ADAPTING AND ENHANCING HYBRID COMPUTATIONAL METHODS FOR RNA SECONDARY STRUCTURE PREDICTION

RA'ED MOHAMMAD ALI AL-KHATIB

UNIVERSITI SAINS MALAYSIA

2011

ADAPTING AND ENHANCING HYBRID COMPUTATIONAL METHODS FOR RNA SECONDARY STRUCTURE PREDICTION

by

RA'ED MOHAMMAD ALI AL-KHATIB

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

December 2011

ACKNOWLEDGEMENTS

All praise and thanks be to Allah, the Lord of the world, for giving me the energy and the talent to finish my research. He guides me and grants me success in my life. I cannot count His bounties on me.

I would also, like to express my deepest gratitude and appreciation to my main supervisor, Assoc. Prof. Dr. Nur'Aini Abdul Rashid, for her invaluable encouragement and guidance. Her support and comments have provided me with adequate strength that enabled me to undertake this challenge. I am also grateful to my co-supervisor, Prof. Dr. Rosni Abdullah for her comments and guidance throughout the period of my study and conducting this research.

In addition, I would like to thank the academic and technical support staff of the School of Computer Sciences, USM, who provided me with the facilities needed to conduct my research. I also wish to extend my gratitude to Universiti Sains Malaysia for granting me a fellowship to pursue my Ph.D.

Last but not least, my sincere thanks to my family; my father, my wives, my children, and my brothers and sisters who have always shown their faithful support during my study. I appreciate their everlasting patience during the long period of my study. Special thanks to my cousin Dr Sami for his continuous support in my study since bachelor's degree and to my friends Alfian Abdul Halin, Alireza Taghizadeh and Dr. Mohammed Al-Betar for their valuable comments, suggestions, and language editing. I sincerely thank my friends: Dr. Mohammed S. AbuRub, Mozaherul Hoque Abul Hasnat, Ibrahim Umar, Muhannad Abu-Hashem, Najihah Binti Ibrahim, Dr. Khalid Jaber, Dr. Hesham Bahamish, Dr. Ali Kattan and Dr. Fazilah Osman for their help and support. Finally, I am very thankful to all my friends and family who always make dua'a (supplications to Allah) for me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

Page

LIST OF ABBREVIATIONS

ABC Artificial Bee Colony

ACS Ant Colony System

AI Artificial Intelligence

API Application Program Interface

ANN Artificial Neural Network

BCPA Bee Collecting Pollen Algorithm

CBR Case-Based Reasoning

DNA Deoxyribonucleic Acid

ILM Iterated Loop Matching

GA Genetic Algorithm

HSA Harmony Search Algorithm

MBO Marriage in Honey-Bees Optimization

MC Monte Carlo

MFE Minimum Free Energy

MIMD Multiple Instruction, Multiple Data

MISD Multiple Instruction, Single Data

mRNA Messenger RNA

NMR Nuclear Magnetic Resonance

NP Non polynomial

OpenMP Open Multi-Processing

PDB Protein Data Bank

PCGSs parallel communicating grammar systems

PSO Particle Swarm Optimization

RNA Ribonucleic Acid

rRNA Ribosomal RNA

HPRna Honey Production Algorithm for RNA Secondary Structure Prediction

SIMD Single Instruction, Multiple Data

SISD Single Instruction, Single Data

SA Simulating Annealing

SI Swarm Intelligence

TAGs Tree Adjoining Grammar Algorithms

tRNA Transfer RNA

PENGADAPTASIAN DAN PENAMBAHBAIKAN KAEDAH-KAEDAH PENGKOMPUTERAN HIBRID UNTUK RAMALAN STRUKTUR SEKUNDER RNA

ABSTRAK

Struktur sekunder RNA berpseudoknot digunakan secara meluas bagi mengesan struktur tertier RNA yang merupakan kunci untuk memahami fungsi-fungsi RNA dan pelbagai kegunaannya dalam penghasilan ubatan untuk penyakit viral. Kaedah-kaedah eksperimen untuk menentukan struktur tertier RNA mengambil masa yang lama dan menjemukan. Oleh itu, pendekatan pengkomputeran ramalan adalah diperlukan. Ramalan struktur sekunder RNA berpseudoknot yang paling tepat dan stabil dari segi tenaga telah dibuktikan sebagai suatu permasalahan *NP-hard*. Tesis ini membentangkan suatu kaedah hibrid untuk meramal struktur sekunder RNA berpseudoknot dengan menggabungkan kaedah-kaedah pengesanan dengan algoritma-algoritma pengaturcaraan dinamik. Kaedah hibrid ini ditambahbaik dengan menggunakan teknik penaakulan berdasarkan kes. Tiga kaedah berbeza dicadangkan: (i) kaedah diinspirasi kecerdasan kawanan (HPRna); (ii) kaedah hibrid adaptif (MSeeker); dan (iii) kaedah selari pantas (FGTSeeker), di mana setiap kaedah merupakan penambahbaikan kepada kaedah-kaedah sebelumnya. Kaedah-kaedah ramalan yang dicadangkan telah dinilai terhadap kaedah-kaedah ramalan sedia ada menggunakan struktur-struktur asli sebenar sebagai faktor utama perbandingan. Keputusan menunjukkan bahawa ketiga-tiga kaedah yang dicadangkan memperoleh struktur sekunder RNA berpseudoknot yang lebih tepat dengan prestasi yang lebih baik, terutamanya dalam meramal turutan-turutan panjang.

ADAPTING AND ENHANCING HYBRID COMPUTATIONAL METHODS FOR RNA SECONDARY STRUCTURE PREDICTION

ABSTRACT

The secondary structure of RNA with pseudoknots is widely utilized for tracing the RNA tertiary structure, which is a key to understanding the functions of the RNAs and their useful roles in developing drugs for viral diseases. Experimental methods for determining RNA tertiary structure are time consuming and tedious. Therefore, predictive computational approaches are required. Predicting the most accurate and energy-stable pseudoknot RNA secondary structure has been proven to be an NP-hard problem. This thesis presents a hybrid method to predict the RNA pseudoknot secondary structures by combining detection methods with dynamic programming algorithms. This hybrid method is further enhanced by adopting the case-based reasoning (CBR) technique. Three different methods are proposed, (i) Bio-inspired swarm intelligence method (HPRna); (ii) Adaptive hybrid method (MSeeker); and (iii) Fast parallel method (FGTSeeker), where each is an improvement to the previous method. The proposed prediction methods were evaluated against other existing prediction methods using the real native structures as the main factor of comparison. Results show that the three proposed methods obtained more accurate pseudoknotted RNA secondary structures with better performance, especially in predicting long sequences.

CHAPTER 1

INTRODUCTION

1.1 Background

Bioinformatics is a new discipline resulting from the combination of two science fields: *Computer Science* and *Biology*. This discipline was coined by Hogeweg (1978) and has been rapidly growing in recent years. Nowadays, bioinformatics has become the foundation in ongoing biomolecular research study (Counsell, 2003; Whitfield et al., 2006).

Basically, bioinformatics research assists biologists in expediting the biological processes through the use of advanced computer algorithms to collect, accumulate, store, analyze and integrate biological data and genetic macromolecules; such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or proteins (Nair, 2007). The DNA contains directions on how to build other cell components, such as proteins and RNA molecules. The RNA on the other hand is a type of nucleic acid that provides a mechanism to copy the genetic information from DNA for protein synthesis (Brenner et al., 1961; Halfmann and Lindquist, 2010).

Numerous efforts have been undertaken by bioinformaticians to address the requirements in many related problems such as biomolecule sequence alignment, gene therapy and finding, gene expression control and drug design and development. Another crucial issue is the study of inferring the various useful RNA functions, especially by predicting the structures of known primary RNA sequences. It is worth stating that the RNA primary structures can easily be determined by gene sequencing techniques in an experimental setting (Ellis et al., 1992). However, these primary structures cannot give sufficient information pertaining to the important

RNA functions (Beebee and Rowe, 2008). According to (Blazewicz et al., 2005), the structure with the most amount of information is the RNA tertiary structure. However, this structure can be obtained and scrutinized by identifying the RNA secondary structure (Nebel, 2003; Tsang and Wiese, 2010).

Consequently, determining the RNA secondary structure is deemed key towards building the tertiary structure and to understand the various functions and roles of RNA molecules (Tinoco et al., 1999). There are only small numbers of known RNA secondary structures compared to the colossal amounts of discovered primary sequences. There is hence a great gap in the research pertaining to the prediction of RNA structures from given primary sequences. Furthermore, this opens the door for the use of computational methods as these methods can potentially be faster compared to structure prediction via experimental methods (Tinoco et al., 1999; Gee et al., 2006).

The field of RNA secondary structure prediction via computational methods has become one of the most active research fields. Thus, this thesis will focus on computationally solving RNA secondary structure prediction, which has been proven to be an NP-hard problem (Lyngso and Pedersen, 2000b; Akutsu, 2000). Recently, many predictive computational approaches have been suggested. Among them are dynamic programming (DP) algorithms such as pknotsRG (Reeder and Giegerich, 2004). Heuristic-based methods were also proposed such as HotKnots (Ren et al., 2005). Lately, heuristic-based methods have been successful in solving the RNA secondary structure prediction problems. Compared to DP methods, which suffer from recursion and drawback that get more complexity when the input RNA is long, heuristic-based methods are more advantageous since they perform prediction in many separate stages. Each stage contains several steps where the input RNA sequence is divided into sub-elements and parts. This results in a more efficient prediction process that executes more quickly with less memory consumption (compared to the DP algorithms). Due to this, the work in this the-

sis will focus on heuristic-based methods, which is further specified to deal with secondary structure of RNA with pseudoknots class.

The proposed approach is basically a novel hybrid model, which combines a KnotSeeker detection method with dynamic programming algorithm. This combination works on the basis of global optimization, which is further enhanced by using the case-based reasoning (CBR) technique as a local optimization method.

1.2 Motivations and Research Problems

The main motivation for building the RNA structure is to understand its various functions. These functions are vital to know the RNA's therapeutic applications such as designing antiviral drugs for malignant diseases (cancer) and for AIDS (Anderson and Kedersha, 2009; Karagiannis and El-Osta, 2005; Eguchi et al., 2009). The exponential growth rate of RNA primary sequence data has motivated bioinformatics researchers to propose efficient approaches that predict the RNA secondary structure for the purpose of understanding their biological functions (Mahen et al., 2010). However, there are many difficulties in determining the pseudoknotted RNA secondary structures. This is worsened by the fact that the prediction process is proven to be an NP-hard problem (Lyngso and Pedersen, 2000b; Akutsu, 2000). As a result, there is a big gap between the colossal number of known RNA sequences and the quantity of the known RNA structures. Figure 1.1[1] illustrates the growth of the biological data in GenBank (Benson et al., 2008), where the zoomed-in sub-illustration depicts the growth of experimental structures showing the different growth rate between the huge number of primary sequences and the limited number of known structures.

The two best known biological experimental methods for determining RNA tertiary struc-

[1] Statistical data from: $http://www.ddbj.nig.ac.jp/documents-e.html$

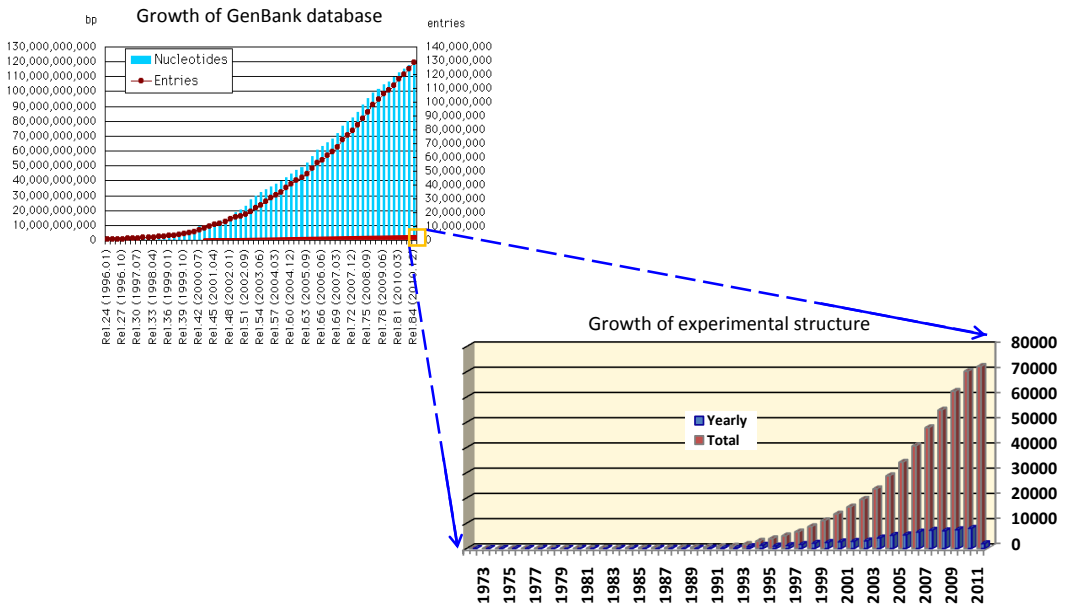
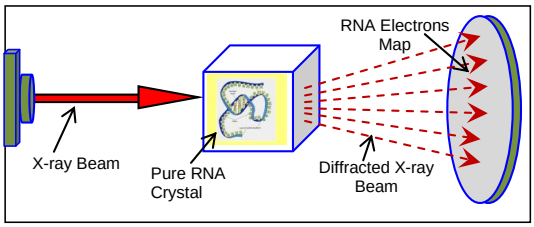


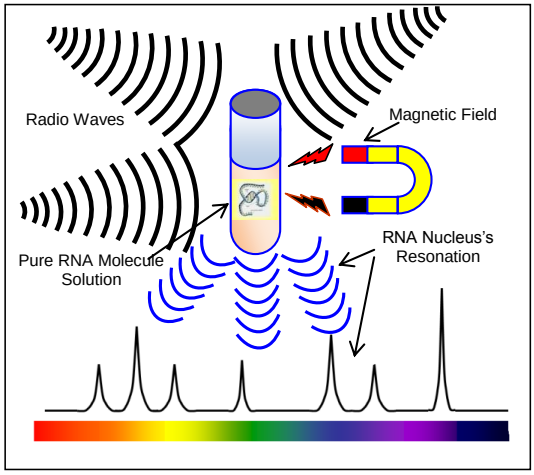
Figure 1.1: Exponential growth of biological data in GenBank and the growth of the known structures (Benson et al., 2008; Golding et al., 2002).

ture are nuclear magnetic resonance (NMR) and X-ray crystallography (XRC), which are shown in Figures 1.2a and b, (Wang et al., 2010; Al-Khatib et al., 2010; Kasprzak et al., 2010). These purification methods however, require lengthy experimental time and special equipments (Cheong et al., 2004; Al-Khatib et al., 2010). Specifically, biological researchers who use the X-ray method (Figure 1.2a) face some serious constraints. For this method to be effective, sufficient RNA pure crystal is required in the diffraction process. However, not all RNA organic molecules can be put in crystal easily. Furthermore, the X-ray beam diffracts when it hits the electrons around the RNA nuclei. This gives the electrons map of the target RNA instead of the real structure and causes the final RNA structure prediction to be less accurate. In the NMR physical method, the resonation of the RNA nuclei is done by bombarding the fixed RNA molecule with radio waves from thousands of different angles, which is an incredibly time-consuming process (see Figure 1.2b).

According to the above explanation, many factors need to be considered when running biological experimental methods. In order to decrease the difficulty in performing such experimental methods, the tertiary structure of RNA molecules can also be scrutinized and ob-



(a) X-ray crystallography method.



(b) Nuclear magnetic resonance (NMR) Method.

Figure 1.2: Experimental methods for RNA tertiary structures determination.

tained much faster by predicting their secondary structures (Bindewald et al., 2008). Therefore, bioinformatics-based computational methods for predicting RNA secondary structure are preferred (Gee et al., 2006).

Predicting the RNA structure by computational methods is faster than determining its structure by experimental methods (Tinoco et al., 1999; Tsang and Wiese, 2010). Generally, the RNA secondary structure is formed quickly. Figure 1.3 shows an example of the computational methods and their tangible contributions to predicting the secondary structure, which assists biologists in scrutinizing the RNA tertiary structure. The most accurate method for predicting the RNA secondary structure is based on the minimum free energy (MFE) model, which is the DP algorithm Mfold (Zuker and Stiegler, 1981; Zuker, 2003).

Although, the pseudoknot RNA secondary structure is difficult to predict and has been proven to be an NP-hard problem (Lyngso and Pedersen, 2000b), it is still important to be

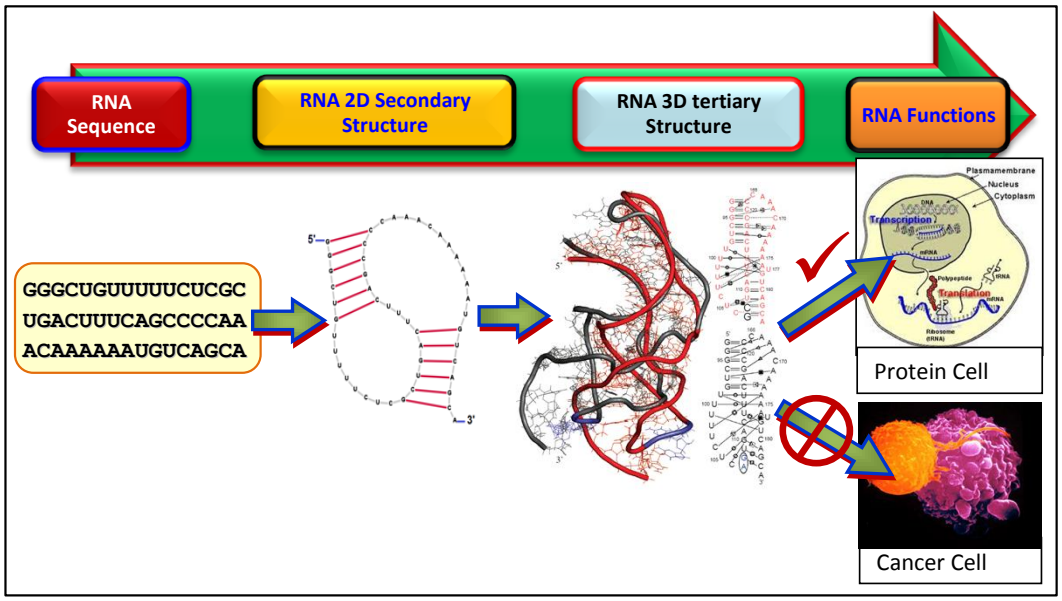


Figure 1.3: RNA structures (primary sequence, secondary structure with pseudoknots, and tertiary structure) of the human telomerase RNA with pseudoknots (Reipa et al., 2007), which includes a wild-type and DKC-mutated pseudoknot structure. The first structure was predicted by HotKnots (Ren et al., 2005) and its image was generated using jViz.Rna (Wiese et al., 2005). The second image is adapted from (Yingling and Shapiro, 2007).

solved computationally. Many DP methods have been proposed to solve the secondary structure of RNA with pseudoknots type such as pknotsRG, which requires $O(n^4)$ for run-time and $O(n^2)$ for space complexity (Reeder and Giegerich, 2004). The DP algorithms give more accurate RNA structural results globally optimizing the predictions of the secondary structure of small RNA input sequences. Particularly, the DP algorithms for pseudoknotted RNA prediction have some drawbacks including recursive difficulties when the length of the input RNA sequences become long. This recursive nature of the DP functional algorithm raises its complexity exponentially. Therefore, the final results of the DP algorithms in predicting the secondary structure of RNA with pseudoknots are less accurate for long RNA sequences. Thus, the DP algorithms are not considered an entirely accurate solution for long RNA primary sequences (Sperschneider and Datta, 2008).

The most prominent methods for solving the difficult problem of secondary structure prediction of RNA with pseudoknots have been based on heuristics or meta-heuristics approaches,

such as HotKnots (Ren et al., 2005), FlexStem (Chen et al., 2008) and DotKnot (Sperschneider and Datta, 2010). The hybrid computational methods, which can be considered as subcategory of the metaheuristic-based methods, provide opportunity to tackle the prediction problem of pseudoknotted RNA secondary structure. These hybrid approaches present balance between the global optimization that combines the strength of detection methods with thermodynamic algorithms, and is further hybridized with CBR as a local optimization method utilizing the power of the similarity-based technique.

CBR is an Artificial Intelligence (AI) methodology that has shown to be successful in problem solving as a local search-based function by using the Nearest Neighbour algorithm (Aamodt and Plaza, 1994; Watson, 1999). Existing state-of-the-art methods have not yet investigated the CBR model for predicting the pseudoknotted RNA secondary structure. The focus of this thesis is to explore and adapt the CBR method towards the development of a new RNA prediction method. The main advantage of the proposed method is to enhance efficiency, performance and accuracy of the final RNA structural results. This research provides a new means for predicting the secondary structure of RNA with pseudoknots in bioinformatics domain.

1.3 Research Questions

This research aims to address and answer the following questions:

1. How can a hybrid algorithm that combines detection and dynamic programming methods be used as a new approach to tackle the secondary structure problem of RNA with pseudoknots?

2. How can the CBR search-based methodology be utilized to enhance the final RNA secondary structural outputs?

3. Can the time to predict accurate secondary structure for long RNA molecules with pseudoknots be reduced by using the parallel methods?

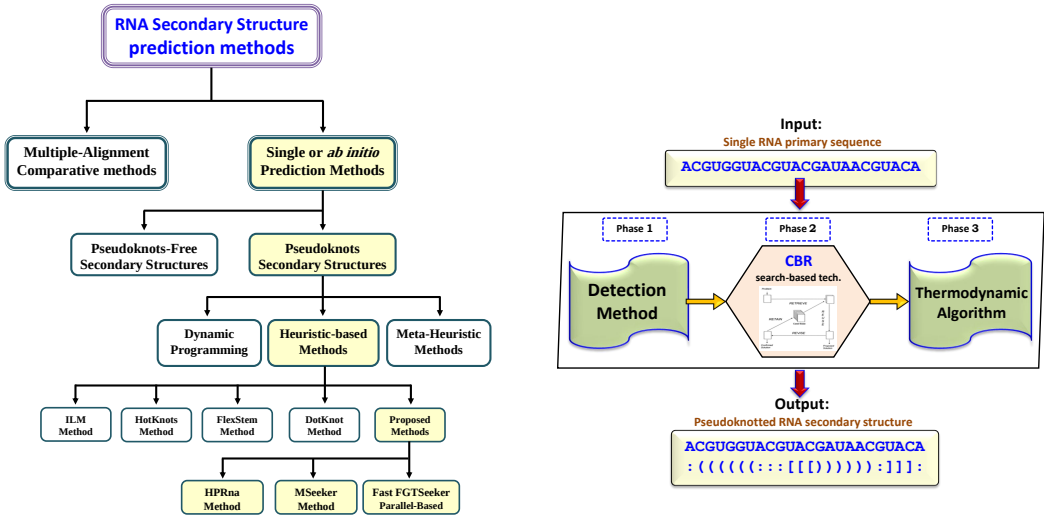
1.4 Research Objectives

The main objective of this dissertation is not merely to propose efficient prediction algorithms for solving pseudoknotted RNA secondary structure prediction problem, but to show that these algorithms can outperform other RNA prediction methods that have already been proposed. Consequently, the new proposed RNA prediction methods are suitably customized to handle the structural problem of long RNA sequences in minimal execution time and with improved accuracy. The objectives of this dissertation are therefore, as follows:

1. To predict the pseudoknotted RNA secondary structure sequences by adapting a bio-inspired swarm intelligence prediction algorithm;
2. To improve and enhance the accuracy of prediction results for RNA secondary structures through the development of a hybrid prediction method; and
3. To reduce the execution time via utilizing parallel-distributed programming models, while also improving the accuracy of the final predicted RNA structure.

1.5 Research Scope

The scope of this research covers the RNA structure prediction problem. RNA structure has four structural levels: primary, secondary, tertiary and quaternary structure. This work focuses on the secondary structure of RNA with pseudoknots. However, there are several groups of computational methods to predict the pseudoknotted RNA secondary structure. Accordingly, this thesis considers the *ab initio* RNA method to predict the secondary structure of RNA with



(a) Flowchart of the research scope.

(b) Flowchart of general hybridized methodology.

Figure 1.4: Scope and general methodology of the research overview.

pseudoknots from a given single sequence. Figure 1.4a, represents the scope of this research that focuses on predicting secondary structure of RNA with pseudoknots. Meanwhile, the *ab initio* RNA structure prediction methods comprise dynamic programming methods, meta-heuristic methods and heuristic-based methods. As illustrated in Figure 1.4b, the research scope of this work concentrates on proposing a new hybrid method that belongs to the group of heuristic-based methods. Particularly, it combines the detection method, CBR technique and thermodynamic algorithm together in this hybrid method, to obtain the final RNA prediction structure.

1.6 Overview of Research Methodology

As explained in previous sections, the main objective of this research is to investigate a hybrid method to predict the secondary structure of RNA with pseudoknots type from a given primary sequence. This section provides an overview of the research methodology used for predicting the secondary structure of RNA with pseudoknots. While the details of this methodology are fully described in Chapter 4. This methodology is presented in order to answer the aforementioned research questions and justifying the research objectives, respectively:

1. For the first objective, the KnotSeeker RNA detection method and UNAFold DP algorithm, are hybridized into HPRna method. This new hybrid method is inspired by swarm-intelligence social behavioral model of honey-bees during nectar collection and honey production (Lu and Zhou, 2008). This research adapts a new bee-inspired algorithm, which is HPRna algorithm, to work as a global optimization model. The advantage of this new bee-inspired algorithm is the adaptation of CBR, which is a prominent AI technique with a history of success in problem solving. The CBR adaptation is meant to enhance the quality of RNA structural results, and to work as a local optimization technique to achieve the final results.

2. For the second objective, two algorithms KnotSeeker and Mfold are combined. This combination is further integrated with CBR to a new predictor termed MSeeker. The MSeeker uses the initial results of RNA pseudoknot elements from the detection algorithm KnotSeeker (Sperschneider and Datta, 2008). Furthermore, a new filtering function is presented to remove the undesirable components that are discovered in KnotSeekers' initial results. Then, the adapted CBR system is used as a local optimization technique for reducing the false positive cases that are discovered in detecting some of the pseudoknot elements. After that, Mfold, which is a more efficient algorithm, predicts the structure of pseudoknot-free parts. Finally, a re-joining function produces the entire predicted target, which is the pseudoknots secondary structure of the input RNA primary sequence.

3. For the final objective, a new version of the parallel-distributed processing framework is proposed to enhance the speed of the hybrid algorithm, which is termed as FGT-Seeker. This parallel version improves the performance by reducing the time of predicting secondary structures for long RNA input sequences. Particularly, this method combines KnotSeeker and GTFold for fast prediction, which works as a global opti-

mization method. This combination is further hybridized with a parallel version of the CBR search-based technique, which works as a local optimization model to enhance the prediction accuracy. Then, this combined parallel method reduce the execution time of prediction process. Its accuracy is further enhanced by adapting more efficient MFE model for pseudoknot-free parts.

In order to evaluate the performance and efficiency of the proposed RNA prediction methods, a series of comprehensive experiments were carried out against other state-of-the-art RNA prediction methods. Figure 1.5 shows the main stages of the research methodology of this thesis, which can be summarized as follows:

Stage 1: Initially, a broad evaluation study for the prominent RNA secondary structural perdition methods was carried out, whose details are covered in Chapters 2 and 3 (i.e. Background and Related Work).

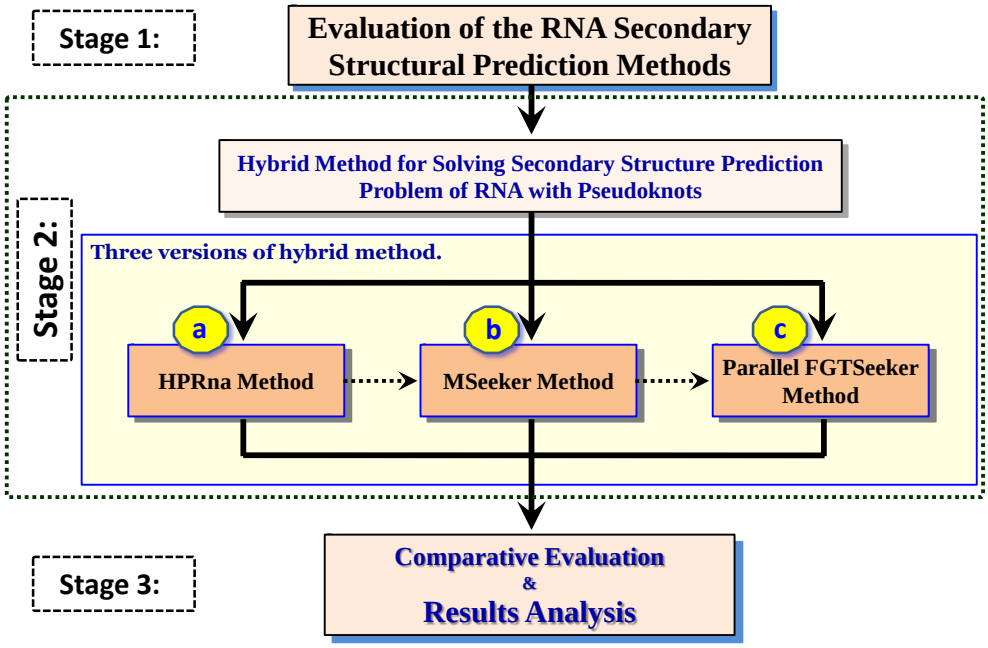


Figure 1.5: Main stages of the research methodology.

Stage 2: In this stage, CBR was adapted from AI and combined with hybrid algorithms to form a new method for predicting secondary structure of RNA with pseudoknots. However, the obtained results from this stage show potential for improvement in order to enhance the accuracy and quality of the algorithm. This improvement is fully explained in the next stage. Thereafter, three different hybrid methods HPRna, MSeeker and FGTSeeker have been sequentially proposed, as depicted at Figure 1.5-Points (a), (b) and (c), respectively. All these methods had a similar objective in mind, which is to solve the secondary structure prediction problem of RNA with pseudoknots. Each new method is an enhancement of the previous one, which is supposed to report improved prediction accuracy. The dotted arrows in Figure 1.5 between the three hybrid methods, denote that the methods were sequentially proposed and each new method is an improvement of the previous one. Furthermore, each new method overcomes the weaknesses of the previous one and produces more accurate RNA structural results. This means that the proposed hybrid methods are developed sequentially, leading to the fulfillment of the all research objectives of this thesis.

Stage 3: The final stage provides a comparative performance evaluation of the three proposed methods in terms of accuracy and efficiency. Improved performances have been obtained where speed up of the computational time for predicting the structure of long RNA sequences was reduced by using fast parallel implementations. Simultaneously, the quality of the final RNA structural results was still impressive.

1.7 Main Contributions

The research in this thesis is inspired by an idea to adapt the CBR search-based methodology for more accurate predictions of the secondary structure of RNA with pseudoknots. The primary topic of this thesis is thus, to present a prediction method for solving the pseudoknotted

RNA secondary structure prediction problem. The research offers contributions in the domain of RNA secondary structure prediction; which can be explained as follows:

1. An adapted CBR method with a new hybrid algorithm to predict the secondary structure of RNA with pseudoknots. This adaptation produces an efficient method by adapting CBR to enhance the secondary structure prediction of RNA with pseudoknots;

2. Three different hybrid RNA prediction methods have been proposed. These three variants were sequentially proposed, to overcome the weaknesses in each previous version. Note that these methods are the three major contributions of this thesis. Each of the contributions can be summarized as follows:

   (a) A novel algorithm based on the bio-inspired swarm intelligence (SI) algorithm with CBR technique, termed as the HPRna predictor. This method can predict the secondary structure of RNA with pseudoknots.

   (b) A new hybrid algorithm with CBR technique called MSeeker is proposed. This method combines KnotSeeker with Mfold, which predict more accurate pseudoknotted RNA structures.

   (c) A fast parallel-distributed algorithm termed FGTSeeker. This method has accelerated the prediction capabilities through the utilization of a new parallel thermodynamic GTFold algorithm. Furthermore, the adapted CBR search-based technique is presented in a new parallel model. FGTSeeker enhances the accuracy of the RNA structures with better performance.

1.8 Organization of Thesis

This thesis is divided into eight chapters and organized as follows. Chapter 2 explains the background of RNA molecules, RNA structures and RNA secondary structure prediction methods.

The background of the CBR method and bee algorithms are also presented in this chapter.

Chapter 3 is divided into two main parts, where part-1 includes a comprehensive review of the current and related works in the domain of RNA secondary structure prediction. It provides a comprehensive discussion of the various methods that have been presented for predicting the secondary structure of RNA with pseudoknots. Part-2 discusses the different methods that have been proposed by imitating the bee colony algorithms. In addition, this part discusses the application of CBR as a search-based method in problem solving.

Chapter 4 describes the main methodology of this research. It also presents a theoretical analysis of the procedures that were adapted. Chapters 5, 6 and 7 introduce the HPRna, MSeeker and FGTSeeker, respectively, which are the three methods proposed in this thesis. Note that each chapter provides a full description of the proposed method and discusses the achieved results to the other state-of-the-art methods. Finally, Chapter 8 provides concluding remarks as well as potential future directions of this work.

CHAPTER 2

BACKGROUND

2.1 Introduction

Bioinformatics is a discipline arising from the combination of computer science and biology (Hogeweg, 1978). Research in this area is rapidly gaining ground, especially with the utilization of advanced computer algorithms, databases, statistical tools and computational theorem, to solve problems relating to management, analysis and retrieval of biological data. Results from Bioinformatics research can be used for crucial practical applications such as the development of therapeutic drugs. Understanding intrinsic biological processes is very important in Bioinformatics research. Computer scientists in particular, need to know important biological terms and concepts. This is important so that proper theoretical computing applications are consequently utilized to perform the proper biological research. In this chapter, the author's intention is to provide the fundamental background and understanding pertaining to biological terms and concepts. Based on the thesis scope mentioned in Section 1.5, the topics being covered will focus on explaining the RNA (Ribonucleic Acid), RNA structure and RNA structure prediction.

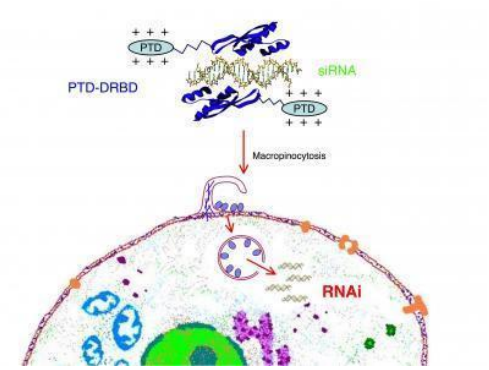
This chapter is divided into three parts. The first part begins with defining various biological terms such as RNA, DNA and protein. Also included are explanations regarding the RNA primary sequence and the various levels of RNA structures. The second part discusses RNA secondary structure prediction and details the major types of prediction methods. The experimental and computational prediction methods are also covered, which are fundamentals

for building the RNA structures. This part also explains the prediction methods for pseudo-knotted RNAs. Finally, the third part presents a background of bio-inspired swarm intelligence (SI), which will be employed in this thesis for solving RNA secondary structure prediction. In addition, an overview of adapting the case-based reasoning (CBR) method to enhance obtained structural results of pseudoknotted RNAs is provided.

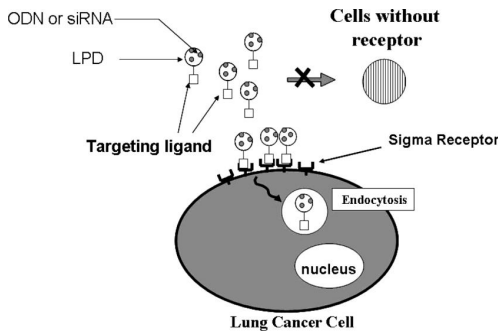
2.2 Basic Biological Data Types

GenBank is a public database housing a myriad of biological data, including nucleotide sequences. This database is constructed mainly via submissions from large-scale projects, where to date, contains data for more than 260,000 known organisms (Benson et al., 2008). The primary sequence (or primary structure) is the main fundamental type of biological data. It is also the easiest to be determined through laboratory experimental methods such as gene sequencing (Ellis et al., 1992; Gray et al., 2005; Bishop et al., 2001). Such primary structures however, do not contain sufficient information about the various roles and the different functions of the biomolecules (i.e RNA, DNA & protein) (Beebee and Rowe, 2008). The secondary and tertiary structures on the other hand contain more information which, can be used to understand the important functions of the RNA biomolecules.

Protein, RNA and DNA are the three main categories of the biological data, which are mostly available in the primary sequences. Protein is an essential component for the living organisms, and is basically a large molecular polymer consisting of amino acid chains linked together by peptide bonds, forming the primary protein sequence. RNA is a single-stranded nucleic acid that carries genetic information for the process of proteins synthesis. DNA on the other hand is a double-stranded nucleic acid that includes genetic instructions for the construction of other components. This section provides detailed discussions of the RNA and RNA structural levels since these are the primary focus of this research. These discussions will



(a) An efficient delivery of siRNA into primary cells to treat cancer.



(b) Targeted delivery of siRNA into Lung cancer cells.

Figure 2.1: A small interference RNA (siRNA) molecule is used to treat and manage cancer disease, adapted from Li and Huang (2006) and Eguchi et al. (2009).

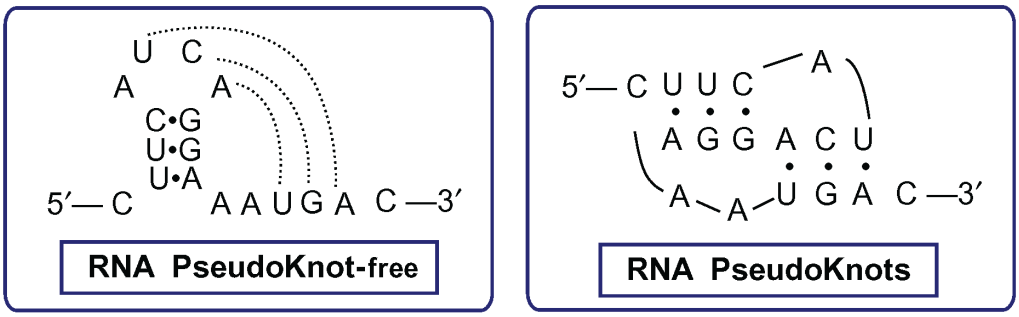
mainly concentrate on the prediction of RNA structures from a given primary sequence.

2.2.1 RNA

Ribonucleic acid (RNA) is one of two nucleic acids that plays a variety of roles in living cells. One type of RNA is the messenger RNA (mRNA), which acts as an intermediary to carry genetic information from DNA for the purpose of protein synthesis (Wang and Shi, 2009). Another type is the small interference RNA (siRNA). The siRNA is delivered into the primary cells by an efficient RNA interference (RNA*i*) system (Figure 2.1) to combat against terminal malignant diseases such as cancer (Li and Huang, 2006; Eguchi et al., 2009).

Recent biological studies have shown that, besides just carrying genetic information for protein synthesis, RNA molecules are also responsible for other useful tasks. These are such as catalyzing biological activities, controlling gene expression, and ribosomal frameshifting (Brierley et al., 2007; Bindewald et al., 2010).

It is important to understand that RNA can mainly be classified into two structural shapes: *pseudoknot-free* and *pseudoknots*. The pseudoknot-free RNA (Figure 2.2a) has the shape of a non-crossing RNA structure motif, which is also known as a stem-loop. Pseudoknots RNA



(a) RNA with pseudoknot-free structural shape. (b) RNA with pseudoknots structural shape.

Figure 2.2: RNA molecules with two main structural shapes (*pseudoknot-free* and *pseudoknots*), adapted from Rivas and Eddy (1999).

(Figure 2.2b) on the other hand has a crossing RNA shape structure, which was discovered by Pleij et al. (1985). The latter RNA has many useful functions where the study of these functions can help in the development and design of antiviral drugs (Andronescu et al., 2010).

RNA is a single-stranded sequence comprising of nucleotides with one of four nucleobases: adenine (A), cytosine (C), guanine (G) and uracil (U). Both DNA and RNA are nucleic acids located in living cells, however with minor differences. For example, RNA is a single-stranded sequence of nucleotide units, whereas DNA is a double-stranded helix of nucleotides that has a thymine (T) nucleobase instead of uracil (U) in RNA. These variations lead to different behavioral roles of RNA and DNA inside living organisms. For instance, DNA builds and stores genetic information, whereas RNA carries this genetic information for protein synthesis. The major differences between RNA and DNA are listed in Table 2.1. The Figure 2.3[1] further demonstrates the basic structures of RNA and DNA, which also illustrates their shapes based on chemical components.

2.2.2 Levels of RNA Structures

Recall that RNA is a single-stranded sequence, which comprises four nucleobases {A, C, G and U}. The RNA structure molecules are classified into the following four hierarchical structural

[1] adapted from *http://www.genome.gov/Pages/Hyperion/DIR/Glossary/Illustration/rna.shtml*

Table 2.1: Basic different variations between the nucleic acids (RNA and DNA) (Osuri, 2003)

RNA	**DNA**
Single-stranded sequence	Double-stranded sequence as a helix
Uracil base instead of thymine	Thymine base instead of uracil
Ribose as a sugar group	Deoxyribose as a sugar group
Uses protein-encoding information	Maintains protein-encoding information
Carries genetic information	Builds and stores genetic information

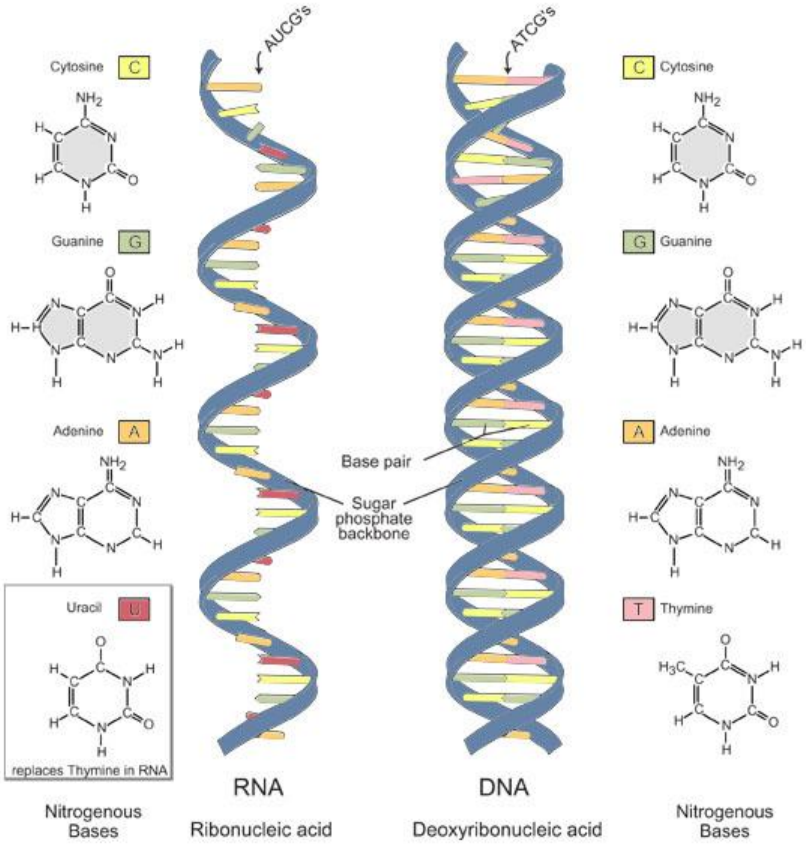
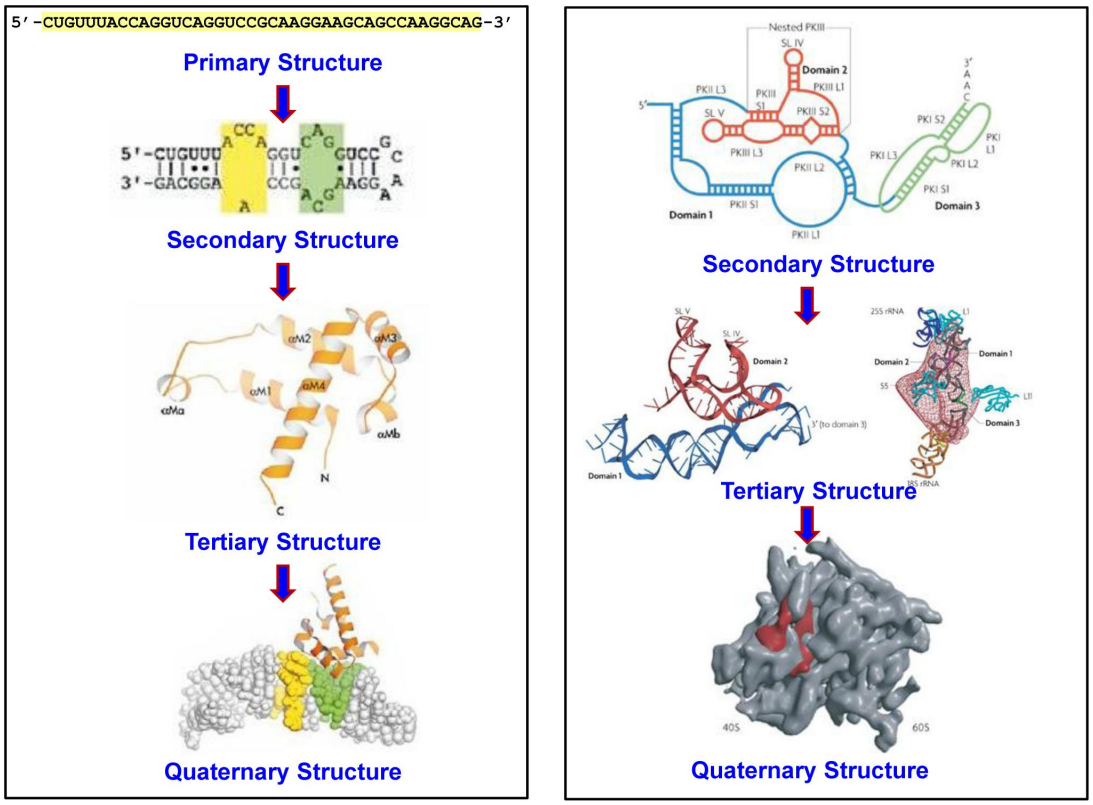


Figure 2.3: RNA and DNA chemical structures.

levels: *primary*, *secondary*, *tertiary* and *quaternary* (Boehringer et al., 2005). These are illustrated in Figure 2.4 and can be described as follows:

1. **RNA primary structure**: This level denotes a linear sequence of RNA bases or nucleobases. It is the basic structural level and they can be easily obtained through the laboratory gene sequencing (Azad and Deacon, 1980). However, the primary structure does not contain much information needed to understand the important roles of the RNA molecule (Kochanek et al., 1996; Beebee and Rowe, 2008).



(a) Escherichia coli SRP RNA molecule.

(b) Hepatitis C virus (HCV) RNA molecule.

Figure 2.4: Illustrative examples representing the four different levels of RNA structures (*primary*, *secondary*, *tertiary* and *quaternary*). Some parts are adapted from Schmitz et al. (1999) and Boehringer et al. (2005).

2. **RNA secondary structure**: This level refers to the two-dimensional (2D) folding structure of the RNA molecule, which occurs when two non-neighboring nucleotides connect through the base pairing of hydrogen bonds (Bauer and Runte, 2000; Al-Khatib et al., 2009). The folding structure shapes the secondary structure of the RNA motif. The bonding is possible based on the following rules: (i) the two Watson-Crick pairs, {C-G} and {A-U}, are the canonical and most stable base pairs (Parisien and Major, 2008); and (ii) the Wobble pair {G-U}, which is a canonical, non-Watson-Crick base pair. Base pairs other than the three canonical pairs {C-G},{A-U} and {G-U}, and their mirrors, are conventionally not allowed (Leontis et al., 2002). An accurate RNA secondary structure is useful as it allows the scrutiny of the RNA's biological functions (Bindewald and Shapiro, 2006). Furthermore, reliable secondary structures can lead to more accurate

tracings of the RNA molecule tertiary structure (Capriotti and Marti-Renom, 2010).

3. **RNA tertiary structure**: This presents the precise three-dimensional (3D) structure, within which, elucidation of the 3D space location of the RNA atoms can be made possible. The RNA tertiary structure describes the global folding of RNA and considers the geometrical and steric limitations to the arrangement of atoms in the RNA molecules. The tertiary structure is important for understanding the functions of RNA molecules, which in turn can be used for the development of therapeutic drugs.

4. **RNA quaternary structure**: This structure refers to the interactions among sub-elements of RNA that consists of the separate units of the molecule. However, this quaternary structure is only used for establishing structural communication between several separate units of sub-elements of RNA like ribosome or spliceosome (Ban et al., 2000).

2.3 RNA Structure Prediction

Determining biomolecular structures is important in order to know biomolecules' crucial functions and myriad roles (Crick, 1970; Anderson and Kedersha, 2009). These structures can be further utilized by biologists and biomedical researchers to develop drugs for diseases (Dass et al., 2008). In general, determining RNA structures can be done in two ways: (i) Biological experimental purification methods, to determine the RNA tertiary structure, or (ii) Computational methods to predict the RNA secondary structure from a given primary sequence (which in turn can be used to find the tertiary structure).

2.3.1 Experimental Methods for Determining RNA Structure

X-ray crystallography and NMR are the two well known experimental methods used by biologists to determine the 3D structures of RNA molecules. From a biological context, these

experimental or biophysical methods are the prominent methods to determine the RNA tertiary structure. However, these methods pose some disadvantages where they consume a considerable amount of time and require special equipments and instrumentations. Due to potentially huge amounts of biological data that need to be processed (i.e. in GenBank), these experimentation methods are inefficient. The following provides comprehensive descriptions of these experimental methods.

2.3.1(a) X-ray Crystallography

X-ray crystallography (XRC) is a diffraction method used to determine the tertiary structures of RNA molecules. During the process, a pure crystal from a single RNA molecule is bombarded with X-ray beams, where the beams are then diffracted to specific locations on a collecting film, as shown in Figure 2.5. The crystallographer then uses the angles and intensities of the diffracted beams to build the 3D depiction as an electron map. Several variables are considered to determine the finalized 3D structure such as the electron density, atom positions and the chemical bonds.

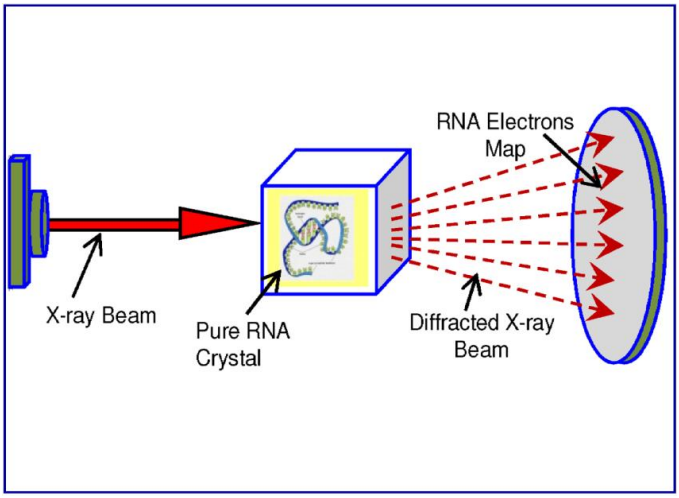


Figure 2.5: Layout of X-ray crystallography workflow as a diffraction method for determining the RNA tertiary structure. Some parts are adapted from Jiang et al. (2008) and Al-Khatib et al. (2010).

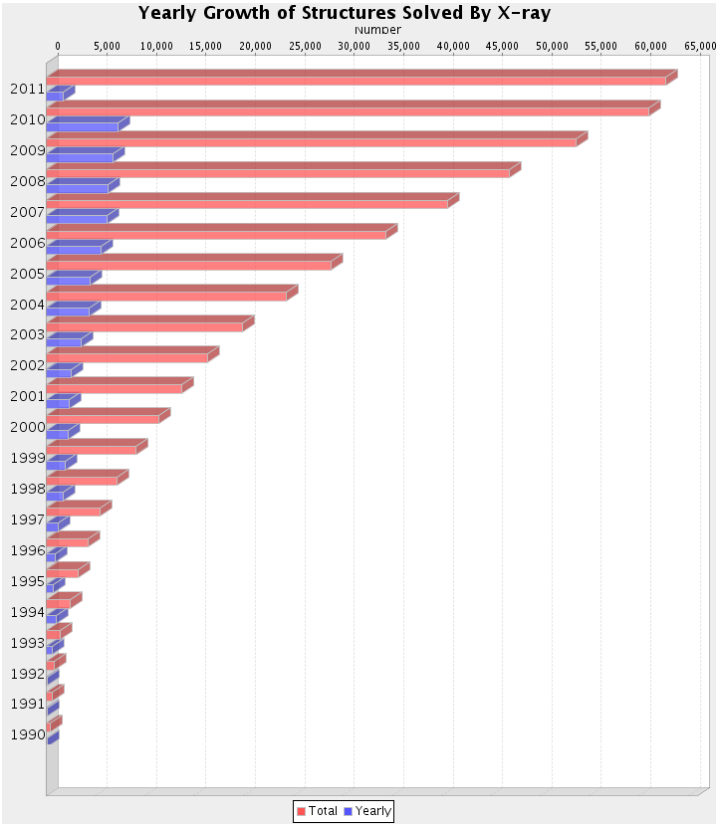
The XRC method is the most popular method to determine the RNA tertiary structure (Westhof and Auffinger, 2000). Figure 2.6a[2] shows approximately 62,750 tertiary structures of molecules determined by XRC from the actual 72,104 structures (as indicated in Figure 2.6b [3]) (Edwards et al., 2009). But the crystallization process of XRC has many limitations that make it a time-consuming, tedious and sometimes practically difficult process. The main constraints are: (i) It is difficult to obtain a pure RNA crystal, and (ii) Large RNA molecules cannot be easily crystallized.

There are many variants of the XRC method. The single-crystal X-ray diffraction method is the most accurate, as shown in Figure 2.5 (Jiang et al., 2008). The success of using the XRC is undeniable where approximately 62,750 tertiary structures from the GenBank were able to be identified (Figure 2.6a). However, there are more than 130-million primary sequence molecules (entries) in the GenBank database (Figure 2.7[4]), from which the tertiary structures still needs to be determined. It is unfeasible for the XRC to cover this gap due to their mentioned limitations. This therefore necessitates the need for alternative methods.

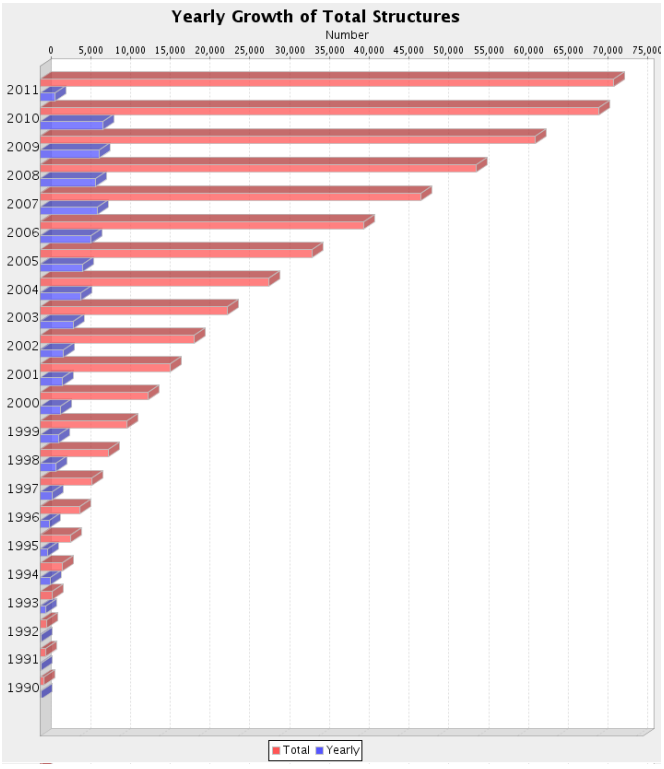
2.3.1(b) NMR Spectroscopy Experimental Method

The nuclear magnetic resonance (NMR) is an alternative method to potentially circumvent the issues faced by XRC. NMR is a powerful experimental method used for determining the tertiary structure of RNA and other biomolecules (Fuertig et al., 2003). This method works on the basic principle that each nucleus in the RNA atoms naturally re-emits absorbed energy from when the RNA sample is fixed and immersed by a magnetic field in nuclear spin process (Figure 2.8). Radio waves from different angles are used to cause resonation of the RNA nuclei. This response is then exploited to identify and build the tertiary structure of RNA molecules by recording the resonation of the nuclei (Kolk et al., 1998).

[2] adapted from $http://www.rcsb.org/pdb/statistics/contentGrowthChart.explMethod-xray\&seqid=100$
[3] adapted from $http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total\&seqid=100$
[4] adapted from $http://www.ddbj.nig.ac.jp/images/breakdown_stats/DBGrowth-e.gif$



(a) Structures determined by X-ray crystallography method.



(b) Structures determined by all experimental methods.

Figure 2.6: Total number of tertiary structures determined by (a) X-ray crystallography method, (b) All experimental methods. (PDB, March 2011 release).