Creating Mass Signatures for the Detection of MicroRNA. *Anal. Methods*, 2012, 4 (10), 3453 – 3459.

2.10 Introduction

MicroRNA is a class of endogenous non-coding RNA which has ~19-25 nucleotides (nt). Currently, there are more than 2,100 human miRNA reported in the literature (Griffiths-Jones et al. 2006), (Griffiths-Jones et al. 2008). With the formation of RNA-induced silencing complex (RISC), specific miRNA can recognize and bind to its corresponding messenger RNAs (mRNA). The miRNA binding can lead to either the degradation of mRNA or the translation of mRNA being blocked (Wiemer 2007). Owing to these regulatory functions of miRNA on protein expression, miRNAs have been used as biomarkers for differentiating cellular activities, especially the activities that are related to diseases (Iorio et al. 2005), (Jiang et al. 2009), (Calin et al. 2005), (Park et al. 2009), (Ferdin et al. 2010), (Calin and Croce 2006), (Calin et al. 2002), (Backes et al. 2010). More recently, miRNAs have also been used as drugs, for example, the restoration of miR-26 levels with miRNA mimics has been proposed in a scheme for the treatment of liver cancer. Numerous methods for identifying miRNA have been developed and tested. Among these methods, the traditional Northern blotting is still being used, which is

mainly due to its availability in many research laboratories (Várallyay et al. 2008). The less labor intensive and more sensitive real-time polymerase chain reaction (PCR) method for measuring specific miRNA has been commercialized (Sharbati-Tehrani et al. 2008). However, due to the small sizes of miRNA, the primer design for PCR has been challenging. For achieving higher throughput, microarray technology and the next generation sequencing method are available for the analysis of multiple miRNAs (Davison et al. 2006), (Motameny et al. 2010), (Lu et al. 2005). Similar to other high throughput analysis, the challenges of using either method are the needs for suitable bioinformatics software to analyze the experimental data as well as the relative high costs on purchasing the equipment and consumable reagents. All the above methods rely on using a reporting molecule that has been attached to a nucleic acid probe or primer to generate a detection signal. In other words, the end-point signal was not directly obtained from miRNA. Theoretically, the direct measurement of miRNA can provide higher accuracy than measuring the reporting molecule. With the development of soft ionization techniques in mass spectrometry (MS), for example matrix-assisted laser desorption/ionization (MALDI), it has become practical to measure the molecular mass of various biomolecules (Bahr et al. 2009). Besides achieving high accuracy on the detection of specific miRNA, the direct MS measurement of miRNA can also be used to discover or detect miRNA modifications, such as nucleotide addition or deletion, conversion of adenosine to inosine, and methylation (Nishikura 2010).

In this report, we introduce a new concept that can further enhance the accuracy on using the direct MS measurements of miRNA. The concept aims at creating a unique

signature in the mass spectrum of miRNA. To achieve this goal, one way is to digest a specific miRNA with an endonuclease prior to the MS measurement. Before the RNase digestion, miRNA was first isolated from a sample by using a complementary DNA probe. The captured miRNA was eluted from the probe and digested with RNase. All digested RNA fragments were then measured by using MS. The result was the creation of a unique pattern of peaks that corresponds to the digested miRNA fragments in the mass spectrum. By using this experimental approach i.e. capture and digest the miRNA of interest, the unique pattern of peaks is defined as a mass signature in this study (Figure 2.1a). Alternatively, a unique mass signature can be created by using the template-dependent activity of DNA polymerase to extend the 3' end of miRNA with a specific type and number of nucleotides (Figure 2.1b). In this way, the extended miRNA has a unique molecular mass which is different from the molecular masses of other miRNA within the same sample.

2.20 Materials and Methods

Two synthetic RNA oligos with the same sequence as the human miR-153 (5' p-UUGCAUAGUCACAAAAGUGAUC-OH 3'), and miR-183-5p (5' p-UAUGGCACUGGUAGAAUUCACU-OH 3') were synthesized and HPLC purified by Integrated DNA Technologies (San Diego, CA, USA). Oligo dT coated magnetic beads and binding buffer (20 mM Tris-HCl, 500 mM LiCl, 0.5% LiDS, 1 mM EDTA, 5 mM DTT at pH 7.5), 2'-deoxynucleoside-5'-triphosphates (dATP, dTTP, dGTP, and dCTP) and Klenow Fragment (3′→5′ exo-) of DNA Polymerase I (5000

units/mL) together with the hybridization buffer (10 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM Dithiothreitol, pH 7.9 at 25°C) were all purchased from New England BioLabs (Ipswich, MA, USA). 2',3'-dideoxynucleoside triphosphate set (ddATP, ddTTP, ddGTP and ddCTP) and RNase T1 from Aspergillus oryzae were bought from Roche (Indianapolis, IN, USA). Ammonium citrate dibasic, triethylammonium acetate buffer (TEAA) and 3-hydroxypicolinic acid (3-HPA), 2',3',4' trihydroxyacetophenone (THAP) were acquired from Sigma Aldrich (St. Louis, MO, USA). Acetonitrile and methanol, both were HPLC grade, were obtained from Fisher Scientific (Pittsburgh, PA, USA). Non-sterile 0.22 µm low protein binding Durapore (PVDF) syringe driven membrane filter units (25mm) were purchased from Millipore Corp. (Bedford, MA, USA). Deionized water obtained from a Barnstead NANOpure Diamond Water Purification System (Dubuque, Iowa, USA) was used for reconstitution and dilution of all oligos.

2.21 Design of DNA capture probes

In order to separate a specific miRNA from a sample, a DNA oligo was used as a probe to capture the miRNA. The design of the DNA probe can be divided into three parts: (1) the complementary DNA sequence for capturing miRNA is strategically placed in the middle of the probe, (2) the downstream sequence of each DNA probe has a poly$(dA)_{20}$ tail, and (3) the upstream sequence has a variable sequence of 15 nucleotides. The two DNA probes that were used to capture the selected miRNAs in this study are shown on Figure 2.2. The underlined section is the complementary sequence for capturing corresponding miRNA. Both DNA probes with hydroxyl groups at both 5' and

3’ ends were ordered from Integrated DNA Technologies (San Diego, CA). The Integrated DNA Technologies OligoAnalyzer software was used to evaluate the cross hybridization of DNA probes with other possible miRNA, and ruled out the formation of hairpins or substantial secondary structures of DNA probes that would otherwise interfere with the hybridization.

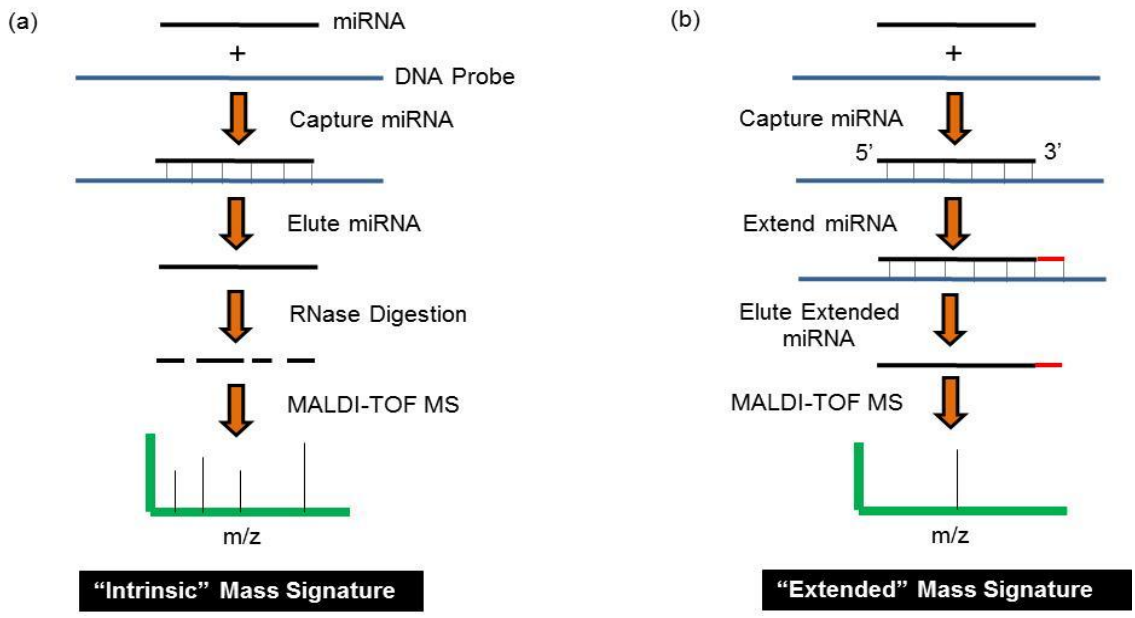


Figure 2.1. Schematic diagram of (a) intrinsic and (b) extended mass signatures

The two synthetic miRNAs (miR-153 and miR-183) and their respective DNA capture probes (miR-153 DNA probe and miR-183 DNA probe) used in this study are shown in Figure 2.2. The sequences underlined on the DNA capture probes are complementary to the specific miRNAs.

2.2 (a)

```
                          5' pUAUGGCACUGGUAGAAUUCACU-OH (miR-153)
                             ||||||||||||||||||||||
 3' HO-AAAAAAAAAAAAAAAAAAAAATACCGTGACCATCTTAAGTGAGATCACAATAGGCGA-OH
(Probe)
```

2.2 (b)

```
                          5' pUUGCAUAGUCACAAAAGUGAUC-OH (miR-183)
                             ||||||||||||||||||||||
 3' HO-AAAAAAAAAAAAAAAAAAAAAACGTATCAGTGTTTTCACTAGCTAGACAATAGGCGA-OH
```

Figure 2.2. (a) miR-183 and (b) miR-183 hybridized onto DNA capture probes.

Table 2.1. Molecular masses of miR-153, miR-183 and their DNA capture probes.

Oligo	Neutral Mass (Da)	Positively charged Mass (Da)
hsa-miR-153	7090.272	7091.280
hsa-miR-183	7083.232	7084.240
miR-153 DNA probe	17660.699	17661.707
miR-183 DNA probe	17654.700	17655.708

2.22 Intrinsic mass signature by digestion of miRNA

20 pmol of miRNA was mixed with 30 pmol of complementary DNA capture probe in a hybridization buffer (10 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM Dithiothreitol, pH 7.9). The total reaction volume was 15 µL. The reaction mixture was heated to 95 °C for 5 min to denature the miRNA and DNA capture probe. This was

followed by cooling the reaction mixture to 37 °C at 2.0^0C/min, during which the miRNA annealed to the DNA capture probe. The miRNA-DNA probe duplex was then immobilized on the surface of oligo dT coated magnetic beads (60 µL) through the interactions between the poly$(dA)_{20}$ tail on the DNA probe and the oligo dT. Before using the beads, they were washed in 100 µL of binding buffer which was supplied with the beads. The beads were resuspended in 100 µL of binding buffer, and added to the solution that contained the miRNA-DNA probe duplex. The mixture was incubated at room temperature for 10 minutes with occasional mixing. The beads were washed twice, each time with 100 µL of 200 mM TEAA buffer at pH 7.0. After that, miRNA was eluted from the beads by adding 5 µL of deionized water to the beads, and the mixture was heated up to 50 °C for 2 minutes. The eluted miRNA was digested with 5 U of RNase T1 at 37 °C for 10 minutes (Douthwaite and Kirpekar 2007). The digested RNA fragments were measured by using MALDI-TOF MS as described later in this section.

2.23 Extended mass signature by miRNA-primed extension

The use of DNA capture probe to capture the corresponding miRNA was carried out by the same protocol as described above. After the annealing of miRNA to the DNA capture probe, the 3' end of miRNA was enzymatically extended by adding 10 µL of a reaction cocktail that contained a suitable 2',3'-dideoxyribo-nucleotide triphosphate (ddNTP) or mixture of suitable 2'-deoxyribo-nucleotide triphosphate (dNTP) and 2',3'-dideoxy- ribonucleotide triphosphate and 6.75 U of Klenow Fragment (3′→5′ exo$^-$) in the hybridization buffer. The final concentration of each nucleotide triphosphate in the

reaction mixture was 33 μM. The mixture was incubated at 37 °C for 20 minutes or 40 minutes to perform single or double nucleotide extension of miRNA, respectively. Following the 3' extension, the miRNA-DNA probe duplex was immobilized on the magnetic beads as described above. After washing the beads, the extended miRNA was eluted in 5 μL of deionized water with heating at 50 °C for 2 minutes. The extended miRNA was measured by using MALDI-TOF MS as described later in this section. For experiments that involved the analysis of two miRNAs, the amount of dNTP or ddNTP, Klenow fragment, magnetic beads, hybridization and washing buffers was doubled.

2.24 Preparation of MALDI-TOF MS samples

3-HPA was selected as the MALDI matrix unless stated otherwise. The 3-HPA solution was prepared by dissolving 35.0 mg of 3-HPA and 8.80 mg of ammonium citrate dibasic in 1.0 mL of 10 % acetonitrile as described by Chiu et al (Chiu 2009). To dissolve the matrix, the solution was vortexed for 1-2 minutes. Before using the matrix solution, it was filtered through a membrane filter and stored at −20 °C for not more than one month. The stainless steel MALDI sample plate was cleaned with water and methanol before use. All MALDI samples were prepared by the thin-layer method. Specifically, 0.3 μL of MALDI matrix solution was spotted on the sample plate and was completely air dried before spotting 0.3 μL of the sample on top of the dried matrix.

2.25 MALDI-TOF MS measurements

A MALDI-TOF mass spectrometer (4700 Proteomics Analyzer, Applied Biosystems, Framingham, MA) was used for all MS measurements. Each sample was

measured by using the linear high mass positive mode in the 4000 Series Explorer Version 3.0 software, unless stated otherwise. The Nd:YAG laser (200 Hz) was set at 5,000 arbitrary units with maximum setting at 7,900 arbitrary units. The focus mass was the expected mass of undigested or extended miRNA in each sample. Molecular ions were extracted from the ion source after 450 ns of delayed time. The accelerating voltage was +20.0 kV and grid voltage was +18.8 kV. The instrument was equipped with a 200 Hz digitizer. The sampling bin size was 0.5 ns with an input bandwidth of 500 MHz, and a vertical full scale was 500 mV. The linear detector voltage was +2.0 kV. The pressure inside the entire instrument was maintained at the level of 10^{-8} Torr. Each spectrum was automatically acquired by accumulating the results of 3,000 shots with random edge-biased positioning of each laser shot. By using the Data Explorer Version 4.6 software, resulting mass spectra were calibrated.

2.26 Optimization MS measurements and enzymatic extensions

Since the introduction of MALDI-MS in 1985, a great deal of effort has been directed towards understanding the mechanisms underlying the process of desorption and ionization. Although these processes are still not fully understood, several key parameters that affect the quality of results obtained such as matrix conditions, laser intensity, number of laser shots and delays in ion extraction have been well documented (Guenther et al. 2010). Sample preparation techniques have been shown to greatly influence the quality of mass spectra obtained in MALDI MS measurements. Factors that affect signal intensity include matrix solvent composition, matrix-sample ratios, pH and the rate or

nature in which the matrix- sample- cocrystals are grown (Cohen and Chait 1996). Our group and others have previously demonstrated that 3-HPA and THAP are the most suitable MALDI matrices for the analysis of nucleic acids. Of the two matrices, 3-HPA is best suited for the measurement of RNA of high molecular weight while THAP is best used for RNA of low molecular weight (Chiu 2009), (Kirpekar et al. 2000). The MALDI measurements were performed using 3-HPA as the MALDI matrix except for the analysis of RNase digested miRNA in which case THAP matrix was used. To optimize the mass spectrometric measurements of miRNA, the factors that affect the signal of measured RNA were considered. In all circumstances, the samples consisted of 0.3 µL of miR-183 and were prepared through the thin layer method using 3-HPA matrix as explained under materials and methods. The samples were then analysed using Applied Biosystems 4700 MALDI-TOF MS in the positive ion mode.

(a) Optimization of MALDI matrix volume, laser intensity, number of laser shots and delay time

miR-183 was separately analysed by MALDI-MS using varying volumes of 3-HPA matrix. As shown in Figure 2.3 (a) the signal/noise ratio (S/N) or the measured miR-183 increased with increasing matrix volume. The highest S/N was observed when 0.3 µL of matrix was used. The S/N slightly reduced with increased matrix volume. Laser intensity was varied from 3000 to 7000 arbitrary units (arb); the highest setting of the laser is 7900 arb. As shown in Figure 2.3 (b), the highest S/N was observed with laser intensity of 5000 arb, after which a steady drop in S/N was observed. This is in

agreement with previous studies that have shown that minimum laser intensity is required to observe the analyte ions, thereafter, the S/N increases as a function of the increasing laser intensity until such a time that the laser intensity starts to cause fragmentation of the analyte (Guenther et al. 2010). The laser energy was held constant at 5000 arb and the number of laser shots used were varied from 500-4000 shots with 500 shot increments. As shown in Figure 2.3 (c) the highest signal was achieved with 3000 laser shots (Holle et al. 2006).

The main cause of mass resolution loss in linear MALDI TOF-MS is attributed to the difference in flight times of ions with identical mass to charge ratios caused by their different initial velocities. This is because this initial velocity component of the ions also contributes to the final velocity that is attained by ions of a particular m/z during their acceleration out of the ion source and into the field free flight tube. There are two main ways of extracting ions; continuous and delayed. For continuous mode the ions are extracted and accelerated towards the detector immediately after their formation, under these conditions, no flight time compensation is made for ions with the same m/z but different initial ion velocities. This leads to ions of same m/z reaching the detector at different times, resulting in wide peaks and a great loss of resolution. In delayed ion extraction, the extraction of ions is discontinuous. A time delay is employed between the formation of ions and their extraction from the source. During this delay time, the ions shift to new positions depending on their velocities. This delay time allows the high density plume to expand thereby reducing collisional kinetic energy. The resulting change in position and potential energy has been shown to compensate for the differences

in initial ion velocity. When potential energy is applied to accelerate the ions towards the detector, ions of the same m/z travel together leading to narrower peaks and better resolution compared to the continuous ion extraction (Juhasz et al. 1996).

In these experiments, the delay time was varied from 200-800 ns and plotted against the resolution of the resulting spectra as shown in Figure 2.3 (d). The highest resolution was observed at 400 ns beyond which the resolution reduced sharply. The drastic reduction in resolution with increasing delay time could be due to the loss of ions during the extended period of time that the ions spend in the ion source before extraction.

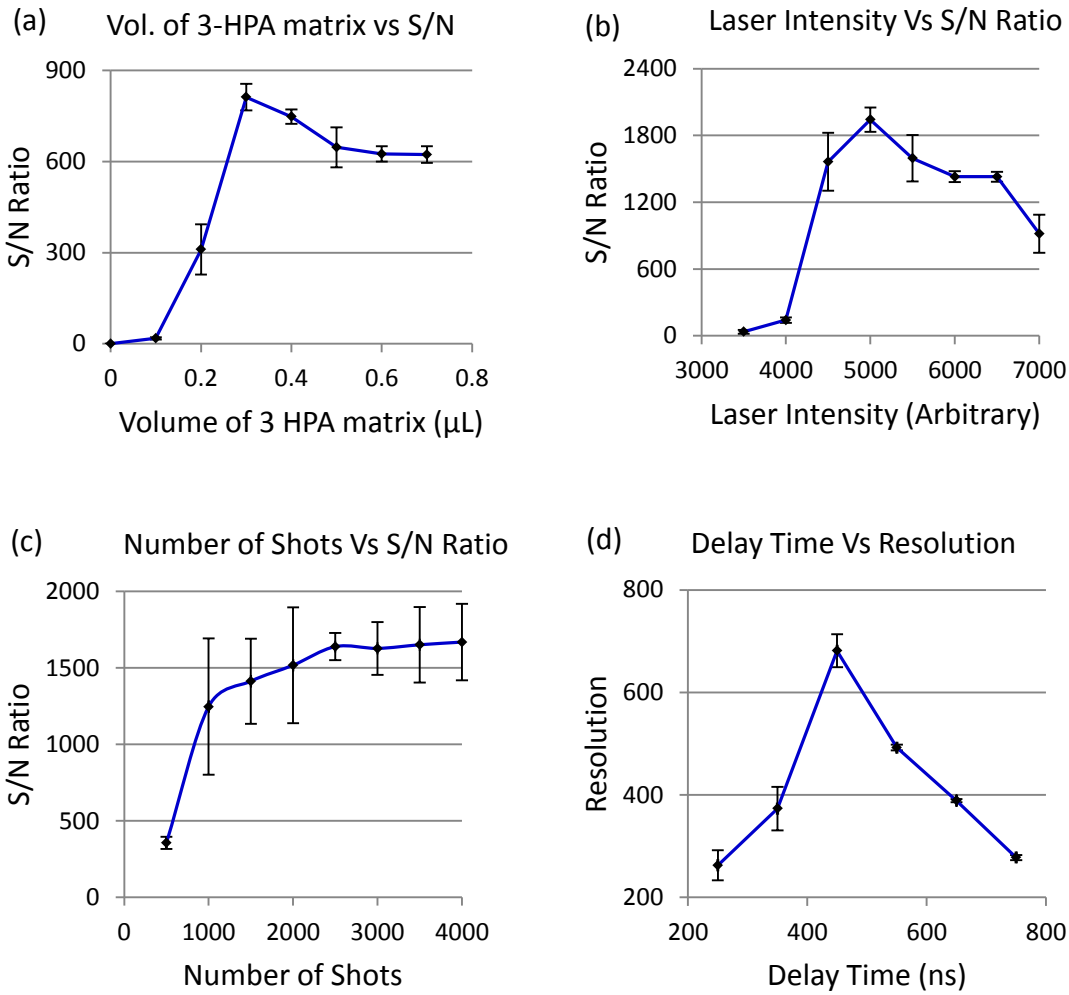


Figure 2.3. Optimization of MALDI matrix volume, laser intensity, number of laser shots and delay time

The results in Figure 2.3 were obtained by analysis of miR-183 and show (a) a graph of volume of 3 HPA matrix vs signal to noise ratio (b) laser intensity vs signal to noise ratio, (c) number of laser shots vs signal to noise ratio and (d) delay time vs peak resolution. The spectra were obtained using thin layer sample preparation technique in positive ion mode.

(b) Optimization of Enzymatic Extension

The Klenow Fragment (3′→5′ exo-) used in these studies can incorporate non-natural and modified nucleotides such as dideoxynucleotides or chemically modified nucleotides in a template depended manner, as long as the nucleotides are good substrates (Verma and Eckstein 1998). The Klenow Fragment used in these studies is a proteolytically cleaved product of E. coli DNA Polymerase I which retains its polymerase activity, but has lost its 5'→3' exonuclease activity (Klenow and Henningsen 1970). A mutation performed on D355A and E357A also abolishes the 3′→5′ exonuclease activity. The exonuclease inactivity of the Klenow Fragment polymerase used in this study was important in order to ensure that modified nucleotides used for miRNA extensions were not excised. The time required to complete substrate incorporation does vary considerably. Several groups that have done similar enzymatic incorporation have reported variable durations required to complete the enzymatic extension depending on the modified nucleotide being incorporated (Munafó and Robb 2010), (Lieu et al. 2005). It was therefore necessary to find optimal time for the enzymatic reaction. The reaction mixture was incubated as outlined in materials for different durations of time and analyzed for the extended miRNA. The results for single nucleotide incorporation are shown in Figure 2.4. 2.4 (a) shows miR-183 before extension 2.9 (b) shows the analysis of the miRNA after a 10 minute incubation period having extended miR-183 and residual unextended miR-183. It was therefore necessary to increase the incubation time in order to achieve a complete extension of the miRNA. Figure 2.4(c) shows the results obtained after a 20 minute incubation period. There was no residual unextended miR-153 observed

and therefore subsequent single nucleotide extension experiments were carried out by incubating the reaction mixture for 20 minutes. The double nucleotide enzymatic extension of miR-183 was optimized using a similar approach as the single nucleotide extension. Figure 2.5(a) and (b) shows the results of 10 and 20 minute reactions. After 10 minutes incubation, the reaction contained a mixture of unextended miR-183 and singly extended miR-183, further incubating the mixture for 20 minutes (2-5b) yielded a mixture of singly and doubly extended miR-183 with no peaks identified from unextended miR-183. Performing the reaction for a duration of 40 minutes enabled completion of the reaction as shown in Figure 2.5(c).

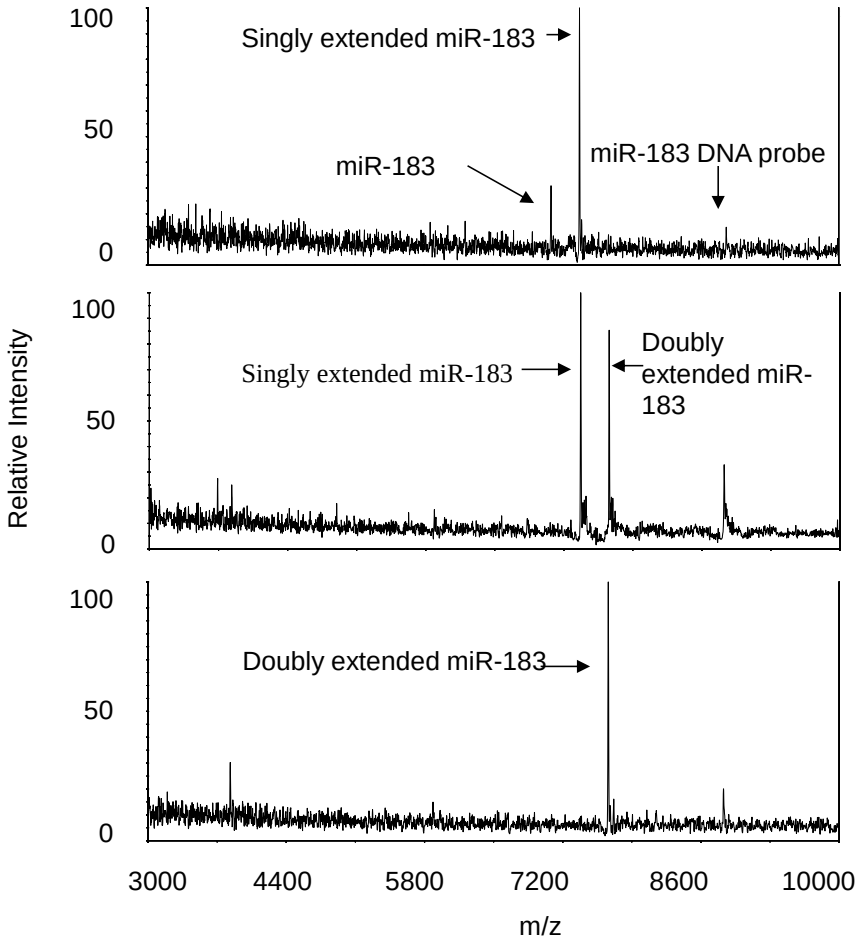


Figure 2.4. Optimization of time for single nucleotide extension of miR-183 using dCTP

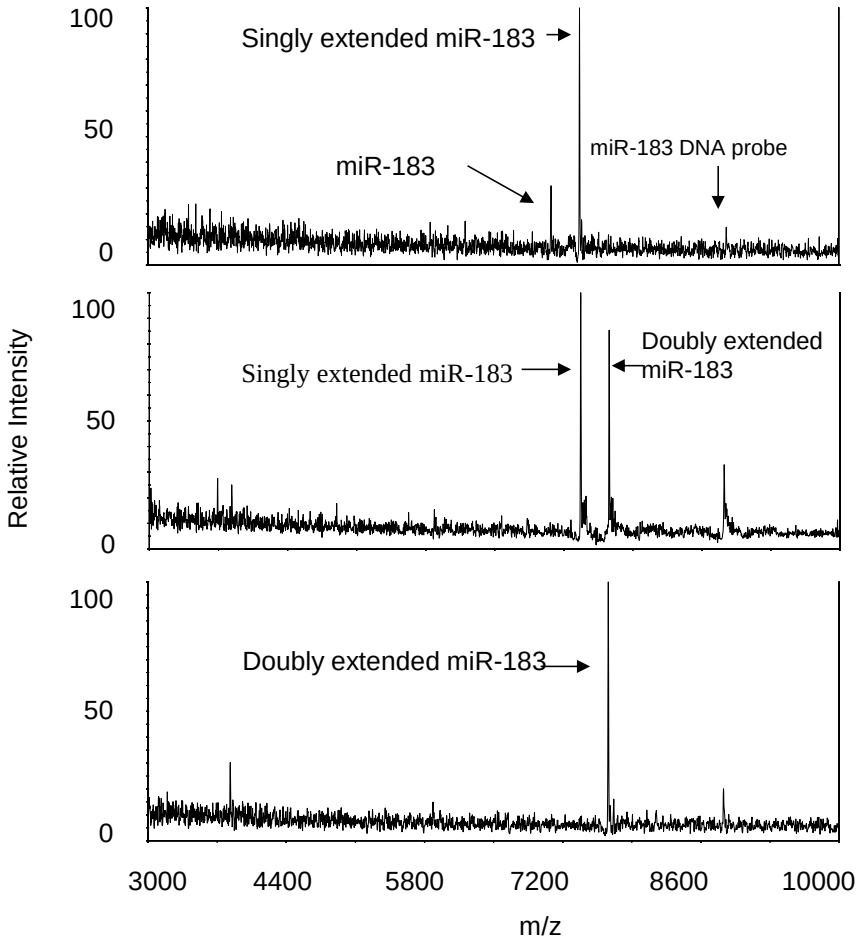


Figure 2.5. Optimization of time for double nucleotide extension of miR-183 using dCTP and ddTTP.

2.30 Results and Discussion

Recently, the use of MALDI-TOF MS to directly measure RNA has been reported by Limbach and his co-workers (Hossain and Limbach 2009), (Banoub et al. 2010). With the direct miRNA measurements, the MS approach is theoretically more accurate than those methods using a detection signal generated from a reporting label to determine the presence of a specific miRNA. However, before carrying out the MS measurements, specific miRNA ought to be separated from its aqueous sample, especially when cellular materials are involved. As demonstrated in a wide variety of molecular biology techniques, one of the most effective ways to isolate a specific miRNA with known RNA sequence is by using a complementary DNA probe to capture the miRNA target. However, with > 2,100 different human miRNA, plus other types of small RNA that may co-exist in the same biological sample, non-specific RNA hybridization to the DNA capture probe is unavoidable. This is further complicated by the fact that there are isobaric miRNA i.e. miRNA with different RNA sequences but identical molecular masses. For example, human miR-183-5p and six other human miRNA share the exact same molecular mass of 7083.13 Da. Thus, even with the use of high resolution mass spectrometry, the molecular ions of isobaric miRNA cannot be resolved. To address these issues and to ensure higher accuracy is achieved when MS is used to measure miRNA, we have developed a new concept which aims to create a unique signature in the mass spectrum of miRNA. To achieve this goal, one way is to digest a specific miRNA with a specific endonuclease prior to the MS measurement (Figure 2.1 a). Before the RNase

digestion, miRNA was first isolated from a sample by using a complementary DNA probe. The miRNA-DNA probe duplex was then immobilized on oligo dT coated magnetic beads and washed. The desalted miRNA was subsequently eluted from the probe and digested with RNase. For the proof of concept, human miR-183 which is up-regulated in glioblastoma was selected as a model in this study, and RNase T1 that cleaves the phosphate backbone at the 3' end of every guanosine was used to digest the selected miRNA. The digested miRNA fragments were then ready to be measured by using MS. Among various MS techniques, MALDI-TOF MS was chosen in this study because of its small sample size (0.3 µL), relative large mass range of up to 1.5 MDa, high sample throughput (~1 sec/sample), automatable data acquisition and the possibility to archive the same MALDI samples that have been measured.

Based on the mechanism for RNase T1 digestion shown in Figure 2.6, the cleavage of a phosphodiester bond on the RNA phosphate backbone would first give rise to a 2',3'-cyclic phosphate on each 5' digested fragment. The 2',3'-cyclic phosphate is considered to be the molecular structure of an intermediate. Following the ring opening of 2',3'-cyclic phosphate, each 5' digested fragment would have a linear 2' phosphate group at its 3' end (Douthwaite and Kirpekar 2007). The results that were obtained from the analysis of a sample that contained miR-183 are shown in Figure 2.7. To minimize the interference from the molecular ions of MALDI matrix and its clusters, the low mass cut-off in the data acquisition was set at 950 m/z. For this reason, the molecular ions of digested guanosine 3' monophosphate were not detected. As indicated in Figure 2.7, all other RNA fragments resulting from the digestion of miR-183 were identified. These

results exemplify the approach of using a DNA probe to capture a miRNA of interest and the subsequent digestion of miRNA with a specific RNase can be used to create a pattern of peaks which is unique to the miRNA being analyzed. Thus, the unique pattern of peaks is a mass signature that can be used to identify the miRNA. In this study, a unique pattern of peaks that correspond to the digested miRNA fragments is defined as an intrinsic mass signature as described in Figure 2.1 (a). This is because the pattern of peaks is dependent on the intrinsic property, i.e. RNA sequence, of a specific miRNA. In comparison to the use of a single peak which corresponds to an undigested miRNA to identify a specific miRNA, the concept of using an intrinsic mass signature, i.e. multiple peaks, is more accurate and robust. Following the isolation of miRNA, other specific endonuclease, like RNase A, can also be used to digest miRNA and create a unique intrinsic mass signature.

The digested RNA fragments that contained either 2',3'-cyclic phosphate or 3' phosphate were detected with an expected mass difference of 18 Da. To induce the ring opening, the incubation step of RNase T1 digestion was extended from 10 min to 20 min, but the two peaks that corresponded to the 2',3'-cyclic phosphate group remained detectable in repeated experiments. From the viewpoint of analyzing the spectral data, the presence of the peaks that correspond to 2',3'-cyclic phosphate group reinforce the identity of the peaks that correspond to the linear 2' phosphate group. Thus, in the case of the selected miR-183, the intrinsic mass signature would include all six peaks as shown in Figure 2.7.

To demonstrate the feasibility of the concept of creating a unique mass signature to accurately identify a specific miRNA, the miRNA-DNA probe dulplex was

immobilized on the surface of magnetic beads by using the simple and well established approach of specific molecular interactions between the poly(dA) tail on the DNA probe and oligo dT on the magnetic beads. It is important to note that other approaches for immobilizing the miRNA-DNA probe duplex can also be used, such as the immobilization of biotinylated DNA probe on streptavidin-coated magnetic beads.

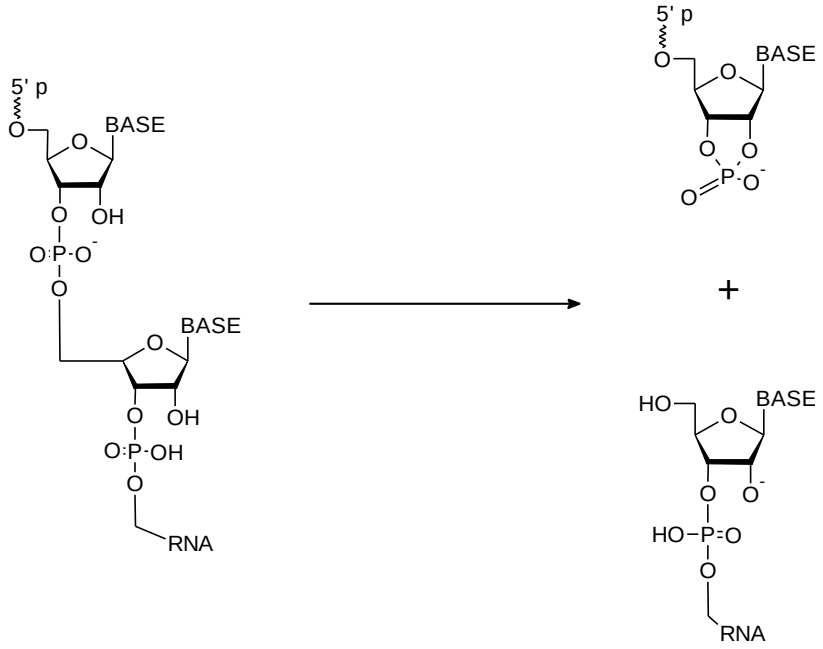


Figure 2.6. Mechanism of RNase hydrolysis.

Table 2.2. Expected RNase T1 digestion fragments of miR-183

Cleavage	Linear phosphate (Da)	2’ 3’ cyclic phosphate Da)	Sequence
U1:G4	1383.743	1365.728	pUAUGp
G5:G5	362.216	344.201	Gp
C6:G10	1607.962	1589.947	CACUGp
G11:G11	362.216	344.201	Gp
U12:G14	997.594	979.579	UAGp
A15:U22	2453.529	-	AAUUCACU

Unxtended miR-183 UAU GGC ACU GGU AGA AUU CAC U ($[M-H]^-$ = 7082.225)
These expected theoretical molecular masses of RNase T1 digestion from unextended miR-183 in negative ion mode MS were calculated using Mongo Oligo Mass Calculator version 2.

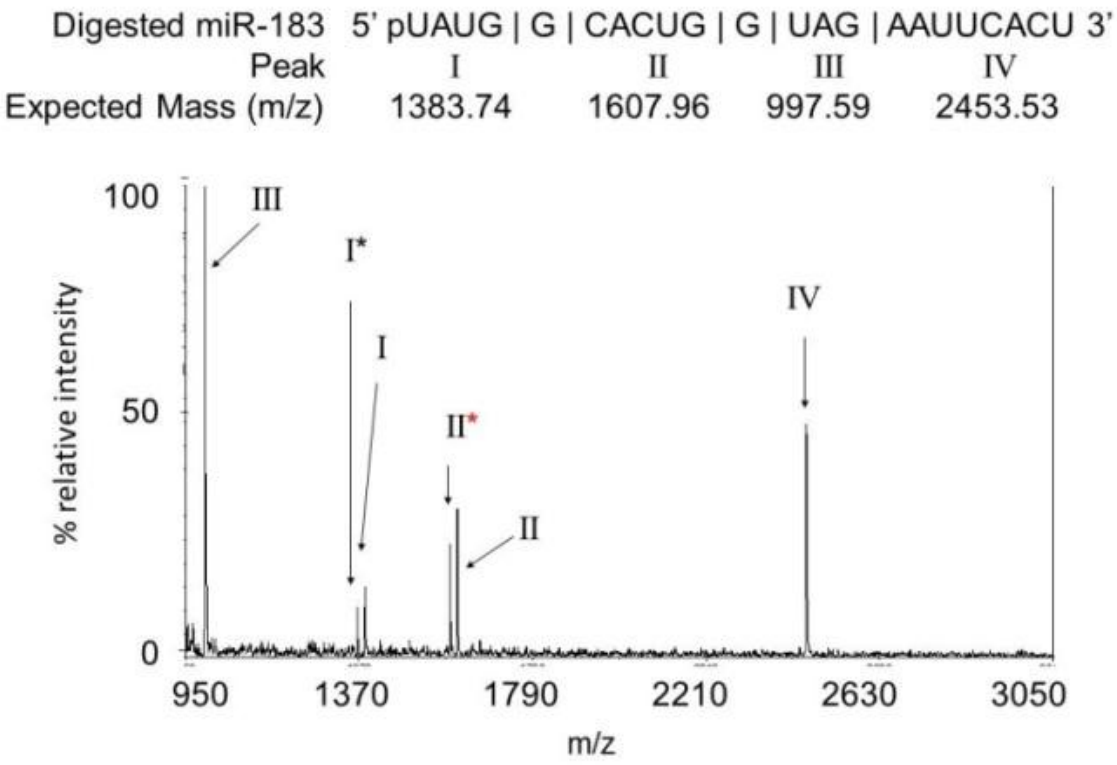


Figure 2.7. MALDI-TOF mass spectrum of miR-183 after RNase T1 digestion

The peak labels without asterisks (*) represent the digested fragment that contains a linear 2' phosphate group, and the peak labels with asterisks represent the digested fragment that contains a 2',3'-cyclic phosphate group. The spectrum was obtained with THAP matrix and linear negative ion mode. The THAP matrix solution was prepared by dissolving 20.0 mg of 2',3',4' THAP and 6.7 mg of ammonium citrate dibasic in 1.0 mL of 50% acetonitrile.

To overcome the limitation on the intrinsic property of miRNA, i.e. the molecular masses of digested miRNA fragments, an alternative approach to create a unique mass signature was developed (Figure 2.1b). The goal is to create a mass signature which is

independent on the intrinsic property of miRNA being analyzed. To achieve this goal, the target miRNA was first captured by using a complementary DNA probe. This was followed by using the upstream sequence of the DNA probe as a single-stranded template and combined with the activity of DNA polymerase to extend the captured miRNA from its 3' end. The outcome of this enzymatic 3' extension include: (1) using the substrate requirements of DNA polymerase to ensure the double-stranded portion of the duplex between miRNA and DNA probe was free from any mismatch of base pairs (Lee et al. 2008); (2) with appropriate selection on the upstream sequence of DNA probe and the complementary deoxyribonucleotide triphosphate substrate that were available to the 3' extension of miRNA, it ensured the annealing of miRNA was at the correct position on the DNA probe, otherwise the 3' extension of miRNA could not occur properly; and (3) the number and type of deoxribonucleotide that would be incorporated to the 3' end of miRNA were controlled by the upstream sequence of DNA probe and the available complementary deoxyribonucleotide triphospate substrate in the 3' extension reaction. As a result, an extended miRNA with a unique molecular mass was created, which is defined as an extended mass signature in this study. The extended mass signature is expected to have a different molecular mass when compare to the molecular masses of any other miRNA within the same sample. Recently, the ability of various DNA polymerases to accept an RNA fragment as a primer and the effectiveness to enzymatically incorporate different modified or labeled nucleotides to the 3' end of the RNA fragment have been examined (Sun et al. 2009), (Purohit et al. 2003). Five different DNA polymerases were evaluated, which included DNA polymerase I, Klenow fragment ($3'\rightarrow5'$ exo^-), Bst DNA

polymerase, Therminator DNA polymerase, and Taq DNA polymerase. Among them, only Taq DNA polymerase could not use RNA as a primer. Whereas, Klenow Fragment (3′→5′ exo$^-$) had shown the highest efficiency for nucleotide incorporation. In this study, Klenow fragment (3′→5′ exo$^-$) was selected to extend the captured miRNA. With the direct MS measurements, the 3' extension of miRNA could be extended with unmodified or unlabeled nucleotides. Following the 3' extension, the miRNA-DNA probe duplex was immobilized on oligo dT coated magnetic beads. After that, all the reagents from 3' extension were washed off, and the extended miRNA was eluted. Although the extended miRNA could be detected by several other MS techniques, MALDI-TOF MS was selected for the same reasons that have already mentioned earlier in this report. For the proof of concept, the same miRNA biomarker for glioblastoma, miR-183, and the corresponding DNA probe were used. The experimental protocol for creating the extended mass signature of miR-183 was systematically optimized. In comparison to a similar protocol for 3' extension that has been reported earlier, the amount of starting material i.e. miRNA was cut down by half to 20 pmol per assay, and the incubation time for 3' extension was reduced from 60 min to 20 min (Sun et al. 2009). For achieving the highest signal-to-noise ratio from using MALDI-TOF MS to measure the extended miR-183, the laser intensity, number of laser shots, and delayed time for ion extraction were optimized. By using the optimized protocol which is described in the section of materials and methods, miR-183 was extended with a single nucleotide. To ensure the 3' extension would be terminated after the incorporation of the first nucleotide, 2',3-dideoxyribonucleotide triphosphate (ddCTP) was used as the substrate for Klenow

fragment. The results are shown in Figure 2.8 (a). miR-183 ([miR-183+H]+ = 7084.24 m/z); (b) miR-183 that was extended with a single nucleotide (nt) of ddC ([miR-183+1nt+H]+ = 7357.44 m/z); and (c) miR-183 that was extended with two nucleotides, dC + ddT, ([miR-183+2nt+H]+ = 7661.65 m/z). In comparison to the MS measurement of an unextended miR-183 (Figure 2.8a), the mass difference between the extended and unextended miR-183 does correspond to the incorporation of an extra nucleotide to miR-183. As shown in 2-5b, there is no detectable unextended miR-183. This confirms that all the miR-183 present was successfully extended. To demonstrate the flexibility of creating an extended mass signature, in a separate experiment, miR-183 was extended with two nucleotides instead of only one nucleotide. This could be achieved by using an appropriate mixture of two different deoxyribonucleotide triphosphate as the substrate for Klenow fragment (dCTP and ddTTP), which was depended on the upstream sequence of the DNA probe for capturing miR-183. The results that were obtained from the extension of miR-183 with two nucleotides are shown in Figure 2.8 (c), in which neither unextended miR-183 nor premature terminated extension was detected.

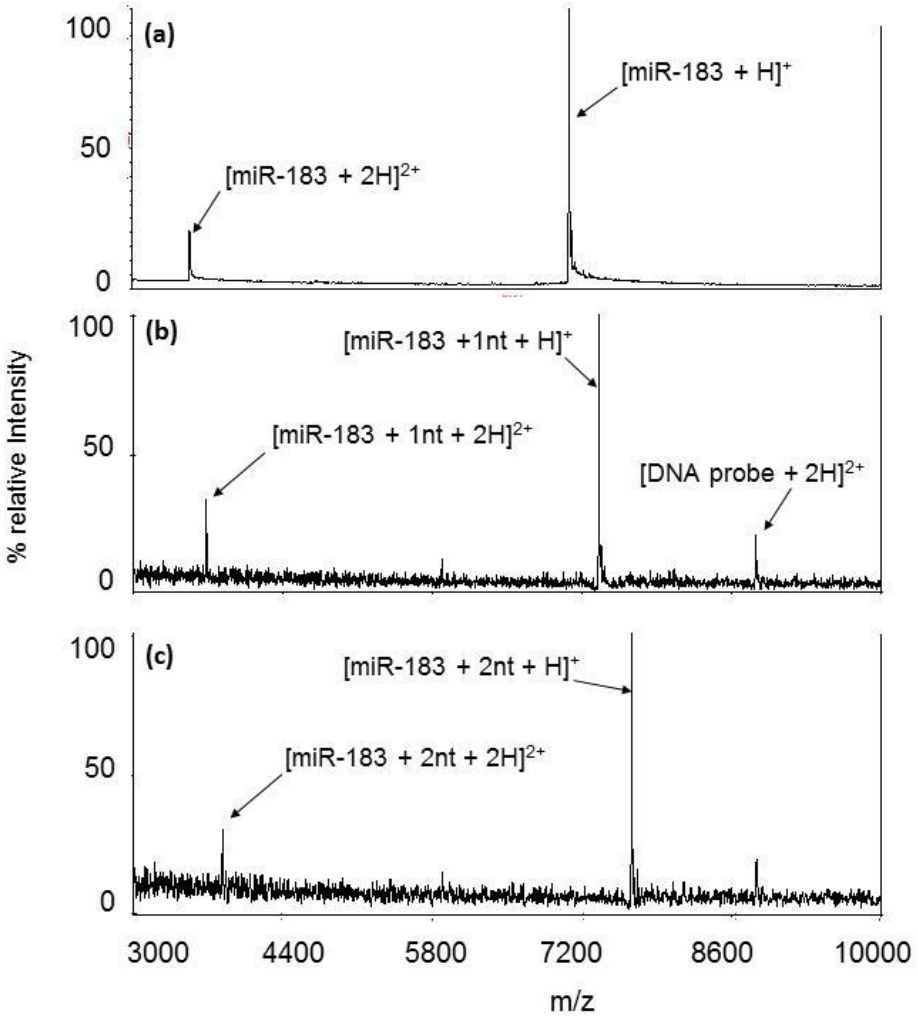


Figure 2.8. MALDI-TOF MS of (a) 1.2 pmol of unextended miR-183 ([miR-183+H]+ (b) miR-183 that was extended with a single nucleotide (nt) of ddC ([miR-183+1nt+H]+; and (c) miR-183 that was extended with two nucleotides, dC + ddT, ([miR-183+2nt+H]+

To confirm the specificity of creating an extended mass signature of miR-183, a sample that contained two different miRNA (miR-153 and miR-183) at equal concentration was analyzed. By using the low resolution linear MALDI-TOF MS to directly measure the mixture of miRNA, the two peaks that corresponded to the molecular ions of unextended miR-153 and unextended miR-183, respectively, could not

be resolved as shown in Figure 2.9 (a). This was due to the fact that the difference between the molecular masses of miR-153 and miR-183 is only 7.04 Da. When the mixture of miRNA was analyzed, both miRNA were captured by their corresponding DNA probe. It is important to note that the upstream sequences of the two DNA probes were designed to be different from each other. When only miR-183 was selectively extended with a single nucleotide (ddC), the same extended mass signature of miR-183 (7357.44 m/z) as in Figure 2.9 (b) was recreated despite the presence of another miRNA and its corresponding DNA probe (Figure 2.9b). In the second case, the same mixture of miRNA was analyzed and only miR-153 was selectively extended with a single nucleotide (ddG). As shown in Figure 2.9 (c), a unique extended mass signature of miR-153 (7404.51 m/z) was created while miR-183 remained unextended.

Table 2.3. Molecular masses of miR-153 and miR-183 before and after extension.

Oligo	Neutral Mass	Positively ion	Single nucleotide	Double nucleotide
hsa-miR-153	7090.272	7091.280	(ddGTP) 7404.509	
hsa-miR-183	7083.232	7084.240	(dCTP) 7373.434	
hsa-miR-183	7083.232	7084.240	(ddCTP) 7357.444	(dCTP+ddTTP) 7661.650

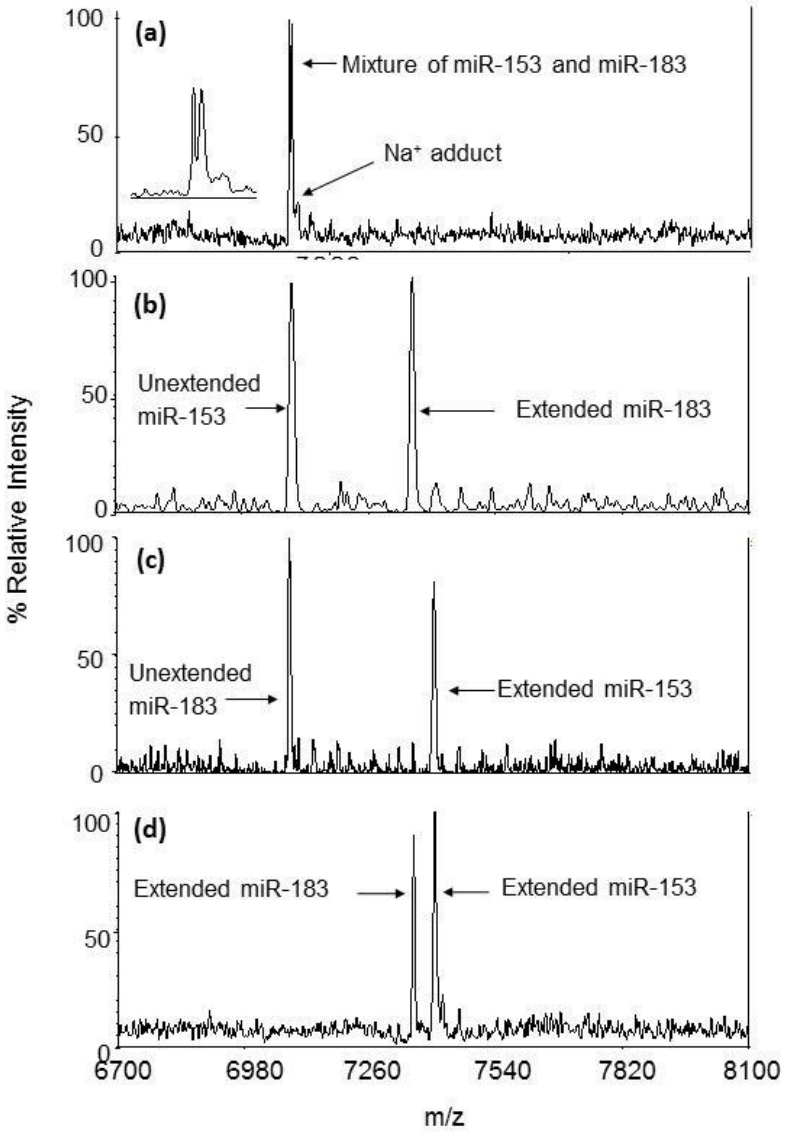


Figure 2.9. MALDI-TOF mass spectrum of (a) mixture of miR-153 and miR-183, (b) single nucleotide extension of miR-183 in the presence of miR-153, (c) single nucleotide extension of miR-153 in the presence of miR-183 and (d) simultaneous single nucleotide extension of both miR-183 and miR-153.

In another experiment, both miR-153 and miR-183 were simultaneously extended, and the same extended mass signatures of miR-153 and miR-183 were recreated despite the presence of more than one particular 2',3'-dideoxyribonculeotide triphosphate substrate in the 3' extension reaction (Figure 2.9d).

Table 2.4. Expected RNase T1 digestion fragments of extended miR-183

Extended miR-183 UAUGGCACUGGUAGAAUUCACUddC $[M-H]^-$ = 7355.433			
Cleavage position	Linear phosphate (Da)	2' 3' cyclic phosphate (Da)	Sequence
U1:G4	1383.743	1365.728	pUAUGp
G5:G5	362.216	344.201	Gp
C6:G10	1607.962	1589.947	CACUGp
G11:G11	362.216	344.201	Gp
U12:G14	997.594	979.579	UAGp
A15:C23	2726.737	-	AAUUCACUddC

These expected theoretical molecular masses of RNase T1 digestion from single nucleotide extended miR-183 in negative ion mode MS were calculated using Mongo Oligo Mass Calculator version 2.

The results in Figure 2.9 b-d have also demonstrated the use of extended mass signature to distinguish a specific miRNA from another miRNA, which has very similar molecular mass. Hence, the same approach can be used to distinguish isobaric miRNA. Furthermore, the above results exemplify the potential of using extended mass signature to analyse different miRNA in one single experiment. The results from the 3' extension of miR-183 was confirmed by digesting the extended miR-183 with RNase T1 as described earlier and the digested fragments were measured by using negative ion mode

MALDI-TOF MS and the results are shown in Figure 2.10. All the major peaks in the spectrum have been assigned to their respective digestion fragments.

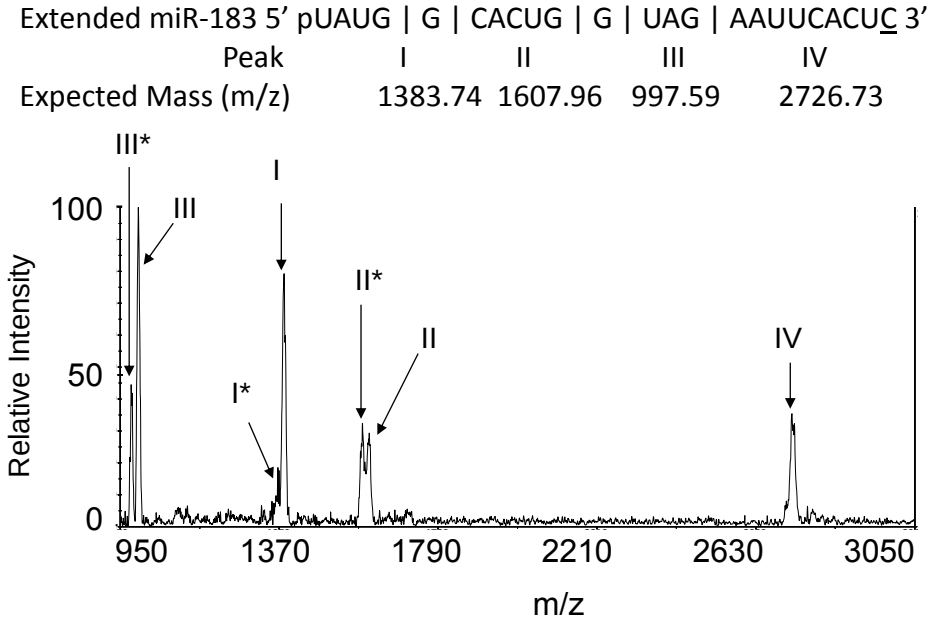


Figure 2.10. MALDI-TOF mass spectrum of extended human miR-183 after RNase T1 digestion.

In Figure 2.10, the peak labels without asterisks (*) represent the digested fragment that contains a linear 2’ phosphate group, and the peak labels with asterisks represent the digested fragment that contains a 2’,3’-cyclic phosphate group. The spectrum was obtained with THAP matrix and linear negative ion mode. The THAP matrix solution was prepared by dissolving 20.0 mg of 2’,3’,4’ THAP and 6.7 mg of ammonium citrate dibasic in 1.0 mL of 50% acetonitrile.

RNase T1 is a Guanine-specific fungal endonuclease that cleaves single stranded RNAs, releasing oligonucleotides from the guanosine 3’ phosphate terminal (Schatz et al.

2004). RNase T1 digestion yields predictable sequences of known oligonucleotides and has been previously used by Hartmer et al to deduce the sequence of a 25 mer RNA by MALDI-TOF MS (Hartmer et al. 2003). As shown on Figure 2.10, RNase T1 digestion of oligonucleotides gives rise to 5' cleavage fragments with a linear phosphate, however, 2'3' cyclic phosphates cleavage fragments that are an intermediate to the linear phosphate fragments are also known to coexist with the linear phosphate fragments (Hahner et al. 1997b). The RNase T1 digestion products of extended miR-183 is shown on table 1. The 2'3' cyclic phosphate intermediates are 18 Da lighter in mass compared to the 3' linear phosphate forms (Douthwaite and Kirpekar 2007). Both 2'3' cyclic phosphate and linear phosphate fragments were observed in the extended miRNA digest as shown in Figure 2.10.

It has been shown that the conversion of the 2'3' cyclic phosphate into the linear phosphate form occurs more readily when the identity of the nucleotide next to the last 3' nucleotide is an adenosine as opposed to a uridine or cytidine. The process of ring opening of the cyclic phosphate can be increased by increasing the time for digestion the RNA or by increasing the enzyme/substrate ratio (Kirpekar et al. 2000). Although 3-HPA is more commonly used for the analysis of RNA, 3-HPA matrix clusters appearing in the low mass window may overlap with analyte peaks interfering with the analysis of spectra. This phenomenon has also been observed when measuring RNA fragments from RNase digestion (Kirpekar et al. 2000). For this reason, 2',3',4' THAP matrix which exhibits matrix clusters in a much lower mass window was used. The peak corresponding to a

single guanine residue however lies within the lower end of the mass window of the spectrum and was therefore not identified.

2.40 Conclusions

Both methods for creating the mass signatures require the use of DNA probe to capture specific miRNA. The benefits of using the DNA probe in the future analysis of miRNA include to: (1) facilitate the separation of specific miRNA from a complex biological sample that may contain different RNA molecules, thus improving the accuracy on qualitative MS detection of specific miRNA; (2) allow multiple samples to be processed in parallel, for example in the microtiter plate format. In comparison to the approach of injecting samples sequentially into an LC-MS system, higher sample throughput can be achieved; (3) allow a sample that contains a low concentration of miRNA to be concentrated prior to the MS measurements. This can be achieved by either eluting the captured miRNA from the beads in smaller volume of water (as described in Materials and Methods) or start with a larger sample size and elute the capture miRNA from the beads in smaller volume of water; (4) no interference with MALDI MS measurements. The DNA capture probe does not interfere with the detection of miRNA. This is because the molecular mass of all DNA probe used in this study are about double the molecular mass of any miRNA that has been reported in the literature.

To improve the direct MS measurements of miRNA biomarkers, the concept of using a unique mass signature of miRNA is introduced in this report. The feasibility of creating a unique mass signature was demonstrated by using either specific RNase

digestion or miRNA-primed extension. In comparison to matching the measured molecular mass of an intact miRNA molecule to the expected molecular mass, the two different approaches to create a unique mass signature have improved the accuracy on the detection of specific miRNA. Since the creation of an intrinsic mass signature has relied on using the DNA probe to isolate a specific miRNA prior to its digestion with the RNase, the concept of intrinsic mass signature is considered to be different from the previous reports in which only RNase digestion was used to create similar pattern of peaks for RNA analysis (Hossain and Limbach 2007), (Hossain and Limbach 2009; Banoub et al. 2010). Furthermore, the mass signature concept has also included an alternative approach to create a mass signature by extending miRNA from its 3' end. The concept of mass signature is compatible to high throughput analysis of a specific panel of miRNA biomarkers. Previously, we have reported that the quantitation of nucleic acids by using MALDI-TOF MS was as accurate as the standard method of ultraviolet (UV) absorption spectrometry (Yang and Chiu 2010). Thus, the two different approaches to create the mass signatures can also be used to determine the amount of miRNA.

CHAPTER III

TOWARDS DE NOVO SEQUENCING OF microRNA USING BOTTOM-UP MALDI MASS SPECTROMETRY

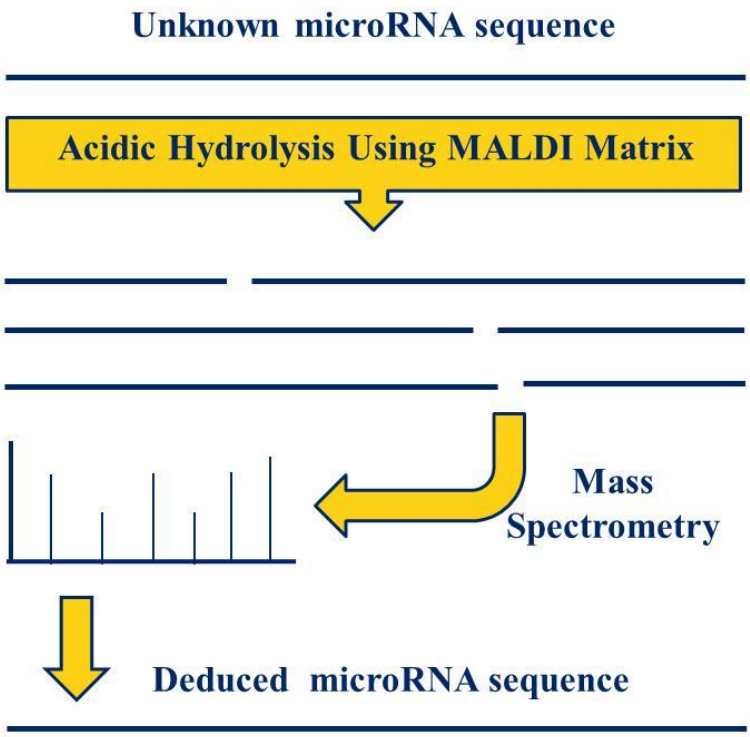


Figure 3.1. Schematic abstract depicting the work flow towards de novo sequencing of microRNA using bottom-up MALDI mass spectrometry.

3.10 Introduction

The increased interest in small non-coding RNA such as microRNA (miRNA) and small interfering RNA (siRNA) as potential targets for development of novel drugs (Park et al. 2009), as well as diagnostic tools for several human diseases (Calin and Croce 2006) has raised the demand for the development of more efficient, accurate and rapid methods for sequencing or sequence verification of the identities of such small endogenous or synthetic RNAs (Wark et al. 2008), (Oberacher and Pitterl 2011). So far, miRNAs have been implicated in the initiation and progression of more than 160 human diseases including numerous cancer types, heart disease and diabetes (Jiang et al. 2009). Sequencing of RNA also has numerous applications in fields such as genotyping, forensics, elucidation of structural features as well as in the analysis of oligonucleotide-bases for therapeutic and diagnostic purposes (Metzker 2005). New methods for sequencing of miRNAs can therefore be valuable tools for the development of miRNA as disease biomarkers or therapeutic products (Izumi et al. 2011).

RNA sequencing can be categorized as re-sequencing, a quality control or confirmatory measure in which the nucleic acid sequence is already known prior to analysis or *de novo* sequencing in which case the nucleic acid sequence is unknown. Traditionally, *de novo* sequencing of RNA has been carried out by Sanger-based sequencing methods, in which the identity of ribonucleotide (A, U, G or C) at each specific position is indirectly determined by measuring a detection label that corresponds to one of the four ribonucleotides (Morozova and Marra 2008). Sanger sequencing has however been deemed too expensive, time consuming, and laborious (Metzker 2005).

In the past few years, several alternate methods for the sequencing of nucleic acids have been developed. Amongst the emerging techniques, mass spectrometry (MS) stands out as a very promising approach for RNA sequencing. Recent advances in mass spectrometry in speed, accuracy and sensitivity as well as the capability for automation and high throughput operation have made MS an ideal tool for nucleic acid sequencing (Chait 2011).. These advances can be exploited for the development of cheaper and more efficient methods for both re-sequencing and de novo sequencing of miRNA. As opposed to the conventional RNA sequencing techniques such as Sanger and next-generation-sequencing that indirectly determine the RNA sequence by analyzing signals generated from secondary sources such as complementary DNA (cDNA) or fluorescence generation, MS offers a more direct RNA analysis by measuring mass which is an intrinsic property of the analyte and therefore circumvents the drawbacks associated with conventional RNA sequencing techniques. Moreover, MS does not require labeling or staining of samples prior to measurement. MS can therefore play a valuable role in the characterization of miRNA. An important step towards realizing the full potential of miRNA as therapeutic or diagnostic agents is the development of new sequencing methods that can be used for identification and characterization for miRNA. The development of such methods will provide scientists with important analytical tools and an opportunity to better understand gene regulatory functions of miRNA. Such tools will also be influential in propelling the breakthroughs in clinical diagnostics (Izumi et al. 2011).

The development of various "soft" ionization techniques in MS, such as matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI), has made it possible to measure the molecular mass of various biomolecules. Compared to other MS techniques, MALDI time-of-flight MS (MALDI-TOF MS) has been the preferred ion source for sequence analysis of RNA. MALDI results in the generation of spectra dominated by singly charged ions that are relatively easy to interpret compared to spectra with multiply charged inns from ionization sources such as ESI (Bahr et al. 2009). MALDI has therefore become an important ion source for the analysis of nucleic acids (Fu et al. 2006).

Here we report an approach to bottom-up sequencing of miRNA based upon partial hydrolysis and MALDI-MS. The temperature and time profile required to achieve partial miRNA hydrolysis with sufficient fragments yield and strong enough mass spectrometric signals to enable 100% sequence assignment were determined. *De novo* sequencing was carried out by using a ladder sequence determined from the mass differences between the monoisotopic peaks of the miRNA fragments combined with calculation of the base compositions of each fragment from exact mass measurements. A third approach is proposed based upon the average isotopic composition of a hypothetical average nucleotide "averageotide", analogous to the "averagine" method described by Senko, Beu and McLafferty for interpreting mass spectra of proteins and peptides (Senko et al. 1995).

3.20 Materials and Methods

Three synthetic RNA oligos shown in Table 3.1 were synthesized and HPLC purified by Integrated DNA Technologies (San Diego, CA, USA).

Table 3.1. Names, accession numbers and sequences of miRNAs used.

miRNA name	Accession number	Sequence (5'→3')
hsa-miR-153	MIMAT0000439	5' p-UUGCAUAGUCACAAAAGUGAUC-OH 3'
hsa-miR-183-5p	MIMAT0000261	5' p-UAUGGCACUGGUAGAAUUCACU-OH 3'
mml-miR-124a	MIMAT0002470	5' p-UUAAGGCACGCGGUGAAUGCCA-OH 3'

2,5 Dihydroxybenzoic acid (2,5 DHB) was purchased from Acros Organics (Morris Plains, NJ, USA). 3-hydroxypicolinic acid (3-HPA), sinapinic acid (SA), ammonium acetate and ammonium citrate dibasic were acquired from Sigma Aldrich (St. Louis, MO, USA). Acetonitrile and methanol, both of HPLC grade were obtained from Fisher Scientific (Pittsburgh, PA, USA). Non-sterile 0.22 µm low protein binding Durapore (PVDF) syringe driven membrane filter units (25mm) were purchased from Millipore Corp. (Bedford, MA, USA). Nanopure water from a Barnstead NANOpure Diamond Water Purification System (Dubuque, Iowa, USA) was used for the reconstitution and dilution of RNA oligos. The deionized water was autoclaved at 250 °F for 30 minutes.

3.21 Preparation of saturated MALDI matrices for hydrolyzing miRNA

The three MALDI matrices namely 3-HPA, SA and 2,5 DHB were separately dissolved in 10% and 50% acetonitrile to saturation by vortexing. The matrices were centrifuged and the saturated supernatant was drawn for hydrolysis of miRNA.. Saturated matrices were freshly before use.

3.22 Preparation of MALDI matrices for measuring hydrolyzed miRNA

MALDI matrices that were used for analysis of the hydrolyzed miRNA were prepared using standard procedures available in literature. 3-HPA was prepared by dissolving 35.0 mg of 3-HPA and 8.80 mg of ammonium citrate dibasic in 1.0 mL of 10% acetonitrile (Chiu 2009). 2,5 DHB was prepared by dissolving 20.0 mg of 2,5 DHB and 8.80 mg of ammonium citrate dibasic in 1.0 mL of 50% acetonitrile (Krause et al. 1999). Sinapinic acid was prepared by dissolving 10.0 mg of SA and 8.80 mg of ammonium citrate dibasic in 1.0 mL of 50% acetonitrile (Onnerfjord et al. 1998). The matrix solutions were vortexed for 1-2 minutes, filtered through 0.22 µm Durapore membrane filter and stored at −20 °C.

3.23 Preparation of hydrolyzed miRNA for MS measurement

For the partial hydrolysis of miRNA, a 2 µL solution containing 20 pmol of miRNA was mixed with 2 µL of saturated MALDI matrix. The mixture was incubated in

a thermocycler at 55, 65, 70 or 75 °C, aliquots of each sample were drawn after 30 and 60 minutes incubation and analyzed by MALDI-MS. The samples were prepared by the thin-layer technique as detailed by Barnes and Chiu (Chiu 2009). The stainless steel MALDI plate was cleaned with water and methanol. 0.3 µL of MALDI matrix was spotted on a MALDI sample plate and allowed to air dry before 0.3 µL of sample that had been hydrolyzed using the same type of matrix laid down on the plate was then spotted on top of the dried matrix and also allowed to air dry. The sample size corresponds to 3 pmol of miRNA.

3.24 Low-resolution MALDI- TOF MS instrument settings

AB 4700 MALDI-TOF/TOF mass spectrometer (Proteomics Analyzer, Applied Biosystems, Framingham, MA) was used for low-resolution MS measurements to develop and optimize the hydrolysis procedure. Each sample was measured by using the linear high mass positive or negative mode in the 4000 Series Explorer Version 3.0 software. The Nd:YAG laser (200 Hz) was set at 5000 arbitrary units. (maximum setting = 7900). The focus mass was the mass of an expected molecular ion from each sample. Molecular ions were extracted from the ion source after 400 ns delay. The accelerating voltage was +20.0 kV and grid voltage was +18.8 kV. The instrument was equipped with a 200 Hz digitizer. The sampling bin size was 0.5 ns with an input bandwidth of 500 MHz, and a vertical full scale of 500 mV. The linear detector voltage was +2.0 kV. The pressure inside the entire instrument was maintained at the level of 10^{-8} Torr. Each

spectrum was automatically acquired by accumulating the results of 3000 shots with random edge-biased positioning of each laser shot. By using the Data Explorer Version 4.6 software, resulting mass spectra were internally calibrated.

3.25 High Resolution JMS-S3000 SpiralTOF instrument settings

A second MALDI-TOF/TOF mass spectrometer (JMS-S3000 SpiralTOF™, JEOL Ltd., Tokyo, Japan) was used for high-resolution MS measurements. The SpiralTOF mass spectrometer is based on a unique multi-turn ion optic design that provides a very long (17-meter) flight path with high ion transmission efficiency (Satoh et al. 2011), (Satoh et al. 2007), (Satoh et al. 2006). The Spiral TOF measurements took place in a separate laboratory (JEOL USA, Inc. 11 Dearborn Road, Peabody, MA 01960, United States) and the samples were coded and not identified by their miRNA name.

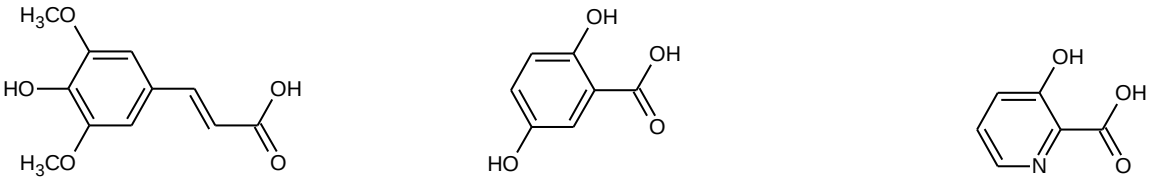
Using the sample preparation method described under materials and methods, samples were spotted onto a special 384-well plate that contains one additional external-reference spot for every four sample spots. A Nd-YLF laser with a wavelength of 349 nm operated at 500 Hz was used to desorb and ionize the samples. Data from 250,000 laser shots were accumulated for each mass spectrum. The extraction delay (250 ns) and the laser power were optimized to maximize the separation of the analyte isotope peaks. A peptide mixture containing Angiotensin I, Angiotensin II, P14R, ACTH 1-17, ACTH 18-39, insulin chain B oxidized and bovine insulin was used as an external calibration standard for exact mass measurements.

JEOL *MS Tornado* software was used to process the Spiral TOF mass spectra. Ladder sequences were determined from the mass differences between successive fragments by using native functions in the *MS Tornado* software and by using *New MS Tools* software (RBC Software) to examine mass spectra exported in text format. Nucleotide compositions were calculated by *New MS Tools* software. A Microsoft Visual Basic program was written to calculate the average nucleotide distribution for the human miRNA sequences in the miRNA database. Averageotide calculations were carried out in a Microsoft Excel spreadsheet. All spectra shown, both from the AB MALDI-TOF and the Spiral TOF were obtained in the positive ion mode.

3.30 Results and Discussion

Analysis of samples by MALDI requires that the samples be co-crystallized with a large molar excess of a matrix. The purpose of the matrix is to strongly absorb light in ultraviolet region (for UV laser MALDI) or infrared region (for IR laser MALDI) in order to desorb the analyte. The matrix also acts as a proton donor or receptor during the positive or negative ionization process of the sample respectively. The most commonly used MALDI matrices for RNA analysis are small weak organic acids such as sinapinic acid (SA), 2,5 Dihydroxybenzoic acid (2,5 DHB) and 3-hydroxypicolinic acid (3-HPA). The structures and pKas of these matrices are shown on Figure 3.2. The acidic nature of MALDI matrices presents a unique opportunity for the development of miRNA sequencing assays. We explored the ability of acidic MALDI matrices to hydrolyze

miRNA in order to create fragments that resemble sequencing ladders prior to measuring the hydrolyzed mixture using MALDI-TOF MS. Usually hydrolysis of RNA has previously been done suing various chemicals such as trifluoroacetic acid, however, these external additives have been shown to affect the ionization efficiency of nucleic acid, leading to the suppression of signal. It is therefore important that solvents and additives that do not affect the MALDI process be used for the success of RNA sequencing (Izumi et al. 2011).



SA (pKa = 4.53) 2,5 DHB (pKa = 3.01) 3-HPA (pKa = 1.14)

Figure 3.2. Chemical structures and pKa values of common MALDI matrices.

Since the developed assay utilizes the traditional MALDI matrix reagents which have been optimized for MALDI-MS measurements, interference from external sources and hence contamination and signal suppression is minimized.

It has been known for many years that RNA is susceptible to acidic hydrolysis through the cleavage of its phosphodiester bonds. The RNA hydrolysis mechanism proceeds through two stages as shown in Figure 3.3. In the first step, the phosphoryl oxygen is rapidly protonated, followed by a nucleophilic attack by the 2' hydroxyl on phosphorus to form an intermediate trigonal-bipyramidal phosphane. In the second step,

transfer of a proton to the 5’ oxygen directs the departure of the 5’ linked nucleotide, causing cleavage of the phosphodiester bond (Perreault and Anslyn 1997). The hydrolysis therefore leads to two sequencing ladders. The first ladder carries the original 5’ phosphate group and a 3' phosphate at the cleavage site; these fragments are termed here as the 5’ ladder, while the other ladder carries the original 3’ hydroxyl group and a 5’ hydroxyl at the cleavage site and are termed here as the 3’ ladder. The hydrolysis products are independent of pH but the yield of products is pH depended (Oivanen et al. 1998).

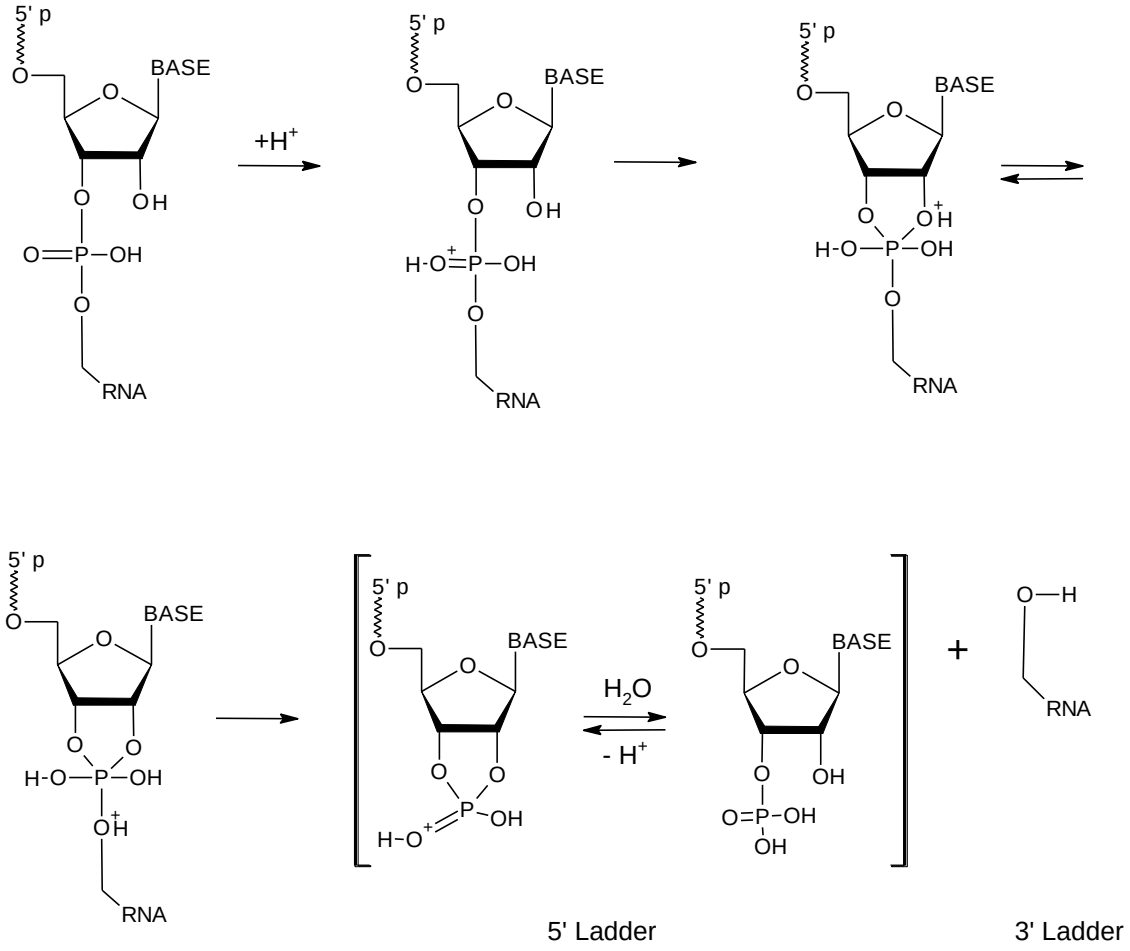


Figure 3.3. Schematic diagram for the acidic hydrolysis of RNA.

3.31 Re-sequencing results obtained using low resolution MALDI MS

For the initial hydrolysis studies, a low resolution mass spectrometer (AB 4700 TOF/TOF) was used. The three matrices (SA, 2,5 DHB and 3-HPA) were separately dissolved in 10% acetonitrile and incubated with miR-153 at 60 °C as explained in materials and methods. To maximize the hydrolysis process, the matrices were dissolved to saturation. After the hydrolysis, the resulting fragments were then analyzed by MS. The mass difference between adjacent peaks were calculated and matched to known

nucleotide masses. The results on sequence coverage achieved by hydrolyzing miR-153 using saturated SA, 2,5 DHB and 3-HPA at 60° C for 30 minutes and 60 minutes are shown in Figure 3.4. Analysis of the hydrolysis products after 30 minutes of incubation for SA, 2,5 DHB and 3-HPA yielded 22.7%, 69.7% and 74.2% sequence identification respectively. The incubation was extended to 60 minutes and SA, 2,5 DHB and 3-HPA yielded 45.5%, 62.1% and 83.3% sequence identification respectively. The matrices were dissolved in 50% acetonitrile and the experiments repeated. The results of sequence identification for SA, 2,5 DHB and 3-HPA at 30 min and 60 min incubation were 18.2%, 25.7%, 65.2% and 36.4%, 27.5%, 72.7% respectively.

There was a notable decrease in sequence identification when the matrices were dissolved in 50% compared to 10% acetonitrile. It is a well-known fact that for aqueous-organic solvents, the composition of the solvent has a profound effect on the degree of dissociation of any dissolved ionizable compounds. In the case of carboxylic acids such as SA, 2,5 DHB and 3-HPA, the pKa of these matrices increases with increasing content of acetonitrile. Consequently, the pH of the saturated matrices increases as the acetonitrile content is increased (Padró et al. 2012). As Oivanen et al noted, the yield of the hydrolysis products is pH dependent (Oivanen et al. 1998). Based on this phenomenon, it was hypothesized that the pKa of the saturated MALDI matrices dissolved in 10% acetonitrile is considerably lower and so is the pH when compared with matrices dissolved in 50% acetonitrile. Subsequently, a higher yield of the miRNA hydrolysis fragments was achieved when using solutions dissolved in 10% acetonitrile translating to a higher sequence coverage. The results shown in Figure 3.4 are

representative of a triplicate measurement in which all three acquired spectra were used to determine the sequence coverage. From Figure 3.4, when comparing the ability of three matrices for the hydrolysis of miRNA to create sufficient 3' and 5' ladders for the sequencing of miRNA 3-HPA hydrolysis gave the highest percent sequence coverage compared to SA and 2,5 DHB.

For the optimization of the yield of hydrolysis products, the hydrolysis temperature and time were also investigated. The temperature profile for miR-153 hydrolysis using 3-HPA was performed on the AB MALDI-TOF by incubating miR-153 with saturated 3-HPA at 55, 60, 65, 70 and 75 °C, and measuring the samples after 30 and 60 minute durations. As shown in Figure 3.5 the sequence identification for the five temperature settings after 30 minute and 60 minutes hydrolysis was 40.9, 74.2, 89.4, 84.9, 77.2% and 59.1, 84.9, 81.8, 74.2, 71.2% respectively.

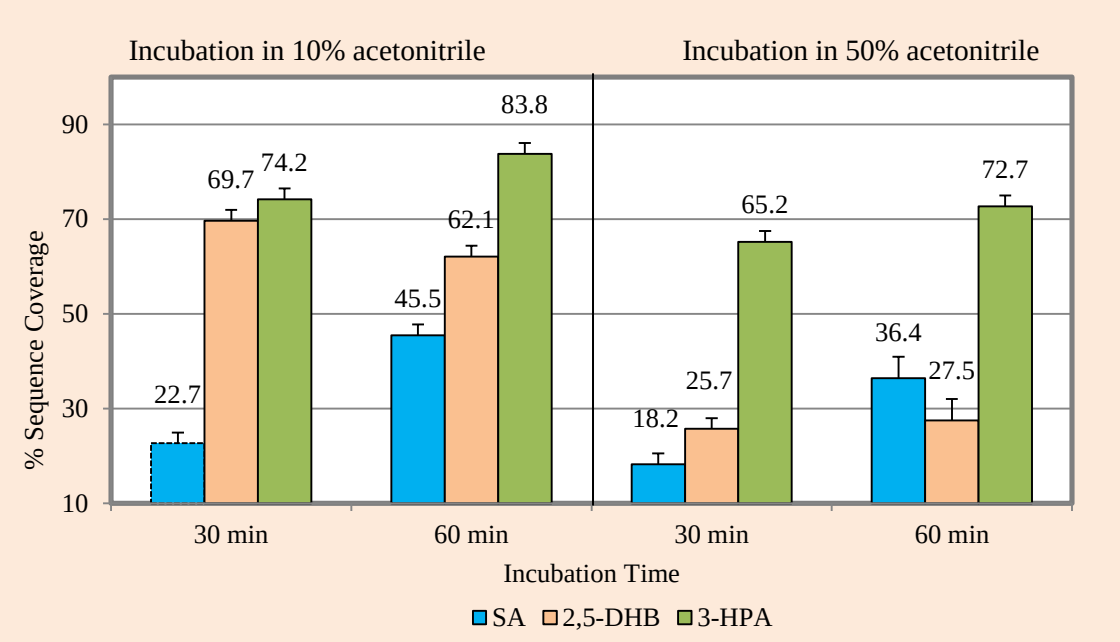


Figure 3.4. Sequence coverage achieved by hydrolyzing miR-153 using saturated SA, 2,5 DHB and 3-HPA at 60° C.

Minimal hydrolysis of the control samples was observed at 55, 60 and 65 °C; however, this rate doubled when the temperature was elevated to 70 and 75 °C. The increased hydrolysis is attributable to the elevated incubation temperature. The results shown in Figure 3.5 are representative of a triplicate measurement in which all three acquired spectra were used to determine the sequence coverage. Hydrolysis of miR-153 using 3-HPA saturated in 10% acetonitrile at 65°C for 30 minutes yielded the highest sequence identification, these conditions were therefore used for all subsequent hydrolysis experiments.

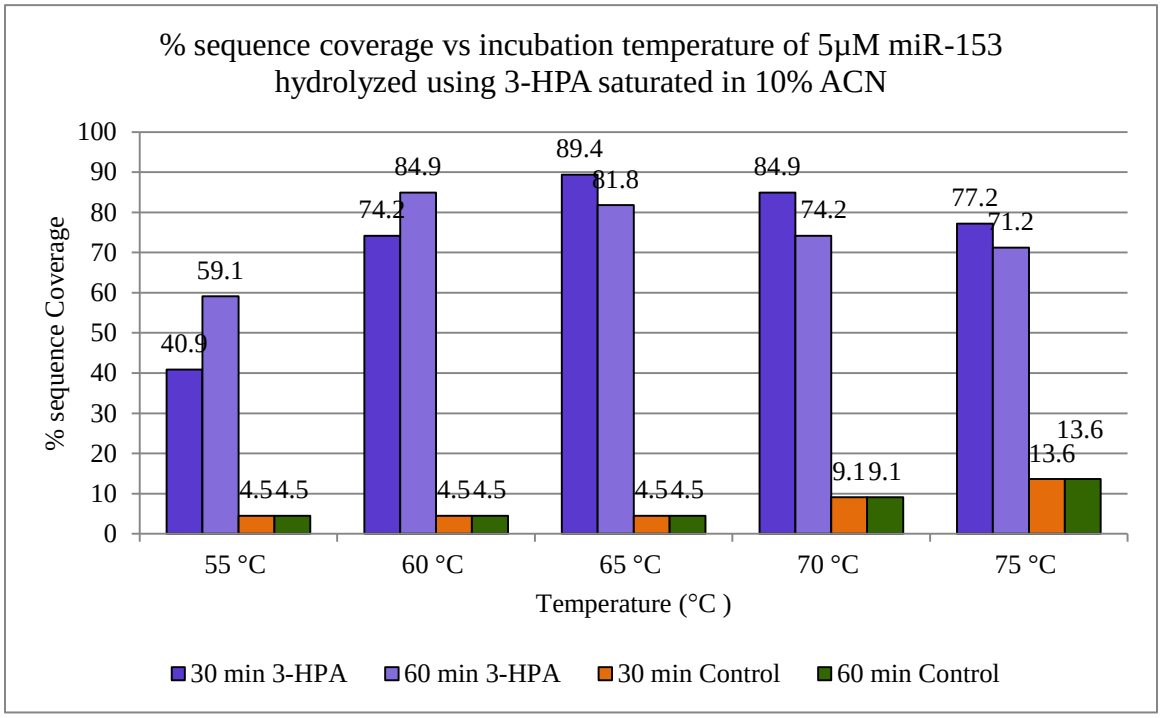


Figure 3.5. Sequence coverage achieved by incubation of miR-153 with 3-HPA in saturated 10% acetonitrile at different temperatures.

From the previous results in figure 3.4, it was observed that higher sequence coverage was obtained when the matrices were dissolved in 10% acetonitrile compared to 50% acetonitrile. Figure 3.6, 3.7 and 3-8 shows mass spectra resulting from the AB MALDI-TOF analysis of miR-153 using SA, 2,5 DHB and 3-HPA dissolved in 10% acetonitrile respectively. The hydrolysis was performed at 60 °C. The solid lines are used to identify the 3' ladder and the dashed lines are used to identify the 5' ladder. The overall sequence coverage achieved using SA, 2,5 DHB and 3-HPA when considering only one spectrum was 72.7%, 81.8% and 81.9% respectively, however, when considering the three replicate spectra from a single measurement, the sequence coverage obtained from SA, 2,5 DHB and 3-HPA was 72.7%, 81.8% and 89.4% respectively. In

deciding the best matrix to use for the hydrolysis of miRNA, several factors were taken into consideration 1) it was noted that spectra from SA and 2,5 DHB had wider peaks and therefore poor mass resolution which presents a challenge to assignment of nucleotide sequences especially for ladders differing by a single U or C nucleotide. 2) the overlap of the two ladders reinforce and support the sequence deduction, and it was noted that although the spectra from both SA and 2,5 DHB had more 5' ladder fragments, they exhibited fewer overall 3' ladders when compared to 3-HPA ladders and 3) spectra resulting from SA and 2,5 DHB hydrolysis showed some peaks that could not be assigned to neither the 5' nor the 3' ladder. It is hypothesized that these unassigned fragments could be as a result of base loss from the miRNA fragments. Based on these facts, 3-HPA was selected as the most suitable matrix for the hydrolysis and sequence determination of miRNA.

5’ p-UUGCAUAGUCACAAAAGUGAUC-OH 3’

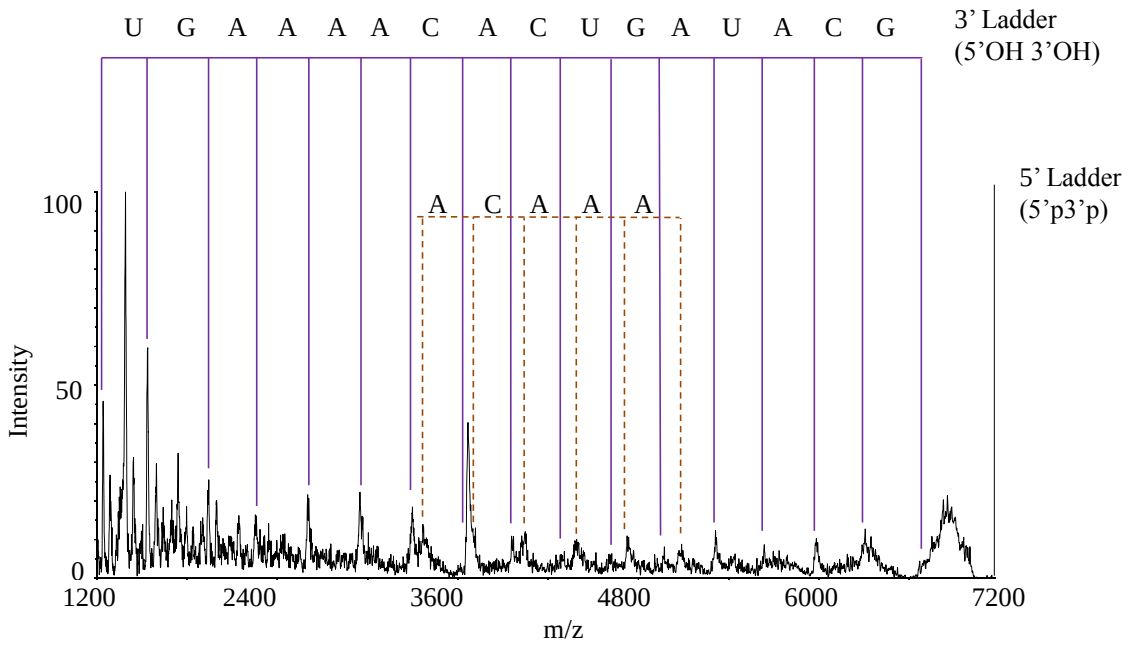


Figure 3.6. Mass spectrum of miR-153 hydrolyzed using saturated SA in 10% acetonitrile at 60 °C.

5’ p-UUGCAUAGUCACAAAAGUGAUC-OH 3’

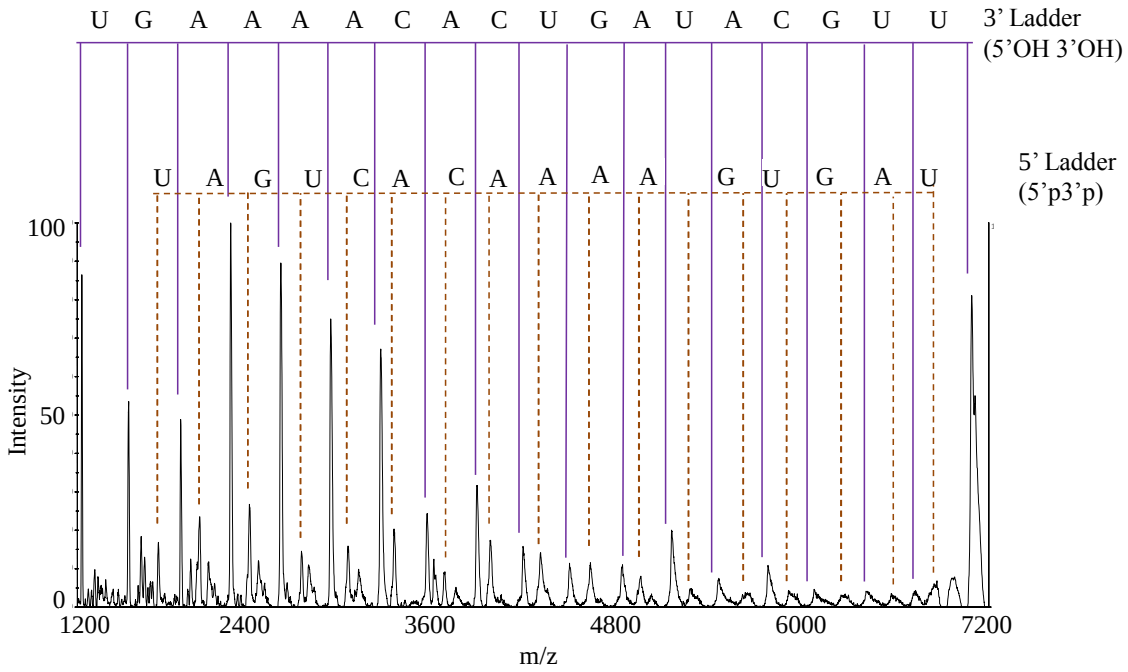


Figure 3.7. Mass spectrum of miR-153 hydrolyzed using saturated 2,5 DHB in 10% acetonitrile at 60 °C.

5’ p-UUGCAUAGUCACAAAAGUGAUC-OH 3’

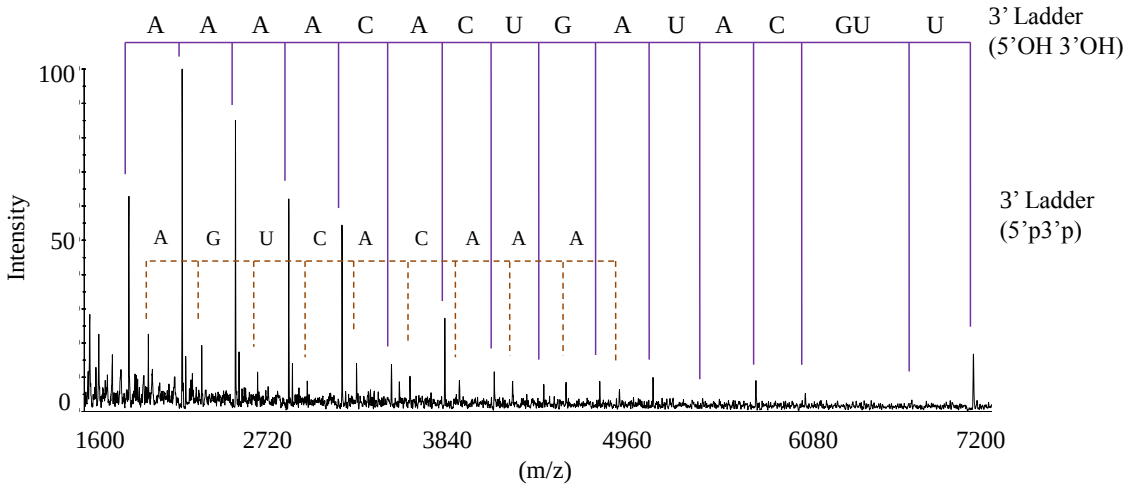


Figure 3.8. Mass spectrum of miR-153 hydrolyzed using saturated 3-HPA in 10% acetonitrile at 60 °C.

Mass accuracy calculations of 5’ ladder obtained from the hydrolysis of miR-153 using 3-HPA saturated in 10% acetonitrile on the AB MALDI-TOF (low resolution) are shown in Table 3.2. The average mass accuracy using the singly charged average mass peaks was 59.65 ppm. It is important to note the mass accuracy achieved when using the AB 4700 TOF/TOF mass spectrometer was not sufficient to distinguish fragments differing by a U or C nucleotide due to the low resolution. Since the sequence of miR-153 was known prior to the, this experiment is considered as a sequence confirmation i.e. re-sequencing.

Table 3.2. Mass accuracy of 3’ ladder fragments from hydrolyzed miR-153

3’ Ladder	Theoretical average *m/z*	Observed average *m/z*	Mass difference	Mass accuracy (ppm)
GAUC	1223.806	1223.7589	0.0471	38.486
UGAUC	1529.975	1529.9674	0.0076	4.967
GUGAUC	1875.184	1875.2865	0.1025	54.661
AGUGAUC	2204.393	2204.4194	0.0264	11.976
AAGUGAUC	2533.602	2533.5078	0.0942	37.180
AAAGUGAUC	2862.811	2862.4785	0.3325	116.144
AAAAGUGAUC	3192.02	3192.3406	0.3206	100.437
CAAAAGUGAUC	3497.203	3497.7473	0.5443	155.638
ACAAAAGUGAUC	3826.413	3826.6326	0.2196	57.390
CACAAAAGUGAUC	4131.597	4131.3413	0.2557	61.888
UCACAAAAGUGAUC	4437.766	4437.344	0.422	95.092
GUCACAAAAGUGAUC	4782.974	4783.127	0.153	31.988
AGUCACAAAAGUGAUC	5112.183	5112.2466	0.0636	12.440
UAGUCACAAAAGUGAUC	5418.352	5418.9697	0.6177	114.001
AUAGUCACAAAAGUGAUC	5747.561	5747.9619	0.4009	69.751
CAUAGUCACAAAAGUGAUC	6052.745	6052.512	0.233	38.494
GCAUAGUCACAAAAGUGAUC	6397.954	6397.594	0.36	56.267
UGCAUAGUCACAAAAGUGAUC	6704.123	-	-	-
pUUGCAUAGUCACAAAAGUGAUC	7090.272	7090.392	0.12	16.924
Average mass accuracy				59.651 ppm

The sequencing ladder was obtained from low resolution AB 4700 MALDI measurements of miRNA hydrolyzed at 60 °C using 3-HPA saturated in 10% acetonitrile.

3.32 De novo sequencing results obtained using JMS-S3000 SpiralTOF

The approach of assigning nucleic acid sequences using mass differences from mass spectral data of nucleic acids was proposed in 1993 by Pieles et al (Pieles et al. 1993) and has become a well-established procedure for interpreting sequence data from

sequencing ladders. The application of mass differences technique to assign RNA sequences has however been much slower compared to its use for DNA sequences. Sequencing of DNA is easier partly because of the large minimum mass difference between the nucleotides adenosine and thymine (>9 Da), compared to the minimum mass difference between RNA nucleotides uracil and cytidine (1 Da). It is therefore easier to unambiguously assign sequence information to fragments of DNA compared to RNA due to limitations in instrument resolving power (Tolson and Nicholson 1998). To overcome this challenge, the JMS-S3000 SpiralTOF MALDI MS which can attain a resolution of >60,000 over a large mass range was used in this study for *de novo* sequencing of miRNA (Satoh et al. 2007). The three miRNAs (miR-183, miR-124 and miR-153) were hydrolyzed using, 3-HPA saturated in 10% acetonitrile at 65 ^{0}C for 30 minutes and the resulting mass ladders analyzed using the JMS-S3000 SpiralTOF. The mass differences between adjacent ladder peaks were calculated and matched against the known mass of ribonucleotides A, U, G and C. Once a match was found, the nucleotide was assigned to the mass difference. This was done for both the 5' and 3' sequence ladders.

Following the establishment of the optimal miRNA hydrolysis conditions using MALDI matrices with low resolution mass spectrometry, the developed method was tested on two coded blind samples of hsa-miR-183 (miR-183) and mml-miR-124 (miR-124). The two miRNAs were analyzed by a separate laboratory (JEOL USA, Inc. 11 Dearborn Road, Peabody, MA 01960, United States). The sequence elucidation of the blind samples is therefore considered *de novo* sequencing of miRNA. In addition to the two blind samples, miR-153 which was used as a model in the low resolution

optimization studies was also analyzed in the same laboratory. To establish the resolution and mass accuracy at high molecular weight, the two intact miRNAs, miR-183 and miR-124 were measured by the JMS-S3000 SpiralTOF. Theoretical simulation of mass spectra for the molecular ions of miR-183 and miR-124 was generated using *msTornado Analysis* software and were compared to the measured spectra. Figure 3.9 (a) and (b) are the measured and theoretical mass spectra of miR-183 while 3-10 (a) and (b) represent the measured and theoretical mass spectra of miR-124. The mass accuracy obtained for the molecular ions of miR-183 and miR-124 is tabulated in table 3.3 and 3-4 respectively. A mass accuracy of 4.1705 ppm and 9.5590 ppm were achieved for miR-183 and miR-124 respectively. In each case, isotopic resolution was achieved with a resolving power of 20,000 full width half maximum (FWHM).

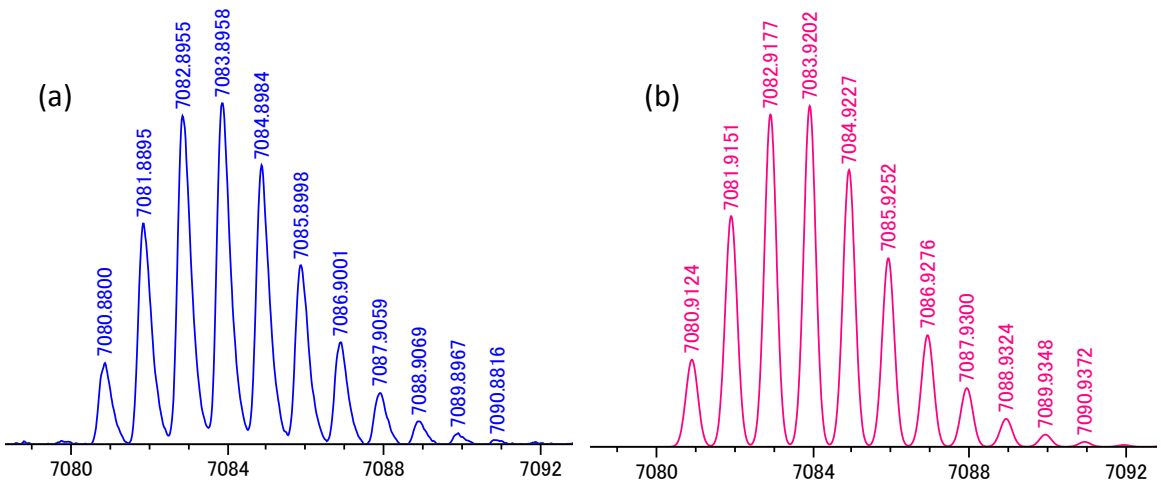


Figure 3.9. (a) Measured and (b) theoretical simulation of isotopic distribution of [miR-183 + H]$^{+}$.

Table 3.3. Mass accuracy attained on the isotopic distribution of miR-24

Measured Mass (M)	Theoretical Mass (T)	M-T	Mass Accuracy
7080.8800	7080.9124	-0.0324	4.5756
7081.8895	7081.9151	-0.0256	3.6148
7082.8955	7082.9177	-0.0222	3.1343
7083.8958	7083.9202	-0.0244	3.4444
7084.8984	7084.9227	-0.0243	3.4298
7085.8998	7085.9252	-0.0254	3.5845
7086.9001	7086.9276	-0.0275	3.8803
7087.9059	7087.9300	-0.0241	3.4001
7088.9069	7088.9324	-0.0255	3.5971
7089.8967	7089.9348	-0.0381	5.3738
7090.8816	7090.9372	-0.0556	7.8409
Average mass accuracy			4.1705 ppm

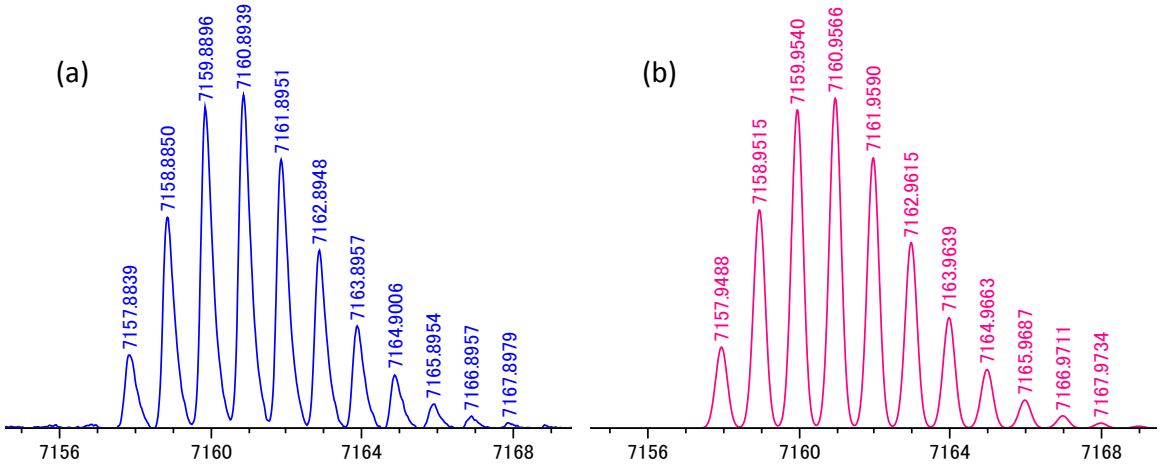


Figure 3.10. (a) Measured and (b) theoretical simulation of isotopic distribution of [miR-124 + H]$^{+}$.

Table 3.4. Mass accuracy attained on the isotopic distribution of miR-124

Measured Mass (M)	Theoretical Mass (T)	M-T	Mass Accuracy
7157.8839	7157.9488	-0.0649	9.0668
7158.8850	7158.9515	-0.0665	9.2890
7159.8896	7159.9540	-0.0644	8.9944
7160.8939	7160.9566	-0.0627	8.7558
7161.8951	7161.9590	-0.0639	8.9221
7162.8948	7162.9615	-0.0667	9.3117
7163.8957	7163.9639	-0.0682	9.5198
7164.9006	7164.9663	-0.0657	9.1696
7165.8954	7165.9687	-0.0733	10.2289
7166.8957	7166.9771	-0.0814	11.3576
7167.8979	7167.9734	-0.0755	10.5329
Average Mass accuracy			9.5590 ppm

Figure 3.11 is a MALDI mass spectrum of partially hydrolyzed miR-183 obtained using the high resolution JMS-S3000 SpiralTOF. 3-10(a) shows the peaks that correspond to the 5’ ladder while 3(b) shows peaks that correspond to the 3’ ladder. In total, 17 unique 5’ fragments and 18 unique 3’ fragments as well as the precursor ion of

miR-183 were identified. In the sequence above Figure 3.11 (a) the fragments identified in the 5' ladder are underlined while those identified in the 3' ladder are overlined. Since the sequencing can done from both the 5' and 3' ladders, the two ladders overlap and complement each other e.g. the bolded region in the sequence above Figure 3.11 represents a 59.1% sequence overlap between the 5' and 3' ladders of miR-183. Besides increasing the confidence of making the correct sequence identification, the intrinsic degeneracy provided by the two overlapping ladders makes it possible to deduce the sequence of a miRNA even in cases where all the fragment peaks are not detected from one sequencing ladder. Isotopic resolution of the hydrolyzed fragments was achieved and the results of the isotopic distribution of selected peaks are shown in Figure 3.12, with the 3' ladder subspectra being marked with *. The resulting hydrolyzed fragments, their theoretical masses, observed mass, difference between theoretical and observed masses as well as the mass accuracy for both 5' and 3' ladders for miR-183 are tabulated in Tables 3-5 and 3-6 respectively. The 5' ladder was detected with an average mass accuracy of 2.11 ppm while for the 3' ladder, an average mass accuracy of 1.95 was attained.

miR-183 5’ pUAUG**GCACUGGUAGAAU**UCACU-3’

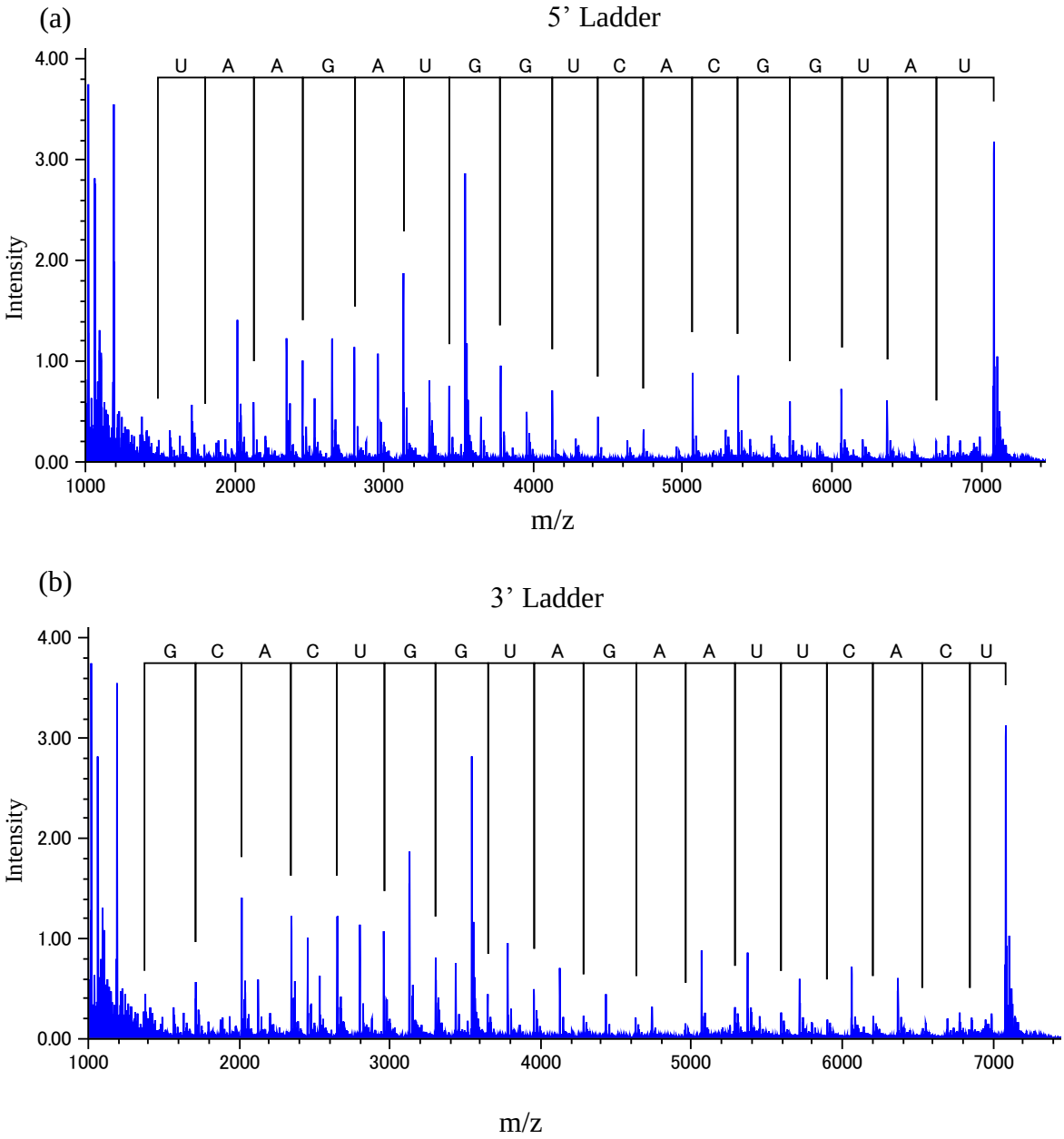


Figure 3.11. (a) 5’ ladder and (b) 3’ ladder of hydrolyzed miR-183 obtained using the JMS-S3000 SpiralTOF.

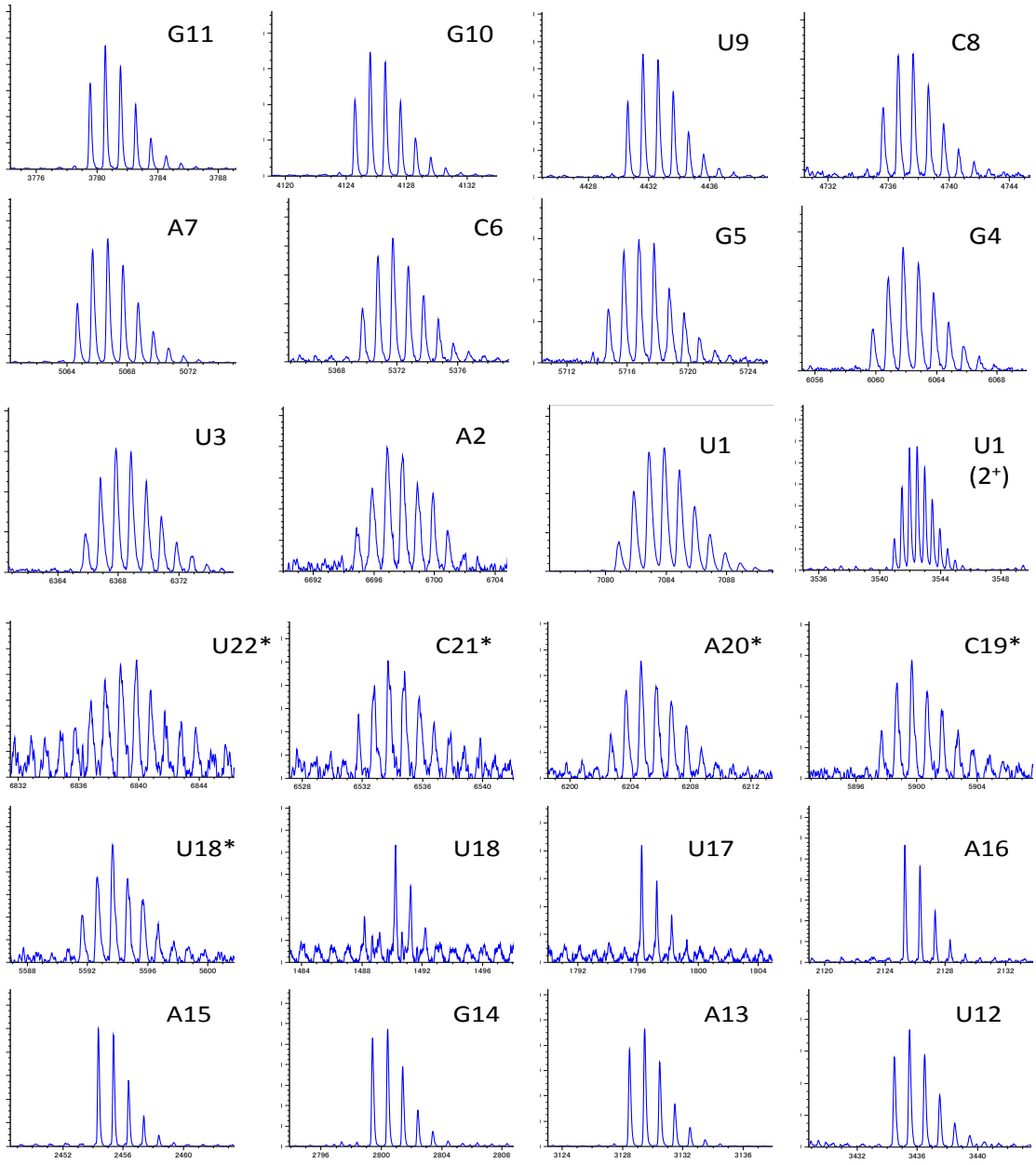


Figure 3.12. Subspectra from 5’ and 3’ ladders of hydrolyzed miR-183 showing the isotopic resolution.

Table 3.5. Mass accuracy of 5’ ladder fragments from hydrolyzed miR-183

5’ Ladder	Theoretical Mass (Da)	Observed Mass (m/z)	Mass Difference (Da)	Mass Accuracy (ppm)
pUAUG	1367.12416	1367.1172	0.007	5.09
pUAUGG	1712.17160	1712.1729	-0.001	-0.76
pUAUGGC	2017.21289	2017.2159	-0.003	-1.49
pUAUGGCA	2346.26540	2346.2705	-0.005	-2.17
pUAUGGCAC	2651.30669	2651.3093	-0.003	-0.98
pUAUGGCACU	2958.33471	2958.3404	-0.006	-1.92
pUAUGGCACUG	3303.38212	3303.3896	-0.007	-2.26
pUAUGGCACUGG	3648.42954	3648.4383	-0.009	-2.40
pUAUGGCACUGGU	3954.45486	3954.4617	-0.007	-1.73
pUAUGGCACUGGUA	4283.50736	4283.5116	-0.004	-0.99
pUAUGGCACUGGUAG	4628.55479	4628.5570	-0.002	-0.48
pUAUGGCACUGGUAGA	4958.60982	4958.6071	0.003	0.55
pUAUGGCACUGGUAGAA	5287.66233	5287.6549	0.007	1.41
pUAUGGCACUGGUAGAAU	5593.68767	5593.6738	0.014	2.48
pUAUGGCACUGGUAGAAUU	5899.71301	5899.6895	0.024	3.98
pUAUGGCACUGGUAGAAUUC	6204.75431	6204.7474	0.007	1.11
pUAUGGCACUGGUAGAAUUCA	6533.80681	6533.7727	0.034	5.22
pUAUGGCACUGGUAGAAUUCAC	6838.84811	6838.8370	0.011	1.62
pUAUGGCACUGGUAGAAUUCACU	7083.92021	7083.8958	0.024	3.45
Average mass accuracy				2.11 ppm

The sequencing ladder was obtained from high resolution MALDI SpiralTOF measurements of miRNA hydrolyzed at 65 °C using 3-HPA saturated in 10% acetonitrile.

Table 3.6. Mass accuracy of 3’ ladder fragments from hydrolyzed miR-183

3’ Ladder	Theoretical Mass (Da)	Observed Mass (m/z)	Mass Difference (Da)	Mass Accuracy (ppm)
UCACU	1490.23721	1490.2265	0.011	7.19
UUCACU	1796.26251	1796.2595	0.003	1.68
AUUCACU	2125.31503	2125.3164	-0.001	-0.64
AAUUCACU	2454.36755	2454.3708	-0.003	-1.32
GAAUUCACU	2800.41769	2800.4129	0.005	1.71
AGAAUUCACU	3129.47019	3129.4777	-0.008	-2.40
UAGAAUUCACU	3435.49552	3435.5030	-0.007	-2.18
GUAGAAUUCACU	3780.54293	3780.5527	-0.010	-2.58
GGUAGAAUUCACU	4125.59035	4125.5964	-0.006	-1.47
UGGUAGAAUUCACU	4431.61567	4431.6252	-0.010	-2.15
CUGGUAGAAUUCACU	4737.65953	4737.6605	-0.001	-0.20
ACUGGUAGAAUUCACU	5066.71203	5066.7152	-0.003	-0.63
CACUGGUAGAAUUCACU	5371.75334	5371.7521	0.001	0.23
GCACUGGUAGAAUUCACU	5716.80076	5716.7911	0.010	1.69
GGCACUGGUAGAAUUCACU	6061.84818	6061.8345	0.014	2.26
UGGCACUGGUAGAAUUCACU	6367.87351	6367.8602	0.013	2.09
AUGGCACUGGUAGAAUUCACU	6696.92076	6696.9123	0.008	1.26
pUAUGGCACUGGUAGAAUUCACU	7083.92021	7083.8958	0.024	3.45
Average mass accuracy				1.95 ppm

The sequencing ladder was obtained from high resolution MALDI SpiralTOF measurements of miRNA hydrolyzed at 65 °C using 3-HPA saturated in 10% acetonitrile.

Figure 3.13 shows the resulting hydrolyzed fragments for miR-124. The 5’ and 3’ ladders are shown in Figure 3.14 (a) and (b) respectively. The bolded region in the sequence above Figure 3.14 represents a 63.6% sequence overlap between the 5’ and 3’ ladders of miR-124, further reinforcing the sequence determined. Isotopic resolution of the hydrolyzed fragments was achieved and the results of the isotopic distribution of

selected peaks are shown in Figure 3.15. The theoretical masses, observed mass, difference between theoretical and observed masses as well as the mass accuracy for both 5' and 3' ladders for miR-124 are tabulated in Tables 3-7 and 3-8 respectively. The 5' ladder was detected with an avreage mass accuracy of 4.32 ppm while for the 3' ladder, an avreage mass accuracy of 4.29 was attained.

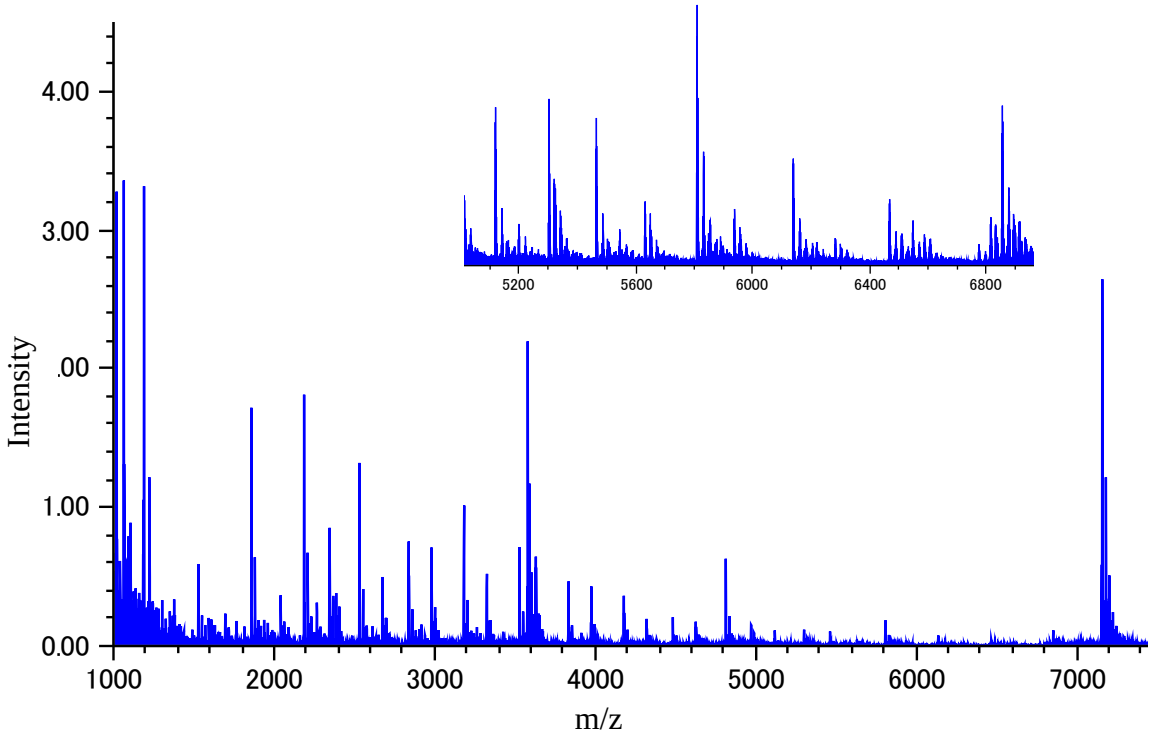


Figure 3.13. Spiral TOF MALDI MS spectrum of 5' and 3' fragments obtained from the hydrolysis of miR-124a using 3-HPA saturated in 10% acetonitrile in postive ion mode.

miR-124 sequence pUUAA**GGCACGCGGUGAAU**GCCA

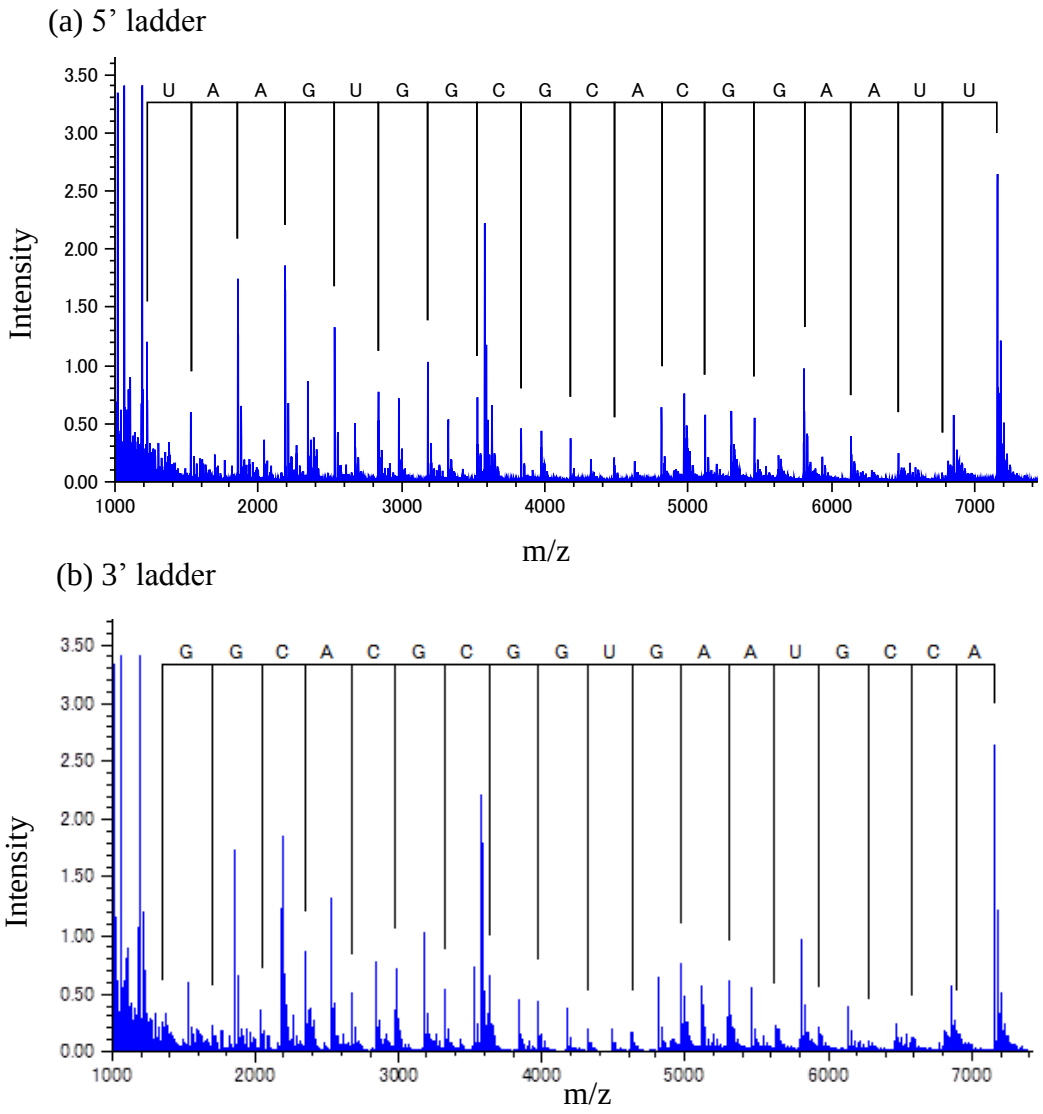


Figure 3.14. Spiral TOF MALDI MS spectrum of miR-124a fragments obtained from the hydrolysis of using 3-HPA saturated in 10% acetonitrile in positive ion mode.

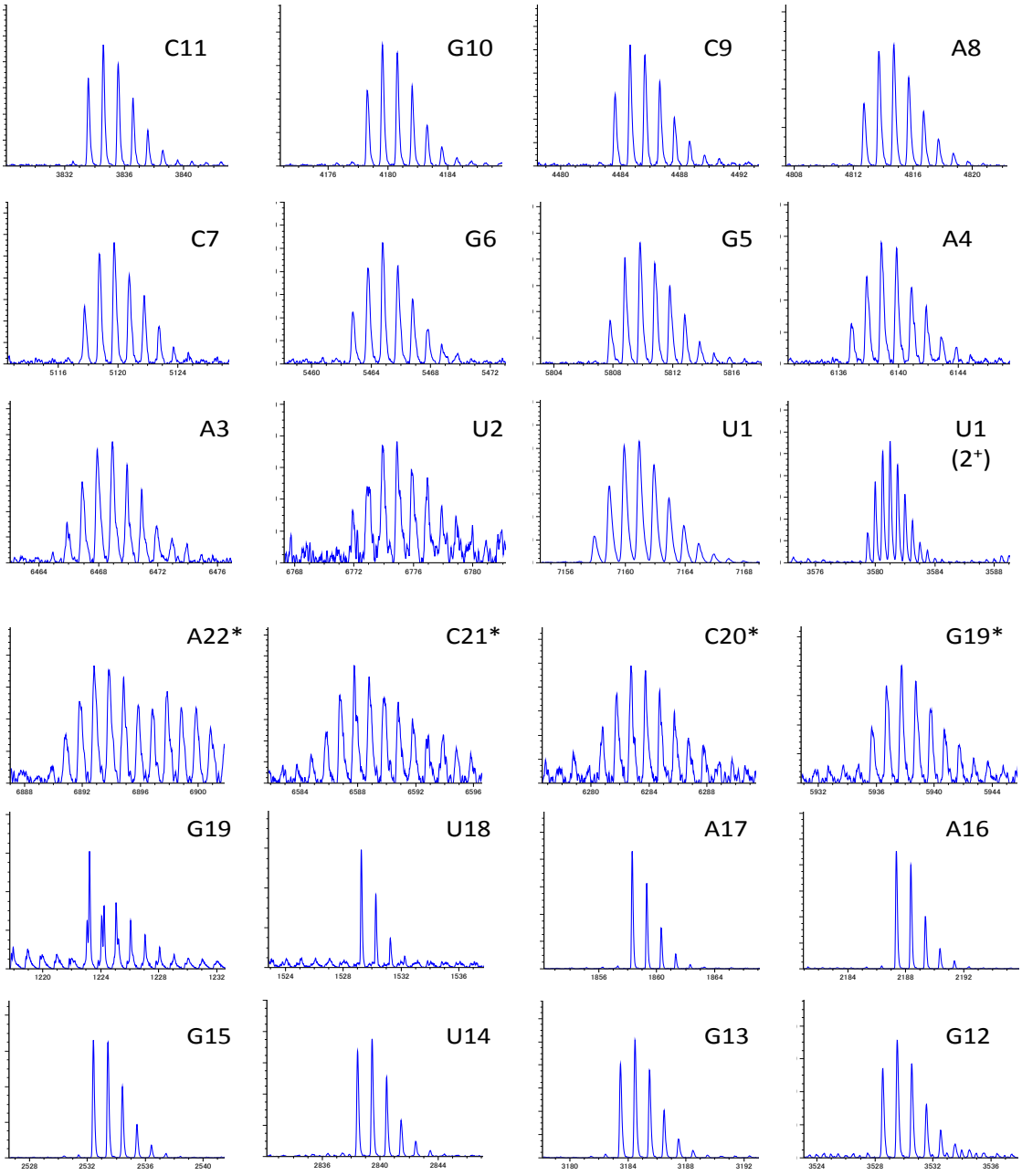


Figure 3.15. Spiral TOF MALDI MS spectrum showing the fragments obtained from the hydrolysis of miR-124.

Table 3.7. Mass accuracy of 3’ ladder fragments from hydrolyzed miR-124.

3’ Ladder	Theoretical *m/z*	Observed *m/z*	Mass difference	Mass accuracy (ppm)
GCCA	1223.23404	1223.2338	0.000	0.20
UGCCA	1529.25934	1529.2593	0.000	0.03
AUGCCA	1858.31186	1858.3117	0.000	0.09
AAUGCCA	2187.36438	2187.3645	0.000	-0.05
GAAUGCCA	2532.41181	2532.4133	-0.001	-0.59
UGAAUGCCA	2839.43977	2839.4394	0.000	0.13
GUGAAUGCCA	3184.48455	3184.4883	-0.004	-1.18
GGUGAAUGCCA	3529.53461	3529.5303	0.004	1.22
CGGUGAAUGCCA	3834.57591	3834.5707	0.005	1.36
GCGGUGAAUGCCA	4179.62334	4179.6175	0.006	1.40
CGCGGUGAAUGCCA	4484.66463	4484.6522	0.012	2.77
ACGCGGUGAAUGCCA	4814.71966	4814.6975	0.022	4.60
CACGCGGUGAAUGCCA	5119.76097	5119.7339	0.027	5.29
GCACGCGGUGAAUGCCA	5464.80839	5464.7706	0.038	6.92
GGCACGCGGUGAAUGCCA	5809.85582	5809.8023	0.054	9.21
AGGCACGCGGUGAAUGCCA	6138.90833	6138.8390	0.069	11.29
AAGGCACGCGGUGAAUGCCA	6467.96085	6467.8988	0.062	9.59
UAAGGCACGCGGUGAAUGCCA	6773.98618	6773.8905	0.096	14.12
pUUAAGGCACGCGGUGAAUGCCA	7160.98035	7160.8939	0.086	12.07
Average mass accuracy				4.32 ppm

The sequencing ladder was obtained from high resolution MALDI SpiralTOF measurements of miRNA hydrolyzed at 65 °C using 3-HPA saturated in 10% acetonitrile.

Table 3.8. Mass accuracy of 5’ ladder fragments from hydrolyzed miR-124.

5’ Ladder	Theoretical *m/z*	Observed *m/z*	Mass difference	Mass accuracy (ppm)
pUUAA	1351.12925	1351.1265	0.003	2.04
pUUAAG	1696.17668	1696.1762	0.000	0.28
pUUAAGG	2041.22412	2041.2213	0.003	1.38
pUUAAGGC	2346.26540	2346.2657	0.000	-0.13
pUUAAGGCA	2675.31792	2675.3159	0.002	0.76
pUUAAGGCAC	2981.36187	2981.3618	0.000	0.02
pUUAAGGCACG	3326.40929	3326.4096	0.000	-0.09
pUUAAGGCACGC	3631.45059	3631.4477	0.003	0.80
pUUAAGGCACGCG	3976.49801	3976.4925	0.006	1.39
pUUAAGGCACGCGG	4321.54543	4321.5400	0.005	1.26
pUUAAGGCACGCGGU	4627.57076	4627.5533	0.017	3.77
pUUAAGGCACGCGGUG	4973.62070	4973.6061	0.015	2.94
pUUAAGGCACGCGGUGA	5302.67321	5302.6512	0.022	4.15
pUUAAGGCACGCGGUGAA	5631.72573	5631.6900	0.036	6.34
pUUAAGGCACGCGGUGAAU	5937.75106	5937.6938	0.057	9.64
pUUAAGGCACGCGGUGAAUG	6282.79849	6282.7237	0.075	11.90
pUUAAGGCACGCGGUGAAUGC	6587.83979	6587.7763	0.063	9.64
pUUAAGGCACGCGGUGAAUGCC	6892.88109	6892.7921	0.089	12.91
pUUAAGGCACGCGGUGAAUGCCA	7160.98035	7160.8939	0.086	12.07
Average mass accuracy				4.29 ppm

The sequencing ladder was obtained from high resolution MALDI SpiralTOF measurements of miRNA hydrolyzed at 65 °C using 3-HPA saturated in 10% acetonitrile.

Using the same protocol of hydrolysis with saturated 3-HPA, analysis of partially hydrolyzed miR-153 was performed and the results are shown in Figure 3.16. The peaks in the mass spectrum that correspond to fragments that end with U or C were identified in

Figure 3.17, enlarged, labeled and presented in Figure 3.18. The rest of the isotopic distribution is shown in Figure 3.19.

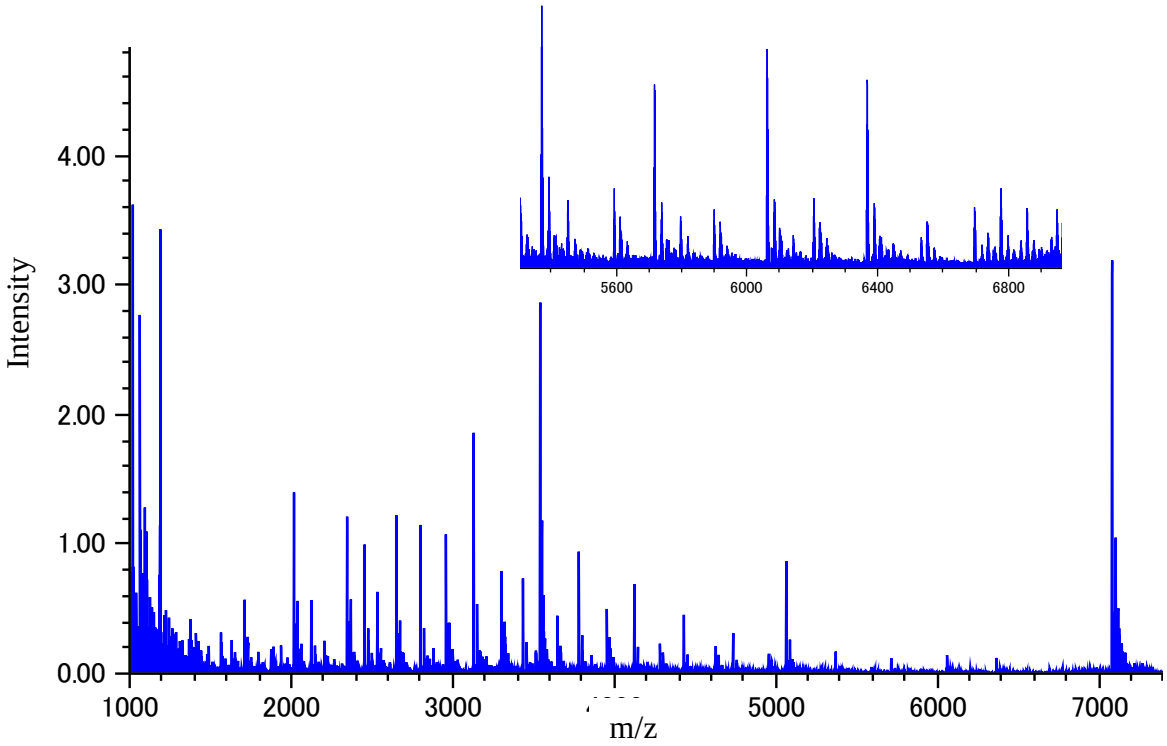


Figure 3.16. Spiral TOF MALDI MS spectrum of miR-153 fragments obtained from the hydrolysis of using 3-HPA saturated in 10% acetonitrile in positive ion mode.

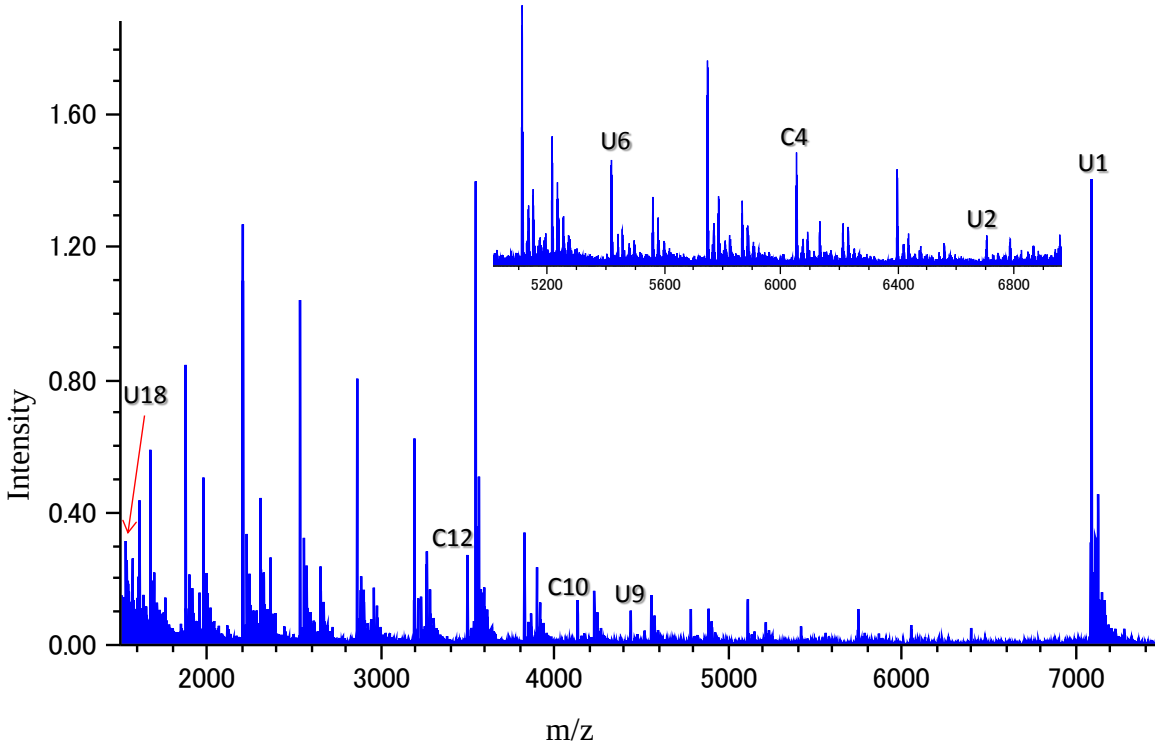


Figure 3.17. Identification of fragments terminated by U or C from a MALDI *SpiralTOF* mass spectrum obtained from the hydrolysis of miR-153.

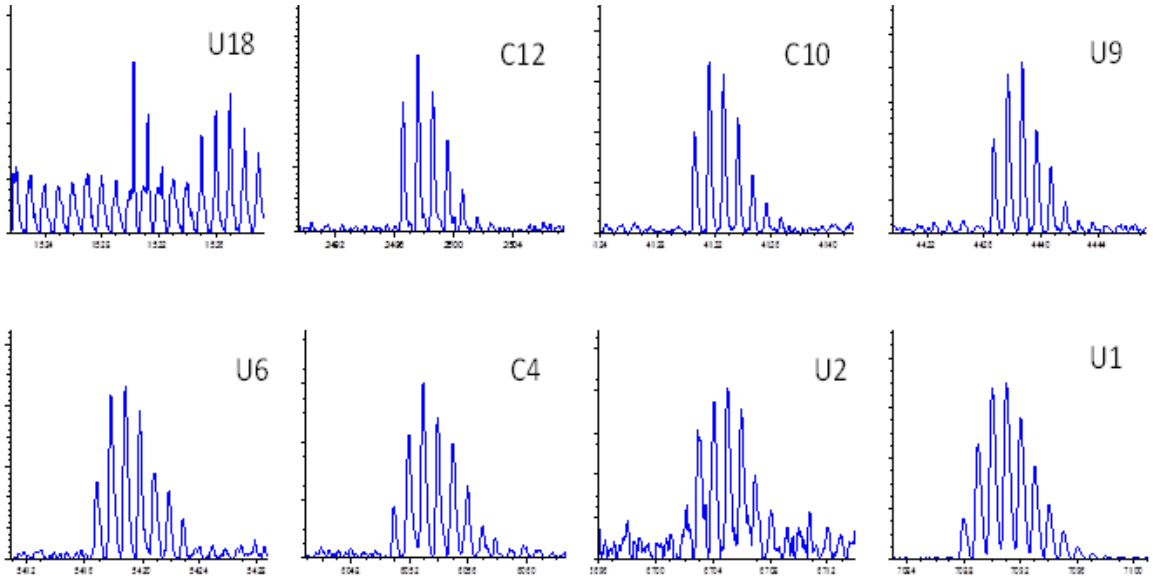


Figure 3.18. Subspectra of isotopic distribution from 3’ ladder of hydrolyzed miR-153 with terminal C or U.

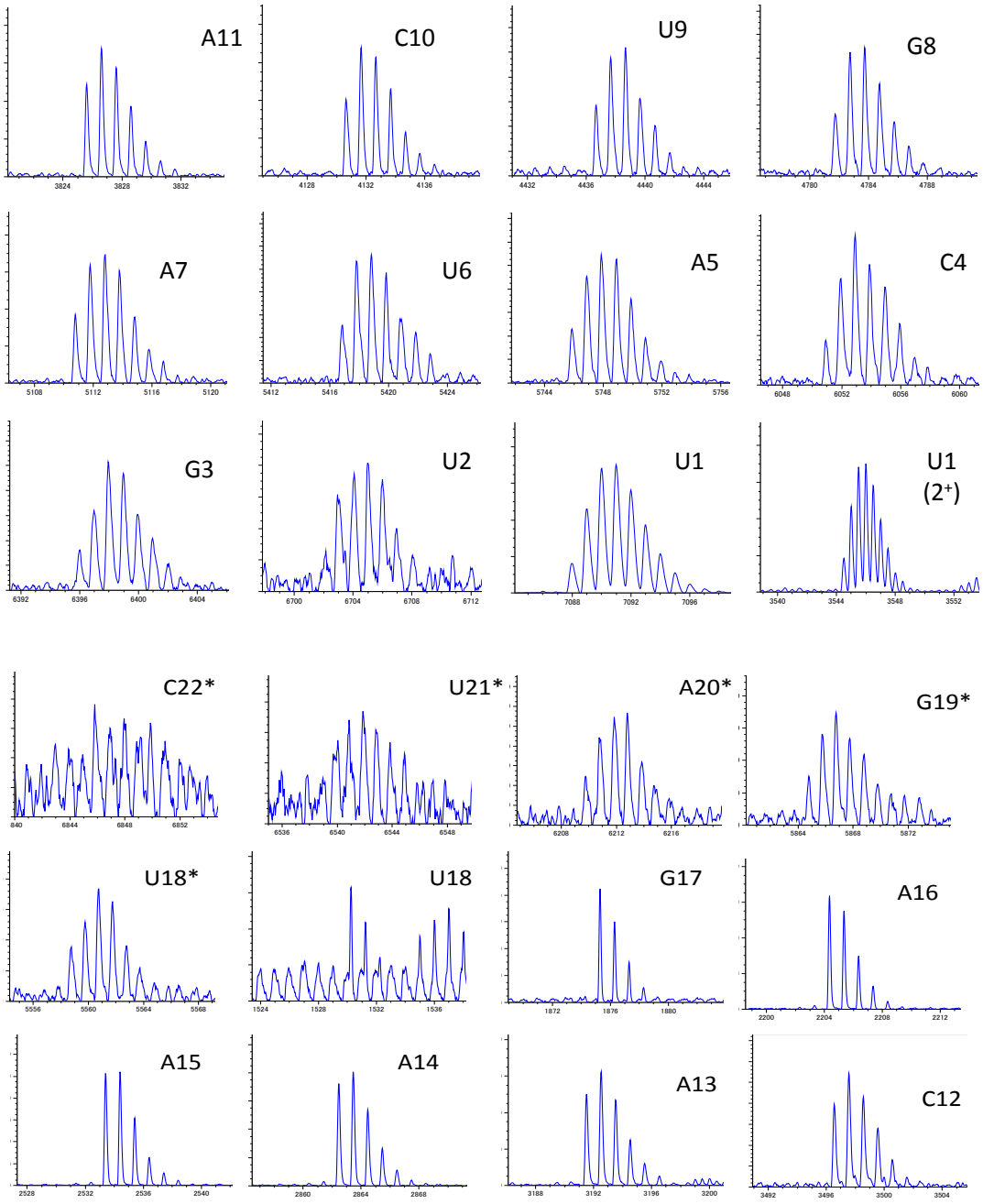


Figure 3.19. (a) and (b) subspectra of isotopic distribution from 3’ ladder of hydrolyzed miR-153

Table 3.9. Mass accuracy of 3’ ladder fragments from hydrolyzed miR-153.

3’ Ladder	Theoretical m/z	Observed m/z	Mass difference	Mass accuracy (ppm)
UGAUC	1530.2434	1530.2483	0.005	3.23
GUGAUC	1875.2908	1875.3005	0.01	5.18
AGUGAUC	2204.3433	2204.3545	0.011	5.08
AAGUGAUC	2533.3958	2533.4084	0.013	4.96
AAAGUGAUC	2863.451	2863.4623	0.011	3.95
AAAAGUGAUC	3192.5035	3192.5215	0.018	5.64
CAAAAGUGAUC	3497.5448	3497.567	0.022	6.35
ACAAAAGUGAUC	3826.5973	3826.6238	0.027	6.93
CACAAAAGUGAUC	4131.6386	4131.6672	0.029	6.92
UCACAAAAGUGAUC	4437.6639	4437.6912	0.027	6.15
GUCACAAAAGUGAUC	4783.7139	4783.7447	0.031	6.44
AGUCACAAAAGUGAUC	5112.7664	5112.8243	0.058	11.33
UAGUCACAAAAGUGAUC	5418.7917	5418.8373	0.046	8.41
AUAGUCACAAAAGUGAUC	5747.8442	5747.9013	0.057	9.93
CAUAGUCACAAAAGUGAUC	6052.8855	6052.9208	0.035	5.83
GCAUAGUCACAAAAGUGAUC	6397.933	6397.992	0.059	9.23
UGCAUAGUCACAAAAGUGAUC	6703.9583	6704.0421	0.084	12.5
pUUGCAUAGUCACAAAAGUGAUC	7090.9525	7091.0328	0.08	11.33
Average mass accuracy				7.19 ppm

The sequencing ladder was obtained from high resolution MALDI SpiralTOF measurements of miRNA hydrolyzed at 65 °C using 3-HPA saturated in 10% acetonitrile.

When using the low resolution AB 4700 TOF/TOF mass spectrometer, it was noted that the percentage of miRNA sequence identified was closely related to the trend in pKa of the matrices. 3-HPA (pKa = 1.14) produced the highest sequence coverage followed by 2,5 DHB (pKa = 3.01) then SA (pKa = 4.53), however, 100% sequence coverage was achieved when using high resolution JMS-S3000 SpiralTOF mass

spectrometer. This was demonstrated by using the same hydrolysis and MALDI sample preparation methods for the hydrolysis of miR-153 followed by measurement of the resultant fragments using both instruments. The sequence identification obtained from a single spectra using the AB MALDI-TOF was 72.7% while 100% sequence identification was obtained when using the JMS-S3000 SpiralTOF. The difference in sequence coverage observed between the AB 4700 TOF/TOF and JMS-S3000 SpiralTOF mass spectrometers may be attributed to the sensitivity difference between the two i.e. the low signal to noise ratio of the JMS-S3000 SpiralTOF enabled the detection of more fragments compared to the AB 4700 TOF/TOF instrument.

Theoretically, the 5' and 3' fragments are produced in equal amounts; however, the spectra from all three miRNAs analyzed consistently showed higher intensity of the 3' fragments that bear hydroxyl groups on both 5' and 3' termini. This phenomenon was also observed by Hahner et al. It has been hypothesized that there is probably a preferential inclusion of fragments in the MALDI matrix based on their termini or more likely, that there is preferential ionization of fragments during the MALDI process. There is however no structural differences in the 5' and 3' fragments that can support the explanation of the preference (Hahner et al. 1997a).

3.40 Alternative Approach to Interpret JMS-S3000 SpiralTOF Data

In the previous section, JEOL *MS Tornado* software was used to process the *Spiral TOF* mass spectra. Ladder sequences were determined from the mass differences between successive fragments by using native functions in the *MS Tornado* software and

by using *New MS Tools* software (RBC Software) to examine mass spectra exported in text format. Although the assay presented here is designed and optimized for miRNA whose average length is 22 nucleotides, it should also be noted that the assay could be applicable to longer RNA samples. It is however expected that due to instrument limitations on resolving power, isotopic distribution may not be achieved in all circumstances. Also, for analytes that are present in low concentrations, the monoisotopic peaks necessary for calculation of mass differences may also not be observed. Under these conditions, the use of mass differences between monoisotopic peaks will not be feasible. To tackle this challenge, an alternative approach to MS data interpretation for sequence elucidation was proposed. The proposed approach is based upon the average isotopic composition of a hypothetical average nucleotide "averageotide", analogous to the "averagine" method described by Senko, Beu and McLafferty for interpreting mass spectra of proteins and peptides (Senko et al. 1995). Nucleotide compositions were calculated by *New MS Tools* software. A Microsoft Visual Basic program was written to calculate the average nucleotide distribution for the human miRNA sequences in the miRBase database. Averageotide calculations were carried out in a Microsoft Excel spreadsheet. The approach uses exact masses to check and confirm these sequence assignments. The calculation was carried out in the same manner that elemental compositions are typically determined from high-resolution, exact-mass measurements. "Pseudoelements" (labeled C,U,A,G) were defined having the isotopic composition (exact masses and relative abundances) of each nucleotide as shown in Tables 3-10 a, b, c, d, e, and f. One pseudoelement (labeled c, u, a, g) corresponding to the terminal

nucleoside is also defined. It does not matter which nucleoside is designated as the terminal group because we are only looking at the total base counts for successive fragments. The nucleotide compositions for each fragment were determined by using the monoisotopic mass to calculate the possible nucleotide compositions. Because the elemental composition constraints are limited to the four RNA nucleotides and the endgroup elements, only a small number of nucleotide compositions are calculated for even relatively large fragments within the error tolerance. Without this constraint, it would be nearly impossible to determine a unique elemental composition (e.g. $C_vH_wN_xO_yP_z$) for the larger fragments based on exact mass measurements alone.

Candidate compositions were then ranked by matching the measured isotopic *m/z* values and relative abundances to the calculated values for each composition. Candidate compositions for successive fragments that differed by more than a single nucleotide were rejected. The resulting nucleotide compositions were cross-checked against the expected compositions determined from the mass differences between the monoisotopic peaks for successive fragments. An example is shown in Table 3.11 for the 5' ladder of miR-183. Both "blind" samples (miR-183 and miR-124) were completely sequenced by this method with 100% accuracy.

Table 3.10. Isotopic composition of "pseudoelements" C,U,A,G and protonated U showing exact masses and relative abundances

C ($C_9H_{12}N_3O_7P$)	
m/z	Rel. abundance
305.04129	100
306.04408	11.247
307.0459	2.016
308.04843	0.179
309.05042	0.017
310.0528	0.001

U ($C_9H_{11}N_2O_8P$)	
m/z	Rel. abundance
306.02531	100
307.02829	10.904
308.02997	2.184
309.03266	0.195
310.03455	0.021
311.03705	0.002

A ($C_{10}H_{12}N_5O_6P$)	
m/z	Rel. abundance
329.05252	100
330.05503	13.029
331.05703	2.02
332.05935	0.189
333.06145	0.017
334.06367	0.001

G ($C_{10}H_{12}N_5O_7P$)	
m/z	Rel. abundance
345.04744	100
346.04995	13.067
347.05193	2.23
348.05426	0.217
349.05634	0.021
350.05857	0.002

c (protonated cytidine $C_9H_{14}N_3O_5$)	
m/z	Rel. abundance
244.09335	100
245.09613	11.194
246.09804	1.599
247.10052	0.132
248.10262	0.01
249.10494	0.001

Table 3.11. Nucleotide compositions calculated from miR-183 3’ fragments.

Calc. m/z	Abund %	mmu	ppm	DBE	Composition	**Difference**
1490.237780	6.042	-11.34	-7.61	4.0	c1C1U2A1	
1796.263090	4.127	-3.57	-1.99	5.0	c1C1U3A1	**(U)**
2125.315610	11.945	0.80	0.37	5.0	c1C1U3A2	**(A)**
2454.368130	20.823	2.72	1.11	5.0	c1C1U3A3	**(A)**
2799.415570	22.450	2.89	1.03	6.0	c1C1U3A3G1	**(G)**
3128.468090	32.772	5.54	1.77	6.0	c1C1U3A4G1	**(A)**
3434.493400	12.240	2.45	0.71	7.0	c1C1U4A4G1	**(U)**
3779.540840	14.741	8.48	2.24	8.0	c1C1U4A4G2	**(G)**
4124.588280	9.731	9.38	2.27	9.0	c1C1U4A4G3	**(G)**
4430.613590	6.114	-3.24	-0.73	10.0	c1C1U5A4G3	**(U)**
4735.654880	4.540	1.37	0.29	11.0	c1C2U5A4G3	**(C)**
5064.707400	10.374	-4.27	-0.84	11.0	c1C2U5A5G3	**(A)**
5369.748690	1.986	-4.55	-0.85	12.0	c1C3U5A5G3	**(C)**
5714.796130	1.452	-19.27	-3.37	13.0	c1C3U5A5G4	**(G)**
5714.758430	1.452	18.43	3.22	18.0	c1C5U5G7	*****
6059.843570	1.338	-20.33	-3.35	14.0	c1C3U5A5G5	**(G)**
6059.805870	1.338	17.37	2.87	19.0	c1C5U5G8	*****
6365.868880	1.154	-2.67	-0.42	15.0	c1C3U6A5G5	**(U)**

* Not a valid composition, out of sequence with the other compositions.
DBE - Double bond equivalent, mmu – milli mass unit/ millidalton.

Table 3.11 shows nucleotide compositions calculated from exact masses measured for monoisotopic peaks of fragment ions having relative abundances >3% in the Spiral TOF mass spectrum of miR-183 (pUAUGGCACUGGUAGAAUUCACU). The labels in parentheses show the nucleotide difference from the preceding composition. These differences reveal the partial sequence shown in bold.

To ensure the correct assignment of the nucleotides C and U, a third approach was proposed for determining the monoisotopic masses for fragments where the monoisotopic peak is weak, absent, or unresolved was also developed. This approach is analogous to the method proposed by Senko, Beu and McLafferty for proteins and peptides in which a hypothetical amino acid "averagine" was defined. An examination of all of the sequences for human RNA in the miRNA database (http://www.mirbase.org/cgi-bin/mirna_summary.pl?org=hsa) showed a nearly equal nucleotide distribution corresponding to C (23.20%), U (26.78%), A (23.52%), and G (26.50%). Based on these values, we define an average nucleotide "averageotide" having the composition $C_{9.5001}H_{11.7321}N_{3.7325}O_{7.0325}P$ and an average molecular weight of 321.6983 Da. Using these values to determine the monoisotopic mass for a 24 kDa miRNA precursor (hsa-mir-328 MI0000804, average mass 24225.4 Da), we obtain an estimated value of 24214.3 Da, which differs from the calculated value of 24214.2 by only 0.1 Da, or 4 ppm.

This approach was not used for the de novo sequencing of the unknown samples reported herein because the monoisotopic peaks for each fragment were observed. However, the "averageotide" concept is expected to be useful for larger fragments or spectra with insufficient resolving power or signal-to-noise ratio. Although we calculated an average nucleotide based upon human miRNA sequences, it may be noted that alternative values could be calculated for other RNA populations or for DNA. All spectra shown, both from the AB MALDI-TOF and the *Spiral TOF* were obtained in the positive ion mode.

3.50 Conclusions

Using hydrolysis and high-resolution MALDI, complete sequence coverage with 100% accuracy for de novo sequencing of miRNA samples was achieved. The method described herein can be automated and does not require any additional reagents other than the conventional MALDI-MS reagents for RNA measurement. The developed assay opens a way for high-throughput and low-cost direct sequencing of miRNA with the potential to detect modifications in mature miRNA such as nucleotide addition or deletion, as well as enzymatic modification of miRNA nucleotides commonly referred to as RNA editing.

CHAPTER IV

MicroRNA MultiTool; A SOFTWARE FOR THE IDENTIFICATION OF MODIFIED AND UNMODIFIED microRNA FROM MASS SPECTROMETRIC DATA

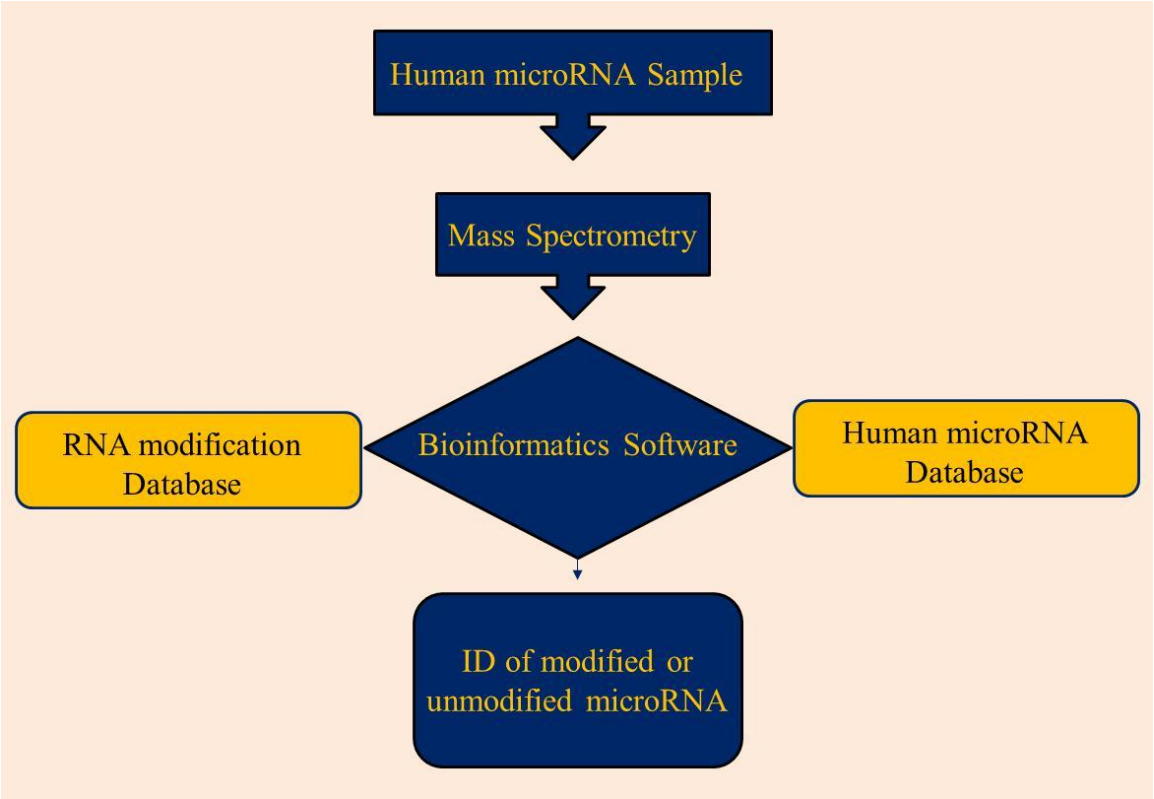


Figure 4.1. Schematic abstract showing the use of bioinformatics in microRNA identification from mass spectrometric data.

4.10 Introduction

MicroRNA (miRNA) are a class of short endogenous non-coding RNA, ~19-25 nucleotides in length that play a crucial role in post-transcriptional gene regulation by either triggering degradation or preventing translation of the target messenger RNAs. With the recent discovery that miRNA can control over 60% of human gene expression, it is not surprising that the field of miRNA biology has quickly gained considerable interest within the last decade (Griffiths-Jones et al. 2006). MiRNA can regulate a variety of cellular processes (Lee and Dutta 2009), (Lu et al. 2008) and have been implicated in the initiation and progression of 160+ human diseases (Jiang et al. 2009). For this reason, miRNAs have been intensely studied for their potential use as disease biomarkers or therapeutic targets (Melo and Esteller 2011).

Numerous methods have been developed for the identification and characterization of miRNAs. These methods include polymerase chain reaction, northern blots, microarrays and mass spectrometry (MS) among others. Recently, MS emerged as an important tool for the analysis of nucleic acids due to its accuracy, sensitivity, efficiency and ability to perform high throughput measurements, furthermore, the ability of MS to measure the intrinsic property of miRNA i.e. mass, gives it an added advantage in miRNA measurements over the alternative molecular biology techniques that rely on measurement of secondary signals. The availability of high throughput assays for analysis of miRNA have made MS a more appealing technique for miRNA measurements (Wambua et al. 2012). We recently demonstrated that MS can be used for the accurate detection and *de novo* sequencing of miRNA (Wambua et al. unpublished data presented

in chapter III). Matrix assisted laser desorption ionization (MALDI) results in the generation of spectra which are dominated by singly charged ions that are relatively easy to interpret is therefore the preferred choice for oligonucleotide analysis (Bahr et al. 2009).

Currently, there are more than 2,100 human miRNA reported in the literature (Griffiths-Jones et al. 2008). Several miRNAs can coexist within the same sample and therefore, parallel analysis of miRNA is unavoidable. Except in cases where only very few miRNAs are analyzed, or only a few mass spectra are involved, the mass spectrometric data generated from miRNA measurements can be complex and manual data processing is tedious and challenging (Oberacher and Pitterl 2009). Recent studies have also revealed that just like other forms of RNA, miRNA can undergo nucleotide modification which plays a key role in miRNA maturation and function (Iida et al. 2009). MiRNA nucleotide modifications are important since modified miRNA are known to target transcripts that are different than those targeted by non-modified miRNA (Burroughs et al. 2010). According to MODOMICS; a comprehensive database resource for RNA modifications, there are more than 100 distinct nucleotide modifications (Czerwoniec et al. 2009). Whereas current molecular biology techniques are unable to identify miRNA modifications, MS has been used for the characterization of different RNA post-transcriptional modifications including identification of type and exact position of the RNA modification (Nakayama et al. 2011), (Mengel-Jørgensen and Kirpekar 2002). The phenomenon of miRNA modification adds a layer of complexity to the miRNA data interpretation, further underlining the importance of developing methods

and appropriate software capable of identifying both modified and unmodified miRNA (Mizuguchi et al. 2011).

With these challenges in mind, the spectral interpretation of mass spectrometric data can quickly turn out to be the bottleneck in miRNA analysis (Griffiths-Jones et al. 2008). The success of MS as a tool for the analysis of miRNA will therefore strongly depend on the development of relevant computational software with the ability to properly interpret and analyze the large-scale, high-dimensional data that is generated from MS experiments (Chandramouli and Qian 2009).

To address the challenges associated with MS data analysis of both post-transcriptionally modified and unmodified miRNA, we present here the development of a user-interactive software, termed as MicroRNA MultiTool. MicroRNA MultiTool has a graphical user interface that allows the user to directly utilize mass spectral data in order to identify miRNA by sequence, name, mass, accession number and RNA modification if any. Since the mass of miRNA can be directly used for its identification, MicroRNA MultiTool eliminates the need of performing tandem mass spectrometry on a miRNA for the sole purpose of its identification.

4.20 MicroRNA MultiTool Functionalities

The human miRNA sequences were obtained from miRBase; a database which serves as a primary repository for published miRNAs (http://www.mirbase.org/) (Griffiths-Jones et al. 2006) while RNA modifications were from MODOMICS; a comprehensive database for RNA modification (http://modomics.genesilico.pl/)

(Czerwoniec et al. 2009). As shown in Figure 4.2, the MicroRNA MultiTool graphical user interface is divided primarily into three features as discussed in sections 4.21) miRNA search and mass calculator, 4.22) modified miRNA mass calculator and 4.23) miRNA fragment search. Help buttons are strategically placed with appropriate information to guide the user on the functions offered by the program.

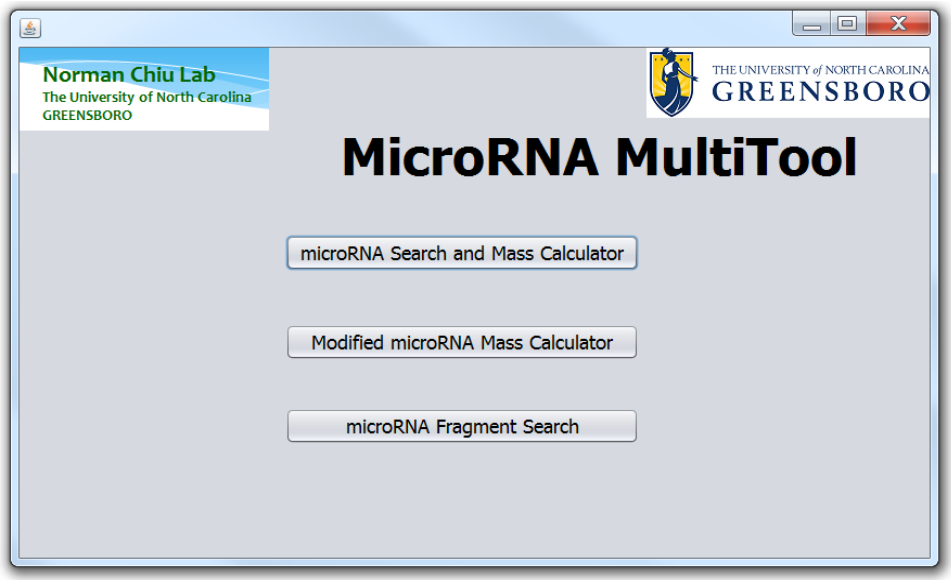


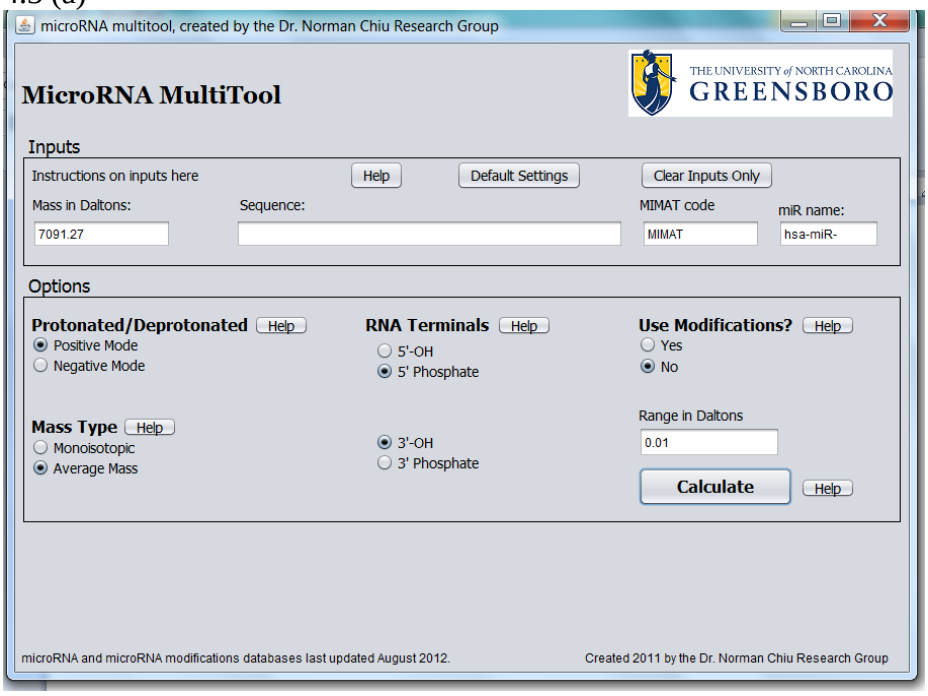
Figure 4.2. MicroRNA MultiTool graphical user interface.

4.21 MiRNA search and mass calculator

The graphical user interface for miRNA search and mass calculator function and its output are shown in Figure 4.3 (a) and (b) respectively. This purpose of this function is to match the data obtained from a mass spectrometric measurement to possible masses of all known human miRNA. This feature is especially important when miRNA are to be identified from mass spectral data with no other information available. In the interface

shown in Figure 4.3 (a) the “Mass in Daltons” field accepts numeric values and is provided for searching modified or unmodified miRNA. To search for the identity of a miRNA, a specific mass from MS data is entered into the Mass in Daltons field, in the “Options” panel the ionization mode used to acquire the MS data i.e. “Protonated” for positive ion mode or “Deprotonated” for negative ion mode is selected, mono isotopic or average mass and the type of 3’ and 5’ terminal groups on the miRNA can also be determined. On the “Options” panel, the option of “Use Modification” is provided to the user if they are searching for modified miRNA, with the default set at “No” under “Use Modifications” for searching unmodified miRNA. Clicking on Calculate returns the result of the analysis. The result of the search is a list of all unmodified miRNAs matching the entered mass together with their names, accession numbers, sequence and molecular masses as shown in Figure 4.3(b).

4.3 (a)



4.3 (b)

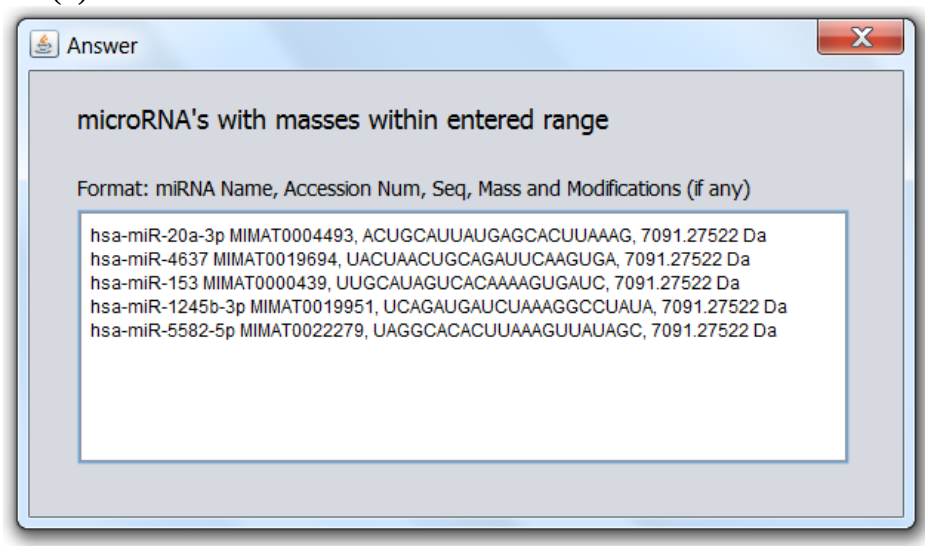


Figure 4.3. (a) Graphical user interface for miRNA search and mass calculator function and (b) its results output.

An especially unique function of the miRNA search and mass calculator feature compared to other RNA mass calculators is the ability to identify miRNA that may carry a modification on one of the nucleotides as well as the possible modification. This is done by first entering the mass as obtained from the mass spectrometric measurement in the "Mass in Daltons field" shown in Figure 4.3 (a), selecting desired options in the Options panel as explained in the previous section, followed by selecting "Yes" under "Use Modifications", then "Calculate". The software algorithm then scans each miRNA in the database and evaluates the sum of the miRNA mass plus the mass of any one of the known miRNA modification. This is done for all the known human miRNA and all RNA modifications. The output consists of the modified miRNA mass, name, accession number and modification as shown in Figure 4.4.

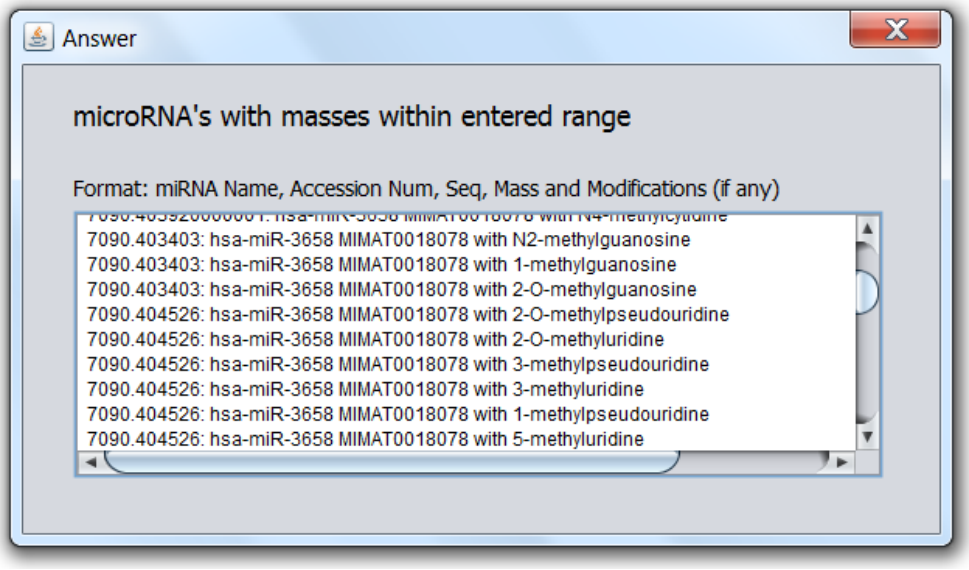
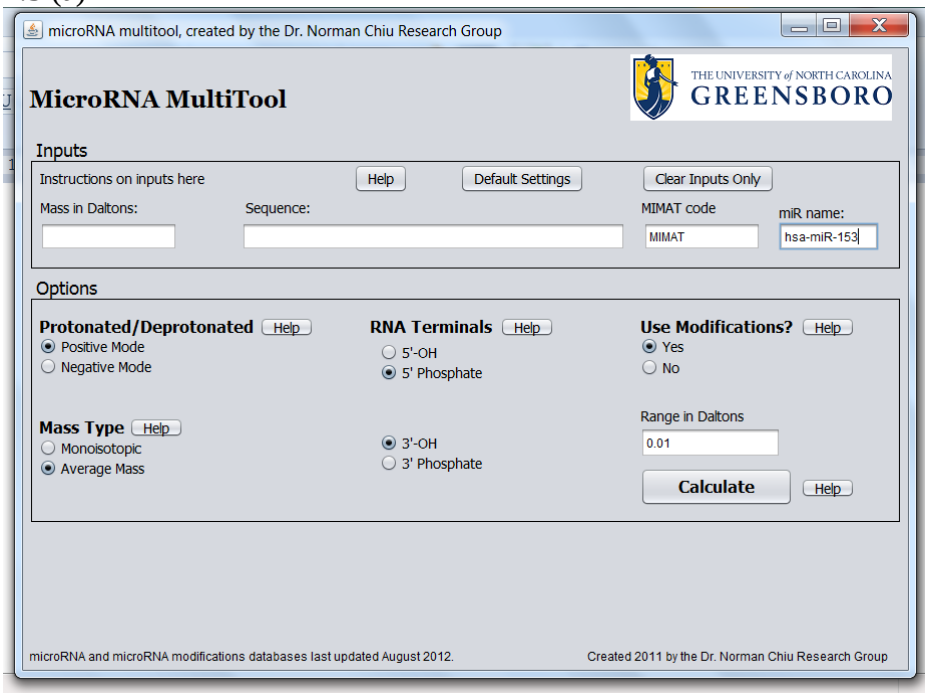


Figure 4.4. Modified miRNA search results output.

Depending on the type of mass spectrometer used, the mass accuracy can vary greatly. In order to account for different accuracy levels of various mass spectrometers, the user is offered the option to define a +/- range in Daltons (Da) around the experimental mass which the program algorithm can use to search for matching miRNA masses. The defaults in the options panel are set at positive ionization mode, average mass, 5' phosphate, 3' hydroxyl, no modification and +/- 0.01 Da range. After all options have been made, the user clicks the "Calculate" button to return the results.

The "Sequence", "MIMAT" and "miR Name" fields of the miRNA search and mass calculator feature accepts RNA sequence (A,G,C,U), accession number and miRNA name entries respectively. If the entry matches any known miRNA, the program returns all parameters associated with the miRNA including its accession number, name, sequence and mass. The choices on the option panel can be selected as previously explained. Spaces entered between inputs are automatically ignored. An example with hsa-miR-153 entered in the name field is shown in Figure 4.5 (a) while the output is shown in Figure 4.5 (b).

4.5 (a)



4.5 (b)

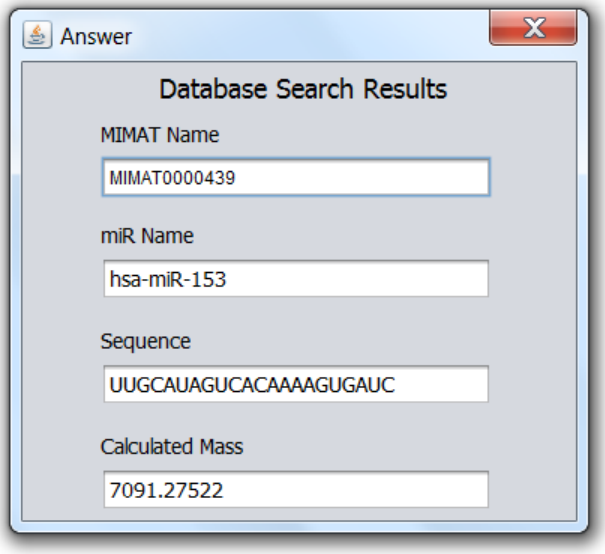


Figure 4.5. (a) miRNA search and mass calculator user interface showing the various fields for data entry and (b) the output panel generated using miRNA name.

4.22 Modified miRNA mass calculator

To provide pre-experimental miRNA sample analysis, MicroRNA MultiTool offers a novel feature whose purpose is to simulate expected mass data of modified miRNA. In Figure 4.6 (a) the data entry field provided in the Modified miRNA mass calculator interface accepts input in the form of miRNA name e.g. hsa-miR-153. A drop down menu with RNA modifications is provided from which the user can select a single modification. The program calculates the total mass of the miRNA and the modification and returns an output as shown in Figure 4.6 (b) consisting of the mass of the unmodified miRNA selected, mass of modified miRNA and the modification chosen.

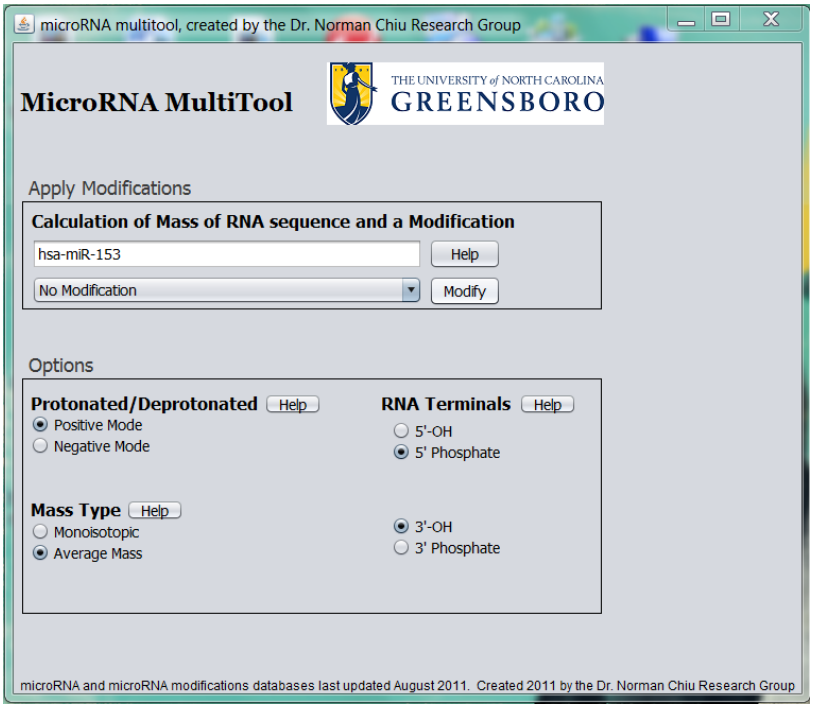


Figure 4.6. (a) MicroRNA MultiTool modified miRNA mass calculator user interface.

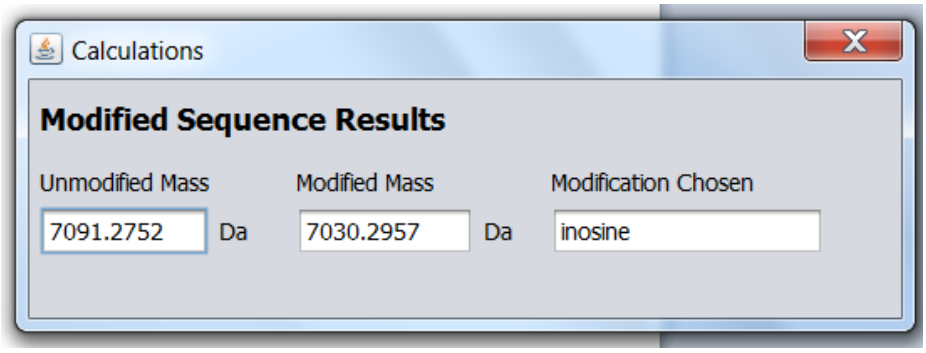
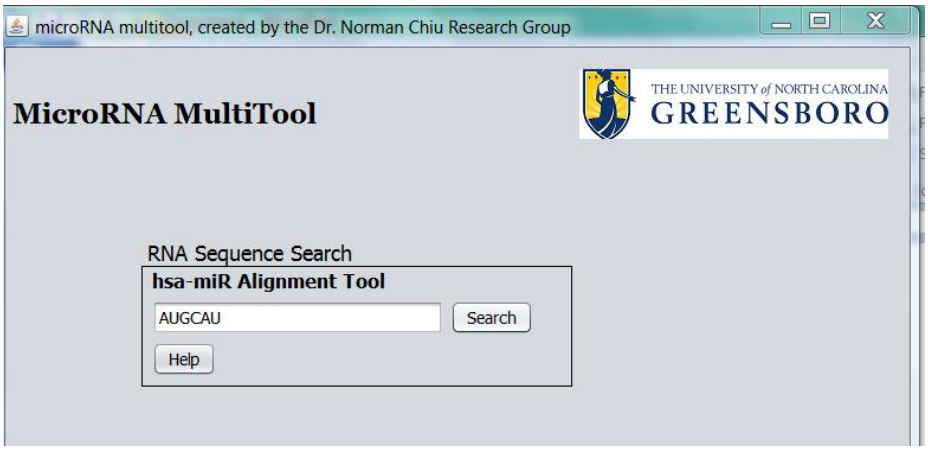


Figure 4.6. (b) MicroRNA MultiTool modified miRNA mass calculator user output.

4.23 MiRNA fragment search

The miRNA fragment search feature shown in Figure 4.7 (a) allows the alignment of RNA sequence of any length against all known human miRNA for a match. Once miRNA match(es) containing the queried sequence are found, MicroRNA MultiTool returns the matching miRNA name, accession number and sequence(s) of all miRNAs containing the input sequence fragment shown in Figure 4.7 (b).

4.7 (a)



4.7 (b)

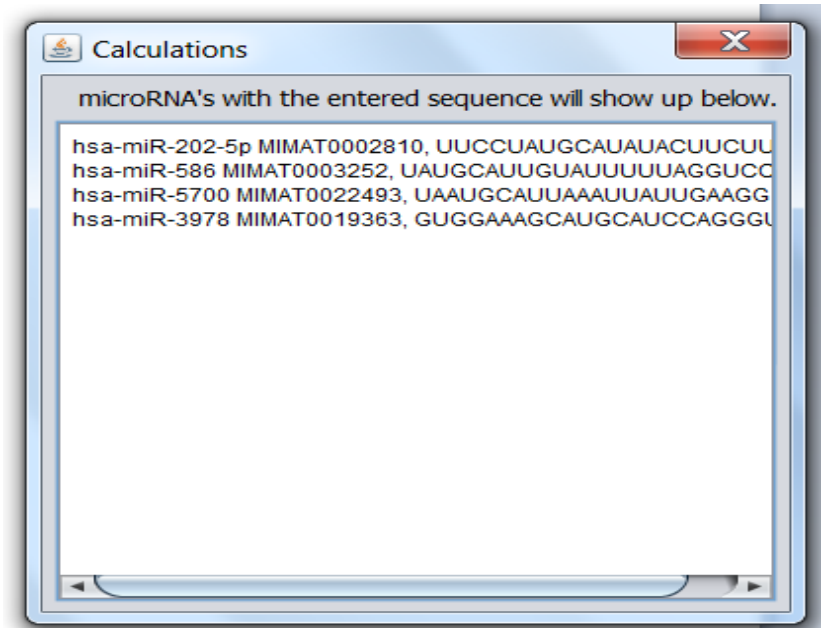


Figure 4.7. (a) MiRNA fragment search feature and (b) its output panel.

4.30 Discussion

The current strategy for identifying miRNA by MS relies on first obtaining the mass of the miRNA, followed by tandem mass spectrometry (MS/MS) in which the miRNA is fragmented through different ways such as collision induced dissociation (CID). The resulting fragment ions are then used to deduce a sequence that can be matched to possible miRNA identity (Izumi et al. 2011). Even though a few groups have recently reported the complete sequence assignment of RNA greater than 19 nucleotides long by MS/MS, this process does not always yield sufficient sequence coverage to enable miRNA identification (Taucher and Breuker 2010), (Taucher et al. 2010), (Huang et al. 2008). The determination of miRNA identity from MS data rather than MS/MS is therefore a simpler, straightforward and more appealing approach to this challenge, however, the bioinformatics tools necessary for the direct identification of miRNA from MS data are lacking.

Currently, there are several bioinformatics tools available for assisting in the interpretation of MS data e.g. McCloskey's group developed the Mongo Oligo Mass Calculator; a multipurpose program for the calculation of oligonucleotide molecular masses, CID fragments as well as endonuclease and exonuclease digestion fragments (Cantara et al. 2011). Simple Oligonucleotide Sequencer (SOS) from the same group is a useful tool for the determination of oligonucleotide sequences from CID MS/MS data (Rozenski and McCloskey 2002). The Compas program developed by Oberacher et al uses comparative sequencing of oligonucleotides for the sequence verification of nucleic acids by matching the predicted CID-MS/MS data with the multiply charged anionic

peaks obtained from an analyzed sample (Oberacher et al. 2004). These current programs have helped in the interpretation of MS data. However, none of these programs is capable of identifying miRNA from MS data. MicroRNA MultiTool offers a simple yet practical solution to this challenge by allowing the user to match molecular masses obtained in a MS experiment against molecular masses of all human miRNA in order to identify the measured miRNA. This approach of identifying miRNA directly from mass spectral data eliminates the need to perform MS/MS experiments for the sole purpose of miRNA identification.

Due to the short nature of miRNAs, isobaric miRNAs i.e. miRNAs with different RNA sequences but identical molecular masses do exist. Under these circumstances, MicroRNA Multitool can be used to narrow down the identity of isobaric miRNA by returning a list of all the isobaric miRNA with the queried mass. For the absolute identification of isobaric miRNAs, currently existing methods of isolation and identification can be used. We have previously demonstrated that isobaric miRNA can be identified by capturing them using complementary probes followed by MS measurement. To help in the identification of miRNA when a partial sequence is available, MicroRNA MultiTool sequence alignment tool can be used. The algorithm searches for miRNA bearing queried partial sequence and matches it to known miRNA. Partial sequences of miR-153 were querried in MicroRNA Multitool and the results of matching number of miRNA tabulated in Table 4.1. From these results, it is clear that as low as 45.4 % sequence information can be used to positively identify hsa-miR-153 whose sequence is 5' pUUGCAUAGUCACAAAAGUGAUC-OH 3'. This is because the partial sequence

[5’ pUUGCAUAGUC 3’] is unique to miR-153. MicroRNA MultiTool therefore provides an opportunity for incomplete sequencing data to be used for miRNA identification.

Table 4.1. Alignment of hsa-miR-153 using MicroRNA MultiTool

5' pUUGCAUAGUCACAAAAGUGAUC-OH 3' (hsa-miR-153)		
Partial sequence	No of miRNAs with partial Seq.	% sequence coverage
pUUGCAUAGUC	1	45.4
pUUGCAUAGU	3	40.9
pUUGCAUAG	3	36.3
pUUGCAUA	5	31.8
pUUGCAU	17	27.2
pUUGCA	58	22.7
pUUGC	174	18.1
pUUG	597	13.6
pUU	1101	9.0
pU	1722	4.5

The presence of post-transcriptionally modified nucleotides in various forms of RNA such as tRNA has been known for many decades. So far, there are more than 100 known RNA modifications (Czerwoniec et al. 2009). The specific roles of most modifications still remains unknown; however, modifications are well conserved across evolution and are thought to be important for the general wellness of the organisms. Just

like other forms of RNA, miRNA are susceptible to different forms of post-transcriptional enzymatic editing and covalent modifications (Jonatha et al. 2007). Nucleotide modifications are capable of altering the way in which miRNA interact with target mRNA molecules (Pantano et al. 2010). This can result to changes in gene regulation (Ebhardt et al. 2009) and the development of various diseases such as cancer (Galeano et al. 2011). Even though most of the RNA modifications are functionally and structurally important, the conventional and currently available molecular biology based techniques such as microarrays, qRT-PCR and northern blots or sequencing methods are incapable of detecting or identifying RNA modifications (Kellner et al. 2010). Except for the "mass silent" pseudouridylation modification which involves the isomerization of uridine to pseudouridine without any molecular mass changes (Taucher et al. 2011), most miRNA modifications result in mass shifts with respect to the unmodified miRNA. Since the mass shifts are characteristic of the modification, MS analysis is a useful tool for the identification of modified miRNA as well as the specific modification involved (Meng and Limbach 2006). For the identification of modified miRNA, the MicroRNA MultiTool algorithm uses a database of all known human miRNA together with known RNA modifications and computes all possible permutations for modified miRNA within the user queried mass. The return includes modified miRNA mass, name, accession number and modification involved. To accurately predict a modified miRNA, the program does not require the sequence, name or accession number of the miRNA, rather, the only input required is the mass of the modified miRNA as obtained from MS experiment. The mass is used to automatically generate a list of all modified miRNA

together with the modification(s) involved. Usually spectra generated from MALDI-MS are easier to analyze since they are mainly dominated by singly charged ions (Bahr et al. 2009) this is also true when using MicroRNA MultiTool, however, in other cases where multiply charged ions are observed such as in electrospray ionization MS, the spectra should be deconvoluted prior to using the MicroRNA MultiTool.

4.40 Conclusions

This study was aimed at simplifying the identification of miRNA from mass spectrometric data. MicroRNA MultiTool provides the ability to search the identity of miRNA using a single parameter e.g. miRNA mass, sequence, accession number or name one at a time and return the other three related parameters. Given that miRNA therapeutic agents are already being explored as possible therapies for numerous diseased conditions, the rapid identification of miRNA will play a key role into the development of such therapies. The development of computational tools for the interpretation of MS data can greatly reduce the complexity associated with analysis of MS data while increasing the speed with which the data can be processed.

REFERENCES

Aitken A. 2005. Identification of proteins by MALDI-TOF MS. pp. 319-324. Humana Press Inc.

Allawi HT, Dahlberg JE, Olson S, Lund E, Olson M, Ma WP, Takova T, Neri BP, Lyamichev VI. 2004. Quantitation of microRNAs using a modified Invader assay. *RNA* **10**(7): 1153-1161.

Backes C, Meese E, Lenhof HP, Keller A. 2010. A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res* **38**(13): 4476-4486.

Bader AG, Brown D, Stoudemire J, Lammers P. 2011. Developing therapeutic microRNAs for cancer. *Gene Ther* **18**(12): 1121-1126.

Bahr U, Aygün H, Karas M. 2009. Sequencing of single and double stranded RNA oligonucleotides by acid hydrolysis and MALDI mass spectrometry. *Anal Chem* **81**(8): 3173-3179.

Bandrés E, Cubedo E, Agirre X, Malumbres R, Zárate R, Ramirez N, Abajo A, Navarro A, Moreno I, Monzó M et al. 2006. Identification by Real-time PCR of 13 mature microRNAs differentially expressed in colorectal cancer and non-tumoral tissues. *Mol Cancer* **5**: 29.

Banoub JH, Limbach PA, Editors. 2010. *Mass Spectrometry Of Nucleosides And Nucleic Acids*. CRC Press.

Bartels C, Tsongalis G. 2009. MicroRNAs: novel biomarkers for human cancer. *Clin Chem* **55**(4): 623-631.

Beavis RC, Chait BT. 1989. Matrix-assisted laser-desorption mass spectrometry using 355 nm radiation. *Rapid Commun Mass Spectrom* **3**(12): 436-439.

Berezikov E, Cuppen E, Plasterk R. 2006. Approaches to microRNA discovery. *Nat Genet* **38 Suppl**: S2-7.

Burroughs AM, Ando Y, de HMJL, Tomaru Y, Nishibu T, Ukekawa R, Funakoshi T, Kurokawa T, Suzuki H, Hayashizaki Y et al. 2010. A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. *Genome Res* **20**(Copyright (C) 2012 American Chemical Society (ACS). All Rights Reserved.): 1398-1410.

Calin G, Croce C. 2006. MicroRNA-cancer connection: the beginning of a new tale. *Cancer Res* **66**(15): 7390-7394.

Calin G, Dumitru C, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K et al. 2002. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* **99**(24): 15524-15529.

Calin G, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik S, Iorio M, Visone R, Sever N, Fabbri M et al. 2005. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* **353**(17): 1793-1801.

Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF. 2011. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res* **39**(Database issue): D195-201.

Carthew RW, Sontheimer EJ. 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**(4): 642-655.

Chait BT. 2011. Mass spectrometry in the postgenomic era. *Annu Rev Biochem* **80**: 239-246.

Chandramouli K, Qian PY. 2009. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics* **2009**.

Chen C, Ridzon D, Broomer A, Zhou Z, Lee D, Nguyen J, Barbisin M, Xu N, Mahuvakar V, Andersen M et al. 2005. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* **33**(20): e179.

Chen H, Li J. 2007. Nanotechnology: moving from microarrays toward nanoarrays. *Methods Mol Biol* **381**: 411-436.

Chen J, Lozach J, Garcia E, Barnes B, Luo S, Mikoulitch I, Zhou L, Schroth G, Fan J. 2008. Highly sensitive and specific microRNA expression profiling using BeadArray technology. *Nucleic Acids Res* **36**(14): e87.

Chiu NHL, Barnes, Charles A. . 2009. Accurate characterization of carcinogenic DNA adducts using MALDI tandemtime-of-flight mass spectrometry. *International Journal of Mass Spectrometry* **279**: 170–175.

Cho WC. 2007. OncomiRs: the discovery and progress of microRNAs in cancers. *Mol Cancer* **6**: 60.

Cissell K, Shrestha S, Deo S. 2007. MicroRNA Detection: Challenges for the Analytical Chemist. *Analytical Chemistry* **79**: 4754-4761.

Cohen SL, Chait BT. 1996. Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins. *Anal Chem* **68**(1): 31-37.

Creighton CJ, Reid JG, Gunaratne PH. 2009. Expression profiling of microRNAs by deep sequencing. *Briefings Bioinf* **10**(Copyright (C) 2012 American Chemical Society (ACS). All Rights Reserved.): 490-497.

Czerwoniec A, Dunin-Horkawicz S, Purta E, Kaminska K, Kasprzak J, Bujnicki J, Grosjean H, Rother K. 2009. MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res* **37**(Database issue): D118-121.

Davison TS, Johnson CD, Andruss BF. 2006. Analyzing micro-RNA expression using microarrays. *Methods Enzymol* **411**: 14-34.

Douthwaite S, Kirpekar F. 2007. Identifying modifications in RNA by MALDI mass spectrometry. *Methods Enzymol* **425**: 1-20.

Durairaj A, Limbach P. 2008. Mass spectrometry of the fifth nucleoside: a review of the identification of pseudouridine in nucleic acids. *Anal Chim Acta* **623**(2): 117-125.

Ebhardt H, Tsang H, Dai D, Liu Y, Bostan B, Fahlman R. 2009. Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res* **37**(8): 2461-2470.

Feliciano A, Sánchez-Sendra B, Kondoh H, Lleonart ME. 2011. MicroRNAs Regulate Key Effector Pathways of Senescence. *J Aging Res* **2011**: 205378.

Ferdin J, Kunej T, Calin GA. 2010. Non-coding RNAs: identification of cancer-associated microRNAs by gene profiling. *Technol Cancer Res Treat* **9**(Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.): 123-138.

Fiorucci G, Chiantore MV, Mangino G, Percario ZA, Affabris E, Romeo G. 2012. Cancer regulator microRNA: potential relevance in diagnosis, prognosis and treatment of cancer. *Curr Med Chem* **19**(4): 461-474.

Friedman R, Farh K, Burge C, Bartel D. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**(1): 92-105.

Fu Y, Xu S, Pan C, Ye M, Zou H, Guo B. 2006. A matrix of 3,4-diaminobenzophenone for the analysis of oligonucleotides by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Nucleic Acids Res* **34**(13): e94.

Galeano F, Tomaselli S, Locatelli F, Gallo A. 2011. A-to-I RNA editing: The "ADAR" side of human cancer. *Semin Cell Dev Biol*.

Garcia BA, Heaney PJ, Tang K. 2002. Improvement of the MALDI-TOF analysis of DNA with thin-layer matrix preparation. *Anal Chem* **74**(9): 2083-2091.

Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, Bertone P, Caldas C. 2010. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* **16**(5): 991-1006.

Griffiths-Jones S, Grocock R, van Dongen S, Bateman A, Enright A. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**(Database issue): D140-144.

Griffiths-Jones S, Saini H, van Dongen S, Enright A. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**(Database issue): D154-158.

Guenther S, Koestler M, Schulz O, Spengler B. 2010. Laser spot size and laser power dependence of ion formation in high resolution MALDI imaging. *Int J Mass Spectrom* **294**(Copyright (C) 2012 American Chemical Society (ACS). All Rights Reserved.): 7-15.

Hahner S, Lüdemann H, Kirpekar F, Nordhoff E, Roepstorff P, Galla H, Hillenkamp F. 1997a. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI) of endonuclease digests of RNA. *Nucleic Acids Res* **25**(10): 1957-1964.

Hahner S, Lüdemann HC, Kirpekar F, Nordhoff E, Roepstorff P, Galla HJ, Hillenkamp F. 1997b. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI) of endonuclease digests of RNA. *Nucleic Acids Res* **25**(10): 1957-1964.

Hartmer R, Storm N, Boecker S, Rodi CP, Hillenkamp F, Jurinke C, van den Boom D. 2003. RNase T1 mediated base-specific cleavage and MALDI-TOF MS for high-throughput comparative sequence analysis. *Nucleic Acids Res* **31**(9): e47.

Holle A, Haase A, Kayser M, Hoehndorf J. 2006. Optimizing UV laser focus profiles for improved MALDI performance. *J Mass Spectrom* **41**(Copyright (C) 2012 American Chemical Society (ACS). All Rights Reserved.): 705-716.

Hossain M, Limbach PA. 2007. Mass spectrometry-based detection of transfer RNAs by their signature endonuclease digestion products. *RNA* **13**(2): 295-303.

Hossain M, Limbach PA. 2009. Multiple endonucleases improve MALDI-MS signature digestion product detection of bacterial transfer RNAs. *Anal Bioanal Chem* **394**(4): 1125-1135.

Huang TY, Kharlamova A, Liu J, McLuckey SA. 2008. Ion trap collision-induced dissociation of multiply deprotonated RNA: c/y-ions versus (a-B)/w-ions. *J Am Soc Mass Spectrom* **19**(12): 1832-1840.

Hunt E, Goulding A, Deo S. 2009. Direct detection and quantification of microRNAs. *Anal Biochem* **387**(1): 1-12.

Hutvágner G, Zamore P. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297**(5589): 2056-2060.

Iida K, Jin H, Zhu JK. 2009. Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in Arabidopsis thaliana. *BMC Genomics* **10**: 155.

Iorio M, Ferracin M, Liu C, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M et al. 2005. MicroRNA gene expression deregulation in human breast cancer. *Cancer Res* **65**(16): 7065-7070.

Izumi Y, Takimura S, Yamaguchi S, Iida J, Bamba T, Fukusaki E. 2011. Application of electrospray ionization ion trap/time-of-flight mass spectrometry for chemically-synthesized small RNAs. *J Biosci Bioeng*.

Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. 2009. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* **37**(Database issue): D98-104.

Jonatha, M, Gott. 2007. *Methods in Enzymology*. Academic Press, San Diego, California, USA.

Jonstrup S, Koch J, Kjems J. 2006. A microRNA detection system based on padlock probes and rolling circle amplification. *RNA* **12**(9): 1747-1752.

Juhasz P, Roskey MT, Smirnov IP, Haff LA, Vestal ML, Martin SA. 1996. Applications of delayed extraction matrix-assisted laser desorption ionization time-of-flight mass spectrometry to oligonucleotide analysis. *Anal Chem* **68**(6): 941-946.

Kellner S, Burhenne J, Helm M. 2010. Detection of RNA modifications. *RNA Biol* **7**(2): 237-247.

Keough T, Baker TR, Dobson RL, Lacey MP, Riley TA, Hasselfield JA, Hesselberth PE. 1993. Antisense DNA oligonucleotides. II: The use of matrix-assisted laser desorption/ionization mass spectrometry for the sequence verification of methylphosphonate oligodeoxyribonucleotides. *Rapid Commun Mass Spectrom* **7**(3): 195-200.

Kirpekar F, Douthwaite S, Roepstorff P. 2000. Mapping posttranscriptional modifications in 5S ribosomal RNA by MALDI mass spectrometry. *RNA* **6**(2): 296-306.

Klenow H, Henningsen I. 1970. Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. *Proc Natl Acad Sci U S A* **65**(1): 168-175.

Kong W, Zhao JJ, He L, Cheng JQ. 2009. Strategies for profiling microRNA expression. *J Cell Physiol* **218**(1): 22-25.

Krause E, Wenschuh H, Jungblut PR. 1999. The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal Chem* **71**(19): 4160-4165.

Kumar P, Johnston BH, Kazakov SA. 2011. miR-ID: a novel, circularization-based platform for detection of microRNAs. *RNA* **17**(2): 365-380.

Lee I, Ajay S, Chen H, Maruyama A, Wang N, McInnis M, Athey B. 2008. Discriminating single-base difference miRNA expressions using microarray Probe Design Guru (ProDeG). *Nucleic Acids Res* **36**(5): e27.

Lee Y, Dutta A. 2009. MicroRNAs in cancer. *Annu Rev Pathol* **4**: 199-227.

Li W, Ruan K. 2009. MicroRNA detection by microarray. *Anal Bioanal Chem* **394**(4): 1117-1124.

Lieu PT, Jozsi P, Gilles P, Peterson T. 2005. Development of a DNA-labeling system for array-based comparative genomic hybridization. *J Biomol Tech* **16**(2): 104-111.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA et al. 2005. MicroRNA expression profiles classify human cancers. *Nature* **435**(7043): 834-838.

Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. 2008. An analysis of human microRNA and disease associations. *PLoS One* **3**(10): e3420.

Maiti M, Nauwelaerts K, Herdewijn P. 2012. Pre-microRNA binding aminoglycosides and antitumor drugs as inhibitors of Dicer catalyzed microRNA processing. *Bioorg Med Chem Lett* **22**(4): 1709-1711.

Melo SA, Esteller M. 2011. Dysregulation of microRNAs in cancer: Playing with fire. *FEBS Lett* **585**(Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.): 2087-2099.

Meltzer P. 2005. Cancer genomics: small RNAs with big impacts. *Nature* **435**(7043): 745-746.

Meng Z, Limbach PA. 2006. Mass spectrometry of RNA: linking the genome to the proteome. *Brief Funct Genomic Proteomic* **5**(1): 87-95.

Mengel-Jørgensen J, Kirpekar F. 2002. Detection of pseudouridine and other modifications in tRNA by cyanoethylation and MALDI mass spectrometry. *Nucleic Acids Res* **30**(23): e135.

Metzker ML. 2005. Emerging technologies in DNA sequencing. *Genome Res* **15**(12): 1767-1776.

Mizuguchi Y, Mishima T, Yokomuro S, Arima Y, Kawahigashi Y, Shigehara K, Kanda T, Yoshida H, Uchida E, Tajiri T et al. 2011. Sequencing and bioinformatics-based analyses of the microRNA transcriptome in hepatitis B-related hepatocellular carcinoma. *PLoS One* **6**(1): e15304.

Morozova O, Marra M. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**(5): 255-264.

Motameny S, Wolters S, Nuernberg P, Schumacher B. 2010. Next generation sequencing of miRNAs - strategies, resources and methods. *Genes* **1**(Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.): 70-84.

Munafó DB, Robb GB. 2010. Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. *RNA* **16**(12): 2537-2552.

Nakayama H, Takahashi N, Isobe T. 2011. Informatics for mass spectrometry-based RNA analysis. *Mass Spectrom Rev* **30**(6): 1000-1012.

Nana-Sinkam SP, Croce CM. 2011. MicroRNAs as therapeutic targets in cancer. *Transl Res* **157**(4): 216-225.

Negrini M, Nicoloso M, Calin G. 2009. MicroRNAs and cancer--new paradigms in molecular oncology. *Curr Opin Cell Biol* **21**(3): 470-479.

Nelson PT, Baldwin DA, Scearce LM, Oberholtzer JC, Tobias JW, Mourelatos Z. 2004. Microarray-based, high-throughput gene expression profiling of microRNAs. *Nat Methods* **1**(2): 155-161.

Nishikura K. 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**: 321-349.

Niu S, Zhang W, Chait BT. 1998. Direct comparison of infrared and ultraviolet wavelength matrix-assisted laser desorption/ionization mass spectrometry of proteins. *J Am Soc Mass Spectrom* **9**(1): 1-7.

Oberacher H, Parson W, Oefner PJ, Mayr BM, Huber CG. 2004. Applicability of tandem mass spectrometry to the automated comparative sequencing of long-chain oligonucleotides. *J Am Soc Mass Spectrom* **15**(4): 510-522.

Oberacher H, Pitterl F. 2009. On the use of ESI-QqTOF-MS/MS for the comparative sequencing of nucleic acids. *Biopolymers* **91**(Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.): 401-409.

Oberacher H, Pitterl F. 2011. Tandem mass spectrometric de novo sequencing of oligonucleotides using simulated annealing for stochastic optimization. *Int J Mass Spectrom* **304**(Copyright (C) 2012 American Chemical Society (ACS). All Rights Reserved.): 124-129.

Oivanen M, Kuusela S, Lönnberg H. 1998. Kinetics and Mechanisms for the Cleavage and Isomerization of the Phosphodiester Bonds of RNA by Brønsted Acids and Bases. *Chem Rev* **98**(3): 961-990.

Omahen DA. 2011. MicroRNA and diseases of the nervous system. *Neurosurgery* **69**(2): 440-454.

Onnerfjord P, Nilsson J, Wallman L, Laurell T, Marko-Varga G. 1998. Picoliter sample preparation in MALDI-TOF MS using a micromachined silicon flow-through dispenser. *Anal Chem* **70**(22): 4755-4760.

Padró JM, Acquaviva A, Tascon M, Gagliardi LG, Castells CB. 2012. Effect of temperature and solvent composition on acid dissociation equilibria, I: Sequenced determination of compounds commonly used as buffers in high performance liquid chromatography coupled to mass spectroscopy detection. *Analytica Chimica Acta* **725**(0): 87-94.

Pantano L, Estivill X, Martí E. 2010. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res* **38**(5): e34.

Park N, Zhou H, Elashoff D, Henson B, Kastratovic D, Abemayor E, Wong D. 2009. Salivary microRNA: discovery, characterization, and clinical utility for oral cancer detection. *Clin Cancer Res* **15**(17): 5473-5477.

Perreault DM, Anslyn EV. 1997. Unifying the current data on the mechanism of cleavage-transesterification of RNA. *Angew Chem, Int Ed Engl* **36**(Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.): 433-450.

Pieles U, Zürcher W, Schär M, Moser HE. 1993. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic Acids Res* **21**(14): 3191-3196.

Purohit V, Grindley ND, Joyce CM. 2003. Use of 2-aminopurine fluorescence to examine conformational changes during nucleotide incorporation by DNA polymerase I (Klenow fragment). *Biochemistry* **42**(34): 10200-10211.

Qavi AJ, Kindt JT, Bailey RC. 2010. Sizing up the future of microRNA analysis. *Anal Bioanal Chem* **398**(6): 2535-2549.

Rozenski J, McCloskey JA. 2002. SOS: a simple interactive program for ab initio oligonucleotide sequencing by mass spectrometry. *J Am Soc Mass Spectrom* **13**(3): 200-203.

Satoh T, Sato T, Kubo A, Tamura J. 2011. Tandem Time-of-Flight Mass Spectrometer with High Precursor Ion Selectivity Employing Spiral Ion Trajectory and Improved Offset Parabolic Reflectron. *Journal of The American Society for Mass Spectrometry* **22**(5): 797-803.

Satoh T, Sato T, Tamura J. 2007. Development of a high-performance MALDI-TOF mass spectrometer utilizing a spiral ion trajectory. *J Am Soc Mass Spectrom* **18**(7): 1318-1323.

Satoh T, Tsuno H, Iwanaga M, Kammei Y. 2006. A new Spiral Time-of-Flight Mass Spectrometer for High Mass Analysis. *J Mass Spectrom Soc Jpn* **541**(1): 11-17.

Schatz P, Dietrich D, Schuster M. 2004. Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF. *Nucleic Acids Res* **32**(21): e167.

Schuette JM, Pieles U, Maleknia SD, Srivatsa GS, Cole DL, Moser HE, Afeyan NB. 1995. Sequence analysis of phosphorothioate oligonucleotides via matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *J Pharm Biomed Anal* **13**(10): 1195-1203.

Senko MW, Beu SC, McLafferty FW. 1995. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of The American Society for Mass Spectrometry* **6**(4): 229-233.

Sharbati-Tehrani S, Kutz-Lohroff B, Bergbauer R, Scholven J, Einspanier R. 2008. miR-Q: a novel quantitative RT-PCR approach for the expression profiling of small RNA molecules such as miRNAs in a complex sample. *BMC Mol Biol* **9**: 34.

Sipova H, Zhang S-L, Dudley AM, Galas D, Wang K, Homola J. 2010. Surface plasmon resonance biosensor for rapid label-free detection of microribonucleic acid at subfemtomole level. *Anal Chem (Washington, DC, U S)* **82**(Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.): 10110-10115.

Sun Y, Gregory K, Golovlev V. 2009. Efficiency and specificity of microRNA-primed nucleotide analog incorporation by various DNA polymerases. *Anal Biochem* **391**(2): 85-90.

Taucher M, Breuker K. 2010. Top-down mass spectrometry for sequencing of larger (up to 61 nt) RNA by CAD and EDD. *J Am Soc Mass Spectrom* **21**(6): 918-929.

Taucher M, Ganisl B, Breuker K. 2011. Identification, localization, and relative quantitation of pseudouridine in RNA by tandem mass spectrometry of hydrolysis products. *Int J Mass Spectrom* **304**(2-3): 91-97.

Taucher M, Rieder U, Breuker K. 2010. Minimizing base loss and internal fragmentation in collisionally activated dissociation of multiply deprotonated RNA. *J Am Soc Mass Spectrom* **21**(2): 278-285.

Thomson J, Parker J, Perou C, Hammond S. 2004. A custom microarray platform for analysis of microRNA gene expression. *Nat Methods* **1**(1): 47-53.

Tolson DA, Nicholson NH. 1998. Sequencing RNA by a combination of exonuclease digestion and uridine specific chemical cleavage using MALDI-TOF. *Nucleic Acids Res* **26**(2): 446-451.

van Rooij E, Olson E. 2007. microRNAs put their signatures on the heart. *Physiol Genomics* **31**(3): 365-366.

Verma S, Eckstein F. 1998. Modified oligonucleotides: synthesis and strategy for users. *Annu Rev Biochem* **67**: 99-134.

Várallyay E, Burgyán J, Havelda Z. 2007. Detection of microRNAs by Northern blot analyses using LNA probes. *Methods* **43**(2): 140-145.

Várallyay E, Burgyán J, Havelda Z. 2008. MicroRNA detection by northern blotting using locked nucleic acid probes. *Nat Protoc* **3**(2): 190-196.

Wambua DM, Tannous BA, Chiu NHL. Creating Mass Signatures for the Detection of MicroRNA. *Anal. Methods, 2012, 4 (10), 3453 – 3459*.

Wark A, Lee H, Corn R. 2008. Multiplexed detection methods for profiling microRNA expression in biological samples. *Angew Chem Int Ed Engl* **47**(4): 644-652.

Wiemer E. 2007. The role of microRNAs in cancer: no small matter. *Eur J Cancer* **43**(10): 1529-1544.

Xu L, Anchordoquy T. 2011. Drug delivery trends in clinical trials and translational medicine: challenges and opportunities in the delivery of nucleic acid-based therapeutics. *J Pharm Sci* **100**(1): 38-52.

Yang W, Chiu NHL. 2010. Correlation between UV spectrophotometry and quantitative MALDI-TOF MS. *Spectroscopy Letters* **43**(7): 602-608.

APPENDIX A

MICRORNA MULTITOOL MAIN JAVA CLASS

MicroRNA MultiTool Main Java Class is written in Java language and the code that determines the algorithm of the calculations is given below.

```
1.  package rna;

2.  import java.io.BufferedReader;
3.  import java.io.FileReader;
4.  import java.io.IOException;
5.  import java.text.DecimalFormat;
6.  import java.util.ArrayList;
7.  import java.util.Collections;
8.  import java.util.List;
9.  import java.util.Scanner;
10. import java.awt.*;
11. import javax.swing.*;
12. import java.awt.event.*;
13. import java.net.URL;

14. /* This program was created by
15. Norman H.L Chiu, Dickson M. Wambua and Zhonghao Cui
16. For the use of calculating the masses of
17. user-input RNA sequences and of calculating
18. the masses of modified strands of RNA,
19. in addition to predicting the presence of modified
20. RNA sequences.
21. *
22. June 9th, 2011
23. */
24. public class RNA {

25. public static String RNAinput;
26. public static double AdeMass = 329.2091;
27. public static double CytMass = 305.1840;
28. public static double UriMass = 306.1687;
29. public static double GuaMass = 345.2084;
30. public static double QueMass = 409.3998;
31. public static double ThyMass = 304.1964;
32. public static double OHmass = 17.0027;
```

```
33.  public static double Phosmass = 95.9793;
34.  public static double positivenegative = 2.01565;
35.  public static double Acount;
36.  public static double Ccount;
37.  public static double Ucount;
38.  public static double Gcount;
39.  public static double Tcount;
40.  public static double Qcount;
41.  public static double five_end;
42.  public static double three_end;
43.  public static double TotalMass;
44.  public static String[] words;
45.  public static String[] microRNAs;
46.  public static double mass;
47.  public static boolean match = false;
48.  public static int wrong;

49.  public RNA() {
50.  // Default, with Avg mass, Positive Mode (Protonated), 5' Phos and 3' OH terms.
51.  AdeMass = 329.2091;
52.  CytMass = 305.1840;
53.  UriMass = 306.1687;
54.  GuaMass = 345.2084;
55.  positivenegative = 2.0157;
56.  }

57.  public RNA(boolean mono_iso, boolean pos_neg, boolean five_prime, boolean
     three_prime) {
58.  if (mono_iso) {
59.  AdeMass = 329.05252;
60.  CytMass = 305.04129;
61.  UriMass = 306.02530;
62.  GuaMass = 345.04744;
63.  QueMass = 409.15974;
64.  ThyMass = 304.04604;
65.  } else if (!mono_iso) {
66.  AdeMass = 329.2091;
67.  CytMass = 305.1840;
68.  UriMass = 306.1687;
69.  GuaMass = 345.2084;
70.  QueMass = 409.3998;
71.  ThyMass = 304.1964;
72.  }
73.  if (pos_neg) {
```

```
74.     positivenegative = 2.01565;
75.     } else if (!pos_neg) {
76.     positivenegative = 0;
77.     }
78.     if (five_prime) {
79.     five_end = OHmass;
80.     } else if (!five_prime) {
81.     five_end = Phosmass;
82.     }
83.     if (three_prime) {
84.     three_end = OHmass;
85.     } else if (!three_prime) {
86.     three_end = Phosmass;
87.     }

88.     }

89.     // Gets an RNA sequence from the user:
90.     public static String getInput(String temp) {
91.     RNAinput = temp;
92.     for (int i = 0; i < RNAinput.length(); i++) {
93.     if (RNAinput.charAt(i) != 'A' && RNAinput.charAt(i) != 'C'
94.     && RNAinput.charAt(i) != 'U' && RNAinput.charAt(i) != 'G' &&
        RNAinput.charAt(i) != 'Q') {
95.     System.out.println("This is not a valid RNA sequence.");
96.     }
97.     }
98.     return RNAinput;
99.     }

100.    public static boolean testInput(String temp) {
101.    boolean valid=true;
102.    temp = temp.toUpperCase();
103.    temp = temp.replaceAll(" ", "");
104.    RNAinput = temp;
105.    for (int i = 0; i < RNAinput.length(); i++) {
106.    if (RNAinput.charAt(i) != 'A' && RNAinput.charAt(i) != 'C'
107.    && RNAinput.charAt(i) != 'U' && RNAinput.charAt(i) != 'G' &&
        RNAinput.charAt(i) != 'Q' && RNAinput.charAt(i) != 'T') {
108.    valid = false;
109.    }
110.    }
111.    return valid;
112.    }
```

```
113.  // Accesses/imports a database of possible RNA modifications
114.  // for use in the program.
115.  public static String[] getModDBfromTextFile() {
116.  List<String> wordList = new ArrayList<String>();
117.  BufferedReader br = null;
118.  try {
119.  br = new BufferedReader(new FileReader("ModDB.txt"));
120.  String word;
121.  while ((word = br.readLine()) != null) {
122.  wordList.add(word);
123.  }
124.  } catch (IOException e) {
125.  e.printStackTrace();
126.  } finally {
127.  try {
128.  br.close();
129.  } catch (IOException ex) {
130.  ex.printStackTrace();
131.  }
132.  }
133.  String[] temp = new String[wordList.size()];
134.  words = temp;
135.  wordList.toArray(words);

136.  return words;
137.  }

138.  // Accesses/imports a database of all mature homo sapien miRNA
139.  // sequences.
140.  public static String[] getRNADBfromTextFile() {
141.  List<String> wordList = new ArrayList<String>();
142.  BufferedReader br = null;
143.  try {
144.  br = new BufferedReader(new FileReader("miRDB.txt"));
145.  String RNAword;
146.  while ((RNAword = br.readLine()) != null) {
147.  wordList.add(RNAword);
148.  }
149.  } catch (IOException e) {
150.  e.printStackTrace();
151.  } finally {
152.  try {
153.  br.close();
```

```
154.    } catch (IOException ex) {
155.    ex.printStackTrace();
156.    }
157.    }
158.    String[] temp = new String[wordList.size()];
159.    microRNAs = temp;
160.    wordList.toArray(microRNAs);
161.    return microRNAs;
162.    }

163.    // Counts number of Adenines in the user-input RNA sequence.
164.    public static double countAdenine() {
165.    for (int i = 0; i < RNAinput.length(); i++) {
166.    if (RNAinput.charAt(i) == 'A') {
167.    Acount++;
168.    }
169.    }
170.    return Acount;
171.    }

172.    // Counts number of Cytosine in the user-input RNA sequence.
173.    public static double countCytosine() {
174.    for (int i = 0; i < RNAinput.length(); i++) {
175.    if (RNAinput.charAt(i) == 'C') {
176.    Ccount++;
177.    }
178.    }
179.    return Ccount;
180.    }

181.    // Counts number of Uridine in the user-input RNA sequence.
182.    public static double countUridine() {
183.    for (int i = 0; i < RNAinput.length(); i++) {
184.    if (RNAinput.charAt(i) == 'U') {
185.    Ucount++;
186.    }
187.    }
188.    return Ucount;
189.    }

190.    // Counts number of Guanine in the user-input RNA sequence.
191.    public static double countGuanine() {
192.    for (int i = 0; i < RNAinput.length(); i++) {
193.    if (RNAinput.charAt(i) == 'G') {
```

```
194.    Gcount++;
195.    }
196.    }
197.    return Gcount;
198.    }

199.    // Counts number of Thymidine in the user-input RNA sequence.
200.    public static double countThymidine() {
201.    for (int i = 0; i < RNAinput.length(); i++) {
202.    if (RNAinput.charAt(i) == 'T') {
203.    Tcount++;
204.    }
205.    }
206.    return Tcount;
207.    }

208.    // Counts number of Queuosines in the user-inout RNA sequence.
209.    public static double countQueuosine() {
210.    for (int i = 0; i < RNAinput.length(); i++) {
211.    if (RNAinput.charAt(i) == 'Q') {
212.    Qcount++;
213.    }
214.    }
215.    return Qcount;
216.    }

217.    // Returns the mass of RNA sequence.
218.    public static double calcTotalMass(String temp) {
219.    getInput(temp);
220.    Acount = 0;
221.    Gcount = 0;
222.    Ccount = 0;
223.    Ucount = 0;
224.    Qcount = 0;
225.    Tcount = 0;
226.    countAdenine();
227.    countCytosine();
228.    countGuanine();
229.    countUridine();
230.    countQueuosine();
231.    countThymidine();
232.    TotalMass = (Acount * AdeMass) + (Ccount * CytMass)
233.    + (Ucount * UriMass) + (Gcount * GuaMass) + (Tcount * ThyMass) + (Qcount *
        QueMass) - Phosmass + five_end + three_end + positivenegative;
```

```
234.  String TotalMassString = Double.toString(TotalMass);
235.  int decimal = TotalMassString.indexOf(".");
236.  TotalMassString = TotalMassString.substring(0,decimal+6);
237.  TotalMass = Double.parseDouble(TotalMassString);
238.  return TotalMass;
239.  }

240.  public static String getHsaName(String sequence) {
241.  String[] rnadb = getRNADBfromTextFile();
242.  for (int i = 1; i < rnadb.length; i += 2) {
243.  if (rnadb[i].equals(sequence)) {
244.  int hsaEnd = rnadb[i-1].indexOf(" ");
245.  return rnadb[i - 1].substring(0,hsaEnd);
246.  }
247.  }
248.  return "";
249.  }

250.  public static String getHsaMimat(String mimat){
251.  String[] rnadb = getRNADBfromTextFile();
252.  for (int i = 0; i < rnadb.length; i += 2) {
253.  int hsaEnd = rnadb[i].indexOf(" ");
254.  if (rnadb[i].substring(hsaEnd+1,rnadb[i].length()).equals(mimat)) {
255.  return rnadb[i].substring(0,hsaEnd);
256.  }
257.  }
258.  return "";
259.  }

260.  public static String getMimat(String sequence){
261.  String[] rnadb = getRNADBfromTextFile();
262.  for (int i = 1; i < rnadb.length; i += 2) {
263.  int hsaEnd = rnadb[i-1].indexOf(" ");
264.  if (rnadb[i].equals(sequence)) {
265.  return rnadb[i - 1].substring(hsaEnd+1, rnadb[i-1].length());
266.  }
267.  }
268.  return "";
269.  }

270.  public static String getSequence(String miRname) {
271.  String[] rnadb = getRNADBfromTextFile();
272.  for (int i = 0; i < rnadb.length; i += 2) {
273.  int hsaEnd = rnadb[i].indexOf(" ");
```

```
274.  if (rnadb[i].substring(0,hsaEnd).equals(miRname)) {
275.  return rnadb[i + 1];
276.  }
277.  }
278.  return "";
279.  }

280.  // Method that returns miRNA sequences with similar weights to
281.  // the one input.
282.  public static String getPossibleRNA(double testMass, double Range) {
283.  ArrayList<String> RNAmatches = null;
284.  StringBuffer sb = new StringBuffer();
285.  String testString = Double.toString(testMass);
286.  RNAmatches = new ArrayList<String>();
287.  getRNADBfromTextFile();
288.  for (int i = 1; i < microRNAs.length - 1; i += 2) {
289.  double datmass = calcTotalMass(microRNAs[i]);
290.  String dataMass = Double.toString(datmass);
291.  int decimal = dataMass.indexOf(".");
292.  dataMass = dataMass.substring(0,decimal+6);
293.  String RNAmass = new String(microRNAs[i - 1] + ", " + microRNAs[i]
294.  + ", " + dataMass)
295.  + " Da";
296.  Double dbmass = calcTotalMass(microRNAs[i]);
297.  if (dbmass >= testMass - Range && dbmass <= testMass + Range) {
298.  RNAmatches.add(RNAmass);
299.  }
300.  }
301.  for (String s : RNAmatches) {
302.  sb.append(s + "\n");
303.  }
304.  String matches = sb.toString();
305.  return matches;
306.  }

307.  // Finds the input modification in the database and retrieves
308.  // the mass of the modification.
309.  public static double getMassofMod(String mod) {
310.  getModDBfromTextFile();
311.  for (int i = 0; i < words.length; i++) {
312.  if (words[i].equals(mod)) {
313.  if (mod.length() <= 4) {
314.  String sr = words[i + 2];
315.  mass = Double.parseDouble(sr);
```

```
316.    return mass;
317.    } else {
318.    String sr = words[i + 1];
319.    mass = Double.parseDouble(sr);
320.    return mass;
321.    }
322.    }
323.    }
324.    System.out.println("The modification you entered is invalid.");
325.    return -1;
326.    }

327.    // Method that returns new mass of a RNA string after modifications
328.    // have been made.
329.    public static double findNewModMass(String temp, char einin, double modimass)
    {
330.    temp = temp.toUpperCase();
331.    getInput(temp);
332.    countAdenine();
333.    countCytosine();
334.    countUridine();
335.    countGuanine();
336.    countQueuosine();
337.    if (einin == 'A' && Acount!=0) {
338.    double newMass = ((Acount - 1) * AdeMass) + (Ccount * CytMass)
339.    + (Ucount * UriMass) + (Gcount * GuaMass) + (Qcount * QueMass) + modimass
340.    Phosmass + five_end + three_end + positivenegative;
341.    return newMass;
342.    }
343.    else if (einin == 'A' && Acount==0){
344.    return calcTotalMass(temp);
345.    }
346.    else if (einin == 'C' && Ccount!=0) {
347.    double newMass = (Acount * AdeMass) + ((Ccount - 1) * CytMass)
348.    + (Ucount * UriMass) + (Gcount * GuaMass) + (Qcount * QueMass) + modimass
349.    Phosmass + five_end + three_end + positivenegative;
350.    return newMass;
351.    }
352.    else if (einin == 'C' && Ccount==0){
353.    return calcTotalMass(temp);
354.    }
355.    else if (einin == 'U' && Ucount!=0) {
356.    double newMass = (Acount * AdeMass) + (Ccount * CytMass)
```

```
357.  + ((Ucount - 1) * UriMass) + (Gcount * GuaMass) + (Qcount * QueMass) +
      modimass
358.  Phosmass + five_end + three_end + positivenegative;
359.  return newMass;
360.  }
361.  else if (einin == 'U' && Ucount==0){
362.  return calcTotalMass(temp);
363.  }
364.  else if (einin == 'G' && Gcount!=0) {
365.  double newMass = (Acount * AdeMass) + (Ccount * CytMass)
366.  + (Ucount * UriMass) + ((Gcount - 1) * GuaMass )+ (Qcount * QueMass) +
      modimass
367.  Phosmass + five_end + three_end + positivenegative;
368.  return newMass;
369.  }
370.  else if (einin == 'G' && Gcount==0){
371.  return calcTotalMass(temp);
372.  }
373.  if (einin == 'Q' && Qcount!=0) {
374.  double newMass = (Acount * AdeMass) + (Ccount * CytMass)
375.  + (Ucount * UriMass) + (Gcount * GuaMass) + ((Qcount-1) * QueMass) +
      modimass
376.  Phosmass + five_end + three_end + positivenegative;
377.  return newMass;
378.  }
379.  return -1;
380.  }

381.  // Method that returns new mass of a RNA string after modifications
382.  // have been made.
383.  public static double apply_new_mod(double startMass, char einin, double
      modmass, String startString) {
384.  double finalMass = 0;
385.  Acount = 0;
386.  Ucount = 0;
387.  Gcount = 0;
388.  Ccount = 0;
389.  Qcount = 0;
390.  startString.toUpperCase();
391.  getInput(startString);
392.  countAdenine();
393.  countGuanine();
394.  countCytosine();
395.  countUridine();
```

```
396.    countQueuosine();
397.    if (einin == 'A' && Acount != 0) {
398.    finalMass = startMass - AdeMass + modmass;
399.    }
400.    if (einin == 'A' && Acount == 0){
401.    finalMass = startMass;
402.    }
403.    if (einin == 'C' && Ccount != 0) {
404.    finalMass = startMass - CytMass + modmass;
405.    }
406.    if (einin == 'C' && Ccount == 0){
407.    finalMass = startMass;
408.    }
409.    if (einin == 'G' && Gcount != 0) {
410.    finalMass = startMass - GuaMass + modmass;
411.    }
412.    if (einin == 'G' && Gcount == 0){
413.    finalMass = startMass;
414.    }
415.    if (einin == 'U' && Ucount != 0) {
416.    finalMass = startMass - UriMass + modmass;
417.    }
418.    if (einin == 'U' && Ucount == 0){
419.    finalMass = startMass;
420.    }
421.    if (einin == 'Q' && Qcount != 0) {
422.    finalMass = startMass - QueMass + modmass;
423.    }
424.    if (einin == 'Q' && Qcount == 0){
425.    finalMass = startMass;
426.    }
427.    return finalMass;
428.    }

429.    // Method that, when given a sequence of RNA, returns all the weights
430.    // of that RNA with one modification. All of the modifications
431.    // in the database are used.
432.    public static ArrayList<String> allModdedMasses(String hsamir) {
433.    ArrayList<String> massList = null;
434.    massList = new ArrayList<String>();
435.    String rnaseq = null;
436.    getModDBfromTextFile();
437.    getRNADBfromTextFile();
438.    for (int i = 0; i < microRNAs.length; i += 2) {
```

```
439.  if (hsamir.equals(microRNAs[i])) {
440.  rnaseq = microRNAs[i + 1];
441.  break;
442.  }
443.  }
444.  for (int i = 3; i < words.length; i += 4) {
445.  double mass = 0;
446.  Acount = 0;
447.  Gcount = 0;
448.  Ccount = 0;
449.  Ucount = 0;
450.  String modificationmass = words[i];
451.  int decimal = modificationmass.indexOf(".");
452.  modificationmass = modificationmass.substring(0,decimal+6);
453.  Double modmass = Double.parseDouble(modificationmass);
454.  String temp = words[i - 2];
455.  String modname = words[i - 1];
456.  int space = temp.indexOf(' ');
457.  char nucleotide = temp.charAt(space + 1);
458.  mass = findNewModMass(rnaseq, nucleotide, modmass);
459.  String massString = Double.toString(mass);
460.  String printString = new String(massString + ": " + hsamir + " with " +
     modname);
461.  massList.add(printString);
462.  }
463.  return massList;
464.  }

465.  public static String findMatchingRNA(String s, double Range) {
466.  String inputString = null;
467.  StringBuffer strbuff = new StringBuffer();
468.  ArrayList<String> possibles = null;
469.  possibles = new ArrayList<String>();
470.  getModDBfromTextFile();
471.  getRNADBfromTextFile();
472.  double inputMass = 0;
473.  if (s.charAt(0) == '1'
474.  || s.charAt(0) == '2'
475.  || s.charAt(0) == '3'
476.  || s.charAt(0) == '4'
477.  || s.charAt(0) == '5'
478.  || s.charAt(0) == '6'
479.  || s.charAt(0) == '7'
480.  || s.charAt(0) == '8'
```

```
481.    || s.charAt(0) == '9'
482.    || s.charAt(0) == '0') {
483.    inputMass = Double.parseDouble(s);
484.    } else if (s.substring(0, 3).equals("hsa")) {
485.    for (int i = 0; i < microRNAs.length; i += 2) {
486.    if (s.equals(microRNAs[i])) {
487.    inputString = microRNAs[i + 1];
488.    break;
489.    }
490.    }
491.    inputMass = calcTotalMass(inputString);
492.   } else if (s.charAt(0) == 'A' || s.charAt(0) == 'C' || s.charAt(0) == 'G' || s.charAt(0)
       == 'U') {
493.    inputString = s;
494.    inputMass = calcTotalMass(inputString);
495.    }
496.    if (inputMass == 0) {
497.    System.err.println("Input Invalid! Please try again with valid input.");
498.    return "";
499.    }
500.    for (int i = 1; i < microRNAs.length; i += 2) {
501.    double DBmass = calcTotalMass(microRNAs[i]);
502.    if ((inputMass - Range) <= DBmass && DBmass <= (inputMass + Range)) {
503.    possibles.add(DBmass + ": " + microRNAs[i - 1]);
504.    }
505.    }
506.    ArrayList<String> oneModList = null;
507.    oneModList = new ArrayList<String>();
508.    for (int c = 0; c < microRNAs.length; c += 2) {
509.    oneModList.addAll(allModdedMasses(microRNAs[c]));
510.    }
511.    ArrayList<String> twoModList = null;
512.    twoModList = new ArrayList<String>();
513.    for (String st : oneModList) {
514.    int index = st.indexOf(':');
515.    String weightstring = st.substring(0, index);
516.    int decimal = weightstring.indexOf(".");
517.    weightstring = weightstring.substring(0,decimal+6);
518.    Double onemodWeight = Double.parseDouble(weightstring);
519.    if ((inputMass - Range) <= onemodWeight && onemodWeight <= (inputMass +
       Range)) {
520.    possibles.add(st);
521.    }
```

```
522.  // 2 Modifications
523.  twoModList.clear();
524.  int secIndex = st.indexOf("with");
525.  String miR_oneMod = st.substring(index);
526.  // NOTE: The weight of the one-modded miRNA is above in onemodWeight.
527.  // NOTE: Sequence is rna_seq
528.  // NOTE: Nucleotide to be modified is nucleo
529.  // NOTE: Mass of mod is mod_mass
530.  // NOTE: miR_oneMod = [Mass]: [hsa-mir] with [mod]
531.  String hsa = st.substring(index + 2, secIndex - 1);
532.  String rna_seq = null;
533.  for (int b = 0; b < microRNAs.length; b += 2) {
534.  if (hsa.equals(microRNAs[b])) {
535.  rna_seq = microRNAs[b + 1];
536.  break;
537.  }
538.  }
539.  String oneListModname = st.substring(secIndex + 5);
540.  int strt_snd_mod = 0;
541.  for (int v = 2; v < words.length; v += 4) {
542.  if (oneListModname.equals(words[v])) {
543.  strt_snd_mod = v;
544.  break;
545.  }
546.  }
547.  strt_snd_mod += 1;
548.  for (int g = strt_snd_mod; g < words.length; g += 4) {
549.  String modif_mass = words[g];
550.  double mod_mass = Double.parseDouble(modif_mass);
551.  String nomenc = words[g - 2];
552.  String mod_name = words[g - 1];
553.  char nucleo = nomenc.charAt(nomenc.indexOf(' ') + 1);
554.  double finalMass = apply_new_mod(onemodWeight, nucleo, mod_mass,
       rna_seq);
555.  String massString = Double.toString(finalMass);
556.  decimal = massString.indexOf(".");
557.  massString = massString.substring(0,decimal+6);
558.  twoModList.add(massString + miR_oneMod + " and " + mod_name);
559.  }
560.  for (String ts : twoModList) {
561.  int colon_index = ts.indexOf(':');
562.  String two_mass_string = ts.substring(0, colon_index);
563.  decimal = two_mass_string.indexOf(".");
564.  two_mass_string = two_mass_string.substring(0,decimal+6);
```

```
565.  Double two_mass_double = Double.parseDouble(two_mass_string);
566.  if ((inputMass - Range) <= two_mass_double && two_mass_double <=
      (inputMass + Range)) {
567.  possibles.add(ts);
568.  }
569.  }
570.  System.out.println(possibles.size());
571.  }
572.  System.out.println("there are: " + possibles.size() + " results");
573.  System.out.println("the range is +/- " + Range);
574.  System.out.println("the input mass was: " + inputMass);
575.  for (String o : possibles) {
576.  strbuff.append(o + "\n");
577.  }
578.  String matches = strbuff.toString();
579.  return matches;
580.  }

581.  // Method if the database-retrieved RNA is any number of spaces longer than
582.  // the input RNA. testRNA is the RNA given by the user to the system to find
583.  // similar RNAs.
584.  // DBRNA will be the various RNAs in the entire miR database.
585.  public static boolean findMatch(String testRNA, String DBRNA) {
586.  StringBuffer strbuff = new
      StringBuffer().append("00").append(DBRNA).append("00");
587.  String newDBRNA = strbuff.toString();
588.  int wrong;
589.  int diff = Math.abs(newDBRNA.length() - testRNA.length());
590.  boolean matchMore = false;
591.  for (int i = 0; i < diff + 1; i++) {
592.  wrong = 0;
593.  for (int j = 0; j < testRNA.length(); j++) {
594.  if (!(testRNA.charAt(j) == newDBRNA.charAt(i + j))) {
595.  wrong++;
596.  }
597.  if (wrong > 0) {
598.  break;
599.  }
600.  if ((j == testRNA.length() - 1) && (wrong <= 0)) {
601.  matchMore = true;
602.  System.out.println(matchMore);
603.  return matchMore;
604.  }
605.  }
```

```
606.    }
607.    System.out.println(matchMore);
608.    return matchMore;
609.    }

610.    // This method determines whether each of the 1733 elements in the database
611.    // match, and if they do,
612.    // it adds the miRNA sequence to an ArrayList of matches.
613.    public String getSimilar(String rna) {
614.    ArrayList<String> matchList = null;
615.    StringBuffer sb = new StringBuffer();
616.    matchList = new ArrayList<String>();
617.    Scanner sc = new Scanner(System.in);
618.    String[] datbase = getRNADBfromTextFile();
619.    for (int i = 1; i < datbase.length; i += 2) {
620.    if (datbase[i].length() > rna.length()) {
621.    boolean ifmatch = findMatch(rna, datbase[i]);
622.    if (ifmatch) {
623.    matchList.add(datbase[i - 1] + ", " + datbase[i]);
624.    }
625.    } else if (datbase[i].length() > rna.length() - 3
626.    && datbase[i].length() < rna.length()) {
627.    boolean ifmatch = findMatch(datbase[i], rna);
628.    if (ifmatch) {
629.    matchList.add(datbase[i - 1] + ", " + datbase[i]);
630.    }
631.    } else if (datbase[i].length() == rna.length()) {
632.    boolean ifmatch = findMatch(rna, datbase[i]);
633.    if (ifmatch) {
634.    matchList.add(datbase[i - 1] + ", " + datbase[i]);
635.    }
636.    }
637.    }
638.    for (String s : matchList) {
639.    sb.append(s + "\n");
640.    }
641.    String matches = sb.toString();
642.    System.out.println(sb.toString());
643.    return matches;
644.    }
645.    }
```