A LARGE MULTISCALE DETAILED MODELLING OF APTAMERS AS ANTICANCER THERAPEUTICS

A thesis submitted in the fulfilment of the requirements for the degree of

MASTERS OF SCIENCE

IN CHEMISTRY

at

RHODES UNIVERSITY, SOUTH AFRICA

Department of Chemistry

Faculty of science

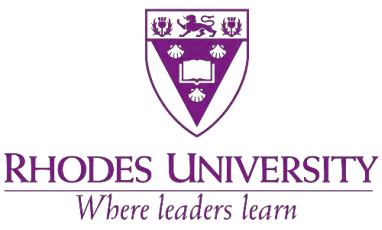
By

KABELO PHUTI MOKGOPA

SUPERVISOR: DR TENDAMUDZIMU TSHIWAWA

CO-SUPERVISOR: PROF. KEVIN LOBB

COMPUTATIONAL MECHANISTIC CHEMISTRY AND DRUG DISCOVERY RESEARCH UNIT



ABSTRACT

Cancer remains a leading cause of death worldwide, characterized by uncontrolled cell growth and spread. The challenge of effectively treating cancer has spurred interest in novel therapeutic strategies that target specific biological or biochemical mechanisms involved in cancer progression. Although many enzymes have been labelled as inducers of cancer development, microRNAs (miRNAs) are also emerging as significant contributors to cancer progression. This is because miRNAs play a crucial role in regulating gene expression, while cancer develops and grows due to genetic mutations, variations, and alterations. Among these miRNAs, miRNA-10b is notable for its involvement in promoting cancer cell proliferation, migration, and metastasis across various cancers, including breast cancer, glioblastoma, and esophageal squamous cell carcinoma. For this reason, we propose inhibiting miRNA-10b using RNA aptamers as a novel and promising approach for developing new anti-cancer therapeutics. RNA aptamers are short, non-coded, synthetic, and single-stranded nucleic acid molecules capable of binding to a wide range of targets, including metal ions, chemical compounds, proteins, cells, and microorganisms. They are used for a range of applications due to their well-known specificity and selectivity, starting from drug delivery to diagnostics. In this project we aimed to design and discover novel RNA aptamers that can effectively inhibit miRNA-10b using advanced computational methods. However, major challenges were encountered due to the lack of databases or tools available to design and predict secondary and tertiary structures of RNA aptamers at a large scale. Furthermore, no tools were available to perform high throughput virtual screening of these aptamers against macromolecular targets at a large scale. Prompted by that, we developed the T_SELEX program, which encompasses the various algorithms and tools dedicated to designing RNA aptamer sequences, predicting their secondary and tertiary structures, and, lastly, virtually screening aptamers. These algorithms and advanced tools are designed to handle the complexities of aptamer selection and virtual screening. By employing virtual screening methods, the aptamer discovery process was streamlined, offering a cost-effective and efficient alternative to traditional SELEX techniques. Prior to the main purpose application, the T_SELEX program was tested by designing aptamers for targeting HIV-1 protease, and a few applications were also done to assess its aptamer design approach. The study explored RNA aptamer sequences, revealing important insights into nucleotide composition, sequence patterns, and their role in aptamer efficacy and design. Analysis of secondary and tertiary structure predictions showed that Minimum Free Energy (MFE) values do not always correlate with structural compactness or complexity, with aptamers of similar MFE values exhibiting variations in attributes like loop size and guanine content. A novel Sequence Similarity Check (SSC) algorithm is introduced focused on internal sequence comparisons and secondary structures, revealing that aptamers with similar base compositions could have distinct folding states and stability. The Base Randomization Algorithm (BRA) generated RNA aptamer libraries was further benchmarked, highlighting a critical threshold for aptamer length and demonstrating Gaussian distribution in base compositions. Virtual screening of aptamers using the T_SELEX program against pre-miRNA-10b and their mature 5p and 3p arm, identified aptamers557 and 899 as effective binders for the 3p and 5p arms, respectively. Extensive quantum mechanical and molecular dynamics simulations confirmed the stability of the aptamer-RNA complexes. Due to the understanding of the flexibility of these RNA-RNA complexes, we further proposed the stability matrices method as a calculus-based method to evaluate the relative stability of the complexes without being biased during MD analysis. MM-GBSA calculations supported docking results, identifying aptamers like aptamers557, aptamer274 and aptamer734 as strong inhibitors of the 3p arm. Overall, this project has proposed novel approaches for aptamer *in silico* design and validation, particularly in targeting miRNA-10b for cancer therapy.

DEDICATION

This thesis is dedicated to my loving late mother,

GRACE MMAPHUTI MOKGOPA,

whom I lost while working on this project this year. I know she was proud of the effort I put into this work, and her unwavering support and belief in me continue to inspire my every step.

"The righteous perish, and no one takes it to heart;
the devout are taken away, and no one understands
that the righteous are taken to be spared from evil.
Those who walk uprightly enter into peace;
they find rest as they lie in death."
– Isaiah 57:1-2

ACKNOWLEDGEMENTS

I begin by giving all glory, honour, and praise to **God**, who has been the ultimate source of strength, wisdom, and guidance for choosing to work through me to bring this project to completion. I trust in His plan, and I give Him the praise for bringing me this far.

I would like to express my deepest gratitude to my late mother, **Grace Mmaphuti Mokgopa**, whose love, guidance, and encouragement were the foundation of my strength. Though she is no longer with me, her spirit continues to inspire and propel me forward.

My heartfelt thanks go to the **National Research Foundation (NRF)** and **Centre for High-Performance Computing (CHPC) for providing** funding and the state-of-the-art computing infrastructure that was crucial for this research.

I am incredibly thankful to my supervisor, **Dr. Tendamudzimu Tshiwawa**, for entrusting me with such a huge and challenging task. Your support, guidance, kindness, understanding, and patience have been a constant source of strength throughout this journey. I truly appreciate your belief in me and your dedication to helping me achieve this milestone. Your mentorship has been invaluable, and I could not have completed this work without your encouragement and expertise.

A special thank you goes to my co-supervisor, **Prof. Kevin Lobb**, whose contributions have greatly enriched this research. From teaching me how to write clean, efficient code to inspiring me to automate every task I encountered, your insights and approach to problem-solving have been instrumental. Your guidance not only helped me overcome technical challenges but also instilled in me a deeper understanding of research methodology and the importance of innovation.

I would also like to extend my heartfelt thanks to Dr Shina D Oloniiju for his constructive feedback and thoughtful suggestions. A special thank you to my friend, Mofeli Leoma, for support during the highs and lows of this journey. Another special word of thanks goes to my research group, the **Computational Mechanistic Chemistry and Drug Discovery** research group members. Special recognition and gratitude also go to my **family** for their emotional support and understanding, especially during this challenging journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

2D	Two Dimension
3D	Three Dimension
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
T_SELEX	Theoretical Automated Systematic Evolution of Ligands by Exponential Enrichment
MFE	Minimum Free Energy
SSC	Sequence Similarity Check
BRA	Base Randomization Algorithm
MP-NNTA	Matrix Based Python Nearest Neighbour Thermodynamics Algorithm
SMO-IBM	Simple Mutation Optimization Interactions Based Method
miRNA	MicroRNA
3p	3' arm of a microRNA
5p	5' arm of a microRNA
miR-10b-3p	MicroRNA-10b-3p (3' arm of miR-10b)
miR-10b-5p	MicroRNA-10b-5p (5' arm of miR-10b)
PCA	Principal Component Analysis
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
Rg	Radius of Gyration
MM-GBSA	Molecular Mechanics Generalized Born Surface Area
MD	Molecular Dynamics

HOMO-LUMO	Highest Occupied Molecular Orbital - Lowest Unoccupied Molecular Orbital
GC	Guanine-Cytosine (content)
PDB	Protein Data Bank
H-L Gap	HOMO-LUMO Gap (energy gap)
SD	Standard Deviation
VDW	Van der Waals (forces)
SWA	Stepwise Assembly
SWM	Stepwise Monte Carlo
ResNet	Residual Convolutional Network
CAPRI	Critical Assessment of Predicted Interactions
HDOCK	Hybrid DOCKing
NPT	Isothermal-Isobaric Ensemble (constant Number of particles, Pressure, and Temperature)
PBC	Periodic Boundary Conditions
REMD	Replica-Exchange Molecular Dynamics
SASA	Solvent Accessible Surface Area
SAV	Solvent Accessible Volume
SA	Simulated Annealing
PBE	Poisson-Boltzmann Equation
BAT	Bond/Angle/Torsion (coordinate models)
AMBER	Assisted Model Building with Energy Refinement
miR	Simplified standard code for miRNAs

Chapter 1

Introduction and Literature review

1.1 Overview

Chapter 1 provides a thorough exploration of cancer literature, including types of cancer, risk factors, and the complex biological mechanisms driving cancer development. It looks into various contributors to cancer development such as occupational exposures, smoking, pharmaceuticals, viruses, and natural or non-viral infectious agents. This chapter provides detailed information about cancer biology, covering multistage carcinogenesis, carcinogen activation, oncogenes, tumour suppressor genes, and critical processes like cell cycle regulation, telomere dynamics, apoptosis, and metastasis. A significant focus is on the role of microRNAs (miRNAs) in cancer, particularly their biogenesis, regulation, and the consequences of their dysregulation on tumour progression. This chapter emphasizes the therapeutic potential of targeting specific miRNAs, such as miR-10b, exploring ongoing clinical trials and the need for optimizing delivery methods to enhance therapeutic outcomes. In addition to its focus on cancer and miRNAs, this chapter includes a discussion on computational methods relevant to molecular modelling, where it covers RNA secondary structure prediction techniques and tertiary structure predictions. It also goes into macromolecular docking and molecular dynamics, explaining key concepts such as force fields, potential functions, and numerical integration. This chapter further explores the development of parameters, focusing on the AMBER force fields and their application to nucleic acids modelling.

1.2 Cancer

Cancer ranks second globally amongst diseases with the highest mortality rate, accounting for over 9.6 million deaths according world health organisation (WHO), as reported in 2018 [1]. The development of cancer is no longer a mystery due to availability of literature on impactful research progress being done on this subject. Although the word “cancer” is an overall term used, there are multiple forms this disease [2]. Cancer is caused by the uncontrollable cell division processes [3]. These processes involve single or a group of cells dividing uncontrollably which leads to formation of tumour. These tumours may be classified as benign (non-cancerous) or malignant (cancerous) [3,4]. Unlike malignant, benign soft tissue do not pose any thread as they do not spread to the other parts of the body. On the other hand, malignant tumours pose a huge threat as they metastasise (they subsequently spread to the neighbouring body parts) [5]. Cancer can develop from any type of organ or tissue including lungs, liver, breast, colon and prostate. It is important to highlight that normal cells do undergo cell division. However, what distinguishes normal cells from cancerous cells is that, normal cells divide controllably, they go through cell apoptosis (cell death) and they signal and respond to the neighbouring cells [6-8].

1.2.1 Types of cancer

The most common type of cancer is carcinomas which originate in epithelial tissues [9]. These tissues cover both the external and internal surfaces of the body, including organs and body cavities [9,10]. Carcinomas can manifest in various forms, such as squamous cell carcinoma, adenocarcinoma, transitional cell carcinoma, and basal cell carcinoma, depending on the type of epithelial cells affected [11]. Another type on cancer is sarcomas which arise from connective tissues, which provide structural support throughout the body [12]. They are relatively rare and can affect tissues like bone, cartilage, muscle, or blood vessels. Bone sarcomas and soft tissue sarcomas are the main subtypes of sarcomas [12]. Leukaemia is another type of cancer that originates in the bone marrow, where excessive production of abnormal white blood cells occurs [13]. These cancerous cells impair the normal functioning of the immune system. Leukaemia is a less common form of cancer but is the most prevalent type of cancer among children [14]. Lymphomas and myeloma are the types of cancer that affect the lymphatic system, a crucial component of the body's defence against infections [15,16]. Lymphomas start in lymph glands or lymphatic cells, while myeloma begins in plasma cells, a type of white blood cell responsible for producing antibodies. These cancers disrupt the body's ability to fight infections effectively [17]. Lastly, brain and spinal cord cancers,

collectively known as central nervous system (CNS) cancers, originate from the cells of the brain or spinal cord [18]. Gliomas, arising from glial cells that support nerve cells is the most common type. While some brain tumours are benign and grow slowly, others are malignant and tend to spread rapidly [19].

1.2.2 Risk factors and causes

Causes of cancer have been a hot topic for generations dating back to 1950 World Health Organization symposium [20]. It was posed that, understanding the causes of cancer may lead to proposition of safety measures of prevention. The symposium's findings suggested that environmental exposures rather than inherited genetic factors were primarily responsible for cancer development [20]. This revelation prompted the establishment of the International Agency for Research on Cancer (IARC) in 1965, tasked with investigating the causes of human cancers [21]. IARC focused on the on epidemiological evidence which includes cohort, case-control, and cross-sectional studies [22, 23]. Cohort studies follow a defined group over time to assess the relationship between exposure to an agent and their possible health outcomes by estimating risk differences [24]. In case-control studies, selection of individuals is based on their health status, then compare the odds of exposure between cases and controls to determine an odds ratio, providing information regarding the association between exposure and disease [26]. Cross-sectional studies collect exposure and health data simultaneously, to examine the associations between exposure and health effects without consideration of temporal sequence, utilizing measures like prevalence ratios or odds ratios [27]. With those methods there were still missing data then, later the IARC incorporated experimental evidence into its assessments. This led to the discoveries of how occupational, pharmaceutical, infectious, and natural agents are linked to cancer [28-30]. In this section, possible causes of cancer are reviewed.

1.2.2.1 Occupation

Historically, specific occupations were associated with increasing the risks of cancer. For instance, Bernardino Ramazzini in 1713 observed elevated breast cancer rates among nuns due to their celibate lifestyle [31]. Percivall Pott in 1775 linked scrotal cancer among chimney sweepers to heavy exposure to soot [32]. Additionally, reports from Richard von Volkmann and Joseph Bell in the late 19th century highlighted scrotal and bladder cancers was most common among coal tar distillers and shale oil workers, respectively [33, 34]. Furthermore, Ludwig Rhen documented bladder cancer cases among long-term dye workers in Germany [35]. The discovery of X-rays by Wilhelm Conrad Röntgen in 1895 led to the emergence of radiodermatitis among early radiologists, followed by reports of skin cancers among

radiologists [36]. Most of these discoveries were based on the epidemiological studies. In 1880, theories regarding the cause of cancer emerged, mostly pointing at the occupation. In order prove occupation induced cancer theories, several attempts to induce cancer in experimental animals were done, but with limited success. However, Ellermann and Bang's findings with leukaemia in chickens and Peyton Rous's discovery of sarcoma production in chickens using cell-free filtrate provided significant insights [37, 38]. Despite ongoing efforts to replicate these experiments, scepticism remained due to challenges in reproducibility. Katsusaburo Yamagiwa's ground breaking experiment in inducing skin cancer in rabbits with tar application marked a significant advancement in cancer research [39]. In the subsequent decades, additional evidence emerged linking synthetic agents to cancer, such as coal tar fractions identified by Ernest Kennaway, further demostrating the importance of experimental studies [40]. Additionally, occupational studies highlighted associations between certain exposures and cancer, as seen in the case of radium exposure among watch dial painters [41].

1.2.2.2 Smoking

Studies since the 1986 IARC Monograph establish a causal connection between cigarette smoking and various cancers, including nasal cavities, paranasal sinuses, nasopharynx, stomach, liver, kidney (renal cell carcinoma), uterine cervix, adenocarcinoma of the oesophagus, and myeloid leukaemia [42-44]. These add to the previously identified cancers associated with smoking, such as lung, oral cavity, pharynx, larynx, oesophagus, pancreas, urinary bladder, and renal pelvis [45]. Sophisticated statistical methods were developed to uncover additional cancer causes, exemplified by the pioneering work of Austin Bradford Hill and Richard Doll in linking cigarette smoking to lung cancer [46, 47]. Despite initial skepticism, their findings were eventually confirmed through numerous studies, marking a pivotal moment in understanding the link between smoking and cancer [48, 49].

1.2.2.3 Pharmaceutics

Beginning in the late 1960s, pharmaceuticals emerged as significant contributors to carcinogenesis, with various drugs implicated in elevated cancer risks [50]. High doses of analgesic mixtures containing phenacetin were linked to increased rates of carcinoma of the renal pelvis, while organ transplant recipients using azathioprine faced heightened risks of lymphomas [51, 52]. Similarly, women exposed to diethylstilboestrol during pregnancy saw their daughters develop adenocarcinoma later in life, and post-menopausal women on oestrogen replacement therapy were at higher risk of endometrial cancer [53-55]. Even therapeutic agents like chlornaphazine, melphalan, and cyclophosphamide, initially developed

for cancer treatment, were associated with secondary malignancies such as bladder carcinoma and acute nonlymphocytic leukaemia (ANLL) [55-59]. Recent studies investigating various medications by Friedman and co-workers have found nifedipine, nortriptyline, oxazepam, paroxetine, and piroxicam to potentially increase cancer risk, prompting further investigation into their long-term effects [60].

1.2.2.4 Viruses

Early experimental studies on the infectious causes of cancer trace back to classical times, where it was hypothesized that a single infectious agent could cause all types of cancer. Despite extensive research, no confirmed contagious cause was found. In the late 1800s, Pasteur and Koch demonstrated the contagious nature of many diseases, sparking more searches for microbial causes of cancer [61]. The first human virus, the yellow fever virus, was identified in 1900, leading to hypotheses about viral causes of cancer, notably after observations on chicken leukaemia and sarcoma [62]. Experimental studies from 1920 to 1950 saw Rous's work with the chicken virus, but controversies and skepticism hindered progress [63,64]. Although some viruses inducing cancer were identified in animals during this period, industrial agents like chimney soot and coal tar were more convincingly linked to human cancer, diverting attention from viral studies [65, 66]. Between 1950 and 1980, animal studies flourished, revealing endogenous viruses in mice and reports of retroviruses integrating into host DNA [66]. This led to the notion that similar agents might induce leukaemia in other species, fuelling further interest in viral causes of human cancer. However, extensive efforts to find human oncogenic retroviruses yielded few results, leading to scepticism similar to the early 1900s [67]. From 1980 to the present, significant advances occurred in establishing the role of viruses in human cancer. Improved laboratory methods and epidemiological analyses improved research efforts. The discovery of human T cell leukaemia virus (HTLV-1) and human immunodeficiency virus (HIV) highlighted the link between retroviruses and specific cancers [68]. Similarly, hepatitis B virus (HBV) and human papillomavirus (HPV) were implicated in liver and cervical cancers, respectively, leading to the development of vaccines [68]. Recent research has elucidated mechanisms by which these viruses induce cancer. For instance, viruses such as HPV have been found to cause cancer by integrating their genomes into host DNA, disrupting normal cellular functions [69]. Viral oncoproteins such as E6 and E7 of HPV and LMP1 of EBV can promote cell proliferation, inhibit apoptosis, and evade immune responses. Chronic inflammation and tissue injury induced by viruses like HCV and HBV contribute to hepatocarcinogenesis [70].

1.2.2.5 Natural factors and non-viral infectious agents

In the early 1900s, cancer rates in various occupations led to the belief that synthetic agents were the primary cause. However, research since the 1980s has established viruses as another significant cause of cancer, alongside natural factors and non-viral infectious agents [71]. This shift in understanding was marked by the identification of hormones like estrone, found in pregnant women's urine, as potential carcinogens [71]. Studies administering estrone to male mice resulting in mammary adenocarcinoma highlighted the complexity of naturally occurring carcinogens, challenging previous assumptions about cancer causation [72]. Occupational patterns of disease observed by Bernardino Ramazzini in the 18th century hinted at the role of natural factors in cancer [73]. Studies like those by Lane-Claypon and MacMahon et al. emphasized childbirth and breastfeeding's impact on breast cancer risk [74, 75]. Similarly, observations of lymphoma distribution in malaria-endemic regions by Dr. Burkitt suggested an environmental link to cancer [76]. These early findings laid the groundwork for understanding how factors like ultraviolet radiation, parasites like *Schistosoma haematobium* and *Opisthorchis viverrini*, as well as fungal toxins like aflatoxin, contribute to cancer development [77-80]. Paul Unna's work on skin diseases linked ultraviolet radiation to cancer, while Ferguson and others associated parasitic infections with bladder and bile duct cancers [77,78]. Burkitt's research in Africa provided insights into the connection between infectious diseases and cancer, particularly lymphomas [81]. Additionally, studies on fungal toxins like aflatoxin demonstrated their carcinogenic potential, especially in regions with high contamination levels. These findings highlight the diverse and global nature of natural carcinogens and their impact on public health [82-86].

1.2.3 Cancer development and mechanism (Biology)

Tumour development involves a complex and gradual process of malignant transformation caused by genetic alterations within cells. This progression is not sudden but occurs over several years through successive generations of cells, evolving from benign lesions to malignant tumours. Key players in this process include environmental agents such as carcinogens, radiation, and viruses, which initiate genetic changes, and genetic mutations affecting oncogenes and tumour suppressor genes, contributing to cellular instability and uncontrolled growth, as discussed in section 1.2 [4]. Carcinogenesis involves multiple stages: initiation, where irreversible genetic changes occur; promotion, leading to tumour progression, and finally, progression, which includes tumour growth, genetic instability, and angiogenesis [87]. Angiogenesis is the formation of new blood vessels that support tumour expansion and

metastasis [88]. This complex process of metastasis involves various steps, including the invasion of surrounding tissues, and dissemination *via* the blood or lymphatic systems. Key genes, including suppressor genes, oncogenes, and DNA repair genes, undergo mutations, leading to genetic instability and loss of differentiation [87]. The most essential steps in cancer development are multistep carcinogenesis and oncogene activation. **Figure 1.1** below shows the schematic representation of the process of cancer development *via* multistep carcinogenesis.

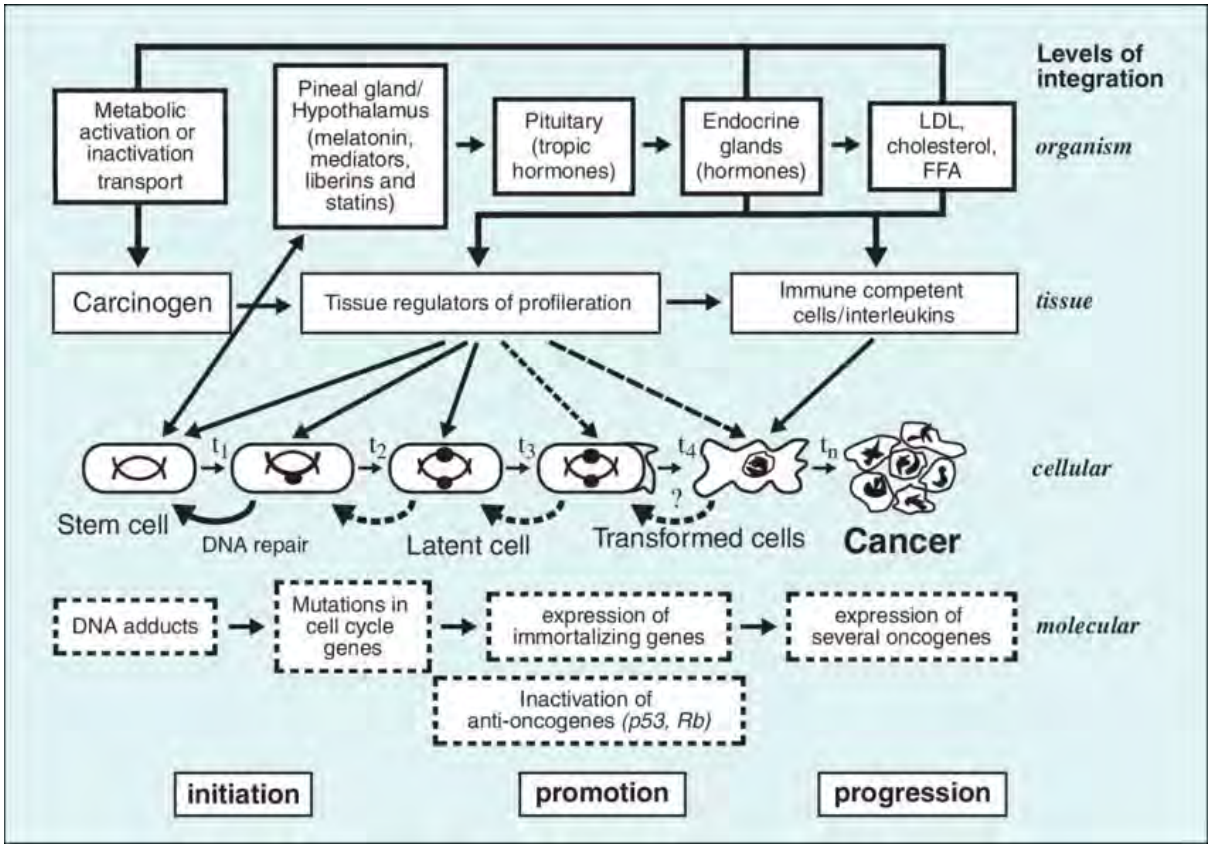


Figure 1.1: Essential framework of cancer development *via* multistage carcinogenesis. Adapted from Anisimov VN [89].

1.2.3.1 Carcinogenesis

Cancer develops through successive generations of cells, transitioning from benign lesions to malignant tumours [88]. This progression is driven by sequential genetic alterations that affect cellular proliferation, differentiation, and genetic integrity [88, 89]. Environmental agents such as chemical carcinogens, radiation, and viruses, as well as inherited genetic factors, play a crucial role in initiating cancer development [21]. Carcinogenesis involves two main phases: initiation, marked by irreversible changes in the growth potential of exposed cells and progression, where cells transition to autonomous cancerous growth and eventually metastasize

to distant sites [21, 90]. At the molecular level, significant breakthroughs have identified specific genetic changes associated with various stages of cancer [91]. Studies have linked histological stages of cancer with successive genetic alterations, including oncogene activation and the loss of tumour suppressor genes [91]. While most genetic changes critical to carcinogenesis are acquired during an individual's lifetime, some cancers may result from inherited mutations [92, 93]. These genetic alterations often arise from DNA damage caused by environmental factors or natural processes like reactive oxygen species production [94, 96]. Aging further complicates cancer development, as the risk of cancer increases with age growth. Usually, normal cells undergo senescence, which is a state where they stop to divide and acquire differentiated functions, potentially acting as an anti-cancer mechanism. However, carcinogenic chemicals or oncogenic viruses can induce cells to bypass senescence, leading to indefinite proliferation and the acquisition of transformed characteristics [97, 98]. In response to the complex process of multistage carcinogenesis, chemoprevention has been developed as a strategy to intervene before malignancy arises. This approach focuses on identifying and targeting precursor lesions, such as colorectal adenomas, which are validated as indicators of subsequent malignancy [99]. Removal of these adenomas has been shown to reduce the risk of colorectal cancer, highlighting their role in cancer prevention. Additionally, other phenotypic and genotypic markers, including histological and genetic changes, are considered in chemoprevention trials [100].

1.2.3.2 Carcinogen activation

Carcinogenesis involves the metabolic activation of carcinogens, which generates reactive intermediates that modify cellular macromolecules, including DNA and proteins [101]. This process highlights the multistage nature of carcinogenesis, where chemical carcinogens initiate and promote cancer through a series of sequential steps. The identification of initiators and promoters was a significant advancement in the 1940s, following earlier experiments showing that applying tar to rabbit ears induced skin tumours [67]. Most chemical carcinogens undergo metabolic activation, primarily by cytochromes P450, producing electrophilic metabolites that bind to DNA and form DNA adducts [102]. Techniques like immunoassays, 32P-postlabeling, and mass spectrometry have enabled sensitive detection of these DNA adducts, which serve as biomarkers for carcinogen exposure and cancer risk [103-105]. Genetic variations influencing enzyme activities involved in carcinogen metabolism affect adduct formation and subsequently, cancer risk. Studies linking genetic polymorphisms, adduct levels, and cancer risk highlight the importance of gene-environment interactions [106, 107]. Additionally,

mutations in specific genes, such as those in the HPRT gene, can indicate exposure to genotoxic agents, while oncogenes and tumour suppressor genes like RAS and p53 often show mutation patterns related to past carcinogen exposure [108, 109]. These "signature" mutations, resulting from specific DNA adducts, provide compelling evidence of the causal link between chemical exposure and cancer development [105].

1.2.3.3 DNA repair

As indicated in the preceding section, chemical and environmental exposures can lead to the formation of DNA adducts, which may lead to damage or mutation of the original DNA. Efficient DNA repair processes are vital for correcting these adducts and preventing their accumulation, which can otherwise contribute to genomic instability and malignant cell growth [110]. To prevent the accumulation of mutated DNA that could lead to cancer, cells utilize various DNA repair mechanisms. These mechanisms are essential for removing and replacing damaged or inappropriate bases. An example of DNA repair mechanism is an excision repair. There are two primary excision repair pathways: base excision repair (BER) and nucleotide excision repair (NER) [111, 112]. BER primarily addresses modifications caused by endogenous agents, whereas NER targets lesions induced by environmental mutagens, including UV light [111]. UV light is a notable environmental mutagen, and the NER pathway plays a critical role in protecting against UV-induced carcinogenesis. This significance is illustrated by conditions like xeroderma pigmentosum, where patients lack specific NER enzymes and exhibit extreme susceptibility to skin cancer following sunlight exposure [113, 114]. In NER, damaged DNA is first recognized, the damaged nucleotide is excised, and the gap is filled and ligated [112]. Conversely, in BER, a damaged base is recognized and removed by specific glycosylases, followed by similar gap filling and ligation steps. While these repair pathways effectively address most forms of DNA damage, complex damage such as double-strand breaks require alternative mechanisms.

Disorders like ataxia telangiectasia and Nijmegen breakage syndrome, characterized by extreme sensitivity to ionizing radiation, provide insights into the repair enzymes involved in handling such damage [115, 116]. Another crucial repair pathway, mismatch repair, corrects errors in DNA replication, such as base mispairing. Defects in mismatch repair genes are associated with hereditary nonpolyposis colorectal cancer, highlighting the importance of this pathway in maintaining genomic stability [117]. To promote DNA repair, various drugs have been developed or exploited with different mechanisms. For instance, Bleomycin (BLM) induces complex DNA damage, including strand breaks and 3’-phosphoglycolate (3’PG)

residues, which can be targeted by inhibiting tyrosyl-DNA phosphodiesterase 1 (Tdp1) to enhance therapy [118]. Other drugs include Quadruplex stabilizers which inhibit telomerase by stabilizing guanine quadruplexes (G-4s) at telomeres and gene promoters, while PARP inhibitors and DNA-PKcs inhibitors exploit synthetic lethality in cancer cells with defective DNA repair pathways [119].

1.2.3.4 Oncogenes

If the DNA repair process fails, it can lead to successful mutations in proto-oncogenes, resulting in their activation as oncogenes [120]. Proto-oncogenes are normal cellular genes involved in regulating cell growth and division in response to mitogenic signals. These genes encode various components of molecular cascades, including growth factors, receptors, signalling molecules, and transcription factors [120-123]. For example, the SRC gene discovered in the Rous sarcoma virus genome was the first identified oncogene [124]. Commonly activated oncogenes in human cancers include ERBB2, RAS, and MYC. ERBB2 activation often results from gene amplification, leading to overexpression of its receptor and constitutive activation of growth signals, making it a target for therapeutic antibodies and kinase inhibitors [125]. RAS genes which amplify signals from cell surface receptors have mutations causing continuous activation of downstream pathways [126, 129]. MYC, a transcription factor crucial for cell cycle entry is frequently activated in cancers through amplification or chromosomal translocations [127]. Additionally, BCL2 identified in B cell lymphomas regulates mitochondrial membrane permeability to inhibit apoptosis, thus promoting cell survival [130]. Several drugs have been developed to target oncogenes, including imatinib, which specifically inhibits the BCR-ABL fusion protein in chronic myeloid leukaemia (CML), and other agents that inhibit receptor activity such as small RNA molecules that target oncogene expression and drugs that block downstream signalling proteins [131].

1.2.3.5 Tumour Suppressor Genes

The discovery of tumour suppressor genes (TSGs) complemented the understanding of oncogenes by providing insights into the mechanisms underlying the loss of growth control in cancer cells. Unlike oncogenes, which promote cell proliferation when activated, TSGs act as brakes on cell division and prevent the development of cancer. The study of large DNA viruses and familial tumour syndromes played crucial roles in identifying TSGs [132]. For instance, the two-hit hypothesis proposed by Knudsen elucidated the inheritance pattern of retinoblastoma, highlighting the importance of biallelic inactivation of the RB1 (Retinoblastoma 1) gene in tumorigenesis [132]. This theory laid the groundwork for the

concept of recessive TSGs. Several TSGs have been identified through the study of familial cancer syndromes. These genes, when mutated or inactivated, predispose individuals to various cancers. For example, mutations in the APC (Adenomatous Polyposis Coli) gene are associated with familial adenomatous polyposis, leading to increased cell proliferation and the development of colon cancer [133]. Similarly, BRCA1 and BRCA2 mutations increase the risk of breast and ovarian cancers [134]. The discovery of these genes has revolutionized our understanding of hereditary cancer predisposition and provided opportunities for early detection and intervention. The p53 gene, often described as the "guardian of the genome," is one of the most studied TSGs. Its protein by-product plays a critical role in regulating cell cycle progression, DNA repair, apoptosis, and differentiation [134]. Mutations in the p53 gene are commonly found in various cancers and are associated with poor prognosis. Notably, unlike other TSGs, p53 is frequently mutated *via* point mutations within its DNA-binding domain, leading to loss of function [134]. Accumulation of mutant p53 protein in cancer cells serves as a diagnostic marker and a potential target for therapeutic interventions [134,135]. Another pivotal TSG is CDKN2A, which encodes two distinct proteins, p16INK4A and p14ARF, through alternative reading frames. These proteins regulate key cell cycle pathways by inhibiting cyclin-dependent kinases and modulating p53 stability, respectively [134]. Alterations in the CDKN2A locus, including loss of alleles, mutations, and hypermethylation, contribute to carcinogenesis across various cancer types.

1.2.3.6 Cell Cycle Regulation and Cancer

The cell cycle controls the ordered progression of molecular and cellular events, ensuring the proper replication and segregation of genetic material [136]. Divided into phases like M (mitosis) and S (synthesis), it encompasses critical checkpoints like (Growth 1) G1 and (Growth 2) G2 which ensures DNA fidelity and proper cellular division. Additionally, it governs cellular fate decisions post-mitosis, including quiescence, differentiation, or programmed cell death. The molecular architecture of the cell cycle has been greatly facilitated by studies in model organisms like *Xenopus laevis* and yeast [137,138]. Key discoveries include the identification of essential regulators like cyclin-dependent kinases (CDKs) and cyclins, which coordinate cell cycle progression through sequential activation and inactivation [139]. Checkpoint mechanisms, initially elucidated in yeast, serve as safeguards against aberrant cell division [137]. Failure to pass these checkpoints can lead to cell cycle arrest or aberrant division, as seen in diseases like cancer [140]. Notably, the transition from G2 to M phase in mammalian cells is tightly regulated, ensuring proper DNA replication and spindle

formation. Cell cycle dysregulation is a hallmark of cancer, with mutations in key regulators like p53, CDKIs, and cyclins leading to uncontrolled proliferation. Loss of tumours suppressors or activation of oncogenes disrupts the delicate balance of cell cycle control, contributing to tumorigenesis.

1.2.3.7 Telomeres and Cancer

Telomeres are the protective caps at the ends of eukaryotic chromosomes, play a crucial role in cellular aging and carcinogenesis. These repetitive DNA sequences, such as TTAGGG in vertebrates, undergo gradual shortening with each cell division [141]. As telomeres reach a critical length, cells exit the cell cycle, acting as a barrier against uncontrolled proliferation and carcinogenesis [141]. The enzyme telomerase, expressed by over 85% of cancers, counteracts this process by synthesizing new telomeric DNA, enabling cancer cells to bypass the proliferation barrier [142]. While telomerase assays hold promise for cancer diagnosis and prognosis, their routine clinical use remains under exploration. Studies suggest telomerase activity levels in urine sediments may aid in diagnosing urinary tract cancer, while predicting outcomes in neuroblastoma [143]. The discovery of hTERT (the catalytic subunit of human telomerase) in 1997 paved the way for potential therapeutic interventions targeting telomerase activity [144]. Inhibiting telomerase in tumours cells limits their proliferation and often leads to cell death, suggesting telomerase inhibitors could be a valuable cancer therapy, especially when integrated with other treatments. However, a challenge in telomerase research arises from cancers utilizing alternative mechanisms, such as Alternative Lengthening of Telomeres (ALT), to maintain telomere length [145]. This mechanism is independent of telomerase and it poses problems for the effectiveness of telomerase inhibitors in all cancer types. Developing drugs targeting telomerase activity requires careful consideration and integration with other therapeutic approaches, particularly for cancers employing ALT to sustain telomere length.

1.2.3.8 Apoptosis

Apoptosis or programmed cell death plays a critical role in various physiological processes, including cancer development and immune response. It is a tightly regulated mechanism that involves distinct phases: regulation, effector, and engulfing [153]. Dysregulation of apoptosis can contribute to disorders such as cancer. The identification of genes involved in apoptosis, particularly through studies in model organisms like *Caenorhabditis elegans*, has provided valuable insights into its mechanisms [154]. There are two major apoptotic signalling pathways that have been identified, 1) the extrinsic pathway, which is initiated by death receptors on the cell surface, and 2) the intrinsic pathway, which involves mitochondrial function and can be

triggered by various stimuli such as DNA damage [154]. Both pathways converge on the activation of caspases, a family of proteases that orchestrate the morphological and biochemical changes characteristic of apoptosis. The B-cell leukemia/lymphoma 2 (BCL2) family of genes plays a central role in regulating apoptosis. Members of this family can either suppress or induce apoptosis, depending on their function. For example, BCL-2 and BCL-xL suppress apoptosis, while Bax promotes apoptosis [155]. Dysregulation of BCL2 family genes has been implicated in cancer development, highlighting the importance of apoptosis in tumorigenesis. Apoptosis induction is a key strategy in cancer therapy. Various approaches are being explored to selectively induce apoptosis in tumour cells, including targeting specific signalling pathways involved in apoptosis regulation [156]. For instance, small molecules that interfere with BCL2 family proteins or activate death receptors are being investigated as potential anticancer agents [157].

1.2.3.9 Metastasis

Metastasis, the process by which cancer cells spread from the primary tumour to distant sites in the body, is a pivotal step in cancer progression [158]. It is often the cause of treatment failure and mortality in cancer patients. While benign tumours remain localized, malignant tumours have the ability to metastasize, making them far more dangerous [158]. The spread of cancer cells can occur through various routes, including dissemination *via* the blood or lymphatic system, spread in the cerebrospinal fluid, or trans coelomic passage [159 -163]. Despite advances in diagnostic imaging techniques like CT (Computed Tomography) scans and MRI (Magnetic Resonance Imaging), detecting micro metastases remains a challenge due to their limited sensitivity. The identification of genes associated with metastasis, particularly metastasis suppressor genes, has become a focus of research. Techniques such as laser capture microdissection and serial analysis of gene expression have enabled the comparison of gene expression profiles between invasive cancer cells and normal cells from the same patient [164]. These efforts have shed light on the complex genetic changes that drive metastasis, although many of the identified genes are also involved in tumour growth and development. The process of metastasis involves several sequential steps, including changes in cell-cell and cell-matrix adhesion, alterations in cell shape and motility, invasion of surrounding tissues, access to lymphatic or vascular channels, dissemination, survival in the circulation, extravasation, and colonization of secondary sites [163]. Numerous molecular and cellular events contribute to each of these steps, making metastasis a complex and multifaceted process. One critical aspect of metastasis is the role of angiogenesis, the formation of new blood vessels, which is essential

for tumour growth beyond a certain size. Genetic changes associated with malignant progression often induce an angiogenic phenotype by upregulating cytokines such as vascular endothelial growth factor (VEGF-A) [165]. Hypoxia, a characteristic feature of solid tumours, also stimulates the expression of VEGF-A [165]. Additionally, activation of signalling pathways like the epithelial growth factor receptor (EGFR) pathway can promote both vascular and lymphatic invasion, further facilitating tumour spread [166]. Changes in cell adhesion molecules, particularly E-cadherin, play a significant role in promoting metastasis. Loss or inactivation of E-cadherin, a tumour suppressor gene, is commonly observed in metastatic cancer cells. Other genes involved in cell adhesion, such as APC and DCC (deleted in colorectal cancer), are frequently mutated in cancer and contribute to the disruption of normal cell-cell and cell-matrix interactions [166]. Integrins, a family of cell adhesion receptors, also play a crucial role in mediating interactions between tumour cells and the extracellular matrix. In addition to cell adhesion molecules, other factors such as selectins, immunoglobulin superfamily members, and thrombospondin are implicated in cancer progression and metastasis [167]. These molecules facilitate interactions between tumour cells, endothelial cells, and other components of the microenvironment, promoting processes like tumour cell arrest, extravasation, and colonization of secondary sites.

1.2.4 MicroRNA in human Cancer

MicroRNAs (miRNAs) are small non-coding RNAs that play a crucial role in regulating various biological processes, including carcinogenesis. Discovered initially in *Caenorhabditis elegans*, miRNAs like lin-4 and let-7 were found to regulate gene expression, hinting at their significance in cancer development. Subsequent research revealed the abundance and conservation of miRNAs across species [168]. In humans, the involvement of miRNAs in cancer was first evidenced by studies on B-cell chronic lymphocytic leukaemia, where miR-15a and miR-16-1, located at chromosome 13q14, were identified as tumour suppressors by targeting the anti-apoptotic protein Bcl-2 [168,167]. Deletion or downregulation of these miRNAs was associated with leukaemia development. Further investigations using mouse models confirmed their tumour-suppressive role [169]. Subsequent profiling studies demonstrated widespread dysregulation of miRNA expression in various cancers, paving the way for their potential use as diagnostic, prognostic, and therapeutic biomarkers.

1.2.4.1 Mechanisms of miRNA biogenesis

MiRNA biogenesis initiates with the transcription of a gene into a large primary transcript (pre-miRNA), typically mediated by RNA polymerase II, although some pre-miRNAs are generated by RNA polymerase III [170]. The pre-miRNAs are then processed by a microprocessor complex consisting of RNA-binding protein DGCR8 and type III RNase Drosha, forming a ~ 85-nucleotide stem–loop structure called precursor miRNA (pre-miRNA) [170,171]. Subsequently, pre-miRNAs are transported from the nucleus to the cytoplasm by the Ran/GTP/Exportin 5 complex [171]. In the cytoplasm, pre-miRNAs undergo further processing by the RNase III enzyme Dicer to form a ~ 20–22-nucleotide miRNA/miRNA duplex. After unwinding, the mature miRNA strand is incorporated into the RNA-induced silencing complex (RISC), guiding RISC to target mRNA (massager RNA) [171]. In miRNA-mediated gene regulation, interactions between miRNAs and target mRNAs are often facilitated by the seed region, a 6 to 8-nucleotide fragment at the 5′-end of the miRNA [172]. However, recent studies using techniques like CLASH (cross-linking, ligation, and sequencing of hybrids) have identified non-canonical binding clusters independent of the seed region [172]. Regardless of interaction complexity, miRNAs can cause translational repression or target mRNA degradation upon binding to their targets, depending on complementarity. Beyond their traditional role in mRNA regulation, miRNAs have been found to function as ligands activating signalling pathways. For instance, tumour cell-secreted miR-21/miR-29a was discovered to bind directly to Toll-like receptor 7 or 8, inducing a premetastatic inflammatory response [173,174]. The miRNAs can also affect signalling pathways in immune cells, such as the nuclear factor κB pathway in natural killer cells, through direct interaction with Toll-like receptors as ligands [174].

1.2.4.2 Regulation of miRNA biogenesis

The biogenesis of miRNAs is tightly controlled at multiple levels, including transcription, processing by Drosha and Dicer, transportation, RISC binding, and miRNA decay [170]. Various proteins, such as DEAD-box RNA helicases, SMAD protein, KH-type splicing regulatory protein (KSRP), and methyltransferase-like 3, play roles in miRNA maturation [175-177]. KSRP, for example, acts as a component of both Drosha and Dicer complexes, regulating the biogenesis of a subset of miRNAs in mammalian cells [176]. Additionally, methyltransferase-like 3 regulates miRNA biogenesis by methylating pre-miRNAs, marking them for recognition and processing by DGCR8 to yield mature miRNAs [179].

1.2.4.3 Mechanisms of miRNA dysregulation in cancer

Abnormal expression of miRNAs in cancer often arises from genomic alterations such as copy number variations and changes in gene locations, such as amplification, deletion, or translocation events [180]. For example, the loss of the miR-15a/16-1 cluster gene on chromosome 13q14 is frequently observed in patients with B-cell chronic lymphocytic leukaemia, whereas amplification of the miR-17–92 cluster gene is seen in B-cell lymphomas and lung cancers [181]. These genomic changes significantly influence miRNA expression levels and contribute to the development of cancer. The miRNA expression is tightly regulated by transcription factors, and dysregulation of key transcription factors can lead to aberrant miRNA expression in cancer [182]. For instance, c-Myc, often upregulated in malignancies, activates the transcription of oncogenic miRNAs such as the miR-17–92 cluster while repressing tumour-suppressive miRNAs [183]. Similarly, the p53-miR-34 regulatory axis plays a crucial role in tumour suppression by modulating miRNAs involved in cell-cycle regulation and apoptosis [184]. Epigenetic modifications, including DNA methylation and histone alterations, also play a significant role in miRNA dysregulation in cancer. For instance, in acute myeloid leukaemia, miR-223 expression is silenced by AML1/ETO through CpG methylation [185]. Conversely, DNA demethylation and inhibition of histone deacetylases can activate the expression of miRNAs with tumour-suppressive functions. These epigenetic changes in miRNA genes contribute to tumorigenesis and may serve as biomarkers for cancer diagnosis and prognosis. Defects in the miRNA biogenesis machinery, such as mutations or abnormal expression of components like Drosha, Dicer, DGCR8, Argonaute proteins, and Exportin 5, are also implicated in abnormal miRNA expression in cancer [170]. Dysregulation of these proteins can impact miRNA processing and RNA-silencing mechanisms. Inactivation mutations in Exportin 5, for example, impair the export of pre-miRNAs from the nucleus to the cytoplasm, thereby disrupting miRNA processing and contributing to tumorigenesis [186]. These mechanisms collectively contribute to the dysregulation of miRNA expression in cancer, influencing critical aspects of tumour biology such as cell proliferation, apoptosis, and metastasis

1.2.4.4 Importance of Dysregulated miRNA expression in tumours

Abnormal expression of microRNAs (miRNAs) significantly impacts cancer initiation and progression by influencing critical processes such as evading growth suppressors and sustaining proliferative signalling [187]. MiRNAs can function as oncogenes or tumour suppressors by targeting and disrupting various cell proliferation pathways. For instance, the

miR-17–92 cluster regulates E2F proteins that are crucial for cell cycle progression, affecting the balance between promoting and suppressing cell growth [180]. Additionally, dysregulated miRNAs, such as miR-221/222, target cyclin-dependent kinase (Cdk) inhibitors like p27Kip1, thereby promoting uncontrolled cell-cycle progression [188-190]. MiRNAs also affect other signalling pathways, like miR-486, which, when downregulated in lung cancer, disrupts cell proliferation and migration by targeting the Insulin-like Growth Factor-1 (IGF -1) and phosphatidyl inositol 3-kinase (PI3K) signalling pathways [191]. In cancer, evasion of apoptosis is a key hallmark, and miRNAs play a pivotal role in modulating this process by targeting anti-apoptotic factors and influencing cell survival. Loss of p53 tumour suppressor function, which is a common event in tumours, is often associated with miRNA dysregulation [192]. MiRNAs such as miR-192, miR-194, and miR-215 are activated by p53 to stabilize the p53 protein and promote apoptosis [193]. Conversely, miR-122 reduces p53 activity and contributes to chemoresistance in hepatocellular carcinoma. Other miRNAs, like miR-15a and miR-16-1, target Bcl-2 to induce apoptosis, while miR-221/222 inhibits apoptosis by targeting the pro-apoptotic gene PUMA [194-197]. Additionally, miRNAs modulate apoptotic pathways by affecting components of the extrinsic apoptotic pathway, such as Fas ligand and death receptors [198].

Metastasis, a critical process in cancer progression, involves epithelial-mesenchymal transition (EMT), where miRNAs play essential roles in regulating cell motility and invasion. TGF-β-regulated miRNAs, such as miR-155 and miR-200, are crucial for EMT and metastatic progression [199]. MiR-155 promotes EMT by targeting RhoA GTPase, while miR-200 and miR-203 regulate EMT by inhibiting transcriptional repressors of E-cadherin [199]. Additionally, miRNAs like miR-10b, induced by TWIST, enhance migration and invasion in metastatic breast cancer cells [200]. Other miRNAs, such as miR-9 and miR-212, are also involved in metastasis; miR-9 promotes motility and invasion by downregulating E-cadherin, while miR-212 inhibits cell migration and metastasis in colorectal cancer by targeting MnSOD [201]. Angiogenesis, necessary for tumour growth and metastasis, is influenced by miRNAs through their regulation of hypoxia-induced factors like HIF [202]. MiRNAs such as miR-210 and miR-424 promote angiogenesis by targeting anti-angiogenic factors and stabilizing HIF1α, which enhances VEGF expression [203]. Conversely, miR-20b and miR-519c negatively regulate angiogenesis by targeting VEGF and/or HIF1α [204]. MiR-107 inhibits HIF1β expression, contributing to angiogenesis under hypoxic conditions. Moreover, exosomal miRNAs from cancer cells, like miR-135b, modulate the tumour microenvironment by

suppressing factors that inhibit HIF1, thereby fostering angiogenesis through the HIF-FIH signalling pathway [204-206].

1.2.5 Targeting miRNA-10b

MicroRNAs (miRNAs) have emerged as critical regulators of gene expression, influencing various biological processes, including cancer development and progression. Among these, miR-10b has gained significant attention due to its dual role as an oncogene and a tumour suppressor in different cancer types. Multiple studies have been done that delves into the multifaceted roles of miR-10b in cancer, highlighting its function and therapeutic potential. In oesophageal squamous cell carcinoma (ESCC), miR-10b-3p has been identified as a crucial player under hypoxic conditions [207]. Yang *et al*. in 2022, demonstrated that hypoxia induces the expression of miR-10b-3p, which in turn promotes cell proliferation, migration, invasion, and tumour growth [207]. Mechanistically, miR-10b-3p exerts its oncogenic effects by targeting the tumour suppressor gene TSGA10 [208]. These findings suggest that inhibiting miR-10b-3p could serve as a potential therapeutic approach for ESCC, particularly in hypoxic tumour microenvironments. Further supporting the role of miR-10b in ESCC, Hemmatzadeh *et al*. found that miR-10b is overexpressed in various types of cancer, including ESCC [209]. Tian *et al*., also reported that miR-10b expression correlates with cell motility and aggressiveness in various human ESCC cell lines [209]. They identified KLF4 (Krüppel-like factor 4), a known tumour suppressor gene, as a direct target of miR-10b. KLF4 suppresses oesophageal cancer cell migration and aggression by regulating p21 expression and mediating p53-dependent G1/S cell cycle arrest in response to DNA damage [210]. These studies collectively highlight the oncogenic function of miR-10b in ESCC, suggesting that targeting miR-10b could be a therapeutic strategy for this cancer type. miRNA sequencing data from patients with uterine leiomyosarcoma (ULMS) and myoma have revealed significant differential expression of miR-10b-5p. Wang *et al.* noted that miR-10b-5p is downregulated in ULMS compared to myoma, implicating its tumour-suppressive role [211]. Functional analyses by Liu *et al*. confirmed that overexpression of miR-10b-5p in LMS cells inhibits cell proliferation, migration, and invasion [212]. These findings highlight the potential of restoring miR-10b-5p expression as a therapeutic target for ULMS and warrant further investigation into its underlying mechanisms. In contrast to ULMS, miR-10b-5p is significantly upregulated in glioblastoma multiforme (GBM) tissues compared to normal brain tissues [213]. This upregulation is associated with increased tumour aggressiveness and poor patient survival.

Recent studies by Chen *et al*. in 2023 highlighted the role of the miR-10b-5p/TET2/PD-L1 axis in regulating GBM cell aggressiveness and immune evasion. miR-10b-5p promotes GBM progression by downregulating TET2, leading to increased PD-L1 expression and subsequent immune evasion [214]. These insights offer a potential therapeutic strategy targeting the miR-10b-5p/TET2/PD-L1 pathway in GBM. Therefore, inhibiting miR-10b-5p could be beneficial in treating GBM. Further emphasizing the potential of miR-10b as a therapeutic target, Yoo *et al* explored the use of miR-10b-directed nano therapy in brain metastases from breast cancer [215]. They developed an RNAi-targeted therapeutic, termed MN-anti-miR10b, which effectively inhibited miR-10b in primary tumours and metastases [216]. Their study demonstrated that MN-anti-miR10b accumulated in metastatic lesions following intravenous injection and inhibited metastatic progression in a model of breast cancer metastatic to the brain. These results suggest that inhibiting miR-10b could be a promising strategy for treating brain metastases from breast cancer.

Additionally, Skupin-Mrugalska in 2019 reported the use of liposome-based drug delivery systems targeting miR-10b in lung cancer [216]. They developed liposomes loaded with antagomir-10b and paclitaxel (PTX), which significantly hindered the migration of 4T1 cells, reduced lung metastases, and induced apoptosis and cell death. Zhang *et al.* further enhanced this delivery system by modifying the surface of the liposomes with an antimicrobial peptide [D]-H6L9 (D-Lip), facilitating effective endosomal escape and increased delivery in an acidic environment [217]. This approach demonstrated potential in reducing the required dose of antagomir-10b and mitigating side effects, highlighting a novel therapeutic strategy for lung cancer. The differential roles of miR-10b in various cancers highlight the complexity of miRNA-based therapies. In cancers where miR-10b acts as an oncogene, such as ESCC and GBM, strategies to inhibit miR-10b could be beneficial. Conversely, in cancers like ULMS, where miR-10b functions as a tumour suppressor, restoring its expression might be a viable therapeutic approach. In this context, the effects of miR-10b necessitate precise and targeted therapeutic strategies to maximize efficacy and minimize off-target effects. Several preclinical studies have explored the therapeutic potential of miR-10b modulation. Ma *et al.* investigated antagomirs targeting miR-10b and showed promise in reducing tumour growth and metastasis in animal models of breast cancer [218]. Additionally, clinical trials assessing the safety and efficacy of miR-10b inhibitors in cancer patients are underway. Current research is focused on optimizing delivery methods, minimizing side effects, and understanding the long-term impacts of miR-10b modulation. These studies are crucial for developing effective and targeted

therapies, as miR-10b has been implicated in various aspects of cancer progression, including metastasis and tumour growth. With ongoing advancements, targeting miR-10b remains a prominent topic in drug discovery, reflecting a broader trend in therapeutic research aimed at miRNAs. The ability to precisely modulate miRNA expression presents a promising avenue for innovative treatments across multiple diseases. Researchers are exploring various approaches to enhance the specificity and efficacy of miRNA-based therapies, including the development of novel delivery systems and the use of combination therapies to address potential challenges such as off-target effects and resistance [219]. The continued exploration of miRNAs as therapeutic targets is expected to contribute significantly to the development of next-generation treatments, offering new hope for patients with complex and difficult-to-treat conditions.

1.3 Aptamers

Aptamers, first introduced in 1990 by Ellington and Szostak, they are short synthetic single-stranded nucleic acid sequences capable of binding to a wide range of targets, including metal ions, chemical compounds, proteins, cells, and microorganisms [220]. Numerous aptamers have been developed to target amino acids, proteins, small metal ions, organic molecules, bacteria, viruses, cells, and animals [221-225]. These versatile aptamers find applications in analytical, bioanalytical, imaging, diagnostic, and therapeutics [126]. Factors influencing their binding affinity include hydrogen bonding, structural compatibility, aromatic ring stacking, electrostatic interactions, hydrophobic interactions, and van der Waals forces. Nucleic acid aptamers offer several advantages over protein antibodies across various parameters. Firstly, they are composed of nucleic acids rather than polymer peptides, rendering them less immunogenic, a crucial benefit as aptamers elicit non-humoral response compared to antibodies which can trigger immune reactions [127]. Moreover, aptamers are produced in vitro, contrasting with antibodies that requires in vivo production, thus reducing production costs significantly [128]. Additionally, aptamers exhibit stability, whereas antibodies are prone to degradation [128]. This enhanced stability ensures the reliability of aptamers for various applications. Furthermore, the specificity and affinity of both molecules are comparable, with aptamers demonstrating high specificity and affinity similar to antibodies. Aptamers also offer broader potential targets compared to antibodies, which are primarily limited to immunogenic molecules [129]. The generation time for aptamers is notably shorter, approximately ranging from 3 to 7 weeks, while antibodies typically require around 6 months for development. Lastly,

aptamers provide more flexibility for modification, offering convenience in tailored applications [128, 129].

Despite their advantages over antibodies, including those mentioned, aptamers face challenges in clinical use. Over 2000 aptamers have been generated in the past 30 years, but only one pegaptanib (Macugen; Pfizer/Eyetech) has received clinical approval [230]. However, this Vascular Endothelial Growth Factor (VEGF)-targeted aptamer still lags behind anti-VEGF monoclonal antibodies like ranibizumab (Lucentis; Genentech) and bevacizumab (Avastin; Genentech) in therapeutic efficacy [231, 232]. Two main barriers hinder aptamer development: time-consuming SELEX processes and uncertainty about in vivo functionality. The selection of aptamers is developed through an iterative procedure known as Systematic Evolution of Ligands by Exponential enrichment (SELEX). While the original SELEX process is time-consuming, various adaptations and integrations of advancements in material sciences and analytical techniques have led to SELEX variations aimed at reducing processing time, generating aptamers with novel designs and functions, and increasing process throughput. Today, with the integration of progress in material sciences and analytical techniques, some groups have significantly reduced the time for aptamer development from months required for conventional SELEX to several hours [233]. These advancements have led to the development of various aptamers, such as hydrophobic SOMAmers, thioaptamers, and X-Aptamers, with applications ranging from basic research to clinical trials for multiple oncology indications [234-236]. Despite significant progress, the number of aptamers showing strong binding capacity and specificity remains limited [236]. Improving the SELEX procedure and developing aptamers with higher binding capacity and specificity are ongoing challenges in both basic and applied studies.

1.3.1 SELEX

SELEX (Systematic Evolution of Ligands by Exponential Enrichment) has evolved significantly since its inception in 1990, with several modified methods developed to enhance its efficiency and effectiveness. The traditional SELEX process involves three main steps: selection, partitioning, and amplification as illustrated in **Figure 1.2** [231]. It begins with synthesizing a diverse library of oligonucleotides, which are then exposed to target molecules. Bound sequences are separated from unbound ones and amplified to create a new sub-pool for further selection rounds. Despite its success, conventional SELEX is time-consuming, often requiring weeks to months to identify specific aptamers and suffering from low hit rates [321]. To address these issues, various modified SELEX techniques have been introduced to

streamline the process and improve outcomes [240-248]. **Figure 1.2** illustrate the normal workflow procedure of the SELEX technique.

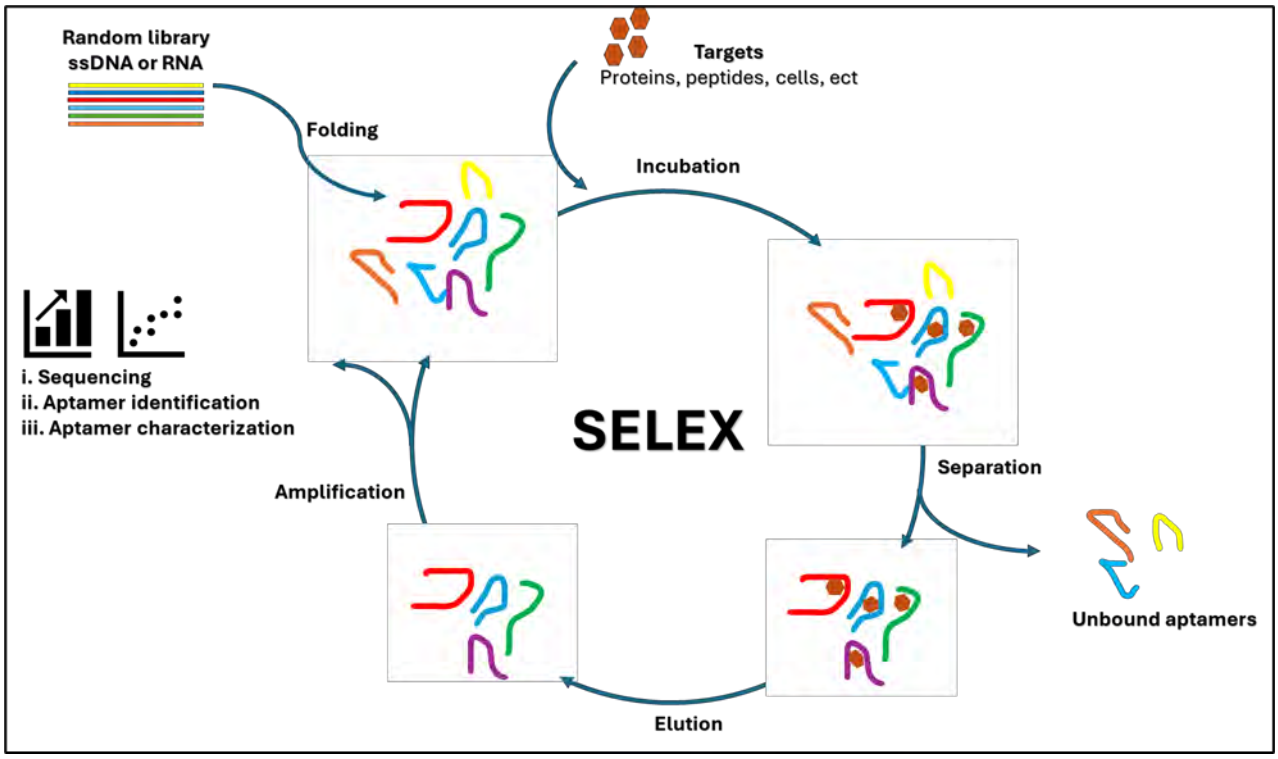


Figure 1.2: Illustration of the general SELEX procedure (adapted from [231]).

Negative SELEX, introduced by Ellington and Szostak in 1992, was aimed to improve aptamer specificity by eliminating non-specific bindings through negative selection with purification support agarose [237]. This method enhances the affinity and effectiveness of aptamers compared to those obtained without negative selection. Counter SELEX follows a similar goal but employs structurally-similar target molecules during selection cycles to further discriminate non-specific oligonucleotides. This additional step helps in selecting more specific aptamers by reducing cross-reactivity with similar molecules [238]. Capillary Electrophoresis SELEX (CE-SELEX), was also developed in 2004, improves upon traditional SELEX by using capillary electrophoresis to separate target-bound sequences, thus reducing the number of selection rounds needed and enhancing aptamer quality [239]. This technique significantly shortens the selection time and increases efficiency. Microfluidic SELEX, which integrates traditional SELEX with microfluidic systems, also was aimed to enhance efficiency [240]. There are other techniques like M-SELEX and sol-gel SELEX which offer rapid and automated aptamer selection while addressing challenges related to aptamer purity, recovery,

and protein stability [241,242]. More interestingly, the Cell SELEX was also designed, which targets the whole live cells, offering advantages over in vitro SELEX by selecting aptamers that bind to native conformational targets on cell surfaces, improving their applicability for diagnostic and therapeutic purposes [243]. Modified methods such as hybrid SELEX and TECS SELEX further optimize aptamer screening success rates [244]. Lastly *in-vivo* SELEX, on the other hand, generates aptamers within animal models under physiological conditions, facilitating the identification of aptamers with functional activity in real biological contexts, such as tumor targeting or blood-brain barrier penetration [245]. Aptamers have shown great potential as therapeutics across various domains. In eye disorders, aptamers targeting VEGF (Vascular endothelial growth factor), like pegaptanib, have demonstrated efficacy in treating conditions such as age-related macular degeneration (AMD) [246]. For thrombosis and vascular diseases, aptamers targeting coagulation factors and platelet aggregation, such as Ch-9.3t and NU172, have shown promise in preclinical and clinical trials [247, 248].

1.3.2 Cancer: aptamer-based therapies

Aptamers offers unparalleled specificity, and they are emerging as potent tools in the arsenal of targeted cancer therapies. Among these, nucleolin-targeting aptamer AS1411 represents a pioneering example, demonstrating safety and efficacy in clinical trials across various tumor types [249]. AS1411's capacity for specific cellular internalization has spurred its application as a versatile delivery agent for diverse payloads, from nanoparticles for imaging and therapy to chemotherapeutic drugs and splice-switching oligonucleotides [249]. Another notable target is prostate-specific membrane antigen (PSMA), exploited for its specificity in prostate cancer [250]. PSMA aptamers have shown promise in inhibiting tumour growth and metastasis in preclinical models. Furthermore, PSMA aptamers have been harnessed for the delivery of therapeutic cargoes such as small interfering RNAs (siRNAs) and microRNAs, highlighting their potential for enhancing the efficacy of conventional chemotherapy [250]. Human epidermal growth factor receptor 2 (HER-2) which is a well-established target in breast cancer, has also been subject to aptamer-based interventions. Aptamers against HER-2 have exhibited inhibitory effects on tumour growth and have been engineered to deliver cytotoxic payloads selectively to HER-2-expressing cancer cells, offering a novel approach to targeted therapy [251]. AXL (tyrosine kinase receptor), which is less recognized, has emerged as a promising target in solid tumours. Aptamers against AXL have demonstrated efficacy in inhibiting tumour growth and have been conjugated with therapeutic agents to enhance their specificity and efficacy [252]. Mucin 1 (MUC1), carcinoembryonic antigen (CEA), and protein tyrosine

kinase 7 (PTK7) are additional tumour-associated targets exploited for aptamer-based interventions [253-255]. Aptamers against these targets have shown potential in imaging, drug delivery, and inhibition of tumour growth in preclinical models, offering versatile platforms for precise cancer therapy. Immunotherapeutic applications of aptamers have also gained attraction, with aptamers targeting immune checkpoint molecules such as cytotoxic T cell antigen-4 (CTLA-4) showing promise in enhancing antitumor immune responses [256]. Additionally, aptamers targeting T cell receptors OX40 and 4–1BB have demonstrated agonistic activity, stimulated T cell activation and promoting tumour rejection in preclinical models [257]. Moreover, aptamers targeting angiopoietin-2 (Ang2) have shown efficacy in inhibiting tumour angiogenesis, complementing the existing armamentarium of anti-angiogenic agents in cancer therapy [258].

1.4 Theory and computational methods

Although experimental SELEX approaches have significantly contributed to the discovery of aptamers and their binding properties, the cost and time associated with these methods slow down advancement in the aptamer development. For this reason, molecular modelling offers an alternative faster approach, while reducing cost and time. Molecular modelling is a vital tool in modern chemistry, leveraging advancements in high-performance computing and software to complement traditional laboratory experiments. This computational approach enables *in silico* simulations to validate experimental findings, predict new materials, and explore properties of known substances [258]. Accurate simulations rely on effective representation of molecular systems and the use of specialized tools. Key areas within molecular modelling include areas such as cheminformatics and bioinformatics. Cheminformatics organizes and analyses vast chemical datasets to extract meaningful insights, which employs algorithms to model molecular structures and interactions [259]. Bioinformatics integrates computational methods with biological data to enhance understanding of biomolecules. This study focuses on modelling of biomolecules, specifically ribonucleic acids (RNA).

1.4.1 RNA representations (A, U, G, and C)

RNA, or ribonucleic acid, is a vital molecule found in all living cells, playing crucial roles in various biological processes such as protein synthesis, gene regulation, and cell signalling [260]. Structurally, RNA is composed of nucleotides, which are the building blocks of the molecule. Each nucleotide consists of three components: a ribose sugar molecule, a phosphate group, and one of four nitrogenous bases [adenine (A), uracil (U), guanine (G), or cytosine (C)]. The ribose sugar and phosphate group form the backbone of the RNA strand, while the

nitrogenous bases extend from this backbone, providing the sequence information essential for RNA's function.

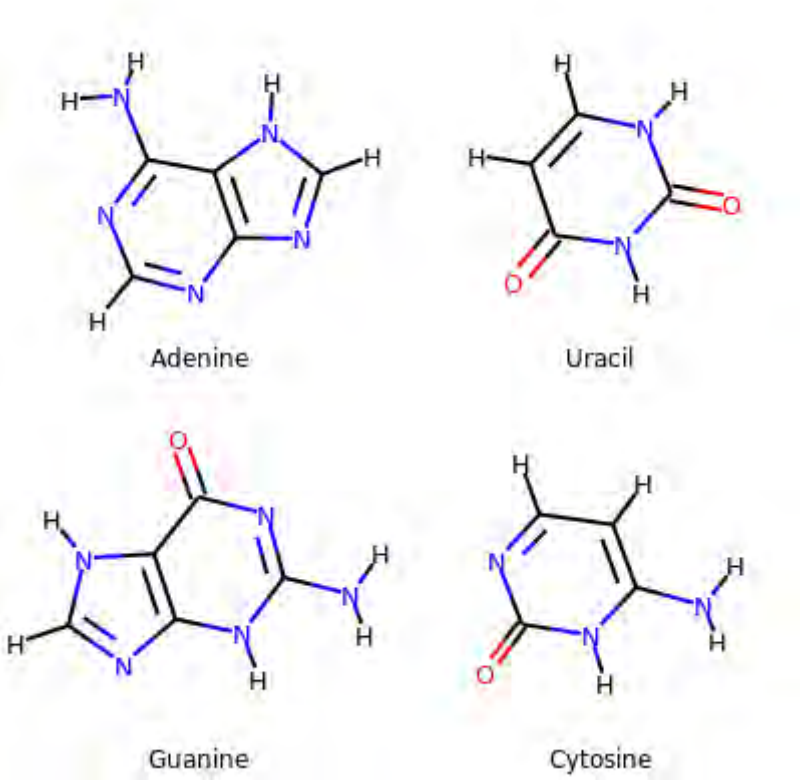


Figure 1.3: Chemical Structures of RNA Nucleobases: Adenine, Uracil, Guanine, and Cytosine Generated Using RDKit.

Adenine, uracil, guanine, and cytosine, as RNA nucleobases, share similarities such as heterocyclic ring structures containing functional groups crucial for hydrogen bonding interactions and genetic information storage [261]. However, they differ in their chemical compositions, with adenine and guanine being purines, while uracil and cytosine are pyrimidines. This distinction impacts their base pairing specificity, with adenine pairing with uracil in RNA and guanine with cytosine.

1.4.2 RNA Structure: Primary to Tertiary

RNA molecules, synthesized as single strands of ribose nucleic acids, exhibit complex structures beyond just nucleotide sequences. The process of RNA structure formation can be delineated into two successive stages: Firstly, the primary structure, comprising the sequence of bases, transitions into a pattern of complementary base pairings known as the secondary structure. Secondly, the secondary structure undergoes distortion to adopt a three-dimensional spatial arrangement, termed the tertiary structure. Predicting RNA secondary structures poses

a challenge due to the high number of degrees of freedom within the RNA chain, surpassing number of degrees of freedom available for proteins [263]. However, converting the secondary structure of RNA from a coarse-grained approach to the tertiary structure is supported by several factors [263]. These factors include conventional base pairing and stackings which covers a significant portion of the folding free energy, while the secondary structure provide a scaffold of distance constraints guiding tertiary structure formation. Moreover, unlike proteins, the RNA secondary structure is well-defined, assigning all bases to specific secondary structure elements [263]. Its conservation across evolution highlight its ability in interpreting RNA function and reactivity. RNA secondary structure arises from the aggregation of planar complexes or base pairs formed by purine (Adenine and Guanine) and pyrimidine (Cytosine and Uracil) bases [263]. Complementary bases (G-C, A-U) engage in strong hydrogen bonds, with weaker interactions possible between G and U, referred to as "wobble" base pairs [264]. The tertiary structure, depicted as a three-dimensional structure is characterized by hydrogen bonding or stacking interactions between structural elements.

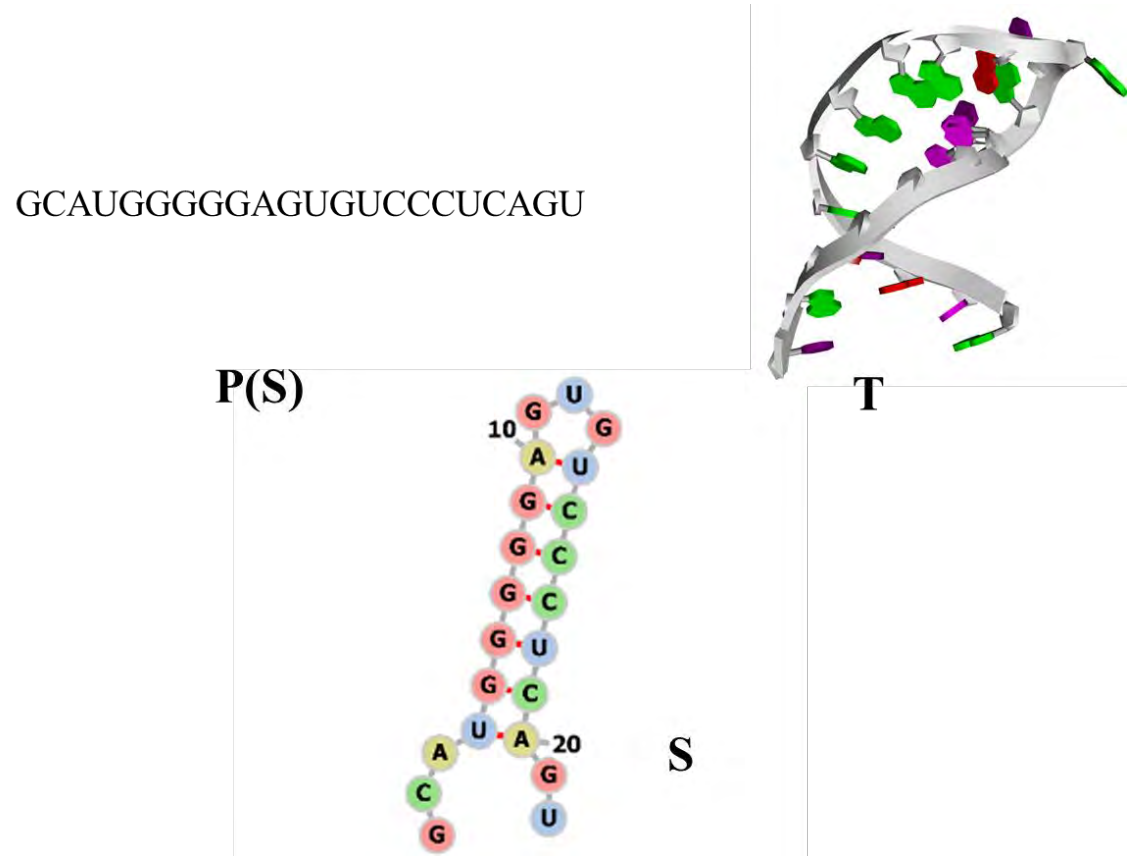


Figure 1.4: An example representation of RNA sequence (P(S)), secondary structure (S) and tertiary structure (T).

Figure 1.4 illustrates a hypothetical RNA sequence (P(S)) alongside its secondary structure (S) and tertiary conformation (T). The RNA sequence, randomly selected from our dataset, exemplifies the diversity inherent in RNA molecules. In the primary sequence (P(S)), represented by a linear string of nucleotides (A, U, G, and C), each letter denotes a specific nucleotide base. The secondary structure (S) depicts the folding pattern of the RNA molecule, showcasing base pairing interactions between complementary nucleotides, forming stem-loop structures and other motifs. This structure is crucial for RNA function and stability. The tertiary structure (T) demonstrates the three-dimensional arrangement of the RNA molecule, highlighting its complex folding and spatial organization.

1.4.2.1 RNA Secondary Structure: Definition and Representation

Before delving into the theory of RNA folding and secondary structures, it is essential to acknowledge the remarkable contributions of the Theoretical Biochemistry Group (University of Vienna) led by the esteemed Ivo Hofacker, who has made significant advancements in RNA research (https://www.tbi.univie.ac.at/). Their pioneering work and innovative tools, such as RNAfold [265], have been instrumental in unravelling the complexities of RNA folding, providing invaluable insights into the prediction of secondary and tertiary structures. Formally, an RNA secondary structure denoted S, encompasses all base pairs (i, j), where $i<j$, subjected to specific conditions [265]. Firstly, each nucleotide can participate in at most one base pair, ensuring $i=k$ if and only if $j=l$. Secondly, knots or pseudoknots are prohibited, enforcing conditions on the arrangement of base pairs to maintain a planar graph representation [265]. The recursive relationship for enumerating the number of secondary structures compatible with a specific RNA sequence can be described in equation 1.1.

$$S_{l,n+1} = S_{l,n} + \sum_{k=l}^{n} S_{l,k-1} \cdot S_{k+1,n} \cdot \Pi(\sigma_k, \sigma_{n+1}) \tag{1.1}$$

$S_{p,q}$ represents the number of structures compatible with the substring $[\sigma_p...\sigma_q]$ of the RNA sequence. $S_{l,n+1}$ represents the number of structures compatible with the entire RNA sequence [265]. The equation recursively calculates the number of compatible structures by considering all possible partitions of the RNA sequence summing them into substrings.

While tertiary interactions like base triplets and G-quartets defy the single pairing condition, pseudoknots, formed by base pairing between hairpin loops and single-stranded stretches,

challenge the planarity constraint [266]. Although important in natural RNAs, pseudoknots are often treated as tertiary interactions for simplicity. The compatibility of a secondary structure with a given RNA sequence relies on specific base pairing rules, typically involving Watson-Crick (AU and GC) and GU pairs [267]. Efficient computation of compatible secondary structures often involves dynamic programming algorithms, constrained by knot-free structures and limited to permissible base pairings. The resulting secondary structures reveal helical regions connected by single-stranded regions, forming various motifs like hairpins, bulges, and loops, showcasing remarkable complexity similar to protein structures. Secondary structures can be represented as strings using symbols such as open brackets, close brackets and full stops [266]. These symbols correspond to nucleotides that are paired with a partner toward the 3' end, toward the 5' end, or remain unpaired. When parentheses match, they indicate base pairs. For example, a short hairpin structure with a 4-loop and a helix of length 3 would be written as (((…))) [266].

1.4.2.2 RNA Secondary Structure: Prediction

1.4.2.2(a) Dynamic programming

In the field of RNA folding, Dynamic Programming stands as a fundamental technique, systematically breaking down complex folding problems into smaller, manageable subproblems. This method, extensively utilized in the ViennaRNA Package [268,269] and associated programs, deconstructs an RNA secondary structure into sequential stretches of sequence. Each stretch, denoted as X, is built from a shorter stretch X_1 and a base i, where i can either remain unpaired or form a base pair within the sequence. The approach hinges on the assumption of strictly additive energies and not context-sensitive interactions, an approach essential for algorithms like the "Nussinov" method, which maximizes base pairing to optimize RNA structure [270].

$$N_{i,j} = \max\left(N_{i-1,j}, \max_{i<k\leq j}\left\{N_{i+1,k-1} + N_{k+1,j} + \delta(i,k)\right\}\right) \tag{1.2}$$

The "Nussinov" algorithm prioritizes maximizing the number of base pairs ($N_{i,j}$) or hydrogen bonds in an RNA molecule [270]. This is achieved through a scoring system where different base pair types receive different scores. The recursive equations (1.3 to 1.6) for the minimum free energy in this model entail determining the maximum number of base pairs attainable, considering both intra- and inter-molecular interactions. While this method is efficient in

theory, its computation complexity grows cubically with sequence length, posing computational challenges for larger RNA molecules [266].

$$F_{i,j} = \min\left\{F_{i+1,j}, \min_{i<k\leq j}\left[B_{i,k} + F_{k+1,j}\right]\right\} \quad (1.3)$$

$$B_{i,j} = \min\left\{H(i,j), \min_{i<p<q<j}\left[I(ij,pq)B_{p,q}\right], \min_{i<k<j}\left[\alpha + \beta + M_{i,k} + M1_{k+1,j-1}\right]\right\} \quad (1.4)$$

$$M_{i,j} = \min\left\{M_{i+1,j} + \gamma, \min_{k}\left[B_{i,k} + M_{k+1,j} + \beta(i,j)\right], M1_{i,j}\right\} \quad (1.5)$$

$$M1_{i,j} = \min\left\{M1_{i,j-1} + \gamma, B_{i,j} + \beta(i,j)\right\} \quad (1.6)$$

Determining the minimum free energy of RNA structures. $F_{i,j}$ represents the minimum free energy on a stretch between positions i and j. It evaluates the energy of different segments of the RNA sequence and aids in determining the overall folding stability. $B_{i,j}$ corresponds to the minimum free energy subject to the constraint that positions i and j form a base pair (i,j) [266]. This term plays a pivotal role in assessing the stability of base pairing within the RNA structure.

Additionally, $M_{i,j}$ denotes a multi-loop contribution, capturing the energetic effects when there is at least one stack somewhere between positions i and j. It accounts for the energy contributions arising from complex structural arrangements within the RNA molecule. $M1_{i,j}$ represents a specific type of multi-loop contribution, indicating a scenario where there is exactly one stack, and base i is paired [266]. These contributions are essential for accurately modelling the energy landscape of RNA structures and predicting their stability.

1.4.2.2(b) Dangling ends contributions

Although the loops are fundamental for decomposing secondary structures, their predictive accuracy is limited. In the early 1970s, it was recognized that unpaired bases adjacent to base pairs contribute to stabilizing secondary structures [271]. Recent research attributes this effect to the exclusion of water from the hydrogen bonds of closing pairs in a stack or to the stacking of single bases onto a base pair, involving interactions of the π-electron-systems of the bases [271]. Incorporating these "Dangling Bases" and their measured energy contributions enhances the predictive power of loop-based energy models. The energy contributions vary depending on the type of base dangling and the base pair. Notably, stacking onto the 3' end of a base pair

tends to be more stabilizing in RNA molecules compared to stacking onto the 5' end [266]. This directional preference is less pronounced in DNA molecules [272]. While dangling energies are typically considered only for bases directly adjacent to a base pair, recent findings suggest a stabilizing effect for stretches of up to four bases. This effect can result in a significant difference in energy between structures with one and four dangling bases. However, incorporating this long-range effect into folding algorithms poses challenges due to its complexity and the need for computational efficiency [273].

1.4.2.2(c) Accuracy of dynamic programming

The accuracy of predicting RNA secondary structures through physics-based dynamic programming is notably influenced by the size of the molecules. Larger molecules tend to yield less accurate predictions due to several factors. Firstly, the model relies on various assumptions, which may become less reliable for longer sequences [268]. Secondly, inaccuracies in the energy parameters used in the model can end up affecting the overall predictive performance. Additionally, the presence of pseudo-knots, which become more likely as sequence length increases, further complicates the prediction process. Empirical data from Mathews *et al*. demonstrated the performance of advanced prediction methods across various RNA molecules [274]. Shorter RNAs, such as tRNA and Group II introns, typically demonstrate higher prediction accuracy, with correct base pair percentages ranging around 80%. In contrast, longer RNAs like 16S and 23S RNA exhibit lower accuracy rates, with correct base pair percentages around 50-60% [274]. The challenging prediction of RNAse P, attributed to the high occurrence of pseudo knotted base pairs, highlight the complexity of longer RNA structures. However, other studies, such as that by Doshi *et al*. have employed stricter criteria for evaluating prediction accuracy. They found that while shorter RNAs still fare relatively well, longer RNAs like 16S and 23S rRNAs exhibit lower correct prediction percentages, around 40-45% [275]. Importantly, the study highlights that the accuracy of predictions tends to decrease as the distance between base pairs, known as the “contact distance” increases.

1.4.2.2(d) Partition function

In physiological conditions, RNA molecules exhibit dynamic behaviour rather than being constrained to a rigid secondary structure. Due to the comparable energies involved in base pair formation and thermal fluctuations, base pairs continuously form and break. While tertiary structure interactions and interactions with other molecules can stabilize certain secondary

structures, RNA molecules generally adopt multiple possible conformations. To explore the variability in RNA secondary structures, various computational approaches have been developed. One approach involves computing a range of suboptimal structures in addition to the minimum free energy (MFE) structure [275]. For instance, Manfred Zuker's algorithm identifies the best secondary structure containing each base pair, while Stefan Wuchty's algorithm predicts all structures within a specified energy range of the MFE [276,277]. Another strategy, inspired by McCaskill's work, involves computing the partition function of RNA secondary structures, providing insights into the ensemble of possible conformations [278].

An ensemble of secondary structures *S(x)* on a sequence x is defined as the collection of all possible secondary structures that sequence *x* can assume, in accordance with the principles governing the formation of secondary structures. Each structure *s* in this ensemble is characterized by an energy value *E(s,x)* which corresponds to the stability of structure s when adopted by sequence *x*. In order to calculate the likelihood of observing sequence *x* in structure *s* at a specific moment, the Boltzmann coefficient is employed [278].

$$p(s|x) \propto e^{-\frac{E(s,x)}{kT}} \qquad (1.7)$$

Here, k represents Boltzmann's constant, and T denotes the absolute temperature. Since the probability $p(s|x)$ represents the likelihood of sequence x adopting structure s, following Boltzmann statistics [272]. Then every sequence x must adopt some structure, so the sum of probabilities over all structures in the ensemble S is equal to 1, as shown in Equation (1.8).

$$p(s|x) = \frac{e^{-\frac{E(s,x)}{RT}}}{\sum_{t \in S} e^{-\frac{E(t,x)}{RT}}} \qquad (1.8)$$

Where the $$\sum_{t \in S} e^{-\frac{E(t,x)}{RT}} = Q \qquad (1.9)$$

Therefore

$$p(s|x) = \frac{e^{-\frac{E(s,x)}{RT}}}{Q}$$

Here, Q, known as the partition function, sums over the Boltzmann factors of all possible structures in the ensemble S.

The partition function in RNA secondary structure prediction offers a powerful tool for assessing the likelihood of different structural features. By examining the probability distribution of structures within the ensemble, particularly focusing on the most probable structure (MFE structure), where one can gauge the confidence in its accuracy [271]. However, as RNA molecules grow longer, the probability of the MFE structure diminishes due to the exponential increase in the number of possible secondary structures with sequence length [275]. Despite this, the partition function approach allows for the exploration of a broader set of potential structures, offering insights beyond solely relying on the MFE prediction. Moreover, analysing the pair probabilities derived from the partition function provides information about the likelihood of specific base pairs forming within the ensemble. By summing the Boltzmann weights of structures containing a given base pair and normalizing by the partition function, one can estimate the probability of observing that base pair in a randomly chosen RNA molecule from the ensemble [273]. These pair probabilities are not independent, as the likelihood of a base pair forming may be influenced by its ability to stack with existing base pairs [265]. When it comes to the partition function, dangling ends are also considered, adding complexity to the calculations. Dangling ends where unpaired bases stack onto base pairs are pivotal in RNA secondary structure formation. Helices can stack onto each other, facilitating interactions between their closing base pairs' π-electron systems, known as coaxial stacking. These phenomena while tertiary in nature is seamlessly integrated into minimum free energy (MFE) prediction algorithms. The ViennaRNA package offers several options for considering dangling end energies, from prioritizing minimum energy stacking to mimicking coaxial stacking [279]. However, incorporating minimum energy dangling or coaxial stacking into partition function calculations would introduce complexities, such as distinguishing between different dangling configurations. Despite not being explicitly defined in secondary structure, accounting for dangling ends ensures accurate energy estimations for RNA molecules.

1.4.2.2(e) Computing the partition function

The partition function emerges as a powerful tool for scrutinizing the RNA secondary structure ensemble, and its implementation in dynamic programming algorithms proves to be relatively straightforward. In the context of the Nussinov algorithm, which operates without any energy

considerations, computing the partition function essentially entails tallying the number of possible secondary structures [270]. When incorporating energy contributions, such as exp*(i, j)* for a base pair *i, j*, the computation is slightly modified. Specifically, the partition function $Q(i,j)$ is recursively calculated using the formula:

$$Q(i,j) = Q(i-1,j) + \sum_k Q(i+1,k-1) \cdot \exp(i,k) \cdot Q(k+1,j) \qquad (1.10)$$

The initiation values $Q(i+1,i)$ and $Q(i,i)$are both set to 1 in this context, differing from their usage in the minimum free energy computation. The distinction between the minimum free energy and partition function algorithms lies primarily in the use of summation instead of minimization [280]. The computational complexity of computing the partition function remains comparable to that of determining the minimum free energy. It is noteworthy that this transition holds true for loop-based energy models as well [272, 273]. The smooth shift between the minimum free energy and partition function computations owes itself to the unique multi-loop decomposition employed. While not strictly required for minimum free energy calculations, this decomposition proves vital for accurately determining the partition function.

Equation 1.11 outlines the computation of the partition function as implemented in the Vienna RNA package [281], accounting for dangling end energies. Here, $Q(i,j)$ represents the partition function for the stretch from i to j, $Q_B(i,j)$ is the partition function when bases i and j form a base pair, and Q^M and Q^{M1} are the partition function equivalents for multi-loop contributions [272].

$$Q(i,j) = Q(i+1,j) + \sum_k Q_B(i,k) \cdot e^{d(i-1,i;k+1,k)} Q(k+1,j)$$

$$Q_B(i,j) = H(i,j) + \sum_{\cdot i<k<l<j;\, j+i-l-k>X} I(ij,kl) Q_B(k,l) \cdot e^{d(i,i+1;j,j-1)}$$

$$+ \sum_{i+5\le k\le j-5} Q^M(i,k) Q^{M1}(k+1,j) \cdot e^{a(i,j)} \qquad (1.11)$$

The equation system utilizes various terms: $H(i,j)$ for hairpin energy, a for the entropy penalty to close a multi-loop. It maintains O(n^3) complexity by limiting interior-loop lengths to ≤X. Initializations $Q(i,j)$ =1 (representing the open chain) and $Q^M\ (i,i) = Q_B(i,i) = Q^{M1}\ (i,i)$ =0 are employed [265]. Additionally, scaling is introduced to prevent overflows, ensuring accurate computation. This scaling is applied and subsequently removed during free energy computation, where $F = -RT(ln\ Q + n\ ln\ f)$. A scaling factor f is chosen such that $1 \approx Q\ f^n$ and its adequacy is confirmed by checks within the ViennaRNA package [270], providing warnings if necessary, with options to adjust f accordingly.

1.4.2.3 RNA tertiary structure prediction

RNA tertiary structure prediction is a complex computational challenge to the based nature of secondary structure prediction. The problem involves predicting the three-dimensional conformation of RNA molecules, which can be approached using various algorithms categorized into knowledge-based and physics-based methods. Knowledge-based methods typically use existing structural information to predict RNA structures, while physics-based methods rely on computational simulations to explore the energy landscape of potential conformations [281]. Knowledge-based prediction methods include fragment assembly-based and homology-based algorithms. Fragment assembly algorithms, such as MANIP, build RNA structures by combining known 3D motifs based on secondary structure data [282]. Although effective, these methods require specialized knowledge and are less accessible to general users. Homology-based algorithms, like RNABuilder and ModeRNA, use known structures of related RNA sequences as templates for predicting the target RNA structure [283,284]. Their accuracy depends on the quality of the template and alignment, presenting challenges when suitable templates are scarce. Physics-based algorithms focus on finding the lowest energy conformations of RNA structures using dynamic simulation techniques [280]. Key methods in this category include Monte Carlo simulations and molecular dynamics. Notable examples are FARNA [285] and FARFAR [286]. FARNA employs a physically based energy function and Monte Carlo sampling to predict RNA structures without evolutionary data. FARFAR enhances FARNA's approach by incorporating a more accurate all-atom energy function, achieving high prediction accuracy for complex motifs, though limitations in sampling still affect the closeness to natural conformations [286]. The "Stepwise Ansatz," introduced by Rhiju Das, addresses the challenge of incomplete conformational sampling by constructing detailed models incrementally [287]. Stepwise Assembly (SWA), implemented in the Rosetta framework, utilizes this approach to enhance prediction accuracy. SWA outperforms previous

methods like FARFAR in high-accuracy modelling, as confirmed by blind experiments. Stepwise Monte Carlo (SWM) further improves prediction by efficiently exploring energy landscapes and accurately modelling noncanonical base pairs and complex RNA motifs [288]. Recent advancements in deep learning have significantly enhanced RNA tertiary structure prediction. Techniques such as residual convolutional networks (ResNet) and deep learning-based models like DeepFoldRNA [289], RoseTTAFoldNA [290], and RhoFold[291] have improved accuracy in predicting RNA 3D structures. The success of AlphaFold2 [292] has inspired new approaches, including trRosettaRNA [293], which aims to advance RNA structure prediction by leveraging deep learning techniques and automated methods, pushing the boundaries of accuracy and reliability in the field.

1.4.3 Macromolecular Docking

Protein-protein interactions are fundamental to various biological processes, making the determination of protein-protein complex structures essential for understanding protein functions and guiding drug design targeting these interactions. However, experimental determination of such structures is limited due to cost and technical challenges, leading to a reliance on computational methods like protein-protein docking. This approach predicts complex structures by sampling potential binding modes and ranking them based on binding scores, using an energy scoring function [294]. Unlike protein-ligand docking, protein-protein docking often lacks interface information, necessitating global docking strategies, particularly in the early stages of development when experimental data was scarce [295]. Advancements in structural proteomics projects have led to an increase in experimentally determined structures of protein-protein complexes, providing more information about binding interfaces between proteins. Additionally, evolutionary conservation in protein sequences offers information into interacting residues between proteins. Leveraging this information, template-based docking has emerged as a promising approach, integrating binding interface information into the docking process [296]. By searching for homologous experimental complexes and aligning individual protein structures onto them, template-based methods construct complex structures which can be further refined using post-docking approaches like Monte Carlo optimization or molecular dynamics simulations [296]. Template-based docking has demonstrated improved performance compared to free docking, especially when conformational changes in component proteins are significant upon binding [297]. However, its reliance on templates poses limitations regarding template availability and reliability [297]. In contrast, template-free docking method samples all potential binding modes without biological information

restrictions, potentially generating true binding modes [298]. Both approaches have strengths and weaknesses, suggesting the potential benefits of integrating them to develop more robust docking protocols that capitalize on their respective advantages while mitigating weaknesses. The Critical Assessment of Predicted Interactions (CAPRI) provides a platform for blind evaluation of protein-protein docking algorithms and scoring functions, fostering the development and improvement of docking methodologies. In response, the Hybrid DOCKing strategy (HDOCK) has been developed to merge template-based and free docking methods, aiming to increase the strengths of both approaches [299]. Participating in CAPRI challenges across various categories, including oligomer modelling, protein-peptide docking, and prediction of hetero-protein complexes, HDOCK [300] has been demonstrated to be a solution, highlighting its potential in advancing protein docking methodologies. The HDOCK server [301] was initially developed for protein-protein and protein-RNA/DNA docking. However, recently it has improved to perform RNA-RNA, RNA-DNA and DNA-DNA docking, these dockings are also benchmarked in the papers.[302].

1.4.4 Molecular Dynamics

Molecular Dynamics (MD) simulations serve as a powerful tool for exploring the dynamic behaviour of proteins or nucleic acids and their interactions with ligands [303]. By numerically solving Newton's equations of motion for a system of interacting atoms, MD simulations provide insights into atomic-level details that underlie biomolecular processes. However, while MD simulations can capture various properties of molecular systems, not all quantities can be directly calculated, nor can certain quantities obtained in simulations be tracked experimentally [303]. In MD simulations, Newton's equations of motion are utilized to update atomic positions over time, with forces derived from an interatomic potential energy function [303]. The equation commonly used in MD simulations is Newton's second law of motion:

$$m_i \frac{d^2 r_i}{dt^2} = F_i \qquad (1.12)$$

Where m_i represents the mass of particle i, $d^2\boldsymbol{r_i}/dt^2$ represents the acceleration of particle i with respect to time, and $\mathbf{F}i$ represents the force acting on particle i. By iteratively solving these equations for a system of interacting atoms, MD simulations generate trajectories that represent the evolution of the system over time [303]. Through careful control of simulation parameters such as temperature and pressure, equilibrium states can be achieved, allowing for the extraction of macroscopic properties from the simulation output. It is important to acknowledge the limitations of MD simulations. Classical mechanics is employed in MD simulations,

assuming that atoms behave according to classical laws of motion [304]. While this approximation is generally valid for most atoms at normal temperatures, it may not accurately capture quantum mechanical effects, particularly for light atoms like hydrogen or systems with high-frequency vibrations [304]. In MD simulations, the traditional approach employs an all-atom Cartesian model to describe the configuration of the entire system [305]. However, there has been a growing interest in using computationally inexpensive bond/angle/torsion (BAT) coordinate models in internal coordinate molecular dynamics (ICMD) [306]. By utilizing BAT models in constrained ICMD simulations, performance improvements are achieved through a reduction in the degrees of freedom sampled and the use of longer integration timesteps [306]. This allows capturing the dynamics of components undergoing large concerted conformational changes within MD simulation timescales.

In simulations of biomolecular systems, various algorithms are employed for isothermal-isobaric (NPT) ensemble simulations to approximate solutions to the Nosé-Hoover equations during the system's motion updating by the integrator [307]. Configurational-based algorithms in temperature and pressure control have been shown to speed up calculations of heat transfer, enabling simulations to better mimic macroscopic behaviour [308]. The use of periodic boundary conditions (PBC) and complementary distance-dependent treatments of electrostatic and van der Waals non-bonded terms ensure robust simulations with reduced computational cost. In interrogating free energies of systems with smaller degrees of freedom, such as those in conformational sampling and protein folding studies, simulated annealing (SA) methods emerge as notable examples [309]. Unlike methods that generate a collective variable in their sampling, SA methods are optimization algorithms that rely on random sampling of conformations followed by minimizations to identify a range of low-energy conformations [309]. These methods manipulate three main parameters during annealing: the parameter influencing the distribution of torsions displaced, the parameter affecting the probability of accepting a conformation, and the parameter controlling the temperature at which acceptance occurs [309]. Replica-exchange molecular dynamics (REMD) methods are among the most popular non-collective variable-based physical methods that allow broad searching of the phase space [310]. These methods enhance sampling through the cloning of configurations into a set of replicas that have variable temperatures or scaled Hamiltonians of the system [310]. By allowing suitable mixing events across a diverse set of replicas, REMD strategies can enhance the sampling of the canonical ensemble, facilitating protein folding studies dependent on the extraction of sparsely populated portions of phase space [310].

1.4.4.1 Force fields

A force field in molecular simulations comprises two key components:

1. **Potential Functions**: These equations generate potential energies and their derivatives, known as forces, which dictate the behaviour of atoms and molecules in the system. The potential functions describe the interactions between atoms and molecules [311].
2. **Parameters**: These values are used within the potential functions to define specific interactions in the system. It's crucial to ensure that the combination of equations and parameters forms a consistent set to accurately represent the system's behaviour [311].

1.4.4.2 Potential Functions

Since in this study, we used GROMACS [311], we will highlight molecular dynamics approaches that involved in the software. GROMACS utilize the Langevin dynamics method. Stochastic or velocity Langevin dynamics is a method used in molecular dynamics simulations to incorporate friction and random thermal fluctuations into the Newton's equations of motion [311]. This is represented by the equation:

$$m_i \frac{d^2 r_i}{dt^2} = -m_i \gamma_i \frac{dr_i}{dt} + F_i(r) + \xi_i \qquad (1.13)$$

Here, m_i is the mass of particle i, γ_i is the friction constant, $F_i(r)$ represents the conservative force acting on particle i at position r, and ξ_i is a noise term. The noise process ξ_i is characterized by its statistical properties, with a mean of zero and a variance given by $2m_i\gamma_i k_B T\delta(t-s)\delta_{ij}$, where k_B is the Boltzmann constant. When the friction constant is much larger than the characteristic time scales of the system, stochastic dynamics behaves similar to molecular dynamics with stochastic temperature coupling [312]. However, processes with time scales longer than $1/\gamma i$, such as hydrodynamics, are damped. In GROMACS, a simple and efficient implementation of stochastic dynamics is provided. This implementation is equivalent in accuracy to the conventional leap-frog and Velocity Verlet integrators used in

molecular dynamics [311, 313]. The integration scheme involves updating the velocity and position of particles using specific equations, with friction and noise applied as impulses at discrete time steps. The advantage of this scheme is that the velocity-dependent terms act over the entire time step, facilitating the correct integration of forces that depend on both coordinates and velocities [313]. When the friction constant is small compared to the time scales of the system, the dynamics are significantly different from conventional molecular dynamics, but the sampling remains correct. In the limit of high friction, stochastic dynamics reduces to Brownian dynamics, also known as position Langevin dynamics [314]. In this method, inertia effects are negligible, and the dynamics are only determined by the interplay between friction and thermal fluctuations. The equation of motion in this method simplifies to:

$$\frac{dr_i}{dt} = \gamma_i^{-1} F_i(r) + \xi_i \qquad (1.14)$$

GROMACS provides a straightforward integration scheme for Brownian dynamics, where the positions of particles are updated based on the conservative forces and stochastic noise. This approach is suitable for systems where inertia effects are negligible, and large time steps can be used. Constraints are typically enforced using the LINCS algorithm in Brownian dynamics simulations, as the SHAKE algorithm may not converge for large atomic displacements.

1.4.4.3 Numerical integration

In molecular dynamics simulations, the Verlet algorithm and its variations play a crucial role in numerically integrating Newton's equations of motion [313]. One such variation, the Brünger-Brooks-Karplus (BBK) method, integrates the Langevin equation, which incorporates stochastic effects [315]. The BBK method is represented by the following equation:

$$rn{+}1r_{n+1} = r_n + \left(1 - \frac{\gamma \Delta t}{2}\right)\left(1 + \frac{\gamma \Delta t}{2}\right)^{-1}(r_n - r_{n+1}) + \frac{1}{\left(1+\frac{\gamma \Delta t}{2}\right)} \frac{\Delta t}{2}\left[\frac{1}{m} F(r_n) + \frac{r}{2\gamma k_B T} \Delta m Z_n\right] \qquad (1.15)$$

In this equation, r_{n+1} represents the position of the particle at the next timestep, r_n is the position at the current timestep, γ is the damping coefficient, Δt is the timestep, m is the mass of the particle, $F(r_n)$ is the force at the current position, k_B is the Boltzmann constant, T is the temperature, Δm is the change in mass, and Z_n is a Gaussian random variable with zero mean

and unit variance. On the other hand, the velocity Verlet algorithm is another widely used integration scheme [315]. It calculates both position and velocity at each timestep and is represented by the following equations:

$$r_{n+1} = r_n + v\Delta t + \frac{F_n}{2m}\Delta t^2 \qquad (1.16)$$

$$v_{n+1} = v_n + \frac{F_n + F_{n+1}}{2m}\Delta t \qquad (1.17)$$

Here, r_{n+1} and v_{n+1} represent the position and velocity at the next timestep, respectively, while v_n and r_n represent the position and velocity at the current timestep. F_n and F_{n+1} are the forces at the current and next timesteps, m is the mass of the particle, and Δt is the timestep. This algorithm ensures the consistency of updating both position and velocity, contributing to the stability and accuracy of the simulation.

1.4.4.4 Potential energy

In MD simulations, solving Newton's equations of motion requires calculating the potential energy function (U), a critical component governing particle behaviour [313]. The Amber force fields are mostly chosen for their efficacy in nucleic acid simulations. And the potential energy defined $U(r)$ through bonded and non-bonded interactions is illustrated in equation 1.18 [316]:

$$U(r) = \sum_{\text{bonds}} K_b(b - b_0) + \sum_{\text{angles}} K_\theta(\theta - \theta_0) + \sum_{\text{dihedrals}} \left(\frac{V_n}{2}\right)(1 + \cos[n\phi - \delta]) + \sum_{i<j} \epsilon_{ij}\left[\left(\frac{R_0^{ij}}{R^{ij}}\right)^{12} - 2\left(\frac{R_0^{ij}}{R^{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{R^{ij}} \qquad (1.18)$$

Various types of interactions contribute to the overall potential energy function, governing the behaviour of particles within the system. Bonded interactions are characterized by the oscillations of bonds around their equilibrium lengths, represented by the first term in the potential energy equation. This term is parameterized by the force constant K_b and the equilibrium bond length b_0. Angle interactions, denoted by the second term, involve oscillations around equilibrium angles, determined by the force constant K_θ and the equilibrium angle θ_0. Additionally, torsional rotations are governed by the third term, which considers parameters such as amplitude (V_n), periodicity (n), and phase (δ). Non-bonded interactions, depicted in the fourth and fifth terms, encompass Lennard-Jones and Coulombic

potentials, respectively. The Lennard-Jones potential captures both attractive and repulsive forces between atoms, influenced by the equilibrium distance R_0 and the potential well depth ϵ_{ij}. Conversely, the Coulombic potential describes electrostatic interactions between charged atoms (q_i and q_j), separated by a distance R_{ij}. Hydrogen bonding effects are accounted for within these non-bonded potentials, enriching the model's accuracy in representing molecular interactions [317]. To determine the parameters in the potential energy function, a combination of quantum-mechanical calculations and experimental validation is employed [317].

1.4.4.6 Parameters development (AMBER) for nucleic acids

The pivotal components in molecular dynamics simulations, dictating the mathematical expressions that determine how atoms and molecules interact within a system [318]. These parameters encompass various aspects of molecular behaviour, including bond lengths, bond angles, and dihedral angles [319]. Some force fields also include terms for polarization and dispersion effects to enhance accuracy, but the effectiveness of these parameters can vary depending on the specific system being simulated, requiring thorough validation [320]. The development of molecular mechanics force fields within the AMBER suite has seen significant evolution since the early parameterizations provided by Weiner *et al* [321]. These early force fields are now considered obsolete due to limitations in accuracy and methodology. Progress continued with the Cornell *et al.* force field, which improved upon previous models, and the ff94 force field, enhancing performance in solvated systems [322]. However, subsequent versions, such as parm96.dat and parm98.dat, addressed specific biases and improved accuracy, particularly in torsional angle parameters for nucleic acids [322,323]. Despite advances like the Wang *et al.* force field and the introduction of polarizable force fields (ff02 and ff02EP) in 2002, which enhanced the accuracy of intermolecular interactions [324]. Many older parameter sets including those for ions, are now outdated and not recommended for modern simulations. In the domain of nucleic acids, particularly RNA, advancements in force fields have focused on addressing inaccuracies revealed through longer simulations, such as systematic errors in backbone geometries and incorrect loop conformations [325]. Early force fields laid the groundwork with basic parameters for nucleic acids, but more recent developments, like ff99-bsc0, have aimed to correct issues observed in longer simulations [326]. These modifications, including OL adjustments, phosphate parameter refinements, and χ angle corrections, have improved the modelling of canonical DNA and RNA structures. Despite these improvements, challenges remain for non-canonical structures, leading to the development of alternative force

fields by groups like Rochester and DE Shaw to better represent RNA's complex conformational preferences and dynamics [327,328].

1.4.5 Molecular Mechanics Generalized Born Surface

The Molecular Mechanics Generalized Born Surface Area (MMGBSA) method is primarily used to calculate the binding free energies (ΔG_{bind}) of small molecule ligands with large biomolecular receptors [329]. The binding free energy in an aqueous solvent ($\Delta G_{\text{bind/aq}}$) can be approximated by considering the gas-phase molecular mechanical energy change (EMM), the solvation free energy change ($\Delta G_{\text{bind/solv}}$), and the conformational entropy change (−TΔS). EMM includes covalent, electrostatic, and van der Waals energy changes, while $\Delta G_{\text{bind/solv}}$,is divided into polar and non-polar contributions [330]. These energy components are computed via ensemble averaging over a large set of conformations obtained from molecular dynamics (MD) simulations, which are typically performed in an explicit solvent model. The polar solvation term (ΔG_{polar}) is calculated using the Poisson-Boltzmann equation (PBE) or its Generalized Born (GB) approximation, which replaces the explicit solvent with an implicit continuum solvent to speed up calculations [330]. The equations used to compute MM-GBSA are described in equation (1.18 – 1.20).

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S \approx \Delta E_{\text{MM}} + G_{\text{bind/sol}} - T\Delta S \qquad (1.18)$$

$$\Delta E_{\text{MM}} = \Delta E_{\text{internal}} + \Delta E_{\text{electrostatic}} + \Delta E_{\text{vdw}} \qquad (1.19)$$

$$\Delta G_{\text{bind/sol}} = \Delta G_{\text{polar}} + \Delta G_{\text{non-polar}} \qquad (1.20)$$

The GB model considers the electrostatic potential distribution in the solvent and can be linearized for numerical efficiency. Recent developments in GB solvers, such as decomposition and minimization schemes and GPU implementations, have significantly accelerated MM/GBSA computations [331]. Higher solute dielectric constants have been found to improve the accuracy of the free energy calculations, which led to a high correlation with experimental binding affinities, especially for systems with highly charged binding pockets [332]. The non-polar solvation term ($\Delta G_{\text{non-polar}}$) arises from cavity formation and van der Waals interactions between the solute and solvent. Traditionally, it is estimated to be proportional to the solvent accessible surface area (SASA). Modern approaches model cavity formation, and dispersion-free energy separately, using solvent-accessible volume (SAV) for cavity formation and specific coefficients for dispersion. These models provide more accurate binding affinity predictions by reducing root-mean-square deviations from experimental values [333]. Overall,

while the classical and modern approaches yield similar non-polar solvation free energies for small molecules, the modern approach offers better performance for larger systems.

1.5 Aims and Objectives

The primary aim of this project is to design and discover novel aptamers that can effectively inhibit miRNA-10b using computational approaches on a large scale. The focus on miRNA-10b is driven by its well-documented role in promoting cancer cell proliferation, migration, invasion, and metastasis in various cancer types, including breast cancer, glioblastoma, and oesophageal squamous cell carcinoma. Inhibiting miRNA-10b presents a promising therapeutic strategy to combat these aggressive malignancies and improve patient outcomes. Additionally, this project will highlight the opportunity to significantly reduce costs by utilizing virtual screening methods, which are more efficient and less resource-intensive compared to traditional SELEX (Systematic Evolution of Ligands by Exponential Enrichment) processes. By leveraging computational tools, we can accelerate the discovery of effective aptamers, reducing the time and expense typically associated with experimental methods.

In order to achieve these aims, the objectives are:

1. Design and develop a computational tool that can generate RNA aptamer sequences of specified sizes by exploring the entire space of randomly positioned bases and their arrangements on a large scale.
2. Create a tool capable of predicting secondary and tertiary structures of RNA aptamers on a large scale to identify the most stable aptamer configurations.
3. Develop a tool for the large-scale screening of these aptamers against any kind of macromolecular target, ensuring high throughput and accuracy in identifying potential binding candidates.
4. Virtually screen the novel aptamers to identify those that can bind effectively to miRNA-10b. Confirm the aptamer-miRNA-10b complexes of the best binding aptamers using Molecular Dynamics (MD) simulations for detailed interaction analysis and use Molecular Mechanics Generalized Born Surface Area (MMGBSA) calculations to assess the binding free energies and stability of the aptamer-miRNA-10b complexes.

Chapter 2

T_SELEX Program

2.1 Overview

Selection of aptamers is a crucial iterative process which is significant for many applications from biosensors to biomedical applications. The process of selecting aptamers is called systematic evolution of ligands by exponential enrichment (SELEX). This process is costly, time consuming and labour-intensive. In this chapter we present T_SELEX API together with its algorithms and application. This program is an autonomous robust python package specifically engineered for designing aptamers and screening of RNA aptamers against macromolecular targets. The tool is evaluated from generation of sequence library of RNA aptamers, multiscale folding or secondary structure predictions, multiscale tertiary structure predictions, virtual screening and detailed analysis. The application was tested by screening a generated aptamer library against HIV-1 protease. The source code of the T_SELEX tool is available on GitHub [https://github.com/CMCDD/T_SELEX] and can be used directly following instructions on the readme file.

2.2 Introduction to T_SELEX program

Aptamer technology has evolved significantly, offering effective means to isolate nucleotide molecules with high affinity for specific targets. While both *in vivo* and *in vitro* techniques exist, *in vitro* methods are preferred due to technical challenges associated with *in vivo* approaches, particularly when dealing with known and unknown target sequences. *In vitro* methods such as SELEX, EMSA, and DNA foot-printing offer efficient ways to select aptamers [334]. SELEX, in particular, has advanced with the development of aptamer technology, allowing selection against various targets including peptides, proteins, and cells [331]. Modern approaches often combine these techniques to enhance efficiency [333]. However, implementing SELEX can be labour-intensive and costly, prompting exploration into *in silico* approaches that utilize computational tools for aptamer design and selection. Challenges remain in this domain, particularly in generating large libraries and screening macromolecules.

Computational tools like Hdock [299], Zdock [335], and pyDock [336] play key roles in understanding macromolecular interactions, however they are not designed for high-throughput screening. In this study, we introduce the Theoretical-SELEX API (T_SELEX). This API is dedicated for the design of virtual RNA aptamer libraries, and it accomplishes this by starting with sequence generation working through to final tertiary structure predictions at a large scale. Additionally, it screens these virtual aptamers using the HDock algorithm designed by Hueng [299, 300]. The primary goal of this chapter is to introduce T_SELEX as an innovative computational pipeline for the rational design and screening of RNA aptamers.

2.3 Implementation

The T_SELEX API works within a python 3 environment and is recommended to be used within an Anaconda environment. There are some dependencies, and one external tool is required. There is a helper tool (install_dependencies.py) that aids with the installation of the dependencies which include Selenium, ViennaRNA [265], RNAComposer [337], and IntaRNA [338] and ensures all required components are correctly installed and configured. The external tool is HDOCK Lite [300] which needs to be installed separately. To fully set up the T_SELEX_API, after downloading and placing the T_SELEX files in the "*site-packages*" directory of the conda environment, these dependencies are installed and T_SELEX is then fit for purpose. This package functions similarly to any other Python library.

2.4 The Algorithm

To illustrate how to use T_SELEX as an API, sample code that demonstrates its basic functionality is shown in **Code1**. This example shows how to make the basic API calls. Additionally, it is possible to run a full autonomous program that handles the complete workflow, from simple sequence generation to virtual screening as shown in **Code2**. The steps from **Code1** are highlighted and explained in the following section.

Code1: T_SELEX program API snippet sample code

```
import pandas as pd
import T_SELEX_program
from T_SELEX_program import
gen_aptamers,fold_and_composition,tertiary_structure,Mol_docking_calc,
import os
aptmers_sequences = gen_aptamers(1,33,10)
print(p)
df = fold_and_composition(aptamers_sequences)
tertiary structures = tertiary_structure(df['Aptamer'],df['MFE structure'] )

home_directory = os.path.expanduser( '~' )
sofware_path = 'software/HDOCKlite-v1.1'
r = '/home/s1800206/Downloads/6zsl.pdb'
l = '/home/s1800206/hdock_test'
s = '/home/s1800206/software/HDOCKlite-v1.1'
virtual_screening = Mol_docking_calc(data_frame= a,MFE_column ='Minimum free
Energy',receptor_name= "6zsl",receptor=r,ligands_directory=l,directory_path=
s,Ap_folded=True)
```

Code2: Full automated sample command for T_SELEX program

```
:~$ t_selex_program --gen_aptamers_seed 1 --gen_aptamers_length 33 --gen_aptamers_width 10
--receptor_file /home/s1800206/Downloads/6zsl.pdb --ligands_directory
/home/s1800206/hdock_test --software_path software/HDOCKlite-v1.1
```

Step1: RNA Aptamer library generation

The first crucial step involves either generating a library of RNA sequences or acquiring an existing one. If the user already possesses a list of sequences, they may proceed directly to Step 2. The program offers two methods for generating the RNA aptamer library, both of which are implemented in the ***secondary.py*** script. The first option involves generating random sequences using the Base Randomization Algorithm (BRA), while the second option entails downloading RNA sequences from the Aptamer Base [339].

Step1a: Generating Aptamer library using Base Randomization Algorithm

To generate random sequences, the Base Randomization Algorithm (BRA) is employed. Although BRA ensures the unique positioning of bases within each sequence, it does not guarantee perfect randomness in the selection of the actual bases. To generate random sequences using BRA, the ***gen_aptamers(*)*** function must be called, which requires three arguments (* is symbol that shows that these functions take in augments). The first argument is ***seed***, a critical parameter that sets the initial value for the random number generator [340]. If the seed argument is set to ***None***, the seed value will change with each execution of the

function, resulting in different lists of aptamers each time the function is run. Conversely, if a specific seed value is provided, the function will use this seed value, allowing for consistent output and easier tracking of sequences.

The second argument is ***length***, which specifies the length of the sequences. If length is set to "*randomized"* the function will generate sequences with random lengths ranging from 16 to 60 nucleotides. For generating sequences longer than this range, minor modifications to the source code in the ***secondary.py*** script will be required. If a specific length is desired, the length parameter can be set to that value; however, all generated sequences will then be of the same length. The third argument, ***aptamers_num***, denotes the number of sequences to be generated. The function is designed to produce unique sequences with no repeats in the dataset. Users should exercise caution when selecting parameters. If the product of the number of bases and the sequence length exceeds the desired number of sequences, the function may not yield a valid output. For instance, choosing ***length=3*** and ***aptamers_num=5000*** may lead to an insufficient number of unique combinations due to the limited possible arrangements of 4 bases in sequences of length 3.

```
def gen_aptamers(seed,length, aptamers_num):

    if seed == None:
        if length == "randomize":

            aptamers = set()

            while len(aptamers)< aptamers_num:

                p= list(range(16,60))

                l = random.choices(p)

                aptamer = ''.join(random.choices('ACUG', k=l[0]))

                if aptamer not in aptamers:

                    aptamers.add(aptamer)
        else:

            aptamers = set()

            aptamer = ''.join(random.choices('ACUG', k=length))

            while len(aptamers)< aptamers_num:

                aptamer = ''.join(random.choices('ACUG', k=length))

                if aptamer not in aptamers:

                    aptamers.add(aptamer)
```

```
    else:
        random.seed(0)
        if length == "randomize":
            aptamers = set()
            while len(aptamers)< aptamers_num:
                p= list(range(16,60))
                l = random.choices(p)
                aptamer = ''.join(random.choices('ACUG', k=l[0]))
                if aptamer not in aptamers:
                    aptamers.add(aptamer)
        else:
            aptamers = set()
            aptamer = ''.join(random.choices('ACUG', k=length))
            while len(aptamers)< aptamers_num:
                aptamer = ''.join(random.choices('ACUG', k=length))
                if aptamer not in aptamers:
                    aptamers.add(aptamer)
    return list(aptamers)
```

Step1b: Downloading Aptamer library

The second option involves downloading RNA aptamer sequences from AptamerBase, which is available on GitHub [339]. To perform this task, the ***aptamerbase(*)*** function, located in the secondary.py script, should be called. This function requires a single argument, ***n_type***, which specifies the nucleic acid type. The ***n_type*** argument can be set to either "DNA" or "RNA," depending on the user's preference. When ***n_type*** is set to "DNA," the algorithm filters out any sequences in the dataset that contain uracil (U or u) bases. Conversely, when n_type is set to "RNA," the function removes sequences containing thymine (T or t) bases. Unlike the ***gen_aptamers(*)*** function, which returns a list of sequences, the ***aptamerbase(*)*** function returns the results as a data frame. Additionally, the ***aptamerbase(*)*** function performs default analyses, including sequence length analysis for the generated dataset. The results of this

analysis are presented as plots, which include a bar graph and a Cumulative Empirical Distribution Function (CEDF) plot.

```
def aptamerbase(n_type):

  import seaborn as sns

  import matplotlib.pyplot as plt

  from statsmodels.distributions.empirical_distribution import ECDF

  import numpy as np

  # this data base is downloaded from aptmerbase projects
  # the article to reference when using this dataset
  #(http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3308162/)
  # this function preprocess the data, by removing empty columns

  url_data =
(r'https://raw.githubusercontent.com/micheldumontier/aptamerbase/master/d
ata/aptamerbase_aptamers.csv')

  df = pd.read_csv(url_data)

  nan_indices_column_A = df[df['sequence'].isna()].index

  nan_indices_column_A

  drop_list =[]

  for i in nan_indices_column_A:

    #print(i)

    drop_list.append(i)

  df = df.drop(df.index[drop_list])


  if n_type == 'RNA':

    df['sequence'] = df['sequence'].astype(str)

    df = df[~df['sequence'].str.contains('T|t', na=False)]

    df = df.reset_index(drop=True)

    df['sequence'] = df['sequence'].str.upper()


  elif n_type == 'DNA':
```

```
    df['sequence'] = df['sequence'].astype(str)

    df = df[~df['sequence'].str.contains('U|u', na=False)]

    df = df.reset_index(drop=True)

    df['sequence'] = df['sequence'].str.upper()


  else:
      error1 =print("You entered wrong n_type")

      error2 =print("for RNA aptamers .... aptamerbase(n_type ='RNA')")

      error3 =print("for DNA aptamers .... aptamerbase(n_type ='DNA')")

      df = {error1,
            error2,
            error3}

  #count plot for leagth analysis

  lengths_seq = []

  sequences = df['sequence'].astype(str).tolist()

  for sequence in sequences:

    length_of_sequence = len(sequence)

    lengths_seq.append(length_of_sequence)

  df['sequence_length'] = lengths_seq
```

Step2: RNA folding and secondary structure predictions

Once the sequence datasets are generated in Step 1, the next step involves predicting the Minimum Free Energy (MFE) of the secondary structures of the RNA sequences. To accomplish this, the ***fold_and_composition(*)*** method must be used. This method takes a single input: the list of sequences generated in Step 1. If the RNA sequences were downloaded using the second option in Step 1 (from AptamerBase), only the sequence column from the data frame should be passed to ***fold_and_composition(*).*** This is because ***fold_and_composition(*)*** does not accept data frames directly, so methods such as ***.tolist(*)*** from the pandas module can be used to extract the sequence column. A critical dependency for this process is the RNA module from the ViennaRNA Suite [265] (Python interface). This module includes the

RNAfold method, which processes the sequences to return both the MFE value and the MFE structure.

```
def fold_and_composition(aptamers_list):
  import RNA
  data =[]
  for i, aptamer in enumerate(aptamers_list):
      G = 0
      A = 0
      C = 0
      U = 0

      for char in aptamer:
          if char == "G":
              G += 1
          if char == "A":
              A += 1
          if char == "C":
              C += 1
          if char == "U":
              U += 1
      (MFE_structure,MFE_energy) = RNA.fold(aptamer)



      data.append([i+1, aptamer, U, G, A, C, MFE_structure,MFE_energy ])

  # Define the column headers
  headers = ['Number', 'Aptamer', 'Us', 'Gs', 'As', 'Cs','MFE
structure','Minimum free Energy']
  import csv
  with open('data.csv', 'w', newline='') as csvfile:
      writer = csv.writer(csvfile)
      writer.writerow(headers)
      for row in data:
          writer.writerow(row)

  import pandas as pd
  df = pd.read_csv('data.csv')
  return df
```

The MFE structures are provided as pseudoknots, which indicate the regions of base pairing within each sequence. The ***fold_and_composition(*)*** method utilizes this capability to fold sequences on a large scale. Additionally, it calculates the composition of each base within the sequences and returns the results in a data frame. To facilitate the identification and avoid confusion, each sequence is assigned a unique ID using an enumeration algorithm. This step is relatively efficient, with the algorithm capable of folding approximately 20,000 RNA sequences of varying lengths (16-60 nt) in about 2 minutes.

Step3: Tertiary structures prediction

The third step involves, predicting and generating the tertiary structures from the secondary structures obtained in step 2. This step relies on the RNAcomposer webserver, which acts as a third-party service to perform this task [337]. The function the tertiary structure is ***tertiary_structure(*)*** method. This method takes in only two arguments which are list of aptamers sequences (*aptamer_list*) and the list of the secondary structures (*secondary_structure*). It is important to keep in mind that this ***tertiary_structure()*** is an automated tool using the Selenium package (https://pypi.org/project/selenium/) that is mostly used for data scraping. If step 2 is omitted and there are no secondary structures associated with the individual sequences, alternative options exist for folding to aid in tertiary structure prediction without the need for secondary structure information. In the first option, *secondary_structure = "default",* the RNAfold algorithm will be deployed to the sequences in order to generate the 3D structures. The second option, ***secondary_structure = "CF"*** , for CentroidFold algorithm to help with folding prior 3D structure prediction. The other options includes ***secondary_structure = "Context"*** for ContextFold algorithm, ***secondary_structure = "CONTRA"*** for CONTRAfold, ***secondary_structure = "IPknot"*** for IPknot, and *secondary_structure = "RNAstructure"* RNAstructure algorithm.

But if the secondary structures were obtained using step 2, then ***secondary_structure*** argument will be equal in length with the list of the secondary structures generated. Caution must be observed, as this method will be running through those lists in parallel, and therefore it is important to make sure that both lists are equal and to ensure that each aptamer and its secondary structure have same position in those lists. This specific method relies heavily on the computer's internet connectivity and its speed since to generate these tertiary structures since connection to the RNAComposer webserver [337] has to requested. We extended the potential duration for the formation of one aptamer tertiary structure to a maximum of 200 seconds. This adjustment allows for situations where the user's connectivity may need to be improved. The output files of this method are pdb format files and the text file that contains possible motifs recognized during the calculation.

```
def tertiary_structure(aptamer_list, secondary_structure):

    # Set up the driver (in this case, Chrome)

    driver = webdriver.Chrome()

    if secondary_structure == "default":


        name = 0


        for aptamer in aptamer_list:

            # nevogate through RNAcomposer

            driver.get("https://rnacomposer.cs.put.poznan.pl/")
            time.sleep(20)
            name += 1

            search_box = driver.find_element(By.NAME, "content")

            search_box.clear()

            search_box.send_keys(f'>aptamer{name}')

            search_box.send_keys(Keys.RETURN)

            search_box.send_keys(aptamer)

            search_box.send_keys(Keys.RETURN)

            search_box.send_keys("RNAfold")

            search_box.send_keys(Keys.RETURN)

            #nevigate and click compose button

            compose = driver.find_element(By.NAME,"send")

            compose.click()

            # Wait for the search results to load

            time.sleep(20)

            # Print the search results

            results = driver.find_elements(By.CLASS_NAME, "task-log")

            output_file_name = f"output_aptamer{name}.txt"

            with open(output_file_name, "w") as file:
```

```
                for result in results:
                    file.write(result.text + "\n")
            # download pdb file
            pdb = driver.find_element(By.PARTIAL_LINK_TEXT,
"aptamer").click()
        # Close the driver
        driver.quit()
```

Step 4: RNA – RNA interactions prediction

This step is somewhat different as it requires only sequences, not 3D structures, to make interaction predictions. This task is performed using the ***intarna(*)*** method from the interactions module. The ***intarna(*)*** method employs the recent IntaRNA algorithm [338], which is faster, more accurate, and takes seeding into consideration. The seedBP is set to four, representing the number of base pairs with the seed, and the output is detailed in a log file has the same basename as that of the input provided. The ***intarna(*)*** method accepts five arguments. The first argument is the name of the CSV file containing the aptamer sequences, for instance, ***csv_file*** = "aptamers.csv". The second argument specifies the name of the column that contains the sequences. The third argument is the target sequence of interest, provided as a string in capital letters. The fourth argument is the name assigned to the target sequences, which is used for naming the output files. Finally, the fifth argument is analysis, which can be set to True or False. If ***analysis*** is set to True, the algorithm will generate and save log files and create plots to aid in overall analysis. It is recommended to set analysis to True for large-scale analyses, as setting it to False will only print the energy values and not save the detailed output files. This method returns a data frame saved as a CSV file.

Step 5: Macromolecular docking.

This is the most crucial and time-consuming step. Molecular docking in this package is performed using the Hdock algorithm [299] as implemented in the external tool Hdock [299]. The Hdock algorithm was chosen for its robustness and accuracy [299]. This calculation is performed by calling ***Mol_docking_calc(*)*** from ***Docking.py*** script, and it requires seven arguments. The function takes several arguments to facilitate the docking process of aptamers. The **data_frame** argument represents a DataFrame that contains aptamer sequences along with their Minimum Free Energy (MFE) structures. The **MFE_column** specifies the name of the column in the DataFrame that holds the MFE values of the aptamers. The **receptor_name**

argument defines the name of the receptor file to be used for the docking process, while the **receptor** argument provides the path to the receptor PDB file, which contains the 3D structure of the receptor. The **ligands_directory** argument points to the directory containing the ligand files, which typically include the aptamers. The **directory_path** refers to the path to the HDOCK software directory [300], which will be used for docking calculations. Lastly, the **Ap_folded** argument is a boolean flag that determines whether to consider folded aptamers (True) or non-folded aptamers (False) during the docking procedure. These arguments collectively enable the function to process aptamers, receptors, and docking software to carry out the docking simulations. Molecular docking will produce a CSV file containing aptamers, best docked model docking score, Fitness quality, and confidence score.

Step 6: Docking Results analysis

The CSV file generated after molecular docking calculations is necessary for aiding with the analysis. This entails conducting an in-depth and thorough analysis of the complete dataset, focusing on aspects such as fitness quality, docking score RMSD fluctuation values for each ligand across all models, and producing confidence scores. Regression analysis is also conducted, taking into account different models to assess the correlation between the best model and the other models generated for each docked ligand. These evaluations are visualized through plots. When dealing with multiple targets, the dataset of ligands or aptamers must remain consistent. This analysis calculates Z-scores to assess statistical significance, and full comparisons are made across targets using multiple plots. The detailed T_ SELEX workflow is shown in **Figure 2.1**.

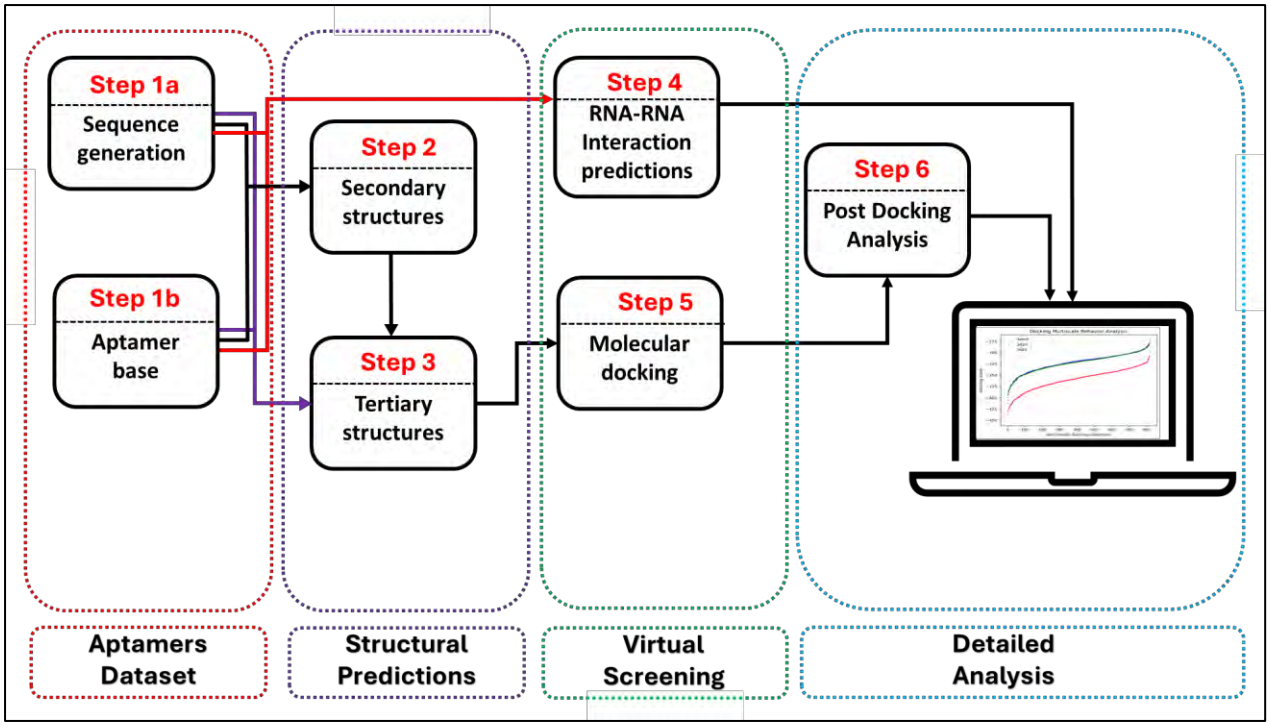


Figure 2.1: Simplified workflow of the T_SELEX algorithm.

Table 2.1: Summary and description of some methods in the T_SELEX package.

Modules	Length (lines)	Methods	Descriptions	Dependencies
Secondary	1 432	*gen_aptamers()*	Generates library of RNA sequences using Base randomization algorithm.	-
		Aptamer_base()	Uploads RNA and/or DNA sequences from the Aptamer base data base from GitHub.	-
		fold_and_composition()	Folds large library of sequences and provide the their secondary structure pseudoknots, MFEs and calculates base composition.	ViennaRNA [265]
		thermodynamics_properties()	Calculates the RNA helix stability by determining melting point using the recent Nearest neighbour parameters from David Mathews research [341].	-
		mass()	Approximate the mass of the aptamers or RNA sequences based on given sequences.	-
		find_similar_aptamers()	The algorithm compares sequences of aptamers to find similar ones based on a given similarity threshold “n”.	-
		tertiary_structure()	Takes multiple sequences together with their secondary structure if given, and predicts the 3D structures using the machine translation principle	RNAComposer web server[337]

			and operates on the RNA FRABASE database acting as the dictionary relating RNA secondary structure and tertiary structure elements[337].	
Interactions	1 782	*intarna()*	Predicts the interaction between RNAs, and further perform pairwise alignment between a query sequences and target RNA sequence, where the target RNA sequence provided.	Intarna [338] and Biopython
Docking	597	*Mol_docking_calc()*	Performs larger scale molecular docking simulations and returns hundred models for each given macromolecule ligands or aptamers.	Hdock[300]
		Sep()	Filter folded aptamers from unfolded aptamers based on the MFE.	-
PDA	625	*DMBA()*	Docking Multiscale Behaviour Analysis, it generates plots for comparing the aptamers docking scores given that there were 6 or less targets.	-
		BMA()	Calculates fit quality and iterate within 100 models of the 10 best aptamers and generate, RMSD ,Docking scores and fit quality plots.	-

		PDCA()	Compute the Zscores of the docked aptamers and their models and generate 2 and plots for analysis.	-
VRNA	672	*FullFold()*	Predict equilibrium pair probabilities using partition function and pair probabilities using maximum expected accuracy.	ViennaRNA [265]
		RNAcofold()	RNA-RNA interaction predictions using partition function and pairing probabilities	ViennaRNA [265]
		RNAeval()	Reveal parameters of a given sequences and secondary structure.	ViennaRNA[265]
		RNAup()	Compute the thermodynamics of RNA-RNA interaction and predict the accessibility of the target region.	ViennaRNA [265]
		barriers()	Compute the RNA folding kinetics as RNA approach biological active state by exploring large energy land scape.	ViennaRNA [265]

2.5. Mathematical representations for some novel algorithm.

Although T_SELEX is an automation package, there are some algorithms that may be considered novel in this package, this includes BRA and ***find_similar_aptamers()*** and others that are worth unpacking and clarifying.

2.5.1 Base randomization algorithm (BRA)

Base randomization algorithm method generates RNA sequence aptamers with randomized bases. This algorithm is derived from pseudo number generators which makes use of seed initialization method to produce sequences that appear randomised.

2.5.2 Sequences Similarity Check Algorithm (SSCA)

This algorithm identifies similar aptamers within a dataset of sequences by using a sequence comparison approach, similar to conventional sequence alignment methods, with some key differences. It employs loops to iterate through each sequence in the dataset, comparing it against all other sequences. Similarity is then scored based on the comparison results. This approach systematically evaluates the relationships between sequences to identify those that are similar based on a defined scoring criterion. For similar aptamers (sa) the algorithm can be thought of as equation 2.1.

$$sa = \{(t_i, t_j, s(a_i, a_j)) \mid s(a_i, a_j) \geq n, \forall i, j) \ \} \qquad (2.1)$$

For any two aptamer sequences a_i and a_j $s(a_i, a_j)$, with sequence tags t_i and $t_j$are retained to assist each identifying those aptamers in the list. The *sa* equation return a set of two tagged aptamers together with its similarly score given that the similarity score exceed a given threshold n. Our similarity score function is defined shown in equation 2.2.

$$s(a_i, a_j) = \sum_{k=1}^{L_{\min(i,j)}} \delta(a_i[k], a_j[k]) \qquad (2.2)$$

Similarity score function is the summation of counts for similar bases in same positions. The equation for each base similarity count follows Kronecker delta function be expressed in equation 2.3.

$$\delta(a_i[k], a_j[k]) = \begin{cases} 1 & if \ \ a_i[k] = a_j[k] \\ 0 & if \ a_i[k] \neq a_j[k] \end{cases} \qquad (2.3)$$

Where $a_i[k]$ and $a_j[k]$ represent the elements at position k in sequences a_i and a_j, respectively. Therefore, this function simplifies to 1, if the element is sequence a_i and a_j are equal, but if that's not the case delta function is equal to 0.

2.5.3 MP-NNTA (Matrix Based Python Nearest Neighbour Thermodynamics Algorithm)

For nearest neighbour model algorithm, which is essential for calculating the melting point of the RNA helix. As mentioned before, since the experimental parameters were obtained from the recent paper by Mathews published in 2022 [341]. In that manuscript, they further emphasis the need of additional terminal parameters to predict increase melting point accuracy of the RNA helix. The introduction of those special handlings influenced by difficulty into implementing this into the algorithm. But here we provide mathematical expression of how our algorithm surpasses those challenges. For a given sequence *S*, we first compute the reverse complementary sequences S' as illustrated in equation 2.4 and 2.5

$$S = [B_1, B_2, B_3, \ldots\ldots, B_{N-1}, B_N] \qquad (2.4)$$

$$S' = [B_N, B_{N-1}, \ldots\ldots\ldots, B_3\ B_2, B_1] \qquad (2.5)$$

Where B is the bases of the sequence of *N* length, of course this algorithm is written assuming the main sequence is also self-complementary reversed as shown to be the case by David Mathews [341]. From there we computed RNA helix as ($\boldsymbol{\vartheta}$)matrix of 2 × N.

$$\boldsymbol{\vartheta} = \begin{bmatrix} S \\ S' \end{bmatrix} = \begin{bmatrix} B_1\ B_2\ B_3\ \ldots\ldots\ldots\ B_{N-1}\ B_N \\ B_N\ B_{N-1}\ \ldots\ldots\ldots\ B_3\ B_2\ B_1 \end{bmatrix} \qquad (2.6)$$

Once the matrix computed, then we search through the matrix ($\boldsymbol{\vartheta}$) for any pairs (η_i), using function expressed in equation 7.

$$\eta_i = \begin{bmatrix} S_i & S_{i+1} \\ S_i' & S_{i+1}' \end{bmatrix} \qquad (2.7)$$

Where S_i and S_{i+1} are elements of S while S_i' and S_{i+1}' elements of S', moreover i can be a position of the base on the sequence. The energy changes (enthalpy, entropy, and Gibbs free energy) associated with these pairings are computed using the following equations:

$$\Delta H^{o}_{pairings}(\eta_i, \gamma_i) = \begin{cases} \Delta H^{o}_{i}, \eta_i = \gamma_i \\ 0, \ \eta_i \neq \gamma_i \end{cases} \quad \gamma_i \in \varsigma \tag{2.8}$$

$$\Delta S^{o}_{pairings}(\eta_i, \gamma_i) = \begin{cases} \Delta S^{o}_{i}, \eta_i = \gamma_i \\ 0, \ \eta_i \neq \gamma_i \end{cases} \quad \gamma_i \in \varsigma \tag{2.9}$$

$$\Delta G^{o}_{pairings}\ (\eta_i, \gamma_i) = \begin{cases} \Delta G_{i}, \eta_i = \gamma_i \\ 0, \ \eta_i \neq \gamma_i \end{cases} \quad \gamma_i \in \varsigma \tag{1.10}$$

Since we use the parameters, let (γ_i) pairing bases parameters to be 2 x 2 matrix of base letter presentations. And for every γ_i is associated with entropy and enthalpy and Gibbs free energy. These parameters are stored in properly as dictionary ς as shown in in equation 2.11.

$$\varsigma = \begin{pmatrix} \gamma_1 = \varrho_1, \Delta G°37 = x_1, \Delta H° = y_1, \Delta S° = z_1 \\ \vdots \\ \gamma_n = \varrho_n, \Delta G°37 = x_n, \Delta H° = y_n, \Delta S° = z_n \end{pmatrix} \tag{2.11}$$

Where γ_i represents the base pairings, and x_i, y_i, and z_i correspond to the Gibbs free energy, enthalpy, and entropy values at 37°C. It is worth noting that the parameters were obtained from Matthews [41] as they are only pairing bases were converted to matrix ϱ_i in this case for easier recognition.

$$\varrho_i = \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix} \tag{2.12}$$

The algorithm sums up the thermodynamic parameters associated with pairings and special cases for Gibb's energy, enthalpy, and entropy as shown in equation 2.13 to 2.15.

$$\Delta G^{o}_{tot} = \Delta G^{o}_{symmetry} + \Delta G^{o}_{intiation} + \sum \Delta G^{o}_{pairings} + 2 \times \Delta G^{o}_{start/end} \tag{2.13}$$

$$\Delta S^{o}_{tot} = \Delta S^{o}_{symmetry} + \Delta S^{o}_{intiation} + \sum \Delta S^{o}_{pairings} + 2 \times \Delta S^{o}_{start/end} \tag{2.14}$$

$$\Delta H^{o}_{tot} = \Delta H^{o}_{intiation} + \sum \Delta H^{o}_{pairings} + 2 \times \Delta H^{o}_{start/end} \tag{2.15}$$

Of course, the equations were generalised based on correction NNM by adding the terminals ($\Delta G^{o}_{start/end}$, $\Delta H^{o}_{start/end}$, $\Delta S^{o}_{start/end}$) on last work of David [341]. These end terminals parameters are computed in similarly to the pairing parameters since they are sensitive to the last base pair composition. $\Delta G^{o}_{symmetry}$, $\Delta H^{o}_{symmetry}$, and $\Delta S^{o}_{symmetry}$ are constants that introduce if the complementary of the parent sequences is itself reversed. Those parameters

are taken further to calculate melting point of RNA helix as expressed in equation 2.16. Where $[NA]$ is the concentration of nucleic acid [341].

$$T_m = \frac{\Delta H^o_{tot}}{(\frac{\Delta H^o_{tot} - \Delta G^o_{tot}}{310.15} + R\ \frac{[NA]}{1})} - 273.15 \qquad (2.16)$$

2.5.4 Simple Mutation Optimization Interactions Based Method (SMO-IBM)

This algorithm implemented right after interactions predictions are performed. First it starts with classification of aptamers sequences based on the interaction energy (E_{in}). The classification is categorized in five (best, good, better, poor and rest). This classification method follows equation 2.17.

$$S = \begin{cases} Best\ , & E_{in} < -8 \\ Good\ , & -8 < E_{in} < -6 \\ better, & -6 < E_{in} < -4 \\ poor, & -4 < E_{in} < -2 \\ rest, & E_{in} > -2 \end{cases} \qquad (2.17)$$

Where S is a given sequence with the interaction E_{in}. Once the classification of this aptamers is made, they are all put in list based on their classification category. Within the best interaction list of aptamers, we further look for the most occurring 4 bases in a sequence using the N-gram approach. We map and iterate through each sequence look for a union of for bases next to each other as expressed in equation 2.18.

$$\varphi = \bigcup_{i=1}^{n} \bigcup_{j=0}^{N_i-4} S_i\,[j{:}\,j+4\,] \qquad (2.18)$$

Where n is the length of the list of best aptamers, N_i is the length of the sequence S_i, which is any aptamer in an $Best$ aptamers list. $S_i\,[j{:}\,j+4\,]$ denotes a 4-gram extracted from the *i*-th aptamer, starting at index *j* and ending at index $j+4$. The φ returns list of the 4-n gram bases. The most frequent n-gram bases determined using the equation expressed as equation 2.19.

$$\Im = sd(\mathcal{L})[:k] \qquad (2.19)$$

$\Im$ is the function that sort and select the top common 4-gram base on given threshold (k) of how many to be retrieved based on the which 4-gram bases occurs the most. Where sd denote

sort in descending and $\mathcal{L}$ is counter method that count those n-gram bases as expressed in equation 2.20.

$$\mathcal{L} = counter(\varphi) \quad (2.20)$$

Once the n-gram bases are obtained, the best interacting sequence, is taken and mutated using those ten most occurring n-gram bases. The simplified mutation equations are given below from 2.21 to 2.25.

$$S'_{in} = S[1 : i - 1] + Y + S[i :] \quad (2.21)$$

$$S'_{ad} = Y + S \quad (2.22)$$

$$S'_{ms} = S[1 : i - 1] + Y + S[i + |X| :] \quad (2.23)$$

$$S'_{ts5} = Y + S[1 + |X| :] \quad (2.24)$$

$$S'_{ts3} = S[1 : i - 1] + Y \quad (2.25)$$

Let X be the bases to be replaced for every substitution mutation, where S' is mutated sequence, and S is the original best interacting sequence against the target RNA. Moreover, i is the position where mutation is occurring and $Y \in \Im$ which are the top ten n-gram bases. For insertion mutation (S'_{in}) equation is expressed in 2.21, addition mutation (S'_{ad}) is 2.22, and for the middle substitution (S'_{ms}) the equation is represented in 2.23. Lastly the terminal substitution mutations for 5' (S'_{ts5}) and 3' (S'_{ts3}) are expressed in equation 2.24 and 2.25 respectively.

2.5.5 Post Docking Analysis (PDA)

For the establishment of the PDA statical methods where employed to help with the detailed analysis of the results obtained from the molecular docking which made use of a blind docking approach. For every ligand (RNA/DNA aptamer) docked against the target there are 100 output PDB files generated as models, from model 1 to model 100. Those models are ranked based on docking scores which is generated based on shape-based pairwise scoring function [299]. The mean of those scores is calculated using equation 2.26.

$$FQ = \frac{1}{M}\sum_{i=1}^{M} G_i \quad (2.26)$$

Where G_i the docking score and M is the number of the models. The Mean FQ is used to calculate the Zscore in equation 24 which measures the deviation of the model 1 form the rest of the hundreds model.

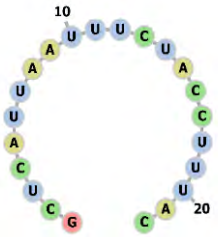
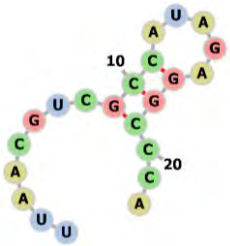
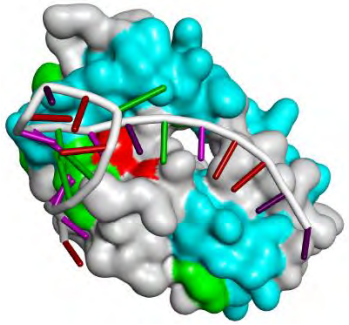
$$Z_{score} = \frac{G_{min} - FQ}{\sigma} \qquad (2.27)$$

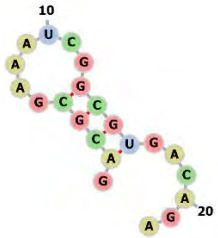
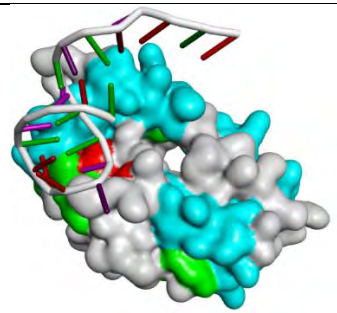
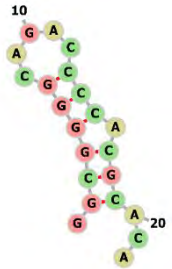
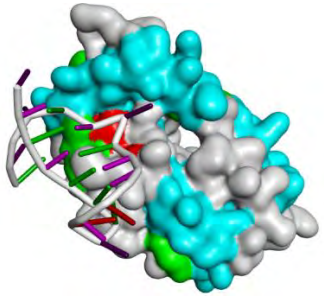
Where G_{min} is the model 1 docking score and σ is the standard deviation. The standardization and Z-score are related concepts used to bring data to a common scale, making it easier to compare and analyse. This normalization of data is crucial for accurate analysis and models performance.

2.6 Application and case study

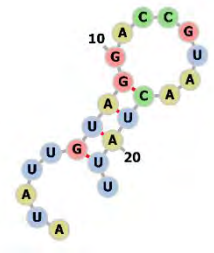
To evaluate the effectiveness of the T_SELEX program, over 1100 sequences with length of 22 nt were generated randomly using BRA approach. Generating random sequences can be susceptible to repeats but as mentioned before, the BRA algorithm does not allow any repeats. Crucially following step 2, secondary structures were generated. Where those sequences, together with their secondary structure, were further taken for predictions of 3D structures using ***fold_and_composition()*** and ***tertiary_structure()*** functions, respectively. The predicted structures were virtually docked against the HIV-1 protease (PDB ID: 3S09). It is worth noting that the protein was prepared externally prior screening by fixing bonds, removing the native ligand and water. The results are presented in **Figure 2.2**.

Figure 2.2 shows the sequences generated by T_SELEX, predicted secondary structures together with their MFE, tertiary structure and docked complexes (model1). **Figure 2.2** shows randomly selected aptamers without taking the ranking system into account, to illustrate the overall results. For example, aptamer tag (2) did not fold, and was not taken further for docking. As previously mentioned in step 6, if the AP_folded augment is set to True, unfolded aptamers will not be taken further for virtual screening. Although the (2) is unfolded, the bases are held together by stacking of the π regions of bases, hence the semi-circular shape. Secondary structure aptamer's stability relies on lower MFE and not on more pairings within the structure. Even though these aptamers show to have different tertiary structures they seem to have their model 1 unquestioningly docking at the same region of the HIV-1 protease as shown **Figure 2.2**. And the region showed to be very close to the active site where the native ligand was previously located. Based on the overall, results aptamer are promising inhibitors for HIV-1 protease.

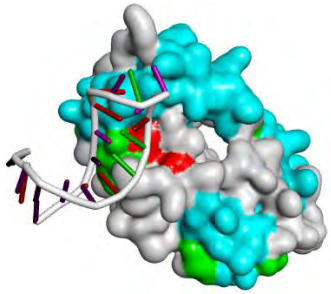
Sequences (tag)	Predicted secondary structures	Predicted tertiary structures	Docked complex (model_1)
GCUCAUUAA UUUCUACCUUUAC (2)	 ..... MFE = 0.00 kcal/mol		-ND
UUAACGUCGC CAUAGAGGCCCA (17)	 .....(((.....)))... MFE = -3.60 kcal/mol		 Docking score = -326.03 RSMD = 34.95 Å Zscore =5.564

GACGCGAAAU CGGCGUGACAGA (140)	 .((((.....)))))..... MFE = - 5.10 kcal/mol		 Docking score = -293.58 RSMD = 30.78 Å Zscore =3.654
UCGCGGGUAC ACCCGUCGUGUG (236)	 .((((((....)))))).)).... MFE = -7.2 kcal/mol		 Docking score = -334.82 RSMD = 35.45 Å Zscore =3.148

AUAUUGUAGGA
CCGUAACUAUU
(712)



.....((((.....))))).
MFE = -1,79kcal/mol



Docking score = -307.91
RSMD = 35.91 Å
Zscore =1.326

Figure 2.2: Simplified workflow of the T_SELEX algorithm, from aptamer dataset generation to detailed analysis post virtual screening calculations.

2.6.1 Ranking aptamers

By default, the T_SELEX program also provides you with the ranked docking results table, and in **Table 2.2**, we show a snippet of these results. All the ligands in this table are aptamers, and the docking results include docking scores, RMSD values, fitness quality, and confidence scores. It is important to note that the RMSD values are directly obtained from HDOCK results and should not be used as a ranking metric, as stated in the HDOCK publication [299]. Therefore, you may choose to omit these results when evaluating the aptamers. The fitness quality, on the other hand, is a calculated metric that accounts for the docking scores of over 100 models generated for each aptamer-target complex. These models represent different docking poses of the aptamer to the target, with varying scores that reflect the aptamer's binding position. The fitness quality provides a more comprehensive evaluation of the docking performance. The confidence score in molecular docking plays a crucial role in assessing the likelihood and reliability of the predicted interaction between a ligand and a protein. A higher confidence score indicates a higher degree of certainty that the predicted binding pose is accurate and reliable.

Looking at the aptamers in **Table 2.2**, aptamer41 has the lowest docking score of -411.73, indicating the strongest binding affinity. Following aptamer41, aptamerd383 has a docking score of -400.13, which is slightly higher (indicating a weaker interaction), and aptamer365 follows with a docking score of -393.44. Interestingly, the fitness quality values show a different trend. While aptamer41 has the highest fitness quality score of -268.33, indicating a strong docking pose, aptamer383 and aptamer365 have slightly lower fitness scores of -271.76 and -268.12, respectively, even though their docking scores are somewhat higher. This suggests that the fitness quality accounts for additional factors, such as the stability of the docking pose and the consistency of the interaction across the generated models, which may not always align directly with the raw docking score. Despite these variations, all aptamers show high confidence scores, suggesting reliable docking predictions across the board.

Table 2.2: Ranked Docking Results for aptamers against HIV-1 protease with Docking Scores, RMSD, Fitness Quality, and Confidence Scores.

Ligand	Docking score	RMSD	Fitness quality	Confidence Score (%)
aptamer41	-411.73	33.53	-268.33	99.47
aptamer383	-400.13	44.28	-271.76	99.33
aptamer365	-393.44	28.66	-268.12	99.24
aptamer225	-386.95	42.51	-274.04	99.13
aptamer496	-385.40	22.41	-258.71	99.11
aptamer466	-384.20	45.50	-276.55	99.08
aptamer705	-383.81	34.58	-254.81	99.07
aptamer918	-383.09	25.49	-250.40	99.06
aptamer142	-382.91	30.25	-264.82	99.06

In examining the docking poses of the two best-performing aptamers, aptamer41 and aptamer383, we can observe key structural features that highlight their binding interactions with the HIV-1 protease. In **Figure 2.3A1**, where aptamer41 is docked with the HIV-1 protease in its unmodified form, the single-stranded tail of the aptamer obstructs the active site of the protease, preventing access to the catalytic site. Meanwhile, the stem region of aptamer41 is positioned at the top of the protease, forming a stable interaction. For aptamer383, shown in **Figure 2.3B1**, the stem of the aptamer is closely bound to the active site of the HIV-1 protease, allowing for better potential interaction with the protease's catalytic region. The single-stranded tail of aptamer383, however, appears to be less engaged with the protease in this pose. It is important to note that these docking poses represent the first models of the best-docked complexes, and further refinement or additional docking models may be essential to review prior conclusions. However, here we provide an overview of the docking poses of the best-bound aptamers.

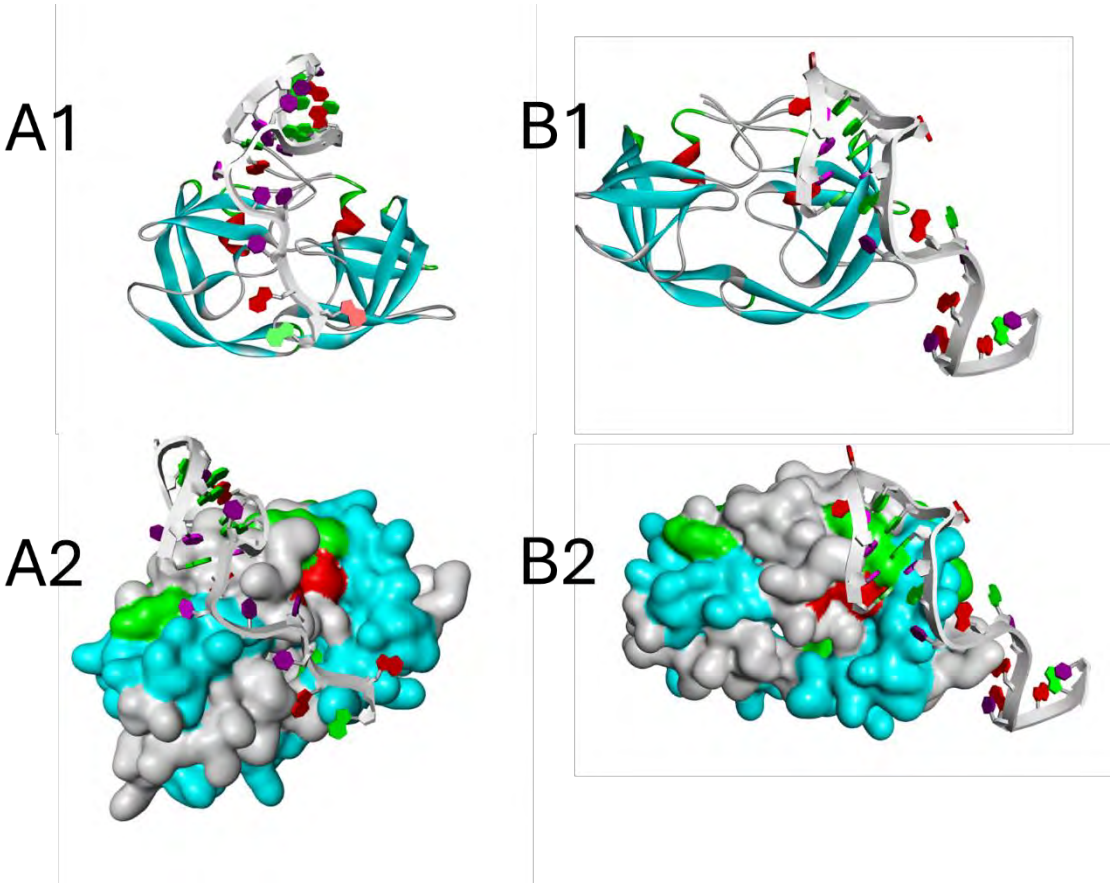


Figure 2.3: Docking Poses of the Best-Bound Aptamers with HIV-1 Protease where (A) is for apamer41 and (B) is for aptamer383. *Where A1 and B1 shows the aptamer-HIV-1 protease complexes where the HIV-1 protease not mashed and A2 and B2 shows the aptamer-HIV-1 protease complexes where the HIV protease is mashed.*

2.6.2 Files management

T_SELEX efficiently handles the organization and naming of files during the docking process. Upon initiating a virtual screening run, the program first creates a main "Docking" folder. Within this folder, it generates subdirectories based on the target name specified by the user in the input arguments. If the *Folded* argument is set to true, T_SELEX will create an additional folder named "folded" to accommodate the docking results of folded aptamers.

Once the folded directory is created, T_SELEX organizes the results by aptamer names. Each aptamer will have its own subdirectory within the "folded" folder. Within these aptamer-specific directories, the program generates and stores over 100 PDB files for each docking model corresponding to that particular aptamer and its interaction with the target. These PDB files represent the various docking poses and binding configurations for the aptamer-target complex.

For example, as shown in **Figure 2.4** and **Figure 2.5**, the folder structure created by T_SELEX is illustrated. **Figure 2.4** shows how the main Docking folder is created, followed by the target-specific directory, and, if applicable, the folded folder is generated based on the user's input. **Figure 2.5** shows how T_SELEX further organizes the results within the folded directory by aptamer name, creating individual folders for each aptamer, and stores the corresponding docking models as PDB files.

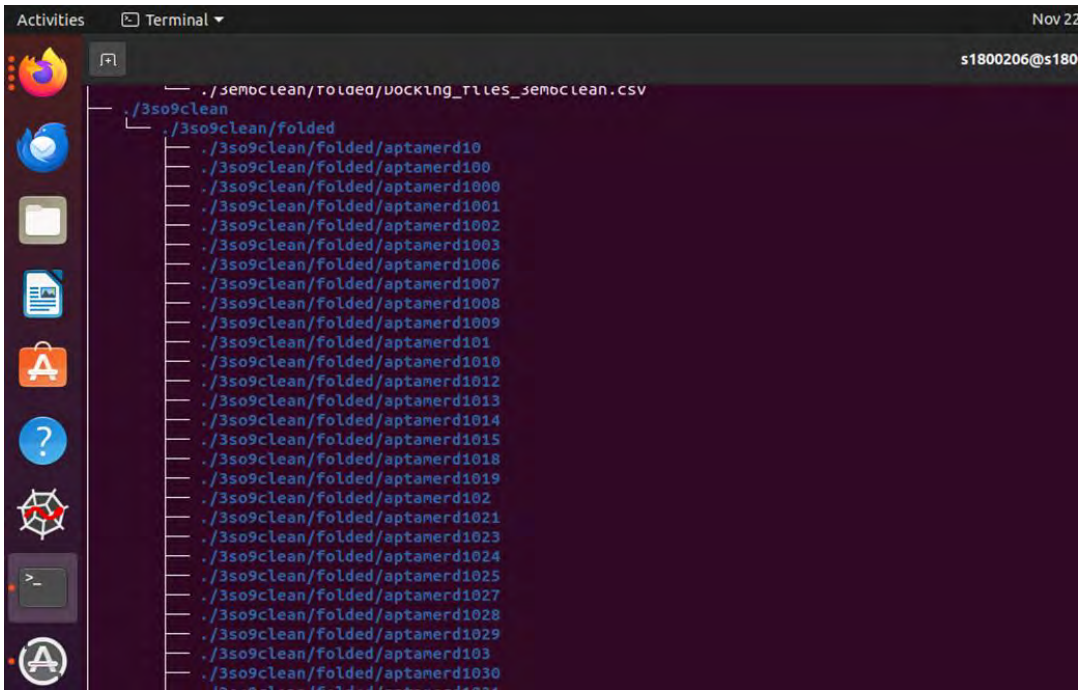


Figure 2.4: Folder Structure Generated by T_SELEX for Docking Results

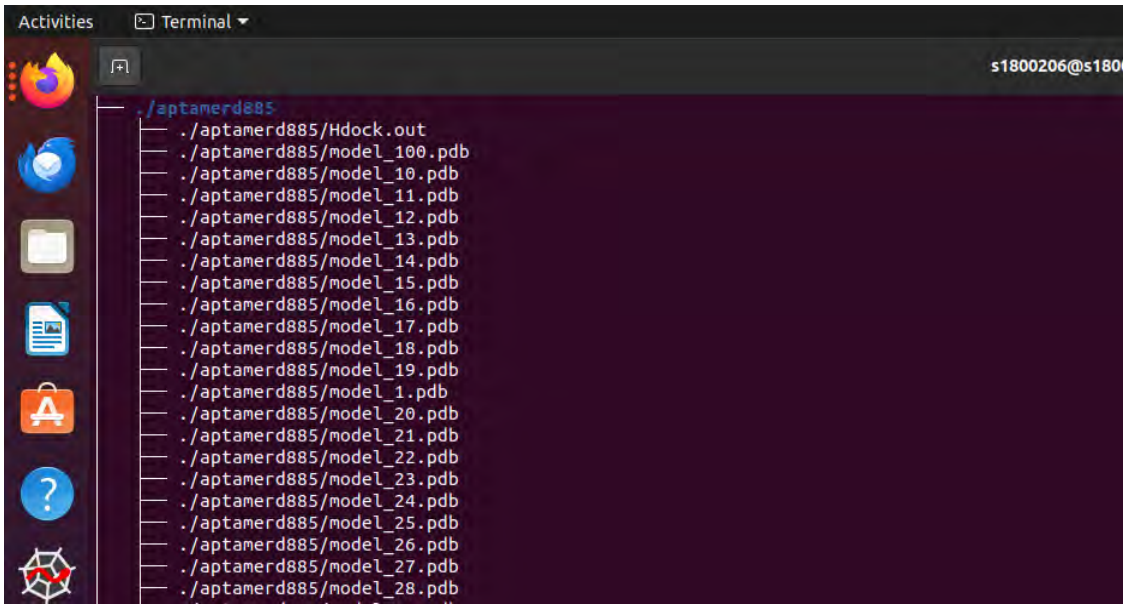


Figure 2.5: Organization of Docking Results for Individual Aptamers

Although we demonstrated this process using 1 100 aptamers, T_SELEX is designed to scale effectively for much larger virtual screening experiments. If the goal is to generate over 1 million aptamers, the framework can accommodate the increased computational load. In such cases, each docking calculation generates 100 output models per aptamer, meaning that screening 1 million aptamers will result in the creation of 100 million docking models. This substantial volume of data requires significant computational power and storage, but T_SELEX efficiently handles it. The program can analyse the docking results, providing an overall summary that includes fitness scores and other key parameters. Furthermore, its automated file management system ensures that all docking results are systematically organized and easily accessible for further analysis, effectively streamlining the management of large datasets generated during virtual screening simulations.

2.7 Conclusion

This chapter introduces the T_SELEX program developed in our research group for macromolecular high throughput virtual screening, more specifically for aptamers screening. The modules, algorithms, methods and performance of this package are discussed in details under the section Algorithm. This programme is aimed at reducing the time and financial cost associated with experimental macromolecular research, especially for SELEX experiments. It provides an automated computational approach for creating a large virtual dataset of aptamers and performing high throughput virtual screening on these large datasets against other macromolecular targets such as proteins and nucleic acids. This limits the number of macromolecules that need to be synthesised and tested for biological activities experimentally, as only those that show theoretical activity can be taken further for experimental work. ***(This work is submitted to journal of Cheminformatics)***

Chapter 3

T_SELEX Application 1: Dataset Generation and Sequence composition evaluation.

3.1 Overview

Understanding patterns within big data is crucial for making predictions based on observed data. Sequences are fundamental representations of nucleic acid compositions, offering insights into distinguishing between RNA and DNA molecules. In this chapter, we explore the sequence composition of a dataset generated by the T_SELEX program, which will serve as a foundation for subsequent analyses. Using straightforward statistical methods, we delve into the composition of sequences to uncover patterns and trends. Our analysis focuses on examining the composition of uracil (Us), guanine (Gs), cytosine (Cs), and adenine (As) within the dataset. We present the results through line graphs/noise plots and count plots to illustrate the distribution of nucleotides. Additionally, empirical cumulative distribution plots are used to further analyse the distribution of nucleotide compositions. Furthermore, we conduct a one-way ANOVA test to validate the assumption of similar means for Us, Gs, Cs, and As. The results confirm that the means of nucleotide compositions are similar, providing valuable insights into the dataset's characteristics. Finally, we perform pair composition analysis to explore the distribution of base pair compositions. Overall, our study provides comprehensive insights into the sequence composition of the T_SELEX-generated dataset, laying the groundwork for further analyses and predictions based on observed data.

3.2 Methodology

A dataset of 1100 sequences each comprising of 21 nucleotides was generated using the T_SELEX program. This was accomplished by importing several functions from the T_SELEX program. After generating the dataset using the ***gen_aptamers()*** function from the T_SELEX.secondary module, the returned list was passed through a function to calculate the composition of each individual nucleotide. This function, ***fold_and_composition()***, takes the list of aptamers as input and computes the composition. Here is the snippet of code:

```
from T_SELEX import gen_aptamers, fold_and_composition

p = gen_aptamers(seed=1, length=21, aptamers_num=1100)
comp = fold_and_composition(aptamers_list=p)
```

3.3 Sequence composition for T_SELEX generated aptamers

3.3.1 Composition analysis

It is essential to highlight the composition of aptamers in the dataset since this theoretical aptamer library was generated "randomly", implying that the aptamers have random numbers of different nucleotides. Unlike aptamers generated using combinatorial synthesis methods, this adds complexity to the study since everything is randomized or pseudo-randomized, with no assumed relationships [342]. On the bright side, this randomization increases the chances of identifying better inhibitors, as it selects 1100 aptamers randomly from 21 factorial (21!) possible sequences, given that all aptamers are 21 nucleotides long. The base composition of RNA molecules plays a significant role in their stability and folding; for instance, G-rich RNAs are associated with high stability [343]. Therefore, this section focuses on unravelling the base composition of aptamer sequences in the dataset, a detail often overlooked. The base compositions for all aptamers generated using the T_SELEX program were evaluated and reported in noise plots in **Figure 3.1.**

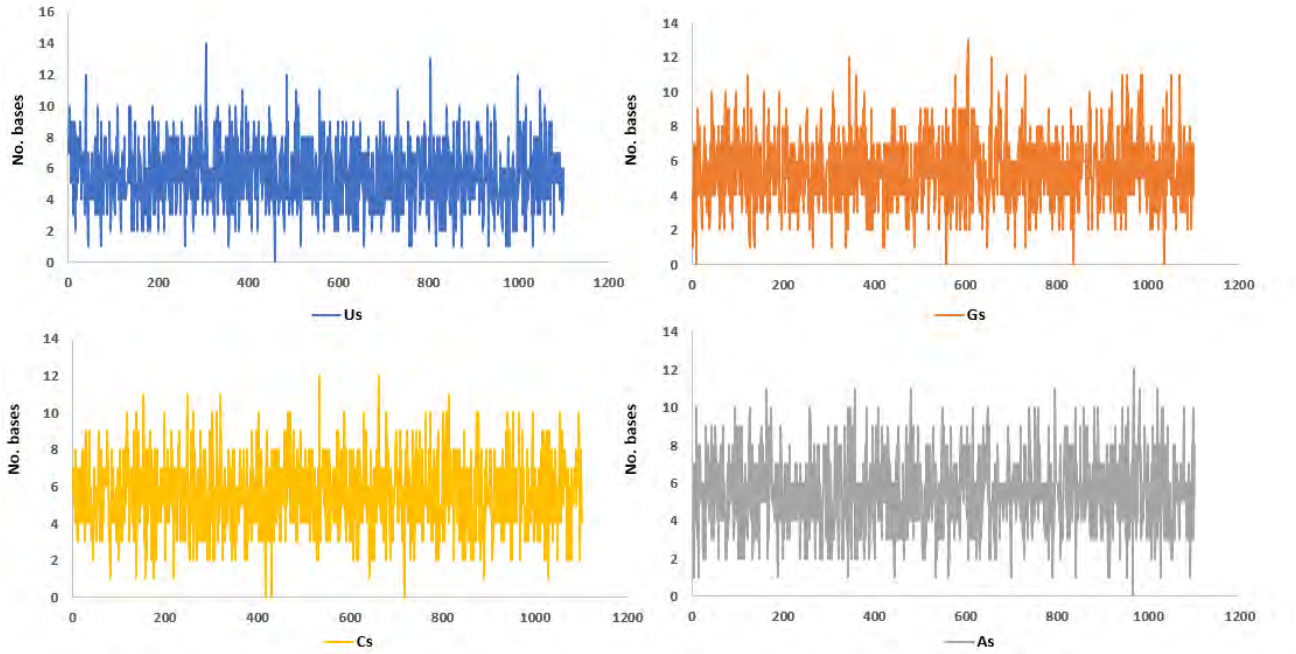


Figure 3.1: Base composition for theoretical generated aptamers library.

The four plots in **Figure 3.1** represents the composition of each nucleotide present in the RNA aptamer molecules. All aptamers were named using enumeration methods. For example, first aptamer sequence was given the name "aptamer 1", second aptamer sequence "aptamer 2", hundredth aptamer sequence "aptamer 100" and so forth until the one thousand one hundred aptamer sequence whose name is "aptamer 1100". In **Figure 3.1**, the x-axis denoted or represented aptamers in numbers, where each number represent an aptamer. The Us, Gs, As, and Cs on the axis labels indicate the nucleotide of interest (Uracil, Guanine, Adenine and Cytosine respectively).

- ***Us***

Looking at the uracil (U) composition plot (blue), most aptamers have a uracil composition between 2 and 10. A closer examination reveals that about 15 aptamers contain only one uracil in their sequences, and more interestingly, only one aptamer has no uracil at all, which is aptamer 458 (3' GGCGGGGCAGACCCCACGCACA 5'). The sequence of aptamer 458 suggests a more stable secondary structure, as it is rich in guanine (G) and cytosine (C). Shifting focus momentarily, the U composition plot shows that the highest uracil content detected in an aptamer is 14, observed in aptamer 305 (3' UUUUUUUAUUAUCCUCUAAUGU 5'). This implies that 63.63% of aptamer 305 sequence is composed of uracil. Notably, aptamer 305 is not only U-rich but also has most of the Us adjacent to each other, with only four adenines (A) positioned where it may be challenging for them to form stable A-U pairings necessary for

RNA folding. This could result in a very unstable folded structure or no folding at all when using Zuker's algorithm-based tools for secondary structure predictions. Zuker's algorithms compute two different energies for each subsequence S_ij (nucleotides from i to j) for any given RNA sequence, where all pairs (i, j) satisfy $1 < i < j < N$ [344,345]. The second U-rich aptamer, based on **Figure3.1**, is aptamer 804 (3' AUAUUUAUUGUACUUUCUGUCCAUUGUUUUCUUUUUCAUCAAGU 5'), which has 13 Us (59.09% of its composition), followed by aptamer 38 (3' AUAUUUAUUGUACUUUCUGUCC 5'), aptamer 485 (3' UCAUCUGUCUAUGUUUGAGUUU 5'), and aptamer 997 (3' UUUUACUUUAGGUCUUGACUUG 5'), each with 12 Us (54.54% of their composition).

- ***Gs***

Closer analysis of the guanine (G) composition plot (orange) in **Figure 3.1** reveals that most aptamers have a guanine composition ranging from 3 to 7. Interestingly, the plots also show that there are four aptamer sequences in the dataset that do not contain guanine nucleotides. These aptamers are aptamer 9 (3' AUAAUUACCCAUUACAUAUAAU 5'), aptamer 558 (3' CAUUUUUUACCAUUCUACUACU 5'), aptamer 837 (3' CUUUCUAUCUCCAAUUACCAAC 5'), and aptamer 1035 (3' ACACAUAACCCCUUUCUUUUCA 5'). Naturally, doubts may arise about the possibility of RNA sequences completely lacking guanines. The simple answer is that it is not naturally possible for RNA molecules to lack guanine entirely since guanine plays a crucial role in folding. However, since we are dealing with aptamers, which are synthetic RNAs, it may be possible in the future to intentionally design aptamers without guanine, although few have been reported or shown to exist thus far [346]. A pertinent question would be why one would design an aptamer without guanine, given its important role in the stability and folding of RNA molecules [347-349]. One reason could be to generate unfolded aptamers. This leads to further questions such as, “Why non-folding aptamers? For what applications?” Hopefully, more answers will become relevantly available to address these issues as the questions continue to emerge.

It would be misleading to ignore the contribution of guanine and G-C pairing to molecular stability. However, in this study, we will focus on designing theoretical aptamers that could be used therapeutically and understanding the roles and differences of using both unfolded and

folded aptamers. Returning to the G composition plot, it shows that about 12 aptamers have only one guanine. On the other hand, only one aptamer in the dataset has 13 guanines, which is aptamer 605 (3' GUGGGGAGCGGCGCGGGGUUAC 5'). This aptamer is likely to fold better despite having fewer cytosines (Cs) in its sequence when using Zuker's algorithm. Zuker's algorithm acknowledges G-C, A-U, and G-U pairing as admissible base pairing [345,347]. Due to the nature of folding, G-U is a wobble pairing that plays a fundamental role in secondary structure and is found in almost every class of RNA [348,349]. G-U pairs are comparable to G-C and A-U Watson-Crick base pairs since there are two hydrogen bonds between G and U, similar to A-U, except for the orientation of the bases with respect to the phosphodiester backbone, which is easily observed through the glycosidic angle [350]. In the G composition plot, aptamer 345 (3' GUAGGUGGUAGGCGGGGUCAGU '5) and aptamer 657 (3' GGCGAGGCGGCGGAGCUAAGGA '5) are the only aptamers with 12 guanines. They can be expected to behave similarly to aptamer 605, though not exactly the same due to the different positions of the bases.

- ***Cs***

In **Figure 3.1**, the cytosine (C) composition plot (yellow) shows that three aptamers in the dataset have no cytosine. These aptamers are aptamer 416 (3' UUAAGAUUUAUGAGAAUGUGAA 5'), aptamer 430 (3' GAUAAUAGUUUAUGUGGAUGGA 5'), and aptamer 717 (3' GGGAGUUUGAAAGUAAUAGGUU 5'). The absence of Cs in an aptamer implies two significant points: (1) there will be no G-C pairing, and (2) the structure will be less stable [349]. While base stacking can contribute to RNA stability in real life, this has been well-demonstrated for DNA and RNA duplexes but perhaps not yet clearly reported for single RNA strands [351,352]. Logically, the stability of RNA depends on the folding capabilities of the sequence, which are highly dependent on G-C, A-U, and G-U pairings, with G-C being more stable than A-U pairing. The main reason is that G-C pairs are held together by three hydrogen bonds, while A-U pairs are retained by two bonds, similar to G-U pairs. Therefore, the lack of G-C pairing resulting from a lack of C or G in the sequence may lead to less stable RNA.
Furthermore, the C composition plot in **Figure 3.1** illustrates that eight aptamers in the dataset have only one C in their sequence. It can be assumed that these aptamers are likely to be less stable, similar to those with no C at all. The rest of the C plot indicates that most of the aptamers in this library have Cs ranging from 2 to 8. Notably, only two aptamers have 12 Cs: aptamer

533 (3' CUCCCCGAUCGUCUGCCGUCCC 5') and aptamer 661 (3' CCACCGUCCUUUCCGCCAGCCA 5'). Both aptamer 533 and 661 are quite similar, and it can be assumed that they may not fold properly. Unlike guanine, cytosine only pairs with guanine, which implies that if there is more than 50% of C in the sequence, there are fewer Gs available to pair with for a more stable structure. The plot further shows that there are four aptamers with 11 Cs (50% of the sequence): aptamer 153, aptamer 247, aptamer 320, and aptamer 812.

- ***As***

Similarly, looking at the adenine (A) composition plot in **Figure 3.1**, it is evident that only one aptamer lacks adenine bases, which is aptamer 965 (3' GCUCGGUCCCUCCUUCGCUCGG 5'). From a theoretical perspective, aptamer 965 can still fold since G-C and G-U pairings can occur, as mentioned earlier based on the Zuker algorithm [345]. This suggests that the absence of adenine in the sequence may not significantly affect folding if Gs, Cs, and Us are evenly distributed. However, it would be misleading to imply that the absence of adenine in a non-coding RNA sequence is negligible. Adenines play significant biological roles, such as being highly associated with A-to-I RNA editing, where adenosine is converted to inosine by ADAR (adenosine deaminase acting on RNA) enzymes [353]. This editing can also be performed experimentally, and inosine has shown therapeutic potential in treating psychiatric and neurological disorders [354].

Further analysis of the adenine composition plot in **Figure 3.1** reveals that about 15 aptamers have only one A in their sequences. Most aptamers have adenine bases ranging from 2 to 9, which is average given their length of 22 nucleotides. On the other hand, aptamer 967 (3' CAUACAAGAAGGAUAGAAAAUU 5') has the highest number of adenines with 12 A bases. Logically, since adenine pairs only with uracil, this suggests that aptamer 967 is unlikely to fold effectively, as more than 50% of its sequence depends on uracil to form a stable folded structure. Aptamer 967 is followed by aptamers 163, 356, 479, 980, and 1020, each with 11 adenines. In terms of folding, these aptamers are likely to have similar minimum free energy (MFE) values, with differences due to the positions of adenine bases and other bases in the sequences.

3.2 Distribution analysis

To quantify the extent of base compositions and how they are distributed in the dataset, count plots were generated in **Figure 3.2**. Count plots are unique in presenting a brief descriptive distribution of one-dimensional categorical data. A key advantage of understanding the distribution of categorical data is observing the frequency or occurrence of numbers or objects in systems or data of interest. In this current study, as mentioned in the preceding section, the composition of each aptamer was evaluated based on how many Us, As, Gs, and Cs are present in the sequence. Assumptions were made based on the noise plots that the highest frequency for all bases was between 3 and 9, suggesting that most aptamers in the dataset had bases that ranged between 3 and 9. Here, we clarify the frequency of the bases and their occurrence by examining the distribution plots in **Figure 3.2**.

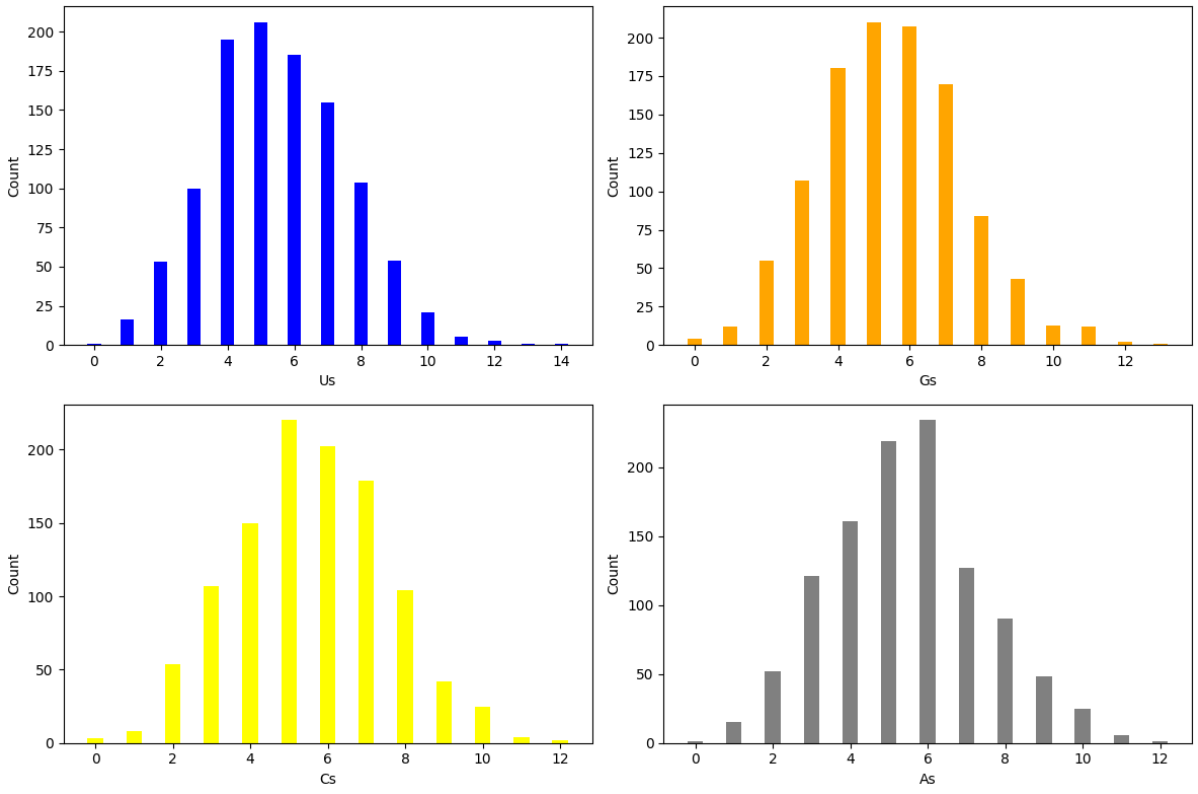


Figure 3.2 Distribution Count plots for base composition of the dataset

Figure 3.2 precisely illustrates that most aptamers in the dataset have uracil ranging from 3 to 8 in their sequences. More specifically, the Us plot shows that more than 200 aptamers have five uracils in their sequences. This is followed by 4 and 6, respectively, indicating that there are more aptamers with 4 uracils in their sequences compared to 6. The same pattern is applied to Gs in Figure 3.2. Through the inspection of the Gs plot, it is seen that most of the aptamers

in the library have guanine ranging from 3 to 9. Moreover, to be more specific, the plot suggests that most of the aptamers have 5 and 6 guanines, with 5 being the highest. This suggests that over 400 aptamers either have 5 or 6 guanines in their sequences, which is almost half of the dataset. Similarly, the Cs plot shows that 5 has the highest count and is above 200. For As, interestingly, 6 is the one with the highest count, followed by 5. It is worth noting that both 6 and 5 are above 200, implying that over 400 aptamer sequences contain either 5 or 6 adenines.

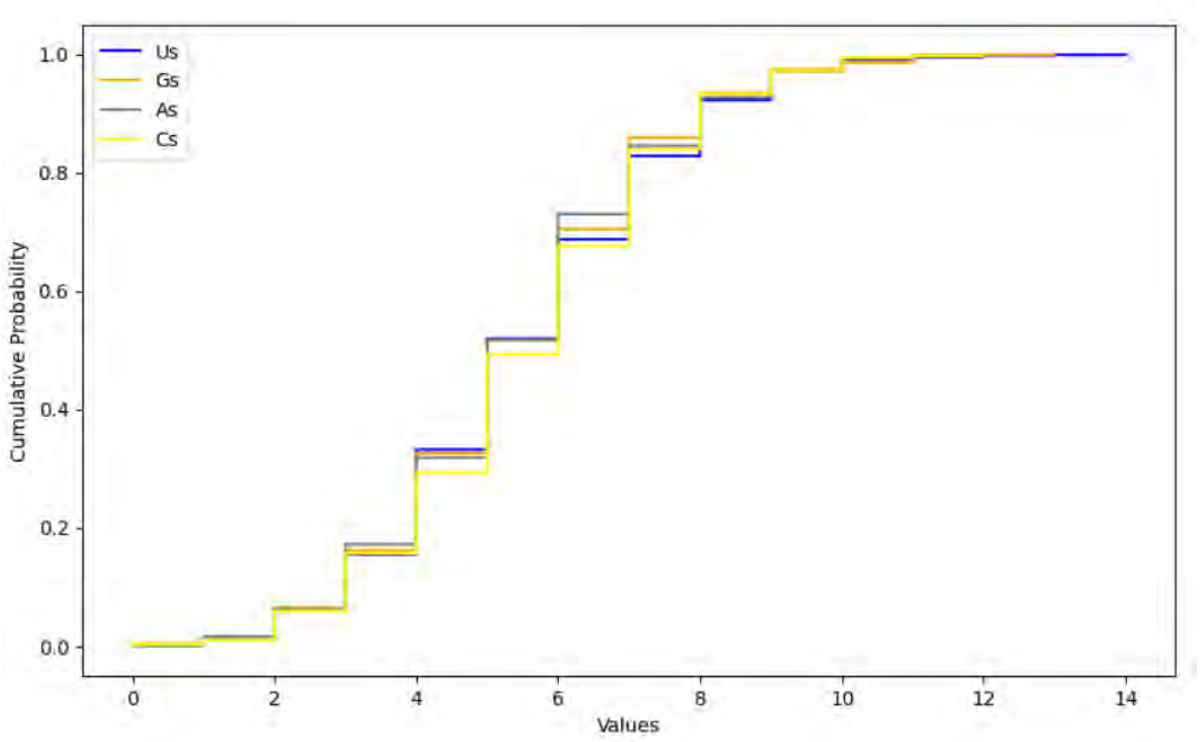


Figure 3.3: Empirical cumulative distribution plots for composition of Us, Gs, As and Cs in the data set.

For more careful examination, the empirical distribution function was applied to the data since, for this instance, it is also worth highlighting that we are dealing with numerical data even though it was treated as categorical data previously. Empirical cumulative distribution functions (eCDF) are step functions which make $1/n$ jumps for n data points [355]. The empirical cumulative plots for Us, Gs, As and Cs are presented in **Figure 3.3**. The plots suggest that for Cs and As the highest composition for sequence twelve, while for Gs is 13 and Us is 14. These plots suggest that there is similarity in the distribution of the bases throughout the dataset even though the random algorithm was used to generate these aptamer sequences.

3.3.2 One-way ANOVA test

In order to validate the assumptions that "Us, Gs, As and Cs have similar means", one-way ANOVA test was performed. Analysis of variance (ANOVA) is a statistical method or approach that is used to compare the mean of the several samples. It is important to note that ANOVA is mostly used to determine the means of the two means independent samples or population [356,357]. Before performing ANOVA procedure this few assumptions must be met, 1) observations are independent from each other, 2) observations in each group must come from normal distribution and lastly, 3) population variance in each group are the same [358,359].

Before going deeply into analysing results, first let us address the concern of dependency or independency of the samples analysed in the studies. Our samples are bases or nucleotides composition and they were generated "randomly" and independent from each other, which simply suggests that they do not follow any simple trend. But there is misconception that since length is fixed this means that the base composition might be dependent on each other, which is not entirely true. If nucleotide bases can be thought of as simple characters that are just simply placed randomly next to each other until certain number of them are placed together. As much as the length of bases may end up playing role in the limit and the probability of certain bases present in a sequence (1/4 multiply by 22) this does not mean that the existence of these bases in a sequence are entirely dependent on each other. Based on those grounds we choose to neglect the dependency of this bases/samples. In one-way ANOVA two hypotheses are evaluated, the hypothesis tested, number one is the null hypothesis [360]:

$$H_0: \mu_{Us} = \mu_{As} = \mu_{Cs} = \mu_{Gs} \qquad (3.1)$$

In null hypothesis we assume the means of As, Us, Gs, and Cs (represented by $\mu_{As}, \mu_{Us}, \mu_{Gs}$ and μ_{cs} respectively) are equal as shown in the equation 3.1. The alternative hypothesis assumes that at least one pair has unequal means [360]:

$$H_1: \mu_i \neq \mu_j \qquad (3.2)$$

Where i, j is equal to As, Us, Gs, and Cs. When looking at the ANOVA test the most important values are F, Critical value and p value.

The summary in **Table 3.1** shows uracil has slightly higher variance value compared to the rest of the bases. This suggest that the Us composition values of the sequences are little spread out

from the mean compared the rest of the bases. Even though Us showed to have high variances, Cs showed to have higher Sum value. Table 3.1 shows that the null hypothesis holds since F critical > F static which implies that the H_0 hold and H_1 should be rejected. Furthermore, p value is greater than the alpha level which is 0.05 or 5% and this emphasize that H_1 is rejected. Prompt by these results, is safe to conclude that Us, As, Gs, and Cs have similar mean, which is approximately 5.

Table 3.1: Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Us	1100	6073	5,520909	4,237051
Gs	1100	6003	5,457273	4,017281
As	1100	5994	5,449091	3,972838
Cs	1100	6130	5,572727	3,911903

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	11,15818	3	3,719394	0,921836	0,429302	2,606929
Within Groups	17736,84	4396	4,034768			
Total	17748	4399				

3.3.3 Pair composition analysis

Pair compositions was performed using the pair composition algorithm from T_SELEX. The algorithm uses a similar approach as that from Thermodynamics Nearest Neighbour model [341]. Instead of calculating percentage composition of an individual base, it calculates for pairs of bases. This approach finds all possible ways RNA bases can be paired next to each other and calculates the composition in each sequence and throughout the data set. The Violins plots in **Figure 3.4** shows the distribution of pair base compositions. Similar to the individual base composition, there is homogeneity of the distribution for almost all pair compositions. For instance, plot AU to CG have same minimum value, first quartile, median, third quartile and maximum value. Their median value showed to be very close to first quartile, this suggest that the data is positively skewed. Therefore, the base pairs compositions have the minimum value

of zero which suggests that there are aptamers that lack either of the base pair compositions. There is a minor difference observed from UU to AA as compared to AU to CG. Inherently, the minimum values in UU to AA are equal to the first quantile, which implies that there is high frequency of minimum value. Even though there is a minor difference, the mean turns to remain the same throughout the plot, from AU to AA.

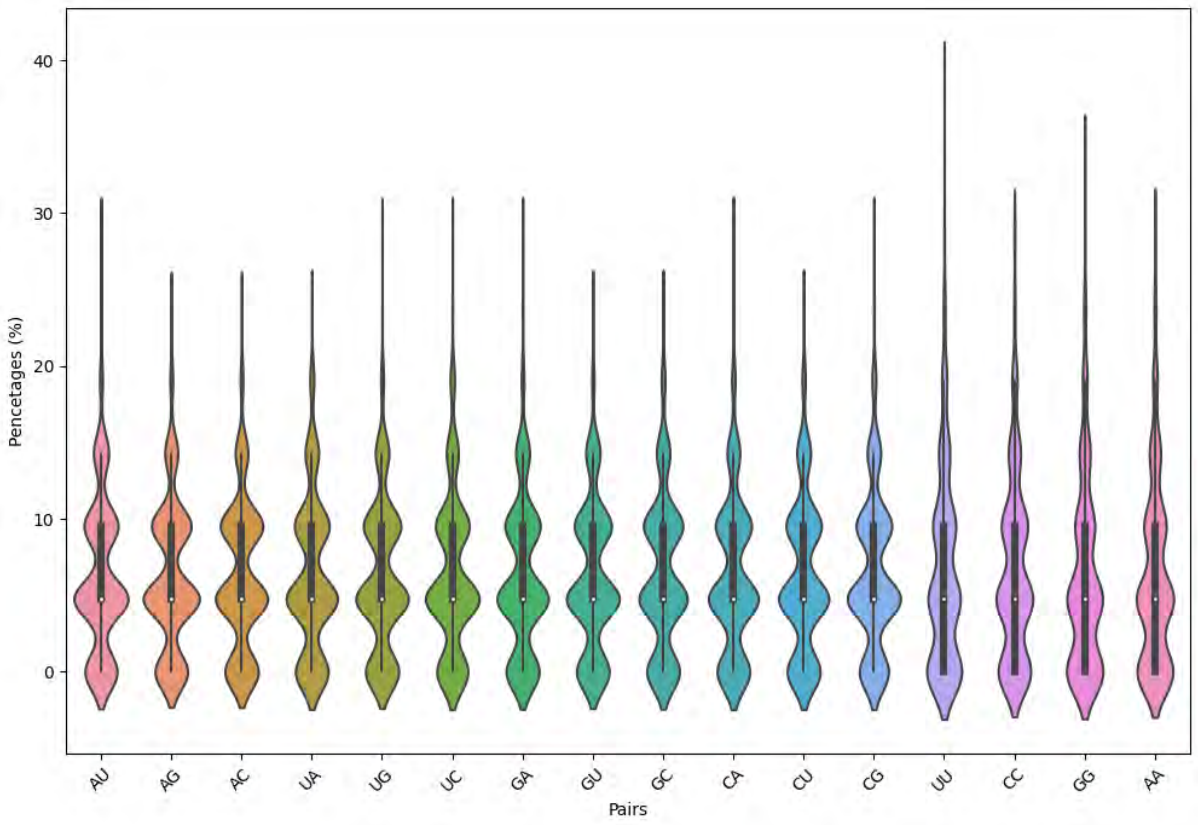


Figure 3.4 Violin plots for the distribution of the adjacent base compositions.

3.4 Conclusion

In this chapter, we demonstrated the analysis of a theoretical aptamer library generated using the T_SELEX program. The dataset comprised 1100 “randomly” generated aptamer sequences, each with a length of 22 nucleotides. Despite the randomization, the analysis aimed to uncover patterns in the base composition of the aptamer sequences, which play crucial roles in stability and folding. Quantitative assessment through count plots offered a descriptive overview of nucleotide distributions, while empirical cumulative distribution plots provided a deeper insight into distribution patterns. The one-way ANOVA test supported the assumption of similar means among nucleotide compositions, reinforcing the observed distribution trends. Pair composition analysis revealed further insights into the distribution of base pair compositions, with violin plots highlighting homogeneity in distributions and frequency analysis elucidating prevalent compositions and variations.

Chapter 4

T_SELEX Application 2: Large scale secondary structure and tertiary structure prediction

4.1 Overview

This chapter evaluates the importance of RNA secondary and tertiary structures, shedding light on aptamer folding dynamics. Through computational modelling, the investigation is focused on the role of pseudoknots which guides the RNA folding and emphasizing their crucial contribution to the secondary and tertiary structural levels. The analysis aims to highlight the significance of stem-loop motifs in RNA folding and stability, determining the influence of sequence composition on structural dynamics.

4.2 Methodology

Following sequence generation in **Chapter 3**, ***fold_and_composition()*** function was employed from T_SELEX program which make use of RNAfold algorithm [265] to predict the Minimum Free Energy (MFE) secondary structures of the randomly generated RNA sequences. This step is essential as it enables the identification of stable conformations based on thermodynamic properties. The comprehensive data frame output containing each RNA sequence along with its corresponding MFE structure was analysed. Thereafter, the comprehensive data frame was passed through the ***tertiary_structures()*** function from the T_SELEX program. The ***tertiary_structure()*** function makes use of RNAComposer web server [337] for highly accurate 3d structures prediction. This made it possible to generate 3D models of the RNA molecules at large scale. A detailed explanation of these T_SELEX program functions is available in Chapter 2.

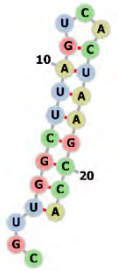
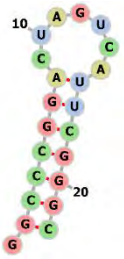
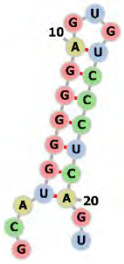
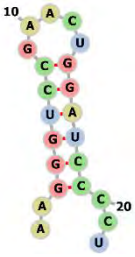
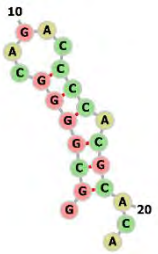
4.3. Results and discussion

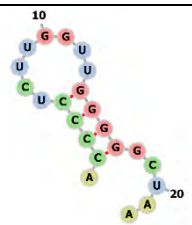
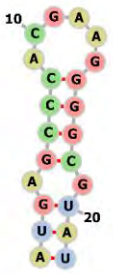
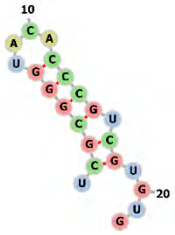
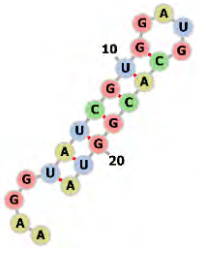
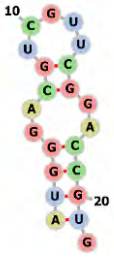
4.3.1 Secondary structures analysis

The library of RNA sequences was passed to ***fold_and_composition()***, which looped through the RNAfold algorithm [265] to obtain secondary structures or pseudoknots. The results were analyzed by looking at the pseudoknots. As expected, most of the aptamers from the dataset folded very well with most folded aptamer having -9.7 kcal/mol. The most stable folded aptamer sequences together with secondary structures are reported in **Table 4.1**. The 2D structures we constructed using FORNA software [361]. The 2D structures were colour-coded based on their nucleotides/sequence, which is the default setting. Guanine (G) is represented with red, Adenine (A) => yellow, Uracil(U) => blue, and Cytosine is represented with green colour. It can be observed that as the MFE decreases, indicating higher stability, the secondary structures tend to become more compact. The compactness of the secondary structures suggests that there are more pairings [362-364]. For instance, in Aptamer1084 with an MFE of -9.5 kcal/mol, the secondary structure is represented as '...(((((((...)))))))', indicating a highly compact structure with extensive base pairing the brackets "(" and ")" represent pairing regions}. Conversely, in aptamer612 with an MFE of -7.2 kcal/mol, the secondary structure is represented as '(((((..((.....))..)))))', showing a less compact structure with more unpaired regions. While Minimum Free Energy (MFE) serves as a useful metric for assessing the stability of RNA secondary structures, it does not always correlate directly with the complexity or compactness of those structures. Take aptamer236, aptamer522, and aptamer612, for example, they all with an MFE of -7.2 kcal/mol. Despite sharing the same MFE value, their secondary structures exhibit varying degrees of pairing and compactness. Aptamer236 and aptamer522 display moderately stable structures with easily recognisable unpaired regions, suggesting flexibility and potential structural rearrangements. Meanwhile, aptamer612 structure features extensive unpaired regions, indicating a less compact conformation despite its similar MFE to aptamer236 and aptamer522. Detailed evaluation their secondary structures reveals variations difference in loop sizes and compositions, which can be influenced by the base composition and position. This can be seen in Aptamer1084, aptamer950, and Aptamer79 which have similar minimum free energies (MFEs) but distinct loop properties. Aptamer1084 have a relatively compact secondary structure with minimal loop regions, indicating a stable fold secondary structure. In contrast, aptamer950 contains a larger loop region which potentially allows flexibility and stronger binding. Aptamer79 shows to exhibit intermediate loop size, suggesting a good balance between structural stability and flexibility. Similarly, the same can

be said about aptamer236, aptamer522, and aptamer612 which possess an MFE of -7.2 kcal/mol, yet their secondary structures differ. Aptamer236 has a relatively compact structure with a small loop, while aptamer522 and aptamer612 feature larger loops. And these structural differences can potentially influence their binding specificity and affinity.

Table 4.1: Best folded aptamers 2D-structure with their sequences, MFE and secondary structure.

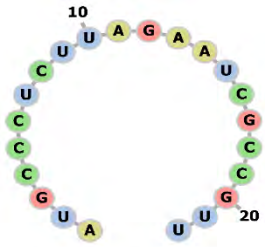
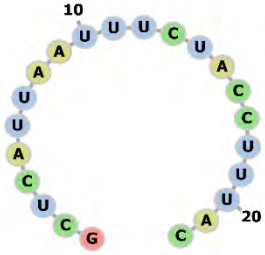
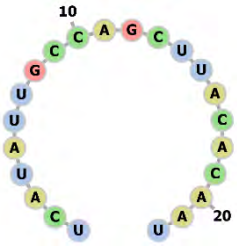
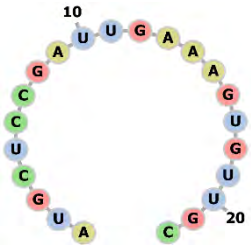
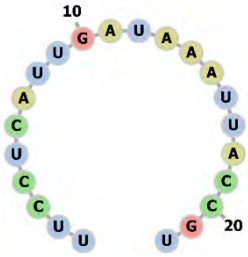
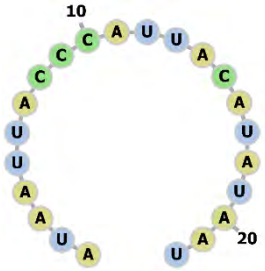
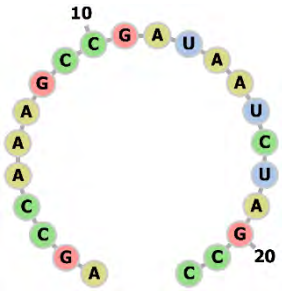
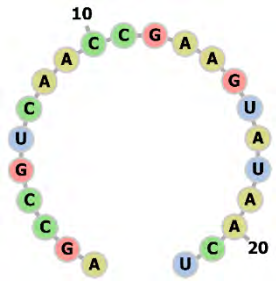
 Aptamer1084: MFE = -9.5 kcal/mol CGUUGGCUUAGUCACUAAGCCA Secondary structure: ...(((((((...)))))))	 Aptamer950: MFE = -9.3 kcal/mol GGCCCGGACUAGUCAUUCGGGC Secondary structure: .(((((((.....)))))))
 Aptamer79: MFE = -9.1 kcal/mol GCAUGGGGGAGUGUCCCUCAGU Secondary structure: ...(((((((...)))))))..	 Aptamer916: MFE = -8.7 kcal/mol AAGGGUCCGAACUGGAUCCCCU Secondary structure: ..((((((.....))))))...
 Aptamer166 : MFE = -8.5 kcal/mol UGUGUUGGGCGGCUCAGCACCC Secondary structure: .(((((((...)))))))..	 Aptamer458: MFE = -8.3 kcal/mol GGCGGGGCAGACCCCACGCACA Secondary structure: .((((((.....))).)))...

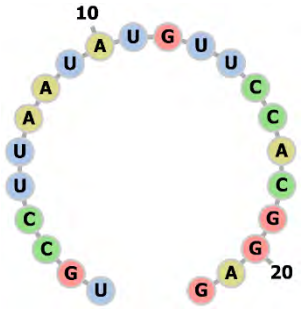
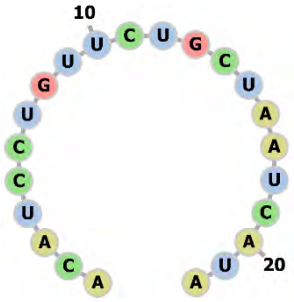
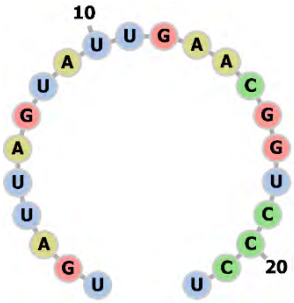
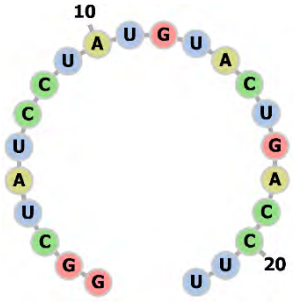
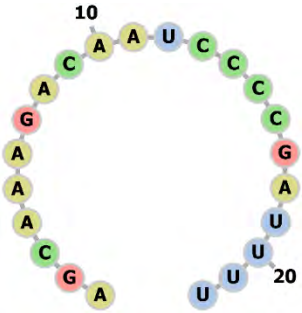
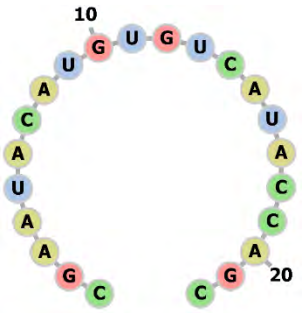
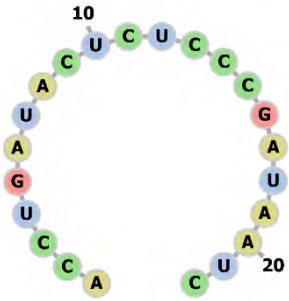
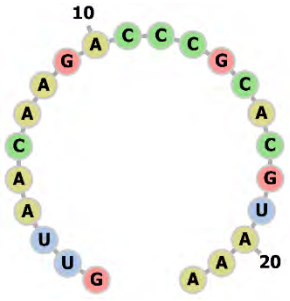
 Aptamer340: MFE = -8.1 kcal/mol ACCCCUCUUGGUUGGGGGCUAA Secondary structure: .((((.....))))....	 Aptamer1010 : MFE = -8.0 kcal/mol UCCAGGAAUCUCCUGGUUAUGU Secondary structure: .(((((((...)))))).....
 Aptamer509 : -7.8 kcal/mol AUGAGCCCACGAAGGGGCGUAU Secondary structure: (((.((((.....)))).)))	 Aptamer589: MFE = -7.4 kcal/mol CCCGAGGUUGUUACUCGGCAAU .(((((.....))))))....
 Aptamer1077 : MFE = -7.4 kcal/mol AACGGGGAUCGCAGUCGCCGCA Secondary structure: ..(((.(((....))).))))..	 Aptamer236 : MFE = -7.2 kcal/mol UCGCGGGUACACCCGUCGUGUG Secondary structure: .((((((....)))).)))....
 Aptamer522 : MFE = -7.2 kcal/mol AAGGUAUCGUGGAUGCACGGUA Secondary Structure: ....(((((((....)))))))	 Aptamer612: MFE = -7.2 kcal/mol AUGGGACGUCGUUCGGACCGUG Secondary Structure: ((((..((.....))..)))).

Although most of the aptamers did fold, some of the aptamers did not fold from the dataset, and this was indicated with an MFE of zero (MFE =0). This suggests that there is no pairing of the bases within the RNA molecule to form stable folded RNA, or the RNAfold [265] algorithms could not find the stable secondary structure based on the MFE approach [365]. Of course, this could be argued that the folding is also based on the algorithm used to fold the system. However, it would be premature to dismiss the folding done using RNAfold [265] or/and Mfold algorithm since they are widely used due to their accuracy. Even though these two algorithms were built upon Zuker-Stiegler algorithm of computing MFE structures using thermodynamic parameters and "nearest neighbour model", there are minor difference that were detected [366]. It was further demonstrated by Gardner and Giegerich that there is no significant difference when it comes to their accuracy [367]. It is very important to note that, for short RNA sequences, the status of equilibrium energy for folding RNA is close to MFE status, which makes the MFE based algorithm highly accurate for folding short RNA sequences [368]. For these reasons, RNAfold [265] or Mfold was best option to be used in this case to fold these aptamers since aptamers are short RNA sequences.

Back to the dataset analysis, 281 aptamers out of 1100 had MFE of zero which indicates that about 26% of the dataset did not fold. This further implies that those aptamers do not have any stems or loops which are highly associated with different functions which include binding properties [369]. For a deeper comprehensive evaluation, a few unfolded aptamers were selected from the dataset with their pseudoknots for 2D visualization. The 2D structure of the unfolded aptamers were reported in **Table 4.2**. Unsurprisingly, all those 16 2D structures illustrated in **Table 4.2** have exactly the same shape, and since they have the same length, this was expected. It is worth noting that all of the aptamers that did not fold have exactly unclosed circle loop-like shapes. But by looking at the sequences of these unfolded one may be convinced that these aptamers should have stable MFE structures. Since all unfolded aptamers seem to have 4 bases present in their sequences. According to Trotta, RNA molecules with less GC content in their sequences are more likely to result in MFE=0 or have no stable MFE structure [368]. In this case, that statement could be debatable since in the dataset, there are sequences that exhibit MFE of zero while having enough GC content for pairing, but nevertheless, it cannot be entirely ignored as some of the aptamers that did not fold showed less content of GC or none at all. It is also important to note that in this case length is also a contributing factor towards unfolded aptamers.

Table 4.2: Some of unfolded aptamers with 2D-structure together with their sequences.

 Aptamer1: AUGCCCUCUUAGAAUCGCCGUU	 Aptamer2: GCUCAUUAAUUUCUACCUUUAC
 Aptamer4: UCAUAUUGCCAGCUUACACAAU	 Aptamer5: AUGCUCCGAUUGAAAGUGUUGC
 Aptamer7: UUCCUCAUUGAUAAAUUACCGU	 Aptamer9: AUAAUUACCCAUUACAUAUAAU
 Aptamer13: AGCCAAAGCCGAUAAUCUAGCC	 Aptamer18: AGCCGUCAACCGAAGUAUAACU

 Aptamer19: UGCCUUAAUAUGUUCCACGGAG	 Aptamer21: ACAUCCUGUUCUGCUAAUCAUA
 Aptamer22: UGAUUAGUAUUGAACGGUCCCU	 Aptamer25: GGCUAUCCUAUGUACUGACCUU
 Aptamer34: AGCAAAGACAAUCCCCGAUUUU	 Aptamer49: CGAAUACAUGUGUCAUACCAGC
 Aptamer61: ACCUGAUACUCUCCCGAUAAUC	 Aptamer69: GUUAACAAGACCCGCACGUAAA

4.3.2 Tertiary structures analysis

Figure 4.1 illustrates examples of the most stable folded tertiary structures of the aptamers from the aptamer library. Different structural motifs within RNA molecules can be observed. Despite these structures being stably folded, the differences in the configuration of stems and loops are still clear. These variations in structure, including differences in position and orientation, shows the unique folding patterns dictated by the primary sequence of each RNA molecule. **Figure 4.1** shows that Aptamer1084 and Aptamer950 have different loop configurations in tertiary structures.

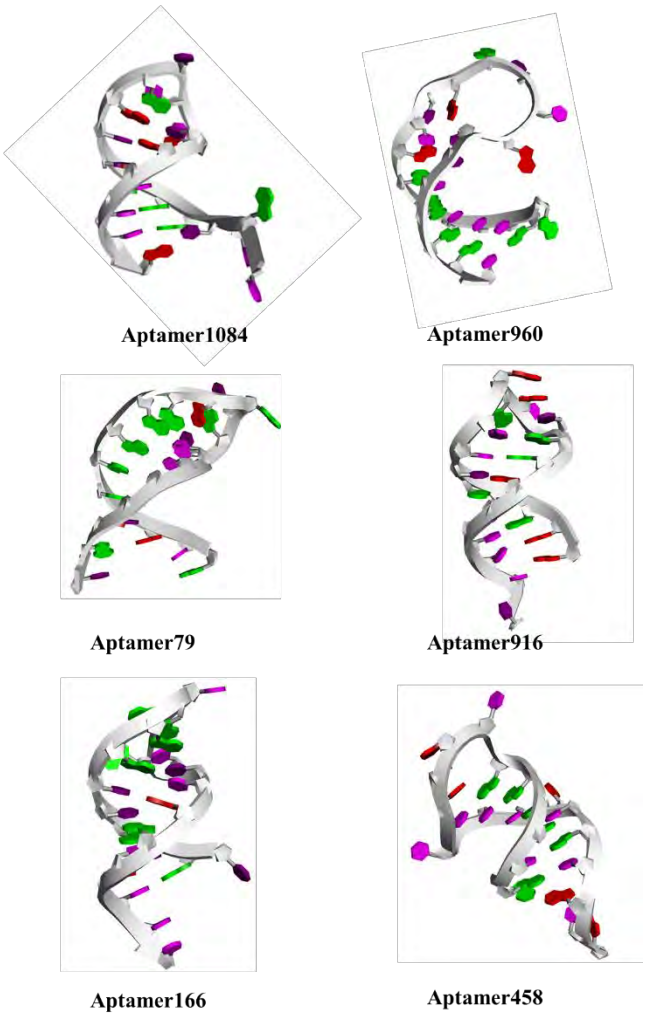


Figure 4.1: Tertiary structures of the best folded aptamers from the dataset.

The tertiary structure depicted in **Figure 4.2** shows a remarkable alignment with the secondary structures, emphasizing the important role played by secondary "pseudoknots" in guiding RNA folding. This alignment is not by mistake but rather a consequence of the complex folding

process orchestrated by pseudoknots, the absence of bulges or additional motifs in both structural levels show the control of pseudoknots in shaping the final folded conformation of RNA aptamers. This simplicity in structural motifs highlights the efficiency of RNA folding, where the primary sequence dictates the formation of specific secondary structures, which, in the end, influences the assembly of the tertiary structure. Despite the shared semi-circular structural motif, the distinct sequences of RNA aptamers in **Figure 4.2** are evident through the differently coloured bases in the tertiary structures.

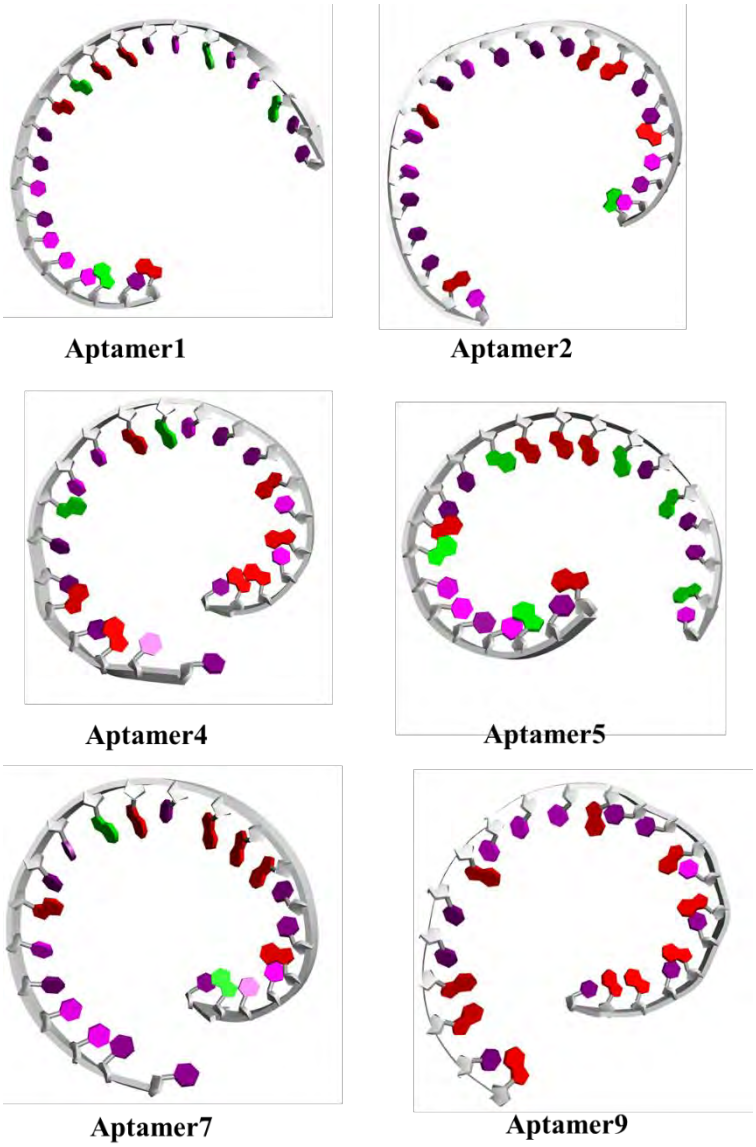


Figure 4.2: Random unfolded aptamer tertiary structures from the dataset

4.3.3 Pairing compositions analysis

The GC and AU pairings for each secondary structure were computed and reported in **Figure 4.3**. In **Figure 4.3**, two scatter plots, labelled (A) and (B), illustrate the relationship between minimum free energy (MFE) and the number of GC and AU pairings in nucleotide sequences, respectively. **Figure 4.3 (A)** depicts the impact of GC pairings on MFE, with the x-axis representing the number of GC pairings and the y-axis showing the corresponding MFE values. Aptamer with zero GC tend to have less or stable or non-folded structure as the showed to have MFE greater -1 kcal/mol. The vertical clusters of blue dots suggest that sequences with fewer GC pairings tend to have higher MFE values, which indicates lesser structural stability. In some cases, as the number of GC pairings increases, the MFE values become more negative, implying that an increase in GC pairings within the aptamers may contributes to greater stability in the nucleotide sequence. The spread of points along y-axis **(Figure 4.3 (A))** also suggests high variance, where in some cases high number of GC pairings seem to have a stronger influence on lowering MFE for some aptamers.

In contrast, **Figure 4.3 (B)** examines the effect of AU pairings on MFE, with the x-axis showing the number of AU pairings and the y-axis indicating MFE values. This plot reveals a broader range of MFE values compared to the GC pairings, as well as more extreme negative values. This figure further demonstrates that there are aptamers with no AU pairings that still possess stable folded structures, indicated by MFE values less than -5 kcal/mol. Interestingly, two of the aptamers in the dataset, which have the maximum number of AU pairings (5), show MFE values above -3 kcal/mol. This suggests that GC pairings play a more significant role in reducing MFE compared to the AU pairings which contribute to highly stable nucleotide structures.

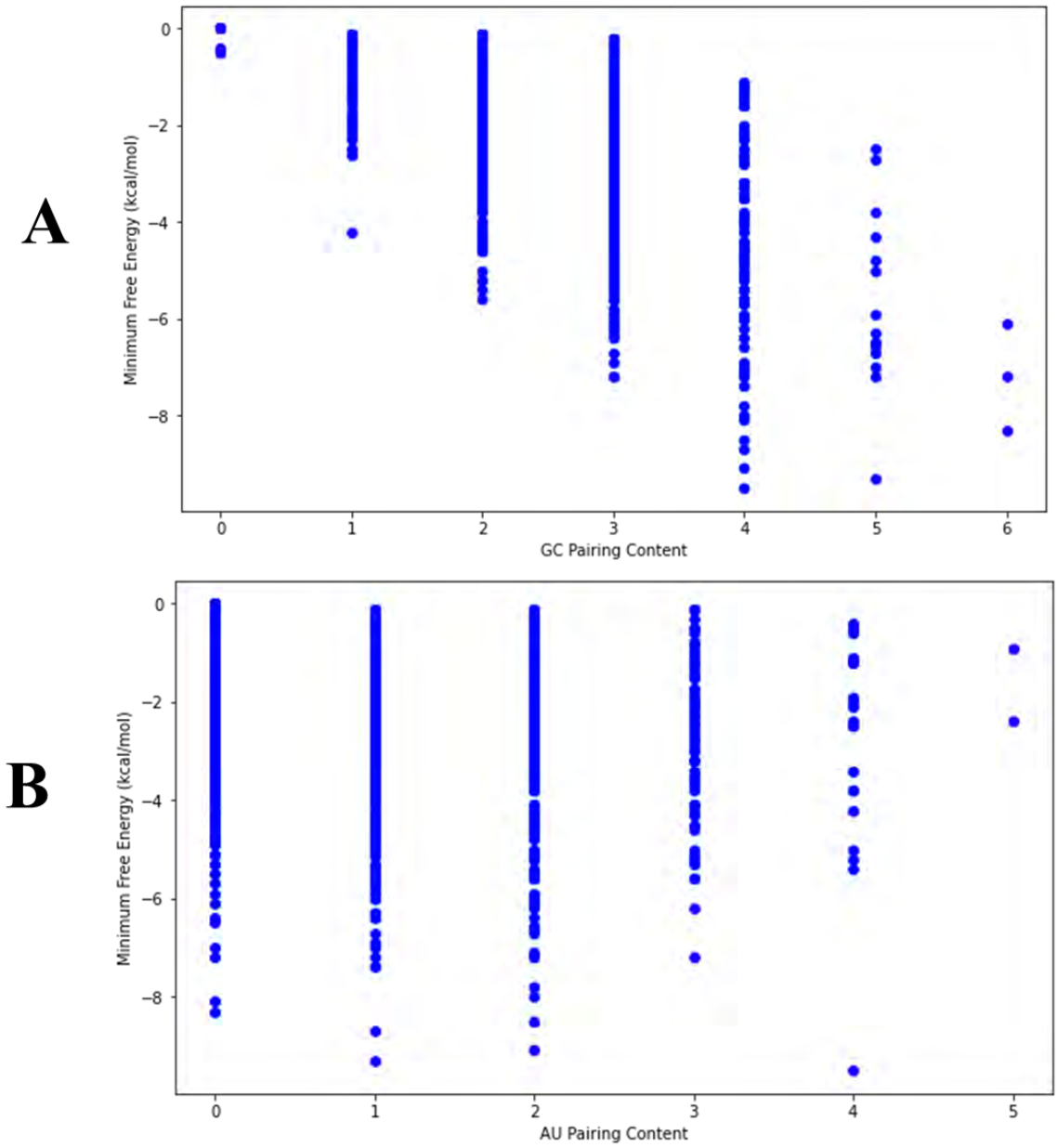


Figure 4.3: A) The Relationship between GC pairings and MFE. B) The relationship between AU pairings and MFE.

In **Figure 4.4**, the relationship between the number of AU and GC pairings in an aptamer and its corresponding minimum free energy (MFE) is shown as 3D surface plot. The colour scale ranges from dark purple (lower MFE) to bright yellow (higher MFE), indicating how the energy values change with different GC and AU pairings. Peaks in yellow represent sequences with higher MFE, meaning fewer stable configurations, while valleys in dark purple represent sequences with lower MFE, indicating more stable structures. The most stable aptamers (lowest MFE) seem to occur in regions with balanced AU and GC pairings. Suggesting that the aptamer with balance GC and AU pairings the aptamers enhance the folded stability. In some instances,

higher GC pairings is generally associated with lower energy, suggesting a trend where the GC-rich sequences contribute more to aptamer stability due to the stronger hydrogen bonds in GC pairs compared to AU pairs. There are fluctuations in the MFE values even for sequences with similar pairing numbers, possibly due to specific arrangements of the pairs or secondary structure effects.

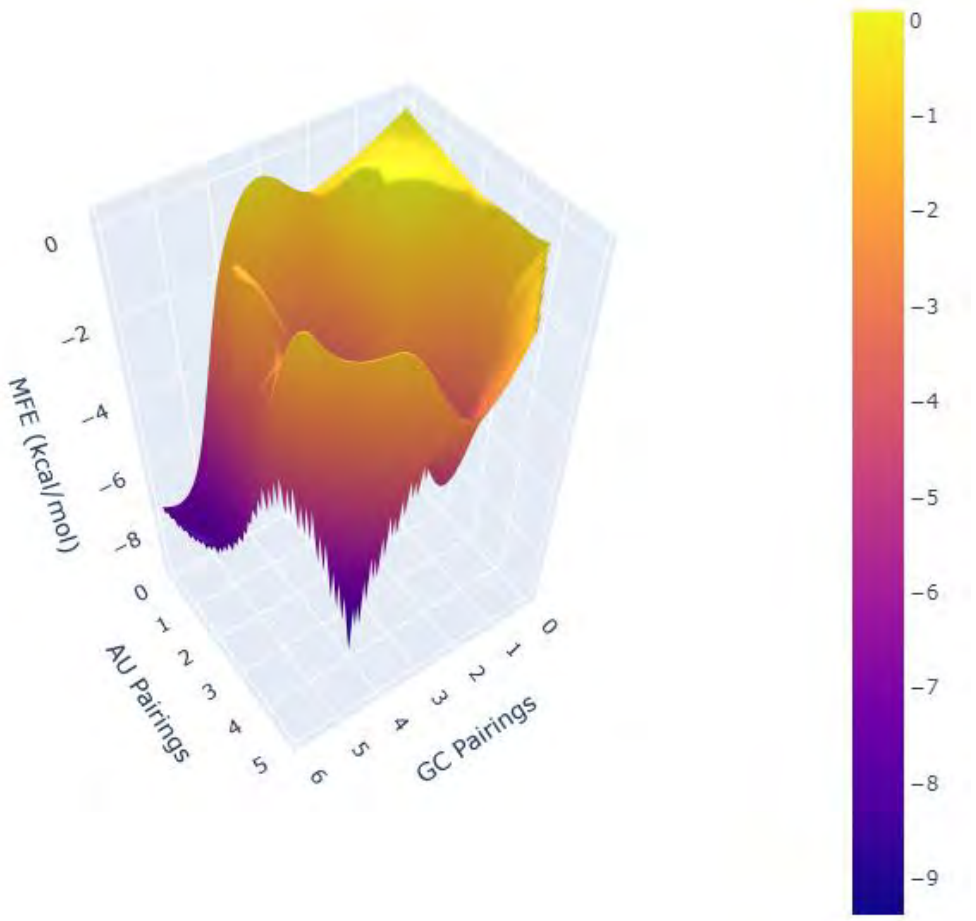


Figure 4.4: 3D Surface Plot of Minimum Free Energy vs GC Pairings and AU Pairings

4.4. Conclusion

In summary, the analysis emphasizes that although Minimum Free Energy (MFE) serves as a valuable indicator of RNA aptamer stability, it may not always directly correspond to structural compactness or folding complexity. Aptamers sharing similar MFE values did show variations in their secondary and tertiary structures. Some aptamers in the datasets did not fold, and they exhibited a semi-circular structural shape. Further, analysis highlights the significant influence of GC and AU pairings on the minimum free energy (MFE) of these aptamers. It reveals that while both types of pairings contribute to structural stability, in some cases the GC pairings are more effective in lowering MFE, thus enhancing the stability of aptamer structures compared to AU pairings.

Chapter 5

T_SELEX Application 3: Introducing Sequence similarity Check algorithm

5.1 Overview

Assessing nucleic acids or protein sequences for similarities is important for finding evolutionary relationships and functional attributes between multiple sequences. Furthermore, it can be used to identify mutations that may have occurred. In this chapter, the Sequence Similarity Check (SSC) algorithm is introduced as a user-friendly algorithm from our T_SELEX program designed to revolutionize sequence similarity analysis. Unlike traditional tools like BLAST, SSC efficiently handles large libraries of sequences with impressive speed while maintaining high precision. SSC features the ability to calculate diversity scores, providing valuable information into data variability, and can also identify similar secondary structures associated with the sequences.

5.2 Theory and methodology

While trying to transform sequence similarity analysis, the Sequence Similarity Check (SSC) algorithm employs a sophisticated strategy. Unlike conventional methods such as BLAST [368], SSC takes a targeted approach by focusing on sequences of identical lengths. This choice not only streamlines the analysis but also ensures comparisons are made on a consistent basis, minimizing unnecessary alignments and enhancing accuracy. Unlike traditional algorithms that scan external biological databases, SSC operates differently. It conducts a thorough comparison within your dataset, systematically examining each sequence against others. This approach offers a unique perspective on sequence relationships, enabling comprehensive analysis without reliance on external databases. Even with extensive datasets, SSC empowers internal comparisons, identifying similarities and differences with precision. Importantly, SSC also considers secondary structures in its analysis, providing a deeper understanding of sequence relationships. Additionally, it calculates a diversity score, offering insights into the variability within your sequences. It is worth noting that SSC handles sequences of different lengths

cautiously. While monitoring similarities, it disregards length disparities and refrains from alignments. This simple approach ensures accurate similarity assessment while accommodating sequence length variations.

5.2.1 Scoring functions

Although the scoring function for the SSC algorithm is clearly explained in chapter 2 and expressed mathematically. It is important to note that the SSC algorithm does much more than just find similarities within sequences. It also takes secondary structures into account. The score function $s(a_i, a_j)$ formula is expressed as follows:

$$s(a_i, a_j) = \sum_{k=1}^{L_{\min(i,j)}} \delta(a_i[k], a_j[k]) \, (\delta(S_i[k], S_j[k]) \,) \qquad (5.1)$$

For any two aptamer sequences a_i and a_j with secondary structures $S_i$and S_j.

5.2.2 Diversity score

Within our SSC method, the diversity score algorithm is also optional. The diversity score provides a quantitative measure of how distinct a set of aptamers is from one another based on their Hamming distances. Hamming Distance is a machine learning approach to measure how different two strings (in this case, aptamer sequences) are from each other. It does this by counting how many positions in the sequences have different characters [370]. The Hamming distance $d_H(A_i, A_j)$ between two aptamers, A_i and A_j is defined as:

$$d_H(A_i, A_j) = \sum_{k=1}^{L} \delta(A_i[k], A_j[k]) \qquad (5.2)$$

Where L is the length of the aptamers. The humming distances are further comprised into a matrix using dissimilarity matrix $D[i][j]$ which is defined as:

$$D[i][j] = d_H(A_i, A_j) \quad \text{for } i \neq j \qquad (5.3)$$

If $i \; is \; eqaul \; to \; j$ then the dissimilarity matrix will be equal to zero, which suggests that the two aptamers are identical. It is important to note that the dissimilarity matrix is symmetric meaning $D[i][j] = D[j][i]$. To find out how distinct the entire set of aptamers is, we can sum all the values in the dissimilarity matrix to find total distances (T_d) and mathematically it can be expressed as:

$$T_d = \sum_{i=1}^{n} \sum_{j=1}^{n} D[i][j] \qquad (5.4)$$

However, since $D[i][j] = D[j][i]$ is symmetric and $D[i][i]= 0$, we considered only the upper triangle of the matrix (or lower triangle) to avoid double counting:

$$T_d = 2 \times \sum_{i=1}^{n} \sum_{j=i+1}^{n} D[i][j] \qquad (5.5)$$

The number of unique pairs of aptamers (i.e., combinations of 2 from n size of dataset) is given by:

$$\binom{n}{2} = \frac{n(n-1)}{2} \qquad (5.6)$$

Therefore, the average pairwise dissimilarity (diversity score S) can be calculated as:

$$S = \frac{T_d}{\binom{n}{2}} = \frac{T_d}{\frac{n(n-1)}{2}} = \frac{2 \times T_d}{n(n-1)} \qquad (5.7)$$

In the implementation we provided, the diversity score is calculated by dividing the total distance by $n(n-1)$ instead of the average (since the total distance is computed directly from the dissimilarity matrix).

5.3 Results and discussion

5.3.1 Sequence similarity analysis

The sequence similarities were investigated based on base search and find algorithm from T_SELEX package, this algorithm is also well explained in the T_SELEX program section chapter 2. This algorithm is slightly different from multiple sequence alignment since it is designed to take account the sequence with same length. As explained in the T_SELEX section, this algorithm goes through every sequence and try to match it with all sequences. In our 1100 library of aptamers that were generated using T_SELEX program, all were searched for similarity and scored based on the number of similar bases and which positions do have those similarities. The top three 4 highest sequence similarities are reported in **Figure 5.1** and **Figure 5.2**.

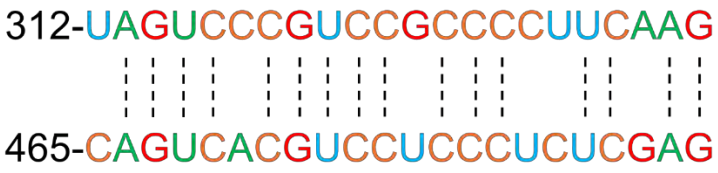


Figure 5.1: Sequences with the highest sequence similarity scores of 16.

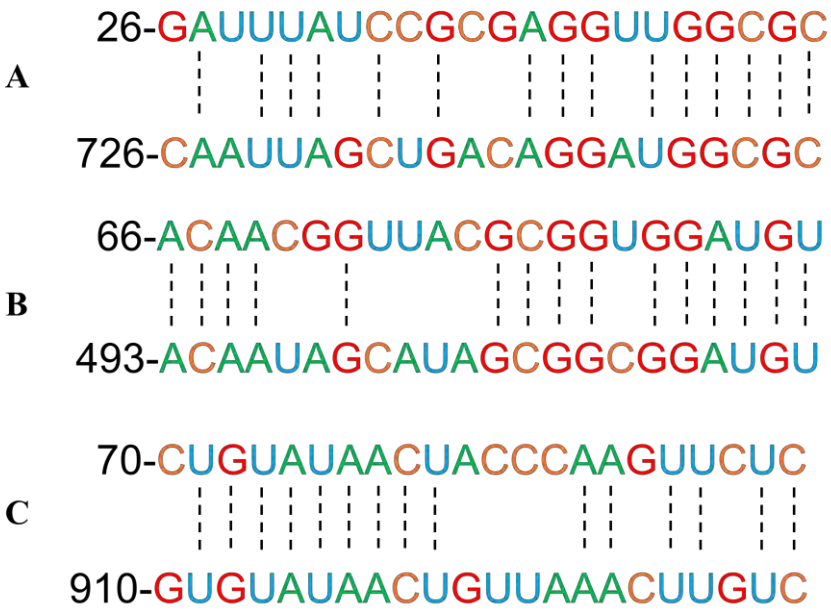


Figure 5.2: Examples of matched sequences with similarity scores of 15.

The base matches from the algorithm were set to generate matches that are above 10. The Table in Excel file format was generated. The results were sorted based on the scores, from the highest value (top) to the lowest score value (below). **Table 5.1** below was obtained as a snippet from the original Excel file that will be attached in the supporting information (https://github.com/KPMOKGOPA/MSc_supplementary_data/tree/main). The Table revealed that from the library, there are only two aptamers that have the highest matching bases of 16. The aptamers with such high similarities were aptamer 312 and aptamer 465, as shown in the Table, with the matching positions of {1|2|3|4|6|7|8|9|10|12|13|14|17|18|20|21}. With a score of 16, this indicates that there are only 6 mismatches among the aptamer bases, as shown in **Figure 5.1** above. The 8 matches from this comparison were C bases, followed by G and C with 3 matches for each, and lastly A with only 2 matches.

Table 5.1: Snippet results for comparing similar sequences using the sequence similarity check without secondary structures consideration.

Aptamer1	Aptamer2	Position	Score	$\lvert MFE_d \rvert$
312-UAGUCCCGUCCGCCCCUUCAAG	465-CAGUCACGUCCUCCCUCUCGAG	1|2|3|4|6|7|8|9|10|12|13|14|17|18|20|21	16	0.5
26-GAUUUAUCCGCGAGGUUGGCGC	726-CAAUUAGCUGACAGGAUGGCGC	1|3|4|5|7|9|12|13|14|16|17|18|19|20|21	15	1.2
66-ACAACGGUUACGCGGUGGAUGU	493-ACAAUAGCAUAGCGGCGGAUGU	0|1|2|3|6|11|12|13|14|16|17|18|19|20|21	15	0.9
70-CUGUAUAACUACCCAAGUUCUC	910-GUGUAUAACUGUUAAACUUGUC	1|2|3|4|5|6|7|8|9|14|15|17|18|20|21	15	0
71-CGCGAUCGUACGGACGCGUCUA	457-CGCGACAGUAUGGCCGCUUAUG	0|1|2|3|4|7|8|9|11|12|14|15|16|18|20	15	3.6
78-CAAGGACGUGGGAUAAGGAGUC	686-UAAGGUAGUGGGAUACCUAGUU	1|2|3|4|7|8|9|10|11|12|13|14|18|19|20	15	4.8
80-AAAGCGCGUAAACCUAGAUUUU	666-AAGGGGAGGAAACCUCGGUUUC	0|1|3|5|7|9|10|11|12|13|14|16|18|19|20	15	4.0
129-AAGUGACUCCUCAAAUCAGCUG	962-CAGUAACUGAUCAACCCAGCGG	1|2|3|5|6|7|10|11|12|13|16|17|18|19|21	15	0.8
135-UUUUUUGUAAUUAGAAAUACCC	755-UUUCUUGUAAUAAGGAUAUGCC	0|1|2|4|5|6|7|8|9|10|12|13|15|20|21	15	1.1
139-AUCGUAAGACAGCCCUAUUGCC	1039-AUCGUAAGCUCGCACUAUGUAC	0|1|2|3|4|5|6|7|11|12|14|15|16|17|21	15	0.2
157-UAGUCUAGAGGGUGAUUGGGAG	311-UAGGCUGCCGGUUGAUCGCGAG	0|1|2|4|5|9|10|12|13|14|15|17|19|20|21	15	2.3
187-CCUUUCUUGUCCGGUGUUAGGU	802-CCUGAAUUUUCCGAUGAUACGU	0|1|2|6|7|9|10|11|12|14|15|17|18|20|21	15	1.4
239-AGCCCUCGCUGGCUAAUGUACU	381-GGCUCUGGCUGAGUAAUGUUCC	1|2|4|5|7|8|9|10|13|14|15|16|17|18|20	15	3.9
299-UAUACGGCGGGGAAAAAACUCA	738-AAGAAGGCGGGAAUAAAAUUCG	1|3|5|6|7|8|9|10|12|14|15|16|17|19|20	15	2.1

Table 5.2: Snippet results for comparing similar sequences using the sequence similarity check with secondary structures consideration.

Aptamer1	Aptamer2	Position	Score	$\lvert MFE_d \rvert$
70-CUGUAUAACUACCCAAGUUCUC	910-GUGUAUAACUGUUAAACUUGUC	1|2|3|4|5|6|7|8|9|14|15|17|18|20|21	15	0.0
566-UUACACGAACUUCUCUUAGAUA	753-CAACACGAACAUCUUUGAUAUU	2|3|4|5|6|7|8|9|11|12|13|15|17|19|20	15	0.0
779-GCCCACGUGUAUCUUCCUAAUG	810-GCCCACCUGAGCCUUCUUAUUA	0|1|2|3|4|5|7|8|12|13|14|15|17|18|20	15	0.0
808-ACCACCGAUUUCCCUUAGUAGC	1064-ACGGCCGACAUCCAUCAUUAGC	0|1|4|5|6|7|10|11|12|14|16|18|19|20|21	15	0.0
7-UUCCUCAUUGAUAAAUUACCGU	937-CUCCUCCUUGAUCUGACUCCGU	1|2|3|4|5|7|8|9|10|11|18|19|20|21	14	0.0
9-AUAAUUACCCAUUACAUAUAAU	50-GUAAUCGCUCACCACAUAGAAG	1|2|3|4|7|9|10|13|14|15|16|17|19|20	14	0.0
71-CGCGAUCGUACGGACGCGUCUA	457-CGCGACAGUAUGGCCGCUUAUG	1|2|3|4|7|8|9|11|12|14|15|16|18|20	14	3.6
84-AAUACGACUACUAUGUCCGAUU	119-AACAGACCUUCUCGGUACGAUU	0|1|3|7|8|10|11|14|15|17|18|19|20|21	14	0.9
171-GCUUUAACCCGAGAGCAGCCCA	183-ACUUUAAACCGGAAGUUGCCAU	1|2|3|4|5|6|8|9|10|13|14|17|18|19	14	2.9
239-AGCCCUCGCUGGCUAAUGUACU	381-GGCUCUGGCUGAGUAAUGUUCC	1|2|5|7|8|9|10|13|14|15|16|17|18|20	14	0.4
302-CGCUUUGCAACUCGUCUCCCUC	750-ACCGUUGCAACCUUGAUCCCUC	2|4|5|6|7|8|9|10|16|17|18|19|20|21	14	0.0
850-UCAUAAAAGACACUUGCAACUG	1057-CCAUUAACGACACCCACAGCAG	1|2|3|5|6|8|9|10|11|12|16|17|19|21	14	0.0
13-AGCCAAAGCCGAUAAUCUAGCC	277-AGCCAAUCGCCAAGACCAAUCC	0|1|2|3|4|5|9|11|14|16|18|20|21	13	0.0
18-AGCCGUCAACCGAAGUAUAACU	154-ACCCGACAAUGCAAAUAGGCCU	0|2|3|4|6|7|8|12|13|15|16|20|21	13	0.2

Table 5.1 and **5.2** shows the difference in Minimum Free Energy (MFE) which was evaluated in comparison to the similarity scores. The difference in MFE was calculated using the formula below:

$$|MFE_d| = MFE_{AP1} - MFE_{AP2} \qquad (5.8)$$

Where $|MFE_d|$ represent an absolute MFE difference, while MFE_{AP1} and MFE_{AP2} represent MFE for aptamer1 and MFE aptamer2 respectively. Analysing both two Tables (5.1 and 5.2) serve as snippets from original Excel files of data comparing aptamers based on their sequences, specifically highlighting differences when secondary structures are taken into account. In **Table 5.1**, which does not consider secondary structures, pairs of aptamers show high scores ranging from 14 to 16, indicating a significant number of matching nucleotides. These pairs typically exhibit smaller differences in Minimum Free Energy (MFE), suggesting that similar sequences tend to fold into stable structures. For example, the highest-scoring pair has a low MFE difference of 0.5, pointing to potential similarities in biological performance.

In contrast, **Table 5.2** which includes the consideration secondary structures, shows a slight drop in scores which highlights that the secondary structures are important when identifying aptamer matches. Some aptamer score around 14 or 15, indicating that while their sequences still show significant similarity, the addition of secondary structures complicates the comparison. Variations in structure can affect the stability of the nucleic acid folds, making it harder to achieve high matching scores. An interesting point is that some pairs have an MFE difference of 0.0. This suggests that, even with differing sequences, these aptamers can maintain similar energy profiles and fold into comparable structures. This kind of stability is crucial in biological settings, as it may indicate similar functional properties or binding affinities to targets. When secondary structures are included in the analysis, the scoring process becomes more detailed. The algorithm now looks for both nucleotide matches and the alignment of structural features. If a nucleotide matches in one aptamer but corresponds to a different structure in another, it may not contribute to the score. As a result, fewer high scores are seen since the criteria for what counts as a "match" are stricter.

5.3.2 3D structural alignment

The 3D structural alignment of aptamers with the best matches from SSC scores from **Table 5.1** and **5.2** were analysed. The differences in folding patterns become strikingly clear when visualizing the 3D structures of the RNAs, as illustrated in **Figure 5.3** and **5.4**. In **Figure 5.3** and **5.4**, the tertiary 3D structures of aptamer 312 and aptamer 465 were obtained and visualized using Discovery Studio [371].

In examining **Figure 5.3**, the stark structural differences between the aptamers become evident. Aptamer 465 appears to fold seamlessly without any discernible motifs, while aptamer312 is notably unfolded, taking on a semi-circular structure. This contrast highlights that, despite a similarity score of 16, the secondary and tertiary structures differ significantly. On the other hand, **Figure 5.4** presents a different scenario. Here, the scoring function incorporates secondary structures, revealing a more cohesive alignment between the aptamers. Both aptamer70, represented in yellow, and aptamer910 appear unfolded and can be superimposed effectively. This visual alignment suggests that when secondary structures are taken into account, the aptamers share a more compatible 3D conformation. However, it is important to note that while this is the case for these two aptamers, it does not imply that all pairs of aptamers with high sequence similarity score will exhibit such precise superimposition, especially if their minimum free energy differences (MFE_d) exceed 0 kcal/mol. When MFE_d values are higher, the structural variations can lead to significant deviations in 3D conformation, making it less likely for aptamers to align closely. Therefore, each aptamer pair may present unique structural characteristics that affect their potential interactions and functionalities.

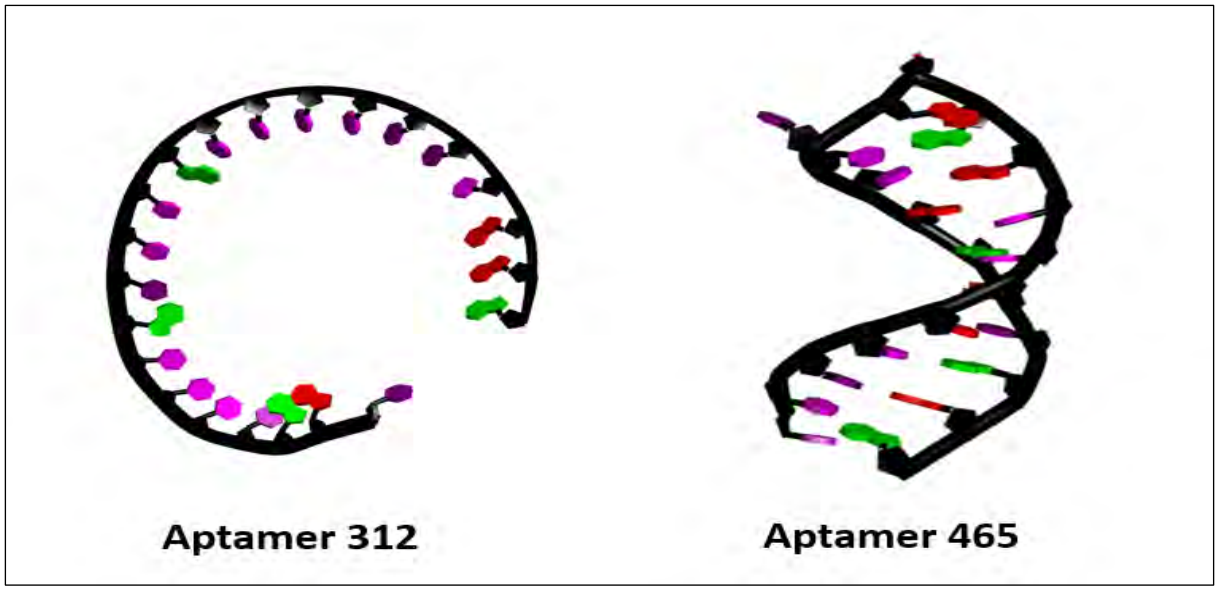


Figure 5.3: Tertiary structures of the best high sequence similarity scores of 16 without considering secondary structures.

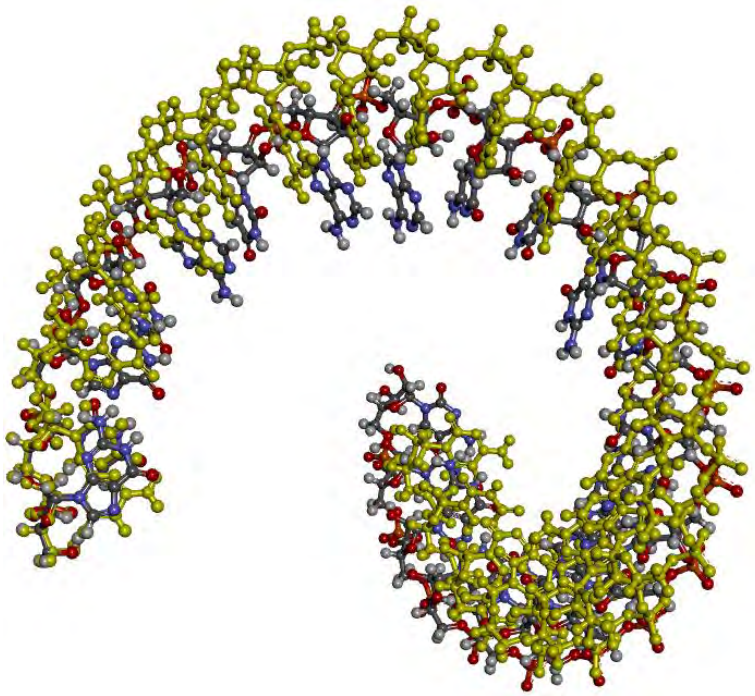


Figure 5.4: Tertiary structures of the best high sequence similarity scores of 16 with considering secondary structures (aptamer910 and aptamer70).

5.3.3 3D Analysis of the Open Semi-circular Shape of Unfolded Aptamers

Although the folded RNA structures have been extensively studied, the unfolded structures have not received as much attention [372]. This analysis focuses on understanding what gives unfolded structures, like aptamer312, their characteristic semi-circular shape. Here, only aptamer 312 is reported in detail, while other unfolded aptamers were also examined. The structural analysis reveals that the shape of aptamer312 is significantly influenced by the intermolecular forces between the bases and their adjacent counterparts along the chain. This dynamic interplay is effectively illustrated in **Figure 5.5** below. Notably, the minimum free energy (MFE) for aptamer312 was calculated to be 0 kcal/mol, aligning with findings related to GC content in sequences exceeding 15 nucleotides. As highlighted by, a zero MFE is often linked to lower GC content, however in this case, there is only one GC pair present at positions 11 and 12 (noting that indexing starts from zero) [373].

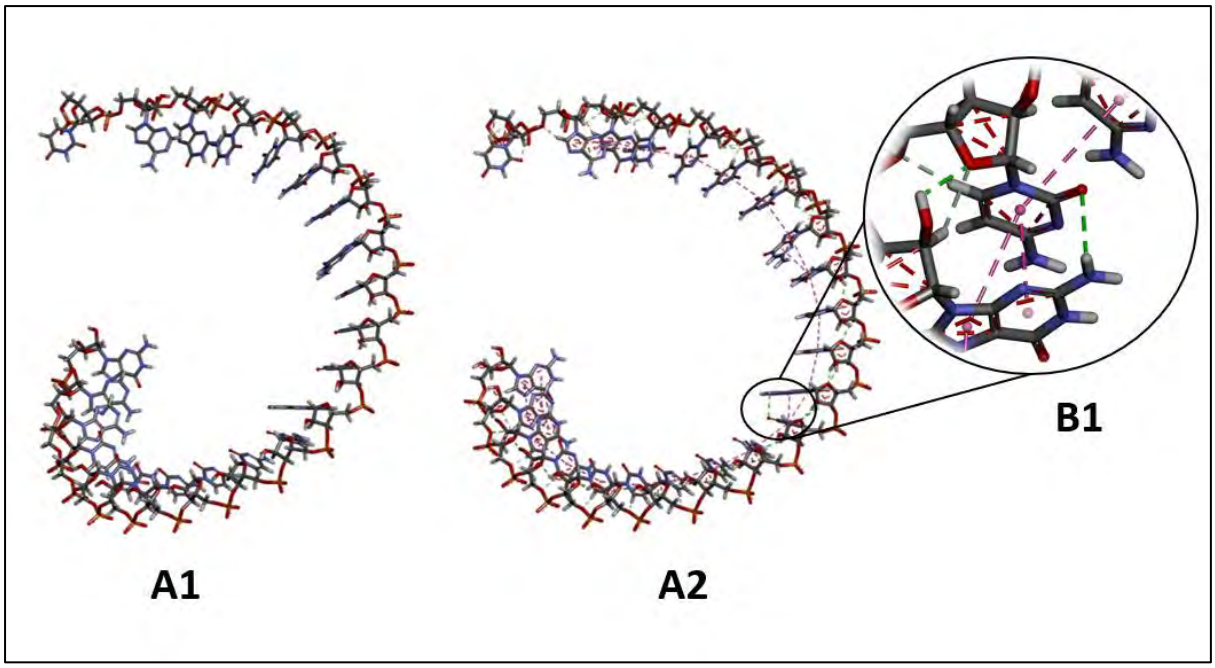


Figure 5.5: 3D structure of the unfolded aptamer 312.

In **Figure 5.5**, the 3D structure of aptamer 312 (**A1**) is shown together with its intramolecular forces (**A2, B1**). From observing the introduction of intramolecular interactions in **A2**, a purple dashed line appears to pass through the pyrimidine rings of Uracil (U), Cytosine (C), and the purine rings of Adenine (A) and Guanine (G). That line indicates the hydrophobic interaction, more specifically the pi-pi stacking. Moreover, looking at **B1**, which zooms in on a certain section of A2's structure, the pi-pi stacking can be easily seen. This suggests that each ring in

this RNA structure forms pi-pi interactions with two rings adjacent to it. This type of interaction is a three-centered pi-pi interaction [374]. Furthermore, looking at **B1** again, the pyrimidine rings show double pi-pi stacking with just one purine ring while still having another pi-pi interaction with a pyrimidine ring. This further suggests that if an RNA sequence has bases comprised of pyrimidine rings adjacent to purine rings, this four-centered pi-pi interaction is more likely to be observed if the MFE is zero kcal/mol or the structure does not fold properly based on the sequence composition. Looking at purine rings adjacent to each other, the same type of interaction is more likely to be observed since each ring distributes the pi-pi interaction across only three adjacent rings. This was observed at positions {19|20|21}, which are comprised of AAG, all of which are purine-type bases.

Hydrogen bonds were further observed between the carbon and hydrogen of the ribose sugar backbone and between the hydrogen of bases and sugar. The bond distances of hydrogen bonds between the carbon and hydrogen of the ribose sugar backbone were shown to be less than 2 Å if they are attached to bases of the same family (pyrimidine and purine). However, if there is a purine and a pyrimidine as the bases adjacent to each other in the backbone, the bond length between the carbon and hydrogen of the ribose sugar backbone increases to above 2 Å. This implies that the strength of the hydrogen bonds weakens, leading to a less folded or more relaxed structure. Of course, molecular dynamics and quantum mechanics simulations are needed for further clarity on this matter. Besides those types of hydrogen bonds mentioned above, looking closely at **B1** of **Figure 5.5**, there is only one hydrogen bond observed between the base pairs themselves. This conventional hydrogen bond occurs at positions {11|12}, where the GC content is present, with a bond length of above 2.75 Å. Theoretically, the hydrogen bond length between G and C is reported to average 2.85 Å, while experimentally, 2.86 Å was reported for the hydrogen bond that occurs between the N-H of Guanine and the O of Cytosine for DNA [375]. As it is well known, conventional hydrogen bonds play a crucial role in stabilizing the structure of biological molecules.

5.3.4 Scores distribution analysis

Figure 5.6 compares the frequency of Sequence Similarity Scores (SSS) from the aptamer dataset for SSS that do not consider the secondary structures and for SSS that take secondary structures into accounts. The graph clearly shows two distributions: one where only the sequence is evaluated (orange line) and another where both sequence and secondary structure are considered (blue line). For the sequence only analysis, the most frequent score is 12, suggesting that over 250 aptamers have this SSS. This indicates that many sequences share a

high degree of similarity when evaluated solely on their primary structure. The scores tend to cluster around higher values, creating a right-skewed distribution. This suggests that most sequences align well at the primary sequence level, but only a few sequences reach the very highest scores, such as 14 or 15. In contrast, when secondary structure is included, the most common score shifts to 9, with just over 200 sequences reaching this value. The scores in this dataset are more evenly spread but tend to cluster around lower values compared to the sequence only graph. This reflects the added complexity that secondary structure introduces, making it harder for sequences to achieve high similarity scores. Both graphs show fewer sequences at the extreme ends of the score range, especially scores below 4 or above 15. Overall, this indicates that very low or very high similarity scores are rare in this dataset, regardless of whether secondary structure is considered.

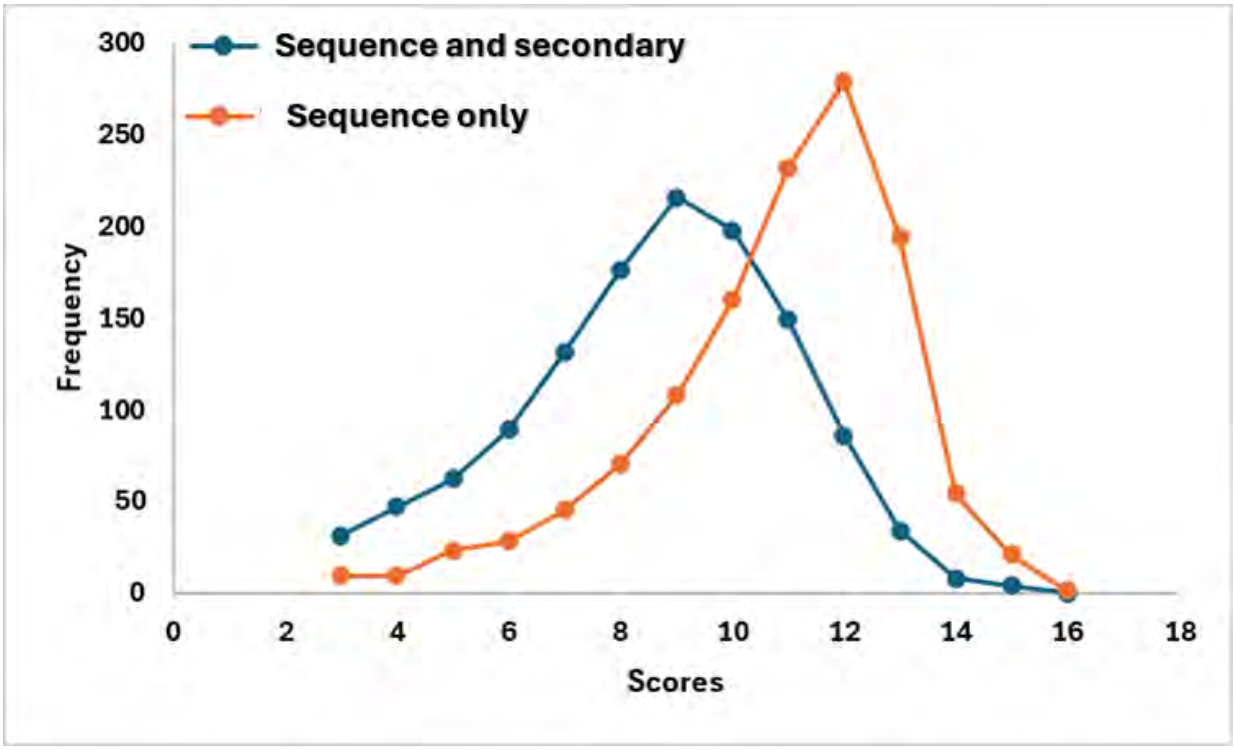


Figure 5.6: Distribution of Sequence Similarity Scores (SSS) where the secondary structures are taken into consideration (orange) and SSS where secondary structures are not taken into accounts (blue).

5.4 Conclusion

In conclusion, Sequence Similarity Check (SSC) was introduced, a novel algorithm from the T_SELEX program, designed to revolutionize sequence similarity check analysis. Through targeted analysis of sequences of identical length and internal comparisons within the dataset while taking into account the secondary structure alignment, SSC offers a unique perspective on sequence relationships, challenging conventional methods. The results indicate that aptamer312 and 465 exhibit the highest match score of 16, with only six mismatches, and a base composition showing predominant C bases. The evaluation of Minimum Free Energy (MFE) differences suggests that even minimal mismatches can lead to significant variations in RNA folding. The 3D structural analysis of aptamer312, which shows incomplete folding and a zero MFE, highlights the role of intermolecular forces like pi-pi stacking and hydrogen bonding in RNA stability. Lastly, the comparison between the distribution of Sequence Similarity Scores (SSS) where the secondary structures are taken into consideration and SSS where secondary structures are not taken into accounts shows that while many sequences align well at the sequence level, incorporating secondary structure results in a broader, more moderate range of SS scores, offering a more detailed view of sequence similarity.

Chapter 6

(Case study 1)

Benchmarking Base Randomization Algorithm (BRA) as a possible tool for the initial step of generating virtual RNA aptamers library.

6.1 Overview

While databases are emerging across various domains from small molecules to genomics and proteins, the availability of aptamer databases remains scarce. The presence of such databases could serve as a comprehensive resource, advancing research, innovation, and the application of aptamer technology across multiple fields. This advancement would likely lead to improvements in healthcare, environmental monitoring, and biotechnology. Furthermore, the establishment of aptamer databases would facilitate molecular modelling and machine learning, opening doors for further advancements in understanding and utilizing aptamers. Prompted by that, here we present and benchmark the Base Randomization Algorithm (BRA) as a potential solution to the scarcity of aptamer databases. BRA generates diverse aptamer sequences, by placing bases randomly in unique positions using pseudorandom number generators. Chapter 3 looked at the sequence composition of an aptamer dataset with sequences of the same length. The results obtained in chapter 3 highlighted that the disadvantage of having aptamers of the same length may result in folding limitation which determines the stability of the aptamers. In an effort to improve the physiochemical properties of the aptamers, we generated an aptamer library consisting of aptamers with randomised length ranging from 16 to 60 nt. Properties such as minimum free energies, sequence compositions and nucleotide arrangement were compared amongst the dataset of aptamers with the same length (1 100 aptamers), those with randomised length (20 000 aptamers), and an aptamer dataset from Aptamer base which was collected from experimental work (904 aptamers) [339].

6.2 Theory and methodology

6.2.1 Base randomization algorithm

Randomization has found application in a variety of fields such as gaming, sampling, simulations and art [376]. There are algorithms/techniques developed to carry this randomization tasks such as Monte-Carlo which is widely used in gaming and computer simulations. Many randomization techniques and algorithms are derived from pseudorandom number generators, which make use of the seed initialization method to produce numbers that appear randomized [376]. The base randomization algorithm presented here makes use of pseudorandom number generation [376]. It generates randomized RNA sequences or "aptamers" where the randomization is both in the bases and in the positions of these bases. The random generation of each single base (or nucleotide) simply follows equation 6.1:

$$RB = Set[index_i] \tag{6.1}$$

In equation 1, *RB* denotes the random base or nucleotide, Set represents a collection of bases and $index_i$ is the position of a base in this collection. $Index_i$ is calculated according to equation 2 where the *Set* is *X*:

$$index_i = \lfloor r \times \text{len}(X) \rfloor \quad \begin{cases} r \in [0,1) \\ X = \{A, U, G, C\} \end{cases} \tag{6.2}$$

Since the Mersenne Twister "random" module from python was used, *r* is the random float value/number within half-closed range between 0 and 1 that is generated randomly using Mersenne Twister as the core generator [377]. The pseudorandom Mersenne Twister is capable of producing 53-bit precision floats with period of 2^{19937}-1[377]. *X* is the *Set* or collection of elements which in this case are the bases {A, U, G, C}, which in terms of python scripting are strings and not numerical values. Since the *X* set, is composed of four strings then *Set* can be mapped to index set {0, 1, 2, 3} to select the base and the $\text{len}(X)$ is denote the length of the *X* set which 4 in this case. It worth noting that we are dealing with RNA hence U (uracil) is present not T (thymine). The main objective is to generate multiple sequences of random

bases/or nucleotides where each individual RNA is unique. A single sequence (seq) may be generated according to equation 6.3:

$$seq = [\,RB_0\,, \quad RB_1\,, RB_2\,\,, \quad \cdots\,, RB_n\,] \qquad where \;\; n \in \mathbb{N}_0$$

$$seq = [Set[\lfloor r_0 \times \text{len}(X)\rfloor],\; Set[\lfloor r_1 \times \text{len}(X)\rfloor] \cdots Set[\lfloor r_n \times \text{len}(X)\rfloor] \qquad (6.3)$$

For a single sequence, the generation is based on the length of the sequence which is denoted as n and this is an element of natural numbers (with zero), since we consider the index starting from 0. For multiple lists of sequences with same length ($M_{seqs[]}$) can be expressed as follows:

$$M_{seqs[]} = \begin{bmatrix} [Set\big[\lfloor r_0^1 \times \text{len(X)}\rfloor\big]] & \cdots & Set\big[\lfloor r_n^1 \times \text{len(X)}\rfloor\big]] \\ \vdots & \ddots & \vdots \\ [Set[\lfloor r_0^m \times \text{len(X)}\rfloor] & \cdots & Set[\lfloor r_n^m \times \text{len(X)}\rfloor]] \end{bmatrix} = \begin{bmatrix} [seq_0] \\ [seq_1] \\ \vdots \\ [seq_m] \end{bmatrix} \qquad (6.4)$$

For multiple sequences of the same length, $M_{seqs[]}$ is represented as matrix since it is a list that contains sub-lists of the same length, where the number of sequences $m \in \mathbb{N}_0$. We can thus denote the position of each of these RNA sequences within $M_{seqs[]}$. For multiple sequences which may be different in length, M_{seqs} can also be expressed similarly but with a few additional conditions. Since we are looking at the randomization also of sequence length we denote each sequence as set rather than an array with equations 6.5 and 6.6 being the equivalent of equation 3 but with sets:

$$seq = \{\,RB_0\,,\; RB_1\,, RB_2\;\,, \cdots\,, RB_n\} \qquad where \;\; n \in \mathbb{N}_0 \qquad (6.5)$$

$$seq = \{Set[\,\lfloor r_0 \times \text{len(X)}\rfloor],\; Set[\,\lfloor r_1 \times \text{len(X)}\rfloor] \cdots Set[\,\lfloor r_n \times \text{len(X)}\rfloor]\} \quad (6.6)$$

n is the last position of a base in one of the sequences which automatically reveals that n is length of that particular sequence. Since this is the case, and the length can be generated

randomly between a given closed range of j and k, we can further continue and denote it as follows:

Let: $n = Set[ind_i]$ (6.7)

Where $ind_i = \lfloor r \times \text{len(S)} \rfloor \begin{cases} r \in [0,1) \\ S = \{j, \cdots, k\} \end{cases} \quad j, k \in \mathbb{N}$ (6.8)

Let M_{seqs} be the main set and seq_i be the subset: $\{seq_i\} \in M_{seqs}$, then for $\exists\, seq_i \in M_{seqs}$ main set can be denoted as follows.

$$M_{seqs} = \begin{Bmatrix} \{Set[\lfloor r_1^1 \times \text{len}(S) \rfloor] & \cdots & Set[\lfloor r_n^1 \times \text{len(S)} \rfloor]\} \\ & \vdots & \\ \{Set[\lfloor r_1^m \times \text{len(S)} \rfloor] & \cdots & Set[\lfloor r_n^m \times \text{len(S)} \rfloor]\} \end{Bmatrix}$$

$$= \begin{Bmatrix} \{seq_0\} \\ \{seq_1\} \\ \vdots \\ \{seq_m\} \\ \{seq_{m+1}\} \end{Bmatrix} \qquad \text{for } n \in [i, j] \qquad (6.)$$

For Multiple sequences with randomized lengths (M_{seqs}), n is the last position of each sequence in the M_{seqs}, therefore, the n values are generated randomly between a closed specified range of j and k. In this study we choose the n to range between 16 and 60. But with that being said, it is not beyond the realm of possibility that during those generation of multiple sequences, the algorithm can generate repeating sequences. This concern can be effectively resolved by applying the 'set' principle, which enables the creation of a distinct and non-repetitive list of unique items.

Pseudocode1: Base Randomization Algorithm

Algorithm: Base Randomization Algorithm (BRA)

Input:

- length: the length of the aptamers to generate (either a specific length or "randomize")
- aptamers numbers: the number of aptamers to generate

Output:

- A list of unique aptamers based on the aptamer number input

Steps:

1. seed (0)
2. Initialize an empty set() aptamers to avoid the repeats in list
3. If length is "randomize", then:
    a. While the size of aptamers is less than aptamers numbers:
        i. Generate a random length number between 16 and 60 (inclusive)
        ii. Generate a random aptamer as an item using characters 'ACUG'
        iii. If the aptamer sequence is not in aptamers, then add it to aptamers
4. If length is a specific value, then:
    a. Generate a random aptamer of the specified length using characters 'ACUG'
    b. While the size of aptamers is less than aptamers numbers:
        i. Generate a random aptamer of the specified length using characters 'ACUG'
        ii. If the aptamer is not in aptamers, add it to aptamers
5. Convert the set aptamers to a list and return the list

The Base Randomization Algorithm (BRA) has a time complexity of $O(m \cdot n)$, where n is the maximum length of the aptamers and m is the number of aptamers to generate. This complexity arises from generating random aptamers, which takes $O(n)$ time, and checking for uniqueness using a set, which has an average case of $O(1)$. In the worst case, particularly when many attempts are needed to find unique sequences, this could lead to $O(m \cdot n)$ iterations. The space complexity is also $O(m \cdot n)$ due to storing unique aptamers in the set. Ultimately, both time and

space complexities reflect the efficiency and potential challenges of generating a specified number of unique aptamers.

The number of possible sequences that can be generated by BRA is determined by length n and k number of different bases given, the formula can be described as:

$$\text{Number of sequences} = \mathrm{k^n} \qquad (6.10)$$

For sequences composed of four nucleotides ('A', 'C', 'U', and 'G'), the number of possible sequences is calculated as by reduce k to 4. To illustrate how the number of possible sequences increases with sequence length, consider the following calculations. For a sequence of length 1, there are 4^1=4 possible sequences. As the sequence length increases to 2, the number of possible sequences grows to 4^2=16 With a length of 3, the number of sequences expands further to 4^3=64. At a length of 4, the number of possible sequences reaches 4^4=256. For a sequence of length 5, the number of sequences increases to 4^5=1024. When the sequence length is extended to 10, the number of possible sequences becomes 4^{10}=10485764. This demonstrates how exponentially the number of possible sequences increases with sequence length, reflecting the vast complexity and variability possible in nucleotide sequences. This exponential growth reflects the combinatorial complexity of variations in nucleotide sequences, indicating that longer sequences can encode a vastly greater number of potential configurations. As the sequence length increases, the number of possible distinct aptamer sequences expands rapidly, providing a larger space for genetic or chemical diversity. This rapid increase in possible aptamer sequences that can be obtained highlights the richness of chemical space available.

Generation of aptamers sequences

Aptamers sequences were generated using "Base randomization algorithm", with the Pseudocode1 written in python programming language. Two lists of aptamers were generated. The first list contained 1 100 aptamer sequences with fixed length ($M_{seqs[]}$) of 22 nt and the second list contained 20 000 aptamer sequences with randomized length ranging between 16 and 60 nt (M_{seqs}). The third dataset was obtained from Aptamer base and in this study the data set is referred as "RNAbase" [336]. This RNAbase dataset contained random RNA and DNA aptamers sequences that are obtained from experimental work together with their properties. The DNA sequences were filtered out, only 904 RNA sequences were left and taken further for composition and structural analysis.

6.2.2 Secondary and Tertiary structure prediction

Single stranded RNAs fold within themselves through base pairing resulting in stable secondary hairpin structures. To address the concern of base pairing regions in RNA molecules, RNA folding is highly required to map the possible base pairing regions that can be conserved within the molecule. In order to fold a biological molecule computationally, certain tools are required, such as RNAfold [265] and Mfold [378]. For this current study secondary structure were predicted using the T_SELEX program that incorporates RNAfold algorithm. As mentioned before, RNAfold algorithm make use of Zuker and Steigler algorithm and John McCaskill's algorithm of partition function [265]. On that note, Zuker and Steigler algorithm in RNAfold enables prediction of Minimum Free Energy (MFE) structures from just a simple given RNA sequences [265]. For 3D (tertiary structure) all the sequences together with their secondary structures in those datasets were submitted RNAComposer [337].

6.3 Results and discussion

6.3.1 Us, Gs, Cs and Us composition analysis

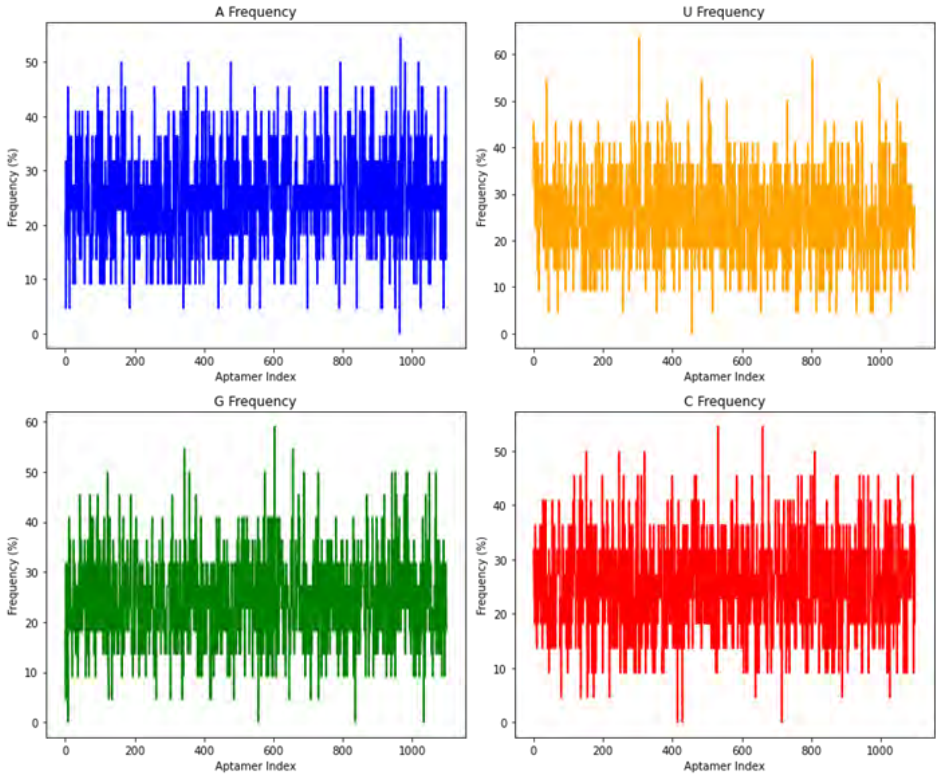
This section is focused on unravelling the base composition of aptamer sequences in the three datasets. Single base composition of the three datasets ($M_{seqs[]}$, dataset, M_{seqs} , and RNA base) are compared in **Figure 6.1**. $M_{seqs[]}$ dataset was generated using BRA with fixed aptamer length of 22 nt, Mseq dataset was generated using BRA with aptamer length ranging between 16 and 60 nt, and RNA base is composed of RNA sequences from aptamer base [336].

In the study of base composition across different datasets, the behaviour of nucleotides in terms of randomness and noise was analysed, as shown in **Figure 6.1**. This figure presents four frequency plots in percentages for each dataset, focusing on specific RNA aptamers. The x-axis numerically labels the aptamers (e.g., "aptamer 1," "aptamer 2," ..., "aptamer 1100") which are referred to as aptamer index, which helps track individual sequences. Each plot corresponds to one of the nucleotides: Uracil (U), Guanine (G), Adenine (A), or Cytosine (C).

Figures 6.1A and **6.1B**, which represent the BRA datasets, reveal that most aptamers in $M_{seqs[]}$ have nucleotide frequency ranging between 15% and 35% for each base, while in the M_{seqs} dataset, the frequency range from 10% to 45%. Many aptamers contain similar amounts of each nucleotide, across these two aptamer datasets. For RNAbase the data is not clear enough to draw definitive conclusions about nucleotide composition frequency for most aptamers. Across all datasets, including RNAbase, some aptamers completely lack certain bases. This is observed within all frequency noise plots with some aptamer having nucleotide frequency of 0%. This

indicates that some aptamers may be synthesized without uracil, guanine, cytosine, or adenine, whether intentionally or unintentionally. And, of course, with the BRA datasets (M_{seqs} and $M_{seqs[]}$), it was not intentional, but still this pattern of missing nucleotide was observed. The absence of guanine (G) in some sequences within three datasets is particularly concerning, as G-rich sequences are known to form stable secondary structures. Chapter 4 discussed how G-C pairing is generally more stable than A-U pairing, indicating that missing G or C could negatively impact the stability of aptamer folding. When examining the RNAbase dataset (**Figure 6.1C**), unusual trends appear, with some aptamers containing 100% of a single nucleotide. This indicates that these aptamers cannot form stable folded structures due to the lack of complementary nucleotides for pairing. This underscores the importance of thoroughly investigating base composition in relation to RNA folding.

A



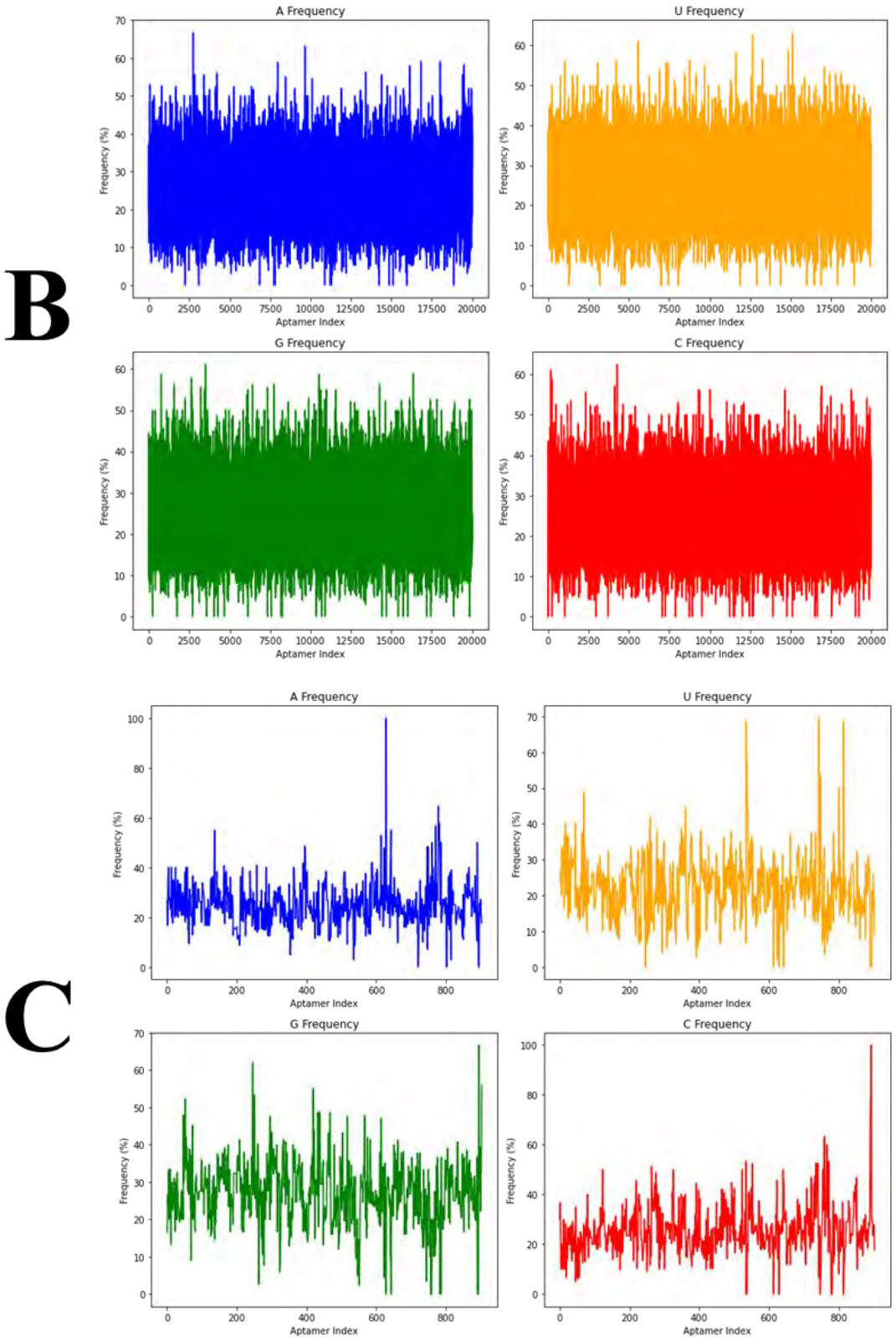


Figure 6.1: A Composite Figure of A, B and C where A is composed individual base composition noise plots of dataset $M_{seqs}[]$, B for M_{seqs} and C for RNAbase.

The distribution counts plots in **Figure 6.2**, shows how often each nucleotide appears in sequence within each dataset. In our study, the focus is on counting Uracil (U), Adenine (A), Guanine (G), and Cytosine (C) in various aptamer sequences across three datasets. The initial noise plots suggested for $M_{seqs[]}$ that most aptamers had nucleotide frequency ranging between 15% and 35% while for M_{seqs} most aptamers exhibit single nucleotide frequency ranging from 10% to 45%. However, **Figure 6.2** provides more clarity, showing that in the $M_{seqs[]}$ dataset, most aptamers have base composition within a sequence ranging from 4 to 8 nt, while in the M_{seqs} dataset, counts vary from 3 to 20 nt. For the RNAbase dataset, pinpointing a specific range is trickier due to its multimodal distribution. Nevertheless, it appears that many aptamer sequences in RNAbase have base counts between 5 and 20 nt, though this is not consistent across all four bases, as illustrated in **Figure 6.2C**. The distribution plots indicate that the $M_{seqs[]}$ dataset base counts follow a normal distribution, whereas the Mseq dataset shows a slight leftward skew. This skew is likely due to the random lengths ranging from 16 to 60 nt. **Figures 6.2A** and **6.2B** reveal similar distributions for all four bases, indicating they have comparable mean values. We confirmed this through a one-way ANOVA, which is detailed in the supplementary data (Table S.1).

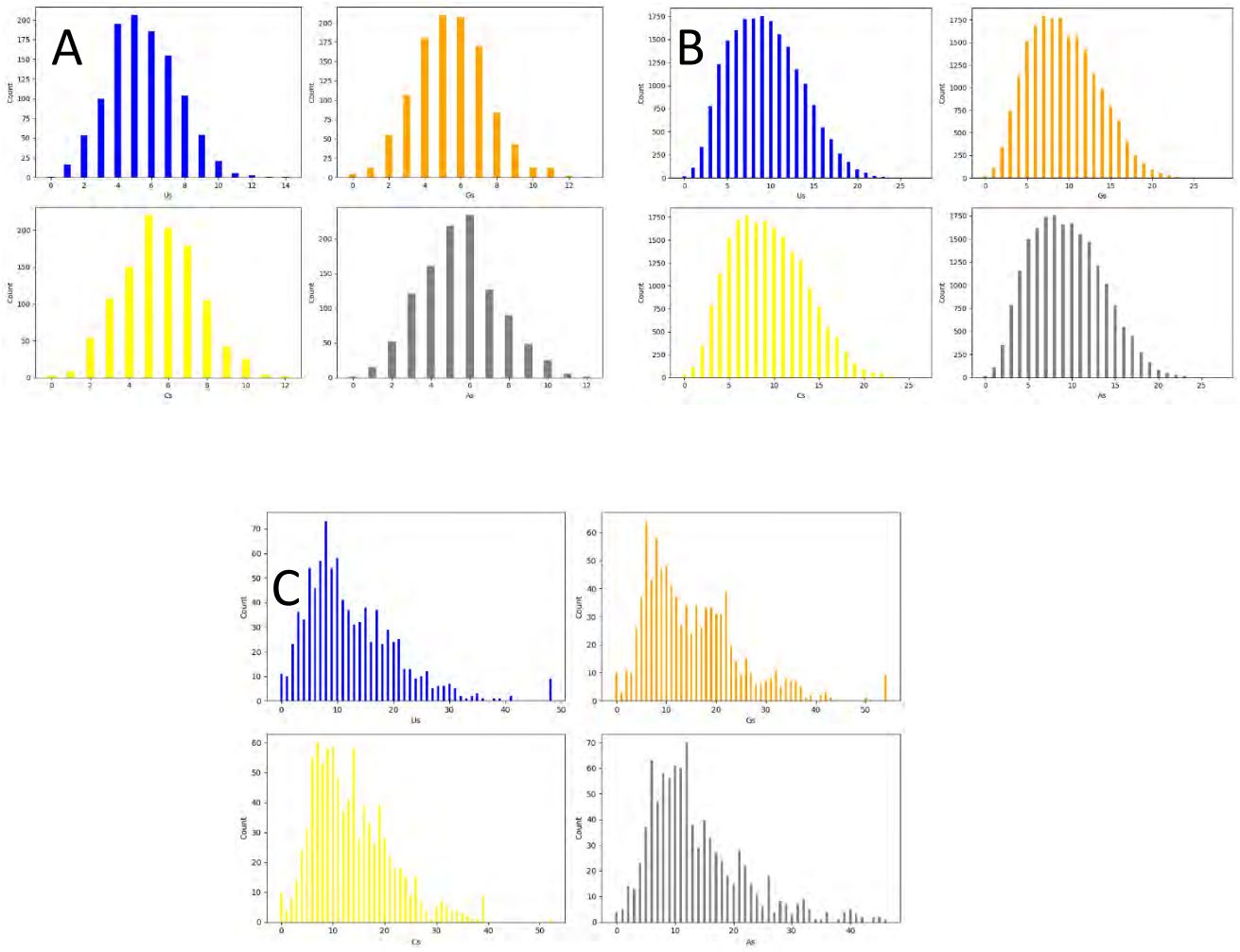


Figure 6.2: A Composite Figure of A, B and C where A is composed of individual base distribution plots within the dataset $M_{seqs[]}$, B for M_{seqs} and C for RNAbase.

6.3.2 Adjacent base composition

The violin plots in **Figure 6.3** shows the distribution of the adjacent base compositions. Like the individual base compositions, the distributions for adjacent base pairs are mostly similar for most pairs within each dataset, as seen in **Figures 6.3A**, **6.3B** and **6.3C**. A difference was expected in terms of the distributions within adjacent base pairs of the RNAbase dataset, unfortunately it was not observed as shown in **Figure 6.3C**. This variation expectation was based on fact that RNAbase (C) dataset contained aptamer sequences from experimental SELEX studies, while for BRA datasets sequences (A and B) were generated theoretically [336]. **Figure 6.3B** shows that adjacent base pairs from AU to CG have similar median values and quartiles, with some differences from UU to AA (which are the last four violin plots in

Figure 6.3B). The highest probabilities for these pairs correspond with their median values. The same is observed in **Figure 6.3A**, where adjacent pairs from AU to CG also share the same minimum, first quartile, median, third quartile, and maximum values. The median is close to the first quartile, indicating a slight positive skew in the data. The most common value peaks around a median of 5, with smaller peaks at 0 and 10. The presence of a minor peak at zero suggests that some aptamers do not have certain adjacent base compositions. All adjacent pair compositions have a minimum value of zero, meaning some aptamers do not include those pairs at all. The distributions from UU to AA show similar trends, although with minor differences. The BRA algorithm rarely places identical bases next to each other but does not completely rule it out. Median values are consistent across the AU to AA range. **Figure 6.2B** shows uniform distributions from AU to CG, with slight variations in the UU to AA range. Interestingly, a similar pattern is observed in **Figure 6.2C** with adjacent base pairs from AU to AA having the same minimum, first quartile, median, third quartile, and maximum values. This shows that although RNAbase exhibit multimodal distribution of the adjacent base pair compositions, the pattern is also similar to that of the BRA algorithm even though the datasets do not have the same size.

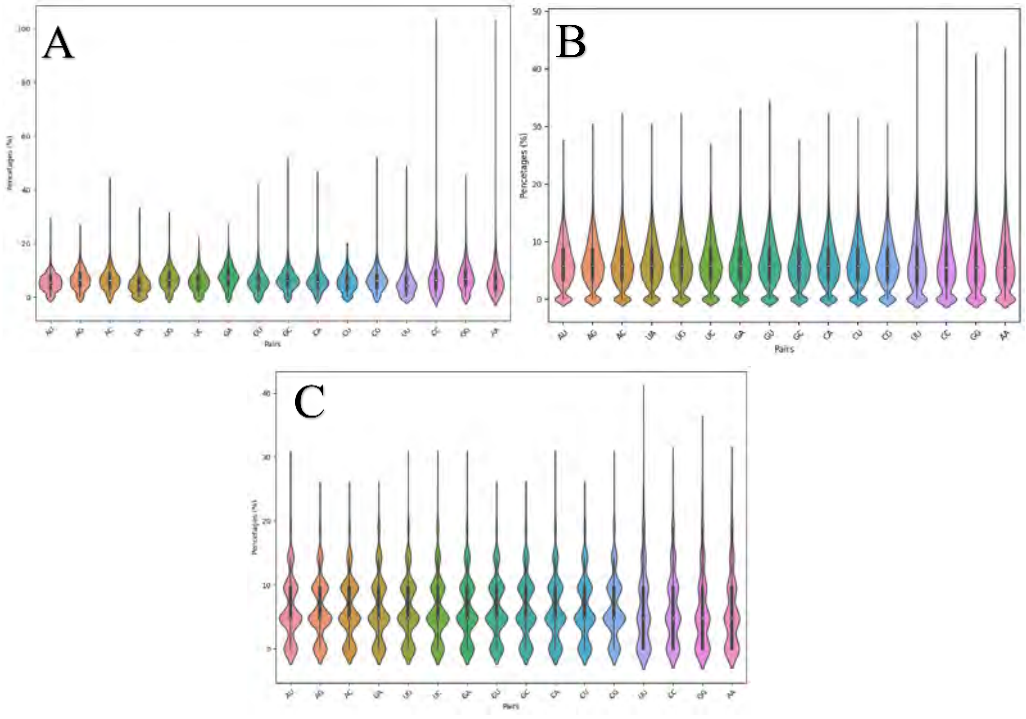


Figure 6.3: A Composite Figure of violin plots for dataset A, B and C, where A is composed of adjacent base composition distribution plots within the dataset $M_{seqs[]}$, B for M_{seqs} and C for RNAbase.

6.3.3 Folding, secondary structure and 3D predictions

Figure 6.4A shows the distribution of MFE values for aptamer sequences across the datasets we investigated. Upon analysis, RNAbase contains the most stable aptamers, with the most stable one reaching an MFE of -80.70 kcal/mol. This trend is further illustrated by the outliers in the box-and-whisker plot for RNAbase in **Figure 6.4A**. Other highly stable aptamers in this dataset have MFEs of -58.00 kcal/mol and -53.29 kcal/mol, along with a notable number of outliers between -46 kcal/mol and -37 kcal/mol. The significant difference between the most and second most stable aptamers suggests that sequence length contributes to variations in MFE. The low MFE of RNAbase aptamers suggests that these aptamers are likely longer, as MFE generally decreases with more base pairings, indicating greater stability. Notably, 61 aptamers in RNAbase have an MFE of zero, accounting for 6.75% of the dataset, which could be due to the relatively short sequences in RNAbase, since aptamer sequence length within this dataset ranges from 3 to 180 nt. Although typical aptamer lengths cited in literature are between 16 and 60 nt, RNAbase includes many shorter sequences, which may not fold into stable structures but could provide more binding surface area for targets. While synthesizing very short nucleic acids can be tricky, they still have practical applications [379].

In the M_{seqs} dataset, the most stable aptamer has an MFE of -26.39 kcal/mol (aptamerd5165) and a length of 54 nt. Although this aptamer is stable, the maximum length in this dataset is 60 nt, indicating that while length is a factor influencing MFE, it is not the only one. Other factors, like base composition (both individually and in pairs) and base positioning, also significantly impact stability. **Figures 6.4A** and **6.4B** show that aptamerd18670, which is 50 nt long, has an MFE of -25.39 kcal/mol, followed by other aptamers with slightly higher MFEs. Out of 20 000 aptamers in the M_{seqs} dataset 1 942 have a MFE of zero suggesting that 9.71% do not fold.

Further examination of the $M_{seqs[]}$ dataset reveals that the most stable aptamer has an MFE of -9.5 kcal/mol (aptamer1084). Since all aptamers in this subset are 22 nt long, therefore length certainly does not influence MFE in this case. The second most stable aptamer has an MFE of -9.3 kcal/mol (aptamer950). Among 1 100 aptamers, 281 have an MFE of zero, meaning 25% do not fold. Overall, $M_{seqs[]}$displays the highest percentage of non-folding aptamers, reinforcing that while length affects MFE, it is not the sole factor to consider.

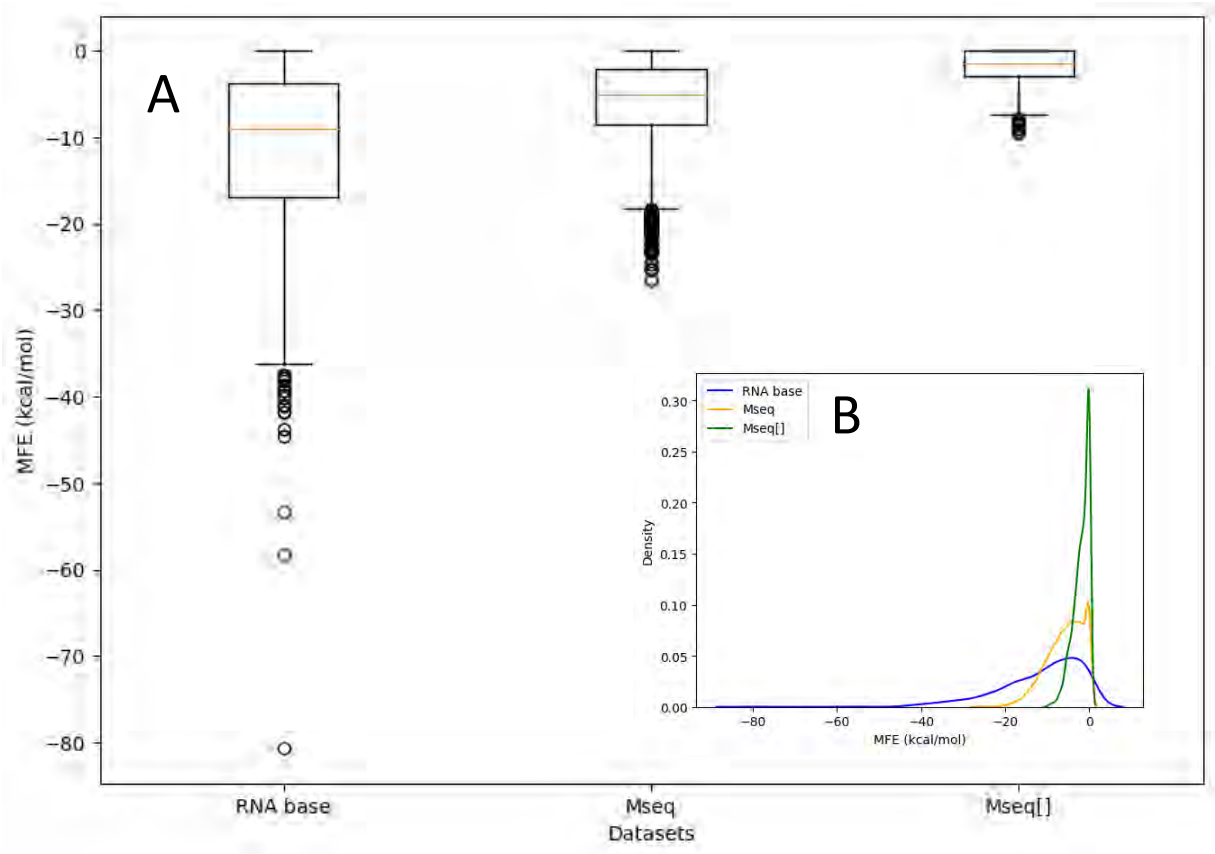


Figure 6.4: A Composite Figure of A and B, where A is composed of box plots of MFE of RNA aptamers within the dataset $M_{seqs[]}$, M_{seqs}, and RNAbase. B is showing the distribution line plots of the RNA aptamers within the dataset $M_{seqs[]}$, M_{seqs} and RNAbase.

Correlation heatmaps were constructed to evaluate and investigate the correlation between length of each sequence in each dataset and their folding behaviour through observing MFE as shown in **Figure 6.5B** and **6.5C**. Prior focused discussion on the folding and correlations, it is important to not overlook compositions correlations. For $M_{seqs[]}$dataset, their correlation is reported in **Figure 6.5A**. There is distinctive correlation of -0.34 to -0.32 amongst the induvial base composition in $M_{seqs[]}$ dataset. This suggest that there is an inverse relationship among the among the bases, even though the correlation is not strong enough. This could be due that the sequences having the same length and if one base were to dominate the other bases have to be reduced thereby ensuring that the combined total remains at 22 nt. For instance, if number of As in sequence is 10, then other bases will have to share the remain twelve composition to make it up to 22, hence the negative correlation. On the contrary the correlation of among bases for M_{seqs} and RNAbase datasets show positive correlation which suggest that for randomized length has significant positive relationship that can be observed amongst the bases.

Figure 6.5B shows a relative strong correlation of -0.73 between length and MFE. Notably, in the RNAbase dataset, this correlation is even stronger at -0.9. This indicates a significant inverse relationship between the stability of RNA molecules and their length. While correlation does not imply causation, these results suggest that length plays a substantial role in RNA folding stability. The inverse relationship emphasises the idea that longer RNA sequences tend to have lower MFE values, implying greater stability. This occurs because longer sequences have more plausible ways of folding through base interactions, which can contribute to more stable structures. However, it is important to note that not all possible folding states are stable. Thus, understanding the stability of RNA involves more than just length. It also requires analysing the composition and positioning of individual bases within the sequence. This highlights the importance of considering both base composition and spatial arrangement in relation to the overall stability of RNA molecules.

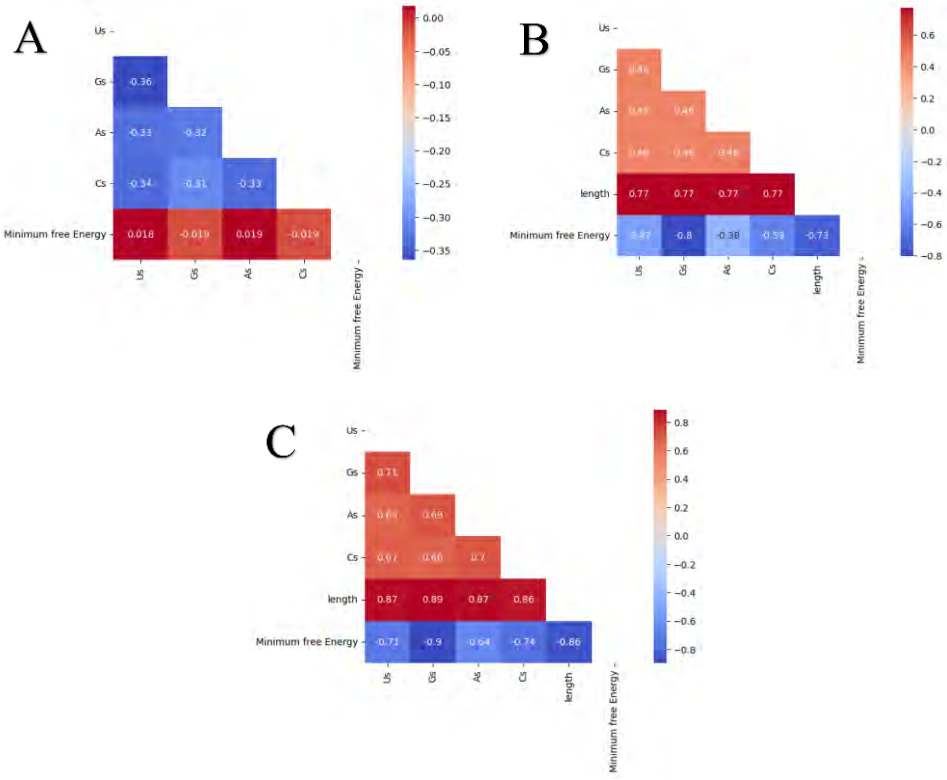


Figure 6.5: Correlation matrices of bases, length of the sequences and the Minimum Free energy of the three datasets, where A is for a dataset $M_{seqs[]}$, *B for* M_{seqs} *and C for RNAbase.*

Because aptamer length, composition, and position or arrangement influences the aptamer MFE, it can also be thought as a product of length dependent factor, aptamer length, composition and arrangement of nucleotides. According Trotta's work, if MFE=a+b•length, then MFE/length=a/length+b in case of perfect linier relationship between MFE, and length [380]. But Trotta further demonstrate the assumption of linear relationship between length and MFE is invalid [380]. Which further justifies that there is more that need to be taken into consideration about the composition and arrangement of nucleotides towards MFE. Although is still a mystery for us to formulate how composition and arrangement deeply affect MFE in order to give a clear and probable hypothesis, we can assume that these factors do contribute since there still much to uncover. We can further denote our hypothesis as $MFE = -(\chi f N)$ where $\chi = 0 \;\; if \; N \leq 7 \; and \; \chi = 1 \; if \; N > 7$. N is the length of sequence and f represent both composition and arrangement factors, even though for now we cannot give a precise equation of how f may be calculated. χ is the length dependent factor which is introduce based on the understanding that all sequences that have any length less or equal to 7 have MFE = 0. Which simply suggest that the arrangement does not matter in that case, the composition and arrangement of nucleotide matters only if the N is greater than the 7. The calculations to back up this claim of χ are provided in this **Figure 6.6**.

The graph in **Figure 6.6** shows the exponential relationship between aptamer length and two factors: the number of stable structures with nonzero minimum free energy and the number of possible arrangements. As aptamer length increases, there's a significant rise in the number of stable structures with nonzero MFE, suggesting longer aptamers are more likely to form stable structures. it is important to note, the red line indicates a sharp increase in the "Number of Nonzero MFE" after the length of aptamers reaches 8. This could be due to the fact that longer sequences have more potential for forming stable secondary structures.

According to Zuker's algorithm, a permissible secondary structure must have a loop that has three free nucleotides and two base pairs [345]. This is because a loop with fewer than three free nucleotides would be too tight to form, and a base pair contributes significantly to the stability of the structure. Moreover, in sequences shorter than 8, there might not be enough nucleotides to form these stable structures with the required loop and base pairs. Hence, a sharp increase in the number of nonzero MFEs after length 7, suggesting longer sequences have more potential for forming these stable structures [381]. Additionally, the number of possible arrangements increases with aptamer length attributing to the exponential increase, indicating a saturation in arrangement diversity. Overall, the trend illustrates the complexity and diversity

of aptamer interactions, with longer aptamers having a higher potential for stable structures and arrangement variety.

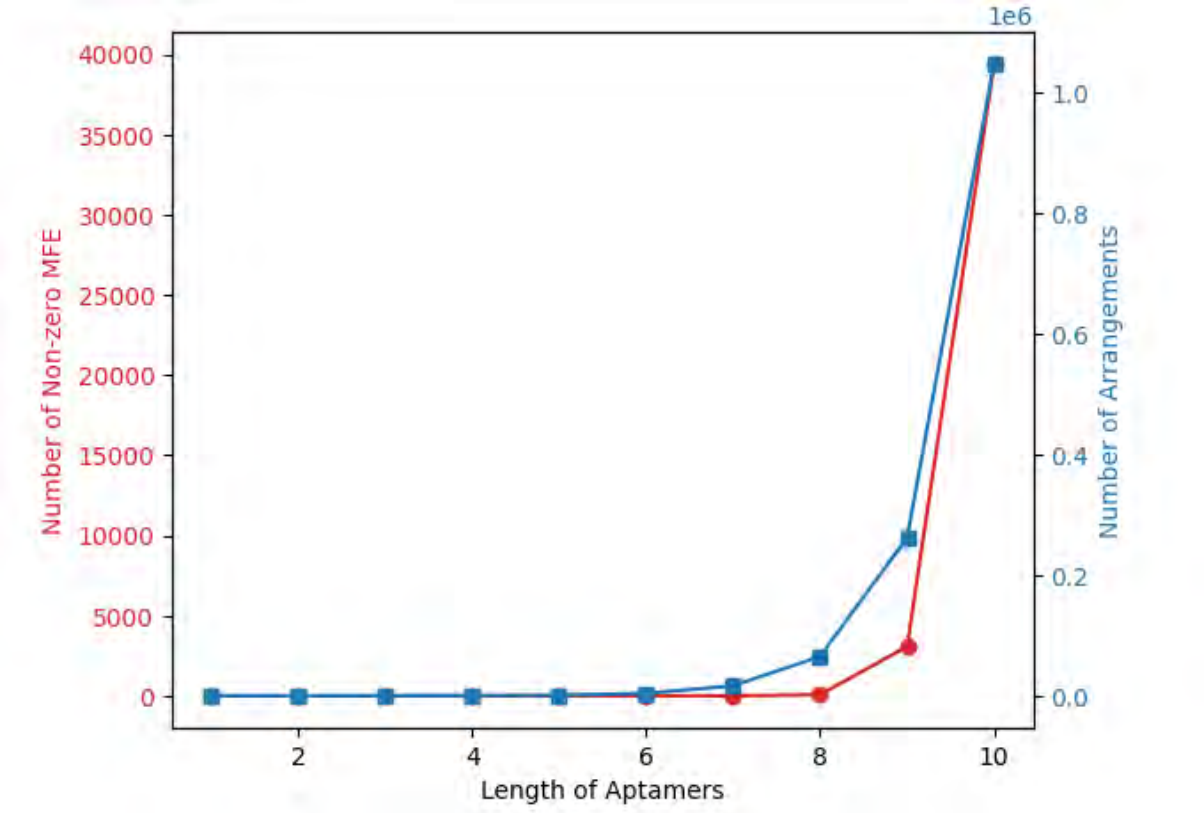


Figure 6.6: Analysis of number of possible bases rearrangements (blue) and number of possible folded aptamers or non-zero MFE aptamers as the length increases using BRA.

To gain insight into how adjacent base compositions influence the minimum free energy (MFE) and thereby contribute to RNA molecule stability, we examined heatmaps in **Figure 6.7**. Notably, GG adjacent base compositions exhibit a consistent, but still not significant, negative correlation with the MFE across all datasets. Intriguingly, other adjacent base compositions such as GC, CG, GU, and AG also display a slight negative correlation with the MFE in all datasets. This suggests that sequences containing GG may favour folding, especially if UU, CC, UC, or CU exist in the sequence. A similar assumption can be made for the other mentioned pair compositions. Despite variations in correlation values, the heatmaps exhibit similar patterns across all three datasets, indicating consistent trends from AA to UA on the y-axis and from GC to CC. However, overall, there is not a strong relationship observed between adjacent base composition and MFE.

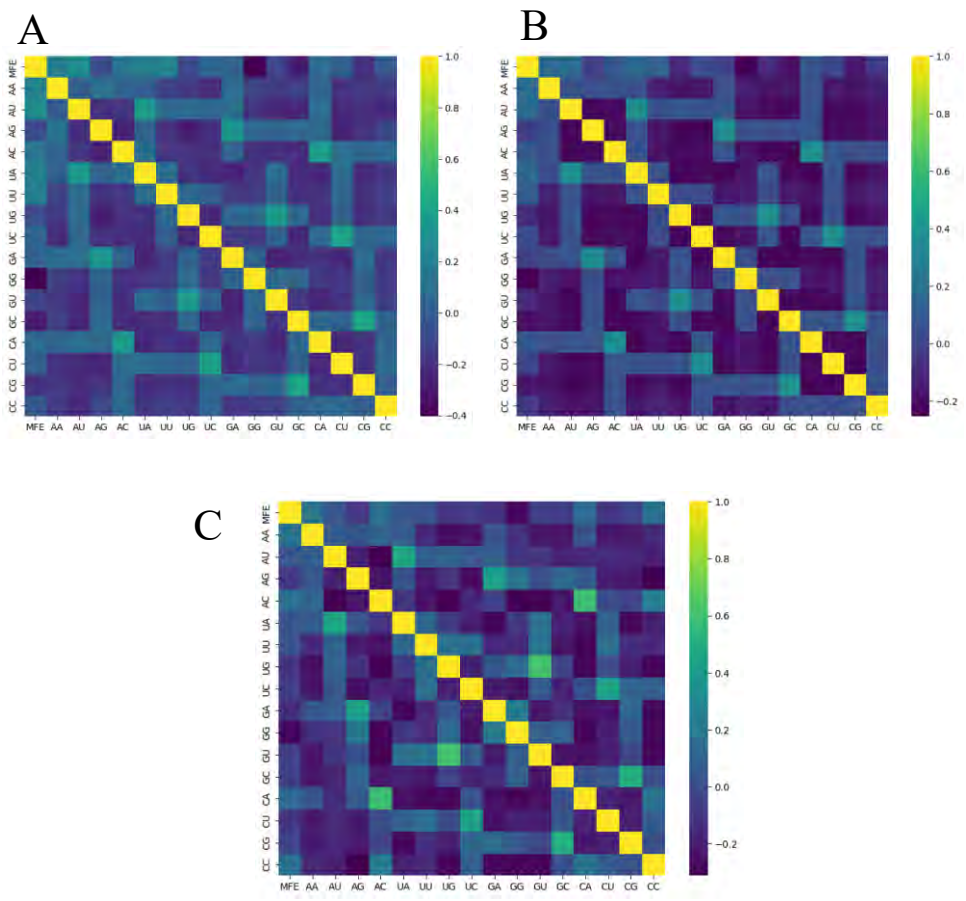
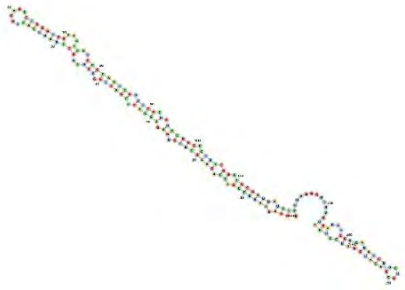
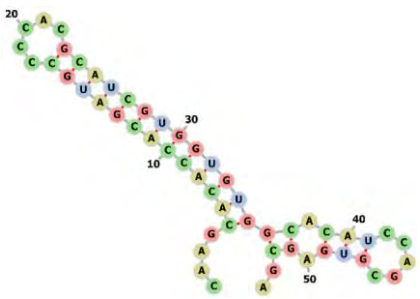
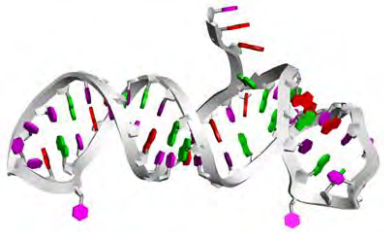
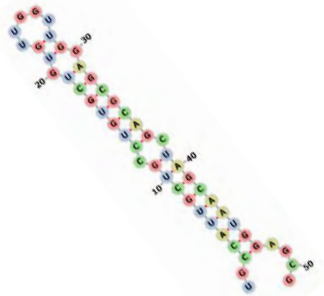
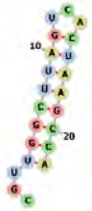


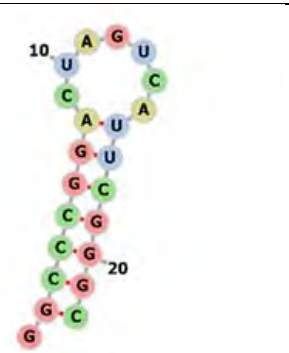
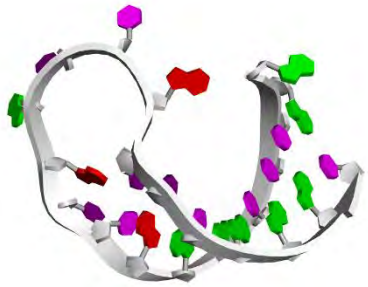
Figure 6.7: Correlation matrices of adjacent base composition within a sequence and the Minimum Free energy (MFE) of the three datasets, where A for a dataset $M_{seqs[]}$, B for M_{seqs}, *and C for RNAbase.*

Table 6.1 shows the two best-folded aptamers from each dataset and compiled their sequences, pseudoknots, MFE secondary structures, and tertiary structures. Discussing the motifs found in these RNAbase aptamers. RNAse69 has a four-way junction or a multi-dimensional loop in the center, suggesting a complex structure that may enhance its binding capabilities. In contrast, RNAse192 features a simpler dodecahedral structure with kinks but lacks a multi-dimensional loop. Despite both sequences being long, they have different secondary structures. For the M_{seqs} aptamers, Aptamer5165 displays a secondary structure that includes a multi-center loop, indicating potential for varied interactions. On the other hand, Aptamer18670 forms a more straightforward structure with good stem-loops and internal loops, but no kinks, suggesting a simpler binding profile. Finally, $M_{seqs[]}$, Aptamer960 and Aptamer1084 show simple secondary structures with no multi-center loops. This likely means their shorter sequences do not have the length needed to form more complicated, stable structures.

Table6.1: The best folded aptamers including sequences, secondary structures and tertiary structures from the three datasets

Aptamer ID	Sequence (5' to 3') and Pseudoknots	MFE 2d structure	Tertiary structure/3d structure
RNAbase			
RNAse69	AUUUCUCUGAGAUGUUCGCAAGCGGGCC AGUCCCCUGAGCCGAUAUUUCAUACCAC AAGAAAUGUGGCGCUCCGCGGUUGGUGA GCAUGCUCGGUCCGUCCGAGAAGCCUUA AAACUGCGACGACACAUUCACCUUGAAC CAAGGGUUCAAGGGUUACAGCCUGCGGC GGCAUCUCGGAGAUUCC ...((((((((((((.(((..((((((.....(((((.(((((((.....)))))))).).))))(((((((..((((....(((((.....)))))....)))).)))))))).....(((((((((...)))))))))......))))))).))))))))))))))....		
RNAse192	GGGAGAAUUCCGACCAGAAGCUUGUGAG ACCAGCCGAGUGGUGUCUGGCUAUUCAC UGGAGCGUGGGUGGAACCCCUGCGCACU CGUUUGGCUGUCCGGGCCUUCGGGCCGG GAUUAUCUCUUUGGGUUUUGUGAUUUGG UCAUAUGUGCGUCUACAUGGAUCCUCA ((((.(((((((.((.(((((((((.((.(((((.((((((.(((..((((((((((.....))))))))...))..)))))))))).))))).).))))))).))))).)).)))))))).)))).......... (((...((((.((((((....))))))))))).)))		
M_{seqs}			

aptamerd 5165	CAAGCACACCACGAUGCCCCA CGCAUCGUGGUGUGGCACAUC CAGCGUGAGCGA ....((((((((((((.....))))))))))))((.(((.....))).))..		
aptamerd 18670	UGCCAUUGCUGCCUGUGCUGU GUUGGUUGGAGCGCAGCUAGC AAUGGAGCG ..(((((((((.(((((((.((.....)).))))))).))))))))....		
$\boldsymbol{M_{seqs}}$[]			
Aptamer1084	CGUUGGCUUAGUCACUAAGCCA Secondary structure: ...(((((((...)))))))		

Aptamer960	GGCCCGGACUAGUCAUUCGGGC Secondary structure: .(((((((.....)))))))		

6.4 Remarks and propositions

Throughout the course of the study, various remarks and propositions emerged, which were identified as promising areas for further exploration. These noteworthy aspects, along with their corresponding mathematical proofs, are documented below. With the two BRA generated datasets, we propose that the mean composition of each base in each dataset can be approximated as follows:

For *N* fixed length of sequences:

$$\overline{x}_B \approx P(x) \times N \qquad x \in X \ \ where\ X = \{A, U, G, C\} \qquad (6.11)$$

Since $P(x) = \frac{1}{n}$ $for\ n\ lengh\ of\ set\ X\ then$:

$$\overline{x}_B \approx \frac{1}{n} \times N$$

For *N* as random length between closed range of i, j; where $i \neq j$ then:

$$\overline{x}_B \approx \frac{1}{n} \times M \qquad where\ M \text{ is median value of } [i, j] \qquad (6.12)$$

Proofs :

1) Statement, equation (6.11)
   Claim: A base x mean of a dataset set can be approximated as the product of probability of base x in set X, and the length of the sequences given that all sequences in dataset have the same length.

$$\overline{x}_B \approx P(x) \times N$$

Proof:

For each sequence, seq whose length is N, for any $x \in X = \{A, U, G, C\}$

$$\frac{\not{k}}{N} \approx P(x)$$

if N is large enough and given it follows gaussian distribution

Where $\hbar$ is the number of x in sequences, *seq* and *P(x)* is the probability of randomly selecting base x.

Therefore, $\hbar$ can be explicitly approximated as:

$$\hbar = P(x) \times N$$

Since the length of the sequence is the same, we expect that the count of base X in each sequence are approximately the same. Then we can average the number of occurrences of the base x.

$$\overline{x}_B = \frac{1}{M}\sum_{i=0}^{M} \hbar_i \approx \frac{M\hbar}{\mathrm{M}} = P(x) \times N$$

Note: $\hbar_1 \approx \hbar_2 \approx \hbar_3 \cdots \approx \hbar_{\mathrm{m}}$

2) Statement equation (12)

   Claim: A base x mean of a dataset set can be approximated as the product of probability of base x in set X, and the mean or median of lengths of the sequences given that sequences in dataset have the random lengths of closed range i and j.

$$\overline{x}_B = \frac{1}{n} \times M$$

Proof :

For m sequences

Assumption:

Each sequence has different length

$$N_1, N_2 N_3, \cdots, N_m$$

For large enough $N_1, N_2 N_3, \cdots, N_m$ then:

$$\frac{\hbar_1}{N_1} \approx P(x), \frac{\hbar_2}{N_2} \approx P(x), \frac{\hbar_3}{N_3} \approx P(x), \quad \cdots, \quad \frac{\hbar_m}{N_m} \approx P(x)$$

The counts:

$\hbar_1 \approx P(x) \times N_1, \quad \hbar_2 \approx P(x) \times N_2, \quad \hbar_3 \approx P(x) \times N_3, \quad \cdots, \hbar_m \approx P(x) \times N_m$

Average number of occurrence:

$$\overline{x}_B = \frac{1}{M}\sum_{i=0}^{M} \hbar_i = \frac{\hbar_1 + \hbar_2 + \hbar_3 + \cdots + \hbar_m}{M}$$

$$= \frac{(P(x) \times N_1) + (P(x) \times N_2) + (P(x) \times N_3) + \cdots + (P(x) \times N_m)}{M}$$

$$= P(x)\frac{(N_3 + N_2 + N_3 + \cdots + N_m)}{M}$$

These proofs lay a solid background for understanding variance and covariance of bases in aptamer libraries. Examining the expected distribution of bases, researchers can assess the diversity of sequences, which is essential for effective target binding. The average occurrences of bases help predict how often they appear in random sequences, guiding the design of experiments and the generation of libraries.

6.5 Conclusion

In conclusion, the analysis revealed diverse base compositions in RNA aptamers, with implications for stability based on the presence or absence of specific nucleotides. The study emphasized the importance of understanding base pairings and compositions for predicting the stability of RNA structures. Through benchmarking BRA, we provide a mathematical aspect of how this algorithm works to generate sequences, wherefore, multiple sequences of the same length ($M_{seqs[]}$) can be denoted as matrices while for sequence with random lengths (and M_{seqs}) they can be thought as main set and subsets. The compositions and arrangement together with MFE of the generated sequences in $M_{seqs[]}$ and M_{seqs} datasets were evaluated and compared to the RNA aptamer sequences from Aptamerbase (RNAbase). As we delved deeper into compositions analysis it was further illustrated that the BRA follows gaussian distributions. Based on composition analysis we propose that the base mean of the dataset can be approximated as $\overline{x}_B \approx P(x) \times N$ for dataset of sequences with same length and $\overline{x}_B \approx P(x) \times M$ for dataset with randomized length that follows gaussian distribution. Finally, we discuss and highlight an important aspect regarding the folding of aptamers generated by the BRA algorithm. Specifically, it is noted that aptamers with lengths equal to or less than 7 nt lack the ability to fold when utilizing RNAfold. This emphasised that aptamers with longer length are more likely to exhibits very low MFE values, suggesting very stable folded aptamers.

Chapter 7

Case study 2 (main case study)

Using RNA aptamers as novel miR-10b inhibitors for anticancer therapeutics

7.1 Overview

This chapter focuses on the investigation of aptamers as microRNA inhibitors for anticancer therapeutics, specifically targeting hsa-miRNA-10b-3p, hsa-miRNA-10b-5p, and Pre-miRNA-10b as a novel approach using the T_SELEX program. The aim is to identify aptamers that can effectively bind to these oncogenic miRNAs. Various computational techniques are explored to predict and analyse these interactions, beginning with interaction predictions that assess binding affinities of aptamers to oncogenic microRNAs. The chapter highlights the process of virtual screening to identify promising candidates, followed by a systematic evaluation of docking results to understand their performance. Additionally, it incorporates quantum mechanical calculations to analyse the structural properties, such as reactivity and kinetic stability of selected aptamers-miRNA complexes. Molecular dynamics simulations are employed to examine the dynamic nature of the aptamer-miRNA complexes over time, offering insights into their interactions and conformational changes under specified physiological conditions. Finally, binding energies are assessed through Molecular Mechanics Generalized Born Surface Area (MMGBSA) calculations, providing average binding energies overtime towards equilibrated aptamers-miRNA complex system.

7.2 Methodology

7.2.1 Interactions predictions

Interaction predictions for both unfolded and folded aptamers were explored. While our primary focus was on hsa-miRNA-10b-3p, Pre_miR10b, and hsa-miRNA-10b-5p, we also examined other well-known oncogenic miRNAs as targets due to the lower computational cost of interaction prediction calculations. This include mmu-miR-21a-5p, hsa-miR-25-3p, hsa-miR-122-5p, hsa-miR-155-5p, Pre_miR10b, Pre_miR155, miR155-3p, Pre_miR122, miR122-3p, hsa-miR-25-5p, and pre-miR-25. The miRNA sequences were obtained from the miRbase database [381]. For RNA-RNA interaction prediction, *interaction.py* module from T_SELEX program was used which utilize of the IntaRNA2.0 algorithm [338]. The number of base pairs for a *seed* was set to four as default settings in T_SELEX program. Furthermore, the default settings were set for *--outMode=D*, to generate a detailed ASCII (American Standard Code for Information Interchange) chart for RNA-RNA interaction together with various interaction details which are saved automatically as individual log files. A snippet example of the code used to call the T_SELEX program is shown in below.

```
"""
Created on Mon May 01 16:01:00 2023

@author: Kabelo Mokgopa
"""

from T_SELEX import interactions
from interactions import intarna

target_seq = "ACAGAUUCGAUUCUAGGGGAAU"
target_seq.upper()
p= intarna('aptamers.csv', 'Aptamer',target_seq,"hsa-miR-10b-3p",True)
print(p)
```

Table 7.1 shows the target sequences of the miRNAs of interest that were used for interaction studies.

Table 7.1: Target names and sequences obtained from mirbase (https://www.mirbase.org).

Target Name	Target Sequences
Pre-hsa-mir-10b	5'- ccagagguuguaacguugucuaua uaUACCCUGUAGAACCGAAUUUGUGugguauccg uauagucACAGAUUCGAUUCUAGGGGAAUauauggu cgaugcaaaaacuuca -3'
hsa-miR-10b-5p	5'- UACCCUGUAGAACCGAAUUUGUG -3'
hsa-miR-10b-3p	5'- ACAGAUUCGAUUCUAGGGGAAU -3'
Pre-hsa-mir-25	5'- ggccaguguugagAGGCGGAGACUUGGGCAAUUG cugga cgcugcccugggCAUUGCACUUGUCUCGGUCUGAca gugccggcc -3'
hsa-miR-25-5p	5'- AGGCGGAGACUUGGGCAAUUG -3'
hsa-miR-25-5p	5'- CAUUGCACUUGUCUCGGUCUGA -3'
Pre-hsa-miR-122	5'- ccuuagcagagcug UGGAGUGUGACAAUGGUGUUUG ugucuaaacuaucaAACGCCAUUAUCACACUAAAUA gcuacugcuaggc -3'
hsa-miR-122-5p	5'- UGGAGUGUGACAAUGGUGUUUG -3'
hsa-miR-122-3p	5'- AACGCCAUUAUCACACUAAAUA -3'
Pre-hsa-miR-155	5'- cugUUAAUGCUAAUCGUGAUAGGGGUU uuugccuccaacugaCUCCUACAUAUUAGCAUUAACAg-3'
hsa-miR-155-5p	5'- UUAAUGCUAAUCGUGAUAGGGGUU-3'
hsa-miR-155-3p	5'- CUCCUACAUAUUAGCAUUAACA -3'
mmu-miR-21a-5p	5'- UAGCUUAUCAGACUGAUGUUGA -3'

7.2.2 Virtual screening

The aptamer dataset of 1 100 aptamers (with the length of 22 nt) used for virtual screening was generated and discussed in chapter 3 and 4. In summary, the sequences were generated using the BRA algorithm from the T_SELEX program. Secondary structure predictions were performed using the RNAfold algorithm [265], which is incorporated in T_SELEX. Tertiary

structures were predicted using the *tertiary_structure()* function from T_SELEX, which leverages the RNAComposer web server [337]. For our study, we specifically focused on the mir-10b precursor and its mature miRNA arms (5p and 3p). In terms of structural preparation, these miRNAs were prepared in a similar manner as aptamers 3D structures. In some cases, miRNA structures exhibit unfolded MFE structures (MFE = 0 kcal/mol) due to their short sequence length. Kinetically there are other folded states that exist, but this folded structures/states are not so stable (local minima with energy > 0 kcal/mol), and when optimized they unfold. The hsa-miR-10b-5p did not fold stably, therefore, we opted to use the folded structure (local minima structure) instead of the MFE structure. The virtual screening process was conducted using the *Mol_docking_calc()* function from the *Docking* module within the T_SELEX program. The provided code snippet below illustrates how to implement the virtual screening calculations in T_SELEX.

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Thu Oct 19 15:25:39 2023

@author: Kabelo Mokgopa
"""
import os
import pandas as pd

from T_SELEX import Docking
from Docking import Mol_docking_calc

home_directory = os.path.expanduser( '~' )
sofware_path = 'software/HDOCKlite-v1.1'
directory_path = os.path.join(home_directory, sofware_path)

df1 = pd.read_csv("Updatedaptamers_with_energy.csv")
r = '/home/s1800206/hdock_test/targets/hsa_miR_10b_5p.pdb'
l = '/home/s1800206/hdock_test/ligands/aptamers'
s = '/home/s1800206/software/HDOCKlite-v1.1'


p = Mol_docking_calc(data_frame= df1,MFE_column
='MFE_energiesRNAfold',receptor_name=
"hsa_miR_10b_5p",receptor=r,ligands_directory=l,directory_path=
s,Ap_folded=True)
print(p)
```

7.2.3 Post Docking Analysis (PDA)

The Post-docking calculations were conducted using the *PDCA* function from the *PDA.py* script of the T_SELEX program. The *PDCA* function follows a systematic approach after

virtual screening through several key steps. First, it initializes lists to collect docking scores from the docking simulations. It then searches in the directories checking for the model files generated during the docking process. Once it locates these files, the *PDCA* function reads their content to extract the docking scores associated with each ligand. Next, it calculates Z-scores for the extracted docking scores values, enabling a statistical assessment of docking performance relative to the entire dataset. Finally, the *PDCA* function generates several scatter plots to illustrate the relationships between docking scores, Z-scores, and fitness quality, providing valuable insights into the efficacy of the ligands.

7.2.4 Quantum Mechanicals calculations

The Single point calculations were performed for the best five models of the four best aptamers that bind efficiently to hsa-miR-10b-5p and hsa-miR-10b-3p. These calculations were carried out using semi-empirical methods due to the large size of these docked complexes. The semi-empirical method of choice was semi-empirical methods from Grimme research group which accounts for dispersion [382]. Only single point calculations were performed since we are interested in evaluating the docked complexes conformation or configuration. The two scripts were written to perform this procedure which is consist of functions designed to automate and summarize quantum mechanical calculations using the GFN2-xTB program package [383] (as shown on the sample below). The first function is responsible for performing xtb calculations on multiple PDB files of these models per aptamer docked complexes. Initially, it loads the necessary module for xtb computations on CHPC sever. Then, it iterates through PDB files named model_1.pdb to model_5.pdb, performing GFN2-xTB calculations on each file and saving the output to corresponding text files (model_1.txt to model_5.txt). The second function, *summary_xtb_results*, from *summary.sh* script summarizes the results of the self-consistent charge (SSC) calculations performed in the previous step. It extracts and show relevant properties such as total energy, HOMO-LUMO gap, dipole moment, and gradients.

```
#name of script :SSC.sh
.....
#!/bin/sh
SSC_xtb_calculations() {
    module add chpc/xtb/intel-2019u5

    for file in model_{1..5}.pdb; do
        xtb "$file" | tee "${file%.pdb}.txt"
    done

    echo "###############################################"
    echo "###############################################"
```

```
    echo "     Calculation perfectly Done Kabelo       "
    echo "################################################"
    echo "################################################"

    echo "################################################"
    echo "                Summary file                    "
    echo "################################################"
    bash summary.sh | tee Summary.log
}

#name of script :summary.sh
……

#!/bin/sh
# Function to summarize SSC calculation results from xtb output files
summary_xtb_results() {
    echo "######################################################"
    echo "######################################################"
    echo "        Summary files for SSC caclations          "
    echo "######################################################"
    echo "######################################################"
    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    echo "                  MODEl_1                            "
    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    grep energy model_1.txt
    grep -m -1 HOMO-LUMO model_1.txt | head -1
    grep -3 dipole model_1.txt
    grep gradient model_1.txt

    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    echo "                  MODEl_2                            "
    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    grep energy model_2.txt
    grep -m -1 HOMO-LUMO model_2.txt | head -1
    grep -3 dipole model_2.txt
    grep gradient model_2.txt

    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    echo "                  MODEl_3                            "
    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    grep energy model_3.txt
    grep -m -1 HOMO-LUMO model_3.txt | head -1
    grep -3 dipole model_3.txt
    grep gradient model_3.txt

    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    echo "                  MODEl_4                            "
    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    grep energy model_4.txt
    grep -m -1 HOMO-LUMO model_4.txt | head -1
    grep -3 dipole model_4.txt
    grep gradient model_4.txt

    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    echo "                  MODEl_5                            "
    echo "::::::::::::::::::::::::::::::::::::::::::::::::::::::"
    grep energy model_5.txt
```

```
    grep -m -1 HOMO-LUMO model_5.txt | head -1
    grep -3 dipole model_5.txt
    grep gradient model_5.txt

    echo "::::::::::::::::::::::::::::::::::::::::::::::::::"
    echo "                summary file completed             "
    echo "::::::::::::::::::::::::::::::::::::::::::::::::::"
}
```

7.2.5 Molecular Dynamics

The same complexes that were studied using Quantum mechanics were further studied using molecular dynamics. The complexes were solvated with water molecules under periodic boundary conditions, neutralized and subjected to dynamics calculations using md.py script from the extension program called PySMQM. PySMQM was developed during our studies to perform automated molecular dynamics simulations for RNA-RNA and RNA-protein complexes. We named it PySMQM, which stands for Python Statistical Mechanics and Quantum Mechanics, because it was specifically designed to investigate the folding kinetics of RNA at the quantum mechanical level, guided by partition functions. Since it is not yet published on git hub the source code of PySMQM can be found in the supporting material (C2). The md.py script is full automation procedure based on GROMACS 2018.8 [331].

Using the pySMQM, the RNA structures from the docked complex were processed to generate a topology and structure file using Amber forcefields. Subsequently, the system solvation box size was set to 1.5 angstrom with a cubic shape, followed by a solvation process with the TIP3P solvent model. Ions were added for system neutralization. The simulation commenced with energy minimization, followed by equilibration, and finally the production of MD steps. This comprehensive workflow ensured the setup and execution of a molecular dynamics simulation for the investigated nucleic acids systems. Over 40 model complexes were investigated for 100 ns.

7.2.5 (i) Stability metric (proposition and theory)

RNA complexes are attributed to high RMSD values because of their high degree of fluctuation dynamic structures [384]. Some short RNA showed to have RMSD values above 3 nm, and high RMSD values are due to their hairpin loop and long exterior loop(tail) [385]. Unlike proteins RNAs are not that structural compact since they do not for alpha and beta sheets [386]. RNA RMSD values fluctuate a lot making it difficult to analyse and interpret their stability. To address this, we introduce an algorithm that calculates the degree of stabilization of the RMSD, denoted as τ. The degree of stabilization is defined as the total number of detected changepoints

(significant changes) in the RMSD time series. The algorithm begins by detecting changepoints using the PELT (Pruned Exact Linear Time) algorithm. This method identifies points in the time series where statistical properties change, indicating structural transitions within the RNA complex. In PELT algorithm we utilise a l2 model, implying a least-squares loss function to quantify the fit of the data within each segment [387]. The minimized cost function C in the PELT algorithm combines the goodness of fit of the model within segments and a penalty for adding new segments. Mathematically, this cost function is expressed as [387]:

$$C = \sum_{i=1}^{m} \left(\sum_{t=\tau_{i-1}+1}^{\tau_i} (x_t - \mu_i)^2 \right) + \beta m \qquad (7.1)$$

where x_t are the RMSD values, μ_i is the mean of segment i, τ_i are the changepoints, m is the number of segments, and β is the penalty term. The penalty parameter β determines the trade-off between model fit and complexity, with higher values resulting in fewer detected changepoints. By detecting changepoints and calculating τ, the algorithm provides a quantitative measure of the system's stability. The detected changepoints are visualized on the RMSD plot, marking significant structural transitions. The degree of stabilization τ summarizes the stability behaviour of the RNA complex, with more changepoints indicating more frequent transitions or phases within the simulation.

It is important to note that the degree of stabilization is related to significant changes in RMSD values as stabilization occurs. However, this approach does not account for the high RMSD values. To account the high RMSD values in our metric approach, we propose calculating the RMSD area, which represents the cumulative deviation over time. This metric integrates the RMSD values across the entire simulation providing a comprehensive measure of how much the biomolecular structure fluctuates. For RNA, in particular, the RMSD value reflects the dynamic nature of its structure. Lower RMSD values indicate less deviation from the reference structure, suggesting greater stability and closer adherence to the intended conformation. Conversely, higher RMSD values signify more movement or conformational changes, which could indicate dynamic structural transitions or flexibility. To calculate area under the RMSD we used integration methods like Simpson's rule. The Simpson's 1/3 rule is attributed to Thomas Simpson, this method is a numerical integration method based on quadratic interpolation [388]. The basic Simpson's 1/3 rule is evaluated based n=2. This rule approximates the integral of a function $f(x)$ over an interval $[a,b]$ as follows [388]:

$$\int_a^b f(x)\ dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

$$= \frac{1}{3}h[f(a) + 4f(a+h) + f(b)] \qquad (7.2)$$

However, the downside of the using the normal 1/3 Simpson rule is inaccuracy when dealing with oscillating function, this Simpson's rule may yield poor results. To address this, the interval [*a*,*b*] is divided into *n* subintervals (*n*>2), and Simpson's rule is applied to each subinterval. This approach is known as the composite Simpson's 1/3 rule, it combines the results to approximate the integral over the entire interval [389, 390]. For an even number of subintervals, *n*, the composite Simpson's rule is expressed as:

$$A_{rmsd} = \int_a^b f(x)\ dx \approx \frac{1}{3}h\sum_{i=1}^{n/2}[f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})]$$

$$= \frac{1}{3}h[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)]$$

$$= \tfrac{1}{3}h\left[f(x_0) + 4\textstyle\sum_{i=1}^{n/2} f(x_{2i-1}) + 2\sum_{i=1}^{n/2-1} f(x_{2i}) + f(x_n)\right]. \qquad (7.3)$$

Where A_{rmsd} is the area of RMSD of interest when simulated with ligand and/or without the ligand and x is the time throughout the simulation.

The error associated with the composite Simpson's rule is expressed as:

$$-\frac{1}{180}h^4(b-a)f^{(4)}(\xi) \qquad (7.4)$$

Where $h = \frac{b-a}{h}$ denotes the "step length" and ξ is some number between step length. This error is bounded (in absolute value) by [390]:

$$\frac{1}{180}h^4(b-a)\max_{\xi\in[a,b]}\left|f^{(4)}(\xi)\right| \qquad (7.5)$$

Finally, to comprehensively assess structural stability in our RNA molecular dynamics simulations where both high RMSD and significant fluctuation or change in the RMSD, we

propose a combined metric called stability metric (μ) that integrates both the temporal evolution of RMSD and the number of identified changepoints. This metric is defined as:

$$\mu = \alpha \cdot \tau + (1 - \alpha) \cdot A_{\mathrm{RMSD}} \qquad (7.6)$$

Here, α serves as a weighting factor to balance the influence of total number of changepoints τ, indicating structural transitions, and the RMSD area (A_{RMSD}), which represents the cumulative deviation over time. A higher value of stability metric suggests a system with less pronounced stabilization, whereas a lower value may indicate greater structural stability. Notably, for longer RNAs, there is no predefined threshold where RMSD stabilizes due to the inherent complexity and variability in molecular dynamics, making our integrated approach crucial in capturing nuanced structural dynamics.

7.2.5 (ii) Autocorrelation Function (proposition and theory)

An alternative way to determine the degree of stabilization of the RMSD, we can use the Autocorrelation Function (ACF). The ACF quantifies the correlation between values of a time series at different time lags [391]. The autocorrelation function $R(k)$ of a time series $X(t)$ at lag k is calculated using the formula [391]:

$$R(k) = \frac{1}{N-k} \sum_{t=1}^{N-k} (X(t) - \mu)(X(t+k) - \mu) \qquad (7.7)$$

The autocorrelation function $R(k)$ is plotted against lag k. The plot shows how correlated the noise is at different time lags. If the autocorrelation function decays quickly to zero, it indicates that the noise is uncorrelated, suggesting stabilization in the RMSD. Conversely, if the autocorrelation function remains high at longer lags, it implies persistent correlations in the noise, indicating instability in the RMSD.

7.2.6 MM-GBSA

The MM-GBSA calculations were performed to evaluate binding energies for 40 aptamer models complexes, where miR-10b-3p and miR-10b-5p are the targets using GROMACS [311]. For each model, trajectory, topology, and index files from molecular dynamics simulations were utilized. To enhance accuracy, the dielectric constant was increased to 100 to better accommodate the highly charged backbone of these RNA and account for structural

fluctuations. Energy calculations were based on the last 5000 frames of the trajectory to ensure the results reflected stable and well-equilibrated systems. The analysis comprised several stages: first, the potential energy *in vacuo* was computed, followed by the calculation of polar solvation energy. Non-polar solvation energies were assessed using SASA and SAV models, with additional calculations performed for the WCA model. Finally, a detailed statistical analysis was conducted to summarize the binding affinities and stability of the aptamer-target complexes.

7.3 Results and discussion

7.3.1 Interaction predictions

To understand molecular interactions, RNA-RNA interactions stand as a starting point. In many biological processes, these predictions serve as a roadmap in guidance through the complex network of RNA-RNA interactions. In this study, the outcomes of the investigation based on the interactions between the library of aptamers and the target miRNAs of interest at a large scale are presented.

7.3.1.1 Interaction energies

Interaction energies were analysed and presented as a heatmap as shown in **Figure 7.1**. With regards to the heatmap in **Figure 7.1**, the darker the colour, the more negative the interaction energy. Negative interaction energy typically indicates favourable interactions between two RNA sequences in this context. Furthermore, it suggests that hybrid interactions contribute to stabilizing the system, indicating that the attraction between the aptamer and the target sequence is thermodynamically favourable.

The major contributions to the favourable thermodynamic interactions are the loops energy and hybridization energy [338]. Hybridization quantifies as the overall stability of the RNA-RNA interaction by taking into account the contributions from base pairing, loop formation, and other structural elements [338]. The loop energy represents the free energy associated with the formation of these unpaired or non-canonical base pair regions while the hybridization energy represents the energy associated with the overall hybridization of the target and query (aptamer) sequences. There are other energy contributions, including the seed energies, initiation energy, dangling, and end energies. The quality of base pairing, such as GC and AU pairs, typically contributes more significantly to the interaction energy than the quantity of base pairing. This means that even with fewer base pairs, if those pairs are of high quality (e.g., GC or AU), they

can contribute significantly to the stability of the interaction [338]. Conversely, if there are many mismatches or non-optimal base pairs, they can reduce the overall interaction energy or even result in a net energy of zero or unfavourable energy. Mismatches or non-canonical base pairs can disrupt the formation of stable secondary structures or the hybridization of two RNAs and weaken the overall stability of the interaction. In extreme cases, a high number of mismatches or non-optimal base pairs can lead to the destabilization or complete restraint of the RNA-RNA interaction [338]. While the number of base pairs are important, the quality of those base pairs and the overall structural content (e.g., loop formation, dangling ends) play crucial roles in determining the stability and favourability of RNA-RNA interactions [338]. Relatively weak interaction or structural stability can be seen by light yellow colour on the heatmap. The heatmap shows that the aptamers seem to perform very well against hsa-miR-25-5p compared to the rest of the targets. It is crucial to understand that each aptamer behaves differently when interacting with different target sequences. This variation becomes evident when examining the heatmap from left to right, where some targets appear as light yellow, while others show darker shades of yellow or other colours. A consistent pattern is not observed for each aptamer across all targets. This lack of consistency comes from the complex nature of the interactions, which are based on the specificity of the aptamers and targets sequences. Since aptamer sequences are generated randomly, each interaction between an aptamer and a target is unique, leading to a lack of observable patterns in the heatmap.

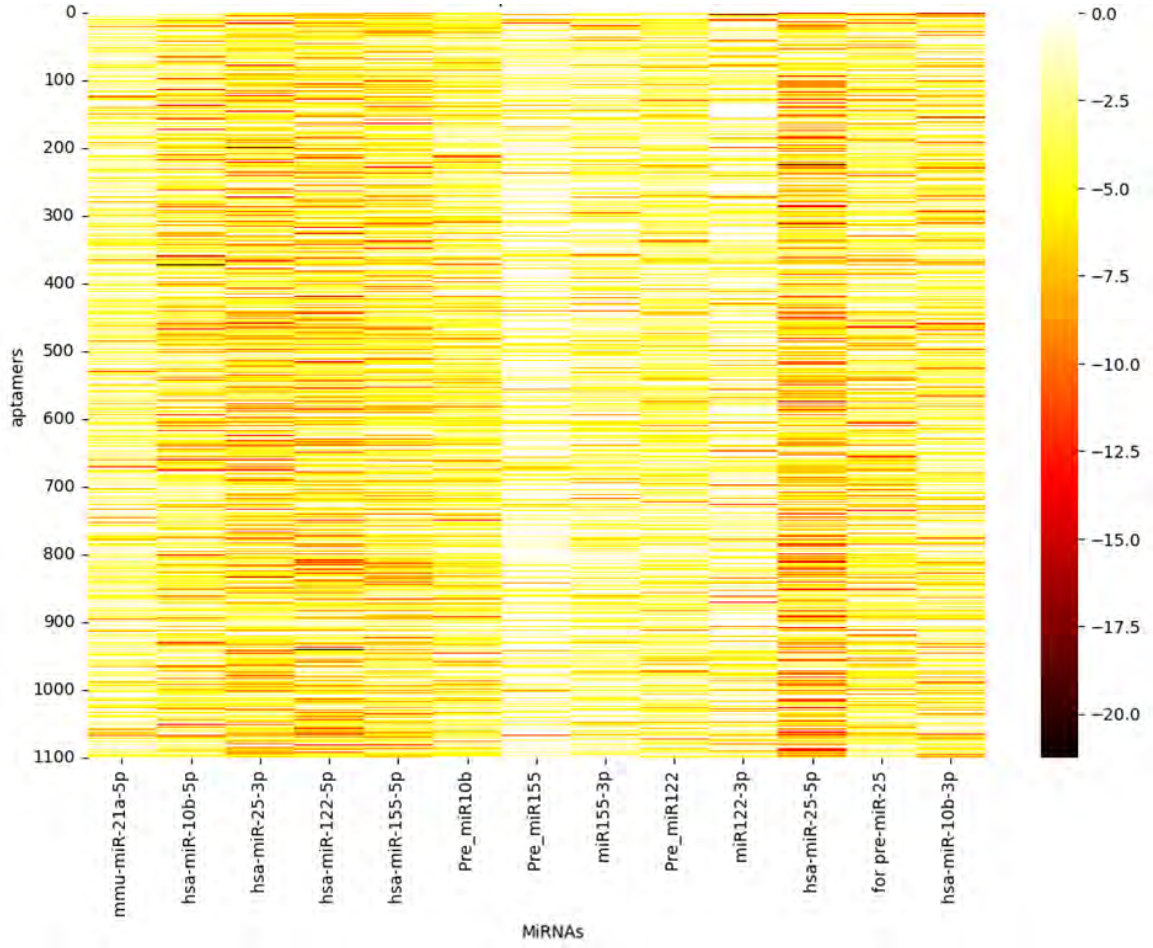


Figure 7.1: Heatmap of large scale predicted interaction energies of aptamers against multiple oncogenic miRNAs, including the premature miRNAs, 5p and 3p mature arms.

It is very difficult to outline which targets follows (where aptamers perform very well) after hsa-miR-25-5p based on the heatmap in **Figure 7.1**. To gain better insight we computed the average interactions energies for each target as reported in **Table 7.2**. Based on the average energies, the aptamers perform exceptionally well against hsa-miR-25-5p, which has the strongest interaction energy at -5.013 kcal/mol. This is closely followed by hsa-miR-122-5p at -4.979 kcal/mol and hsa-miR-155-5p at -4.407 kcal/mol, making them as the top three candidates for further investigation. On the other end of the spectrum, the least favorable interactions were observed with pre_miR155, exhibiting an average energy of -1.676 kcal/mol, followed by mmu-miR-21a-5p at -2.392 kcal/mol and miR122-3p at -2.327 kcal/mol.

Even though 3p arms are slightly complementary to 5p arms (in terms of the bases in the sequences), it is important to highlight that in this dataset, the miRNAs with the 5p generally have low average interaction energies (more negative) compared to those with the 3p. This suggests that the 5p mature miRNA sequences might have stronger binding affinities or more

stable interactions with the most aptamers compared to 3p arms. Since our interest was specifically on mir-10b pre-curser and its mature miRNA arms (5p and 3p). Discussing that, it is evident that Pre-miR10b exhibits a slightly higher energy value (-3.598 kcal/mol), implying a hybridization configuration is somewhat less stable compared to its mature 5p arm counterpart. Moving to the mature miRNA arms, hsa-miR-10b-5p stands out with the lowest average interaction energy value (-3.883 kcal/mol), indicating to strongly interact with aptamers within the dataset interaction. Hsa-miR-10b-3p is ranked between Pre-miR10b and hsa-miR-10b-5p which demonstrates an intermediate average interaction energy value of -3.509 kcal/mol. This suggests that aptamers bind more stably to hsa-miR-10b-3p compared to Pre-miR10b, but less stably compared to hsa-miR-10b-5p.

Although we use the average energies to rank the performances of the aptamers per target, it is important to note this may yet still pose limitations in representing the entire dataset comprehensively. Skewed distributions influenced by outliers can distort the average energy values affecting its accuracy in reflecting the central tendency of the data. Moreover, condensing data into a single value result in a loss of information regarding variability and distribution, potentially masking significant patterns or trends. But in this context the averages tend to completely support the heatmap which give insight into the individual aptamer and its interaction energy.

Table 7.2: The average interaction energies of the aptamer dataset against each target.

Targets	Average interaction energies (kcal/mol)
mmu-miR-21a-5p	-2.391
hsa-miR-10b-5p	-3.882
hsa-miR-25-3p	-5.013
hsa-miR-122-5p	-4.978
hsa-miR-155-5p	-4.407
Pre_miR10b	-3.597
Pre_miR155	-1.676
miR155-3p	-2.467
Pre_miR122	-3.076
miR122-3p	-2.326
hsa-miR-25-5p	-5.643
pre-miR-25	-3.731
hsa-miR-10b-3p	-3.509

7.3.1.2 Correlation analysis

To understand relationships between variables in a dataset, we further looked at the correlation heat maps. The correlation coefficient, ranging from -1.0 to +1.0, serves as a quantitative

measure of the strength and direction of the relationship between two variables [392]. When the correlation coefficient approaches 0, it suggests that there is little to no linear relationship between the variables under consideration. Conversely, as the correlation coefficient moves closer to +1.0 or -1.0, it indicates a stronger linear relationship between the variables. A value near +1.0 suggests a positive correlation, meaning that as one variable increases, the other tends to increase as well. Conversely, a value near -1.0 indicates a negative correlation, signifying that as one variable increases, the other tends to decrease.

In our study, a clustered correlation heatmap analysis as employed to investigate if the targets perform similar in terms of interaction energies within the aptamer dataset as shown in **Figure7.2**. The goal was to find any patterns or connections between the aptamers and how they interact with miRNAs. By creating a visual clustered correlation heatmap (**Figure 7.2**), we can assess whether this aptamer library behaves similarly across the different targets.

The heatmap shows that there is no significant positive or negative correlation. This suggests that there may be no direct linear relationship between the interaction energies within aptamer. This further emphasis that these unique aptamers for the BRA generate library bind uniquely and specific in every target. However, it is important to note that even if there is no significant correlation between variables, it does not mean there is no relationship at all. There was no correlation observed between the miRNAs that is above 0.7. Most are even less than 0.5. Furthermore, even though 3p and 5p arms have sightly complementary sequence to each other, there is no close relationship, suggesting that the aptamer sequences behave even differently in the 5p and 3p arms of the same pre-miRNA to some extent.

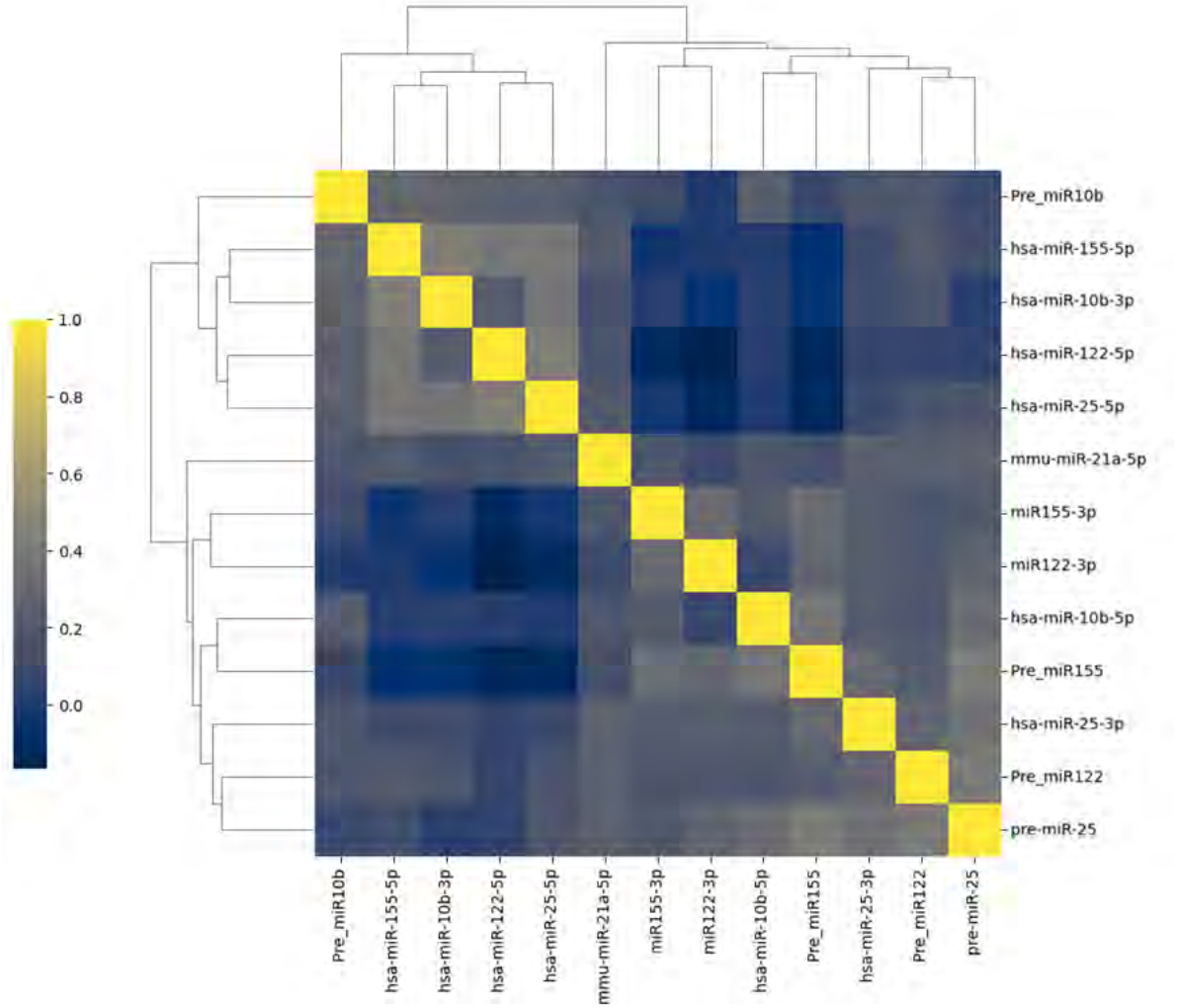


Figure 7.2: Determination of correlation heatmap of miRNAs based on interaction energies with aptamers.

7.3.1.3 Clustering of the aptamers within the dataset

Principal Component Analysis (PCA) was done to simplify a large dataset while retaining significant patterns and trends. PCA reduces the dimensionality of data by transforming a set of correlated variables into a smaller set of uncorrelated variables called principal components [392,393]. While the previous hendrical clustered heatmap was looking at the at the clustering based on the targets, with PCA method the aptamers are clustered based on their interaction energies across all targets. This provide a possibility to evaluate if there are aptamers that behave in similar manner in terms of interaction energies across all 13 targets. To further improve the clustering, KMeans was used on the PCA-transformed data to reduce dimensionality and enhance cluster separation. By leveraging the principal components that capture the most significant variance in the data, KMeans can more effectively identify clusters and assign data points. **Figure 7.3 (A)** and **(B)** shows two different ways of grouping data

points in a aptamer dataset based on the interaction energies with multiple different miRNA targets. Using Principal Component Analysis (PCA) in **Figure 7.3 (A)**, three clusters are observed in the top-left scatter plot, labelled as Cluster 0, Cluster 1, and Cluster 2. Looking at Cluster 2, the data points are scattered far away from each other, suggesting a greater variability or dispersion within this cluster. Additionally, Cluster 2 has fewer data points compared to the rest, indicating that there are less aptamers in this cluster from the dataset. Conversely, Cluster 0 and Cluster 1 exhibit data points that are grouped closer together, suggesting higher cohesion within these clusters. Both Cluster 1 and Cluster 2 have their data points overlapping with Cluster 0, but Cluster 1 and Cluster 2 do not overlap with each other, indicating distinct patterns or characteristics separating these aptamers from each other. This suggests that while aptamers in Cluster 0 share some similarities in terms of interaction energy with both Cluster 1 and Cluster 2, Cluster 1 and Cluster 2 have unique features or interaction energies with the targets that differentiate them from each other and from Cluster 0.

KMeans was further utilized to expand the number of clusters from 3 to 5 to extend the partitioning of overlapping data points observed in the initial PCA clustering into new groupings (**Figure 7.3 (B)**). This suggests that the aptamer dataset possesses additional underlying interaction patterns that were not fully captured by the original clustering with three clusters. For KMeans, cluster three exhibits a notable abundance of data points found in region 1 (R1). Despite the addition of two extra clusters with KMeans, overlapping data points still persist indicating complex interaction relationships within the aptamer dataset to the targets. It is important to note the presence of region 1 and region 2 (R2), suggesting that if the objective was to obtain only two clusters, the points within these regions would likely belong to distinct clusters. This suggests that the data points within region 1 and region 2 tend to follow a somewhat linear pattern, indicating a potential underlying interaction correlation between for some aptamers. Finally, although the all aptamers did not display a linear relationship through correlation analysis, it cannot be disputed that there exists a relationship, even if it is not linear. The bottom-left bar graph (**Figure 7.3 (C)**) shows how the data points are distributed among the three clusters from the PCA clustering, while the bottom-right bar (**Figure 7.3 (D)**) graph illustrates the distribution across all five clusters from the KMeans clustering.

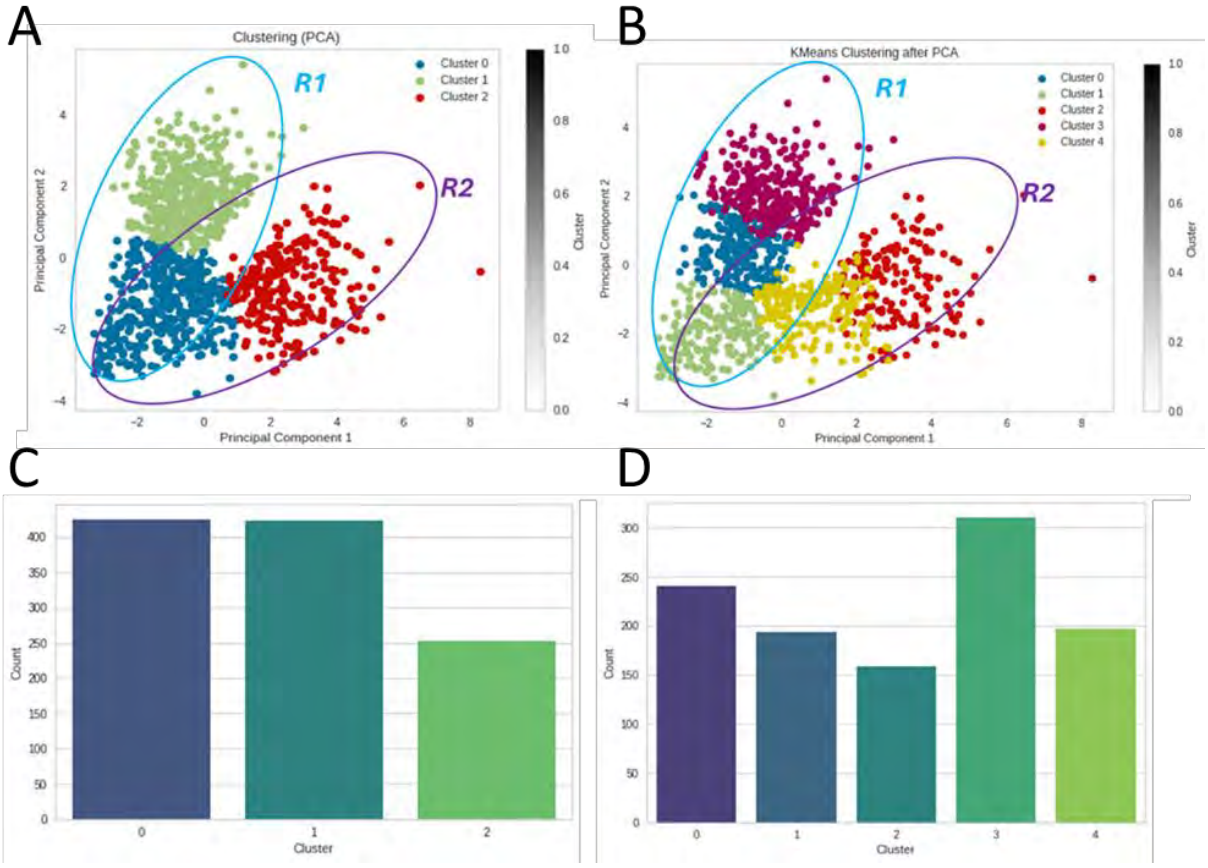


Figure 7.3: Principal Component Analysis (PCA) and KMeans clustering analysis of aptamers based on the on the interaction energies with multiple oncogenic miRNAs.

7.3.1.4 Evaluating best interacting aptamers for the mir-10b pre-curser and its mature miRNA arms

The interactions were analysed and reported for the four best ranked aptamers against mir-10b pre-curser and its mature miRNA arms (5p and 3p). The aptamer raking system was based on the interaction energies against the individual targets. Starting with the pre-hsa-mir-10b, the best four aptamers that form more stable complex are aptamer813, 212, 893 and 750 with the interaction energy of -12.23 kcal/mol, -11.9 kcal/mol, -11.71 kcal/mol and -11.51 kcal/mol respectively.

In comparing the interactions of pre-hsa-mir-10b with aptamer813 and aptamer212 in **Figure 7.4**, both interactions predominantly exhibit Watson-Crick base pairs, signifying strong sequence complementarity conducive to stable RNA-RNA interactions. However, the interaction of Pre-hsa-mir-10b with aptamer 813 demonstrates an extended stretch of consecutive Watson-Crick base pairs, indicating less disruptions and high complementary base pairing. Alternatively, in the interaction of pre-hsa-mir-10b with aptamer212, while Watson-

Crick base pairs are prevalent, small differences such as a shorter overlapping region and potential structural distortions are observed. This may possibly be the ones contributing to variations in interaction stability. Additionally, both interactions feature a few mismatched nucleotides, suggesting occasional mispairing events that could compromise overall stability. Moreover, the presence of wobble pairings such as G-U pairs, introduces flexibility and influences the structural dynamics of RNA-RNA duplexes in both interactions. The same could be said about aptamer893 as its complementary base pairs are fewer compared to aptamer212, and more interesting, there are no wobble base pairs detected. Aptamer750 has a guanine (G) nucleotide in the target sequence that creates a mismatch during complementary base pairing, causing a bend or disruption in the structure. Which may result in a decrease in the interaction energy.

The top 4 aptamers interacting with the hsa-mir-10b 5p and the 3p arms in **Figure 7.4** form more stable duplex complex with the target (their interaction energies are lower). The best interacting aptamers against the 3p arm are aptamer155, 1, 533 and 1027 with the interaction energies of -15.5 kcal/mol, -15.1 kcal/mol, -12.63kcal/mol and 9.86 kcal/mol respectively. On the other hand, for the 5p arm the best interacting aptamers are aptamer373, 1083, 144 and 360 with the interaction energies of -19.36 kcal/mol, -15.63 kcal/mol, -15.29 kcal/mol, and -13.46 kcal/mol respectively. A similar pattern is also noted concerning the interactions, as evidenced in 3p. As the number of complementary base pairings decreases, the interaction energies also decrease. Moreover, the occurrence of mismatches during pairing leads to a significant decrease in interactions energies, as these disruptions within the pairings may weaken the hybridization of the two sequences.

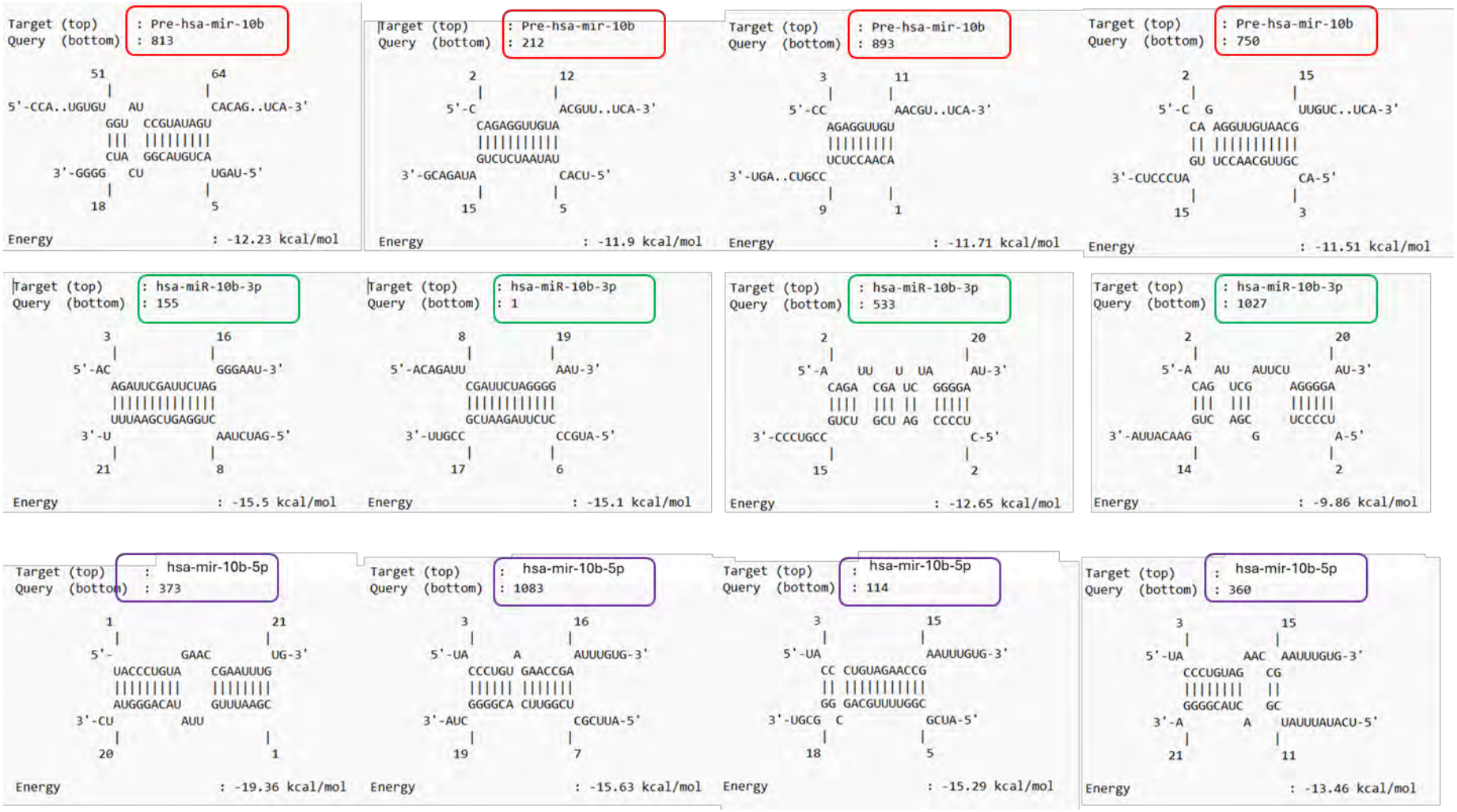


Figure 7.4: Detailed interactions and energies of the top four aptamers with pre-mature miR-10b, miR-10b-3p, and miR-10b-5p.

In order to understand the thermodynamics of RNA-RNA interactions that are essential. We further visualize the heatmap to indicate the minimal energy for each intermolecular index pair in RNA-RNA interactions predicted by IntaRNA2.0 [338]. This heatmaps are reported in **Figure 7.5**. This approach allows for the identification of exclusive or overlapping regions where alternative interactions can occur. Considering the heuristic nature of IntaRNA [338], the provided energies serve as close upper bounds, with only interactions featuring a seed considered for visualization. Energies below or equal to 0 kcal/mol are represented in the heatmap, while missing data are assigned an energy value of 0 kcal/mol (red).

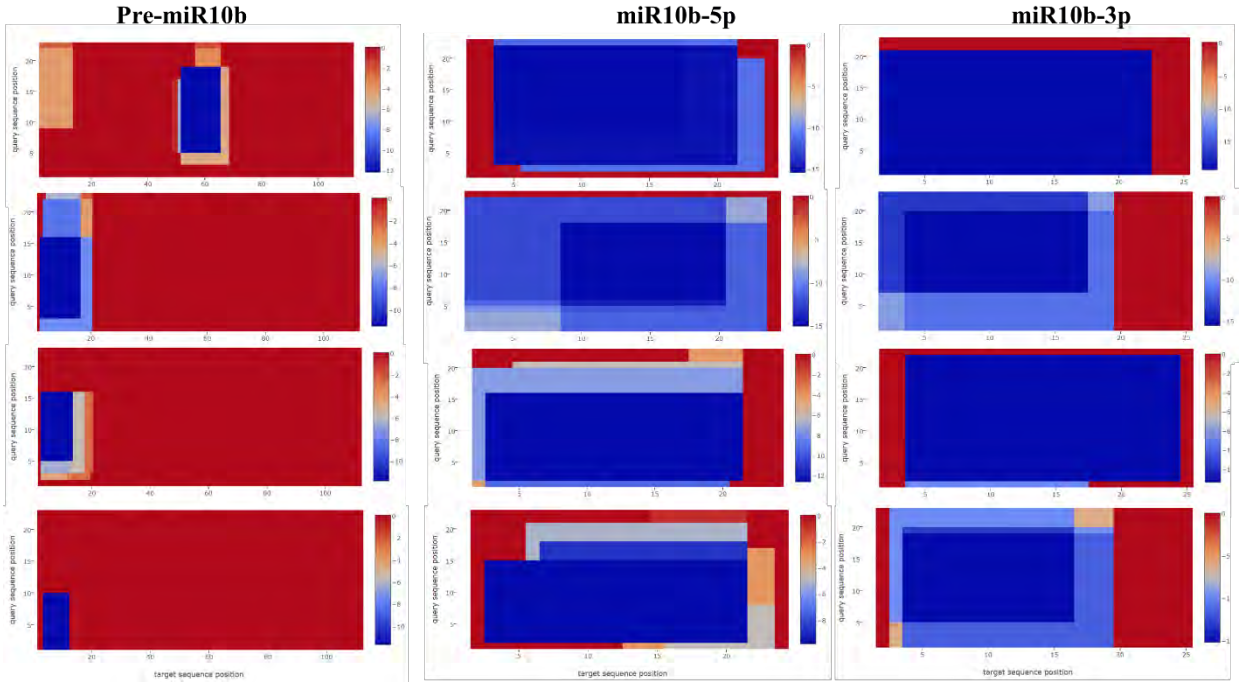


Figure 7.5: Position-wise minimal energy profiles/heatmaps of the top four aptamers with pre-mature miR-10b, miR-10b-3p, and miR-10b-5p.

Looking at the blue sections in the heatmaps reveals areas with predominantly negative energy values. In the pre-miR-10b heatmap, these blue areas appear relatively small. This is attributed to the large length of pre-miRNA sequences compared to the 22-nucleotide aptamer sequences. Therefore, there are fewer bases within the pre-miRNA sequence available for interaction with the aptamer, resulting in smaller blue rectangles surrounded by larger red regions. Conversely, in the heatmaps for the 5p and 3p arms, the blue regions are larger and occupy a larger portion of the heatmap, with fewer adjacent red areas. This observation suggests the presence of missing data, which has been filled with zeros. Upon closer evaluation of pre-miRNA-10b, it appears that aptamer813 interacts primarily with the middle bases (51 to 64 nt) of the pre-miRNA, while other aptamers interact with the initial bases spanning from 3 to 23.

Additionally, apart from the distinct blue and red regions, there are intermediate-coloured regions indicating varying degrees of energy contributions between the extremes of blue and red. These intermediate-coloured regions play a significant role in the overall interactions of the system. For example, when examining the 3p arm, it is observed that aptamer 1 exhibits a small dark region compared to the other regions, while the surrounding areas appear to be of a mid-blue shade. This observation further highlights the importance of these mid-coloured regions in contributing to the overall interaction of the system. Finally, it is evident in both the 3p and 5p arms that almost the entire sequences contribute to the interactions, as there are fewer missing values or data (represented by red regions). This observation highlights the comprehensive nature of the interactions across these sequences since the aptamers have the same number of bases as the targets (3p and 5p).

7.3.2 Virtual screening

7.3.2.1 Docking scores

Virtual screening of molecules for a large dataset is crucial for several reasons. Firstly, it allows for the rapid and cost-effective identification of potential drug candidates, accelerating the drug discovery process [394]. Secondly, virtual screening enables the exploration of a vast chemical space, providing access to a diverse range of compounds that may exhibit therapeutic effects [395]. Thirdly, it allows researchers to prioritize molecules with the highest likelihood of binding to a specific target, thereby increasing the efficiency of subsequent experimental validation [396]. Here, we focus on the virtual screening of aptamers against miRNAs. All the aptamers were docked against the miR-10b precursor and its mature RNAs, and the docking scores of the first model were compared across the targets to assess the behaviour of aptamers across different miRNA targets. The results are reported in **Figure 7.6**

It is important to highlight a few things as we deal with RNA aptamers docked against RNA targets. This is an interesting phenomenon, as both targets and ligands are RNAs. The tricky part about RNAs or nucleic acids is that they lack an active site to dock the ligand in the traditional sense observed with proteins. This suggests that the interaction between RNA molecules and ligands is more complex and may involve different mechanisms compared to protein-ligand interactions. Moreover, RNAs can adopt diverse secondary and tertiary structures, further complicating the docking process.

Figure 7.6 shows the plots based on the docking scores of the first models of each aptamer, it can be seen that aptamers perform variably well across all targets. However, what is important to note is that first models shows that the aptamers follow the same trend in all targets. We do not fully understand the factors contributing to this consistent trend. This trend mimics as simple cubic function such as $f(x)=x^3$, this indicates a nonlinear relationship between the docking score and the aptamer dataset of interest. The curvature of the plot suggests that the docking score responds in a non-proportional manner to changes in the aptamers ranked from the best to the least. Further investigation is needed to uncover the underlying mechanisms driving the observed trends in aptamer performance across these three different targets. In terms of docking scores, the lower the score, the better the binding affinity between the aptamer and its target, indicating a stronger interaction. This suggests that aptamers perform remarkably well against mature miRNAs, demonstrating high binding affinity. Specifically, aptamers have a great performance against 3p arms, followed by 5p arms, with the precursor miRNA showing the least favourable binding. This contradicts the observation made during the interaction studies where aptamers seem to interact very well with 5p arm over 3p arm. It is important to note that with interaction studies the conformation of these aptamer is not taken into account. This surprising finding also contradicts the expectation that the larger precursor miRNA structure might be more susceptible to binding due to potential larger structure that can server as huge surface area for binding. However, the results indicate that the arms of the precursor miRNA are actually more conducive to binding with aptamers. For instance, the best-docked aptamer against miR-10b_3p demonstrated a docking score of approximately -550, while for miR-10b_5p, the best aptamer achieved a score of approximately -500. It is intriguing to note that while 3p is slightly complementary to 5p, the complex RNA folding dynamics lead to these significant structural disparities between their 3D conformations despite their similar sequences. These differences in 3D structure results from various factors, including variations in base pairing, and loop configurations affecting their functionality.

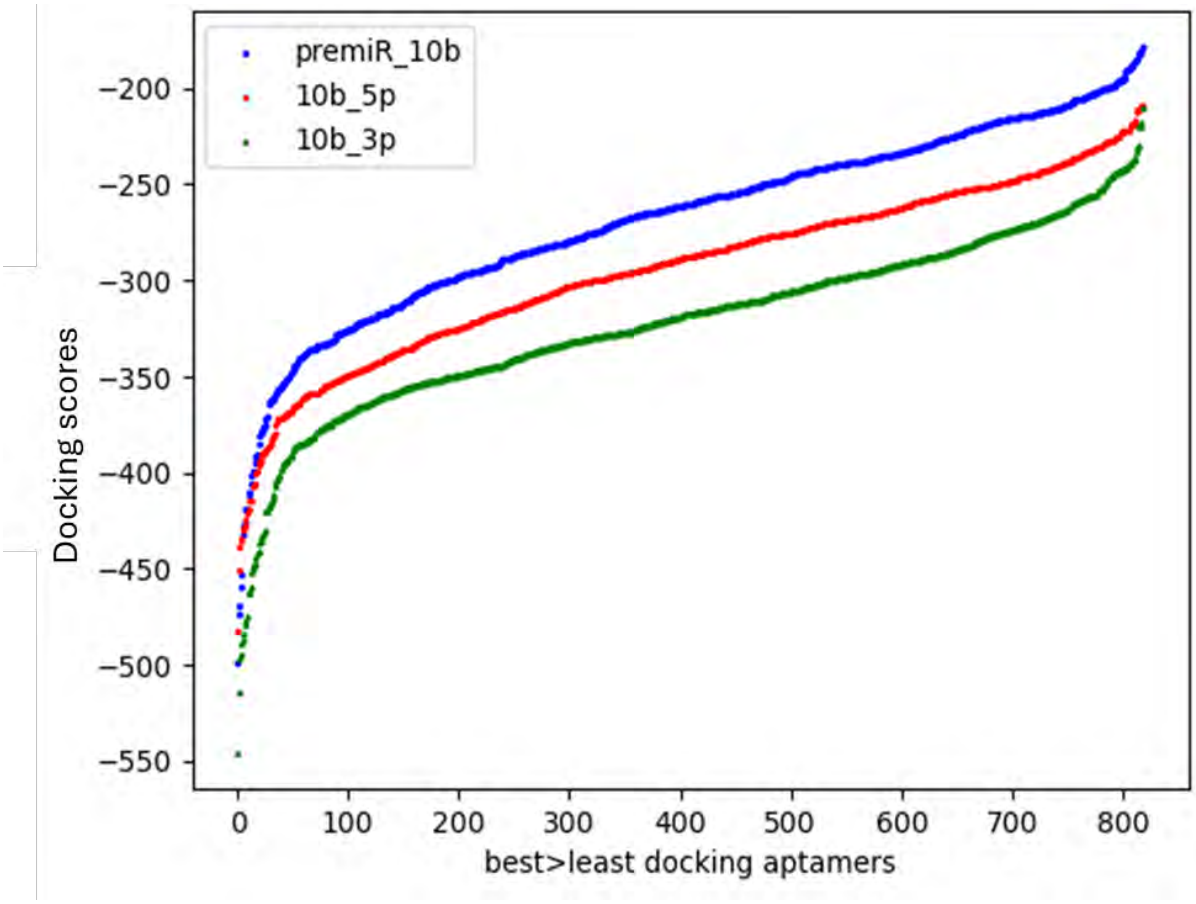


Figure 7.6: The docking scores results from best to least aptamer against the pre-mature miR-10b, miR-10b-3p, and miR-10b-5p.

The snippet molecular docking results, ranked based on the docking score of the first model, are presented in **Table 7.3** for pre-MiR10b and its mature RNA 5p and 3p. When examining the 3p arm, aptamer557 appears to be the most stably docked aptamer, with a docking score of -545.96. This suggest that the aptamer577 model performs very well against the target mir-10b-3p arm. This is followed by aptamer274, 274, 279, and 128, with docking scores of -514,52, -496.96, -494.57, and -489.86 respectively. Interestingly, there is a significant difference between the top-ranked aptamer and the second-ranked one in terms of docking scores, suggesting that the first ranked aptamer is outperforming the rest of the aptamers, including the second-ranked. However, when considering fitness scores, they do not proportionally correlate with the docking scores. For example, aptamer274 has a lower docking score than aptamer734, yet their fitness quality scores are opposite. This suggests that the overall models of aptamer734 perform better compared to those of aptamer274, despite the first model ranking suggesting otherwise. The confidence scores were also reported in **Table 7.3**. The confidence score in molecular docking plays a crucial role in assessing the likelihood

and reliability of the predicted interaction between a ligand and a protein [301]. When the confidence score is above 0.7, it indicates a strong probability of binding with the predicted binding pose considered highly reliable [301]. In cases where the score falls between 0.5 and 0.7, there is a moderate likelihood of binding, although some uncertainty remains regarding the accuracy of the binding pose [301]. Conversely, a low confidence score (below 0.5) suggests a reduced probability of binding, with the predicted binding pose potentially being less dependable [301]. Looking at the confidence scores from the docking results from **Table 7.3**, they are all above 0.99 (99%), suggesting the high reliability and accuracy of the predicted binding interactions and poses.

Moving to miR-10b-5p arms the best performing aptamer is aptamer899 with the docking scores of -482.55 followed by aptamer536, 413, 332, and 278 with the docking scores of -450.55, -439.15, -435.22, and -434.11 respectively. In the 3p arm, a clear distinction exists between the highest-ranked aptamer and the second-ranked one in terms of docking scores. Interestingly, this difference persists even in the 5p arm. Notably, the second-ranked aptamer (aptamer536) for 5p have significantly lower fitness quality compared to both the top-ranked (aptamer899) and third-ranked (aptamer413) aptamers. This suggests that while aptamer 536 initially ranks second based on the docking score, its overall performance across various models is better than that of aptamer899. Additionally, aptamers536 and 413 demonstrate similar fitness quality score, even though fitness quality scores are central mean values that generalize individual model variations into a uniform behaviour metric. Upon evaluating the confidence scores of the 5p arm, it becomes evident that they are slightly lower when compared with one of the ligands against the 3p arm. This difference arises because there is a consistent correlation between the docking scores and the confidence scores. Reason being that confidence scores are calculated from the docking scores. This can be seen as docking scores of the ligand docked against the 3p arm are lower than those against the 5p arm.

The most effective aptamers for the pre-causer miRNA are ranked as follows; aptamer895, followed by aptamer465, and then Aptamer1085. It is worth noting that **Table 7.3** provides only a partial view of the results, while the complete table can be accessed on GitHub (https://github.com/KPMOKGOPA/MSc_supplementary_data/tree/main/Docking_results).

Table 7.3: The molecular docking results of the top 14 best docked aptamers against the pre-mature miR-10b, miR-10b-3p, and miR-10b-5p.

10b_3p				10b_5p				Pre_miR10b			
ligand	**Docking Score**	**Fitness quality**	**Confidence Score (%)**	**ligand**	**Docking Score**	**Fitness quality**	**Confidence Score (%)**	**ligand**	**Docking Score**	**Fitness quality**	**Confidence Score (%)**
aptamer577	-545,96	-336,6	99,964	aptamer899	-482,55	-207,37	99,871	aptamer895	-499,35	-227,14	99,908
aptamer274	-514,52	-319,87	99,932	aptamer536	-450,5	-269,19	99,755	aptamer465	-473,31	-220,22	99,845
aptamer734	-496,96	-327,26	99,903	aptamer413	-439,15	-207,55	99,693	aptamer1085	-469,92	-225,65	99,834
aptamer279	-494,57	-315,98	99,898	aptamer331	-435,22	-267,66	99,668	aptamer437	-459,24	-210,47	99,794
aptamer128	-489,36	-303,52	99,887	aptamer278	-434,11	-200,28	99,661	aptamer154	-452,49	-216,6	99,765
aptamer107	-486,63	-326,36	99,881	aptamer526	-430,02	-249,18	99,632	aptamer734	-432,3	-273,27	99,648
aptamer920	-483,18	-300,06	99,873	aptamer58	-428,78	-240,16	99,623	aptamer260	-427,78	-221,48	99,615
aptamer569	-478,85	-321,81	99,861	aptamer59	-428,09	-202,23	99,617	aptamer104	-425,72	-200,73	99,599
aptamer348	-476,99	-315,45	99,856	aptamer577	-424,08	-306,13	99,585	aptamer26	-419,04	-227,21	99,542
aptamer924	-475,35	-307,16	99,851	aptamer924	-420,81	-282,56	99,558	aptamer938	-418,87	-232,56	99,54
aptamer719	-462,6	-277,02	99,808	aptamer569	-419,53	-310,21	99,546	aptamer303	-411,69	-294,59	99,47
aptamer242	-461,51	-242,83	99,803	aptamer618	-415,11	-249,7	99,504	aptamer727	-410,38	-211,98	99,456
aptamer584	-460,03	-333,68	99,798	aptamer348	-414,53	-295,28	99,499	aptamer603	-406,36	-212,91	99,41
aptamer299	-451,42	-285,96	99,76	aptamer260	-414,35	-252,11	99,497	aptamer96	-401,59	-180,62	99,352

7.3.2.2 Interaction bonds

Interaction bonds are vital for the stability and integrity of complex molecular structures. Hydrogen bonds, which form between hydrogen atoms and electronegative atoms in neighbouring molecules, play a crucial role in stabilizing biomolecular complexes like proteins and DNA [397]. Other bonds found mostly in complex includes all types of pi bonds. Pi-pi stacking is also a crucial interaction observed in RNA and other biomolecules, it involves attractive forces between aromatic rings. The overlap of π-electron clouds stabilizes the stacking arrangement of adjacent nucleobases (adenine, guanine, cytosine, and uracil) in RNA, impacting its sequence-dependent structure. Pi-pi stacking occurs not only within RNA strands but also between different RNA molecules, promoting favourable stacking arrangements [398]. Additionally, these interactions contribute to RNA's complex tertiary structures which is essential for functions like riboswitches and ribozymes [399]. In protein-RNA complexes, aromatic residues in proteins engage in pi-pi interactions with RNA bases, influencing binding and function [400]. We utilized a Perl code (appendix C2) to analyse potential hydrogen bonds and pi-pi bonds within the complexes. Specifically, we focused on the top 2 best aptamers for each target, along with five their models. The obtained results were then plotted in **Figure 7.7**.

Hydrogen bonds were initially monitored at a criterion distance of 2.5 Å between two chains, where Chain A represents the target and Chain B serves as the aptamer. Subsequently, this criterion was extended by 0.5 Å, reaching 3 Å, allowing for a broader exploration of potential interactions. This adjustment aimed to enhance the detection of hydrogen bonding events between the two RNA chains, providing insights into their binding affinity and specificity. Following the assessment of hydrogen bonds, pi-pi interactions were also monitored. The first-ranked aptamers consistently exhibit more hydrogen bonds compared to the second-ranked aptamers across all five models across all targets. However, there is an exception in target miR-10b-5p, where aptamer899 and aptamer536 have the same number of hydrogen bonds under the 2.5 Å criterion. Although these numbers of hydrogen bonds are observed for the stationary complex without considering physiochemical parameters such as solvents and ions, MD simulations are still required to assess the stability and dynamics of these hydrogen bonds under more realistic, biologically relevant conditions. Regarding pi-pi stacking, most miRNA-aptamer complexes appear to lack such interactions. This could be attributed to the fact that RNA-RNA nucleotides are generally not susceptible to form pi-pi interactions, especially considering the angles and potential contacts between two folded RNA molecules. Notably, the

highest number of pi-pi interactions (five) was observed in model 2 of aptamer895 against pre-mature miRNA.

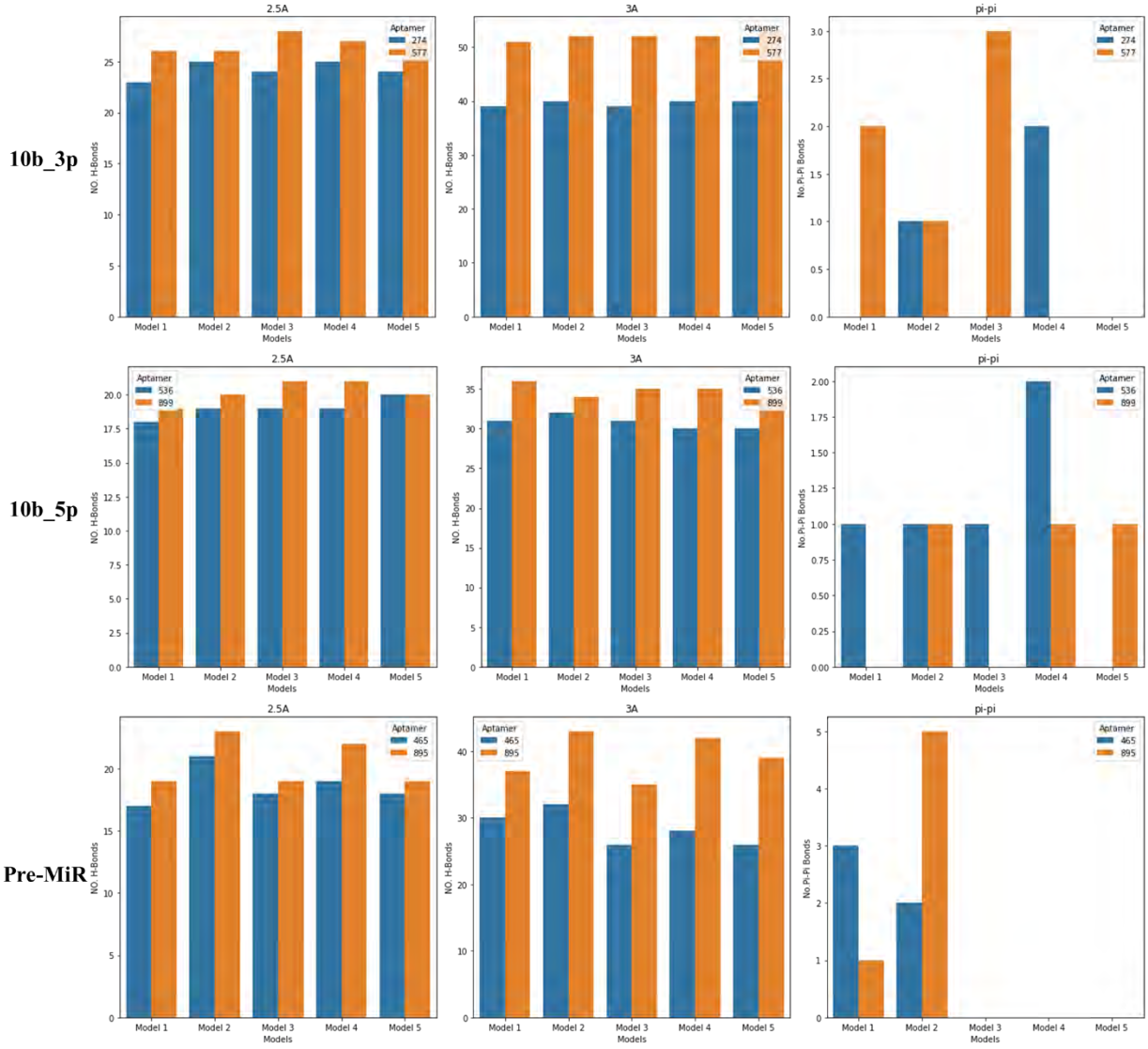


Figure 7.7: Monitored Bonds interactions between the two best-performing aptamers together with their five best models. The orange indicates the first best-performing aptamer, and the blue colour shows the second-best aptamer.

7.3.2.2 Docking poses

The docking poses of the 10 models from two best docking aptamers against miR-10b-3p, and miR-10b-5p are reported in **Figure 7.8** to **Figure 7.11**. **Figure 7.8** show how aptamer899 binds to miR-10b-5p in different ways depicted from 10 models, demonstrating the virous possibility and flexibility of the aptamer when binding with its target. Each of the ten models (M1-M10) represents a unique colour [Brown (model 1), Purple (Model 2), Green (Model 3), Yellow (Model 4), Grey (Model 5), Pink (Model 6), Blue (Model 7), Black (Model 8), Teal (Model 9), and Red (Model 10)]. Since miRNA lacks an active site and there are currently no available aptamer-based drugs, there will be no established standard for comparison. In **Figure 7.8**, unique binding poses of aptamer899 are observed across models 1 to 10. Notably, aptamer899 binds to the target through its stem in models 1, 3, 4, 7, 8, and 9. While these poses are not identical, they demonstrate a consistent interaction pattern. There are also slight variations in the binding positions, although these differences may not be immediately clear. This suggests that aptamer899 has a stronger preference for the groove region of miR-10b-5p. In contrast, models 2, 5, 6, and 10 show aptamer899 binding through its hairpin loop to miR-10b-5p. This indicates that the most optimal binding sites for aptamer899 are either the hairpin loop or the stems, with the binding configurations being somewhat similar across these models.

In **Figure 7.9**, aptamer536 exhibits a significantly different 3D structure compared to aptamer899. Unlike aptamer899, which features well-formed Watson-Crick stems, aptamer536 has a small stem and a large single-stranded region or tail at its terminal end. Aptamer536 binds to miR-10b-5p in the groove position, similar to the binding sites observed for aptamer899 using single-stranded region tail. However, the configuration of the binding style of the tail to the grooves of miR-10b-5p differs from that of aptamer899. Notably, the hairpin loop of aptamer536 is not involved in the binding across any of the models, indicating a distinct binding mechanism compared to aptamer899.

In **Figures 7.10** and **7.11**, both aptamer274 and aptamer577 exhibit similar 3D structures, characterized by having long single-stranded tail at the terminal end and a small stem. However, to distinguish between them, aptamer274 features a loop made of four bases, while aptamer577 has a loop composed of three bases. Despite this minor structural difference, both aptamers tend to bind to miR-10b-3p in a similar manner and at comparable positions across all 10 models. This suggests that miR-10b-3p binding efficiency to the aptamers that have single-stranded tail at the terminal, indicating that this structural feature is conducive to effective binding of miR-10b-3p.

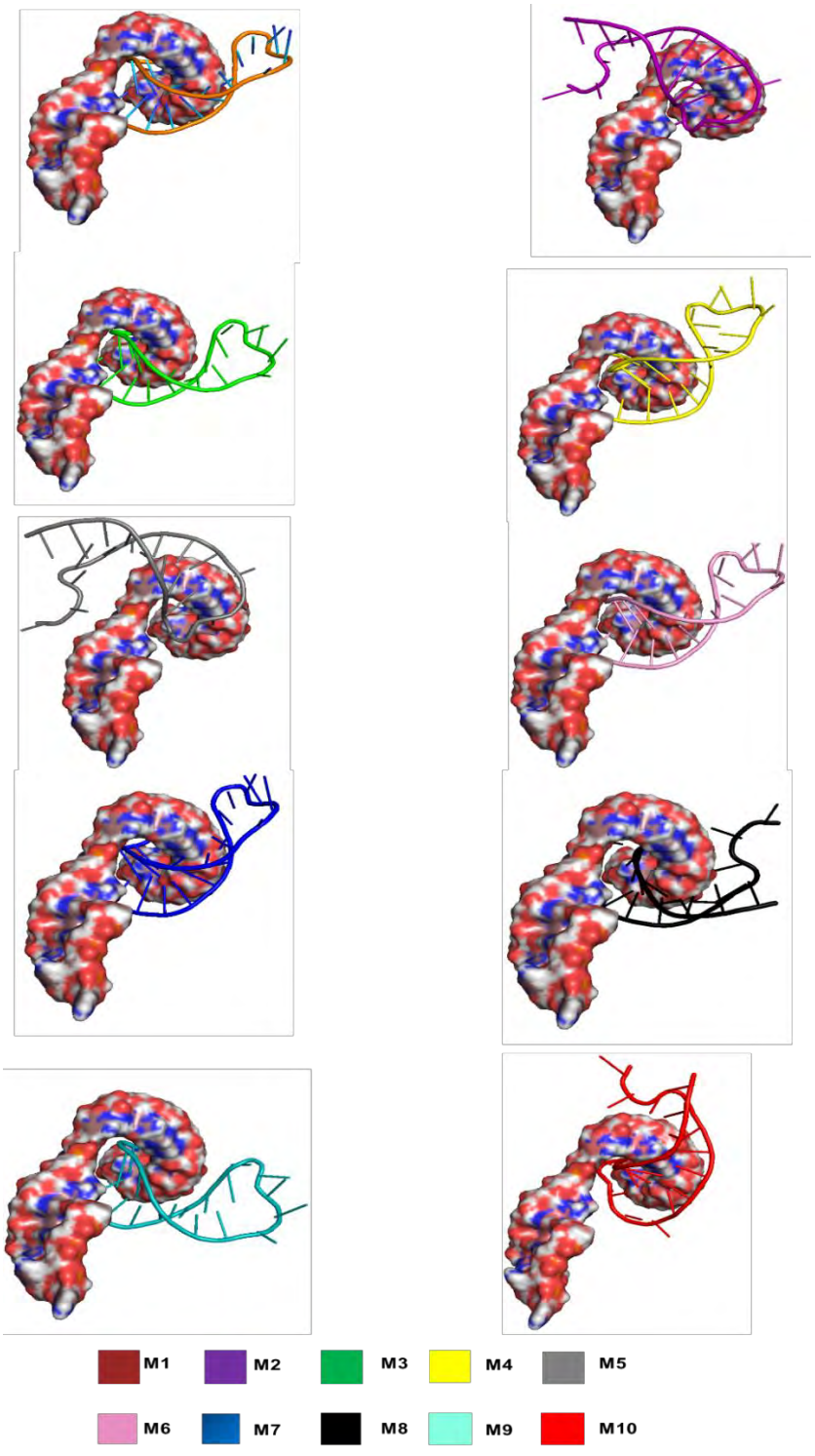


Figure 7.8: Best docked models complex of aptamer899-miR-10b-5p. Where the target miR-10b-5p is meshed with surfaces and M1 – M10 represents model1 to model10, the colour code is given in the figure.

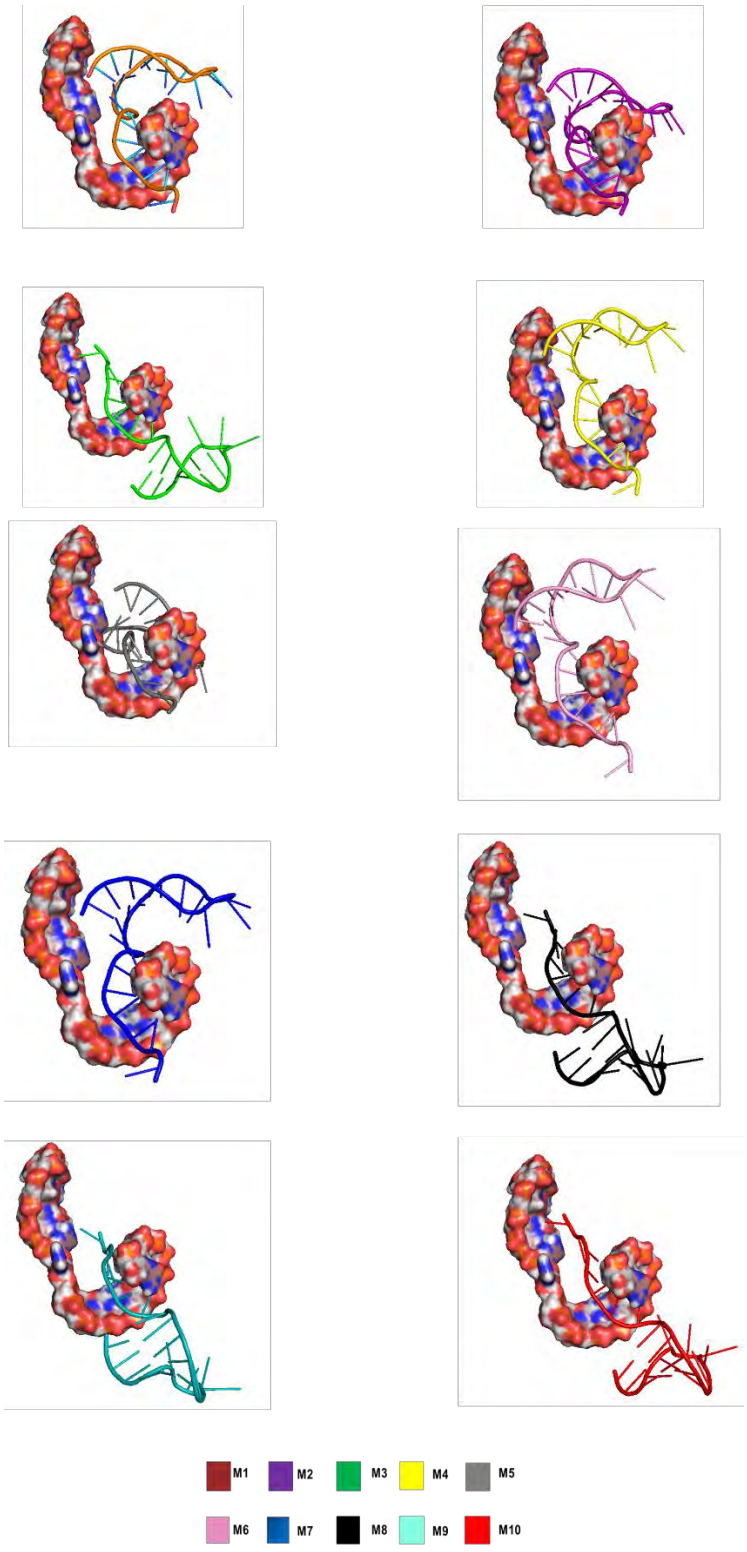


Figure 7.9: Best docked models complexes of aptamer536-mir10b-5p. Where the target mir10b-5p is meshed with surfaces and M1 – M10 represents model1 to model10, the colour code is given in the figure.

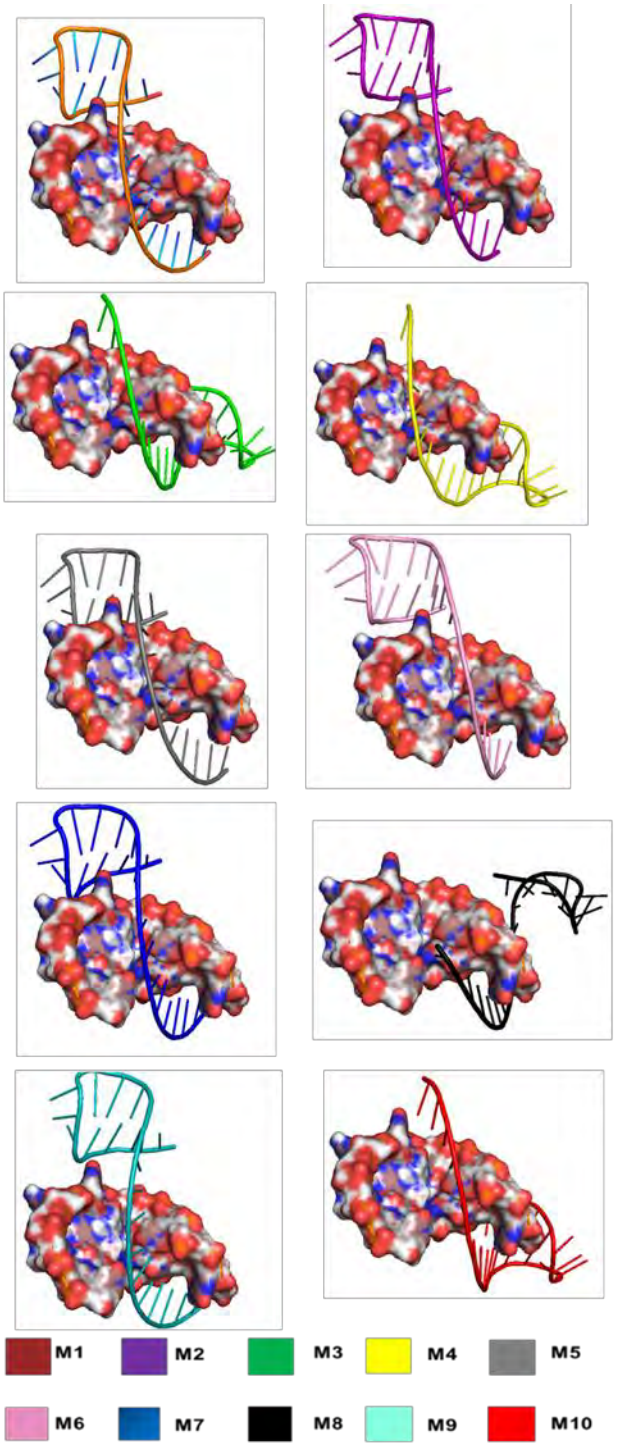


Figure 7.10: Best docked models complex of aptamer577-miR-10b-3p. Where the target miR-10b-3p is meshed with surfaces and M1 – M10 represents model1 to model10, the colour code is given in the figure.

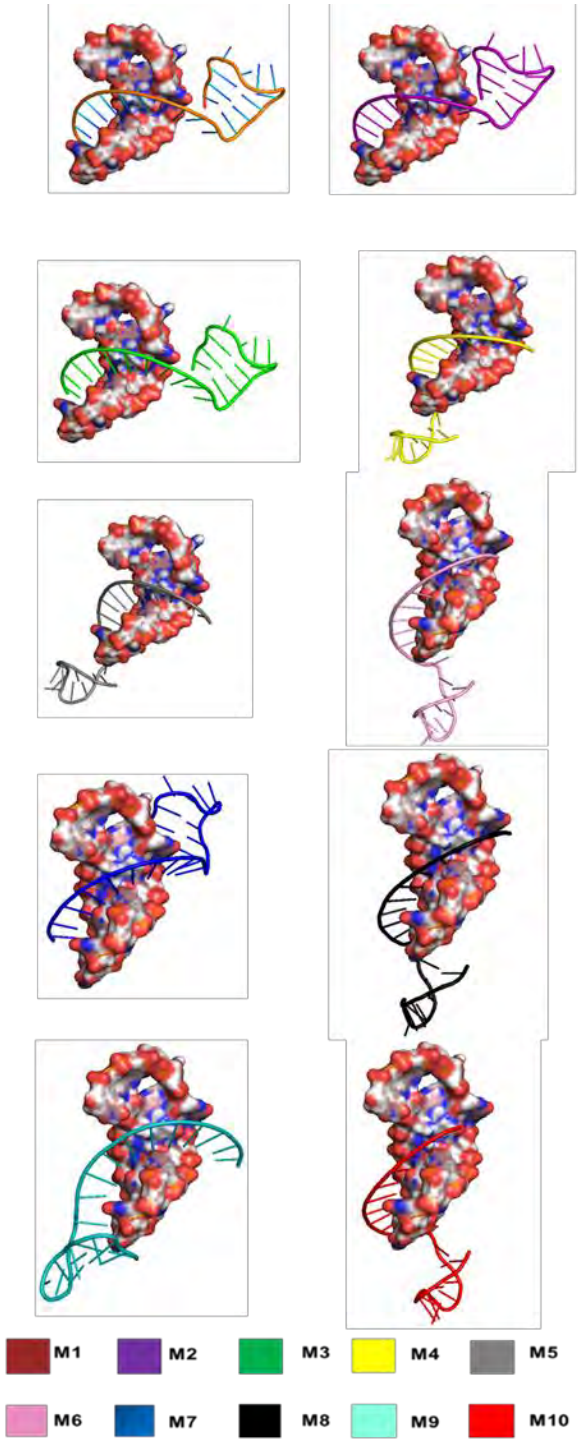


Figure 7.11: Best docked models complex of aptamer274-miR-10b-3p. Where the target miR-10b-3p is meshed with surfaces and M1 – M10 represents model1 to model10, the colour code is given in the figure.

7.3.2.4 Hydrogen bonds

Since the ligands are macromolecules, it is difficult to represent their interactions in a 2D format. Therefore, a code was written in Perl for Discovery Studio to monitor not only the number of hydrogen bonds but also which atoms in Chain_A (miRNA) interact with which atoms in Chain_B (aptamers). The data was further translated into tables as shown in **Tables 7.4**, **7.5**, and **7.6**. The data presented in these tables reflect the results of docking simulations from three different models of aptamer899 against miR-10b-5p, identified as model1, model2, and model3. The tables contain detailed information about the molecular interactions, specifically the distances between bonded atoms in different chains.

The data in **Tables 7.4**, **7.5**, and **7.6** shows that while many bonds are consistently observed across all models, there are slight variations in the bond distances. These differences are indicative of the dynamic nature of molecular interactions, which can vary due to the specific conditions and parameters of each model. For instance, bond 4 (G3 -U22) shows a slight difference in distance between model1 (2.865 Å) and Model 3 (2.864 Å). Such minor variations, although small, can be significant in understanding the flexibility and conformational changes of the docked aptamers throughout the models. Additionally, some bonds are unique to specific models, such as bond 8 (C6-U18) in model3, highlighting the potential for distinct interaction sites within this model1. These slight differences in distances and the occasional unique bond suggest variations in the poses across these different models of the same aptamer899-miR-10b-5p docked complex.

Table 7.4: Intermolecular bonds of model1 between the target miR-10b-5p (Chain A) and aptamer899 (Chain B).

Bond	Distance (Å)	Chain_A	Chain_B
1	2,607	A1:HO2'	C2:OP2
2	2,624	C2:HO2'	G3:O4'
3	2,819	G3:H21	G21:N3
4	2,865	G3:H21	U22:O4
5	2,647	G3:H22	U20:O2
6	2,524	G3:H1	U20:O2
7	2,885	G3:HO2'	G4:OP1
8	1,796	G4:H1	U20:O2
9	1,694	A5:H62	U19:O4
10	2,04	C6:H42	G18:O6
11	2,968	C6:HO2'	U7:O5'
12	2,425	C6:HO2'	U7:O4'
13	2,196	U7:H3	G17:O6
14	2,842	U7:HO2'	C8:O5'
15	2,094	C8:H42	G16:O6
16	2,983	G12:H22	C13:O2
17	2,907	G12:H1	C13:O2
18	2,164	G12:HO2'	C15:O4'
19	1,692	U14:HO2'	C15:O5'
20	2,043	C15:H42	U9:O2
21	2,055	C15:HO2'	C13:OP1
22	2,76	C15:HO2'	C15:O4'
23	1,758	G16:H22	C8:O2
24	2,18	G16:H1	C8:O2
25	1,965	G16:H1	C8:N3
26	1,71	G17:H1	U7:O2
27	1,68	G18:H22	C6:O2
28	2,345	G18:H1	C6:O2
29	1,926	G18:H1	C6:N3
30	1,75	U19:H3	A5:N1
31	2,285	U20:H3	G4:O6
32	2,554	U20:H3	G21:O6
33	2,429	G21:H22	G4:N7
34	2,64	G21:H1	G3:O6
35	2,783	G21:H1	G4:N7
36	2,629	G21:HO2'	U22:OP2

Table 7.5: Intermolecular bonds of model2 between the target miR-10b-5p (Chain A) and aptamer899 (Chain B).

Bond	Distance (Å)	Chain_A	Chain_B
1	2,608	A1:HO2'	C2:OP2
2	2,625	C2:HO2'	G3:O4'
3	2,865	G3:H21	U22:O4
4	2,885	G3:HO2'	G4:OP1
5	1,796	G4:H1	U20:O2
6	1,693	A5:H62	U19:O4
7	2,04	C6:H42	G18:O6
8	2,969	C6:HO2'	U7:O5'
9	2,425	C6:HO2'	U7:O4'
10	2,196	U7:H3	G17:O6
11	2,841	U7:HO2'	C8:O5'
12	2,094	C8:H42	G16:O6
13	2,745	G12:H22	U19:O2
14	2,983	G12:H22	C13:O2
15	2,908	G12:H1	C13:O2
16	2,164	G12:HO2'	C15:O4'
17	1,692	U14:HO2'	C15:O5'
18	2,043	C15:H42	U9:O2
19	2,054	C15:HO2'	C13:OP1
20	2,759	C15:HO2'	C15:O4'
21	1,757	G16:H22	C8:O2
22	2,18	G16:H1	C8:O2
23	1,963	G16:H1	C8:N3
24	1,709	G17:H1	U7:O2
25	1,68	G18:H22	C6:O2
26	2,345	G18:H1	C6:O2
27	1,927	G18:H1	C6:N3
28	1,75	U19:H3	A5:N1
29	2,286	U20:H3	G4:O6
30	2,555	U20:H3	G21:O6
31	2,429	G21:H22	G4:N7
32	2,64	G21:H1	G3:O6
33	2,783	G21:H1	G4:N7
34	2,629	G21:HO2'	U22:OP2

Table 7.6: Intermolecular bonds of model3 between the target miR-10b-5p (Chain A) and aptamer899 (Chain B).

Bond	Distance (Å)	Chain_A	Chain_B
1	2,607	A1:HO2'	C2:OP2
2	2,624	C2:HO2'	G3:O4'
3	2,864	G3:H21	U22:O4
4	2,188	G3:H1	U20:O2
5	2,885	G3:HO2'	G4:OP1
6	1,796	G4:H1	U20:O2
7	1,694	A5:H62	U19:O4
8	2,845	C6:H41	U18:O2
9	2,04	C6:H42	G18:O6
10	2,969	C6:HO2'	U7:O5'
11	2,425	C6:HO2'	U7:O4'
12	2,197	U7:H3	G17:O6
13	2,841	U7:HO2'	C8:O5'
14	2,094	C8:H42	G16:O6
15	2,983	G12:H22	C13:O2
16	2,907	G12:H1	C13:O2
17	2,164	G12:HO2'	C15:O4'
18	1,692	U14:HO2'	C15:O5'
19	2,043	C15:H42	U9:O2
20	2,053	C15:HO2'	C13:OP1
21	2,76	C15:HO2'	C15:O4'
22	1,757	G16:H22	C8:O2
23	2,18	G16:H1	C8:O2
24	1,964	G16:H1	C8:N3
25	1,71	G17:H1	U7:O2
26	1,68	G18:H22	C6:O2
27	2,346	G18:H1	C6:O2
28	1,927	G18:H1	C6:N3
29	1,75	U19:H3	A5:N1
30	2,285	U20:H3	G4:O6
31	2,554	U20:H3	G21:O6
32	2,43	G21:H22	G4:N7
33	2,639	G21:H1	G3:O6
34	2,783	G21:H1	G4:N7
35	2,629	G21:HO2'	U22:OP2

7.3.3 Post Docking Analysis

Post-docking analysis was done to ensure thorough examination of all models for each aptamer, thus preventing oversight. The fitness quality of each aptamer was determined by computing the average docking scores across the 100 models generated during docking calculations. In our study, this metric serves as a proxy for overall performance models associated with the aptamer being docked. Additionally, to assess how the first model compared to the subsequent 100 models, we calculated the Z-score. This statistical measure indicates the deviation of a first model docking score from the mean or the rest of the models docking scores. A high Z-score away from zero indicates significant deviation, suggesting pronounced differences between the first model and the rest. Conversely, a Z-score close to zero signifies minimal deviation, indicating similar performance across all models.

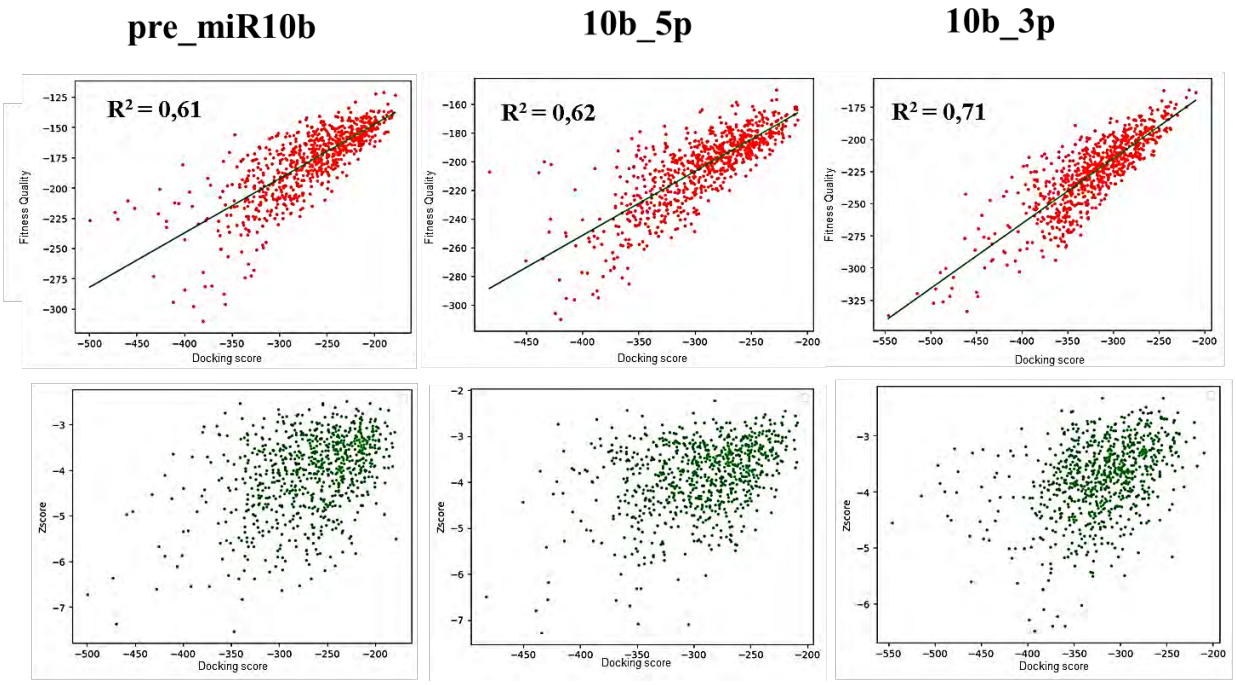


Figure 7.12: The relationship between the fitness quality, Z-scores and docking scores of first models for precursors miR10b and their mature miRNAs (miR-10b-3p, and miR-10b-5p).

The scatter plots in **Figure 7.12** were generated to assess the relationship between the docking score of first models and fitness quality for precursors miR10b and their mature RNAs, marked with red dots. The plots revealed no significant relationships, as indicated by R-squared values below 0.75. The R-squared values were 0.61 and 0.62 for the 3p arm for pre_miR-10b and 5p,

respectively, which are quite similar. However, for the 3p arm, the R-squared value was higher at 0.71, suggesting a relatively stronger linear relationship compared to pre-miR-10b and 5p.

Upon examining pre-miR-10b and the 5p arm, we observed that some of the first models with low docking scores have a predominantly higher fitness quality. This implies that the remaining models might not be performing as well, considering the central tendency of the data is far from the first model docking score. Conversely, for the 3p arm, the trend was different, with the first models having lower docking scores also having lower fitness scores. This implies that when the first model performs well in the 3p arm, the subsequent 100 models tend to have comparable performance. This is illustrated by the central tendency of the data, which closely aligns with the docking scores of the first model. Similarly, when examining the Z-score plots, a similar trend emerges. For the pre-miR-10b and 5p scatter plots, the Z-scores associated with lower docking scores are notably distant from zero, indicating a significant deviation. This suggests that there is huge variance in terms of docking score of the first model and the docking scores rest models. Conversely, for the 3p scatter plot, the Z-scores are closer to zero, indicating less deviation. This suggests a higher degree of consistency in performance of aptamers across all the models against miR-10b-3p.

7.3.4 QM calculations

Quantum mechanical calculations were conducted to assess the stability and reactivity of the 5 models for each of the top 4 performing aptamers docked to 5p and 3p arms. It is important to note that the name model refers to the docked complex where both aptamer and miRNA are present. Only semi-empirical single-point calculations were performed to evaluate the stability of these docking complexes obtained from molecular docking. Total energy (E_{tot}) and HOMO-LUMO energy gaps (H-L gap) were obtained, with results detailed in **Table 7.7** to **7.8**, **Figure 7.13**. Total energy serves as a crucial indicator of complex stability, whereas H-L gap elucidates the electronic properties.

The analysis of total energies for the aptamer-miR-10b-3p complexes reveals varying trends in stability across different models. For aptamer128miR-10b-3p, the second model exhibits lower total energy, suggesting increased stability, with the energy decreasing progressively towards model5. However, this trend is not consistent across all complexes. For aptamer734-, aptamer274-, and aptamer557-miR-10b-3p complexes, the first models have lower total energy values than the second models, indicating no improvement in stability with the second model. Interestingly, aptamer128-miR-10b-3p and aptamer279-miR-10b-3p complexes display

distinct trends in total energy compared to the other aptamers, highlighting variability in stability trends among the complexes. When examining the total energies values of the aptamers-miR-10b-5p complexes, it becomes challenging or even impossible to identify any possible trend. Unlike what was observed for the aptamers-miR-10b-3p complexes, some first models are not stable compared to the rest of the models. This discrepancy could be attributed to significant binding position of an aptamer to miR-10b-5p target which slightly changes across all models.

Table 7.7 : Total energy of the aptamers-miR-10b-3p complexes and their models

models	aptamer577-miR-10b-3p (Eh)	aptamer274 miR-10b-3p (Eh)	aptamer734-miR-10b-3p (Eh)	aptamer279-miR-10b-3p (Eh)	aptamer128-miR-10b-3p (Eh)
1	-3001.38	-2898.73	-2896.73	-2894.59	-2809.03
2	-3007.37	-2879.61	-2967.8	-2943.57	-2912.02
3	-2747.68	-2536.69	-2807.73	-2379.19	-2941.86
4	-1681.32	-1595.91	-497.5	-2776.39	-2939.53
5	-2851.74	-2806.09	-2681.6	-2752.34	-2930.38

Table 7.8 : Total energy of the aptamers-miR-10b-5p complexes and their models

models	Aptamer899-miR-10b-5p (Eh)	aptamer536-miR-10b-5p (Eh)	aptamer413-miR-10b-5p (Eh)	aptamer331-miR-10b-5p (Eh)	aptamer278-miR-10b-5p (Eh)
1	-2801.74	-2587.32	-3014.18	-2723.09	-1416.57
2	-2403.22	-2665.17	-2821.57	-2938.7	-2851.6
3	-2800.52	-2654.75	-2823.34	-2881.13	-2977.7
4	-2773.27	-2874.83	-2539.03	-2818.55	-2998.05
5	-1913.83	-1900.63	-2363.36	-2895.67	-2105.12

To evaluate the electronic properties of the complexes, analysed at the HOMO-LUMO energy gap for the aptamers-miR-10b-3p/5p as reported in **Table 7.9** and **7.10**. It is worth mentioning that the HOMO- LUMO energies gaps are significantly influenced by the size of the aptamers-miRNAs models we used in our simulations. As a result, these values should be considered to highlight general trends. The first model often exhibited a lower or smaller energy gap, while others have higher energy gap up until model4 or model5. Subsequently, the HOMO-LUMO gaps tended to increase until the third model for most aptamers, after which they decreased

again. However, this trend was not consistent across all aptamers. This suggests that electronic properties of the aptamers-miRNA complexes change across all 5 models. Similarly, for aptamers-miR-10b-5p the HOMO-LUMO (H-L) gap values do not exhibit a clear trend or pattern across 5 models. For instance, it cannot be argued that model1 of aptamer278-miR-10b-5p complex, followed by model3 of aptamer899-miR-10b-5p complex, are the least stable since they have small H-L gaps. Conversely, the most stable complex appears to be model1 of aptamer899-miR-10b-5p, followed by model1 of aptamer331-miR-10b-5p.

Table 7.9: HOMO-LUMO energy gap of the aptamers-miR-10b3p complexes and their models

models	aptamer577-miR-10b-3p (eV)	aptamer274-miR-10b-3p (eV)	aptamer734-miR-10b-3p (eV)	aptamer279-miR-10b-3p (eV)	aptamer128-miR-10b-3p (eV)
1	0.001129	0.000305	0.000901	0.000047	0.000654
2	0.00228	0.00609	0.00158	0.00267	0.00105
3	0.0047	0.0148	0.00675	0.0025	0.00684
4	0.00323	0.00037	0.00591	0.00061	0.00821
5	0.00956	0.00473	0.000542	0.0023	0.00019

Table 7.10: HOMO-LUMO energy gap of the aptamers-miR-10b-5p complexes and their models

models	aptamer899-miR-10b-5p (eV)	aptamer536-miR-10b-5p (eV)	aptamer413-miR-10b-5p (eV)	aptamer331-miR-10b-5p (eV)	aptamer278-miR-10b-5p (eV)
1	0.00295	0.0314	0.00105	0.0168	0.000084
2	0.00622	0.0112	0.00292	0.00134	0.0089
3	0.000094	0.0159	0.00721	0.00361	0.00129
4	0.00173	0.00574	0.00436	0.000801	0.00133
5	0.00257	0.00352	0.0131	0.00991	0.00851

Based on these observations, it is evident that the HDock [300] scoring function does not adequately consider the stability of the complex or take the dispersion into account. This is further illustrated by examining the aptamer899-miR-10b-5p model1 complex in **Figure 7.13**. As the docking position of the aptamers change, the total energy and H-L gap changes as well from just looking at only 5 models, and it could be thought that the same will happen when looking at rest of the 100 models. This indicates that the interactions between the aptamers and miRNAs with the models can mainly be viewed as dispersion-driven physisorption, which is characterized by relatively weak interactions.

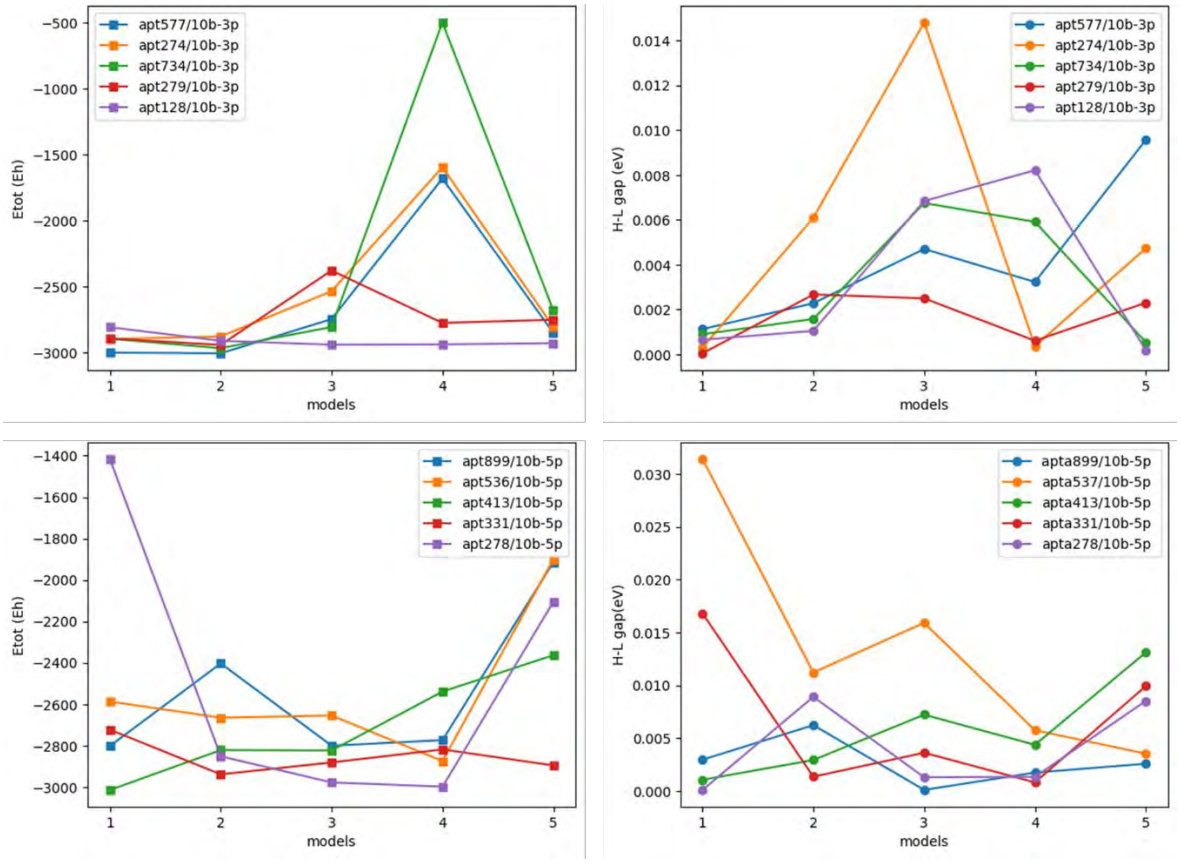


Figure 7.13: QM result for the for the aptamer docked complex against miR-10b-3p, and miR-10b-5p targets.

7.3.5 Molecular Dynamics

7.3.5.1 RMSD for Complexes

In molecular dynamics (MD) simulations, analysing the Root Mean Square Deviation (RMSD) is crucial for assessing how molecular structures evolve over time relative to a reference frame or structure. RMSD measures the average displacement of specific atoms or groups of atoms throughout the simulation trajectory [401]. It serves as a fundamental metric to evaluate the convergence and stability of molecular structures, indicating whether the system has equilibrated and how much it deviates from the initial reference structure. By analysing RMSD plots, we can determine patterns of stabilization in molecular backbones and assess the overall dynamics and structural integrity of the simulated aptamer-miRNA system. Over 40 molecular dynamics simulations were conducted, focusing on the mature miR-10b-5p and miR-10b-3p as the targets. For each target, we selected the top 4 binding aptamers along with their respective 5 models to assess stabilization across different docked poses. In this section, we analyse the RMSDs of the best complexes to understand how these aptamer-RNA complexes evolve structurally over time. The RMSD plots for the complexes involving miR-10b-5p are illustrated in **Figure 7.14**, while those for the miR-10b-3p complexes are illustrated in **Figure 7.15**.

Looking on the RMSD plots illustrated in **Figure 7.14** for 5 models of aptamer536, 413, 331, and 899-miR-10b-5p complexes, it is evident that all complexes stabilize below the 3 nm over the course of the simulations. For instance, in model1 of aptamer899-miR-10b-5p, the RMSD fluctuates up to 80 ns but still remains below 3 nm, indicating relative stability despite structural confirmational changes. These fluctuations are attributed to the dynamic nature of RNA structures, which can adopt various conformations influenced by available substates [402]. This analysis highlights the inherent flexibility of RNA molecules compared to proteins and suggests that the observed fluctuations primarily reflect the potential conformations adopted by the RNA complexes. Thus, while the RMSD values fluctuate, they consistently remain within a stable range below 3 nm, indicative of the stability of the RNA complexes throughout the simulation period. Moreover, upon examining the overall RMSDs for the miR-10b-5p-aptamer complexes, it is apparent that models of aptamer899-miR-10b-5p stabilizes more consistently compared to most aptamer complexes. Notably, model1 of aptamer899-miR-10b-5p shows high fluctuation beyond 20 ns, unlike the other models that stabilize within this timeframe. Despite these fluctuations, it is essential to note that the HDock algorithm's [300] scoring function, while potentially requiring refinement as indicated by the deviations in models of aptamer899-miR-10b-5p, has demonstrated remarkable capability compared to other

algorithms [302]. The algorithm's ability to generate RNA-RNA docked model complexes that thermally equilibrate suggests its robustness in predicting stable RNA-RNA 3D interactions. Therefore, while improvements may be warranted, the HDock algorithm remains a valuable tool in computational biology for studying RNA-RNA complexes. The models of aptamer331-miR-10b-5p stabilizes under 2 nm and is ranked fourth based on molecular docking studies. This suggests that the models of aptamer331-miR-10b-5p exhibit a stable structure during the simulations, indicating that the system remains closer to a reference structure compared to the other aptamers. Furthermore, model of aptamer331-miR-10b-5p shows less RMSD fluctuation, suggesting greater stability and reduced structural variability over time.

Since measuring which complex stabilizes better based only on the fluctuation of the RMSD is difficult, the degree of stabilization (τ) was also calculated for each plot and reported in the plot legend. The plot legends in **Figure 7.14**, shows that the highest degree of stabilization is observed in model4 of aptamer536-miR-10b-5p and model1 of aptamer413-miR-10b-5p with τ =19 and τ =15, respectively. This is followed by the second model of aptamer413-miR-10b-5p, the fifth model of aptamer413-miR-10b-5p, and the first model of aptamer899-miR-10b-5p complex, all with having τ =11. This suggests that these systems undergo significant structural transitions throughout the simulations, indicating less stability. The lowest degree of stabilization is observed in model1 of aptamer331-miR-10b-5p and model2 of aptamer899-miR-10b-5p with τ =2, suggesting fewer structural changes and more stability. Overall, based on the degree of stabilization, the aptamer331- miR-10b-5p models complexes exhibit low degrees of stabilization, with values below 6. This confirms that the RMSD of the aptamer331-miR-10b-5p models are the most stable. In contrast, most models of aptamer536- miR-10b-5p and aptamer413-miR-10b-5p show higher degrees of stabilization, suggesting that these models undergo more structural transitions and are therefore less stable.

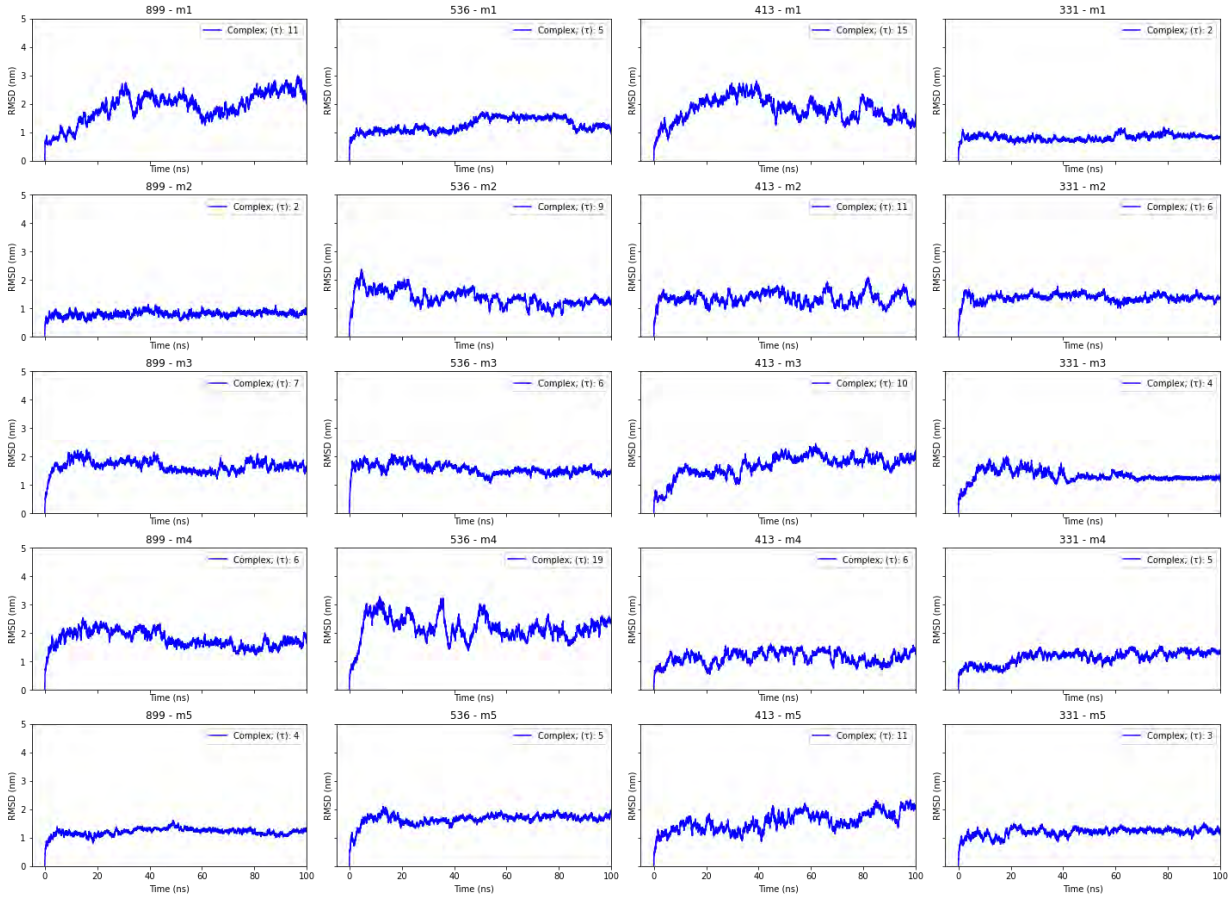


Figure 7.14: The RMSD plots for 5 model complexes of aptamer536, 413, 331, and 899 docked against miR-10b-5p. For each complex the model is denoted as m.

In **Figure 7.15** where the target is miR-10b-3p against the five models of best aptamers (aptamer557, 274, 734, and 279), it is evident that all complexes seem to stabilize below 2 nm except for model5 of aptamer274-miR-10b-3p and model1 of aptamer734-miR-10b-3p, with degrees of stabilization of 18 and 26, respectively. This suggests that these complexes are generally stable; however, model5 of aptamer274-miR-10b-3p and model1 of aptamer734-miR-10b-3p are notable exception. The RMSD of model5 of aptamer274-miR-10b-3p shows chaotic behaviour and does not stabilize, reaching up to 6 nm. This indicates that the complex is undergoing major structural transitions and possibly having either miR-10b-3p or aptamer274 unfolding, or even worse both of them unfolding. Of course, it would be an overstatement to conclude that the complex is unfolding based solely on RMSD, more analyses are needed to confirm this hypothesis. Overall, the aptamer557 models have degrees of stabilization below 4, suggesting that the 557 models, which were ranked top 1 based on molecular docking, have the most stable model complexes compared to the rest.

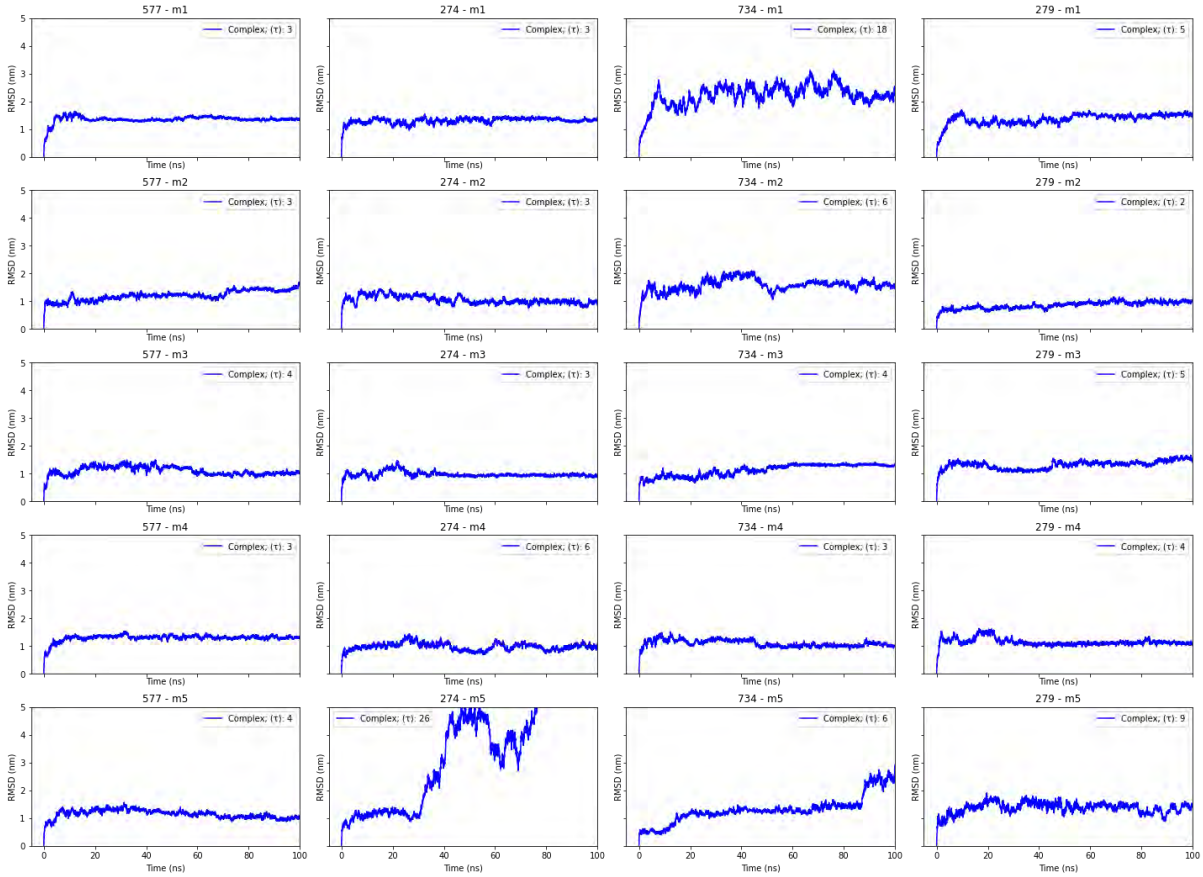


Figure 7.15: The RMSD plots for 5 models of aptamer279, 274, 734, and 577 docked against miR-10b-3p. For each complex the model is denoted as m.

7.3.5.2 Stability metric analysis

To further validate the stability of the complex based on RMSD data while accounting for potential bias and fluctuations towards the end of the simulation, we employed our proposed stability metric method. This approach helps mitigate the impact of RMSD variations and provides a clearer picture of stability. According to the stability metrics presented in the **Table 7.11**, aptamer331 models emerge as the most stable overall when interacting with the miR-10b-5p arm. The stability metric (μ) for aptamer331 models consistently show lower values, ranging from 1 to 3 which indicating a higher degree of stability compared to the other aptamers. In contrast, aptamer899 models exhibit a wider range of μ values from 1 to 5.5, suggesting greater difference in stability. Aptamer536 models also demonstrate generally lower μ values, with values as low ranging from 2.5 and 4.5, implying notable stability but not as stable as Aptamer 331. Lastly, aptamer413 models show μ values between 3 and 7.5, reflecting moderate stability. Therefore, Aptamer 331 models stand out as the most stable aptamer-miR10b-5p based on the stability metric, demonstrating the most consistent and reliable performance.

In the analysis of RMSD of aptamer models interacting with miR10b-3p, which are presented in **Table 7.12**, aptamer279 models consistently show the highest stability. Specifically, model2 of aptamer279-miR-10b-3p complex stands out with the lowest stability metric (μ) of 1, indicating optimal stability among other models. In contrast, aptamer577 models, such as model 3 to model 5, have μ values ranging from 1.5 to 2, suggesting lower stability compared to other Aptamer279-miR-10b-3p models. Aptamer734-miR-10b-3p models also show higher μ values, with models like model 1 to model 5 having μ values from 3 to 9, which indicate greater variability and reduced stability. Therefore, Aptamer279-miR-10b-3p, and particularly model2, exhibit the most consistent and favourable stability in these complexes. The error of the area for the RMSD shows to be zero, this suggest that here little to no error whencomputing the RMSD using the composite Sampsion's rule assuming that the error is bounded.

Tabe 7.11: The stability metric results for 5 models of aptamer536, 413, 331, and 899 docked against miR-10b-5p

Aptamer Name	Model	RMSD Area	Degree of Stabilization (τ)	Stability Metric (μ)	Error for Area
899	m1	190,09	11	5,5	0
899	m2	80,77	2	1	0
899	m3	166,6	7	3,5	0
899	m4	176,95	6	3	0
899	m5	120,6	4	2	0
536	m1	124,65	5	2,5	0
536	m2	159,16	10	4.5	0
536	m3	155,59	6	3	0
536	m4	217,16	19	9,5	0
536	m5	165,15	5	2,5	0
413	m1	179,03	15	7,5	0
413	m2	135,9	11	5,5	0
413	m3	166,89	10	5	0
413	m4	113,37	6	3	0
413	m5	156,38	11	5,5	0
331	m1	81,55	2	1	0
331	m2	138,4	6	3	0
331	m3	132,13	4	2	0
331	m4	113,76	5	2,5	0
331	m5	119,34	3	1,5	0

Table 7.12: The stability metric results for 5 models of aptamer279, 274, 734, and 577 docked against miR-10b-3p

Name	Model	RMSD Area	Degree of Stabilization (τ)	Stability Metric (μ)	Error for Area
577	m1	135,13	3	1,5	0
577	m2	120,99	3	1,5	0
577	m3	109,82	4	2	0
577	m4	128,82	3	1,5	0
577	m5	114,71	4	2	0
274	m1	130,54	3	1,5	0
274	m2	105,88	3	1,5	0
274	m3	96,81	3	1,5	0
274	m4	96,94	6	3	0
274	m5	385,89	26	13	0
734	m1	221,29	18	9	0
734	m2	158,11	6	3	0
734	m3	112,3	4	2	0
734	m4	108,94	3	1,5	0
734	m5	130,74	6	3	0
279	m1	136,25	5	2,5	0
279	m2	86,21	2	1	0
279	m3	131,37	5	2,5	0
279	m4	113,92	4	2	0
279	m5	136,92	9	4,5	0

7.3.5.3 RMSD for individual aptamers and miRNA targets

The RMSD of a complex alone does not provide a complete picture of individual RNA chain contribution to the overall RMSD of the complex. Here, we assess the stability of each chain based on their individual RMSD plots in relation or referencing to the whole system. In **Figure 7.16** and **7.17**, Chain A (represented in blue, corresponding to target miR-10b-5p) and Chain B (depicted in yellow, correspond ligands which in this case are aptamers) are examined separately. In **Figure 7.17**, the RMSD of Chain A (target miR-10b-5p) start stabilizing around 1.5 nm after 10 ns throughout the simulations. However, significant fluctuations are observed among the models of aptamer899, 536, 413, and 331-miR-10b-5p bound to Chain A,

suggesting that their binding slightly influences the structural conformation of the target. Interestingly, variations in stabilization are noted within Chain A when bound to the same aptamer at different positions, evident from model1 (m1) to model5 (m5) during the simulation. Chain B stabilizes below 1 nm in all models of aptamer889-miR-10b-5p, indicating tight binding to the target with minimal structural conformational transitions. Moving to aptamer536-miR-10b-5p complex, a similar pattern is observed in model1 (m1), while model2 to 5 (m2 to m5) show RMSD values for Chain B that superimpose to those of Chain A, suggesting dense binding of all models of aptamer536 to miR-10b-5p. Notably, in models of aptamer413-miR-10b-5p, significant changes in Chain B RMSD are observed, where it stabilizes above Chain A in some instances. It is essential to remember that our targets are miRNAs without active sites, and our ligands are RNA molecules capable of substantial fluctuation. Therefore, fluctuations in ligand RMSD above the targets RMSD do not necessarily indicate detachment from the target pocket. Instead, these fluctuations reflect inherent conformational changes in RNA structures, sometimes associated with a reduction in hydrogen bonds. In models of aptamer413-miR-10b-5p, such as model2 and model5, Chain B's RMSD briefly exceeds that of Chain A before returning beneath it. Finally, across all models of aptamer331-miR-10b-5p, the abrupt increase in RMSD suggests less tight binding to the target but without complete detachment. In **Figure 7.17**, all RMSDs of Chain B abruptly increasing after 10 ns suggesting that the aptamers within these complexes have high fluctuations. This is because the best ranked aptamer bonded to miR-10b-3p have longer single terminal chain that can easily fluctuate without any restrictions. The RMSD plots in **Figure 7.16** and **7.17**, indicate that the main contributors to the overall RMSD of the complex discussed in **section 3.1.1** is chain B (aptamers).

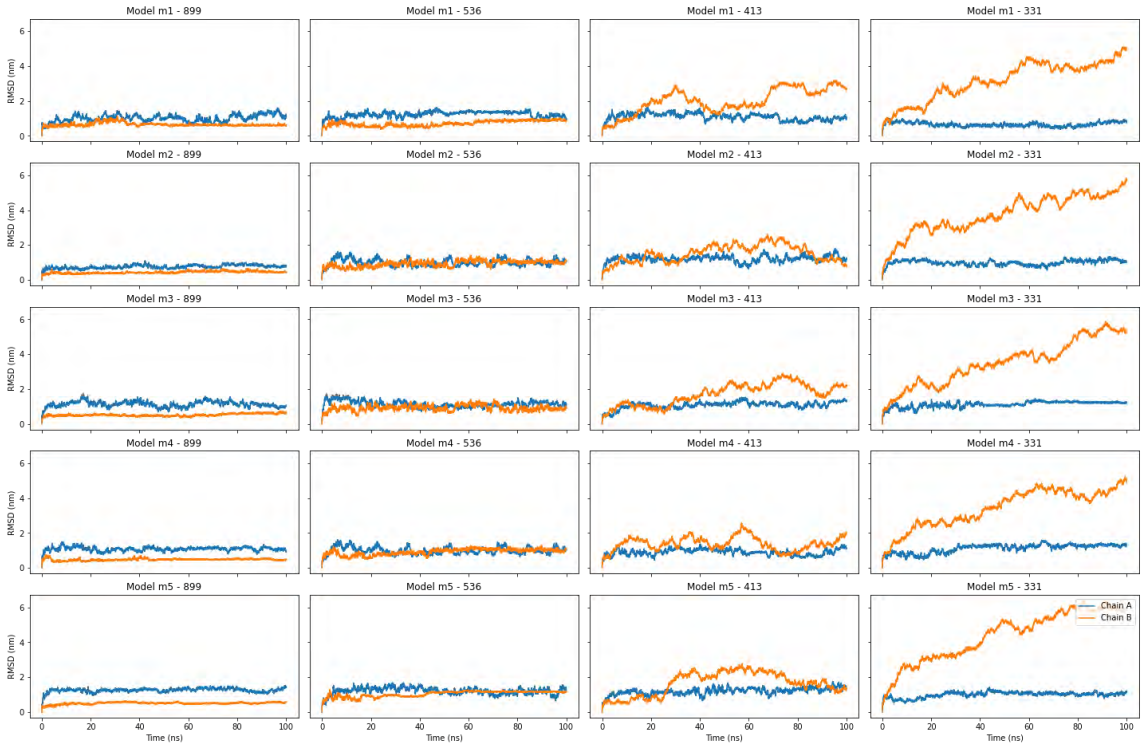


Figure 7.16: Chains RMSD result for the models of aptamer docked complex against miR-10b-5p targets. Where Chain A denoted in Blue is target and Chain B is the aptamer. For all models associated with the complexes are label m1 to m5.

Figure 7.17: Chains RMSD result for the models of aptamer docked complex against miR-10b-3p. Where Chain A denoted in Blue is target and Chain B is the aptamer. For all models associated with the complexes are label m1 to m5.

7.3.5.5 RMSF of an aptamers and miRNAs

The Root Mean Square Fluctuation (RMSF) is a critical measure in molecular dynamics that provides detailed insights into the flexibility and dynamic behaviour of individual residues within RNA molecules. Unlike RMSD, which measures the overall structural deviation from a reference structure at specific time points, RMSF focuses on the average deviation of each residue over the entire simulation period. This makes RMSF particularly valuable for identifying the most mobile or flexible regions of RNA, which is essential for understanding RNA function and interactions. In this discussion, we are reporting the RMSF values for atom indices of 40 molecular dynamics simulations, where the first 20 correspond to the target miR-10b-5p and the remaining 20 correspond to miR-10b-3p. By analysing these RMSF values, we can pinpoint which areas of the RNA exhibit significant fluctuations. The RMSF plots for miR-10b-5p and miR-10b-3p are reported in **Figure 7.18** and **Figure 7.19**, respectively, with the red line representing Chain A (miR-10b-5p/3p) and Chain B (aptamers). In **Figure 7.18**, it is evident that Chain A, which is the target, tends to have atoms that fluctuate significantly in most cases except for model1 and model2 of aptamer331-miR-10b-3p. This suggests that the target miR-10b-5p tends to have different fluctuation patterns when interacting with aptamers at different positions. For aptamer899-miR-10b-5p, in all models, the predominant fluctuating atoms are from 50 to 150 and 600 to 732 atoms, with model 1 of aptamer899-miR-10b-5p having those atoms fluctuating above 1.25 nm. This suggests a less interaction in these regions, leading to higher flexibility. For aptamer536-miR-10b-5p, some models do not show significant fluctuations in atoms 600-732, such as model1. This suggests that the interaction in these models may restrict the movement of these atoms. In some models, we observe that the middle atoms start to fluctuate from atom index 250-450, which includes model3 of this aptamer536-miR-10b-5p complex. This suggests that specific interactions of the aptamer with miR-10b-3p can activate or allow these motif atoms to fluctuate, while others may inhibit this movement. This behaviour is also observed in other aptamer models (413 and 331 aptamer docked complexes). Regarding Chain B, aptamer899-miR-10b-5p seems to have fluctuating atoms from 0-125, 300-470, and 670-702. For aptamer536-miR-10b-5p, fluctuations are observed from 300-620 and/or 0-100 atoms, but this observation is not consistent throughout the models. For aptamer413-miR-10b-5p, the atoms fluctuating throughout the models seem to be 250-500 atoms, except for model4(413-m4). This suggests that there might be a specific interaction or structural feature in model4(413m-4) that stabilizes these regions. Lastly, for aptamer331-miR-10b-5p, prominent fluctuating atoms throughout the models are from atom 190-580, except for model3|(331-m3). This suggests that in model3 of atamer331, the

interaction or structural conformation of the aptamer is different, leading to reduced flexibility in this region.

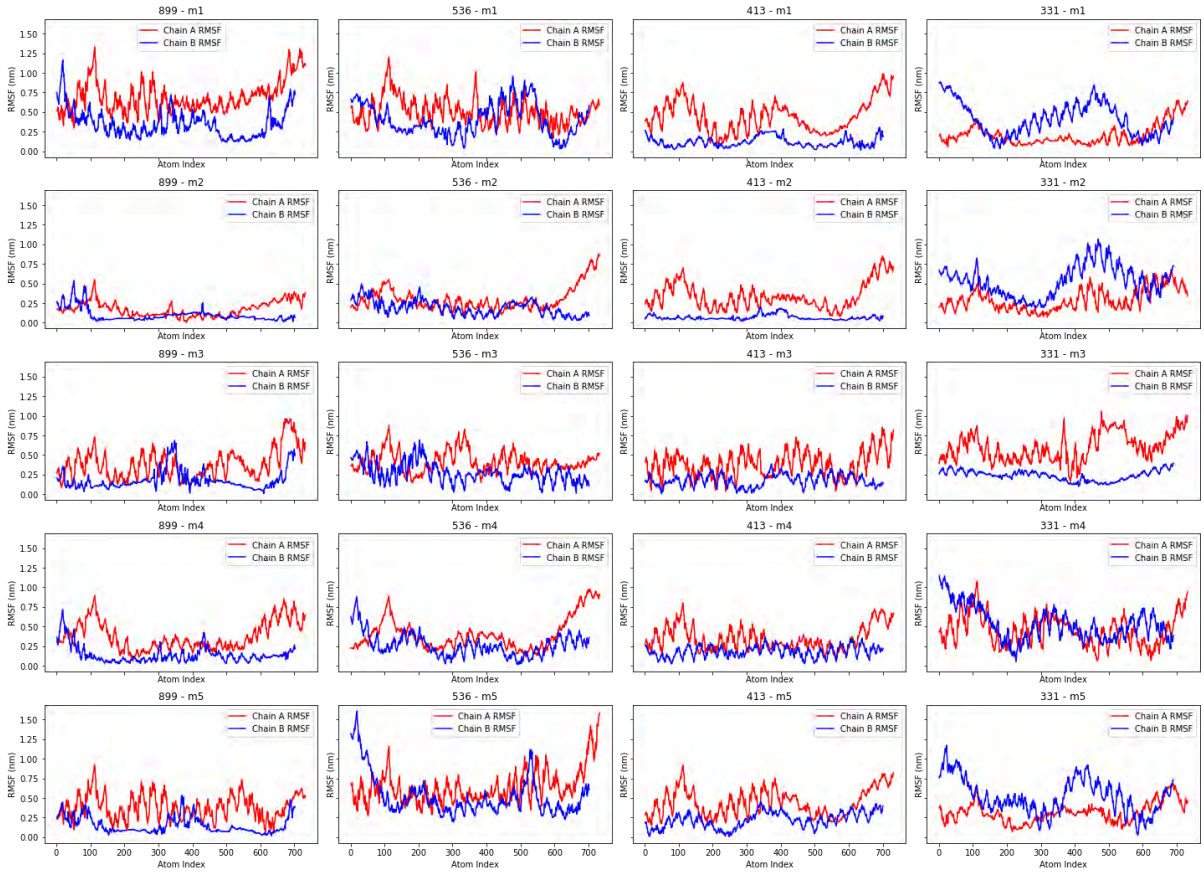


Figure 7.18: Chains RMSF result for the models of aptamer docked complex against miR-10b-5p targets, with the red line representing Chain A (miR-10b-5p) and blue representing Chain B (aptamers).

In **Figure 7.19**, Chain A representing the target miR-10b-3p, shows considerably less fluctuation of atoms compared to **Figure 7.18**, indicating regions of lower flexibility. Exceptions are model1 of aptamer734-miR-10b-3p and aptamer577-miR-10b-3p, where the fluctuations are relatively higher, suggesting that these models may have interactions that allows them to destabilize the target RNA, increasing its flexibility. Specifically, in the aptamer577-miR-10b-3p models, Chain A exhibits relatively low atomic fluctuations across all models. However, there is an emergence of high fluctuating atom indices from 280-360 in models3 to 5, indicating regions that become more flexible in response to specific interactions with the aptamer in the models. Chain B which represents the aptamers, shows a variety of fluctuating atoms with notable fluctuations in 274-m4. Where atoms from indices 300-600 fluctuate above 1.5 nm, suggesting significant conformational changes in the aptamer. In the 274 models, Chain B has fluctuating atoms at one end of the RNA except for model4 of aptamer274-miR-10b-3p, indicating that interactions in this model might stabilize one terminal

while allowing flexibility in the other. Chain B shows a range of fluctuating atoms, with model4 of aptamer274-miR-10b-3p(274-m4) displaying immense fluctuations above 1.5 nm from atom indices 300-600, suggesting that in this model, the aptamer adopts a conformation that enhances flexibility significantly. For aptamer734-miR-10b-3p models, Chain A does not exhibit significant fluctuating atoms, indicating that the binding of the aptamer734 inhibits the fluctuation of the terminal atoms and middle regions, suggesting strong stabilizing interactions. Chain B shows fluctuating atoms from indices 1-220 and 400-690, indicating that the aptamer has flexible regions not directly involved in stabilizing interactions with the RNA.

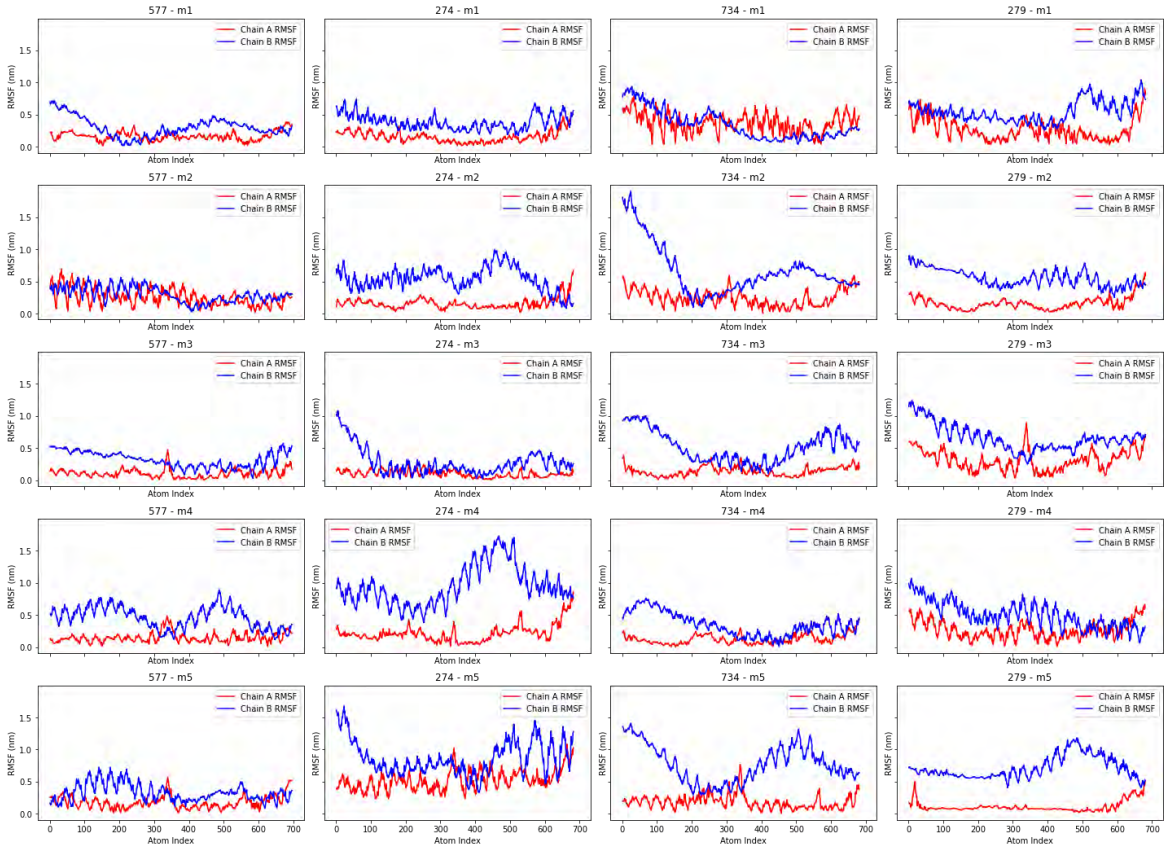


Figure 7.19: Chains RMSF result for the models of aptamer docked complex against miR-10b-3p target, with the red line representing Chain A (miR-10b-5p) and blue representing Chain B (aptamers).

In aptamer279-miR-10b-3p models, noticeable fluctuations occur in the terminal atom regions (0-40 and 630-690) across all models of Chain A, indicating that the interactions of the aptamer279 do not inhibit the fluctuation of the terminal ends but may stabilize other RNA regions. The loop regions remain stable, suggesting that the stable interaction with this aptamer is localized to these regions, allowing the terminal ends to remain flexible. With all mentioned, it is clear and important to highlight that during these MD simulation, miR-10b-3p atoms seem

to fluctuate less compare to the atoms miR-10b-5p, indicating that the miR-10b-3p is more stable within the simulations compared to the miR-10b-5p.

7.3.5.4 Rg for Complexes

The radius of gyration (Rg) is defined as the root mean square distance of the molecule's atoms from its centre of mass or from the axis of rotation, where the entire mass of the molecule is assumed to be concentrated [403]. In RNA MD simulations, Rg is used to assess the degree of compactness and folding of RNA systems. A constant Rg values throughout the simulation period signifies RNA folding stability, whereas an increase in Rg values suggests a loosening or unfolding of the structure, while a decrease indicates a tightening [404]. Although Rg is traditionally used for proteins, in this case, we used it to monitor the structural stability of these RNA complexes. The degree of compactness and folding of the RNA complexes was monitored through the Rg plotted against time, as shown in **Figure 7.20**. Rg was used to track changes in RNA complex structures relative to their native folded state, providing information about the folding and unfolding of the RNA structures during the 100 ns simulations. The results show similar trends as observed in the RMSD plots. Most of the Rg plots fluctuate significantly and gradually. Unlike proteins, RNAs are generally less tight or compact because they can often form only one loop stabilized by a few hydrogen bonds. while proteins, on the other hand are more compact due to the alpha and beta sheet structural formations, leading to more fluctuations in RNA compared to typical protein complexes. Additionally, these fluctuations can be attributed to thermal motion or influenced by the force fields used in the simulations. This suggests that most aptamer-miRNA complexes are more likely to slightly unfold and refold during the simulations. However, it is important to note that some Rg plots stabilize quite well under 2.5 nm, including model2 and 5 of aptamer899-miR-10b-5p, model1 of aptamer536-miR-10b-5p, and model1, 4, and 5 of aptamer 331-miR-10b-5p. This indicates that these RNA complexes are more stable compared to the rest. Additionally, there are some Rg plots that start with a very high value but eventually the Rg plot values decreases towards the end of simulation, this includes model4 of aptamer899-miR-10b-5p, and model3 of aptamer331-miR-10b-5p. This suggests that some of these aptamer-miR-10-5p complexes become more compact during the simulation.

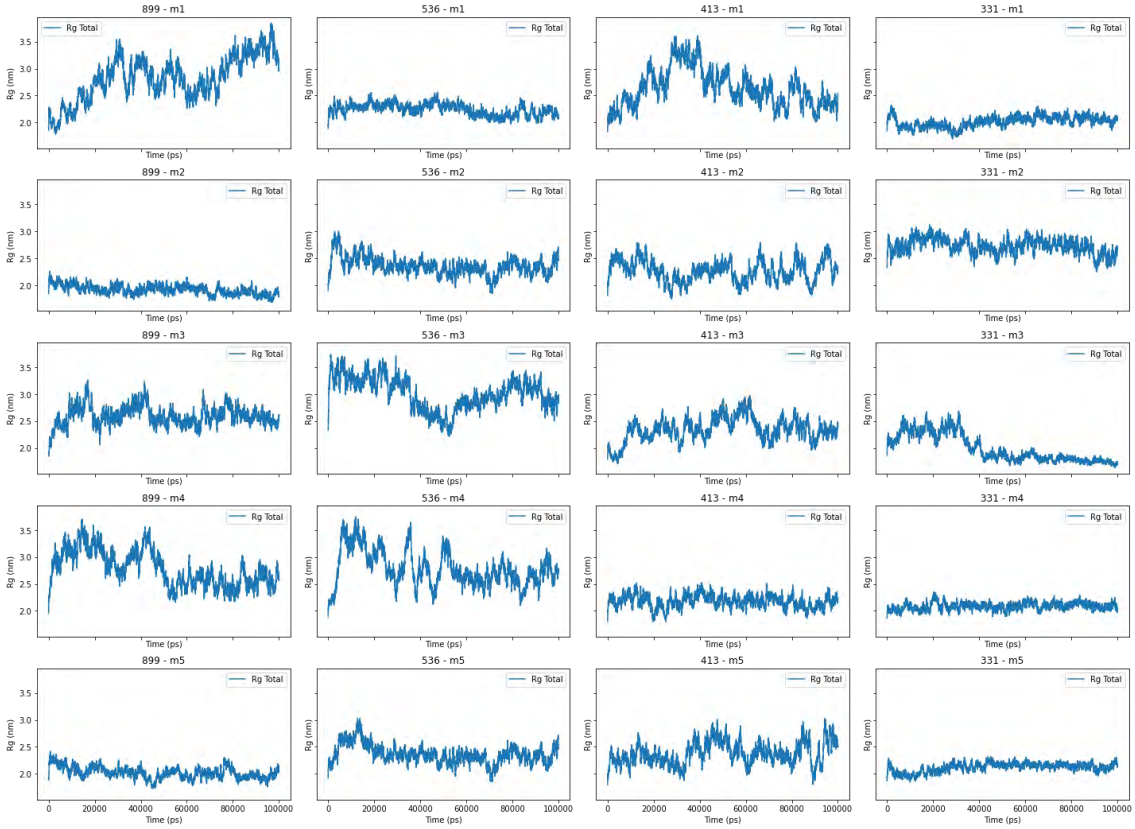


Figure 7.20: Rg results for the models of aptamer docked complex against miR-10b-5p targets (the models are denoted as m, where m1 indicate model1 of the complexes).

The Rg plots where the miR-10b-3p is the target are shown in **Figure 7.21**. Similar trends are observed, however, the actual fact is that their Rg values are more stabilized compared to those of miR-10b-5p, as they stabilize below 2 nm. Despite being better compared to the miR-10b-5p targets, there are still a few with high fluctuations, indicating that their complexes slightly tightness and loosens during the simulations. This includes model1, model2, and model5 of aptamer734-miR-10b-3p, and model5 of aptamer297. Notably, model5 of aptamer297-miR-10b-3p appears chaotic, similar to its RMSD behaviour the Rg increases abruptly. When examining the trajectories throughout the simulation, it appears that the two chains detach from each other while the other chain is unfolding at the same time (not fully complete detachment). This explains the increase in RMSD up to 6 nm.

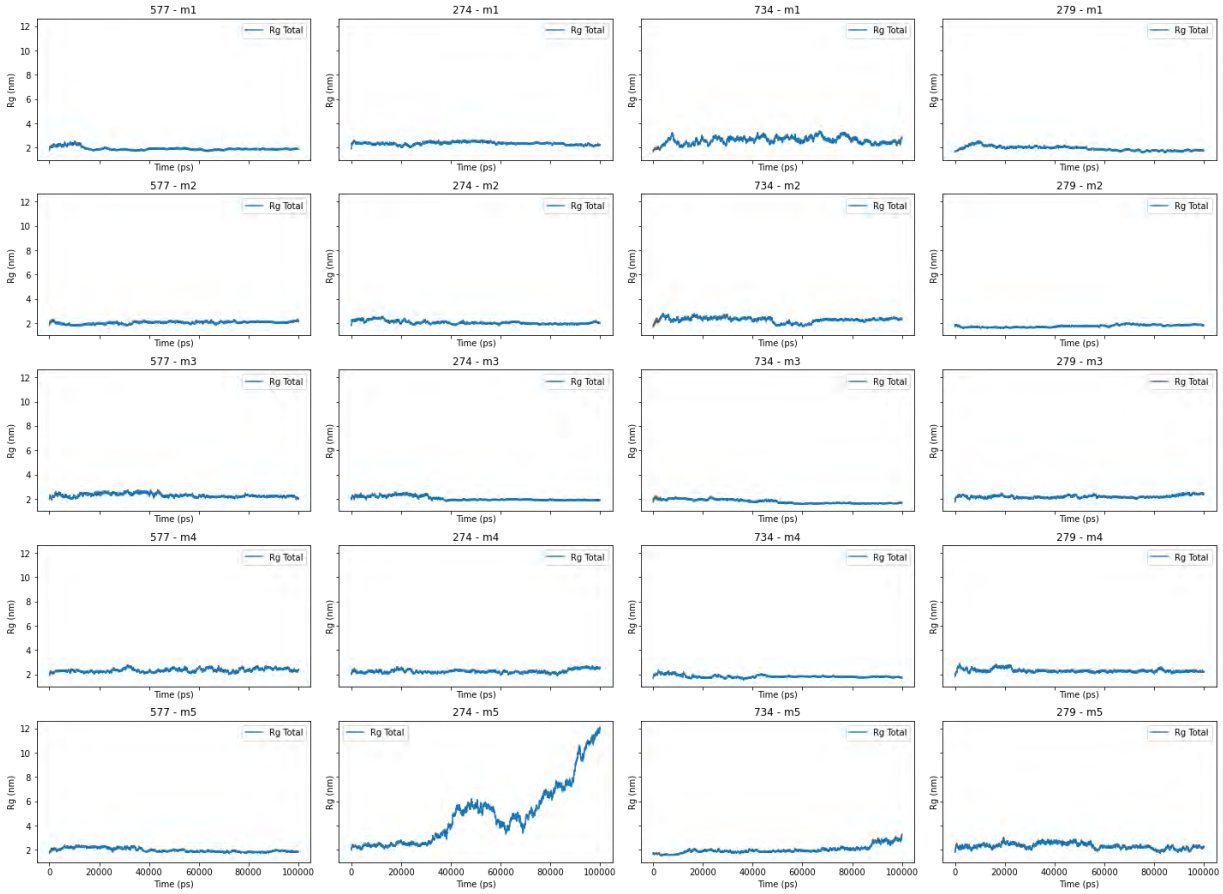


Figure 7.21: Rg result for the models of aptamer docked complex against miR-10b-5p targets

7.3.5.3 PCA for Complexes

Principal Component Analysis (PCA) is an advanced analytical tool used to identify essential motions in molecular dynamics (MD) simulations [405]. This technique operates on the variance-covariance matrix of a molecule, obtained from MD simulations. PCA is valuable for distinguishing different parts of the energy landscape sampled during the simulations, even if the process is not stationary [405]. In the context of proteins, the variance-covariance matrix is constructed from the positional fluctuations of each atom, particularly the Cα or Cβ atoms [405]. After superimposing the protein coordinates on a reference structure, such as the initial or average coordinates, the displacement vector for each residue is calculated at each time point. The covariance matrix is then diagonalized to yield eigenvectors and eigenvalues [311]. Eigenvectors indicate the directions in a high-dimensional conformational space, describing the primary motions of the molecule, while eigenvalues measure the mean square fluctuation along these directions [311]. Only a few eigenvectors with large eigenvalues are typically significant in describing the overall motions. In the context of RNA, PCA helps identify the

most prominent movements of the RNA backbone by projecting the original data onto the first two eigenvectors to create the principal components (PC1 and PC2), which contain the maximum motions.

The motions were divided into five parts to identify motions throughout every 20 ns of the simulation of these RNA complexes. Analysing these motions at each 20 ns interval is important because it allows us to observe the progression and stability of structural changes over time, and to identify any significant transitions or fluctuations within these intervals. The 2D plots we generated, with the projection of eigenvector 1 plotted against the projection of eigenvector 2, are shown in **Figure 7.22** for both miR-10b-5p and miR-10b-3p targets. In these plots, red indicates the first part of the motions (frames 0-2000), green indicates the second part (frames 2001-4000), black indicates the third part (frames 4001-6000), blue indicates the fourth part (frames 6001-8000), and purple indicates the fifth and last part (frames 8001-10000).

Examining the motions for both miR-10b-5p-aptamers and miR10b-3p-aptamers in **Figure 7.22**, it is evident that different systems display distinct patterns of motion. Some maintain clockwise motions, such as model1 of aptamer536-miR-10b-5p(536-m1), model1 and 2 of apatmer413-miR-10b-5p (413-m1, and 413-m3), and model2 of aptamer331-miR-10b-5p (331-m2) of the 5p arm target (from here on abbreviations will be used to refer to the models of these complexes to avoid being wordy), and 557-m3, 734-m2, and 279-m2 of the 3p arm target. In contrast, most systems tend to retain closed motions in both 5p and 3p targets, as observed in **Figure 7.22**. This includes 889-m3, 899-m4, 536-m3, and 331-m4 of the 5p target, and 557-m5, 274-m1, 274-m2, and 279-m3 of the 3p target. In some instances, certain systems exhibit no obvious motions or maintain similar motions throughout the simulations, suggesting that they are more compact. Others exhibit more erratic or complex trajectories, indicating variations in structural stability and dynamic behaviour across different RNA complexes, such as 273-m5, which appears chaotic. While the typical range for both PC1 and PC2 for proteins is between -10 and 10, in this instance, since the complexes are composed of two RNAs, the range is between -50 and 50 for both PC1 and PC2. This further emphasizes that RNA complexes are less compact compared to proteins. These rapid conformational changes throughout the five parts correspond to the RMSD and Rg results, reinforcing the dynamic nature of RNA structures.

5p

3p

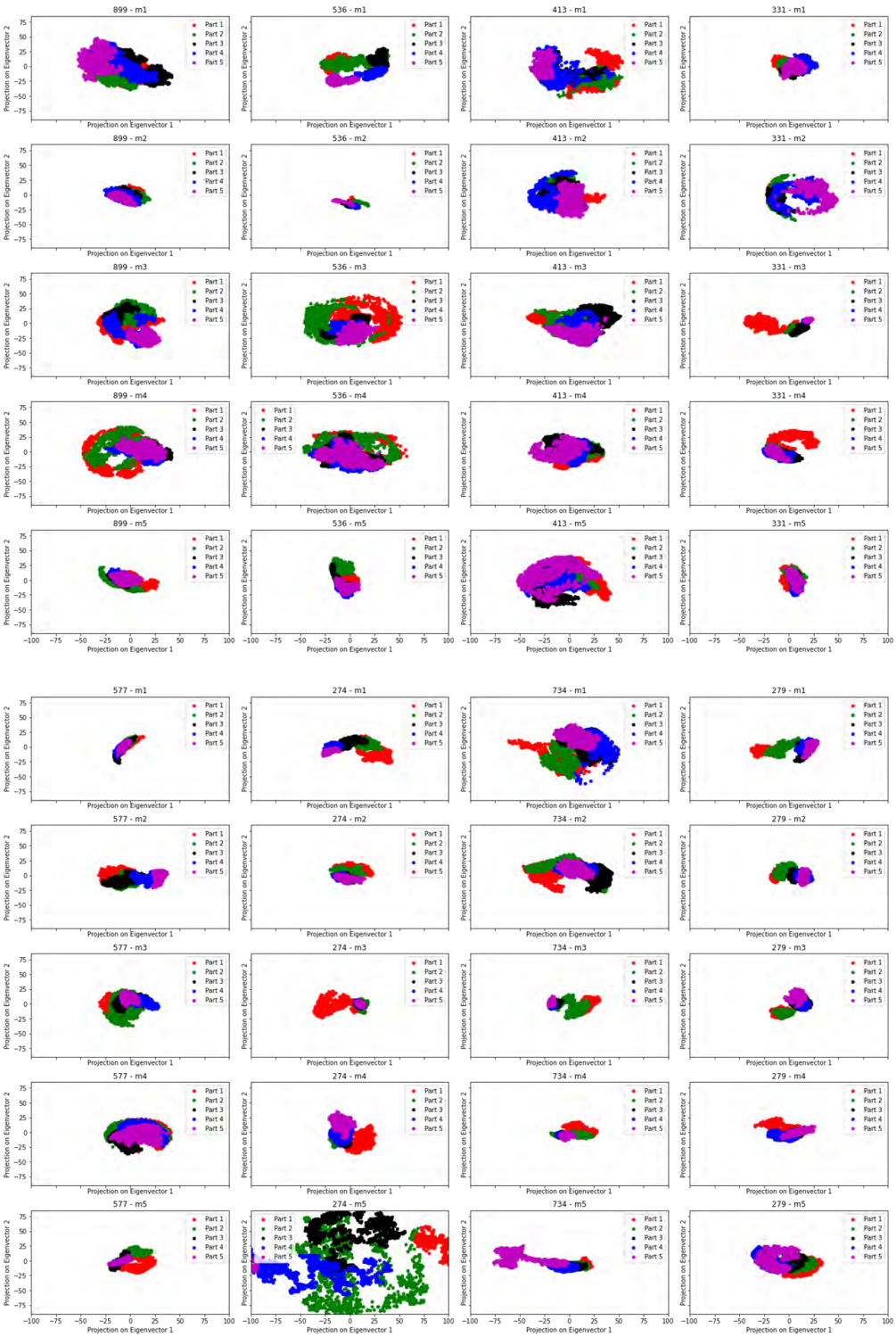


Figure 7.22: PCA result for the models of aptamer docked complex against miR-10b-3p, and miR-10b-5p targets

7.3.5.6 Hydrogen bonds

Hydrogen bonding is a significant intermolecular force that forms a special type of dipole-dipole attraction when a hydrogen atom, bonded to a highly electronegative atom, exists in the vicinity of another electronegative atom with a lone pair of electrons. These interactions, although weaker than covalent and ionic bonds, are generally stronger than ordinary dipole-dipole and dispersion forces [406,407]. Hydrogen bonds play a critical role in determining the properties and behaviours of many biological molecules, influencing their structure, stability, and function [408]. In proteins, hydrogen bonds are essential for the formation of secondary structures such as alpha helices and beta sheets. These structures are stabilized by hydrogen bonds between the backbone amide and carbonyl groups, contributing to the protein's overall three-dimensional conformation and functional dynamics. Hydrogen bonds are equally important in RNA molecules, particularly in the formation and stabilization of their secondary structures, such as hairpins, loops, and stems [409]. The hydrogen bonds between complementary base pairs (adenine-uracil and guanine-cytosine) are pivotal for the formation of double-stranded RNA regions, ensuring the proper folding and function of the molecule. In this section, we study the hydrogen bonds between two RNA molecules (aptamer and miR-10b 5p/3p) over time for the course of the simulation (100 ns). This study is important because understanding the dynamics of hydrogen bonds in RNA-RNA interactions can provide insights into the stability and conformational changes of RNA complexes. Aptamers, which are RNA molecules that bind to specific targets with high affinity, often rely on hydrogen bonding to maintain these affinities and specificities. Monitoring the hydrogen bonds between the aptamer and miR-10b-5p/3p during the simulation allows us to observe how these interactions evolve, which can influence the binding affinity and specificity of the RNA complex. Changes in hydrogen bonding patterns can indicate structural rearrangements, and change in binding stability. Again, before we get into the discussion remember that from here on abbreviations will be used to refer to the models of these complexes to avoid being wordy. Model1 to model5 are referred as m1 to m5, and the number before these models is the aptamer id associated with the complex o interest.

In **Figure 7.23**, most complexes show significant changes in the number of hydrogen bonds throughout the simulations. However, some complexes have more hydrogen bonds as the simulation progresses towards 100 ns. These include 899-m5 (model 5 of aptamer899-miR-10b-5p), model 3536-m1, 536-m3, 331-m1, and 331-m5, with the number of hydrogen bonds ranging between 10 and 20 towards the end of the simulation. This suggests that these aptamer

models tend to bind strongly to miR-10b-5p. Furthermore, this suggests that these specific models may have more stable interactions with miR-10b-5p, possibly due to favourable structural conformations or complementary binding sites/positions. Most of the other complexes, exhibit a steady range of hydrogen bonds throughout the simulation, which suggests that their binding affinity remains relatively constant. This steady interaction indicates a stable binding conformation that does not significantly fluctuate too much, highlighting the potential reliability of these aptamers in maintaining their binding affinity over time. On the other hand, there are some complexes that exhibit a very low number of hydrogen bonds, below 10. These include model1 to model4 of aptamer899-miR-10b-5p, 413-m1, and 413-m3. This suggests that these aptamer with these models may not form strong or stable interactions with miR-10b-5p. The low number of hydrogen bonds indicates weaker binding, which might be due to less favourable structural conformations or a lack of complementary binding sites between the aptamer and the target RNA.

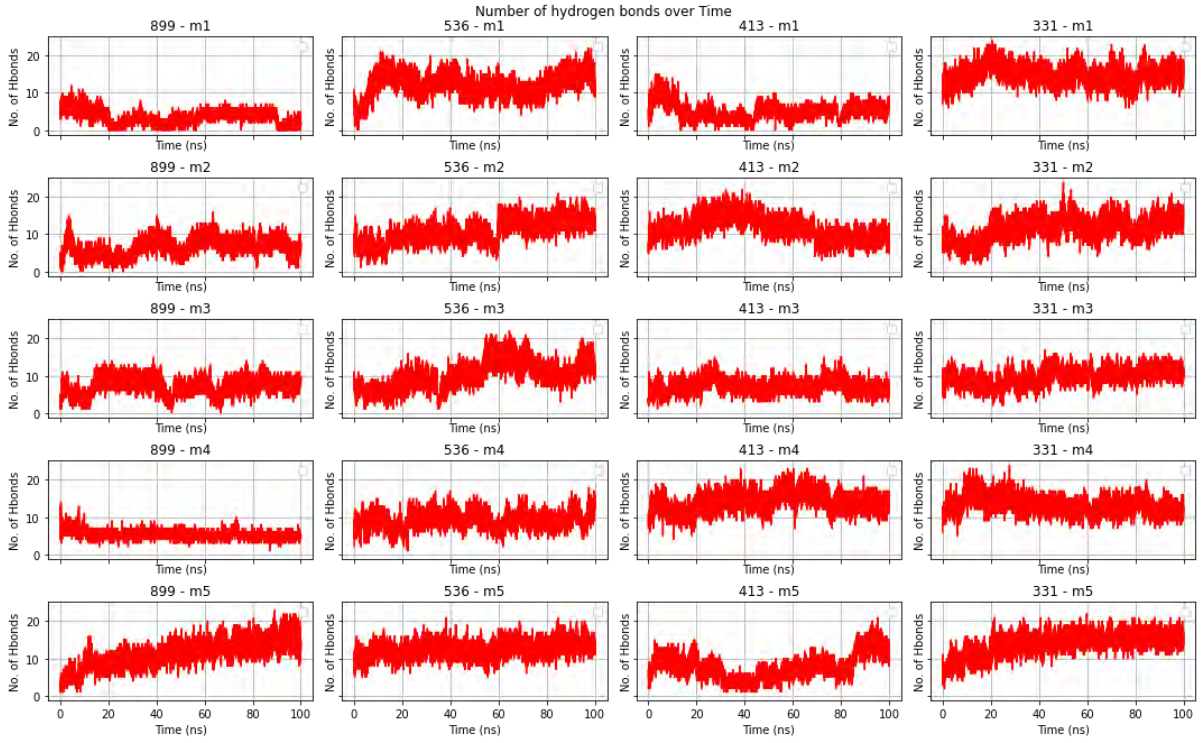


Figure 7.23: Hydrogen bonds for the models of aptamer/ miR-10b-5p docked complex. Where model1 to model5 are referred to as m1 to m5, and the number before these models is the aptamer ID associated with the complex of interest.

In **Figure 7.24**, similar trends can be observed for complexes with miR-10b-3p as the target. Some aptamers within the models shows an increase in hydrogen bonds over time, suggesting strong binding affinity towards the end of the simulation. These include 577-m1, 577-m2, 577-

m5, 274-m1, 274-m2, 734-m1, 279-m1, 279-m2, and 279-m4. These observations provide critical insights into the binding dynamics of the aptamer/miR-10b-3p complexes, highlighting which models are more likely to form stable and strong interactions. A few complexes exhibit a number of hydrogen bonds above 20, such as 577-m1, 577-m5, 274-m2, and 279-m1. This suggests that these aptamer models form particularly strong and stable interactions with miR-10b-3p. These aptamers are likely to maintain their interaction with miR-10b-3p, providing reliability in binding. Some complexes maintain a similar range of hydrogen bonds throughout the simulation, indicating a steady interaction with the target RNA. The ability of an aptamer to maintain or increase its hydrogen bonding over time can be indicative of its potential effectiveness in binding to its target miRNA, making it a valuable candidate for further development. This stability suggests that these aptamers can consistently interact with miR-10b-3p , which is beneficial for applications requiring long-term binding stability. Conversely, some complexes show a reduction in the number of hydrogen bonds throughout the simulation. This includes 274-m5 and 734-m2. This observation coincides with the PCA results, which indicated that both aptamers and the miR-10b-3p target undergo significant structural conformations. These changes may suggest that the RNAs are falling apart or losing their optimal binding configurations. Aptamers that show weak or unstable binding, as indicated by a low or fluctuating number of hydrogen bonds, may require modifications or may be less suitable for inhibition of these miRNAs. Weak binding interactions can undermine the efficacy of the aptamer, necessitating further optimization to enhance binding affinity and stability.

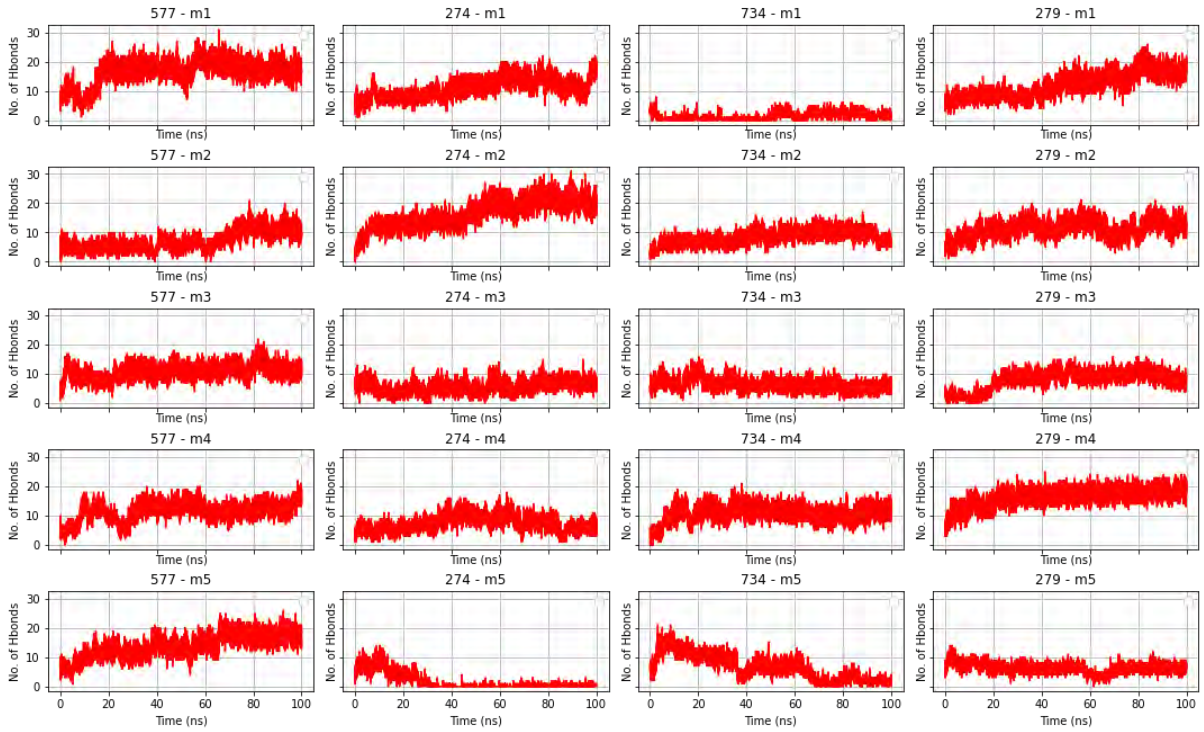


Figure 7.24: Hydrogen bonds result for the models of aptamer docked complex against miR-10b-3p target. Where model1 to model5 are referred to as m1 to m5, and the number before these models is the aptamer ID associated with the complex of interest.

7.3.6 MM-GBSA

MM-GBSA calculations were conducted on the 40 complexes for which molecular dynamics simulations were performed. Of these, 20 systems correspond to aptamers/miR-10b-3p complexes, while the remaining 20 systems are associated with aptamers/miR-10b-5p complexes. The results of these calculations are reported in this section. **Figure 7.25** provides a detailed comparison of binding energies across different models of four aptamer899, 536, 413, and 331, against the target miR-10b-3p. Aptamer899/miR-10b-3p complex models shows a mixed energy profile, model2 and 3 display relatively negative binding energies of -364.93 kJ/mol and -253.099 kJ/mol respectively, this indicates moderate to strong binding affinities. In contrast, model1 and 5 exhibit binding energies close to zero or slightly positive, with model4 notably showing a higher binding energy of -39.101 kJ/mol. This variability suggests that while certain models of aptamer899 have strong binding interactions with their targets, others may have weaker or less favourable binding characteristics. Aptamer536 demonstrates a similar trend, where model2 and 3 display negative binding energies ranging from -342.43 kJ/mol to -476.642 kJ/mol, indicating robust binding affinities. However, model2 stands out with a significantly negative binding energy of -364.93 kJ/mol, implying stronger binding affinity compared to the other models of aptamer536/miR-10b-5p. This disparity in binding

energies across models suggests variations in structural features or conformational dynamics that influence their interaction strengths with target molecules. Moving to aptamer413/miR-10b-5p, model1 and 5 show slightly negative binding energies, indicative of moderate binding affinities. In contrast, model2 exhibits a positive binding energy of 20.329 kJ/mol, suggesting less stable binding interactions or no binding at all. This indicates that while certain models of aptamer413 may effectively bind their targets, others may face challenges due to structural or energetic factors affecting their binding capabilities. Finally, aptamer331/miR-10b-5p complexes display lower binding energies in model1 and 5, -660.282 kJ/mol and -544.148 kJ/mol respectively, highlighting very strong binding affinities.

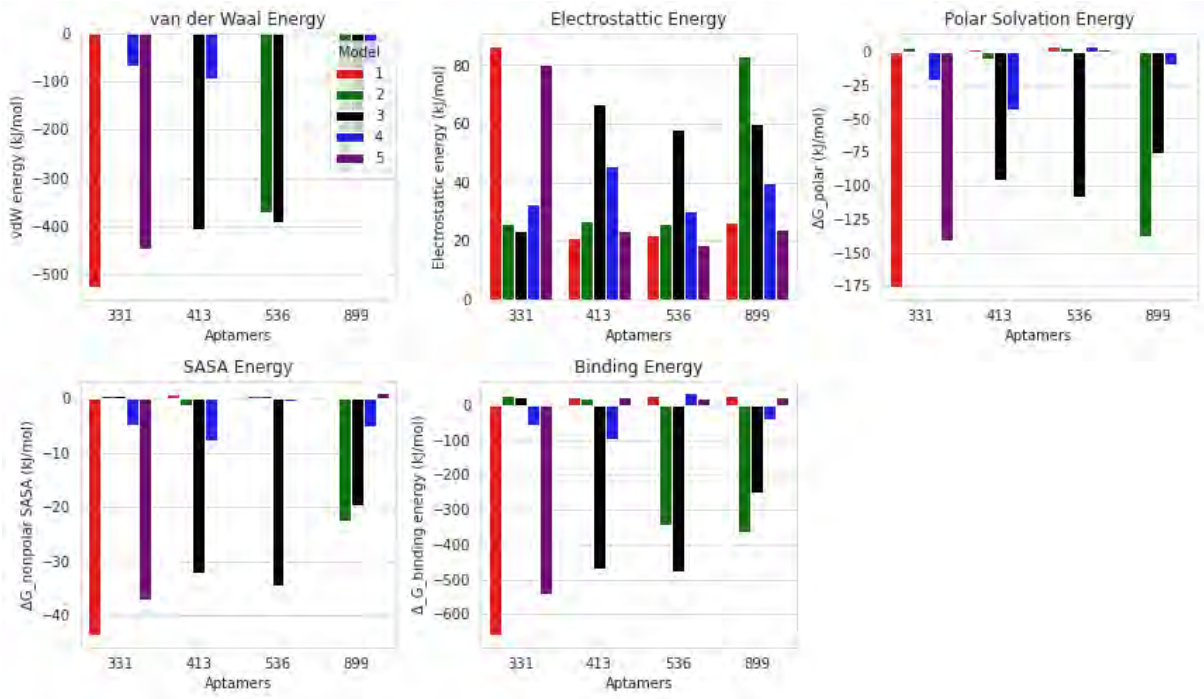


Figure 7.25: MMGBSA results for the models of aptamer docked complex against miR-10b-5p target

The electrostatic energies in **Figure 7.25** are positive across several aptamer models, this reflects the repulsive interactions within the aptamer molecules themselves or with their surroundings. The positively charged backbone of RNA aptamers and the dynamic fluctuations in their conformations can contribute significantly to these higher electrostatic energies. This suggests that the aptamers may experience substantial energy penalties due to their charged nature and the variability in their shapes, which could impact their binding affinities and stability. The positive electrostatic energies and variable binding energies highlight the complexity of aptamer-target interactions. aptamers with higher positive electrostatic energies may face challenges in forming stable complexes due to repulsive forces or less favourable

molecular orientations with their targets. In contrast, aptamers with more negative binding energies are likely better suited for strong and specific binding interactions.

Figure 7.26. shows the MM-GBSA results where miR-10b-3p is the target molecule. One standout from **Figure 7.26** is aptamer274/ miR-10b-3p, model1, which exhibits the most negative binding energy at -701.202 kJ/mol. This indicates exceptionally strong binding affinity, suggesting a robust interaction with miR-10b-3p. Such a low negative value signifies a stable complex formation, likely driven by significant contributions from van der Waals and electrostatic interactions. In comparison, aptamer734/ miR-10b-3p, model2, also shows notable strength with a binding energy of -460.337 kJ/mol, making it one of the strongest binders among the listed aptamer models. This aptamer exhibits a favourable interaction profile, indicating efficient recognition and binding to its target. The differences observed between models within each aptamer series (e.g., aptamer734 model1-5) highlight variations in binding affinities possibly due to structural differences or conformational changes and bind positions of the aptamer to the miR-10b-3p. Across aptamer577/miR-10b-3p and aptamer 279/miR-10b-3p complexes, different models also display varying degrees of binding energies. For instance, aptamer577miR-10b-3p, model2 shows a notable negative binding energy of -604.52 kJ/mol indicating strong binding, even though slightly less than the top performers from aptamers274 and 734. Similarly, aptamer279/ miR-10b-3p demonstrates variability among its models, with differences in binding affinities observed between model1 and model2-5. All other energies, such as van der Waals interactions, electrostatic energy, polar solvation, and SASA (Solvent Accessible Surface Area) energy, contribute to the overall binding energy.

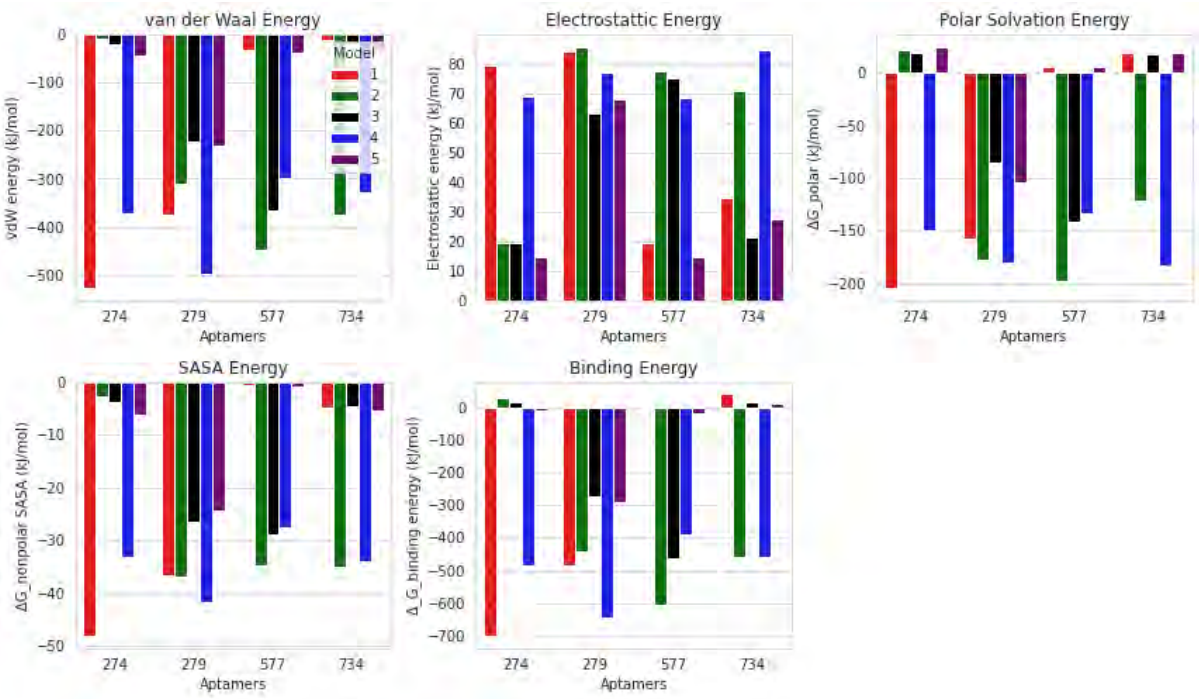


Figure 7.26: MM-GBSA results for the models of aptamer docked complex against miR-10b-3p target

7.4 Conclusion

This study provided an extensive evaluation of miRNA-aptamer interactions by assessing the interaction energies of aptamers against various miRNA targets, specifically pre-miR10b and its mature arms (5p and 3p). Heatmap analysis and average interaction energies demonstrated that aptamers exhibited the strongest interactions with hsa-miR-25-5p. Molecular docking results revealed aptamer557 as the most effective binder to the 3p arm, significantly outperforming other aptamers with a docking score of -545.96. Aptamer274, 279, and 128 also demonstrated strong binding, though with notable variations in scores. For the 5p arm, aptamer899 led with a docking score of -482.55, followed by aptamer536, 413, 332, and 278, each displaying significant binding potential but with varying fitness quality. High confidence scores (above 0.99) further validated the reliability of these docking predictions. To further evaluate the stability and electronic properties of these aptamers-miRNA complex, quantum mechanical (QM) calculations were performed. Semi-empirical single-point calculations assessed the stability through total energy (E_{tot}) and electronic properties *via* the HOMO-LUMO energy gap (H-L gap). Regarding the H-L gap energies and total energies, no significant trend was observed for all models of the aptamer-miR-10b-3p/5p complexes. Molecular dynamics simulation results of aptamer-miR-10b-3p/5p complexes were analysed by integrating multiple metrics such as hydrogen bonding, stability metric, Rg, RMSF, and RMSD for a comprehensive understanding of their stability, flexibility, and conformational changes. The RMSF data for aptamers-miR-10b-5p complexes revealed significant fluctuating atoms of the target RNA chain in many models, indicating regions of high flexibility and weaker interactions with the aptamers. This observation was supported by the hydrogen bonding data, where several models of some complex such as 899-m1 to 899-m4 and 413-m1, exhibited fewer hydrogen bonds, suggesting weaker and less stable interactions. Similarly, the RMSD data for aptamer-miR-10b-5p complexes showed that many models have the RMSD thermally equilibrating at high RMSD values, reinforcing the notion of significant conformational changes and less stable interactions. In contrast, the RMSF data for aptamers-miR-10b-3p complexes indicated generally lower fluctuations, reflecting more stable interactions with the aptamers. This stability is further supported by the hydrogen bonding data, where models such as 577-m1, 577-m5, 274-m2, and 279-m1 exhibited a high number of hydrogen bonds throughout the simulations, indicating strong and stable interactions. The RMSD data for miR-10b-3p also showed lower deviations, suggesting more stable structures. The lower RMSF and RMSD values, combined with the higher hydrogen bond counts,

highlight the more stable and rigid RNA structural complex due to stronger aptamer binding in miR-10b-3p compared to the more flexible and less stable interactions observed in miR-10b-5p. MM-GBSA calculations provided valuable insights into the binding energies of the aptamers, complementing the molecular docking and MD results. Based on the binding energy, aptamers show to perform very well against miR-10b-3p compared to miR-10b-5p. ***(This work to be submitted to Nucleic Acids Research)***

Chapter 8

Conclusion and future work

This project aimed to design and discover novel aptamers that effectively inhibit miRNA-10b and their mature miRNAs (3p and 5p) through computational approaches, leveraging the role of miRNA-10b in cancer progression as a therapeutic target. By focusing on miRNA-10b, which is implicated in promoting cancer cell proliferation, migration, and metastasis, we sought to identify promising aptamers that could potentially improve patient outcomes. The project emphasized the efficiency and cost-effectiveness of virtual screening methods compared to traditional SELEX processes. We developed tools and algorithms for generating and designing RNA aptamer sequences, predicting their secondary and tertiary structures, and virtually screening these aptamers for effective binding to miRNA-10b. The final step involved validating aptamer-miRNA-10b complexes using Molecular Dynamics (MD) simulations and Molecular Mechanics Generalized Born Surface Area (MMGBSA) calculations to assess binding affinities and stability.

Chapter 2 introduced the T_SELEX program, an advanced tool designed for aptamer selection and virtual screening. This program integrates various algorithms to handle RNA aptamer design, structure prediction, and virtual screening efficiently. The program was successfully tested against HIV-1 protease, demonstrating its robustness and versatility.

Chapter 3 detailed the generation and analysis of RNA aptamer sequences using T_SELEX program, highlighting how nucleotide composition and statistical methods inform the design and prediction of aptamer efficacy. Findings revealed critical insights into nucleotide distributions and sequence patterns, establishing a foundation for further *in silico* aptamer development.

Chapter 4 focused on large-scale secondary and tertiary structure predictions. We emphasized that while Minimum Free Energy (MFE) provides an initial measure of RNA stability, it does not always correlate directly with structural compactness or complexity. Aptamers with similar MFE values were shown to have notable variations in their secondary structure arrangement,

influenced by factors such as loop dimensions and nucleotide composition, particularly guanine and cytosine content and placement.

Chapter 5 introduced the Sequence Similarity Check (SSC) algorithm, highlighting its innovative approach to analysing sequence relationships within datasets. SSC challenged conventional methods by focusing on internal comparisons rather than relying on external databases since the algorithms have the capacity to account for the secondary structures. The findings demonstrated that aptamers with similar base compositions and minimal mismatches could exhibit significant variations in folding and stability, providing new perspectives on aptamer design and optimization.

Chapter 6 focused on benchmarking the Base Randomization Algorithm (BRA) for generating virtual RNA aptamer libraries. The study revealed that BRA-generated aptamers followed Gaussian distributions in base compositions and highlighted the importance of aptamer length in determining folding aptamer behaviour. Aptamers shorter than 7 nucleotides were found to lack effective folding capacity, suggesting a 7 nt as a critical threshold for aptamer length in order to produce a folded aptamer.

Chapter 7 provided an in-depth case study on aptamers targeting pre-miR10b and its mature arms (5p and 3p) as a novel anticancer therapeutics approach. RNA-RNA interaction predictions for aptamers against 16 known oncogenic miRNAs, both premature and mature miRNAs, were performed. Results from RNA-RNA interaction predictions revealed that aptamers showed the strongest binding with hsa-miR-25-5p, with aptamer557 emerging as the most effective binder to the 3p arm. Docking results highlighted aptamer274, 279, 128, and 899 as strong candidates, with varying binding strength based on the docking scores. MD simulations further confirmed the stability of these aptamer-miRNA complexes, with complexes such as aptamer331-miR-10b-3p and aptamer536-miR-10b-3p demonstrating consistent stabilization and low Root Mean Square Deviation (RMSD). Principal Component Analysis (PCA) and Root Mean Square Fluctuation (RMSF) analyses illustrated the dynamic nature of RNA structures and the influence of aptamer binding on RNA flexibility. MM-GBSA calculations supported the docking results, identifying aptamer274 and 734 as particularly robust binders with highly negative binding energies.

Overall, this project has advanced the understanding of aptamer design and validation, particularly in targeting miRNA-10b for cancer therapy. The integration of computational tools and algorithms has streamlined the process of aptamer discovery, offering a more efficient and

cost-effective approach to developing novel therapeutic agents. The comprehensive findings across all chapters provide valuable research contributions into the design, stability, and efficacy of RNA aptamers, paving the way for future research and application in targeted cancer therapies, specifically targeting the miRNAs, which are one of the main contributors to cancer development.

Building on the success of this project, future work will expand the scope of aptamer research by designing and screening the aptamers against additional oncogenic miRNAs, such as miR-135b, miR-155, miR-20b, and miR-519c. This expanded focus aims to identify effective aptamers for a broader range of cancer-related miRNA targets, enhancing our ability to develop targeted therapies for various malignancies. Additionally, we plan to design and implement a comprehensive database to store and provide access to tertiary structures of aptamers and their properties. This database will address the current gap in the accessibility of high-quality aptamer data. With the aptamers database available, this will attract the users of machine learning and other advanced computational techniques, which will support the growing field of aptamer research and contribute to the advancement of novel therapeutic strategies.

REFERENCES

[1] WHO, “Global Country Profiles on Burden of Cancer a to k,” no. 2019, p. 196, 2020, [Online]. Available: https://www.who.int/docs/default-source/documents/health-topics/cancer/global-country-profiles-on-burden-of-cancer-a-to-k.pdf

[2] K. Takabe and M. G. K. Benesch, “Types of Cancer and Research Covered in World Journal of Oncology,” World J. Oncol., vol. 13, no. 6, pp. 325–328, 2022, doi: 10.14740/WJON1558.

[3] Ehrlich, M., 2009. DNA hypomethylation in cancer cells. Epigenomics, 1(2), pp.239-259. https://www.tandfonline.com/doi/abs/10.2217/epi.09.33

[4] Moreno-Sánchez, R., Rodríguez-Enríquez, S., Marín-Hernández, A. and Saavedra, E., 2007. Energy metabolism in tumor cells. The FEBS journal, 274(6), pp.1393-1418.

[5] Jain, S.L., 2013. Malignant: How cancer becomes us. Univ of California Press. https://books.google.co.za/books?hl=en&lr=&id=t6kwDwAAQBAJ&oi=fnd&pg=PA1&dq=MALIGNANT&ots=Lameyj54Ad&sig=fh935einFhmiqhcG2WLD91cs5xk&redir_esc=y#v=onepage&q=MALIGNANT&f=false

[6] Oegema, K. and Hyman, T., 2006. Cell division. WormBook: The Online Review of C. elegans Biology [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK19681/

[7] Scholey, J.M., Brust-Mascher, I. and Mogilner, A., 2003. Cell division. Nature, 422(6933), pp.746-752. https://www.nature.com/articles/nature01599

[8] Preston-Martin, S., Pike, M.C., Ross, R.K., Jones, P.A. and Henderson, B.E., 1990. Increased cell division as a cause of human cancer. Cancer research, 50(23), pp.7415-7421. https://aacrjournals.org/cancerres/article/50/23/7415/496102/Increased-Cell-Division-as-a-Cause-of-Human

[9] Ciaramella, A. and Poli, P., 2001. Assessment of depression among cancer patients: the role of pain, cancer type and treatment. Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer, 10(2), pp.156-165. https://onlinelibrary.wiley.com/doi/abs/10.1002/pon.505?casa_token=YsFXQT4XMb0AAAAA:qQ-yO-U-nks0rPBYv0RJcbVQFlQgay2wIbS3NIzb05QHIbPH-X0eFrh8XDzXUooij13w72nB9tBBKzA8

[10] Vogelstein, B., Fearon, E.R., Kern, S.E., Hamilton, S.R., Preisinger, A.C., Nakamura, Y. and White, R., 1989. Allelotype of colorectal carcinomas. Science, 244(4901), pp.207-211. https://www.science.org/doi/abs/10.1126/science.2565047

[11] Breuninger, H., Black, B. and Rassner, G., 1990. Microstaging of squamous cell carcinomas. American journal of clinical pathology, 94(5), pp.624-627. https://academic.oup.com/ajcp/article-abstract/94/5/624/1779660

[12] HaDuong, J.H., Martin, A.A., Skapek, S.X. and Mascarenhas, L., 2015. Sarcomas. Pediatric Clinics, 62(1), pp.179-200. https://www.pediatric.theclinics.com/article/S0031-3955(14)00191-6/abstract

[13] Pejovic, T. and Schwartz, P.E., 2002. Leukemias. Clinical Obstetrics and Gynecology, 45(3), pp.866-878. https://journals.lww.com/clinicalobgyn/citation/2002/09000/leukemias.33.aspx

[14] Gilliland, D.G., Jordan, C.T. and Felix, C.A., 2004. The molecular basis of leukemia. ASH Education Program Book, 2004(1), pp.80-97. https://ashpublications.org/hematology/article/2004/1/80/18694/The-Molecular-Basis-of-Leukemia

[15] Nayak, L.M. and Deschler, D.G., 2003. Lymphomas. Otolaryngologic Clinics of North America, 36(4), pp.625-646. https://www.oto.theclinics.com/article/S0030-6665(03)00033-1/abstract

[16] Barlogie, B., Shaughnessy, J., Tricot, G., Jacobson, J., Zangari, M., Anaissie, E., Walker, R. and Crowley, J., 2004. Treatment of multiple myeloma. Blood, 103(1), pp.20-32. https://ashpublications.org/blood/article/103/1/20/17583/Treatment-of-multiple-myeloma

[17] Alexanian, R. and Dimopoulos, M., 1994. The treatment of multiple myeloma. New England Journal of Medicine, 330(7), pp.484-489. https://www.nejm.org/doi/full/10.1056/NEJM199402173300709

[18] Black, P.M., 1991. Brain tumours. New England Journal of Medicine, 324(22), pp.1555-1564. https://www.nejm.org/doi/full/10.1056/NEJM199105303242205

[19] Herholz, K., Langen, K.J., Schiepers, C. and Mountz, J.M., 2012, November. Brain

tumours. In Seminars in nuclear medicine (Vol. 42, No. 6, pp. 356-370). WB Saunders. https://www.sciencedirect.com/science/article/pii/S0001299812000517?casa_token=_1YtQUtl3D8AAAAA:bOMjzs55SJXfDCntAkHZU-6Ygj2zfHRNbPTp5bZiwtDJc8xksd0fQKTwWfj7vO4dR4_JFE7_Y2Ev

[20] Blackadar, C.B., 2016. Historical review of the causes of cancer. World journal of clinical oncology, 7(1), p.54.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4734938/

[21] Higginson J, Muir CS, Munoz M. Introduction to epidemiology. In: Human Cancer: Epidemiology and Environmental Causes., editor. Cambridge, England: Cambridge University Press; 1992. pp. xvii–xxv,

[22] IARC (International Agency for Research on Cancer) Background and purpose of the IARC programme on the evaluation of the carcinogenic risk of chemicals to man. In: IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man., editor. Volume 1. Lyon: International Agency for Research on Cancer; 1972. pp. 8–14, esp 11. Available from: http://www.iarc.fr/en/publications/list/monographs/index.php or http://www.iarc.fr.

[23] Euser, A.M., Zoccali, C., Jager, K.J. and Dekker, F.W., 2009. Cohort studies: prospective versus retrospective. Nephron Clinical Practice, 113(3), pp.c214-c217. https://karger.com/nec/article/113/3/c214/831273/Cohort-Studies-Prospective-versus-Retrospective

[24] Vandenbroucke, J.P. and Pearce, N., 2012. Case–control studies: basic concepts. International journal of epidemiology, 41(5), pp.1480-1489. https://academic.oup.com/ije/article/41/5/1480/713573

[25] Pandis, N., 2014. Cross-sectional studies. American Journal of Orthodontics and Dentofacial Orthopedics, 146(1), pp.127-129. https://www.ajodo.org/article/S0889-5406(14)00443-0/fulltext

[26] Shimkin MB. Contrary to Nature: being an Illustrated commentary on Some Persons and Events of Historical Importance in the Development of Knowledge concerning Cancer. NIH Publication No. 76-720. Washington, (DC): US Department of Health, Education and Welfare; 1977.

[27] Sirica AE. Introduction: chronology of significant events in the study of neoplasia. In: Sirica AE, editor , The Pathobiology of Neoplasia, editors. New York: Plenum Press; 1989. pp. 1–24.

[28] Gross L. Oncogenic viruses, second edition. Oxford: Permagon Press; 1970.

[29] Mehta RG, Pezzuto JM. Discovery of cancer preventive agents from natural products: from plants to prevention. Current oncology reports. 2002 Dec;4:478-86.

[30] Mukhtar E, Adhami VM, Mukhtar H. Targeting microtubules by natural agents for cancer therapy. Molecular cancer therapeutics. 2014 Feb 1;13(2):275-84.

[31] CARNEVALE F. LA" FORTUNA" IMMEDIATA E DI LUNGA DURATA DEL DE MORBIS ARTIFICUM DIATRIBA (1700-1713) DI BERNARDINO RAMAZZINI. Medicina nei Secoli: Arte e Scienza. 2011 Aug 1;23(2).

[32] Brown JR, Thornton JL. Percivall Pott (1714-1788) and chimney sweepers' cancer of the scrotum. British journal of industrial medicine. 1957 Jan;14(1):68.

[33] Heldwein FL, Rhoden EL, Morgentaler A. Classics of urology: a half century history of the most frequently cited articles (1955-2009). Urology. 2010 Jun 1;75(6):1261-8.

[34] Young RH, Eble JN. The history of urologic pathology: an overview. Histopathology. 2019 Jan;74(1):184-212.

[35] Dietrich, H.G. and Golka, K., 2012. Bladder tumors and aromatic amines–Historical milestones from Ludwig Rehn to Wilhelm Hueper. Frontiers in Bioscience-Elite, 4(1), pp.279-288.

[36] Hall EJ, Brenner DJ. Cancer risks from diagnostic radiology. The British journal of radiology. 2008 May 1;81(965):362-78.

[37] Ellermann V. A new strain of transmissible leucemia in fowls (strain H). The Journal of experimental medicine. 1921 Mar 3;33(4):539.

[38] Becsei-Kilborn E. Scientific discovery and scientific reputation: the reception of Peyton Rous’ discovery of the chicken sarcoma virus. Journal of the History of Biology. 2010 Feb;43(1):111-57.

[39] Panchbhai AS. Wilhelm Conrad Röntgen and the discovery of X-rays: Revisited after centennial. Journal of Indian academy of oral medicine and radiology. 2015 Jan

1;27(1):90-5.

[40] Haddow A. Sir Ernest Laurence Kennaway FRS, 1881-1958: chemical causation of cancer then and today. Perspectives in biology and medicine. 1974;17(4):543-91.

[41] Sharpe WD. The New Jersey radium dial painters: a classic in occupational carcinogenesis. Bulletin of the History of Medicine. 1978 Dec 1;52(4):560-70.

[42] Levitz, J.S., Bradley, T.P. and Golden, A.L., 2004. Overview of smoking and all cancers. Medical Clinics, 88(6), pp.1655-1675.

[43] Thomas, X. and Chelghoum, Y., 2004. Cigarette smoking and acute leukemia. Leukemia & lymphoma, 45(6), pp.1103-1109.

[44] Underwood, J.M., Townsend, J.S., Tai, E., White, A., Davis, S.P. and Fairley, T.L., 2012. Persistent cigarette smoking and other tobacco use after a tobacco-related cancer diagnosis. Journal of Cancer Survivorship, 6, pp.333-344.

[45] Kuper, H., Boffetta, P. and Adami, H.O., 2002. Tobacco use and cancer causation: association by tumour type. Journal of internal medicine, 252(3), pp.206-224.

[46] Doll R. Conversation with Sir Richard Doll. Br J Addict. 1991;**86**:365–377.

[47] Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. Br Med J. 1950;**2**:739–748.

[48] Royal College of Physicians of London. Smoking and Health: Summary and Report of the Royal College of Physicians of London on Smoking in Relation to Cancer of the Lung and Other Diseases. London: Pitman; 1962.

[49] Talley C, Kushner HI, Sterk CE. Lung cancer, chronic disease epidemiology, and medicine, 1948-1964. J Hist Med Allied Sci. 2004;**59**:329–374.

[50] Bengtsson U, Angervall L, Ekman H, Lehmann L. Transitional cell tumours of the renal pelvis in analgesic abusers. Scand J Urol Nephrol. 1968;**2**:145–150.

[51] Thiede T, Christensen BC. [Bladder tumours induced by chlornaphazine treatment] Ugeskr Laeger. 1975;**137**:661–666.

[52] IARC (International Agency for Research on Cancer) Chlornaphazine. In: IARC

Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man Volume 100A: A Review of Human Carcinogens: Pharmaceuticals., editor. Lyon: International Agency for Research on Cancer; 2012. pp. 333–335. Available from: http://www.iarc.fr/en/publications/list/monographs/index.php or http://www.iarc.fr.

[53] Gilman A, Philips FS. The Biological Actions and Therapeutic Applications of the B-Chloroethyl Amines and Sulfides. Science. 1946;**103**:409–436.

[54] Kyle RA, Pierre RV, Bayrd ED. Multiple myeloma and acute myelomonocytic leukemia. N Engl J Med. 1970;**283**:1121–1125.

[55] Stott H, Fox W, Girling DJ, Stephens RJ, Galton DA. Acute leukaemia after busulphan. Br Med J. 1977;**2**:1513–1517.

[56] Osgood EE. Contrasting incidence of acute monocytic and granulocytic leukemias in p32-treated patients with polycythemia vera and chronic lymphocytic leukemia. J Lab Clin Med. 1964;**64**:560–573.

[57] Dahlgren S. Thorotrast tumours. A review of the literature and report of two cases. Acta Pathol Microbiol Scand. 1961;**53**:147–161.

[58] Garshick E, Laden F, Hart JE, Rosner B, Smith TJ, Dockery DW, Speizer FE. Lung cancer in railroad workers exposed to diesel exhaust. Environ Health Perspect. 2004;**112**:1539–1543.

[59] Steenland K, Schnorr T, Beaumont J, Halperin W, Bloom T. Incidence of laryngeal cancer and exposure to acid mists. Br J Ind Med. 1988;**45**:766–776

[60] Martland HS, Conlon P, Knef JP. Some unrecognized dangers in the use and handling of radioactive substances, with special reference to storage of insoluble products of radium and mesothorium in the reticulo-endothelial system. JAMA. 1925;**85**:1769–1776

[61] Cogliano VJ, Baan R, Straif K, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Freeman C, et al. PrevenTable exposures associated with human cancers. J Natl Cancer Inst. 2011;**103**:1827–1839.

[62] Koch R. Die Aetiologie der Tuberculose. Berliner Klinische Wochenschrift. 1882;**19**:221–230 [In German] Translated and edited in reference

#155.

[63] Rous P. Transmission of a malignant new growth by means of a cell-free filtrate. JAMA. 1911;**56**:198.

[64] Rous P. A Sarcoma of the fowl transmissible by an agent separable from the tumor cells. J Exp Med. 1911;**13**:397–411.

[65] Ruska E. Nobel lecture. The development of the electron microscope and of electron microscopy. Biosci Rep.

[66] Bittner JJ. Some possible effects of nursing on the mammary gland tumor incidence in mice. Science. 1936;**84**:162.

[67] Lucké B. Carcinoma in the leopard frog: its probable causation by a virus. J Exp Med. 1938;**68**:457–468.

[68] Gottlieb MS, Schanker HM, Fan PT, Saxon A, Weisman JD, Pozalski I. Pneumocystis pneumonia--Los Angeles. MMWR Morb Mortal Wkly Rep. 1981;**30**:250–252.

[69] Blumberg BS, Alter HJ, Visnich S. A "new" antigen in leukemia sera. JAMA. 1965;**191**:541–546.

[70] Klein G. Perspectives in studies of human tumor viruses. Front Biosci. 2002;**7**:d268–d274.

[71] Doisy EA, Veler CD, Thayer SA. Folliculin from the urine of pregnant women. Am J Physiol. 1929;**90**:329–330.

[72] Ewing J. The General Pathological Conception of Cancer. Can Med Assoc J. 1935;**33**:125–135.

[73] Ramazzini B. Chapter XX. Wet nurses. and reprinted in 1964. New York. In: Diseases of Workers, translated from the Latin text De Morbis Artificum of 1713 by Wilmer Cave Wright, with an introduction by George Rosen, M.D., Ph.D. Translation first published in; 1940. pp. Hafner Publishing Company, 1713: 167–201, esp 191.

[74] Lane-Claypon JE. A further report on cancer of the breast with special reference to its associated antecedent conditions. London: HMSO; 1926.

[75] MacMahon B, Cole P, Lin TM, Lowe CR, Mirra AP, Ravnihar B, Salber EJ, Valaoras VG, Yuasa S. Age at first birth and breast cancer risk. Bull World Health

Organ. 1970;**43**:209–221.

[76] Burkitt DP. The discovery of Burkitt's lymphoma. Cancer. 1983;**51**:1777–1786.

[77] Unna PG. Carcinoma of the sailor's skin. In: The Histopathology of the Diseases of the Skin by Dr P, et al., editors. G. Unna- translated from the German with assistance of the author, by Norman Walker. New York: Macmillan & Co; 1896. pp. 719–724.

[78] Ferguson AR. Associated bilharziosis and primary malignant disease of the urinary bladder, with observations on a series of forty cases. J Pathol Bacteriol. 1911;**16**:76–98.

[79] Stewart MJ. Precancerous lesions of the alimentary tract. Lancet. 1931;**218**:669–675.

[80] Hollander AW. Development of dermatopathology and Paul Gerson Unna. J Am Acad Dermatol. 1986;**15**:727–734.

[81] Burkitt D. Determining the climatic limitations of a children's cancer common in Africa. Br Med J. 1962;**2**:1019–1023.

[82] Allcroft R. Chapter IX. Aflatoxicosis in farm animals. In: Goldblatt LA, editor , Aflatoxin , scientific background, control , et al., editors. New York: Academic Press; 1969. pp. 237–264.

[83] Moss SF. The rediscovery of H. pylori bacteria in the gastric mucosa by Robin Warren, and implications of this finding for human biology and disease. Dig Dis Sci. 2013;**58**:3072–3078.

[84] Warren JR. Helicobacter: the ease and difficulty of a new discovery (Nobel lecture) ChemMedChem. 2006;**1**:672–685.

[85] Acheson ED, Cowdell RH, Hadfield E, Macbeth RG. Nasal cancer in woodworkers in the furniture industry. Br Med J. 1968;**2**:587–596.

[86] Vandenberg LN, Colborn T, Hayes TB, Heindel JJ, Jacobs DR, Lee DH, Shioda T, Soto AM, vom Saal FS, Welshons WV, et al. Hormones and endocrine-disrupting chemicals: low-dose effects and nonmonotonic dose responses. Endocr Rev. 2012;**33**:378–455.

[87] Smittenaar, C.R., Petersen, K.A., Stewart, K. and Moitt, N., 2016. Cancer incidence and mortality projections in the UK until 2035. British journal of cancer, 115(9),

pp.1147-1155.

[88] Loeb, L.A., Emster, V.L., Warner, K.E., Abbotts, J. and Laszlo, J., 1984. Smoking and lung cancer: an overview. Cancer research, 44(12_Part_1), pp.5940-5958.

[89] Anisimov VN. The relationship between aging and carcinogenesis: a critical appraisal. Critical reviews in oncology/hematology. 2003 Mar 1;45(3):277-304.

[90] Pitot, H.C., 1993. The molecular biology of carcinogenesis. Cancer, 72(S3), pp.962-970.

[91] Vogelstein, B., Kinzler, K., Lengauer, C., Kastan, M., Tomlinson, I., Markowitz, S., Greider, C. and DePinho, R., 1. RECENT RESEARCH ADVANCES Genetic instability as an underlying mechanism of cancer.

[92] Shuen, A.Y. and Foulkes, W.D., 2011. Inherited mutations in breast cancer genes—risk and response. Journal of mammary gland biology and neoplasia, 16, pp.3-15.

[93] Lønning, P.E., 2004. Genes causing inherited cancer as beacons to identify the mechanisms of chemoresistance. Trends in molecular medicine, 10(3), pp.113-118.

[94] Hasty, P., 2005. The impact of DNA damage, genetic mutation and cellular responses on cancer prevention, longevity and aging: observations in humans and mice. Mechanisms of ageing and development, 126(1), pp.71-77.

[95] Pritchard, C.C., Mateo, J., Walsh, M.F., De Sarkar, N., Abida, W., Beltran, H., Garofalo, A., Gulati, R., Carreira, S., Eeles, R. and Elemento, O., 2016. Inherited DNA-repair gene mutations in men with metastatic prostate cancer. New England Journal of Medicine, 375(5), pp.443-453.

[96] Knudson, A.G., 1986. Genetics of human cancer. Annual review of genetics, 20(1), pp.231-251.

[97] Huff, J., 1995. Mechanisms, chemical carcinogenesis, and risk assessment: cell proliferation and cancer. American journal of industrial medicine, 27(2), pp.293-300.

[98] Correa, P. and Miller, M.J., 1998. Carcinogenesis, apoptosis and cell proliferation. British medical bulletin, 54(1), pp.151-162.

[99] Tsao, A.S., Kim, E.S. and Hong, W.K., 2004. Chemoprevention of cancer. CA: a cancer journal for clinicians, 54(3), pp.150-180.

[100] Cuny, M., Kramar, A., Courjal, F., Johannsdottir, V., Iacopetta, B., Fontaine, H., Grenier, J., Culine, S. and Theillet, C., 2000. Relating genotype and phenotype in breast cancer: an analysis of the prognostic significance of amplification at eight different genes or loci and of p53 mutations. Cancer research, 60(4), pp.1077-1083.

[101] Yamagiwa, K. and Ichikawa, K., 1918. Experimental study of the pathogenesis of carcinoma. The Journal of Cancer Research, 3(1), pp.1-29.

[102] Heidelberger, C., 1975. Chemical carcinogenesis.

[103] RetinoidsandCancer, M., 1983. Role of retinoids in differentiation and carcinogenesis. Cancer Research, 43, pp.3034-3040.

[104] Siemiatycki, J. and Thomas, D.C., 1981. Biological models and statistical interactions: an example from multistage carcinogenesis. International journal of epidemiology, 10(4), pp.383-387.

[105] Farmer, P.B., Brown, K., Tompkins, E., Emms, V.L., Jones, D.J.L., Singh, R. and Phillips, D.H., 2005. DNA adducts: mass spectrometry methods and future prospects. Toxicology and applied pharmacology, 207(2), pp.293-301.

[106] Kinghorn, A.D., Su, B.N., Jang, D.S., Chang, L.C., Lee, D., Gu, J.Q., Carcache-Blanco, E.J., Pawlus, A.D., Lee, S.K., Park, E.J. and Cuendet, M., 2004. Natural inhibitors of carcinogenesis. Planta medica, 70(08), pp.691-705.

[107] Garner, R.C., 1998. The role of DNA adducts in chemical carcinogenesis. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 402(1-2), pp.67-75.

[108] Hunter, D.J., 2005. Gene–environment interactions in human diseases. Nature reviews genetics, 6(4), pp.287-298.

[109] Berrozpe, G., Schaeffer, J., Peinado, M.A., Real, F.X. and Perucho, M., 1994. Comparative analysis of mutations in the p53 and K-ras genes in pancreatic cancer. International journal of cancer, 58(2), pp.185-191.

[110] Schärer, O.D., 2003. Chemistry and biology of DNA repair. Angewandte Chemie International Edition, 42(26), pp.2946-2974.

[111] Robertson, A.B., Klungland, A., Rognes, T. and Leiros, I., 2009. DNA repair in

mammalian cells: Base excision repair: the long and short of it. Cellular and molecular life sciences, 66, pp.981-993.

[112] Seo, Y.R. and Jung, H.J., 2004. The potential roles of p53 tumor suppressor in nucleotide excision repair (NER) and base excision repair (BER). Experimental & molecular medicine, 36(6), pp.505-509.

[113] Krokan, H.E. and Bjørås, M., 2013. Base excision repair. Cold Spring Harbor perspectives in biology, 5(4), p.a012583.

[114] Kim, Y.J. and M Wilson III, D., 2012. Overview of base excision repair biochemistry. Current molecular pharmacology, 5(1), pp.3-13.

[115] Feller, L., Wood, N.H., Motswaledi, M.H., Khammissa, R.A., Meyer, M. and Lemmer, J., 2010. Xeroderma pigmentosum: a case report and review of the literature. J prev med hyg, 51(2), pp.87-91.

[116] Cadet J, Bourdat AG, D'Ham C, Duarte V, Gasparutto D, Romieu A, Ravanat JL (2000) Oxidative base damage to DNA: specificity of base excision repair enzymes. Mutat Res, 462: 121-128.

[117] McGregor DB, Rice JM, Venitt S, eds (1999) The Use of Short- and Medium-Term Tests for Carcinogens and Data on Genetic Effects in Carcinogenic Hazard Evaluation (IARC Scientific Publications No. 146), Lyon, IARCPress.

[119] Lindahl T (2000) Suppression of spontaneous mutage nesis in human cells by DNA base excision-repair. Mutat Res, 462: 129-135.

[120] Pedroni M, Sala E, Scarselli A, Borghi F , Menigatti M, Benatti P, Percesepe A, Rossi G, Foroni M, Losi L, Di Gregorio C, De Pol A, Nascimbeni R, Di Betta E, Salerni B, de Leon MP, Roncucci L (2001) Microsatellite instability and mismatch-repair protein expression in hereditary and sporadic colorectal carcinogenesis. Cancer Res, 61: 896 899

[121] Fearon ER, Vogelstein B (1990) A genetic model for colo rectal tumorigenesis. Cell, 61: 759-767

[122] Weinberg RA (1995) The molecular basis of oncogenes and tumor suppressor genes. Ann N Y Acad Sci, 758: 331 338.

[123] Savelyeva L, Schwab M (2001) Amplification of onco genes revisited: from expression profiling to clinical appli cation. Cancer Lett, 167: 115-123.

[124] Hanahan D, Weinberg RA (2000) The hallmarks of can cer. Cell, 100: 57-70.

[125] Hunter T (1991) Cooperation between oncogenes. Cell, 64: 249-270.

[126] Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM, vanTuinen P, Ledbetter DH, Barker DF, Nakamura Y (1989) Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science, 244: 217-221.

[127] Birch JM (1994) Li-Fraumeni syndrome. Eur J Cancer, 30A: 1935-1941.

[128] Zheng L, Li S, Boyer TG, Lee WH (2000) Lessons learned from BRCA1 and BRCA2. Oncogene, 19: 6159 6175.

[129] Hainaut P, Hollstein M (2000) p53 and human can cer: the first ten thousand mutations. Adv Cancer Res, 77: 81-137

[130] Bresalier RS (1997) The gatekeeper has many keys: dissecting the function of the APC gene. Gastroenterology, 113: 2009-2010.

[131] Chial, H., 2008. "Proto-oncogenes to oncogenes to cancer." Nature Education, 1(1), pp. 33.

[132] Sherr CJ (2000) The Pezcoller lecture: cancer cell cycles revisited. Cancer Res, 60: 3689-3695.

[133] Henriksson M, Luscher B (1996) Proteins of the Myc network: essential regulators of cell growth and differenti ation. Adv Cancer Res, 68: 109-182.

[134] Knudson AG, Jr. (1971) Mutation and cancer: statisti cal study of retinoblastoma. Proc Natl Acad Sci U S A, 68: 820-823

[135] Hung MC, Lau YK (1999) Basic science of HER-2/neu: a review. Semin Oncol, 26: 51-59.

[136] Sherr CJ, Roberts JM (1999) CDK inhibitors: positive and negative regulators of G1-phase progression. Genes Dev, 13: 1501-1512.

[137] Hartwell LH, Weinert TA (1989) Checkpoints: controls that ensure the order of cell cycle events. Science, 246: 629-634.

[138] Strohmaier H, Spruck CH, Kaiser P, Won KA, Sangfelt O, Reed SI (2001) Human F-box protein hCdc4 targets cyclin E for proteolysis and is mutated in a breast cancer cell line. Nature, 413: 316-322.

[139] Hainaut P, Hollstein M (2000) p53 and human cancer: the first ten thousand mutations. Adv Cancer Res, 77: 81 137.

[140] Kinzler KW, Vogelstein B (1997) Cancer-susceptibility genes. Gatekeepers and caretakers. Nature, 386: 761, 763

[141] Fearon, E.R. and Vogelstein, B., 1990. A genetic model for colorectal tumorigenesis. Cell, 61(5), pp.759-767.

[142] Kinzler, K.W. and Vogelstein, B., 1997. Cancer-susceptibility genes: Gatekeepers and caretakers. Nature, 386(6627), pp.761-763.

[143] Heldin, C.H., 1996. Protein tyrosine kinase receptors. Cancer Surveys, 27, pp.7-24.

[144] Krutovskikh, V. and Yamasaki, H., 1997. The role of gap junctional intercellular communication (GJIC) disorders in experimental and human carcinogenesis. Histology and Histopathology, 12(3), pp.761-768.

[145] Hirohashi, S., 1998. Inactivation of the E-cadherin-mediated cell adhesion system in human cancers. American Journal of Pathology, 153(2), pp.333-339.

[146] Birchmeier, E.J. and Behrens, J., 1994. Cadherin expression in carcinoma: Role in the formation of cell junctions and prevention of invasiveness. Biochimica et Biophysica Acta, 1198(1), pp.11-26.

[147] Bruzzone, R., White, T.W. and Paul, D.L., 1996. Connections with connexins: the molecular basis of direct intercellular signalling. European Journal of Biochemistry, 238(1), pp.1-27.

[148] Loewenstein, W.R. and Kanno, Y., 1966. Intercellular communication and the control of tissue growth: lack of communication between cancer cells. Nature, 209(5026), pp.1248-1249.

[149] Trosko, J.E., Chang, C.C., Madhukar, B.V. and Klaunig, J.E., 1990. Chemical, oncogene and growth factor inhibition gap junctional intercellular communication: an integrative hypothesis of carcinogenesis. Pathobiology, 58(5), pp.265-278.

[150] Yamasaki, H., Omori, Y., Zaidan-Dagli, M.L., Mironov, N., Mesnil, M. and Krutovskikh, V., 1999. Genetic and epigenetic changes of intercellular communication genes during multistage carcinogenesis. Cancer Detection and Prevention, 23(3), pp.273-279.

[151] Stoker, M.G., 1967. Transfer of growth inhibition between normal and virus-transformed cells: autoradiographic studies using marked cells. Journal of Cell Science, 2, pp.293-304. [152]

[152] Behrens, J., Jerchow, B.A., Wurtele, M., Grimm, J., Asbrand, C., Wirtz, R., Kuhl, M., Wedlich, D. and Birchmeier, W., 1998. Functional interaction of an axin homolog, conductin, with beta-catenin, APC, and GSK3beta. *Science*, 280(5363), pp.596-599.

[153] Yoshida, B.A., Sokoloff, M.M., Welch, D.R. and Rinker-Schaeffer, C.W., 2000. Metastasis-suppressor genes: a review and perspective on an emerging field. *Journal of the National Cancer Institute*, 92(21), pp.1717-1730.

[154] Simone, N.L., Paweletz, C.P., Charboneau, L., Petricoin, E.F. and Liotta, L.A., 2000. Laser capture microdissection: Beyond functional genomics to proteomics. *Molecular Diagnosis*, 5(4), pp.301-307.

[155] Fidler, I.J., 2000. Angiogenesis and cancer metastasis. *Cancer Journal for Scientists*, 6 Suppl 2, pp.S134-S141.

[156] Eccles, S.A., 2000. Cell biology of lymphatic metastasis: The potential role of c-erbB oncogene signalling. *Recent Results in Cancer Research*, 157, pp.41-54.

[157] Stracke, M.L. and Liotta, L.A., 1992. Multi-step cascade of tumor cell metastasis. *In Vivo*, 6(4), pp.309-316.

[158] Ridley, A., 2000. Molecular switches in metastasis. *Nature*, 406(6795), pp.466-467.

[159] Berman, A.E. and Kozlova, N.I., 2000. Integrins: structure and functions. *Membrane and Cell Biology*, 13(2), pp.207-244.

[160] McCawley, L.J. and Matrisian, L.M., 2000. Matrix metalloproteinases: multifunctional contributors to tumor progression. *Molecular Medicine Today*, 6(4), pp.149-156.

[161] Bergers, G., Brekken, R., McMahon, G., Vu, T.H., Itoh, T., Tamaki, K., Tanzawa, K., Thorpe, P., Itohara, S., Werb, Z. and Hanahan, D., 2000. Matrix metalloproteinase-9

triggers the angiogenic switch during carcinogenesis. *Nature Cell Biology*, 2(9), pp.737-744.

[162] Brodt, P., 1991. Adhesion mechanisms in lymphatic metastasis. *Cancer Metastasis Reviews*, 10(1), pp.23-32.

[163] Fidler, I.J., 1999. Critical determinants of cancer metastasis: rationale for therapy. *Cancer Chemotherapy and Pharmacology*, 43 Suppl, pp.S3-S10.

[164] McCawley, L.J., and Matrisian, L.M., 2000. Matrix metalloproteinases: multifunctional contributors to tumor progression. Molecular Medicine Today, 6(4), pp.149-156.

[165] Berman, A.E. and Kozlova, N.I., 2000. Integrins: structure and functions. Membrane and Cell Biology, 13(2), pp.207-244.

[166] Ridley, A.J., 2000. Molecular switches in metastasis. *Nature*, 406(6795), pp.466-467.

[167] Stracke, M.L. and Liotta, L.A., 1992. Multi-step cascade of tumor cell metastasis. *In Vivo*, 6(4), pp.309-316.

[168] Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H. Clustering and conservation patterns of human microRNAs. Nucleic acids research. 2005 Jan 1;33(8):2697-706.

[169] Schotte D, Chau JC, Sylvester G, Liu G, Chen C, Van Der Velden VH, Broekhuis MJ, Peters TC, Pieters R, Den Boer ML. Identification of new microRNA genes and aberrant microRNA profiles in childhood acute lymphoblastic leukemia. Leukemia. 2009 Feb;23(2):313-22.

[170] Borchert GM, Lanier W, Davidson BL. RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol 2006; 13: 1097–1101

[171] Peng Y, Croce CM. The role of MicroRNAs in human cancer. Signal transduction and targeted therapy. 2016 Jan 28;1(1):1-9.

[172] Macfarlane LA, Murphy PR. MicroRNA: biogenesis, function and role in cancer. Curr Genomics 2010; 11: 537–561

[173] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009; 136: 215–233

[174] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA

interactome by CLASH reveals frequent noncanonical binding. Cell 2013; 153: 654–665.

[175] Fabbri M, Paone A, Calore F, Galli R, Gaudio E, Santhanam R et al. MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response. Proc Natl Acad Sci USA 2012; 109: E2110–E2116.

[176] Fukuda T, Yamagata K, Fujiyama S, Matsumoto T, Koshida I, Yoshimura K, Mihara M, Naitou M, Endoh H, Nakamura T, Akimoto C. DEAD-box RNA helicase subunits of the Drosha complex are required for processing of rRNA and a subset of microRNAs. Nature cell biology. 2007 9(5):604-11.

[177] Fuller-Pace FV. DEAD box RNA helicase functions in cancer. RNA biology. 2013 J 1;10(1):121-32.

[178] Korchynskyi O, Landström M, Stoika R, Funa K, Heldin CH, ten Dijke P, Souchelnytskyi S. Expression of Smad proteins in human colorectal cancer. International journal of cancer. 1999;82(2):197-202.

[179] Trabucchi M, Briata P, Filipowicz W, Ramos A, Gherzi R, Rosenfeld MG. KSRP promotes the maturation of a group of miRNA precursors. Adv Exp Med Biol. 2010;700:36-42.

[180] Visone R, Croce CM. MiRNAs and cancer. The American journal of pathology. 2009 Apr 1;174(4):1131-8.

[181] O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT. c-Myc-regulated microRNAs modulate E2F1 expression. Nature 2005; 435: 839–843.

[182] Croce CM. Causes and consequences of microRNA dysregulation in cancer. Nature reviews genetics. 2009 ;10(10):704-14.

[183] Kuo G, Wu CY, Yang HY. MiR-17-92 cluster and immunity. Journal of the Formosan Medical Association. 2019;118(1):2-6.

[184] Rokavec M, Li H, Jiang L, Hermeking H. The p53/miR-34 axis in development and disease. Journal of molecular cell biology. 2014; 6(3):214-30.

[185] Fazi F, Racanicchi S, Zardo G, Starnes LM, Mancini M, Travaglini L, Diverio D, Ammatuna E, Cimino G, Lo-Coco F, Grignani F. Epigenetic silencing of the

myelopoiesis regulator microRNA-223 by the AML1/ETO oncoprotein. Cancer cell. 2007;12(5):457-66.

[186] Melo SA, Moutinho C, Ropero S, Calin GA, Rossi S, Spizzo R et al. A genetic defect in exportin-5 traps precursor microRNAs in the nucleus of cancer cells. Cancer Cell 2010; 18: 303–315.

[187] Amin AR, Karpowicz PA, Carey TE, Arbiser J, Nahta R, Chen ZG, Dong JT, Kucuk O, Khan GN, Huang GS, Mi S. Evasion of anti-growth signaling: A key step in tumorigenesis and potential target for treatment and prophylaxis by natural compounds. InSeminars in cancer biology 2015 (Vol. 35, pp. S55-S77). Academic Press.

[188] Joshua A. Role of p27kip1 in the regulation of miR-223 following contact inhibition.

[189] Tang X, Ren Y, Zeng W, Feng X, He M, Lv Y, Li Y, He Y. MicroRNA-based interventions in aberrant cell cycle diseases: Therapeutic strategies for cancers, central nervous system disorders and comorbidities. Biomedicine & Pharmacotherapy. 2024; 177:116979.

[190] Zacharias F, George D, Michail D, Ioannis P, Marianna T, Arzou B, Dimitra A, Athanasios P, Emmanuel KN. MicroRNAs determining carcinogenesis by regulating oncogenes and tumor suppressor genes during cell cycle. MicroRNA. 2020;9(2):82-92.

[191] Peng Y, Dai Y, Hitchcock C, Yang X, Kassis ES, Liu L et al. Insulin growth factor signaling is regulated by microRNA-486, an underexpressed microRNA in lung cancer. Proc Natl Acad Sci USA 2013; 110: 15043–15048

[192] Zilfou JT, Lowe SW. Tumor suppressive functions of p53. Cold Spring Harbor perspectives in biology. 2009;1(5):a001883.

[193] Georges SA, Biery MC, Kim SY, Schelter JM, Guo J, Chang AN, Jackson AL, Carleton MO, Linsley PS, Cleary MA, Chau BN. Coordinated regulation of cell cycle transcripts by p53-Inducible microRNAs, miR-192 and miR-215. Cancer research. 2008 Dec 15;68(24):10105-12.

[194] Jiang X, Liu Y, Zhang G, Lin S, Wu J, Yan X, Ma Y, Ma M. Aloe-Emodin Induces Breast Tumor Cell Apoptosis through Upregulation of miR-15a/miR-16-1 That Suppresses BCL2. Evidence-Based Complementary and Alternative Medicine.

2020;2020(1):5108298.

[195] Aqeilan RI, Calin GA, Croce CM. miR-15a and miR-16-1 in cancer: discovery, function and future perspectives. Cell Death & Differentiation. 2010;17(2):215-20.

[196] Li HQ, Pan ZY, Yang Z, Zhang DB, Chen Q. Overexpression of microRNA-122 resists oxidative Stress-Induced human umbilical vascular endothelial cell injury by inhibition of p53. BioMed Research International. 2020;2020(1):9791608.

[197] Zhao H, Li X, Yang L, Zhang L, Jiang X, Gao W, Chen P, Cheng Y, Wang F, Liu J. Isorhynchophylline relieves ferroptosis-induced nerve damage after intracerebral hemorrhage via miR-122-5p/TP53/SLC7A11 pathway. Neurochemical research. 2021 Aug;46(8):1981-94.

[198] Nagata S. Fas and Fas ligand: a death factor and its receptor. Advances in immunology. 1994;57:129-4

[199] Iser IC, Pereira MB, Lenz G, Wink MR. The epithelial-to-mesenchymal transition-like process in glioblastoma: an updated systematic review and in silico investigation. Medicinal Research Reviews. 2017;37(2):271-313.

[200] Cheng GZ, Chan J, Wang Q, Zhang W, Sun CD, Wang LH. Twist transcriptionally up-regulates AKT2 in breast cancer cells leading to increased migration, invasion, and resistance to paclitaxel. Cancer research. 2007 ;67(5):1979-87.

[201] Liu S, Kumar SM, Lu H, Liu A, Yang R, Pushparajan A, Guo W, Xu X. MicroRNA-9 up-regulates E-cadherin through inhibition of NF-κB1–Snail1 pathway in melanoma. The Journal of pathology. 2012;226(1):61-72.

[202] Lou W, Liu J, Gao Y, Zhong G, Chen D, Shen J, Bao C, Xu L, Pan J, Cheng J, Ding B. MicroRNAs in cancer metastasis and angiogenesis. Oncotarget. 2017;8(70):115787.

[203] Wang Q, Zhang F, Lei Y, Liu P, Liu C, Tao Y. microRNA-322/424 promotes liver fibrosis by regulating angiogenesis through targeting CUL2/HIF-1α pathway. Life Sciences. 2021 Feb 1;266:118819.

[204] Gallach S, Calabuig-Fariñas S, Jantus-Lewintre E, Camps C. MicroRNAs: promising new antiangiogenic targets in cancer. BioMed research international. 2014;2014(1):878450.

[205] Jiang R, Zhang C, Liu G, Gu R, Wu H. MicroRNA-107 promotes proliferation, migration, and invasion of osteosarcoma cells by targeting tropomyosin 1. Oncology research. 2017;25(8):1409.

[206] Kiriakidis S, Henze AT, Kruszynska-Ziaja I, Skobridis K, Theodorou V, Paleolog EM, Mazzone M. Factor-inhibiting HIF-1 (FIH-1) is required for human vascular endothelial cell survival. The FASEB Journal. 2015 Jul;29(7):2814-27.

[207] Tregub PP, Ibrahimli I, Averchuk AS, Salmina AB, Litvitskiy PF, Manasova ZS, Popova IA. The role of microRNAs in epigenetic regulation of signaling pathways in neurological pathologies. International journal of molecular sciences. 2023, 17;24(16):12899.

[208] Chen L, Huang K, Yi K, Huang Y, Tian X, Kang C. Premature MicroRNA-based therapeutic: A “one-two punch” against cancers. Cancers. 2020, 18;12(12):3831.

[209] Tian H, Li Z, Peng D, Bai X, Liang W. Expression difference of miR-10b and miR-135b between the fertile and infertile semen samples (p). Forensic Science International: Genetics Supplement Series. 2017 Dec 1;6:e257-9.

[210] Tian Y, Luo A, Cai Y, Su Q, Ding F, Chen H, Liu Z. MicroRNA-10b promotes migration and invasion through KLF4 in human esophageal cancer cell lines. Journal of Biological Chemistry. 2010 Mar 12;285(11):7986-94.

[211] Wang J, Yan Y, Zhang Z, Li Y. Role of miR-10b-5p in the prognosis of breast cancer. PeerJ. 2019 ;7:e7728.

[212] Liu, Z.; Guo, H.; Wu, J.; Zavadil, J.; Ghanny, S.; Ghuo, S.; Wei, J. Differential Expression of MiRNAs in Uterine Leiomyoma and Adjacent Myometrium of Different Races. Am J Clin Exp Obstet Gynecol. 2015, 2, 45–256.

[213] Yoshida K, Yokoi A, Kitagawa M, Sugiyama M, Yamamoto T, Nakayama J, Yoshida H, Kato T, Kajiyama H, Yamamoto Y. Downregulation of miR-10b-5p facilitates the proliferation of uterine leiomyosarcoma cells: A microRNA sequencing-based approach. Oncology Reports. 2023 May 1;49(5):1-9.

[214] Du W, Chen D, Wei K, Yu D, Gan Z, Xu G, Yao G. MiR-10b-5p Impairs TET2-mediated inhibition of PD-L1 transcription thus promoting immune evasion and tumor progression in glioblastoma. The Tohoku Journal of Experimental Medicine.

2023;260(3):205-14.

[215] Yoo B, Ross A, Pantazopoulos P, Medarova Z. MiRNA10b-directed nanotherapy effectively targets brain metastases from breast cancer. Scientific reports. 2021, 2;11(1):2844.

[216] Yoo B, Greninger P, Stein GT, Egan RK, McClanaghan J, Moore A, Benes CH, Medarova Z. Potent and selective effect of the mir-10b inhibitor MN-anti-mir10b in human cancer cells of diverse primary disease origin. PLoS One. 2018, 20;13(7):e0201046.

[217] Zhang Q, Tang J, Ran R, Liu Y, Zhang Z, Gao H, He Q. Development of an anti-microbial peptide-mediated liposomal delivery system: a novel approach towards pH-responsive anti-microbial peptides. Drug Delivery. 2016 May 3;23(4):1163-70.

[218] Balacescu O, Visan S, Baldasici O, Balacescu L, Vlad C, Achimas-Cadariu P. MiRNA-based therapeutics in oncology, realities, and challenges. Antisense Ther. 2019 Nov 20;15:1-27.

[219] Ellington AD, Szostak JW. Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures. Nature. 1992 Feb 27;355(6363):850-2.

[220] Ellington AD, Szostak JW. Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures. Nature. 1992 Feb 27;355(6363):850-2.

[221] Gold L, Janjic N, Jarvis T, Schneider D, Walker JJ, Wilcox SK, Zichi D. Aptamers and the RNA world, past and present. Cold Spring Harbor perspectives in biology. 2012 Mar 1;4(3):a003582.

[222] Bayat P, Nosrati R, Alibolandi M, Rafatpanah H, Abnous K, Khedri M, Ramezani M. SELEX methods on the road to protein targeting with nucleic acid aptamers. Biochimie. 2018 Nov 1;154:132-55.

[223] Darmostuk M, Rimpelova S, Gbelcova H, Ruml T. Current approaches in SELEX: An update to aptamer selection technology. Biotechnology advances. 2015 Nov 1;33(6):1141-61.

[224] Cossu J, Ravelet C, Martel-Frachet V, Peyrin E, Boturyn D. Peptide-based CE-SELEX enables convenient isolation of aptamers specifically recognizing CD20-expressing cells. Bioorganic & Medicinal Chemistry. 2024 Aug 1;110:117831.

[225] Kärkkäinen RM. Production of DNA aptamers with specificity for bacterial food pathogens.2012.

[226] Proske D, Blank M, Buhmann R, Resch A. Aptamers—basic research, drug development, and clinical applications. Applied microbiology and biotechnology. 2005 Dec;69:367-74.

[227] Bauer M, Strom M, Hammond DS, Shigdar S. Anything you can do, I can do better: Can aptamers replace antibodies in clinical diagnostic applications?. Molecules. 2019 Nov 30;24(23):4377.

[228] Agnello L, Camorani S, Fedele M, Cerchia L. Aptamers and antibodies: rivals or allies in cancer targeted therapy?. Exploration of Targeted Anti-tumor Therapy. 2021;2(1):107.

[229] Toh SY, Citartan M, Gopinath SC, Tang TH. Aptamers as a replacement for antibodies in enzyme-linked immunosorbent assay. Biosensors and bioelectronics. 2015 Feb 15;64:392-403.

[230] Weber PA, LEVEL Study Group. Efficacy and Safety of Maintenance Therapy With Pegaptanib Sodium in Neovascular AMD (NV-AMD)-1-Year Results of the LEVEL Study. Investigative Ophthalmology & Visual Science. 2009 Apr 28;50(13):5227-.

[231] McKeague M, Giamberardino A, DeRosa MC. Advances in aptamer-based biosensors for food safety. InEnvironmental biosensors 2011 Jul 18. IntechOpen.

[232] Rosenfeld PJ, Heier JS, Hantsbarger G, Shams N. Tolerability and efficacy of multiple escalating doses of ranibizumab (Lucentis) for neovascular age-related macular degeneration. Ophthalmology. 2006 Apr 1;113(4):623-32.

[233] Phase H. Genentech discloses safety concerns over Avastin. Nat. Biotechnol. 2004;22:371-4.

[234] Hirota M, Murakami I, Ishikawa Y, Suzuki T, Sumida SI, Ibaragi S, Kasai H, Horai N, Drolet DW, Gupta S, Janjic N. Chemically modified interleukin-6 aptamer inhibits development of collagen-induced arthritis in cynomolgus monkeys. Nucleic acid therapeutics. 2016;26(1):10-9.

[235] Yang X, Gorenstein DG. Progress in thioaptamer development. Current Drug Targets. 2004 Nov 1;5(8):705-15.

[236] He W, Elizondo-Riojas MA, Li X, Lokesh GL, Somasunderam A, Thiviyanathan V, Volk DE, Durland RH, Englehardt J, Cavasotto CN, Gorenstein DG. X-aptamers: a bead-based selection method for random incorporation of druglike moieties onto next-generation aptamers for enhanced binding. Biochemistry. 2012 Oct 23;51(42):8321-3.

[237] Zhuo Z, Yu Y, Wang M, Li J, Zhang Z, Liu J, Wu X, Lu A, Zhang G, Zhang B. Recent advances in SELEX technology and aptamer applications in biomedicine. International journal of molecular sciences. 2017 Oct 14;18(10):2142.

[238] Zhu C, Feng Z, Qin H, Chen L, Yan M, Li L, Qu F. Recent progress of SELEX methods for screening nucleic acid aptamers. Talanta. 2024 Jan 1;266:124998.

[239] Dong L, Tan Q, Ye W, Liu D, Chen H, Hu H, Wen D, Liu Y, Cao Y, Kang J, Fan J. Screening and identifying a novel ssDNA aptamer against alpha-fetoprotein using CE-SELEX. Scientific reports. 2015 Oct 26;5(1):15552.

[240] Hybarger G, Bynum J, Williams RF, Valdes JJ, Chambers JP. A microfluidic SELEX prototype. Analytical and bioanalytical chemistry. 2006 Jan;384:191-8.

[241] Lou X, Qian J, Xiao Y, Viel L, Gerdon AE, Lagally ET, Atzberger P, Tarasow TM, Heeger AJ, Soh HT. Micromagnetic selection of aptamers in microfluidic channels. Proceedings of the National Academy of Sciences. 2009 Mar 3;106(9):2989-94.

[242] Bae H, Ren S, Kang J, Kim M, Jiang Y, Jin MM, Min IM, Kim S. Sol-gel SELEX circumventing chemical conjugation of low molecular weight metabolites discovers aptamers selective to xanthine. nucleic acid therapeutics. 2013 Dec 1;23(6):443-9.

[243] Mencin N, Šmuc T, Vraničar M, Mavri J, Hren M, Galeša K, Krkoč P, Ulrich H, Šolar B. Optimization of SELEX: comparison of different methods for monitoring the progress of in vitro selection of aptamers. Journal of pharmaceutical and biomedical analysis. 2014 Mar 25;91:151-9.

[244] Uemachi H, Kasahara Y, Tanaka K, Okuda T, Yoneda Y, Obika S. Hybrid-Type SELEX for the selection of artificial nucleic acid aptamers exhibiting cell internalization activity. Pharmaceutics. 2021 Jun 15;13(6):888.

[245] Ishihama A, Shimada T, Yamazaki Y. Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. Nucleic acids research. 2016 Mar 18;44(5):2058-74.

[246] Ng EW, Shima DT, Calias P, Cunningham Jr ET, Guyer DR, Adamis AP. Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease. Nature reviews drug discovery. 2006 Feb 1;5(2):123-32.

[247] Li W, Zhang S, Zhao M, Lan X. Aptamers for Thrombotic Diseases. Aptamers for Medical Applications: From Diagnosis to Therapeutics. 2021:279-318.

[248] Troisi R, Napolitano V, Spiridonova V, Russo Krauss I, Sica F. Several structural motifs cooperate in determining the highly effective anti-thrombin activity of NU172 aptamer. Nucleic Acids Research. 2018 Dec 14;46(22):12177-85.

[249] Yazdian-Robati R, Bayat P, Oroojalian F, Zargari M, Ramezani M, Taghdisi SM, Abnous K. Therapeutic applications of AS1411 aptamer, an update review. International journal of biological macromolecules. 2020 Jul 15;155:1420-31.

[250] Lupold SE. Aptamers and apple pies: a mini-review of PSMA aptamers and lessons from Donald S. Coffey. American Journal of Clinical and Experimental Urology. 2018;6(2):78.

[251] Gijs M, Penner G, Blackler GB, Impens NR, Baatout S, Luxen A, Aerts AM. Improved aptamers for the diagnosis and potential treatment of HER2-positive cancer. Pharmaceuticals. 2016 May 19;9(2):29.

[252] Hwang JA, Hur JY, Kim Y, Im JH, Jin SH, Ryu SH, Choi CM. Efficacy of newly discovered DNA aptamers targeting AXL in a lung cancer cell with acquired resistance to Erlotinib. Translational Cancer Research. 2021 Feb;10(2):1025.

[253] Ferreira CS, Matthews CS, Missailidis S. DNA aptamers that bind to MUC1 tumour marker: design and characterization of MUC1-binding single-stranded DNA aptamers. Tumor biology. 2006 Oct 27;27(6):289-301.

[254] de Melo MI, Correa CR, da Silva Cunha P, de Góes AM, Gomes DA, de Andrade AS. DNA aptamers selection for carcinoembryonic antigen (CEA). Bioorganic & medicinal chemistry letters. 2020 Aug 1;30(15):127278.

[255] Sicco E, Mónaco A, Fernandez M, Moreno M, Calzada V, Cerecetto H. Metastatic and non-metastatic melanoma imaging using Sgc8-c aptamer PTK7-recognizer. Scientific Reports. 2021 Oct 7;11(1):19942.

[256] Huang BT, Lai WY, Chang YC, Wang JW, Yeh SD, Lin EP, Yang PC. A CTLA-4

antagonizing DNA aptamer with antitumor effect. Molecular Therapy-Nucleic Acids. 2017 Sep 15;8:520-8.

[257] Mascarelli DE, Rosa RS, Toscaro JM, Semionatto IF, Ruas LP, Fogagnolo CT, Lima GC, Bajgelman MC. Boosting antitumor response by costimulatory strategies driven to 4-1BB and OX40 T-cell receptors. Frontiers in cell and developmental biology. 2021 Jun 30;9:692982.

[258] White RR, Roy JA, Viles KD, Sullenger BA, Kontos CD. A nuclease-resistant RNA aptamer specifically inhibits angiopoietin-1-mediated Tie2 activation and function. Angiogenesis. 2008 Dec;11:395-401.

[259] Nicolotti O, Benfenati E, Carotti A, Gadaleta D, Gissi A, Mangiatordi GF, Novellino E. REACH and in silico methods: an attractive opportunity for medicinal chemists. Drug Discovery Today. 2014 Nov 1;19(11):1757-68.

[260] Wegner JK, Sterling A, Guha R, Bender A, Faulon JL, Hastings J, O'Boyle N, Overington J, Van Vlijmen H, Willighagen E. Cheminformatics. Communications of the ACM. 2012 Nov 1;55(11):65-75.

[261] Faber C, Attaccalite C, Olevano V, Runge E, Blase X. First-principles GW calculations for DNA and RNA nucleobases. Physical Review B—Condensed Matter and Materials Physics. 2011 Mar 15;83(11):115123.

[262] Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC bioinformatics. 2010 Dec;11:1-9.

[263] El-Shafei M. Computational methods in Bioinformatics: Introduction, Review, and Challenges.

[264] Varani G, McClain WH. The G· U wobble base pair. EMBO reports. 2000, 1;1(1):18-23.

[265] Andreas Gruber, Strategies for computational noncoding RNA detection. [Doctoral thesis,University of Vienna] u:theses Server . 2010; https://www.tbi.univie.ac.at/newpapers/pdfs/TBI-t-2011-1.pdf

[266] Michael Wolfinger, The Energy Landscape of RNA Folding. [Maters thesis,University of Vienna] u:theses Server . 2010, https://www.researchgate.net/publication/227594207_Energy_Landscapes_of_Biopoly

mers

[267] Waterman MS, Smith TF. RNA secondary structure: A complete mathematical analysis. Mathematical Biosciences. 1978 Dec 1;42(3-4):257-66.

[268] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms for molecular biology. 2011 Dec;6:1-4.

[269] Hofacker IL. Vienna RNA secondary structure server. Nucleic acids research. 2003 Jul 1;31(13):3429-31.

[270] Nussinov R, Shapiro B, Le SY, Maizel Jr JV. Speeding up the dynamic algorithm for planar RNA folding. Mathematical biosciences. 1990 Jun 1;100(1):33-47.

[271] Hofacker IL, Stadler PF. Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics. 2006 May 15;22(10):1172-6.

[272] Bommarito S, Peyret N, Jr JS. Thermodynamic parameters for DNA sequences with dangling ends. Nucleic acids research. 2000 May 1;28(9):1929-34.

[273] Amman F, Bernhart SH, Doose G, Hofacker IL, Qin J, Stadler PF, Will S. The trouble with long-range base pairs in RNA folding. InAdvances in Bioinformatics and Computational Biology: 8th Brazilian Symposium on Bioinformatics, BSB 2013, Recife, Brazil, November 3-7, 2013, Proceedings 8 2013 (pp. 1-11). Springer International Publishing.

[274] Mathews DH, Moss WN, Turner DH. Folding and finding RNA secondary structure. Cold Spring Harbor perspectives in biology. 2010 Dec 1;2(12):a003665.

[275] Zuker M. On finding all suboptimal foldings of an RNA molecule. Science. 1989 Apr 7;244(4900):48-52.

[276] Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers: Original Research on Biomolecules. 1999 Feb;49(2):145-65.

[277] Lorenz R, Hofacker IL, Stadler PF. RNA folding with hard and soft constraints. Algorithms for Molecular Biology. 2016 Dec;11:1-3.

[278] McCaskill JS. The equilibrium partition function and base pair binding probabilities for

RNA secondary structure. Biopolymers: Original Research on Biomolecules. 1990 May;29(6-7):1105-19.

[279] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms for molecular biology. 2011 Dec;6:1-4.

[280] Flamm C, Fontana W, Hofacker IL, Schuster P. RNA folding at elementary step resolution. Rna. 2000 Mar;6(3):325-38.

[281] Zhang J, Fei Y, Sun L, Zhang QC. Advances and opportunities in RNA structure experimental determination and computational modeling. Nature methods. 2022 Oct;19(10):1193-207.

[282] Parsons RJ, Forrest S, Burks C. Genetic algorithms, operators, and DNA fragment assembly. Machine Learning. 1995 Oct;21:11-33.

[283] Flores SC, Wan Y, Russell R, Altman RB. Predicting RNA structure by multiple template homology modeling. InBiocomputing 2010 2010 (pp. 216-227).

[284] Rother M, Rother K, Puton T, Bujnicki JM. ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic acids research. 2011 May 1;39(10):4007-22.

[285] Alam T, Uludag M, Essack M, Salhi A, Ashoor H, Hanks JB, Kapfer C, Mineta K, Gojobori T, Bajic VB. FARNA: knowledgebase of inferred functions of non-coding RNA transcripts. Nucleic acids research. 2017 Mar 17;45(5):2838-48.

[286] Das R, Karanicolas J, Baker D. Atomic accuracy in predicting and designing noncanonical RNA structure. Nature methods. 2010 Apr;7(4):291-4.

[287] Sripakdeevong P, Kladwang W, Das R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. Proceedings of the National Academy of Sciences. 2011 Dec 20;108(51):20573-8.

[288] Yang Y, Liu Z, Lv X, Li D, Chen X, Li X. RNA Tertiary Structure Prediction Algorithm at Atomic Accuracy. In2021 17th International Conference on Computational Intelligence and Security (CIS) 2021 Nov 19 (pp. 55-59). IEEE.

[289] Chen L, Li Q, Nasif KF, Xie Y, Deng B, Niu S, Pouriyeh S, Dai Z, Chen J, Xie CY. AI-Driven Deep Learning Techniques in Protein Structure Prediction. International journal

of molecular sciences. 2024 Aug 1;25(15):8426.

[290] Baek M, McHugh R, Anishchenko I, Jiang H, Baker D, DiMaio F. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. Nature methods. 2024 Jan;21(1):117-21.

[291] Bernard C, Postic G, Ghannay S, Tahi F. State-of-the-RNArt: benchmarking current methods for RNA 3D structure prediction. NAR Genomics and Bioinformatics. 2024 Jun 1;6(2):lqae048.

[292] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. Nature communications. 2022 Mar 10;13(1):1265.

[293] Wang W, Feng C, Han R, Wang Z, Ye L, Du Z, Wei H, Zhang F, Peng Z, Yang J. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. Nature Communications. 2023 Nov 9;14(1):7266.

[294] Moal IH, Moretti R, Baker D, Fernández-Recio J. Scoring functions for protein–protein interactions. Current opinion in structural biology. 2013 Dec 1;23(6):862-7.

[295] Janin J. Protein–protein docking tested in blind predictions: the CAPRI experiment. Molecular BioSystems. 2010;6(12):2351-62.

[296] Muratcioglu S, Guven-Maiorov E, Keskin Ö, Gursoy A. Advances in template-based protein docking by utilizing interfaces towards completing structural interactome. Current opinion in structural biology. 2015 Dec 1;35:87-92.

[297] Szilagyi A, Zhang Y. Template-based structure modeling of protein–protein interactions. Current opinion in structural biology. 2014 Feb 1;24:10-23.

[298] Zheng J, Hong X, Xie J, Tong X, Liu S. P3DOCK: a protein–RNA docking webserver based on template-based and template-free docking. Bioinformatics. 2020 Jan 1;36(1):96-103.

[299] Yan Y, Wen Z, Wang X, Huang SY. Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking. Proteins: Structure, Function, and Bioinformatics. 2017 Mar;85(3):497-512.

[300] Yan Y, Tao H, He J, Huang SY. The HDOCK server for integrated protein–protein docking. Nature protocols. 2020 May;15(5):1829-52.

[301] Yan Y, Tao H, He J, Huang SY. The HDOCK server for integrated protein–protein docking. Nature protocols. 2020 May;15(5):1829-52.

[302] Yan Y, Huang SY. RRDB: a comprehensive and non-redundant benchmark for RNA–RNA docking and scoring. Bioinformatics. 2018 Feb 1;34(3):453-8.

[303] Rincón L. Minimum-energy configurations of metallic clusters obtained by simulated annealing molecular dynamics using an Extended-Huckel Hamiltonian. Revista Mexicana de Física. 2001 Mar 1;47:54-8.

[304] Sutmann G. Classical molecular dynamics. Quantum simulations of complex many-body systems: from theory to algorithms. 2002;10:211-54.

[305] Bayo E, de Jalon JG, Avello A, Cuadrado J. An efficient computational method for real time multibody dynamic simulation in fully Cartesian coordinates. Computer Methods in Applied Mechanics and Engineering. 1991 Nov 1;92(3):377-95.

[306] Vaidehi N, Jain A. Internal coordinate molecular dynamics: A foundation for multiscale dynamics. The Journal of Physical Chemistry B. 2015 Jan 29;119(4):1233-42.

[307] Hoover WG. Nonequilibrium molecular dynamics. Molecular Dynamics. 1986:92-131.

[308] Travis KP, Braga C. Configurational temperature and pressure molecular dynamics: Review of current methodology and applications to the shear flow of a simple fluid. Molecular Physics. 2006 Nov 20;104(22-24):3735-49.

[309] Rutenbar RA. Simulated annealing algorithms: An overview. IEEE Circuits and Devices magazine. 1989 Jan;5(1):19-26.

[310] Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chemical physics letters. 1999 Nov 26;314(1-2):141-51.

[311] Abraham, M., Alekseenko, A., Bergh, C., Blau, C., Briand, E., Doijade, M., Fleischmann, S., Gapsys, V., Garg, G., Gorelov, S., Gouaillardet, G., Gray, A., Irrgang, M. E., Jalalypour, F., Jordan, J., Junghans, C., Kanduri, P., Keller, S., Kutzner, C., Lemkul, J. A., Lundborg, M., Merz, P., Miletić, V., Morozov, D., Páll, Sz., Schulz, R., Shirts, M., Shvetsov, A., Soproni, B., van der Spoel, D., Turner, P., Uphoff, C., Villa, A., Wingbermühle, S., Zhmurov, A., & Hess, B. (2023). GROMACS 2023.3 Manual. https://zenodo.org/records/10017699

[312] Leimkuhler B, Noorizadeh E, Theil F. A gentle stochastic thermostat for molecular dynamics. Journal of Statistical Physics. 2009 Apr;135:261-77.

[313] Janezic D, Merzel F. An efficient symplectic integration algorithm for molecular dynamics simulations. Journal of chemical information and computer sciences. 1995 Mar 1;35(2):321-6.

[314] Van Gunsteren WF, Berendsen HJ. Algorithms for Brownian dynamics. Molecular Physics. 1982 Feb 20;45(3):637-47.

[315] Izaguirre JA, Catarello DP, Wozniak JM, Skeel RD. Langevin stabilization of molecular dynamics. The Journal of chemical physics. 2001 Feb 1;114(5):2090-8.

[316] Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Computer physics communications. 1995 Sep 2;91(1-3):215-31.

[317] Dang LX. Importance of polarization effects in modeling the hydrogen bond in water using classical molecular dynamics techniques. The Journal of Physical Chemistry B. 1998 Jan 15;102(3):620-4.

[318] González MA. Force fields and molecular dynamics simulations. École thématique de la Société Française de la Neutronique. 2011;12:169-200.

[319] Visscher KM, Geerke DP. Deriving force-field parameters from first principles using a polarizable and higher order dispersion model. Journal of chemical theory and computation. 2019 Feb 14;15(3):1875-83.

[320] Ponder JW, Case DA. Force fields for protein simulations. Advances in protein chemistry. 2003 Jan 1;66:27-85.

[321] Weiner SJ, Kollman PA, Nguyen DT, Case DA. An all atom force field for simulations of proteins and nucleic acids. Journal of computational chemistry. 1986 Apr;7(2):230-52.

[322] Jayaram B, Sprous D, Beveridge DL. Solvation free energy of biomacromolecules: Parameters for a modified generalized Born model consistent with the AMBER force field. The Journal of Physical Chemistry B. 1998 Nov 19;102(47):9571-6.

[323] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC,

Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society. 1995 May;117(19):5179-97.

[324] Wang J, Cieplak P, Li J, Wang J, Cai Q, Hsieh M, Lei H, Luo R, Duan Y. Development of polarizable models for molecular mechanical calculations II: induced dipole models significantly improve accuracy of intermolecular interaction energies. The journal of physical chemistry B. 2011 Mar 31;115(12):3100-11.

[325] Krepl M, Zgarbová M, Stadlbauer P, Otyepka M, Banas P, Koca J, Cheatham III TE, Jurecka P, Sponer J. Reference simulations of noncanonical nucleic acids with different χ variants of the AMBER force field: quadruplex DNA, quadruplex RNA, and Z-DNA. Journal of chemical theory and computation. 2012 Jul 10;8(7):2506-20.

[326] Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. Biophysical journal. 2007 Jun 1;92(11):3817-29.

[327] Aytenfisu AH, Spasic A, Grossfield A, Stern HA, Mathews DH. Revised RNA dihedral parameters for the amber force field improve RNA molecular dynamics. Journal of chemical theory and computation. 2017 Feb 14;13(2):900-15.

[328] Tan D, Piana S, Dirks RM, Shaw DE. RNA force field with accuracy comparable to state-of-the-art protein force fields. Proceedings of the National Academy of Sciences. 2018 Feb 13;115(7):E1346-55.

[329] Hou T, Wang J, Li Y, Wang W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. Journal of chemical information and modeling. 2011 Jan 24;51(1):69-82.

[330] Rastelli G, Rio AD, Degliesposti G, Sgobba M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. Journal of computational chemistry. 2010 Mar;31(4):797-810.

[331] El Khoury L, Santos-Martins D, Sasmal S, Eberhardt J, Bianco G, Ambrosio FA, Solis-Vasquez L, Koch A, Forli S, Mobley DL. Comparison of affinity ranking using AutoDock-GPU and MM-GBSA scores for BACE-1 inhibitors in the D3R Grand

Challenge 4. Journal of computer-aided molecular design. 2019 Dec;33(12):1011-20.

[332] Wang E, Fu W, Jiang D, Sun H, Wang J, Zhang X, Weng G, Liu H, Tao P, Hou T. VAD-MM/GBSA: a variable atomic dielectric MM/GBSA model for improved accuracy in protein–ligand binding free energy calculations. Journal of Chemical Information and Modeling. 2021 May 20;61(6):2844-56.

[333] Wang C, Greene DA, Xiao L, Qi R, Luo R. Recent developments and applications of the MMPBSA method. Frontiers in molecular biosciences. 2018 Jan 10;4:87.

[334] Yang L, Bailey L, Baltimore D, Wang P. Targeting lentiviral vectors to specific cell types in vivo. Proceedings of the National Academy of Sciences. 2006 Aug 1;103(31):11479-84.

[335] Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. Bioinformatics. 2014 Jun 15;30(12):1771-3.

[336] Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. Proteins: Structure, Function, and Bioinformatics. 2007 Aug 1;68(2):503-15.

[337] Biesiada M, Purzycka KJ, Szachniuk M, Blazewicz J, Adamiak RW. Automated RNA 3D structure prediction with RNAComposer. RNA Structure Determination: Methods and Protocols. 2016:199-215.

[338] Cruz-Toledo J, McKeague M, Zhang X, Giamberardino A, McConnell E, Francis T, DeRosa MC, Dumontier M. Aptamer base: a collaborative knowledge base to describe aptamers and SELEX experiments. Database. 2012 Jan 1;2012:bas006.

[339] Perry GH, Martin RD, Verrelli BC. Signatures of functional constraint at aye-aye opsin genes: the potential of adaptive color vision in a nocturnal primate. Molecular Biology and Evolution. 2007 Sep 1;24(9):1963-70.

[340] Hull TE, Dobell AR. Random number generators. SIAM review. 1962 Jul;4(3):230-54.

[341] Zuber J, Schroeder SJ, Sun H, Turner DH, Mathews DH. Nearest neighbor rules for RNA helix folding thermodynamics: improved end effects. Nucleic Acids Research. 2022 May 20;50(9):5251-62.

[342] Osborne SE, Ellington AD. Nucleic acid selection and the challenge of combinatorial chemistry. Chemical reviews. 1997;97(2):349-70.

[343] Bose P, Hermetz KE, Conneely KN, Rudd MK. Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. PLoS One. 2014;9(7):e101607.

[344] Zuker M. On finding all suboptimal foldings of an RNA molecule. Science. 1989 Apr 7;244(4900):48-52.

[345] Zuker M, Mathews DH, Turner DH. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. RNA biochemistry and biotechnology. 1999:11-43.

[346] Kiga D, Futamura Y, Sakamoto K, Yokoyama S. An RNA aptamer to the xanthine/guanine base with a distinctive mode of purine recognition. Nucleic Acids Research. 1998 Apr 1;26(7):1755-60.

[347] Noeske J, Richter C, Grundl MA, Nasiri HR, Schwalbe H, Wöhnert J. An intermolecular base triple as the basis of ligand specificity and affinity in the guanine- and adenine-sensing riboswitch RNAs. Proceedings of the National Academy of Sciences. 2005 Feb 1;102(5):1372-7.

[348] Minchin S, Busby S. Location of close contacts between Escherichia coli RNA polymerase and guanine residues at promoters either with or without consensus-35 region sequences. Biochemical Journal. 1993 ;289(3):771-5.

[349] Kankia B. Which came first: the chicken, the egg, or guanine?. RNA. 2023 Sep 1;29(9):1317-24.

[350] Varani G, McClain WH. The G· U wobble base pair. EMBO reports. 2000 Jul 1;1(1):18-23.

[351] Schwinefus JJ, McDevitt J. Urea Destabilization of DNA and RNA Double Helices: Preferential Interactions with Nucleobase Conjugated Pi-Pi-Systems. Biophysical Journal. 2010 Jan 1;98(3):41a.

[352] Martinez CR, Iverson BL. Rethinking the term “pi-stacking”. Chemical Science. 2012;3(7):2191-201.

[353] Eisenberg E, Levanon EY. A-to-I RNA editing—immune protector and transcriptome

diversifier. Nature Reviews Genetics. 2018 Aug;19(8):473-90.

[354] Soares JM, Rocha AJ, Nascimento FS, Santos AS, Miller RN, Ferreira CF, Haddad F, Amorim VB, Amorim EP. Genetic improvement for resistance to black Sigatoka in bananas: A systematic review. Frontiers in plant science. 2021 Apr 21;12:657916.

[355] Wellner JA. Limit theorems for the ratio of the empirical distribution function to the true distribution function. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete. 1978 Mar;45(1):73-88.

[356] Burdick RK, Borror CM, Montgomery DC. A review of methods for measurement systems capability analysis. Journal of Quality Technology. 2003 Oct 1;35(4):342-54.

[357] Ibrahim N, Abdullahi AB. Analysis of Variance (ANOVA) Randomized Block Design (RBD) to Test the Variability of Three Different Types of Fertilizers (NPK, UREA and SSP) on Millet Production. African Journal of Agricultural Science and Food Research. 2023 Feb 2;9(1):1-0.

[358] Kim TK. Understanding one-way ANOVA using conceptual figures. Korean journal of anesthesiology. 2017 Feb 1;70(1):22-6.

[359] Heiberger RM, Neuwirth E, Heiberger RM, Neuwirth E. One-way anova. R through Excel: A spreadsheet interface for statistics, data analysis, and graphics. 2009:165-91.

[360] Fujikoshi Y, Ohmae M, Yanagihara H. Asymptotic approximations of the null distribution of the one-way ANOVA test statistic under nonnormality. Journal of the Japan Statistical Society. 1999;29(2):147-61.

[361] Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015 Oct 15;31(20):3377-9.

[362] Halder S, Bhattacharyya D. RNA structure and dynamics: a base pairing perspective. Progress in Biophysics and Molecular Biology. 2013 Nov 1;113(2):264-83.

[363] Lee JC, Gutell RR. Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. Journal of molecular biology. 2004 Dec 10;344(5):1225-49.

[364] Nowakowski J, Tinoco Jr I. RNA structure and stability. InSeminars in virology 1997

Jan 1 (Vol. 8, No. 3, pp. 153-165). Academic Press.

[365] Langdon WB, Petke J, Lorenz R. Evolving better RNAfold structure prediction. InGenetic Programming: 21st European Conference, EuroGP 2018, Parma, Italy, April 4-6, 2018, Proceedings 21 2018 (pp. 220-236). Springer International Publishing.

[366] Lunter G, Hein J. A nucleotide substitution model with nearest-neighbour interactions. InISMB/ECCB (Supplement of Bioinformatics) 2004 Aug 4 (pp. 216-223).

[367] Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. BMC bioinformatics. 2004 Dec;5:1-8.

[368] Trotta E. Selective forces and mutational biases drive stop codon usage in the human genome: a comparison with sense codon usage. BMC genomics. 2016; 17:1-2.

[369] George TP, Thomas T. Novel approach to analyzing MFE of noncoding RNA sequences. Genomics Insights. 2016 Jan;9:GEI-S39995.

[370] Norouzi M, Fleet DJ, Salakhutdinov RR. Hamming distance metric learning. Advances in neural information processing systems. 2012;25.

[371] Discovry studio, software, https://www.3ds.com/products/biovia/discovery-studio

[372] Draper DE, Grilley D, Soto AM. Ions and RNA folding. Annu. Rev. Biophys. Biomol. Struct.. 2005 Jun 9;34(1):221-43.

[373] Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS genetics. 2010 Sep 9;6(9):e1001107.

[374] Carter-Fenk K, Liu M, Pujal L, Loipersberger M, Tsanai M, Vernon RM, Forman-Kay JD, Head-Gordon M, Heidar-Zadeh F, Head-Gordon T. The Energetic Origins of Pi–Pi Contacts in Proteins. Journal of the American Chemical Society. 2023 Nov 2;145(45):24836-51.

[375] Li X, Cai Z, Sevilla MD. Investigation of proton transfer within DNA base pair anion and cation radicals by density functional theory (DFT). The Journal of Physical Chemistry B. 2001 Oct 18;105(41):10115-23.

[376] Motwani R, Raghavan P. Randomized algorithms. ACM Computing Surveys (CSUR). 1996 Mar 1;28(1):33-7.

[377] Saito M, Matsumoto M. Variants of Mersenne twister suitable for graphic processors.

ACM Transactions on Mathematical Software (TOMS). 2013 Feb 1;39(2):1-20.

[378] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research. 2003 Jul 1;31(13):3406-15.

[379] Hogrefe RI, Midthune B, Lebedev A. Current challenges in nucleic acid synthesis. Israel Journal of Chemistry. 2013 Jun;53(6-7):326-49.

[380] Trotta E. On the normalization of the minimum free energy of RNAs by sequence length. PloS one. 2014 Nov 18;9(11):e113380.

[381] Morgan SR, Higgs PG. Evidence for kinetic effects in the folding of large RNA molecules. The Journal of chemical physics. 1996 Oct 22;105(16):7152-7.

[382] Grimme S, Antony J, Ehrlich S, Krieg H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. The Journal of chemical physics. 2010 Apr 21;132(15).

[383] Bannwarth C, Ehlert S, Grimme S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. Journal of chemical theory and computation. 2019 Feb 11;15(3):1652-71.

[384] Sargsyan K, Grauffel C, Lim C. How molecular size impacts RMSD applications in molecular dynamics simulations. Journal of chemical theory and computation. 2017 Apr 11;13(4):1518-24.

[385] Badu, S.R., Melnik, R., Paliy, M., Prabhakar, S., Sebetci, A. and Shapiro, B.A., 2014. Modeling of RNA nanotubes using molecular dynamics simulation. European Biophysics Journal, 43, pp.555-564.

[386] Šponer J, Krepl M, Banáš P, Kührová P, Zgarbová M, Jurečka P, Havrila M, Otyepka M. How to understand atomistic molecular dynamics simulations of RNA and protein–RNA complexes?. Wiley Interdisciplinary Reviews: RNA. 2017;8(3):e1405.

[387] Wambui GD, Waititu GA, Wanjoya A. The power of the pruned exact linear time (PELT) test in multiple changepoint detection. American Journal of Theoretical and Applied Statistics. 2015 Nov;4(6):581.

[388] Bora T. Fuzzification of Simpson’s 1/3 Rule and Development of its Computer

Program. Journal of Coastal Life Medicine. 2023 Mar 16;11:1047-53.

[389] Rostamian Delavar M, Kashuri A, De La Sen M. On weighted Simpson's 3 8 rule. Symmetry. 2021 Oct 14;13(10):1933.

[390] Memon K, Shaikh MM, Saleem M, Chandio AW. A New and Efficient Simpson's 1/3-Type Quadrature Rule For Riemann-Stieltjes Integral'. Journal Of Mechanics Of Continua And Mathematical Sciences. 2020;15(11):132-48.

[391] Ramsey FL. Characterization of the partial autocorrelation function. The Annals of Statistics. 1974 Nov 1:1296-301.

[392] Kurita T. Principal component analysis (PCA). Computer vision: a reference guide. 2019:1-4.

[393] Roweis S. EM algorithms for PCA and SPCA. Advances in neural information processing systems. 1997;10.

[394] Banegas-Luna AJ, Cerón-Carrasco JP, Pérez-Sánchez H. A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. Future medicinal chemistry. 2018 Nov 1;10(22):2641-58.

[395] Waszkowycz B, Perkins TD, Sykes RA, Li J. Large-scale virtual screening for discovering leads in the postgenomic era. IBM Systems Journal. 2001;40(2):360-76.

[396] Gentile F, Yaacoub JC, Gleave J, Fernandez M, Ton AT, Ban F, Stern A, Cherkasov A. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. Nature Protocols. 2022 Mar;17(3):672-97.

[397] Coulocheri SA, Pigis DG, Papavassiliou KA, Papavassiliou AG. Hydrogen bonds in protein–DNA complexes: Where geometry meets plasticity. Biochimie. 2007 Nov 1;89(11):1291-303.

[398] Kalra K. A Biocomputational Study of Water-Nucleobase Stacking Contacts in Functional RNAs (Doctoral dissertation).

[399] Garst AD, Edwards AL, Batey RT. Riboswitches: structures and mechanisms. Cold Spring Harbor perspectives in biology. 2011 Jun 1;3(6):a003533.

[400] Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein–RNA interactions: a structural analysis. Nucleic acids research. 2001 Feb 15;29(4):943-54.

[396] Gentile F, Yaacoub JC, Gleave J, Fernandez M, Ton AT, Ban F, Stern A, Cherkasov A. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. Nature Protocols. 2022 Mar;17(3):672-97.

[397] Coulocheri SA, Pigis DG, Papavassiliou KA, Papavassiliou AG. Hydrogen bonds in protein–DNA complexes: Where geometry meets plasticity. Biochimie. 2007 Nov 1;89(11):1291-303.

[398] Kalra K. A Biocomputational Study of Water-Nucleobase Stacking Contacts in Functional RNAs (Doctoral dissertation).

[399] Garst AD, Edwards AL, Batey RT. Riboswitches: structures and mechanisms. Cold Spring Harbor perspectives in biology. 2011 Jun 1;3(6):a003533.

[400] Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein–RNA interactions: a structural analysis. Nucleic acids research. 2001 Feb 15;29(4):943-54.

[401] Lee MR, Tsai J, Baker D, Kollman PA. Molecular dynamics in the endgame of protein structure prediction. Journal of molecular biology. 2001 Oct 19;313(2):417-30.

[402] Sponer J, Bussi G, Krepl M, Banáš P, Bottaro S, Cunha RA, Gil-Ley A, Pinamonti G, Poblete S, Jurecka P, Walter NG. RNA structural dynamics as captured by molecular simulations: a comprehensive overview. Chemical reviews. 2018 Jan 3;118(8):4177-338.

[403] Yamamoto E, Akimoto T, Mitsutake A, Metzler R. Universal relation between instantaneous diffusivity and radius of gyration of proteins in aqueous solution. Physical review letters. 2021 Mar 26;126(12):128101.

[404] Zagrovic B, Jayachandran G, Millett IS, Doniach S, Pande VS. How large is an α-helix? Studies of the radii of gyration of helical peptides by small-angle X-ray scattering and molecular dynamics. Journal of molecular biology. 2005 Oct 21;353(2):232-41.

[405] Stein SA, Loccisano AE, Firestine SM, Evanseck JD. Principal components analysis: a review of its application on molecular dynamics data. Annual Reports in Computational Chemistry. 2006 Jan 1;2:233-61.

[406] Sakurai JJ. Theory of strong interactions. Annals of Physics. 1960 Sep 1;11(1):1-48.

[407] Münch D, Schmeichel B, Silbering AF, Galizia CG. Weaker ligands can dominate an odor blend due to syntopic interactions. Chemical senses. 2013 May 1;38(4):293-304.

[408] Bulusu G, Desiraju GR. Strong and weak hydrogen bonds in protein–ligand recognition. Journal of the Indian Institute of Science. 2020 Jan;100(1):31-41.

[409] Subramani A, Floudas CA. Structure prediction of loops with fixed and flexible stems. The Journal of Physical Chemistry B. 2012 Jun 14;116(23):6670-82.

SUPPLEMENTARY MATERIAL

Note 1: To avoid making this thesis longer than it is already, we have put all big data tables from this project on GitHub (https://github.com/KPMOKGOPA/MSc_supplementary_data/tree/main).

Note 2: The T_SELEX program sources code is also available on GitHub (https://github.com/CMCDD/T_SELEX)
Note 3: The portion of code provided in this document from PySM/QM package is not yet released on GitHub.

Figure S.1: $M_{seqs[]}$dataset individual nucleotide compositions.

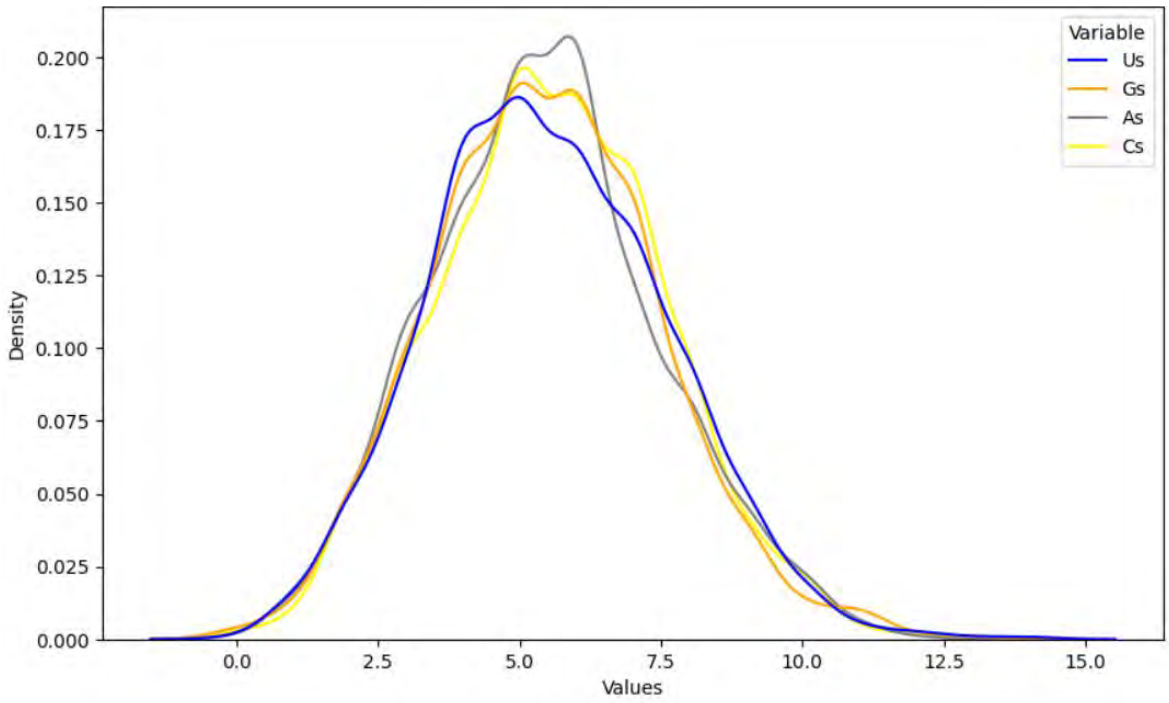


Figure S.2: Distribution line plot of single nucleotide base composition for $M_{seqs[]}$.

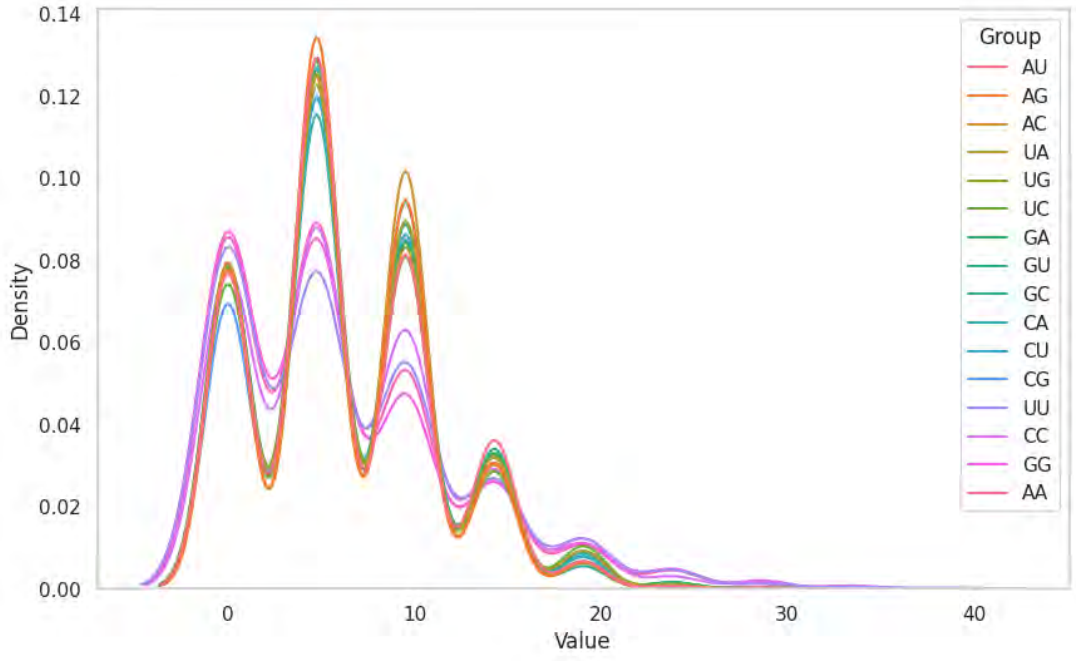


Figure S.3: Distribution line plot of adjacent bases pair composition for thr $M_{seqs[]}$ dataset.

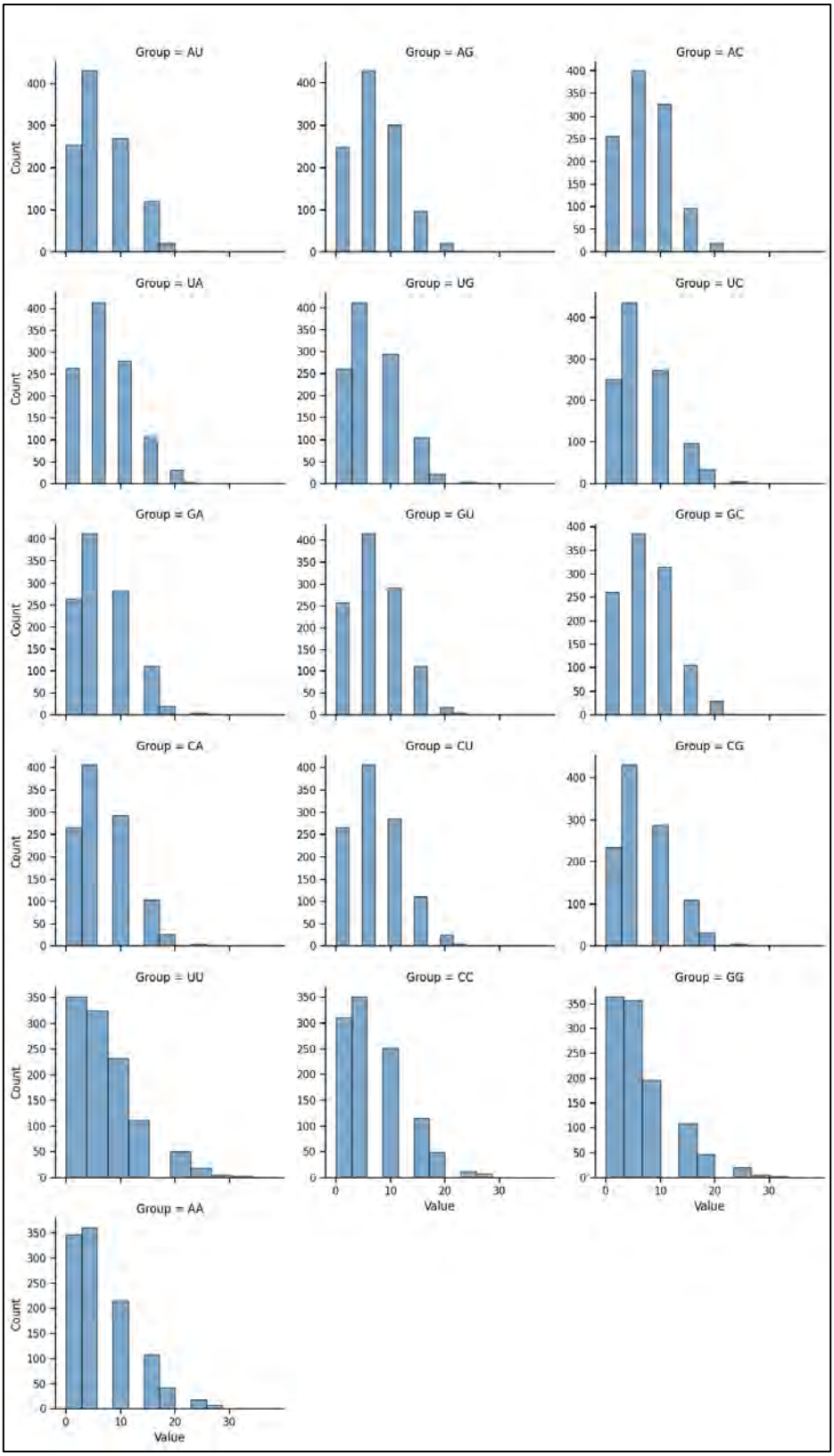


Figure S.4: Count plots of adjacent bases pair composition for aptamer dataset ($M_{seqs[]}$).

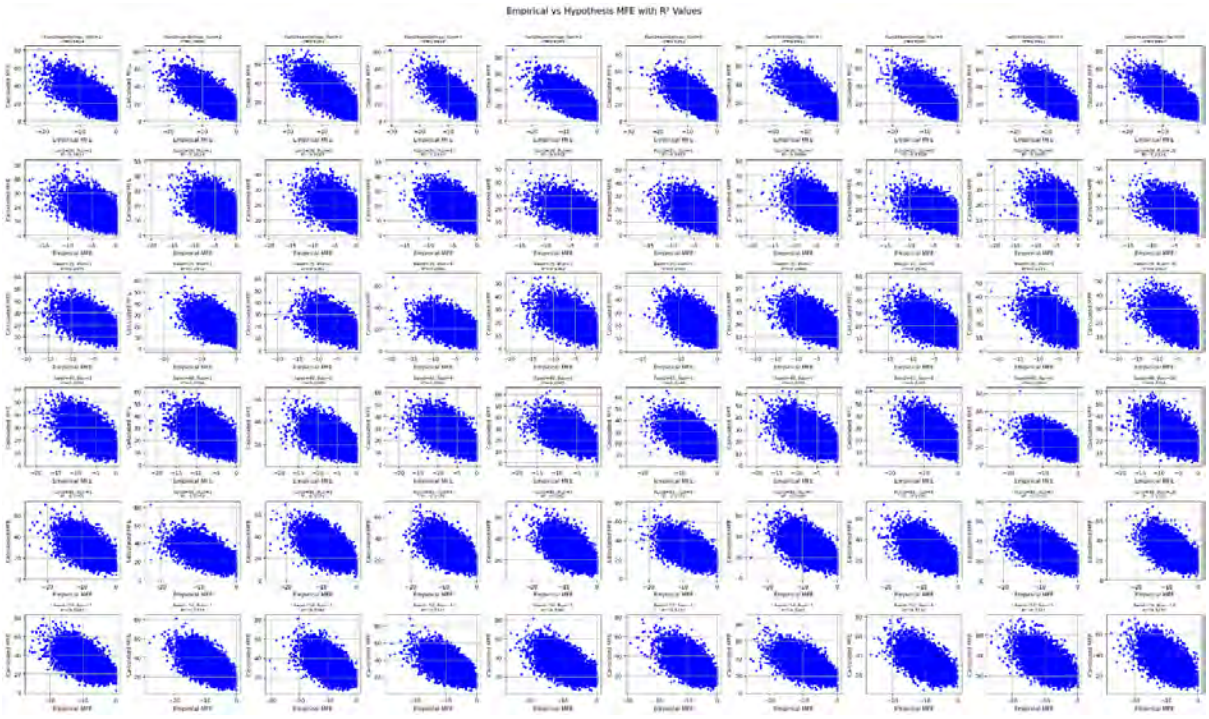


Figure S.5: RNAFold MFE predicted vs the MFE predicted using the $MFE = -(\zeta \int N)$ hypothesis where ∫ is calculated based on the partition function for the 20000 aptamer dataset.

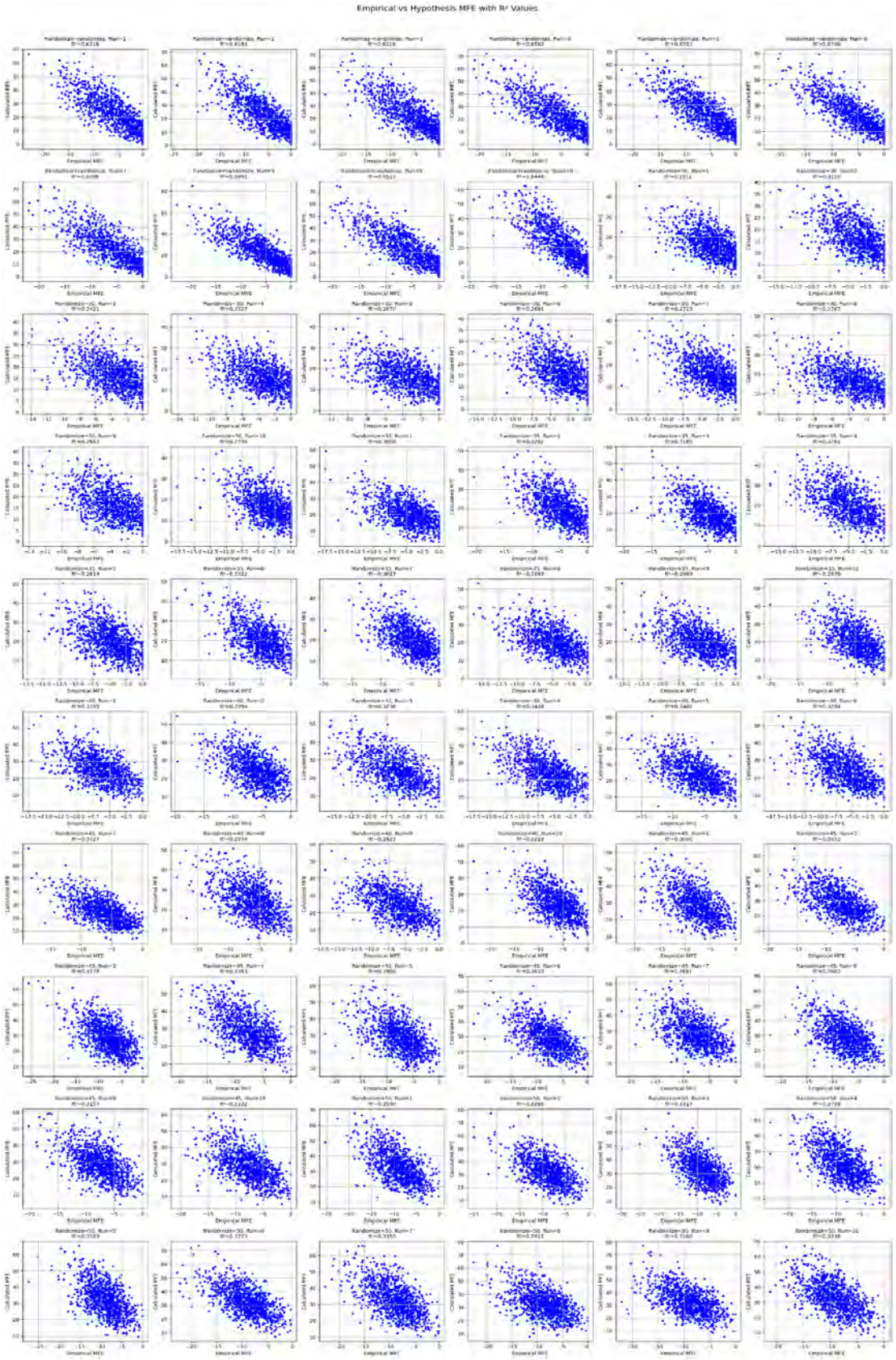


Figure S.6: RNAFold MFE predicted vs the MFE predicted using the $MFE = -(\chi f N)$ hypothesis where f is calculated based on the partition function for the 1000 aptamer dataset.

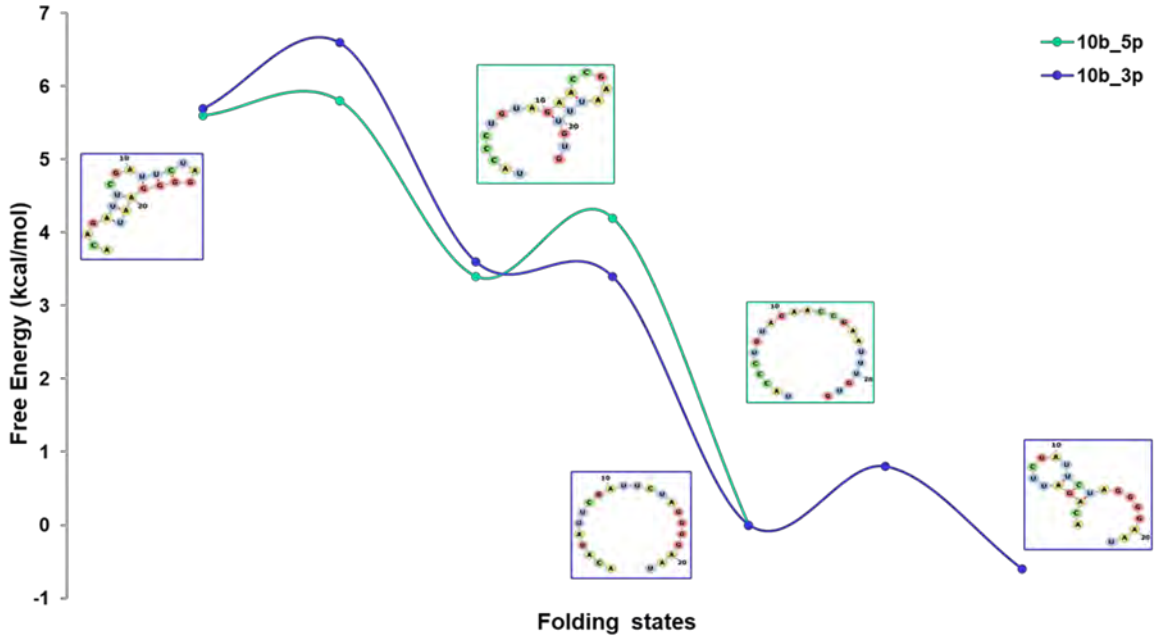


Figure S.7: RNA folding energy landscape for miR-10b-5p and miR-10b-3p calculated using the barrier method.

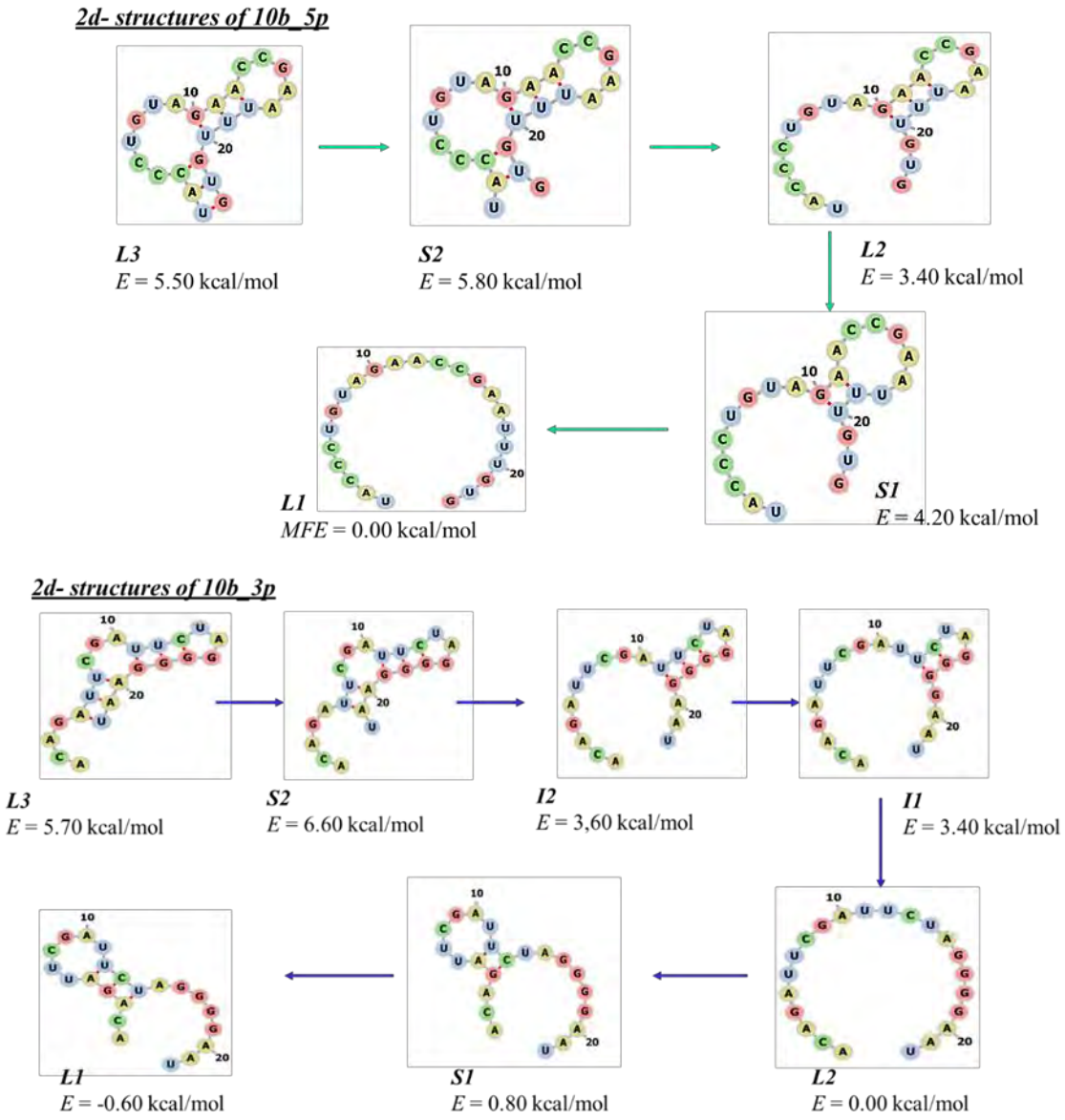


Figure S.8: Minima and maxima folded structure state of miR-10b-5p and miR-10b-3p calculated using the barrier method.

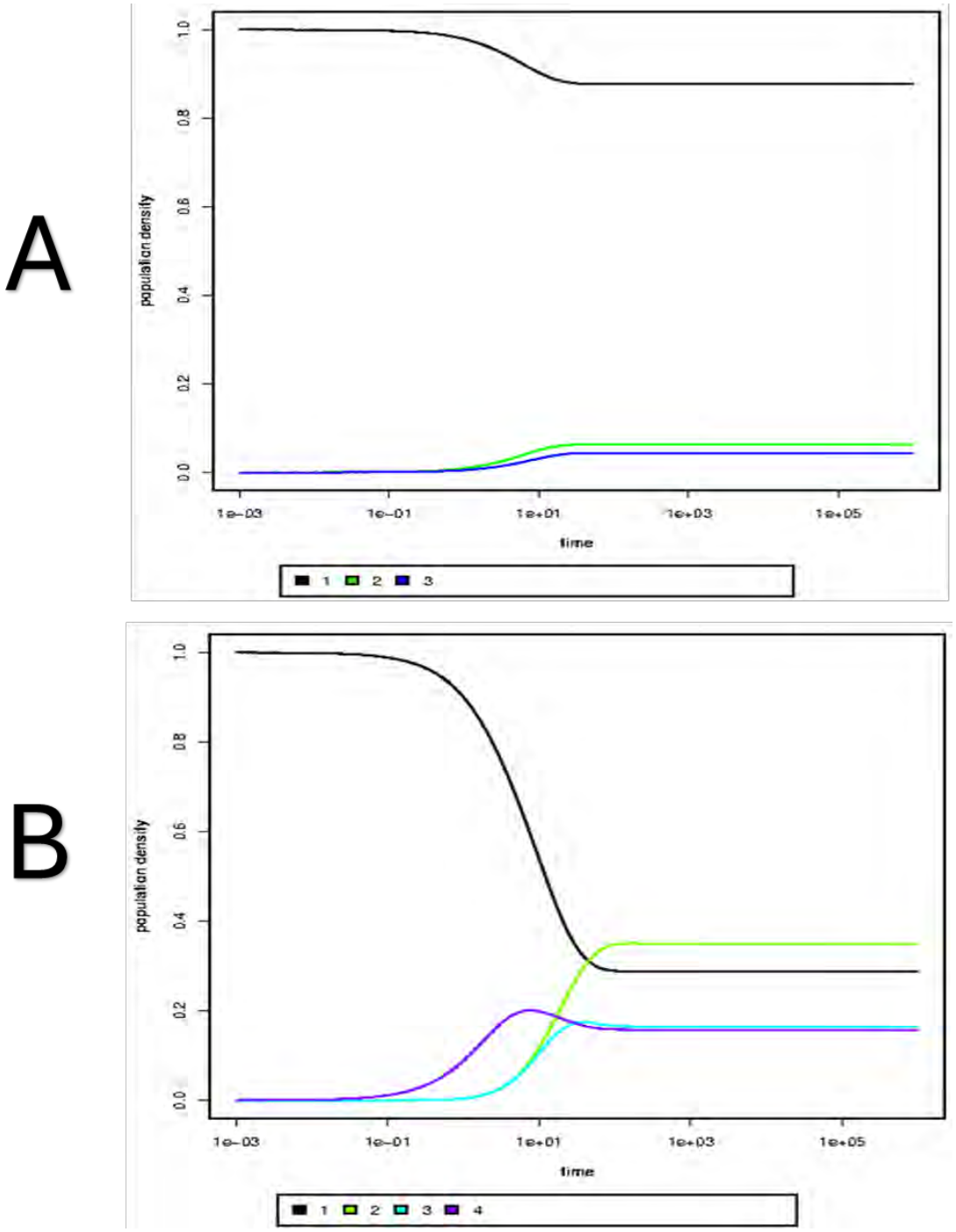


Figure S.9: Population density plots for best-folded minima state of miR-10b-5p (B) and miR-10b-3p (A) calculated using the barrier method.

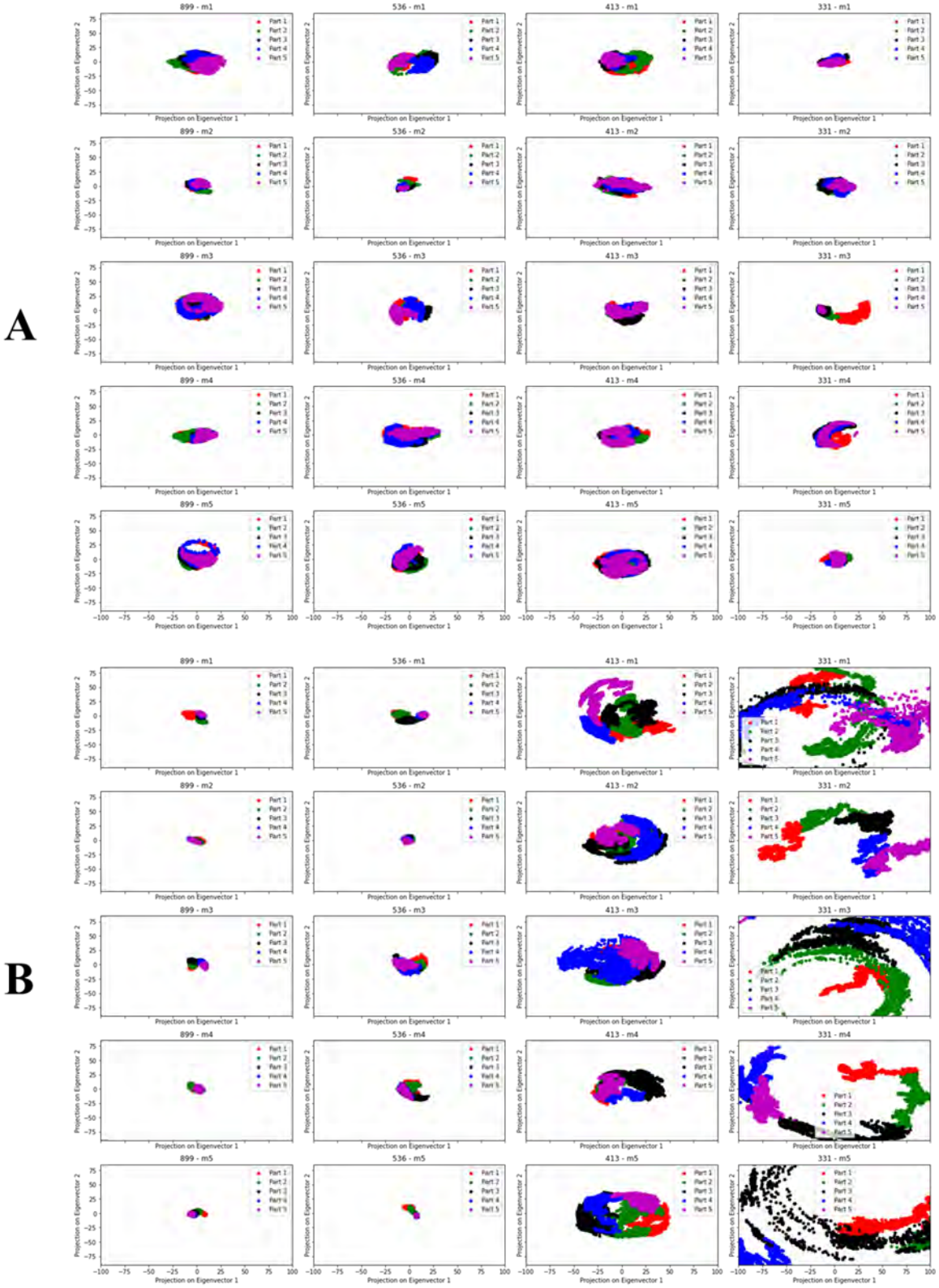


Figure S.10: PCA result for the models of aptamer docked complex against miR-10b-5p for chain A(miR-10b-5p) and chain B (aptamers).

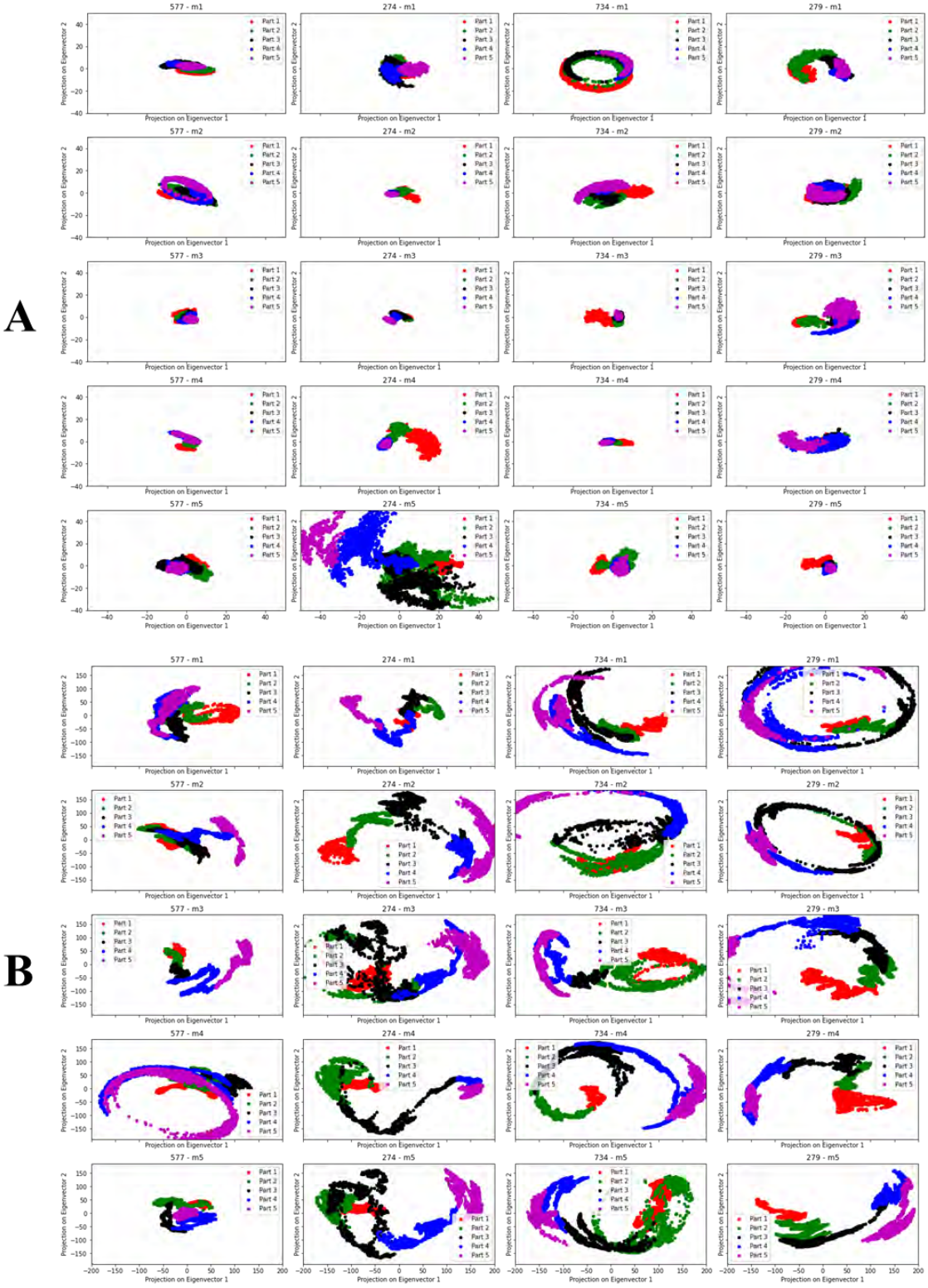


Figure S.11: PCA result for the models of aptamer docked complex against miR-10b-3p for chain A(miR-10b-3p) and chain B (aptamers).

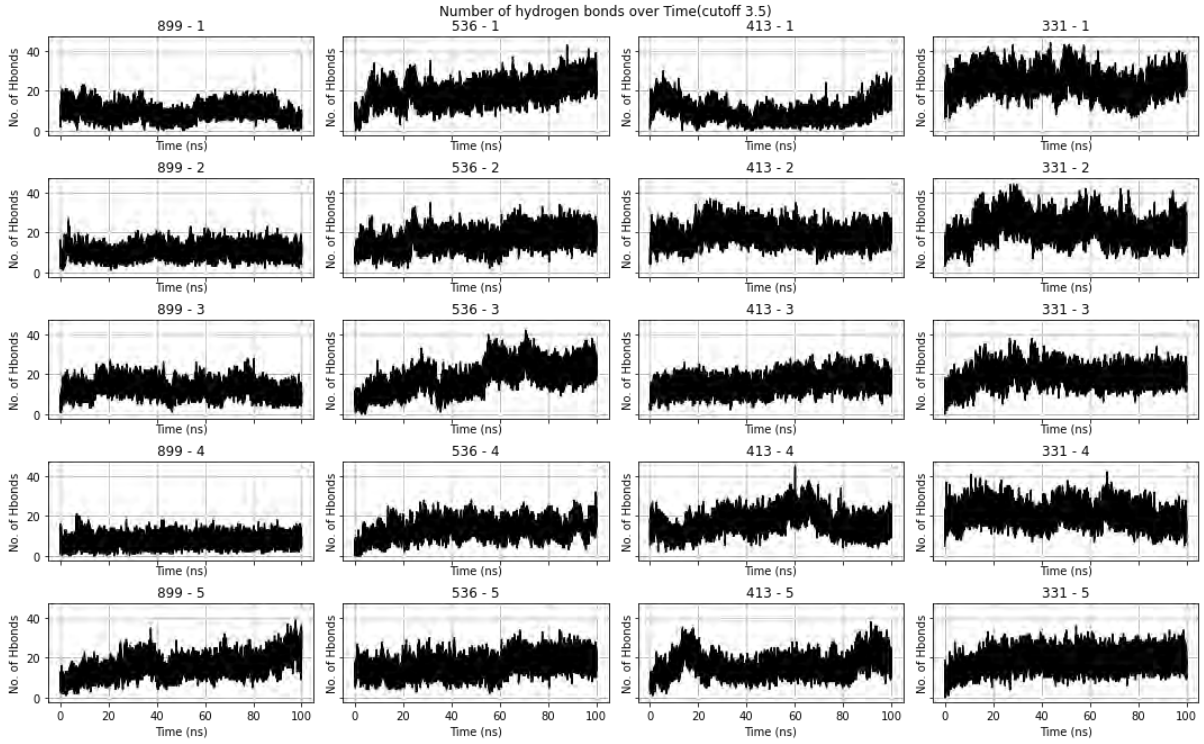


Figure S.12: Hydrogen bonds result for the models of aptamer docked complex against miR-10b-5p target with a cut off distance of 3.5 angstrom.

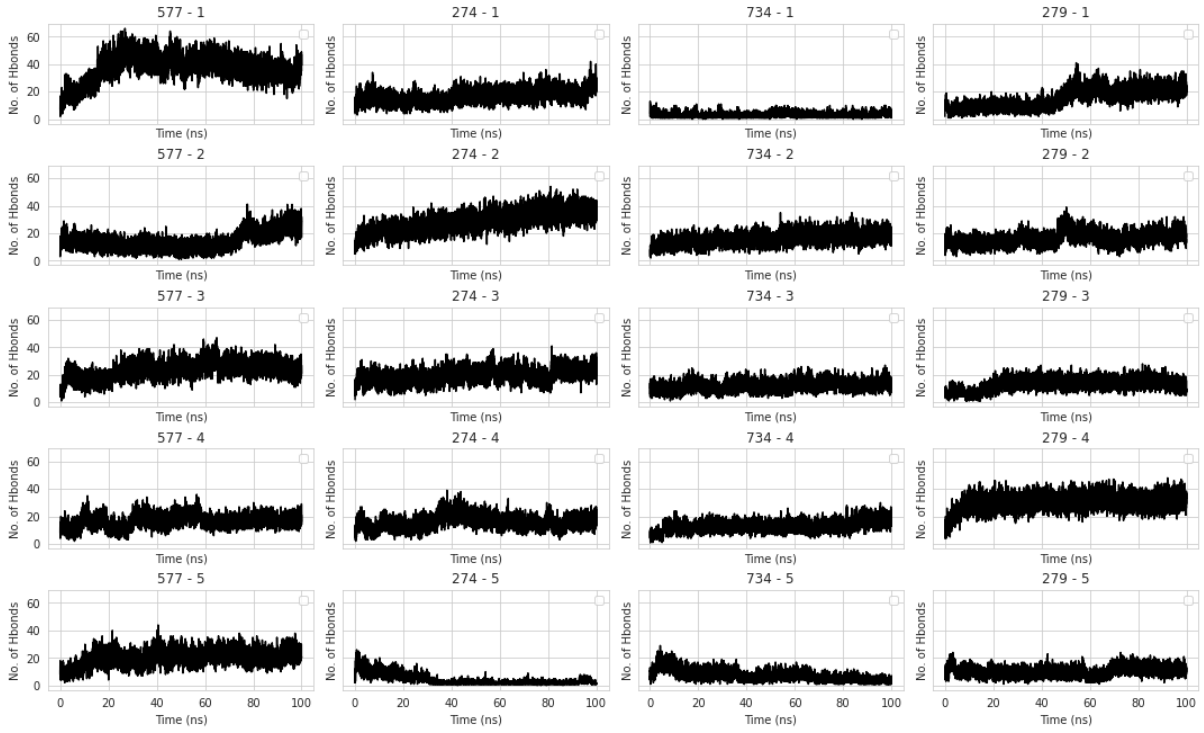


Figure S.13: Hydrogen bonds result for the models of aptamer docked complex against miR-10b-3p target with a cut off of 3.5 angstrom.

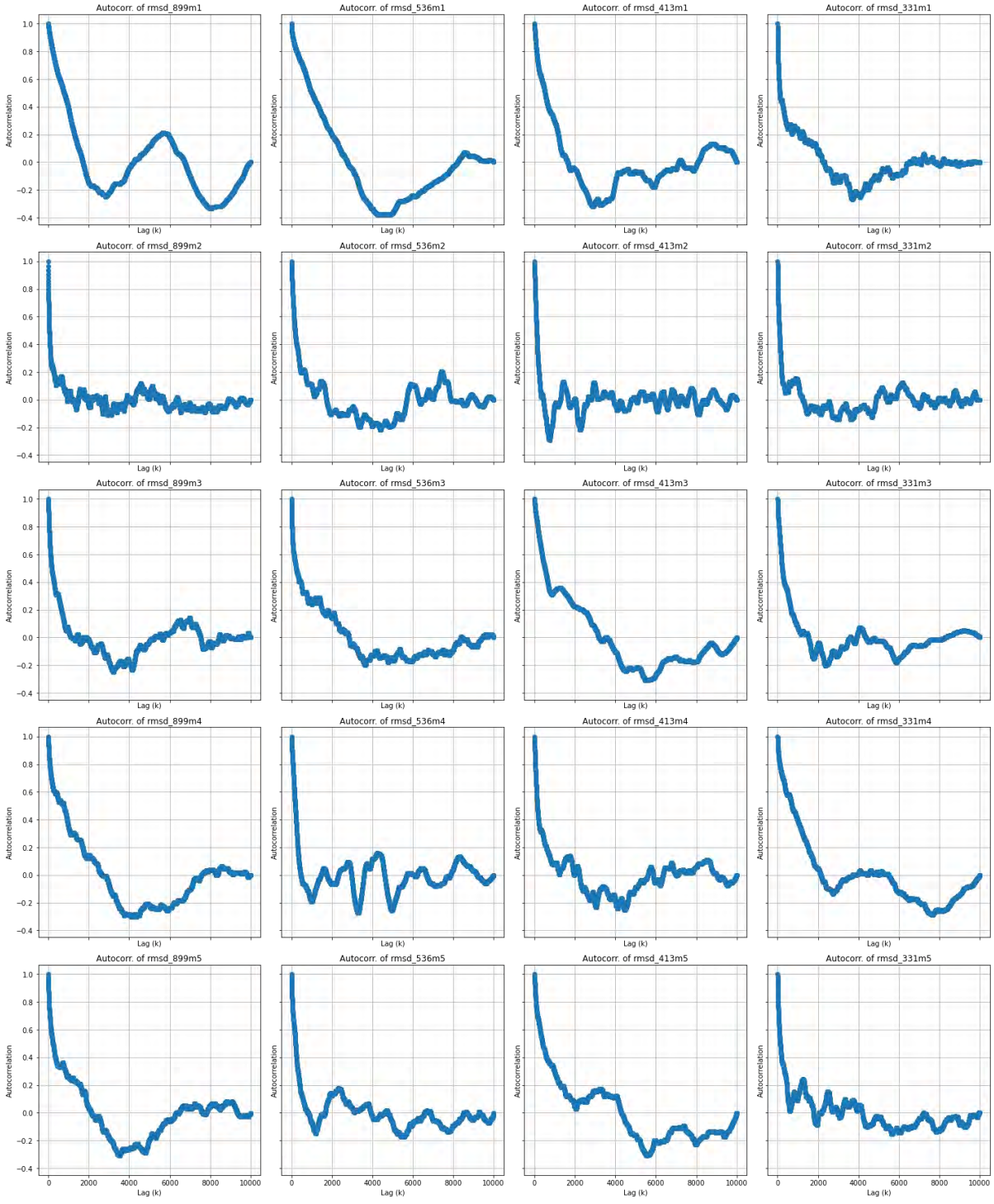


Figure S.15: Autocorrelation plots based on the RMSD of the aptamers-miR-10b-5p for the 5 models of each complex.

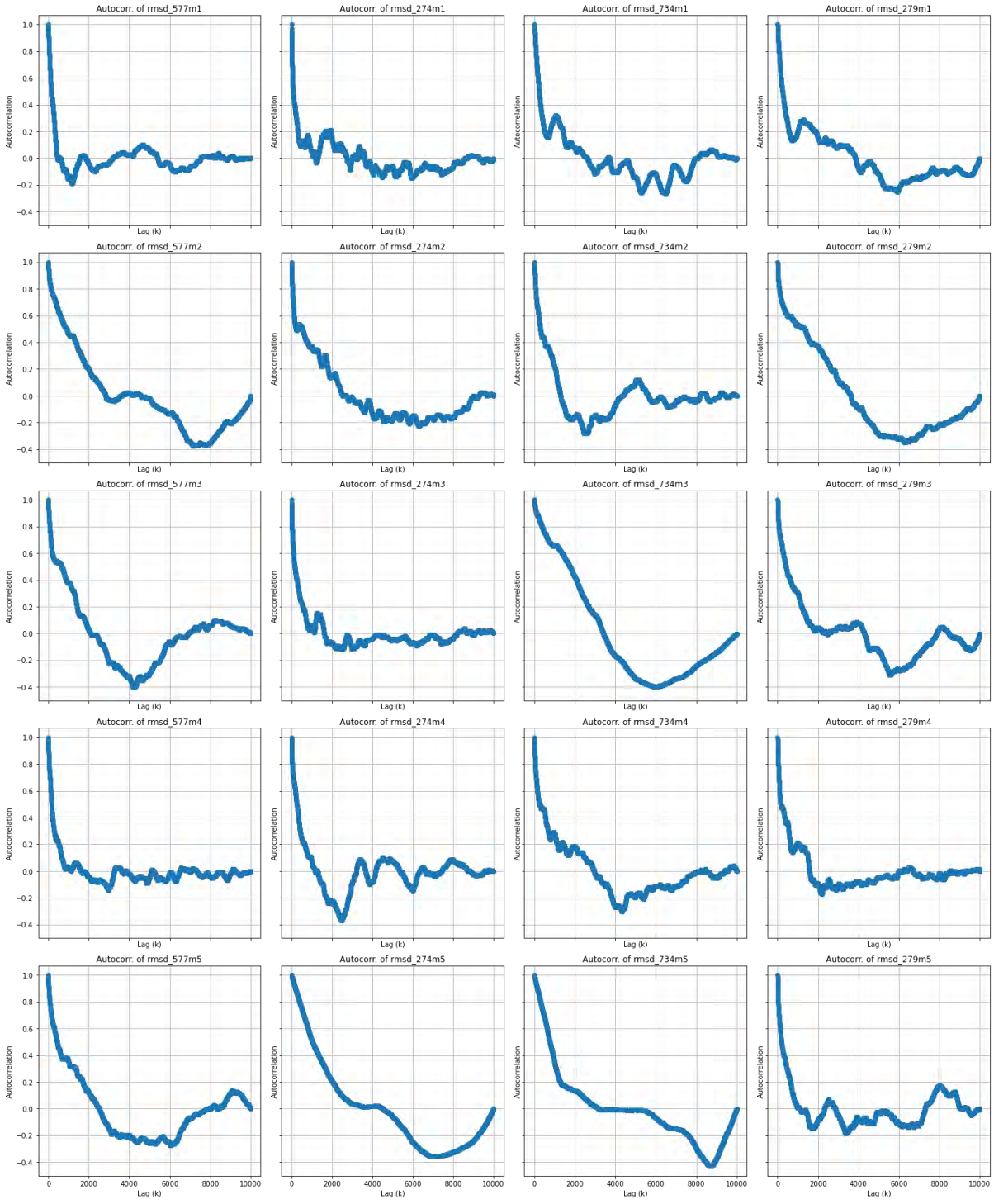


Figure S.15: Autocorrelation plots based on the RMSD of the aptamers-miR-10b-3p for the 5 models of each complex.

Tables S.1: M_{seqs} one-way ANOVA on the single composition

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Us	**20000**	**186997**	**9,34985**	**17,05291**
Gs	**20000**	**187102**	**9,3551**	**17,02896**
As	**20000**	**187605**	**9,38025**	**17,14712**
Cs	**20000**	**187053**	**9,35265**	**17,35696**

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**11,79924**	**3**	**3,933079**	**0,229381**	**0,875989**	**2,60502**
Within Groups	**1371650**	**79996**	**17,14648**			
Total	**1371662**	**79999**				

C1: This is perl script for monitoring possible bonds existing between Chan A and B of complex

```
1.                    ##############################################
2.                    #                                            #
3.                    #  This script is for monitoring Hydrogen    #
4.                    #  Bonds, Pi bonds between the Chain A and   #
5.                    #  B Complex.                                #
6.                    #  It is only written and dedicated for      #
7.                    #  dealing with RNA-RNA, Protein-RNA,        #
8.                    #  Protein-DNA, and DNA-DNA complexes.       #
9.                    #                                            #
10.                   #  This script is written by                 #
11.                   #  Kabelo Phuti Mokgopa                      #
12.                   #                                            #
13.                   ##############################################
14.
15. use MdmDiscoveryScript; # For working with Molecule Window
16. use strict;
17.
18. my $document = DiscoveryScript::LastActiveDocument(MdmModelType);
19.
20. die "No Molecule Window!" unless $document;
21.
```

```
my $molecules = $document->Molecules();

foreach my $molecule (@$molecules)
{
    # Print the name of the molecule
    print $molecule->Name . "\n";
    #print $molecule->Chains . "\n";

    ##########################################################################
    ## !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!#
    ##########################################################################

use strict;
use MdmDiscoveryScript;
use DSCommands;

my $document = DiscoveryScript::LastActiveDocument(MdmModelType);

die "No Molecule Window!" unless $document;


# Creates a hbond monitor between Chains
my $ligands     = $document->Chains->[1];;
my $LigName     = $ligands->Name;
my $ligandAtoms = $ligands->Atoms;
my $monitor     = $document->CreateHydrogenBondMonitor( $ligandAtoms,
    Mdm::allAtomHydrogenBonds, Mdm::allMolecularHydrogenBonds );

# Shows the label for each of the hbonds
$monitor->IsLabelShown(True);
$document->DeselectAll();

printf "Default distance criterion for counting Hbonds: %.3f [A]\n",
  sqrt $monitor->DefaultDistanceCriterion;
printf "%d hydrogen bonds between ligand and receptor.\n\n",
  $monitor->HydrogenBondCount;

my $newHBDist = 3.0;
$monitor->HydrogenBondDistanceSquared( $newHBDist* $newHBDist );

# update view to ensure the correct HB count
$document->UpdateViews();
printf "Set the distance criterion for counting Hbonds: %.3f [A]\n",
  sqrt $monitor->HydrogenBondDistanceSquared;

printf "%d hydrogen bonds between ligand and receptor:\n",
  $monitor->HydrogenBondCount;

```

```
70.  my $hbonds = $document->HydrogenBonds;
71.  my $count = $hbonds->Count;
72.  for ( my $index = 0 ; $index < $count ; ++$index )
73.  {
74.     my $hbond = $hbonds->Item($index);
75.     my $dist  = $hbond->Distance;
76.     my $atom1 = $hbond->Atom1;
77.     my $atom2 = $hbond->Atom2;
78.     my $res   = $atom1->Parent;
79.
80.     if (  $res->Name eq $LigName ) {
81.         $atom2->Parent->Select;
82.     }
83.     else{
84.         $atom1->Parent->Select;
85.     }
86.     printf "%d) Distance: %.3f A  H-bond between Atom %s:%s:%s and Atom %s:%s:%s\n",
87.       $index+1, $dist,
88.       $atom1->Parent->Parent->Name, $atom1->Parent->Name, $atom1->Name,
89.       $atom2->Parent->Parent->Name, $atom2->Parent->Name, $atom2->Name ;
90.  }
91.
92.   $document->FitView();
93.   my $display_style = { DisplayStyle=>'styleAtomStick' };
94.   $document->SetAtomDisplayStyle($display_style);
95.   $document->DeselectAll();
96.
97.
98. print "////////////////////////////////////////////////////////////////////////.\n";
99. print "/            Monitory Pi-bonds in Whole Complex                      /.\n";
100.print "////////////////////////////////////////////////////////////////////////.\n";
101.print "\n";
102.
103.#!/usr/bin/perl -w
104.# pi interaction monitors.
105.use strict;
106.use MdmDiscoveryScript;
107.use DSCommands;
108.my $document = DiscoveryScript::LastActiveDocument(MdmModelType);
109.
110.die "No Molecule Window!" unless $document;
111.
112.my $molecules = $document->Molecules();
113.
114.
115.# Creates pi interaction monitors
116.my $molecule = $document->Molecules->Item(0);
117.
```

```
118. my $pipiMonitor     = $molecule->CreatePiPiMonitor();
119. my $piCationMonitor = $molecule->CreatePiCationMonitor();
120. my $piSigmaMonitor  = $molecule->CreatePiSigmaMonitor();
121.
122. $pipiMonitor->Name     = "Pi-Pi";
123. $piCationMonitor->Name = "Pi-Cation";
124. $piSigmaMonitor->Name  = "Pi-Sigma";
125.
126. my $piPiInteractions     = $pipiMonitor->PiBonds();
127. my $piCationInteraction  = $piCationMonitor->PiBonds();
128. my $piSigmaInteraction   = $piSigmaMonitor->PiBonds();
129. my $piPiInteractionCount = $piPiInteractions->Count();
130. my $piCationCount        = $piCationInteraction->Count();
131. my $piSigmaCount         = $piSigmaInteraction->Count();
132. print "There are "
133.   . $piPiInteractionCount
134.   . " pi-pi interactions in this molecule.\n";
135.
136. print "There are "
137.   . $piCationCount
138.   . " pi-Cation interactions in this molecule.\n";
139.
140. print "There are "
141.   . $piSigmaCount
142.   . " pi-Cation interactions in this molecule.\n";
143.
144.
145. print "////////////////////////////////////////////////////////////////////////.\n";
146. print "/             Monitory Pi-bonds CHAIN A                                /.\n";
147. print "////////////////////////////////////////////////////////////////////////.\n";
148. print "\n";
149.
150. print "\n ";
151.
152. print "        ///////////////Default Parameters/////////////////\n";
153. print "\n"
154. use strict;
155. use MdmDiscoveryScript;
156. use DSCommands;
157.
158.
159. my $document = DiscoveryScript::LastActiveDocument(MdmModelType);
160. die "No Molecule Window!" unless $document;
161.
162. my $molecules = $document->Molecules();
163. my $molecule = $document->Molecules->Item(0);
164. my $ChainA     = $document->Chains->[0];
165.
```

```
166. my $pipiMonitor     = $ChainA->CreatePiPiMonitor();
167. my $piCationMonitor = $ChainA->CreatePiCationMonitor();
168. my $piSigmaMonitor  = $ChainA->CreatePiSigmaMonitor();
169.
170. $pipiMonitor->Name     = "Pi-Pi";
171. $piCationMonitor->Name = "Pi-Cation";
172. $piSigmaMonitor->Name  = "Pi-Sigma";
173.
174. my $piPiInteractions     = $pipiMonitor->PiBonds();
175. my $piPiInteractionCount = $piPiInteractions->Count();
176. print "There are "
177.   . $piPiInteractionCount
178.   . " pi-pi interactions in this molecule.\n";
179.
180. my $piCationInteractions     = $piCationMonitor->PiBonds();
181. my $piCationInteractionCount = $piCationInteractions->Count();
182. print "There are "
183.   . $piCationInteractionCount
184.   . " pi-cation interactions in this molecule.\n";
185.
186. my $piSigmaInteractions     = $piSigmaMonitor->PiBonds();
187. my $piSigmaInteractionCount = $piSigmaInteractions->Count();
188. print "There are "
189.   . $piSigmaInteractionCount
190.   . " pi-sigma interactions in this molecule.\n";
191.
192. print " ";
193.
194. print "        ///////////////Parameters//////////////////\n";
195. print "        /DistanceCutoff             = 5.0          /\n";
196. print "        /MinimumAngle               = 60.0         /\n";
197. print "        /AngleDeviation             = 15.0         /\n";
198. print "        /MaximumCenterDistance      = 7.0          /\n";
199. print "        /MaximumClosestAtomDistance = 4.0          /\n";
200. print "        /MaximumLambda              = 20.0         /\n";
201. print "        /MaximumTheta               = 45.0         /\n";
202. print "        ////////////////////////////////////////////\n";
203. print " ";
204.
205. # Change criteria settings to non-default values.
206. $piCationMonitor->DistanceCutoff = 5.0;
207. $piCationMonitor->MinimumAngle   = 60.0;
208. my $piCationInteractions     = $piCationMonitor->PiBonds();
209. my $piCationInteractionCount = $piCationInteractions->Count();
210. print "There are now "
211.   . $piCationInteractionCount
212.   . " pi-cation interactions in this molecule\n";
213.
```

```
$piSigmaMonitor->DistanceCutoff = 5.0;
$piSigmaMonitor->MinimumAngle   = 30.0;
$piSigmaMonitor->AngleDeviation = 15.0;
my $piSigmaInteractions     = $piSigmaMonitor->PiBonds();
my $piSigmaInteractionCount = $piSigmaInteractions->Count();
print "There are now "
  . $piSigmaInteractionCount
  . " pi-sigma interactions in this molecule.\n";

$pipiMonitor->MaximumCenterDistance      = 7.;
$pipiMonitor->MaximumClosestAtomDistance = 4.0;
$pipiMonitor->MaximumLambda              = 20.;
$pipiMonitor->MaximumTheta               = 45.;
my $pipiInteractions     = $pipiMonitor->PiBonds();
my $pipiInteractionCount = $pipiInteractions->Count();
print "There are now "
  . $pipiInteractionCount
  . " pi-pi interactions in this molecule.\n";

my $piInteractions = $molecules->PiInteractionMonitors();
my $piCount        = $piInteractions->Count();
print "\nThere are " . $piCount . " pi interaction monitors: \n";
for ( my $index = 0 ; $index < $piCount ; ++$index )
{
    my $monitor = $piInteractions->Item($index);
    printf( "%d \t %s \t %s \n",
        $index, $monitor->Name(), $monitor->GetProperty("Interactions") );
}



print "////////////////////////////////////////////////////////////////////////.\n";
print "/             Monitory Pi-bonds CHAIN B                                /.\n";
print "////////////////////////////////////////////////////////////////////////.\n";
print "////////////////////////////////////////////////////////////////////////.\n";

print " \n";
print "        ///////////////Default Parameters//////////////////\n";
print " \n";

use strict;
use MdmDiscoveryScript;
use DSCommands;


my $document = DiscoveryScript::LastActiveDocument(MdmModelType);

```

```
die "No Molecule Window!" unless $document;

my $molecules = $document->Molecules();
#my $molecule = $document->Molecules->Item(1);
my $ChainB     = $document->Chains->[1];

my $pipiMonitor     = $ChainB->CreatePiPiMonitor();
my $piCationMonitor = $ChainB->CreatePiCationMonitor();
my $piSigmaMonitor  = $ChainB->CreatePiSigmaMonitor();

$pipiMonitor->Name     = "Pi-Pi";
$piCationMonitor->Name = "Pi-Cation";
$piSigmaMonitor->Name  = "Pi-Sigma";

my $piPiInteractions     = $pipiMonitor->PiBonds();
my $piPiInteractionCount = $piPiInteractions->Count();
print "There are "
  . $piPiInteractionCount
  . " pi-pi interactions in this molecule.\n";

my $piCationInteractions     = $piCationMonitor->PiBonds();
my $piCationInteractionCount = $piCationInteractions->Count();
print "There are "
  . $piCationInteractionCount
  . " pi-cation interactions in this molecule.\n";

my $piSigmaInteractions     = $piSigmaMonitor->PiBonds();
my $piSigmaInteractionCount = $piSigmaInteractions->Count();
print "There are "
  . $piSigmaInteractionCount
  . " pi-sigma interactions in this molecule.\n";
print " \n";

print "        ///////////////Parameters//////////////////\n";
print "        /DistanceCutoff             = 5.0          /\n";
print "        /MinimumAngle               = 60.0         /\n";
print "        /AngleDeviation             = 15.0         /\n";
print "        /MaximumCenterDistance      = 7.0          /\n";
print "        /MaximumClosestAtomDistance = 4.0          /\n";
print "        /MaximumLambda              = 20.0         /\n";
print "        /MaximumTheta               = 45.0         /\n";
print "        ////////////////////////////////////////////\n";
print " \n";

$piCationMonitor->DistanceCutoff = 5.0;
$piCationMonitor->MinimumAngle   = 60.0;
my $piCationInteractions     = $piCationMonitor->PiBonds();
my $piCationInteractionCount = $piCationInteractions->Count();
```

```
print "There are now "
  . $piCationInteractionCount
  . " pi-cation interactions in this molecule\n";

$piSigmaMonitor->DistanceCutoff = 5.0;
$piSigmaMonitor->MinimumAngle   = 30.0;
$piSigmaMonitor->AngleDeviation = 15.0;
my $piSigmaInteractions     = $piSigmaMonitor->PiBonds();
my $piSigmaInteractionCount = $piSigmaInteractions->Count();
print "There are now "
  . $piSigmaInteractionCount
  . " pi-sigma interactions in this molecule.\n";

$pipiMonitor->MaximumCenterDistance      = 7.;
$pipiMonitor->MaximumClosestAtomDistance = 4.0;
$pipiMonitor->MaximumLambda              = 20.;
$pipiMonitor->MaximumTheta               = 45.;
my $pipiInteractions     = $pipiMonitor->PiBonds();
my $pipiInteractionCount = $pipiInteractions->Count();
print "There are now "
  . $pipiInteractionCount
  . " pi-pi interactions in this molecule.\n";

my $piInteractions = $molecules->PiInteractionMonitors();
my $piCount        = $piInteractions->Count();
print "\nThere are " . $piCount . " pi interaction monitors: \n";
for ( my $index = 0 ; $index < $piCount ; ++$index )
{
    my $monitor = $piInteractions->Item($index);
    printf( "%d \t %s \t %s \n",
        $index, $monitor->Name(), $monitor->GetProperty("Interactions") );
}


print "////////////////////////////////////////////////////////////////////////.\n";
print "/        Monitory Pi-bonds between CHAIN A and B                       /.\n";
print "////////////////////////////////////////////////////////////////////////.\n";
print "////////////////////////////////////////////////////////////////////////.\n";
use strict;
use MdmDiscoveryScript;
use DSCommands;


my $document = DiscoveryScript::LastActiveDocument(MdmModelType);

die "No Molecule Window!" unless $document;

```

```
my $molecules = $document->Molecules();
my $ligands     = $document->Chains->[1];
my $ligandAtoms = $ligands->Atoms;
my $LigName     = $ligands->Name;
my $receptor     = $document->Chains->[0];

my $ligandAtomArray = Mdm::Array::Create();

for ( my $index = 0 ; $index < $ligandAtoms->Count() ; ++$index )
{
    $ligandAtomArray->AddItem( $ligandAtoms->Item($index) );
}

my $pipiMonitor2          = $receptor->CreatePiPiMonitor($ligandAtomArray);
my $pipiInteractions2     = $pipiMonitor2->PiBonds();
my $pipiInteractionCount2 = $pipiInteractions2->Count();

print "Number of pi-pi interactions between Chain"
  . $ligands->Name()
  . " and the rest of the Chain A: "
  . $pipiInteractionCount2 . "\n";


# Creating Pi-cation monitor using ligand atoms
my $pipiMonitor2          = $receptor->CreatePiCationMonitor($ligandAtomArray);
my $pipiInteractions2     = $pipiMonitor2->PiBonds();
my $pipiInteractionCount2 = $pipiInteractions2->Count();

# Printing pi-cation interactions between Chains
print "Number of pi-cation interactions between Chain"
  . $ligands->Name()
  . " and the rest of the Chain A: "
  . $pipiInteractionCount2 . "\n";


# Creating Pi-sigma monitor using ligand atoms
my $pipiMonitor2          = $receptor->CreatePiSigmaMonitor($ligandAtomArray);
my $pipiInteractions2     = $pipiMonitor2->PiBonds();
my $pipiInteractionCount2 = $pipiInteractions2->Count();

# Printing pi-sigma interactions between Chains
print "Number of pi-sigma interactions between Chain"
  . $ligands->Name()
  . " and the rest of the Chain A: "
  . $pipiInteractionCount2 . "\n";

}
```

C2: This is python script for full automated for MD preparation and error handling (including fixing RNA for specific force field) . This design specifically for RNA-RNA complex simulations on GROMACS.(this script will be part of the pySMQM tool which is extension tool for T_SELEX, independent of T_SELEX)

```
import os
import sys

class mdp_files:

    @staticmethod
    def nvt():

        nvt_data='''title                   = Protein-ligand complex NVT equilibration
        define                  = -DPOSRES  ; position restrain the protein and ligand
        ; Run parameters
        integrator              = md        ; leap-frog integrator
        nsteps                  = 50000     ; 2 * 50000 = 100 ps
        dt                      = 0.002     ; 2 fs
        ; Output control
        nstenergy               = 500   ; save energies every 1.0 ps
        nstlog                  = 500   ; update log file every 1.0 ps
        nstxout-compressed      = 500   ; save coordinates every 1.0 ps
        ; Bond parameters
        continuation            = no        ; first dynamics run
        constraint_algorithm    = lincs     ; holonomic constraints
        constraints             = h-bonds   ; bonds to H are constrained
        lincs_iter              = 1         ; accuracy of LINCS
        lincs_order             = 4         ; also related to accuracy
        ; Neighbor searching and vdW
        cutoff-scheme           = Verlet
        ns_type                 = grid      ; search neighboring grid cells
        nstlist                 = 20        ; largely irrelevant with Verlet
        rlist                   = 1.2
        vdwtype                 = cutoff
        vdw-modifier            = force-switch
        rvdw-switch             = 1.0
        rvdw                    = 1.2       ; short-range van der Waals cutoff (in nm)
        ; Electrostatics
        coulombtype             = PME       ; Particle Mesh Ewald for long-range
electrostatics
        rcoulomb                = 1.2       ; short-range electrostatic cutoff (in nm)
        pme_order               = 4         ; cubic interpolation
        fourierspacing          = 0.16      ; grid spacing for FFT
        ; Temperature coupling
        tcoupl                  = V-rescale                     ; modified Berendsen
thermostat
        tc-grps                 = RNA Water_and_ions        ; two coupling groups - more
accurate
        tau_t                   = 0.1   0.1                     ; time constant, in ps
        ref_t                   = 300   300                     ; reference temperature,
one for each group, in K
        ; Pressure coupling
        pcoupl                  = no        ; no pressure coupling in NVT
        ; Periodic boundary conditions
        pbc                     = xyz       ; 3-D PBC
```

```
        ; Dispersion correction is not used for proteins with the C36 additive FF
        DispCorr                 = no
        ; Velocity generation
        gen_vel                  = yes       ; assign velocities from Maxwell distribution
        gen_temp                 = 300       ; temperature for Maxwell distribution
        gen_seed                 = -1        ; generate a random seed

        '''
    return nvt_data

    def npt():
        npt_data = '''title                    = OPLS Lysozyme NPT equilibration
        define                   = -DPOSRES  ; position restrain the protein
        ; Run parameters
        integrator               = md        ; leap-frog integrator
        nsteps                   = 50000     ; 2 * 50000 = 100 ps
        dt                       = 0.002     ; 2 fs
        ; Output control
        nstxout                  = 500       ; save coordinates every 1.0 ps
        nstvout                  = 500       ; save velocities every 1.0 ps
        nstenergy                = 500       ; save energies every 1.0 ps
        nstlog                   = 500       ; update log file every 1.0 ps
        ; Bond parameters
        continuation             = yes       ; Restarting after NVT
        constraint_algorithm     = lincs     ; holonomic constraints
        constraints              = h-bonds   ; bonds involving H are constrained
        lincs_iter               = 1         ; accuracy of LINCS
        lincs_order              = 4         ; also related to accuracy
        ; Nonbonded settings
        cutoff-scheme            = Verlet    ; Buffered neighbor searching
        ns_type                  = grid      ; search neighboring grid cells
        nstlist                  = 10        ; 20 fs, largely irrelevant with Verlet scheme
        rcoulomb                 = 1.0       ; short-range electrostatic cutoff (in nm)
        rvdw                     = 1.0       ; short-range van der Waals cutoff (in nm)
        DispCorr                 = EnerPres  ; account for cut-off vdW scheme
        ; Electrostatics
        coulombtype              = PME       ; Particle Mesh Ewald for long-range
electrostatics
        pme_order                = 4         ; cubic interpolation
        fourierspacing           = 0.16      ; grid spacing for FFT
        ; Temperature coupling is on
        tcoupl                   = V-rescale             ; modified Berendsen thermostat
        tc-grps                  = Protein Non-Protein   ; two coupling groups - more
accurate
        tau_t                    = 0.1     0.1           ; time constant, in ps
        ref_t                    = 300     300           ; reference temperature, one for
each group, in K
        ; Pressure coupling is on
        pcoupl                   = Parrinello-Rahman     ; Pressure coupling on in NPT
        pcoupltype               = isotropic             ; uniform scaling of box vectors
        tau_p                    = 2.0                   ; time constant, in ps
        ref_p                    = 1.0                   ; reference pressure, in bar
        compressibility          = 4.5e-5                ; isothermal compressibility of
water, bar^-1
        refcoord_scaling         = com
        ; Periodic boundary conditions
        pbc                      = xyz       ; 3-D PBC
```

```
        ; Velocity generation
        gen_vel                 = no        ; Velocity generation is off

        '''
        return npt_data


    def ions():
        ions_data = '''; LINES STARTING WITH ';' ARE COMMENTS
        title                   = Minimization ; Title of run

        ; Parameters describing what to do, when to stop and what to save
        integrator     = steep                  ; Algorithm (steep = steepest descent
minimization)
        emtol                   = 2000.0       ; Stop minimization when the maximum force
< 10.0 kJ/mol
        emstep          = 0.01      ; Energy step size
        nsteps                  = 50000                 ; Maximum number of (minimization)
steps to perform

        ; Parameters describing how to find the neighbors of each atom and how to calculate
the interactions
        nstlist                 = 1                ; Frequency to update the neighbor list
and long range forces
        cutoff-scheme   = Verlet
        ns_type                 = grid                   ; Method to determine neighbor
list (simple, grid)
        coulombtype     = cutoff
        rlist                   = 1.0          ; Cut-off for making neighbor list (short
range forces)
        rcoulomb       = 1.0          ; long range electrostatic cut-off
        rvdw                    = 1.0          ; long range Van der Waals cut-off
        pbc             = xyz                  ; Periodic Boundary Conditions
        '''

    def em():
        em_data = '''; LINES STARTING WITH ';' ARE COMMENTS
        title                   = Minimization ; Title of run

        ; Parameters describing what to do, when to stop and what to save
        integrator     = steep                  ; Algorithm (steep = steepest descent
minimization)
        emtol                   = 10000.0      ; Stop minimization when the maximum force
< 10.0 kJ/mol
        emstep          = 0.01      ; Energy step size
        nsteps                  = 10000000              ; Maximum number of (minimization)
steps to perform

        ; Parameters describing how to find the neighbors of each atom and how to calculate
the interactions
        nstlist                 = 1                   ; Frequency to update the neighbor
list and long range forces
        cutoff-scheme   = Verlet
        ns_type                 = grid                       ; Method to determine neighbor
list (simple, grid)
        rlist                   = 1.2             ; Cut-off for making neighbor list
(short range forces)
```

```
        coulombtype     = PME              ; Treatment of long range electrostatic
interactions
        rcoulomb        = 1.2              ; long range electrostatic cut-off
        vdwtype          = cutoff
        vdw-modifier     = force-switch
        rvdw-switch      = 1.0
        rvdw                    = 1.2              ; long range Van der Waals cut-off
        pbc              = xyz                     ; Periodic Boundary Conditions
        DispCorr         = no

        '''
        return em_data

    def ions():
        ions_data= '''; LINES STARTING WITH ';' ARE COMMENTS
        title                   = Minimization ; Title of run

        ; Parameters describing what to do, when to stop and what to save
        integrator      = steep                 ; Algorithm (steep = steepest descent
minimization)
        emtol                   = 2000.0        ; Stop minimization when the maximum force
< 10.0 kJ/mol
        emstep           = 0.01      ; Energy step size
        nsteps                  = 50000                 ; Maximum number of (minimization)
steps to perform

        ; Parameters describing how to find the neighbors of each atom and how to calculate
the interactions
        nstlist                 = 1                ; Frequency to update the neighbor list
and long range forces
        cutoff-scheme   = Verlet
        ns_type                 = grid                  ; Method to determine neighbor
list (simple, grid)
        coulombtype      = cutoff
        rlist                   = 1.0          ; Cut-off for making neighbor list (short
range forces)
        rcoulomb        = 1.0          ; long range electrostatic cut-off
        rvdw                    = 1.0          ; long range Van der Waals cut-off
        pbc              = xyz                  ; Periodic Boundary Conditions

        '''
        return ions_data

    def md():
        md_data = ''' title                    = Protein-ligand complex MD simulation
        ; Run parameters
        integrator              = md        ; leap-frog integrator
        nsteps                  = 50000000  ; 2 * 50000000 = 100000 ps (100 ns)
        dt                      = 0.002     ; 2 fs
        ; Output control
        nstenergy               = 5000      ; save energies every 10.0 ps
        nstlog                  = 5000      ; update log file every 10.0 ps
        nstxout-compressed      = 5000      ; save coordinates every 10.0 ps
        ; Bond parameters
        continuation            = yes       ; continuing from NPT
        constraint_algorithm    = lincs     ; holonomic constraints
        constraints             = h-bonds   ; bonds to H are constrained
```

```
        lincs_iter              = 1         ; accuracy of LINCS
        lincs_order             = 4         ; also related to accuracy
        ; Neighbor searching and vdW
        cutoff-scheme           = Verlet
        ns_type                 = grid      ; search neighboring grid cells
        nstlist                 = 20        ; largely irrelevant with Verlet
        rlist                   = 1.2
        vdwtype                 = cutoff
        vdw-modifier            = force-switch
        rvdw-switch             = 1.0
        rvdw                    = 1.2       ; short-range van der Waals cutoff (in nm)
        ; Electrostatics
        coulombtype             = PME       ; Particle Mesh Ewald for long-range
electrostatics
        rcoulomb                = 1.2
        pme_order               = 4         ; cubic interpolation
        fourierspacing          = 0.16      ; grid spacing for FFT
        ; Temperature coupling
        tcoupl                  = V-rescale                     ; modified Berendsen
thermostat
        tc-grps                 = RNA Water_and_ions        ; two coupling groups - more
accurate
        tau_t                   = 0.1   0.1                     ; time constant, in ps
        ref_t                   = 300   300                     ; reference temperature,
one for each group, in K
        ; Pressure coupling
        pcoupl                  = Parrinello-Rahman             ; pressure coupling is on
for NPT
        pcoupltype              = isotropic                     ; uniform scaling of box
vectors
        tau_p                   = 2.0                           ; time constant, in ps
        ref_p                   = 1.0                           ; reference pressure, in
bar
        compressibility         = 4.5e-5                        ; isothermal
compressibility of water, bar^-1
        ; Periodic boundary conditions
        pbc                     = xyz       ; 3-D PBC
        ; Dispersion correction is not used for proteins with the C36 additive FF
        DispCorr                = no
        ; Velocity generation
        gen_vel                 = no        ; continuing from NPT equilibration
         '''
        return md_data


 class wrt_mdps:

    @staticmethod
    def wrt_nvt():
        m= mdp_files()
        nvt_txt = m.nvt()
        with open("nvt.mdp", "w") as file:
            file.write(nvt_txt)

    def wrt_npt():
        m= mdp_files()
        npt_txt = m.npt()
```

```
        with open("npt.mdp", "w") as file:
            file.write(npt_txt)

    def wrt_em():
        m= mdp_files()
        em_txt = m.em()
        with open("em.mdp", "w") as file:
            file.write(em_txt)

    def wrt_ions():
        m= mdp_files()
        ions_txt = m.ions()
        with open("ions.mdp", "w") as file:
            file.write(ions.txt)

    def wrt_md():
        m= mdp_files()
        md_txt = m.md()
        with open("md.mdp", "w") as file:
            file.write(md_txt)

class extracts_ini:

    def extract_force_fields(file_path):
        with open(file_path, 'r') as file:
            lines = file.readlines()
            for line in lines:
                if line.startswith('#start'):
                    continue
                elif line.startswith('ff'):

                    data = line.split()
                    forceField = data[2]
                    if forceField == 'AMBER03':
                        ff_value =2

                    elif forceField == 'AMBER94':
                        ff_value = 3

                    elif forceField == 'AMBER96':
                        ff_value = 4

                    elif forceField == 'AMBER99':
                        ff_value = 5

                    elif forceField == 'AMBER99SB':
                        ff_value = 6
                    elif forceField == 'AMBER99SB-ILDN':
                        ff_value = 9
                    else:
                        error = '''Fatal error:
                                Incorrect force field (ff) ID or option intered in
                                in the md.ini file, please look for the correct ID in the
documentation
                        '''
                        print(error)
                        break
```

```
        return ff_value

    def extract_solvent(file_path):
        with open(file_path, 'r') as file:
            lines = file.readlines()
            for line in lines:
                if line.startswith('#start'):
                    continue
                elif line.startswith('slv') or line.startswith('solvent'):
                    data = line.split()
                    solvent = data[2]
                    if solvent == 'TIP3P':
                        slv_value =1

                    elif solvent == 'TIP4P':
                        slv_value = 2

                    elif solvent == 'TIP4P-Ew':
                        slv_value = 3
                    elif solvent == 'TIP5P':
                        slv_value = 4

                    elif solvent == 'SPC':
                        slv_value = 5

                    elif solvent == 'SPC/E':
                        slv_value = 6

                    elif solvent == 'None':
                        slv_value = 7
                    else:
                        error = '''Fatal error:
                                Incorrect solvent (slv) ID or option intered in
                                the md.ini file, please look for the correct ID in the
documentation
                        '''
                        print(error)
                        break
            return slv_value


    def cube_details(file_path):
        with open(file_path, 'r') as file:
            lines = file.readlines()
            for line in lines:
                if line.startswith('#start'):
                    continue

                if line.startswith('b_shp'):
                    data = line.split()
                     box_shape= data[2]
                else:
                    box_shape == 'cubic'
                    error = """Warning:
```

```
                                Incorrect box shape was entered in
                                the md.ini file, therefore cubic shape was choosen for this
sumulation
                        """
                        print(error)


                if line.startswith('b_sz'):
                    data = line.split()
                     box_size= data[2]
                else:
                    box_size = 3.0
                    error = """Warning:
                                Incorrect box size was entered in
                                the md.ini file, therefore maximum box size of 3.0 A was
choosen for this sumulation """

                        print(error)
        return float(box_size), box_shape


    def n_steps(file_path):
        with open(file_path, 'r') as file:
            lines = file.readlines()
            for line in lines:
                if line.startswith('#start'):
                    continue

                if line.startswith('n_nvt'):
                    data = line.split()
                    nsteps_nvt = data[2]
                 else :
                  print("Error")

                if line.startswith('n_npt'):
                    data = line.split()
                    nsteps_npt = data[2]
                else:
                    print("Error")

                if line.startswith('n_em'):
                    data = line.split()
                    nsteps_em  = data[2]
                else:
                    print("Error")

                if line.startswith('n_md'):
                    data = line.split()
                    nsteps_md = data[2]
                else:
                    print("Error")

        return float(nsteps_em), float(nsteps_nvt), float(nsteps_npt),float(nsteps_md),


class edit_mdpfiles:
```

```
    def flie_edit():
        p = extracts_ini()
        n1,n2,n3,n4 = p.n_steps()
        os.system("sed '/^nsteps/s/50000/{}/g' nvt.mdp".format(n2))
        os.system("sed '/^nsteps/s/50000/{}/g' npt.mdp".format(n3))
        os.system("sed '/^nsteps/s/10000000/{}/g' em.mdp".format(n1))
        os.system("sed '/^nsteps/s/50000/{}/g' md.mdp".format(n4))

class fixing_pdb_files:
    def fixing_chains (pdb_file):
        os.system("sed '/HEADER lig_1.pdb/,$s/A/B/g' {} >
model_12221.pdb".format(pdb_file))
        os.system("sed -i 's/B B/A B/g' model_12221.pdb")
        os.system("sed -i 's/BTOM/ATOM/g' model_12221.pdb")
        os.system("rm -f working.pdb")
        os.system("cp model_12221.pdb working.pdb")

    def fixing_atoms(working.pdb):
        #atoms name corrections
        #sed -i 's/\x27HO5/ H5T/' working.pdba
        #ATOM      4  H5'   U A   1       2.408  -0.393  -1.179  0.00  0.00           H
        #ATOM      5 'H5'   U A   1       3.869  -1.190  -0.569  0.00  0.00           H
        os.system("sed -i \'s/\\x27HO5/ H5T/\' working.pdb")
        os.system("sed -i \'s/\\x27H5\\x27/H5\\x27"+"2/\' working.pdb")                #
'H5' -> H5'2
        os.system("sed -i \'s/ H5\\x27 /H5\\x27"+"1 /\' working.pdb")                  #
H5' -> H5'1 2
        #os.system("sed -i \'s/'HO2/HO'2/\' working.pdb")                  #  H5' -> H5'1 2
        os.system("sed -i \'s/\\x27HO2/HO\\x27"+"2/\' working.pdb")                  #  H5'
-> H5'1 2
        os.system("sed -i \'s/\\x27HO3/ H3T/\' working.pdb")                  #  H5' -> H5'1
2
        os.system("sed -i \'s/HO2\\x27 /H2\\x27"+"1 /\' working.pdb")     #  HO2' -> H2'1
        os.system("sed -i \'s/\\x27HO3/ H5T/\' working.pdb")

        #os.system("sed -i \'r/H5T/\\\'HO5\' working.pdb")
        print("sed -i \'r/\\\'HO5/H5T/\' working.pdb\n")



import os

class gromacs_automation:

    def initialization(self):
        try:
            os.system("gmx")
            gromacs_location = "without MPI"
            statement = "GROMACS is available without MPI."
            print(statement)
        except FileNotFoundError:
            try:
                os.system("gmx_mpi")
                gromacs_location = "with Open MPI"
                statement = "GROMACS is available with Open MPI."
                print(statement)
```

```
            except FileNotFoundError:
                gromacs_location = "not found"
                statement = "GROMACS is not installed or not found in the system."
                print(statement)
                raise FileNotFoundError("GROMACS is not installed or not found in the
system.")
        return gromacs_location

    def running_gromacs(self, pdb_file,file_path):

        location = self.initialization()
        p = extracts_ini()
        c_d = p.cube_details(file_path)
        solvnt = p.extract_solvent(file_path)
        ffield =p.extract_force_fields(file_path)

        if location == "without MPI":
            os.system('echo -e "{}\n{}\n8\n7"  | gmx pdb2gmx -f working.pdb -o
model_processed.gro -ter'.format(ffield,solvnt))
            os.system(' gmx editconf -f model_processed.gro -o newbox.gro -bt {} -d {} -
c'.format(c_d[1],c_d[0]))
            os.system(' gmx solvate -cp newbox.gro -cs spc216.gro -p topol.top -o
solv.gro')
            os.system('gmx grompp -f ions.mdp -c solv.gro -o ions.tpr -p topol.top')
            os.system('echo 3 | gmx genion -s ions.tpr -o solv_ions.gro -p topol.top -pname
NA -nname CL -neutral')
            os.system('gmx grompp -f em.mdp -c solv_ions.gro -p topol.top -o em.tpr')
            os.system('gmx mdrun -v -deffnm em')
            os.system('gmx grompp -f nvt.mdp -c em.gro -r em.gro -p topol.top -o nvt.tpr')
            os.system('gmx mdrun -deffnm nvt')
            os.system('gmx grompp -f nvt.mdp -c nvt.gro -r nvt.gro -p topol.top -o
npt.tpr')
            os.system('gmx mdrun -deffnm npt')
            #os.system('gmx grompp -f npt.mdp -c nvt.gro -r nvt.gro -p topol.top -o
npt.tpr')
            os.system('gmx grompp -f md.mdp -c npt.gro -t npt.cpt -p topol.top -o
md_0_1.tpr')
            #os.system('gmx mdrun -deffnm md_0_1')

        elif location == "with Open MPI":
            os.system('echo -e "{}\n{}\n8\n7" | gmx_mpi pdb2gmx -f working.pdb -o
model_processed.gro -ter'.format(ffield,solvnt))
            os.system(' gmx_mpi editconf -f model_processed.gro -o newbox.gro -bt {} -d {}
-c'.format(c_d[1],c_d[0]))
            os.system(' gmx_mpi solvate -cp newbox.gro -cs spc216.gro -p topol.top -o
solv.gro')
            os.system('gmx_mpi grompp -f ions.mdp -c solv.gro -o ions.tpr -p topol.top')
            os.system('echo 3 | gmx_mpi genion -s ions.tpr -o solv_ions.gro -p topol.top -
pname NA -nname CL -neutral')
            os.system('gmx_mpi grompp -f em.mdp -c solv_ions.gro -p topol.top -o em.tpr')
            os.system('gmx_mpi mdrun -v -deffnm em')
            os.system('gmx_mpi grompp -f nvt.mdp -c em.gro -r em.gro -p topol.top -o
nvt.tpr')
            os.system('gmx_mpi mdrun -deffnm nvt')
            os.system('gmx_mpi grompp -f nvt.mdp -c nvt.gro -r nvt.gro -p topol.top -o
npt.tpr')
            os.system('gmx_mpi mdrun -deffnm npt')
```

```
            #os.system('gmx_mpi grompp -f npt.mdp -c nvt.gro -r nvt.gro -p topol.top -o
npt.tpr')
            os.system('gmx_mpi grompp -f md.mdp -c npt.gro -t npt.cpt -p topol.top -o
md_0_1.tpr')
            #os.system('gmx_mpi mdrun -deffnm md_0_1')

        elif location == "not found":
            raise FileNotFoundError("GROMACS is not installed or not found in the system.")


class gen_pbs_files:

    def pbs_text():
        pbs_file='''#!/bin/bash
        #PBS -P CHEM0802
        #PBS -N md_0_1
        #PBS -l select=8:ncpus=24:mpiprocs=24
        #PBS -l walltime=48:00:00
        #PBS -q normal
        #PBS -m be
        #PBS -M none
        #PBS -r n
        #PBS -o /mnt/lustre/users/kmokgopa/GromacsMSc/10b_5p/899/model5/md_0_100_output
        #PBS -e /mnt/lustre/users/kmokgopa/GromacsMSc/10b_5p/899/model5/md_0_100_error
        #PBS

        module purge
        module add gcc/6.1.0
        module add chpc/openmpi/3.1.0/gcc-6.1.0
        module add chpc/gromacs/2018.2/openmpi-3.1.0/gcc-6.1.0
        ulimit -s unlimited
        OMP_NUM_THREADS=1

        pushd /mnt/lustre/users/kmokgopa/GromacsMSc/10b_5p/899/model5
        EXE=/apps/chpc/chem/gromacs/2018.2/bin/gmx_mpi
        ARGS="mdrun -deffnm md_0_1 -ntomp ${OMP_NUM_THREADS}"
        mpirun -np 192 ${EXE} ${ARGS}
        popd
        '''

    def wrt_pbs():
        with open("md_0_1.pbs", "w") as file:
            file.write(pbs_file)

        #os.system('qsub md_0_1.pbs')
    def edit_pbs()

if __name__ == "__main__":
    if len(sys.argv) != 2:
        print("Usage: python script.py input_pdb config_file ")
        sys.exit(1)
    #writing mdps
    wrt_mdps.wrt_nvt()
    wrt_mdps.wrt_npt()
    wrt_mdps.wrt_em()
    wrt_mdps.wrt_ions()
    wrt_mdps.wrt_md()
```

```
    edit_mdpfiles.edit_mdpfiles()


    input_pdb = sys.argv[1]
    config_file = sys.argv[2]


    gromacs = GromacsAutomation(input_file, config_file)
    gromacs.running_gromacs()
```

C3: This is bash script for full automated post MD analysis and for computing MMGBSA and analysis.

```
#!/bin/bash


names=('734' '577' '279' '274')
models=('1' '2' '3' '4' '5')

for name in "${names[@]}"; do
    for model in "${models[@]}"; do
        # Construct the base file names
        TRJ_FILE="md_0_1_${name}m${model}.xtc"
        TPR_FILE="md_0_1_${name}m${model}.tpr"
        NDX_FILE="index_${name}m${model}.ndx"
        MDP_POLAR="mmpbsa1.mdp"
        MDP_SASA="sasa.mdp"
        MDP_SAV="sav.mdp"
        MDP_WCA="wca.mdp"

        if [[ -f "$TRJ_FILE" && -f "$TPR_FILE" && -f "$NDX_FILE" ]]; then
            echo "Processing files: $TRJ_FILE, $TPR_FILE, $NDX_FILE"

            # potential energy in Vacuum
            echo 9 10 | g_mmpbsa -f "$TRJ_FILE" -s "$TPR_FILE" -n "$NDX_FILE" -b 99950 -e
100000 -pdie 300 -decomp -incl_14

            mv energy_MM.xvg "energy_${name}m${model}_MM.xvg"
            mv contrib_MM.dat "contrib_${name}m${model}_MM.dat"

            # polar solvation energy
            echo 9 10 | g_mmpbsa -f "$TRJ_FILE" -s "$TPR_FILE" -n "$NDX_FILE" -i
"$MDP_POLAR" -b 99950 -e 100000 -nomme -pbsa -decomp
            mv polar.xvg "polar_${name}m${model}.xvg"
            mv contrib_pol.dat "contrib_pol_${name}m${model}.dat"

            # non-polar solvation energy (SASA-only model)
            echo 9 10 | g_mmpbsa -f "$TRJ_FILE" -s "$TPR_FILE" -n "$NDX_FILE" -i
"$MDP_SASA" -b 99950 -e 100000 -nomme -pbsa -decomp -apol sasa.xvg -apcon sasa_contrib.dat
            mv sasa.xvg "sasa_${name}m${model}.xvg"
```

```
            mv sasa_contrib.dat "sasa_contrib_${name}m${model}.dat"

            # non-polar solvation energy (SAV-only model)
            echo 9 10 | g_mmpbsa -f "$TRJ_FILE" -s "$TPR_FILE" -n "$NDX_FILE" -i "$MDP_SAV"
-b 99950 -e 100000 -nomme -pbsa -decomp -apol sav.xvg -apcon sav_contrib.dat
            mv sav.xvg "sav_${name}m${model}.xvg"
            mv sav_contrib.dat "sav_contrib_${name}m${model}.dat"

            # non-polar solvation energy (WCA model)
            echo 9 10 | g_mmpbsa -f "$TRJ_FILE" -s "$TPR_FILE" -n "$NDX_FILE" -i "$MDP_WCA"
-b 99950 -e 100000 -nomme -pbsa -decomp -apol wca.xvg -apcon wca_contrib.dat
            mv wca.xvg "wca_${name}m${model}.xvg"
            mv wca_contrib.dat "wca_contrib_${name}m${model}.dat"

            # Perform MmPbSaStat analysis
            python MmPbSaStat.py -bs -nbs 2000 -m "energy_${name}m${model}_MM.xvg" -p
"polar_${name}m${model}.xvg" -a "sasa_${name}m${model}.xvg"
            mv full_energy.dat "full_energy_${name}m${model}.dat"
            mv summary_energy.dat "summary_energy_${name}m${model}.dat"

            # Remove PBC
            echo 1 0 | gmx_mpi trjconv -s ${TPR_FILE} -f ${TRJ_FILE} -o md_0_1_noPBC.xtc -
pbc nojump -center


            echo 1 1 | gmx_mpi gyrate -s ${TPR_FILE} -f md_0_1_noPBC.xtc -o gyrate.xvg


#############################################################################################
#########
            # Chain A analysis

#############################################################################################
#########
            echo 9 9 | gmx_mpi rms -s ${TPR_FILE} -f md_0_1_noPBC.xtc -n ${OUTPUT_NDX} -o
rmsdChainA_${name}m${model}.xvg -tu ns

            echo 9 | gmx_mpi gyrate -s ${TPR_FILE} -f md_0_1_noPBC.xtc -n ${OUTPUT_NDX} -o
rgChainA_${name}m${model}.xvg

            echo 9 9 | gmx_mpi covar -s ${TPR_FILE} -f md_0_1_noPBC.xtc -n ${OUTPUT_NDX} -o
motionsChainA_${name}m${model}.xvg


            echo 9 9 | gmx_mpi anaeig -v eigenvec.trr -f md_0_1_noPBC.xtc -s ${TPR_FILE} -n
${OUTPUT_NDX} -comp eigencompChainA_${name}m${model}.xvg -rmsf
rmsfChainA_${name}m${model}.xvg -xvg "xmgrace" -max 24 -proj
motionsChainA_${name}m${model}.xvg -2d 2denvChainA_${name}m${model}.xvg -b 20 -tu ns -first
1 -last 2

            echo 9 9 | gmx_mpi anaeig -v eigenvec.trr -f md_0_1_noPBC.xtc -s ${TPR_FILE} -n
${OUTPUT_NDX} -comp eigencompChainA_${name}m${model}.xvg -rmsf
rmsfChainA_${name}m${model}.xvg -xvg "xmgrace" -max 24 -proj
motionsChainA_${name}m${model}.xvg -2d 3denvChainA_${name}m${model}.xvg -b 20 -tu ns -first
1 -last 3
```

```
#########################################################################################################
#########
            # Chain B analysis

#########################################################################################################
#########

            echo 10 10 | gmx_mpi rms -s ${TPR_FILE} -f md_0_1_noPBC.xtc -n ${OUTPUT_NDX} -o
rmsdChainB_${name}m${model}.xvg -tu ns

            echo 10 | gmx_mpi gyrate -s ${TPR_FILE} -f md_0_1_noPBC.xtc -n ${OUTPUT_NDX} -o
rgChainB_${name}m${model}.xvg

            echo 10 10 | gmx_mpi covar -s ${TPR_FILE} -f md_0_1_noPBC.xtc -n ${OUTPUT_NDX}
-o motionsChainB_${name}m${model}.xvg

            echo 10 10 | gmx_mpi anaeig -v eigenvec.trr -f md_0_1_noPBC.xtc -s ${TPR_FILE}
-n ${OUTPUT_NDX} -comp eigencompChainB_${name}m${model}.xvg -rmsf
rmsfChainB_${name}m${model}.xvg -xvg "xmgrace" -max 24 -proj
motionsChainB_${name}m${model}.xvg -2d 2denvChainB_${name}m${model}.xvg -b 20 -tu ns -first
1 -last 2

            echo 10 10 | gmx_mpi anaeig -v eigenvec.trr -f md_0_1_noPBC.xtc -s ${TPR_FILE}
-n ${OUTPUT_NDX} -comp eigencompChainB_${name}m${model}.xvg -rmsf
rmsfChainB_${name}m${model}.xvg -xvg "xmgrace" -max 24 -proj
motionsChainB_${name}m${model}.xvg -2d 3denvChainB_${name}m${model}.xvg -b 20 -tu ns -first
1 -last 3

        else
            echo "Skipping missing files: $TRJ_FILE, $TPR_FILE, or $NDX_FILE"
        fi
    done
done

echo "All operations completed successfully."
```

C4: This is python script for visualizing post MD and MMGBSA results obtained using C3.

```
import os
import pandas as pd
import seaborn as sns
from scipy.cluster.hierarchy import linkage, leaves_list
```

```
import matplotlib.pyplot as plt

models = ['m1', 'm2', 'm3', 'm4', 'm5']
names = ['577', '274', '734', '279']

class DataExtractor:
    @staticmethod
    def extract_rmsd_data(model_names):
        model_data_chainA = {}
        model_data_chainB = {}

        for name in model_names:
            for model_number in models:
                file_path_chainA = f'rmsdChainA_{name}{model_number}.xvg'
                file_path_chainB = f'rmsdChainB_{name}{model_number}.xvg'

                if os.path.exists(file_path_chainA) and os.path.exists(file_path_chainB):
                    time_chainA, rmsd_chainA =
DataExtractor.extract_single_file(file_path_chainA)
                    time_chainB, rmsd_chainB =
DataExtractor.extract_single_file(file_path_chainB)

                    if name not in model_data_chainA:
                        model_data_chainA[name] = {}
                        model_data_chainB[name] = {}

                    model_data_chainA[name][model_number] = (time_chainA, rmsd_chainA)
                    model_data_chainB[name][model_number] = (time_chainB, rmsd_chainB)
                else:
                    print(f"Warning: Files not found for model {name} {model_number}.
Skipping.")

        return model_data_chainA, model_data_chainB

    @staticmethod
    def extract_single_file(file_path):
        time1 = []
        rmsd = []

        try:
            with open(file_path, 'r') as file:
                lines = file.readlines()
                for line in lines:
                    if line.startswith('@') or line.startswith('#'):
                        continue
                    data = line.split()
                    time1.append(float(data[0]))
                    rmsd.append(float(data[1]))
        except Exception as e:
            print(f"Error extracting data from {file_path}: {str(e)}")

        return time1, rmsd

class DataPlotter:
    @staticmethod
    def plot_rmsd(model_data_chainA, model_data_chainB):
        fig, axs = plt.subplots(1, len(names), figsize=(20, 5), sharey=True)
```

```
        fig.suptitle('Root Mean Square Deviation (RMSD) over Time', y=1.02)

        for i, name in enumerate(names):
            for model_number in models:
                if name in model_data_chainA and name in model_data_chainB:
                    if model_number in model_data_chainA[name] and model_number in
model_data_chainB[name]:
                        time_chainA, rmsd_chainA = model_data_chainA[name][model_number]
                        time_chainB, rmsd_chainB = model_data_chainB[name][model_number]

                        axs[i].plot(time_chainA, rmsd_chainA, label=f'Chain A Model
{model_number}')
                        axs[i].plot(time_chainB, rmsd_chainB, label=f'Chain B Model
{model_number}')
                        axs[i].set_title(f'{name}')

        for ax in axs:
            ax.set_xlabel('Time (ns)')
            ax.legend(loc='upper right')

        axs[0].set_ylabel('RMSD (nm)')
        plt.tight_layout()
        plt.subplots_adjust(top=0.8)
        plt.savefig('combined_rmsd_chainA_chainB_all.png')
        plt.show()



class RMSDAnalysis:
    def __init__(self, file_path, time_step_ns=1):
        self.file_path = file_path
        self.time_step_ns = time_step_ns
        self.data = self.read_xvg()
        if self.data is not None:
            self.time_ns = self.data[:, 0]
            self.rmsd = self.data[:, 1]
        else:
            self.time_ns = None
            self.rmsd = None

    def read_xvg(self):
        data = []
        try:
            with open(self.file_path, 'r') as f:
                for line in f:
                    if not line.startswith(('#', '@')):
                        data.append([float(x) for x in line.split()])
            return np.array(data)
        except FileNotFoundError:
            print(f"File {self.file_path} not found.")
            return None

    def detect_changepoints(self, penalty=12):
        if self.rmsd is None:
            return []
        model = "l2"
        algo = rpt.Pelt(model=model).fit(self.rmsd)
```

```
        changepoints = algo.predict(pen=penalty)
        return changepoints

    def plot_rmsd_with_changepoints(self, ax, changepoints, max_time_ns=100):
        if self.time_ns is None or self.rmsd is None:
            return

        changepoints_ns = [cp * self.time_step_ns for cp in changepoints]
        degree_of_stabilization = len(changepoints_ns)

        ax.plot(self.time_ns, self.rmsd, color='blue', label='RMSD')
        #for cp in changepoints_ns:
            #if cp <= max_time_ns:
                #ax.axvline(x=cp, color='r', linestyle='--', label='Detected Changepoint')
        ax.set_xlabel('Time (ns)')
        ax.set_ylabel('RMSD (Å)')

        legend_text = f'Complex; (τ): {degree_of_stabilization}'
        ax.legend([legend_text])
        ax.set_xlim(-5, max_time_ns)

def analyze_all_files():
    models = ['m1', 'm2', 'm3', 'm4', 'm5']
    names = ['577', '274', '734', '279']
    file_template = 'rmsd_{}{}.xvg'

    fig, axs = plt.subplots(len(models), len(names), figsize=(20, 15), sharex=True,
sharey=True)
    #fig.suptitle('RMSD Analysis of Different Complexes and Models', y=0.92)

    for i, model in enumerate(models):
        for j, name in enumerate(names):
            file_path = file_template.format(name, model)
            analysis = RMSDAnalysis(file_path)
            changepoints = analysis.detect_changepoints()

            if analysis.time_ns is not None and analysis.rmsd is not None:
                analysis.plot_rmsd_with_changepoints(axs[i, j], changepoints)
            else:
                axs[i, j].plot([], [])
                axs[i, j].text(0.5, 0.5, 'File Not Found', horizontalalignment='center',
verticalalignment='center', transform=axs[i, j].transAxes)

            axs[i, j].set_title(f'{name} - {model}')
            axs[i, j].set_xlim(-5, 100)
            axs[i, j].set_ylim(0, 5)
            axs[i, j].set_xlabel('Time (ns)')
            axs[i, j].set_ylabel('RMSD (nm)')

    plt.tight_layout()
    plt.subplots_adjust(top=0.95)
    plt.show()

analyze_all_files()
```

```
class RMSFAnalyzer:
    def __init__(self, file_path):
        self.file_path = file_path
        self.atoms = []
        self.rmsf = []
        self.read_xvg(file_path)

    def read_xvg(self, file_path):
        with open(file_path, 'r') as file:
            for line in file:
                if line.startswith('@') or line.startswith('#'):
                    continue
                if line.startswith('&'):
                    break
                parts = line.split()
                if len(parts) >= 2:
                    self.atoms.append(float(parts[0]))
                    self.rmsf.append(float(parts[1]))

def collect_rmsf_data():
    models = ['m1', 'm2', 'm3', 'm4', 'm5']
    names = ['577', '274', '734', '279']
    file_template_chainA = 'rmsfChainA_{}{}.xvg'

    rmsf_data = {}

    for model in models:
        for name in names:
            file_path_chainA = file_template_chainA.format(name, model)
            if os.path.exists(file_path_chainA):
                analyzer = RMSFAnalyzer(file_path_chainA)
                rmsf_data[f'{name}_{model}'] = analyzer.rmsf

    return rmsf_data


class RMSD_chains:
    @staticmethod
    def plot_rmsd(model_data_chainA, model_data_chainB):
        fig, axs = plt.subplots(len(models), len(names), figsize=(20, 15), sharex=True,
sharey=True)
        fig.suptitle('Root Mean Square Deviation (RMSD) over Time', y=0.92)

        for i, model_number in enumerate(models):
            for j, name in enumerate(names):
                if name in model_data_chainA and name in model_data_chainB:
                    if model_number in model_data_chainA[name] and model_number in
model_data_chainB[name]:
                        time_chainA, rmsd_chainA = model_data_chainA[name][model_number]
                        time_chainB, rmsd_chainB = model_data_chainB[name][model_number]

                        axs[i, j].plot(time_chainA, rmsd_chainA, label=f'Chain A')
                        axs[i, j].plot(time_chainB, rmsd_chainB, label=f'Chain B')
                        axs[i, j].set_title(f'Model {model_number} - {name}')

        for ax in axs[-1, :]:
            ax.set_xlabel('Time (ns)')
```

```
        for ax in axs[:, 0]:
            ax.set_ylabel('RMSD (nm)')

        plt.tight_layout()
        plt.subplots_adjust(top=0.85)
        plt.legend(loc='upper right')
        plt.savefig('combined_rmsd_chainA_chainB_matrix.png')
        plt.show()


# data_extractor = DataExtractor()
# data_plotter = RMSD_chains()


# model_data_chainA, model_data_chainB = data_extractor.extract_rmsd_data(names)


# data_plotter.plot_rmsd(model_data_chainA, model_data_chainB)


#def generate_clustered_heatmap(rmsf_data):
    #rmsf_df = pd.DataFrame(rmsf_data)

    #corr_matrix = rmsf_df.corr()

    #sns.clustermap(corr_matrix, method='average', cmap='coolwarm', annot=False,
figsize=(10, 10))
    #plt.savefig('clustered_correlation_heatmap_rmsf_chainA.png')
    #plt.show()

class GyrationAnalyzer:
    def __init__(self, file_path):
        self.file_path = file_path
        self.time = []
        self.rg_total = []
        self.read_xvg()

    def read_xvg(self):
        with open(self.file_path, 'r') as file:
            for line in file:
                if line.startswith('@') or line.startswith('#'):
                    continue
                parts = line.split()
                self.time.append(float(parts[0]))
                self.rg_total.append(float(parts[1]))

    def plot_rg(self, ax, title):
        ax.plot(self.time, self.rg_total, label="Rg Total")
        ax.set_xlabel('Time (ps)')
        ax.set_ylabel('Rg (nm)')
        ax.set_title(title)
        ax.legend()
        ax.grid(False)

class PCA_analysis:
    def read_xvg(filename):
```

```
        with open(filename, 'r') as file:
            lines = file.readlines()


        data_lines = [line.strip() for line in lines if not line.startswith(('@', '#')) and
line.strip()]

        # Parse the data into a numpy array,
        data = []
        for line in data_lines:
            try:
                values = list(map(float, line.split()))
                data.append(values)
            except ValueError:

                continue

        data = np.array(data)
        return data

    # split data into five parts
    def split_data(data):
        split_data = np.array_split(data, 5)
        return split_data

    # Plot
    def plot_data(ax, data_parts, title):
        colors = ['r', 'g', 'k', 'b', 'm']

        for i, part in enumerate(data_parts):
            ax.scatter(part[:, 0], part[:, 1], label=f'Part {i+1}', color=colors[i], s=20)

        ax.set_title(title)
        ax.set_xlabel("Projection on Eigenvector 1")
        ax.set_ylabel("Projection on Eigenvector 2")
        ax.grid(False)
        ax.set_xlim(-100, 100)
        ax.set_ylim(-90, 85)
        ax.legend()

        # for correlation matrics
        def analyze_all_files():
            models = ['m1', 'm2', 'm3', 'm4', 'm5']
            names = ['577', '274', '734', '279']
            file_template = '2denv_{}{}.xvg'

            all_data = {}

            fig, axs = plt.subplots(len(models), len(names), figsize=(20, 15), sharex=True,
sharey=True)

            for i, model in enumerate(models):
                for j, name in enumerate(names):
                    file_path = file_template.format(name, model)
                    key = f'{name}_{model}'
                    if os.path.exists(file_path):
                        data = read_xvg(file_path)
```

```
                        all_data[key] = data[:, 1]
                        data_parts = split_data(data)
                        plot_data(axs[i, j], data_parts, f'{name} - {model}')
                    else:
                        axs[i, j].set_title(f'{name} - {model}')
                        axs[i, j].text(0.5, 0.5, 'File Not Found',
horizontalalignment='center', verticalalignment='center', transform=axs[i, j].transAxes)
                        axs[i, j].set_xlim(-100, 100)
                        axs[i, j].set_ylim(-90, 85)
                        axs[i, j].set_xlabel("Projection on Eigenvector 1")
                        axs[i, j].set_ylabel("Projection on Eigenvector 2")
                        axs[i, j].grid(False)

            plt.tight_layout()
            plt.subplots_adjust(top=0.95)
            plt.savefig('combined_2d_projection_trajectory_matrix.png')
            plt.show()


            all_data_df = pd.DataFrame(all_data)
            all_data_df.to_csv('PCA_complexes.csv')

            corr_matrix = all_data_df.corr()

            linkage_matrix = linkage(corr_matrix, method='ward')
            ordered_indices = leaves_list(linkage_matrix)
            ordered_corr_matrix = corr_matrix.iloc[ordered_indices, ordered_indices]

            sns.clustermap(ordered_corr_matrix, cmap='viridis', annot=False, fmt=".2f",
linewidths=0, figsize=(10, 8))
            plt.title('Hierarchically Clustered Correlation Heatmap')
            plt.savefig('correlation_heatmap_with_dendrogram.png')
            plt.show()

            corr_matrix.to_csv('correlation_matrix.csv')



class auto_corrAnalysis:
    def __init__(self, file_path, time_step_ns=1):
        self.file_path = file_path
        self.time_step_ns = time_step_ns
        self.data = self.read_xvg()
        if self.data is not None:
            self.time_ns = self.data[:, 0]
            self.rmsd = self.data[:, 1]
        else:
            self.time_ns = None
            self.rmsd = None

    def read_xvg(self):
        data = []
        try:
            with open(self.file_path, 'r') as f:
                for line in f:
                    if not line.startswith(('#', '@')):
                        data.append([float(x) for x in line.split()])
```

```
            return np.array(data)
        except FileNotFoundError:
            print(f"File {self.file_path} not found.")
            return None

    def autocorr(self, x):
        n = len(x)
        mean = np.mean(x)
        var = np.var(x)
        autocorr_func = np.correlate(x - mean, x - mean, mode='full') / (var * n)
        return autocorr_func[n-1:]

    def calculate_autocorrelation_time(self, ax):
        if self.rmsd is None:
            print("RMSD data is not available.")
            return

        autocorr_func = self.autocorr(self.rmsd)
        zero_crossings = np.where(np.diff(np.sign(autocorr_func)))[0]

        if len(zero_crossings) > 1:
            second_zero_crossing = zero_crossings[1]
            correlation_time_second = 1 + 2 * np.sum(autocorr_func[1:second_zero_crossing +
1])
        else:
            correlation_time_second = 1 + 2 * np.sum(autocorr_func[1:])

        correlation_time_ns = correlation_time_second * self.time_step_ns

        lags = np.arange(len(autocorr_func))
        ax.plot(lags, autocorr_func, marker='o', linestyle='-')
        ax.set_title(f'Autocorr. of {os.path.basename(self.file_path).replace(".xvg",
"")}')
        ax.set_xlabel('Lag (k)')
        ax.set_ylabel('Autocorrelation')
        ax.grid(True)

        print(f"Autocorrelation time (ns): {correlation_time_ns} for
{os.path.basename(self.file_path).replace('.xvg', '')}")
        return correlation_time_ns

def analyze_auto_all_files():
    models = ['m1', 'm2', 'm3', 'm4', 'm5']
    names = ['577', '274', '734', '279']
    file_template = 'rmsd_{}{}.xvg'

    fig, axs = plt.subplots(5, 4, figsize=(20, 25), sharex=True, sharey=True)
    #fig.suptitle('Autocorrelation Functions of RMSD Data', y=0.92)

    plot_idx = 0
    for model in models:
        for name in names:
            row = plot_idx // 4
            col = plot_idx % 4
            file_path = file_template.format(name, model)
            analysis = auto_corrAnalysis(file_path)
            if analysis.time_ns is not None and analysis.rmsd is not None:
```

```
                analysis.calculate_autocorrelation_time(axs[row, col])
            else:
                axs[row, col].plot([], [])
                axs[row, col].text(0.5, 0.5, 'File Not Found',
horizontalalignment='center', verticalalignment='center', transform=axs[row,
col].transAxes)
                axs[row, col].set_title(f'{name} - {model}')
            plot_idx += 1

    plt.tight_layout()
    plt.subplots_adjust(top=0.95)
    plt.show()

analyze_auto_all_files()


rmsf_data = collect_rmsf_data()
generate_clustered_heatmap(rmsf_data)
rmsd_df = pd.read_csv('rmsd_data.csv')


corr_matrix = rmsd_df.corr()


linkage_matrix = linkage(corr_matrix, method='ward')
ordered_indices = leaves_list(linkage_matrix)
ordered_corr_matrix = corr_matrix.iloc[ordered_indices, ordered_indices]


plt.Figure(figsize=(10, 8))
sns.heatmap(ordered_corr_matrix, cmap='coolwarm', annot=False, fmt=".2f")
plt.show()
+
data_extractor = DataExtractor()
data_plotter = DataPlotter()

model_data_chainA, model_data_chainB = data_extractor.extract_rmsd_data(names)

data_plotter.plot_rmsd(model_data_chainA, model_data_chainB)
```

----------*End*----------