

**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

Acknowledgements.....	ii
Table of Contents	iii
List of Figures	ix
List of Tables	xviii
List of Abbreviations	xxi
Abstrak.....	xxiii
Abstract	xxiv
CHAPTER 1 – INTRODUCTION	
1.1 Background	1
1.2 Motivations and Research Problems	3
1.3 Research Questions	7
1.4 Research Objectives	8
1.5 Research Scope	8
1.6 Overview of Research Methodology.....	9
1.7 Main Contributions	12
1.8 Organization of Thesis.....	13
CHAPTER 2 – BACKGROUND	
2.1 Introduction	15
2.2 Basic Biological Data Types	16
2.2.1 RNA	17
2.2.2 Levels of RNA Structures	18
2.3 RNA Structure Prediction	21
2.3.1 Experimental Methods for Determining RNA Structure	21
2.3.1(a) X-ray Crystallography	22

2.3.1(b)	NMR Spectroscopy Experimental Method	23
2.4	Computational Methods	27
2.4.1	Multiple Sequences Alignment Methods	28
2.4.2	Single or (<i>ab initio</i>) Prediction Methods	31
2.4.2(a)	Dynamic Programming Methods	31
2.4.2(b)	Pseudoknotted Dynamic Programming Methods	34
2.4.2(c)	Limitation of Dynamic Programming Algorithms	35
2.4.2(d)	Meta-heuristic Methods	36
2.4.2(e)	Heuristic-based Methods	37
2.5	Bio-inspired Swarm Intelligence (SI) Field	39
2.5.1	Honey Bee Colony in Nature	40
2.5.2	Honey Bee Foraging	41
2.5.3	Honey Production Process	42
2.6	Case-Based Reasoning Background	43
2.7	Summary	44
CHAPTER 3 – RELATED WORK		
3.1	Single or <i>ab initio</i> Methods	46
3.2	Dynamic Programming Methods	49
3.2.1	PknotsRE Dynamic Programming Algorithm.....	51
3.2.2	Akutsu’s Algorithm.....	52
3.2.3	NUPACK Algorithm	52
3.2.4	PknotsRG Algorithm	53
3.2.5	Parallel RNA DP Prediction Methods	53
3.3	Metaheuristic-based Methods	56
3.3.1	Genetic Algorithm (GA)	59
3.3.2	Simulated Annealing (SA) and Monte Carlo (MC) Methods	60
3.4	Prominent Heuristic-based Methods	63

3.4.1	ILM Method	63
3.4.2	HotKnots Method.....	64
3.4.3	FlexStem Method	64
3.4.4	DotKnot Method	65
3.5	Heuristic Detection Methods	66
3.5.1	KnotSeeker Detection Method	66
3.6	Swarm Intelligence Algorithms for RNA Structure Prediction.....	67
3.6.1	Particle Swarm Optimization for RNA Structure	69
3.6.2	Ant Colony Optimization Algorithm for RNA Structure	71
3.7	Honey Bees Optimization Algorithms	73
3.7.1	Artificial Bee Colony (ABC) Algorithms	74
3.7.2	Marriage in Honey Bees Optimization (MBO) Algorithms	76
3.8	Case-Based Reasoning Method	78
3.9	Summary	80
CHAPTER 4 – RESEARCH METHODOLOGY		
4.1	Introduction	82
4.2	Schema of the Research Methodology.....	83
4.2.1	The Detailed Research Methodology	84
4.3	Pre-processing of RNA Data	86
4.3.1	Datasets of RNA Molecules.....	87
4.3.2	Primary Structure of RNA Molecules	89
4.4	RNA Canonical Base-pairs and Bonding Rules	91
4.5	MFE Models and Energy Functions	94
4.6	Proposed Prediction Algorithms	97
4.6.1	Bio-Inspired HPRna Method.....	98
4.6.2	Combined MSeeker Method	99
4.6.3	Fast Parallel FGTSeeker Method	100

4.7	Data Comparative Analysis and Evaluation	101
4.7.1	Experimental RNA Datasets	101
4.7.2	Evaluation Measurements and Comparison Process	103
4.7.2(a)	Dataset Comparison Analysis of RNA results	105
4.7.2(b)	Qualitative Comparison Analysis of RNA Results.....	106
4.8	Parallel Computing and Evaluation Measurements	108
4.9	Implementation and Test Platform	110
4.10	Summary	111
CHAPTER 5 – A NATURE-INSPIRED METHOD FOR RNA SECONDARY STRUCTURE PREDICTION		
5.1	Introduction	112
5.2	Bio-inspired Method for RNA Prediction	112
5.2.1	Proposed Honey Production Method for RNA (HPRna)	113
5.2.2	Natural Foraging Process.....	114
5.3	Hybrid HPRna for RNA Prediction	115
5.3.1	Forager Bee (<i>Phase 1</i>).....	117
5.3.1(a)	KnotSeeker Detection Step.....	118
5.3.1(b)	CBR-revision Step.....	120
5.3.1(c)	UNAFold Prediction Step	125
5.3.2	Nurse Bee (<i>Phase 2</i>).....	126
5.4	Experimental Results and Discussions.....	129
5.4.1	Data Set Comparison	129
5.4.1(a)	Tabular Comparison	130
5.4.1(b)	Average Results and ROC Plot.....	134
5.4.2	Qualitative Comparison	136
5.5	Summary	138

CHAPTER 6 – ADAPTIVE HYBRID ALGORITHM FOR RNA SECONDARY STRUCTURE PREDICTION

6.1	Introduction	140
6.2	MSeeker for Pseudoknotted RNA Structure Prediction.....	141
6.2.1	Pseudoknot Detection Stage (<i>Stage 1</i>).....	144
6.2.2	Pseudoknot CBR-revision Stage (<i>Stage 2</i>)	146
6.2.3	Non-Pseudoknot Prediction Stage (<i>Stage 3</i>)	151
6.3	Experimental Results and Discussions.....	153
6.3.1	Data Set Comparison Analysis.....	154
6.3.1(a)	General Discussion of Experiments	155
6.3.1(b)	Tabular Comparisons	156
6.3.2	Qualitative Comparison Analysis	162
6.4	Summary	163

CHAPTER 7 – FAST PARALLEL ALGORITHM FOR RNA SECONDARY STRUCTURE PREDICTION

7.1	Introduction	166
7.2	The Potential Advantages of Parallel Processing.....	167
7.2.1	CBR in Parallel Model	168
7.3	Proposed Parallel FGTSeeker Method.....	170
7.3.1	Phase One (Pseudoknots detection).....	172
7.3.2	Phase Two (CBR-Revision)	175
7.3.2(a)	Parallel of CBR-revision	177
7.3.3	Phase Three (Pseudoknot-free)	180
7.3.3(a)	Preparation Step	181
7.3.3(b)	Parallel Multicore GTFold Algorithm	182
7.3.3(c)	Scenario of GTFold Prediction	184
7.3.3(d)	Re-joining Final Prediction Structure.....	184
7.4	Results and Discussion	185

7.4.1	Tabular Comparison and Discussion.....	186
7.4.2	ROC Comparison and Discussion	190
7.4.3	Parallel Experimental Results	192
7.4.3(a)	Parallel Analysis	193
7.4.3(b)	Experiments and System Requirement	194
7.4.3(c)	Parallel Experiments and Discussion	194
7.5	Summary	197
CHAPTER 8 – CONCLUSION AND FUTURE WORK		
8.1	Summary of Contributions	199
8.2	Contributions versus Research Objectives	203
8.3	Future Research	204
	References	206
	APPENDICES	224
	APPENDIX A – DATA SET OF RNA MOLECULES.....	225
	APPENDIX B – NATIVE STRUCTURES OF RNA MOLECULES	228
	APPENDIX C – PSEUDO CODE OF CBR ALGORITHM	231
	APPENDIX D – QUALITATIVE COMPARISON	233
	APPENDIX E – COMPLEXITY ANALYSIS	246
	APPENDIX F – LIST OF PUBLICATIONS	250

LIST OF FIGURES

		Page
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).	4
Figure 1.2	Experimental methods for RNA tertiary structures determination.	5
Figure 1.2(a)	X-ray crystallography method.....	5
Figure 1.2(b)	Nuclear magnetic resonance (NMR) Method.	5
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).	6
Figure 1.4	Scope and general methodology of the research overview.	9
Figure 1.4(a)	Flowchart of the research scope.	9
Figure 1.4(b)	Flowchart of general hybridized methodology.	9
Figure 1.5	Main stages of the research methodology.	11
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).	17
Figure 2.1(a)	An efficient delivery of siRNA into primary cells to treat cancer.....	17
Figure 2.1(b)	Targeted delivery of siRNA into Lung cancer cells.	17
Figure 2.2	RNA molecules with two main structural shapes (<i>pseudoknot-free</i> and <i>pseudoknots</i>), adapted from Rivas and Eddy (1999).	18
Figure 2.2(a)	RNA with pseudoknot-free structural shape.....	18
Figure 2.2(b)	RNA with pseudoknots structural shape.	18
Figure 2.3	RNA and DNA chemical structures.	19
Figure 2.4	Illustrative examples representing the four different levels of RNA structures (<i>primary</i> , <i>secondary</i> , <i>tertiary</i> and <i>quaternary</i>). Some parts are adapted from Schmitz et al. (1999) and Boehringer et al. (2005).	20
Figure 2.4(a)	Escherichia coli SRP RNA molecule.	20

Figure 2.4(b)	Hepatitis C virus (HCV) RNA molecule.....	20
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).	22
Figure 2.6	Total number of tertiary structures determined by (a) X-ray crystallography method, (b) All experimental methods. (PDB, March 2011 release).	24
Figure 2.6(a)	Structures determined by X-ray crystallography method.	24
Figure 2.6(b)	Structures determined by all experimental methods.....	24
Figure 2.7	Total growth of the biological data in GenBank.	25
Figure 2.8	Determining the RNA structure by the NMR spectroscopy method. Some parts adapted from Fuertig et al. (2003) and Al-Khatib et al. (2010).	26
Figure 2.8(a)	NMR spectroscopy method to Determine the RNA tertiary structures...	26
Figure 2.8(b)	Flow scheme workflow of NMR method for RNA structures.	26
Figure 2.9	Total growth of structures determined by NMR spectroscopy method (PDB, March 2011 release).	27
Figure 2.10	Classification of the two groups of computational RNA prediction methods.	28
Figure 2.11	Workflow of multiple sequence alignment (MSA) methods for predicting RNA secondary structure. Some parts adapted from Hofacker et al. (1998) and Marques-Bonet et al. (2009).	30
Figure 2.11(a)	Scheme of multiple sequence alignment (MSA) methods.	30
Figure 2.11(b)	Multiple sequence alignment process for Hepatitis C virus (HCV) RNA molecule.	30
Figure 2.12	Growth of publications methods inspired by the swarm intelligence of honey bees, adapted from Karaboga and Akay (2009b).	40
Figure 2.13	Illustration of the waggle dance as a kind of bee communication, adapted from Guney and Onay (2007).	41
Figure 2.14	The Case-Based Reasoning (CBR) cycle with four main activities (REtrieve, REuse, REvise and REtain), adapted from Aamodt and Plaza (1994) and Watson (1999).	44
Figure 3.1	The classification of RNA secondary structure prediction methods.	48
Figure 3.2	Dynamic programming algorithms for predicting the secondary structure of RNA molecules.	50

Figure 3.3	A schematic classification of the heuristic-based algorithms for predicting the secondary structure of RNA with pseudoknots.	57
Figure 3.4	Types of Swarm Intelligence (SI) algorithms.	68
Figure 3.5	Two Particle Swarm Optimization (PSO) models of collective social behaviour adapted from Grosan et al. (2006): (a) a school of fish behaves in a dynamic parallel group; (b) a flock of birds in a high parallel group.	69
Figure 3.5(a)	69
Figure 3.5(b)	69
Figure 3.6	Two actual double bridge experiments: (a) real ants exploring the shortest path in the double bridge; (b) most of the ants eventually select the shortest path to conduct foraging (Adapted from Goss et al. (1989) and Dorigo et al. (1999)).	71
Figure 3.6(a)	71
Figure 3.6(b)	71
Figure 3.7	RNA prediction methods based on the Swarm Intelligence algorithms.	72
Figure 4.1	General schema of the research methodology.	84
Figure 4.2	Detailed framework of the research methodology.	85
Figure 4.3	Details of the pre-processing stage applied on the selected Escherichia coli (E.coli) molecule from the tmRNA gene type.	88
Figure 4.4	The available information for RNA molecules in PseudoBase (van Batenburg et al., 2001) and GenBank (Benson et al., 2000, 2008).	90
Figure 4.4(a)	The PseudoBase information about one pseudoknot hit from the TMV (Koenig et al., 2005), RNA molecule.	90
Figure 4.4(b)	The available information in GenBank about the whole TMV RNA molecule with pseudoknots class.....	90
Figure 4.5	Various secondary structural elements form by canonical base pairs in real RNA molecules, the images are produced by jViz.Rna software (Wiese et al., 2005).	92
Figure 4.5(a)	Different structure elements of the RNA stem-loops shapes present in S.cerevisiae 5S rRNA molecule.	92
Figure 4.5(b)	Pseudoknot elements present in the secondary structure of TMVYV viral RNA molecule.....	92

Figure 4.6	Diagrammatic positions relation between different RNA base pairs. (a) Two base pairs in nested fashion. (b) Two base pairs in juxtaposed fashion. (c) and (d) Two pseudoknotted base pairs.	93
Figure 4.7	Six different expected structures from primary RNA sequence {A-G-G-C-C-U-U-C-C-U} by using the canonical base pairs rules. This images was produced by the RNA visualization tools PseudoViewer3 & jViz.Rna from Byun and Han (2009) and Wiese et al. (2005), respectively.	94
Figure 4.8	General workflow of the <i>ab initio</i> RNA prediction algorithm.	95
Figure 4.9	Qualitative comparison analysis of the TMV.R structures. (a) The native secondary structure of the TMV.R molecule. (b) The predicted structure from our proposed methods, which has the highest sensitivity 100% and specificity 97%. (c) The predicted structure from DotKnot with sensitivity of 94% and specificity of 91%. The planar images were generated using PseudoViewer3 (Byun and Han, 2009).	105
Figure 4.10	Qualitative comparison analysis of the TMV.R structures. (a) The native secondary structure of the TMV.R molecule. (b) The predicted structure from our proposed methods, which has the highest sensitivity 100% and specificity 97%. (c) The predicted structure from DotKnot with sensitivity of 94% and specificity of 91%. The planar images were generated using PseudoViewer3 (Byun and Han, 2009).	107
Figure 4.10(a)	Native secondary structure of TMV.R molecule from GenBank.	107
Figure 4.10(b)	Predicted structure of TMV.R by proposed methods.....	107
Figure 4.10(c)	Predicted structure of TMV.R by competitor existing method.....	107
Figure 4.11	General schematic view of the Master-Slave parallel model, adapted from Cantu-Paz (1997).	109
Figure 5.1	Workflow of the bee-inspired HPRna method. Note that the final structure was produced using PseudoViewer3 tools (Byun and Han, 2009).	114
Figure 5.2	A representation of modelling the honey production process.	116
Figure 5.3	Detecting the pseudoknot sub-elements using $M_{\text{KnotSeeker}}$ method in the BSMV (Leathers et al., 1993).	119
Figure 5.4	Removal of the overlapped component that is identified in the BSMV a genomic RNA molecule (Leathers et al., 1993).	119
Figure 5.5	The CBR-revision step of the HPRna method is designed in a naïve search function for revising the false positive cases.	120

Figure 5.6	CBR step revises the false positives that are detected in STMV molecule.	121
Figure 5.7	Flowchart of the similarity function used to revise the false positives by using Nearest Neighbour algorithm.	122
Figure 5.8	(a) The same sub-sequence, "CGUGGUGCAUACGAUAAUGCAU", in two different locations within the same RNA molecule (ORSV) folds into the same pseudoknots. (b) The same sub-sequence, "AGUGUUUUUCCCUCCACUAAAUCGAAGGG", in two different RNA molecules (TMV and STMV) folds into the same pseudoknot shape. The secondary structural images of the three RNA molecules are adapted from (Gulyaev et al., 1994).	124
Figure 5.8(a)	124
Figure 5.8(b)	124
Figure 5.9	The detected pseudoknot hits of STMV RNA molecule are coded by (P_{coded}) procedure in (2), after they are revised by CBR in (1). In (3), the pseudoknot-free structure of the coded sequence is predicted by UNAFold. Finally, in (4), the entire pseudoknotted RNA structures is re-constructed.	128
Figure 5.10	Sensitivity comparative plot of proposed HPRna algorithm and FlexStem (2008) method.	134
Figure 5.11	Sensitivity comparative plot of proposed HPRna algorithm and HotKnots (2005) method.	134
Figure 5.12	Sensitivity comparative plot of proposed HPRna algorithm and pknotsRG (2004) method.	135
Figure 5.13	Sensitivity comparative plot of proposed HPRna algorithm and ILM (2004) method.	135
Figure 5.14	ROC plot chart visualizes the average of accuracy in respect to the sensitivity and specificity, for results of the proposed HPRna method and the other methods.	136
Figure 5.15	Qualitative comparison plots of TMV structures: (a) The known native secondary structure of TMV molecule. (b) Secondary structure predicted by our proposed HPRna algorithm, with highest excellent sensitivity of (92.9%) and specificity (95.6%). (c) Secondary structure predicted by FlexStem (sensitivity of 44.3% and specificity 44.9%). (d) Secondary structure predicted by HotKnots (sensitivity of 67.1% and specificity 81.0%). (e) Secondary structure predicted by pknotsRG (sensitivity of 60.0% and specificity 66.7%). (f) Secondary structure predicted by ILM (sensitivity of 20.0% and specificity 20.6%); the images of TMV structures were generated by using PseudoViewer3 tool (Byun and Han, 2009).	137

Figure 6.1	Workflow of proposed MSeeker method for secondary structure prediction of RNA with pseudoknots; the last imaged structure was produced by jViz.RNA 2.0 Software (Wiese et al., 2005).	142
Figure 6.2	Detecting the pseudoknots in the (a) BSMV and (b) T2 bacteriophage RNA molecules using KnotSeeker and the filtration process to prune the overlapped and overflowed components.	145
Figure 6.2(a)	145
Figure 6.2(b)	145
Figure 6.3	The CBR-revision of the MSeeker method is designed in a simple string search function to revise the false positive cases.	147
Figure 6.4	A flowchart representing the computational CBR-revision stage to revise the false positives by using simple string search function.	148
Figure 6.5	Predicted structure of RNA genome JEV (Firth and Atkins, 2009), by using MSeeker with or without of applying the CBR-revision stage.	149
Figure 6.6	Coding the pseudoknot element, predicting the pseudoknot-free structure of the non-pseudoknot portion and re-joining the revised pseudoknot parts to the pseudoknot-free structure in the T2 bacteriophage RNA molecule.	152
Figure 6.7	Sensitivity plot for the detailed comparison between MSeeker and DotKnot.	157
Figure 6.8	Sensitivity plot for the detailed comparison between MSeeker and FlexStem.	158
Figure 6.9	Sensitivity plot for the detailed comparison between MSeeker and HotKnots.	159
Figure 6.10	Sensitivity plot for the detailed comparison between MSeeker and pknotsRG.	159
Figure 6.11	Sensitivity plot for the detailed comparison between MSeeker and ILM.	160

Figure 6.12	Different qualitative comparison structures of the TMV.R molecule using imaged plots produced by the PseudoViewer software (Byun and Han, 2009). (a) The known native secondary structure. (b) The predicted structure from MSeeker (our proposed method, which has the highest sensitivity (100%) and specificity (97%)). (c) The predicted structure from DotKnot (a sensitivity of 94% and a specificity 91%). (d) The predicted structure from FlexStem (a sensitivity of 73% and a specificity 71%). (e) The predicted structure from HotKnots (a sensitivity of 52% and a specificity 56%). (f) The predicted structure from pknotsRG (a sensitivity of 67% and a specificity 74%). (g) The predicted structure from ILM (a sensitivity of 55% and a specificity 61%). (h) The predicted structure from NUPACK (a sensitivity of 61% and a specificity 63%). (i) The predicted structure from pknotsRE (a sensitivity of 94% and a specificity 94%).	162
Figure 7.1	The linear relationship between the retrieval time and the number of cases in Case-Base library for CBR system.	169
Figure 7.2	The workflow diagram of proposed parallel FGTSeeker method.	171
Figure 7.3	Predicting two different RNA sequences (BSMV & T2 bacteriophage) using the proposed parallel FGTSeeker method: (a) Detecting the pseudoknotted sub-elements (hits); (b) Trimming and pruning the undesirable components; (c) Revising the pseudoknot elements by the CBR-revision process; (d) Coding the pseudoknot hits and predicting the non-pseudoknots structure by the parallel multicore GTFold algorithm; and (e) Re-concatenating and re-joining the pseudoknot hits with non-pseudoknots structure to get the final predicted structure $S^{predicted}$, from a given primary sequence X .	173
Figure 7.4	A schematic flowchart for CBR revision phase using the Master-Slave model.	177
Figure 7.5	Flowchart for logical partition policy of CBR revision phase.	178
Figure 7.6	Master algorithm for parallel MPI similarity function (partition, send and receive task).	178
Figure 7.7	Slave algorithm for parallel MPI of similarity function.	179
Figure 7.8	GTFold scalable multicore thermodynamic programming algorithm. Some parts are adapted from (Mathuriya et al., 2009). (a) Pseudo code of the OpenMP parallel programming model of the GTFold algorithm; and (b) Fork-Join style of OpenMP programming model.	183
Figure 7.8(a)	183
Figure 7.8(b)	183
Figure 7.9	ROC plot displaying the average accuracy in terms of sensitivity and specificity of FGTSeeker and the other state-of-the-art methods.	190

Figure 7.10	The sensitivity comparative plot of FGTSeeker and DotKnot (2010) (Sperschneider and Datta, 2010) methods.	191
Figure 7.11	The sensitivity comparative plot of FGTSeeker and FlexStem (2008) (Chen et al., 2008) methods.	191
Figure 7.12	The sensitivity comparative plot of FGTSeeker and HotKnots (2005) (Ren et al., 2005) methods.	191
Figure 7.13	The sensitivity comparative plot of FGTSeeker and pknotsRG (2004) (Reeder and Giegerich, 2004) methods.	192
Figure 7.14	The sensitivity comparative plot of FGTSeeker and ILM (2004) (Ruan et al., 2004a) methods.	192
Figure 7.15	The speedup plot-chart of the parallel FGTSeeker on tested RNA data.	195
Figure 7.16	The execution time of the parallel FGTSeeker algorithm on predicting the tested RNA dataset.	196
Figure 8.1	Overview of the three contributions and the chapter in which they are explained.	200
Figure D.1	Qualitative comparison structures of the VMV Frameshifting RNA molecule (Pennell et al., 2008).	234
Figure D.2	Qualitative comparison structures of the HDV-It_g molecule (Been and Wickham, 1997).	235
Figure D.3	Qualitative comparison structures of the TMV.R molecule (Belkum et al., 1985; Zhang et al., 2009).	236
Figure D.4	Qualitative comparison structures of the BMV molecule (Pleij et al., 1986).	237
Figure D.5	Qualitative comparison structures of the FMDV-A molecule (Clarke et al., 1987).	238
Figure D.6	Qualitative comparison structures of the NeRVN molecule (Koenig et al., 2005).	239
Figure D.7	Qualitative comparison structures of the FMDV-C molecule (Escarmis et al., 1995).	240
Figure D.8	Qualitative comparison structures of the TMV molecule (Koenig et al., 2005).	241
Figure D.9	Qualitative comparison structures of the BSMV molecule (Kozlov et al., 1984).	242
Figure D.10	Qualitative comparison structures of the Ecoli-tmRNA molecule (Nameki et al., 2000).	243

Figure D.11	Qualitative comparison structures of the ORSV molecule.	244
Figure D.12	Qualitative comparison structures of the STMV molecule (Gulyaev et al., 1994).	245

LIST OF TABLES

		Page
Table 2.1	Basic different variations between the nucleic acids (RNA and DNA) (Osuri, 2003)	19
Table 3.1	A comparison of RNA single (<i>ab initio</i>) prediction methods discussed in related works, along with their tangible contributions	49
Table 3.2	The state-of-the-art DP methods for predicting RNA secondary structures	51
Table 3.3	Heuristic-based methods for secondary structure prediction of RNA with pseudoknots	58
Table 3.4	A summary of the bio-inspired honey bee optimisation algorithms and their tangible contributions	74
Table 4.1	Different RNA organisms are assigned as the test dataset with their relevant characteristic statistical information (organism name, accession number, RNA type, length and number of base-pair in native known structure)	102
Table 5.1	The accuracy metrics of the structural results of base pairs between the HPRna algorithm and the other RNA prediction algorithms	132
Table 5.2	A summary of the comparison between HPRna and the other RNA prediction methods using the average sensitivity, specificity and F-measure	135
Table 6.1	A detailed summary of the comparison between MSeeker and the other methods using the average sensitivity, specificity and F-measure	154
Table 6.2	A comparison of the base-pair structural results for MSeeker and DotKnot using sensitivity (SE), specificity (SP) and the F-measure	156
Table 6.3	A comparison of the base-pair structural results for MSeeker and FlexStem using sensitivity (SE), specificity (SP) and the F-measure	157
Table 6.4	A comparison of the base-pair structural results for MSeeker and HotKnots using sensitivity (SE), specificity (SP) and the F-measure	158
Table 6.5	A comparison of the base-pair structural results for MSeeker and pknotsRG using sensitivity (SE), specificity (SP) and the F-measure	159
Table 6.6	A comparison of the base-pair structural results for MSeeker and ILM using sensitivity (SE), specificity (SP) and the F-measure	160

Table 6.7	A comparison of the base-pair structural results for MSeeker and NUPACK using sensitivity (SE), specificity (SP) and the F-measure	161
Table 6.8	A comparison of the base-pair structural results for MSeeker and pknotsRE using sensitivity (SE), specificity (SP) and the F-measure	161
Table 7.1	Comparison between FGTSeeker and other state-of-the-art methods in terms of accuracy metrics	187
Table 7.2	A summary of the accuracy metrics between FGTSeeker and other state-of-the-art methods	188
Table 7.3	Specifications and Requirements of the Parallel System	194
Table 7.4	The average elapsed time and speedup ratio of the parallel MPI model of the proposed FGTSeeker method on homogeneous cluster	195
Table 7.5	Measuring the parallel performance of FGTSeeker by calculating the parallel metrics	196
Table 8.1	The comparative analysis of the results amongst the proposed methods (HPRna, MSeeker and FGTSeeker)	202
Table E.1	Comparative Table of Performance/Complexities Analysis of the bio-inspired HPRna with other Heuristic RNA Prediction Methods	247
Table E.2	Comparative Table of Performance/Complexities Analysis of the combined MSeeker with other Heuristic RNA Prediction Methods	248
Table E.3	Comparative Table of Performance/Complexities Analysis of the parallel FGTSeeker with other Heuristic RNA Prediction Methods	249

LIST OF ALGORITHMS

	Page
1 The Honey Production Algorithm for RNA Prediction Structure (HPRna) . . .	116
2 The Pseudocode of the Nearest Neighbour Algorithm used by the CBR Method	232

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 (Beebe 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 1.1¹ 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-

¹Statistical data from: <http://www.ddbj.nig.ac.jp/documents-e.html>

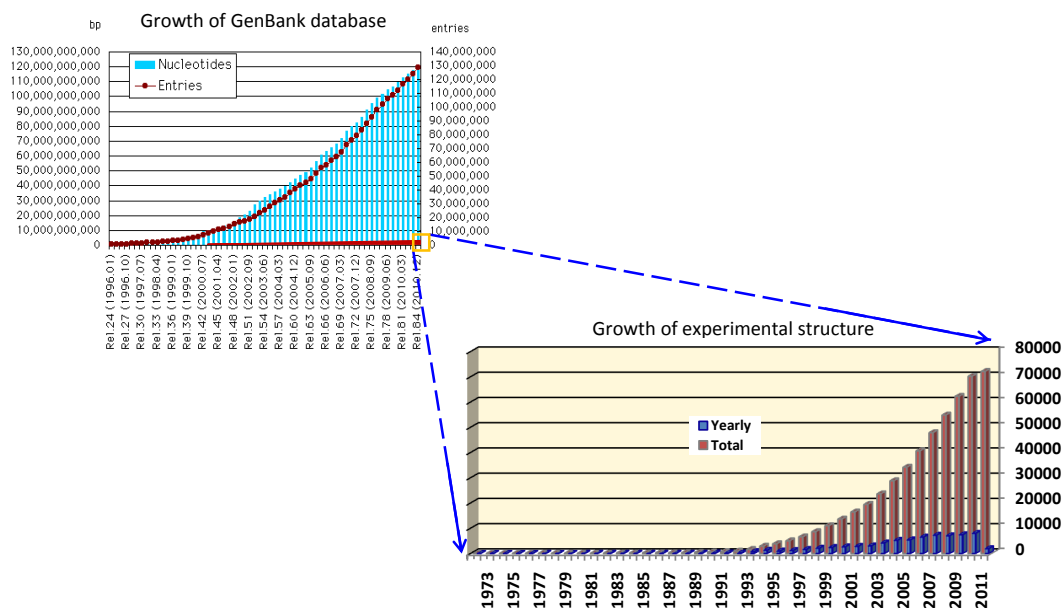
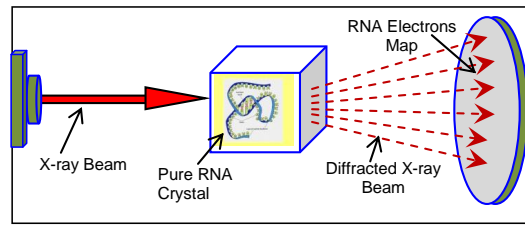


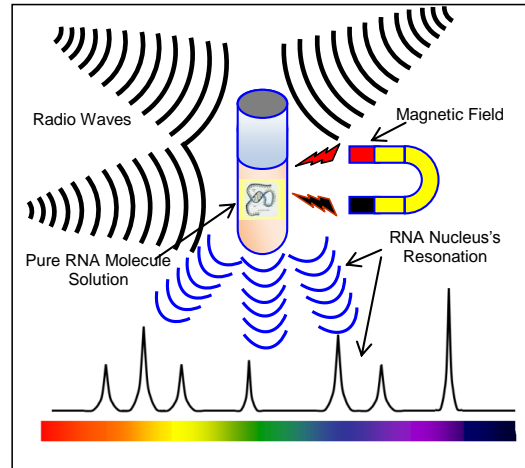
Figure 1.1: Exponential growth of biological data in GenBank and the growth of the known structures (Benson et al., 2008; Golding et al., 2002).

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 resonance 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

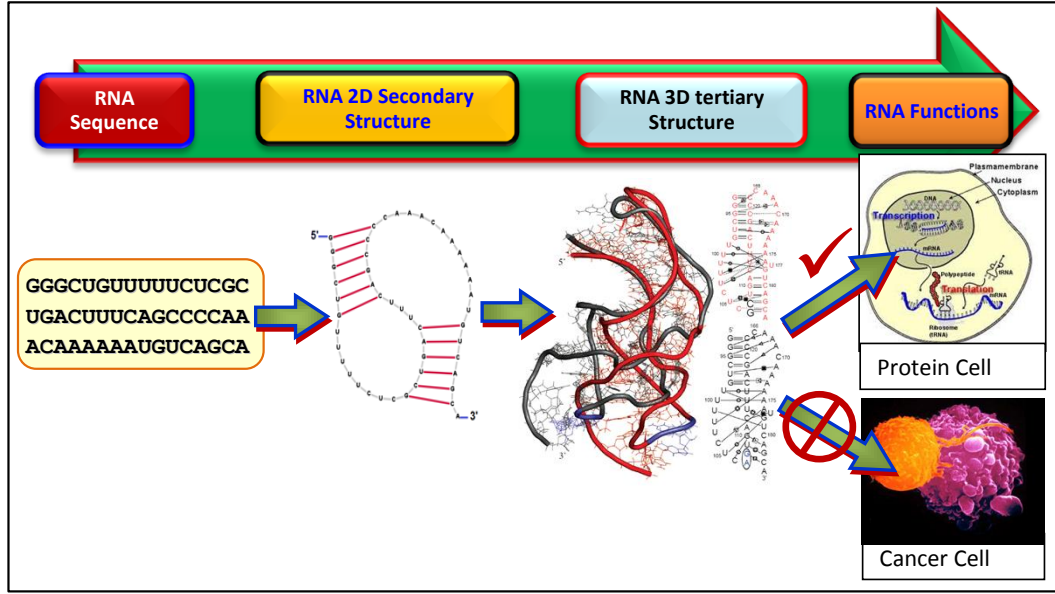


Figure 1.3: RNA structures (primary sequence, secondary structure with pseudoknots, and tertiary structure) of the human telomerase RNA with pseudoknots (Reipa et al., 2007), which includes a wild-type and DKC-mutated pseudoknot structure. The first structure was predicted by HotKnots (Ren et al., 2005) 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

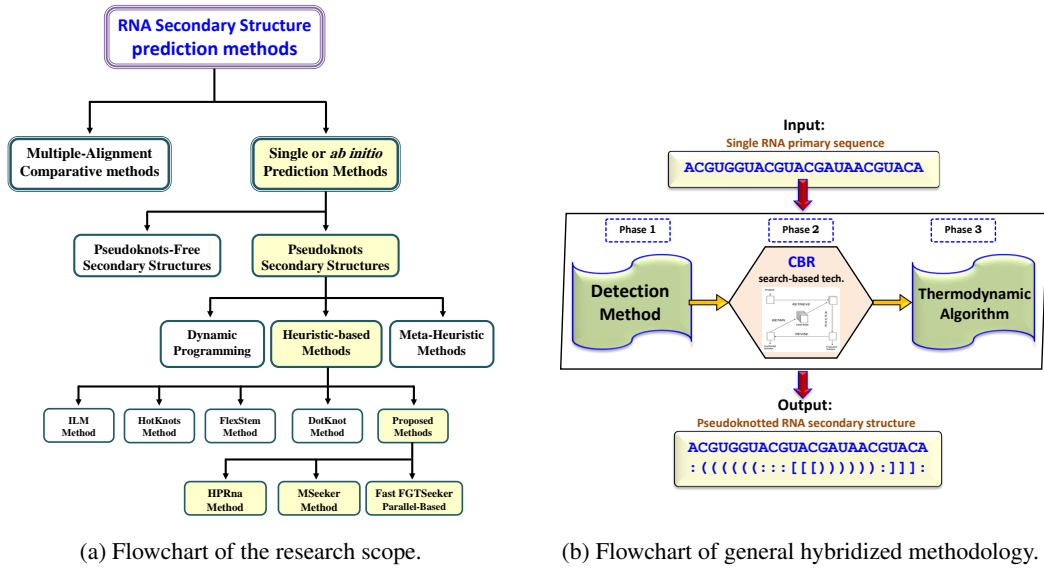


Figure 1.4: Scope and general methodology of the research overview.

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 prediction methods was carried out, whose details are covered in Chapters 2 and 3 (i.e. Background and Related Work).

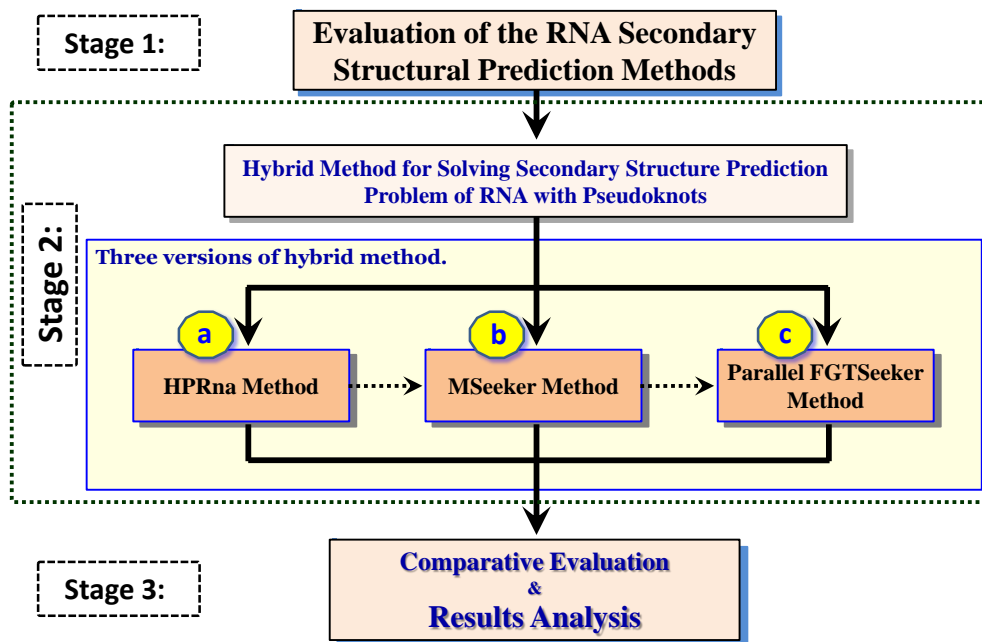


Figure 1.5: Main stages of the research methodology.

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 pseudoknotted 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) (Beebe 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

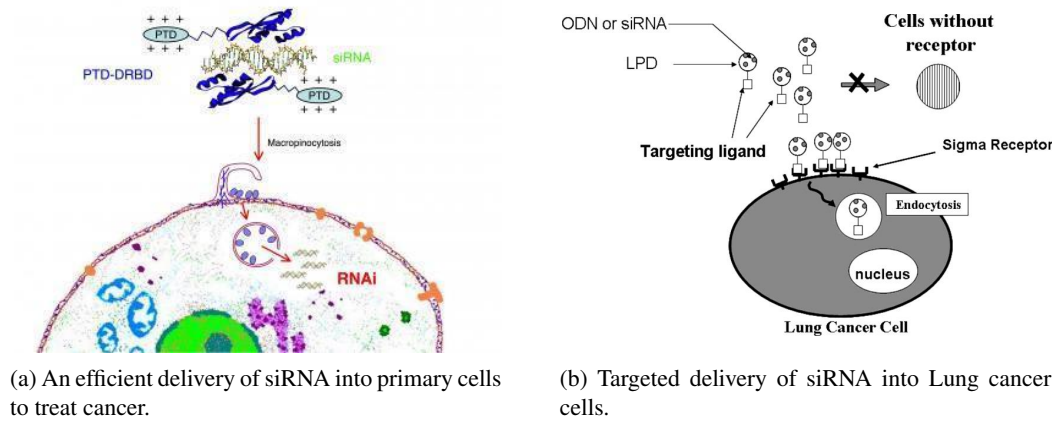


Figure 2.1: A small interference RNA (siRNA) molecule is used to treat and manage cancer disease, adapted from Li and Huang (2006) and Eguchi et al. (2009).

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 (RNAi) 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

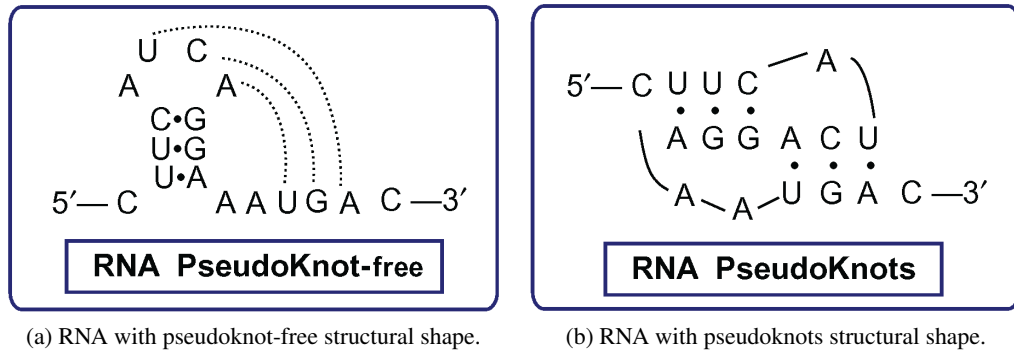


Figure 2.2: RNA molecules with two main structural shapes (*pseudoknot-free* and *pseudoknots*), adapted from Rivas and Eddy (1999).

(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 2.3¹ 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

¹adapted 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

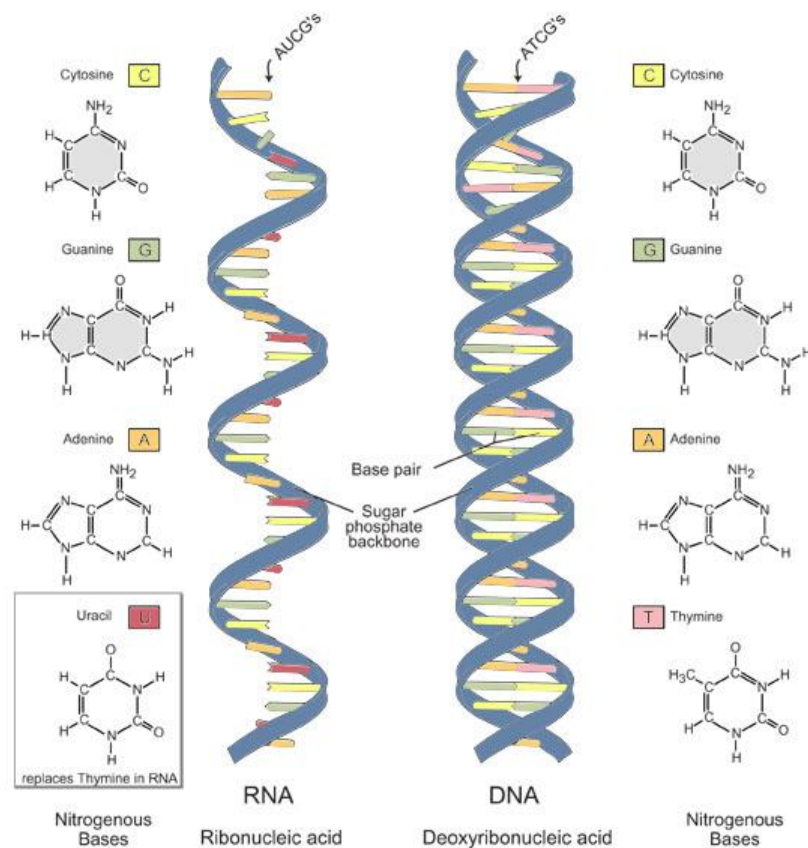
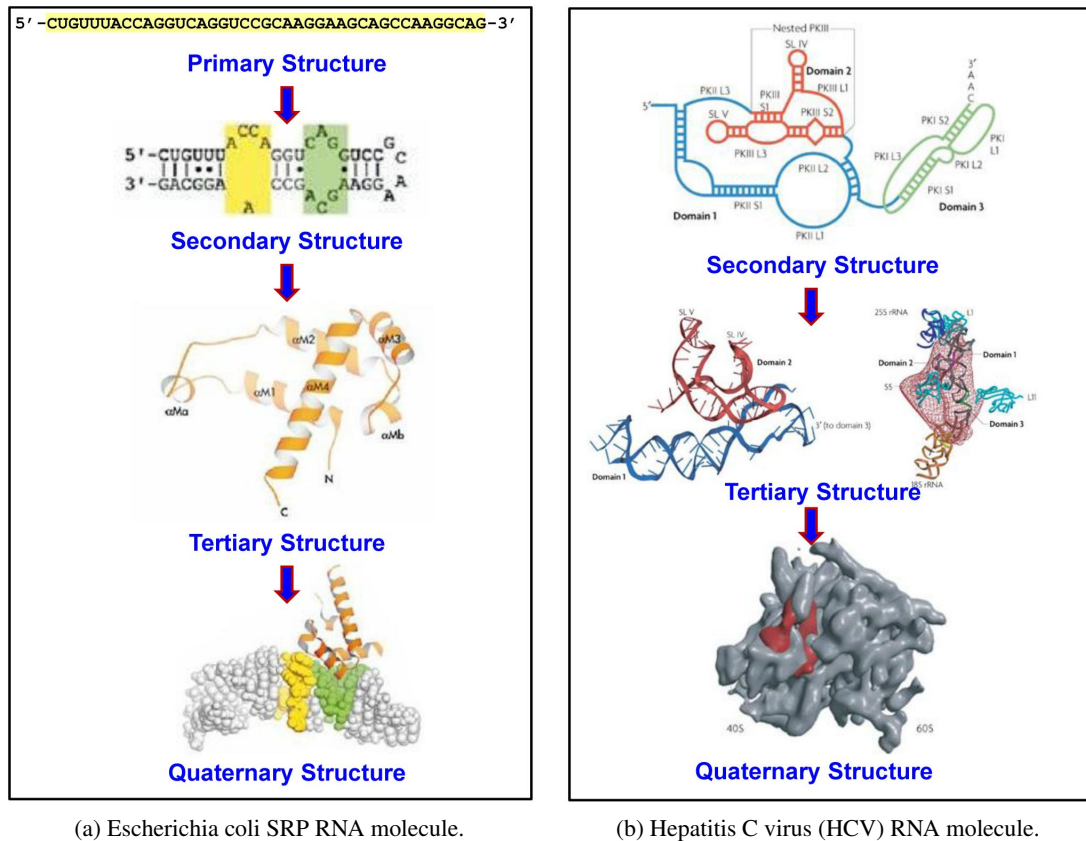


Figure 2.3: RNA and DNA chemical structures.

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).

- 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.

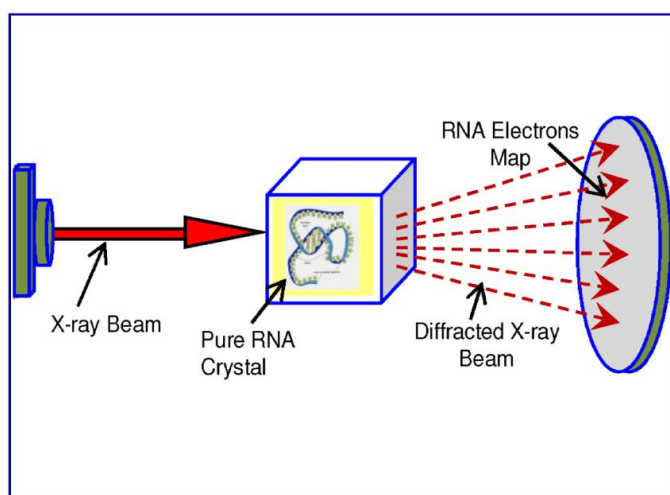


Figure 2.5: Layout of X-ray crystallography workflow as a diffraction method for determining the RNA tertiary structure. Some parts are adapted from Jiang et al. (2008) and Al-Khatib et al. (2010).

The XRC method is the most popular method to determine the RNA tertiary structure (Westhof and Auffinger, 2000). Figure 2.6a² shows approximately 62,750 tertiary structures of molecules determined by XRC from the actual 72,104 structures (as indicated in Figure 2.6b³) (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 2.7⁴), 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.

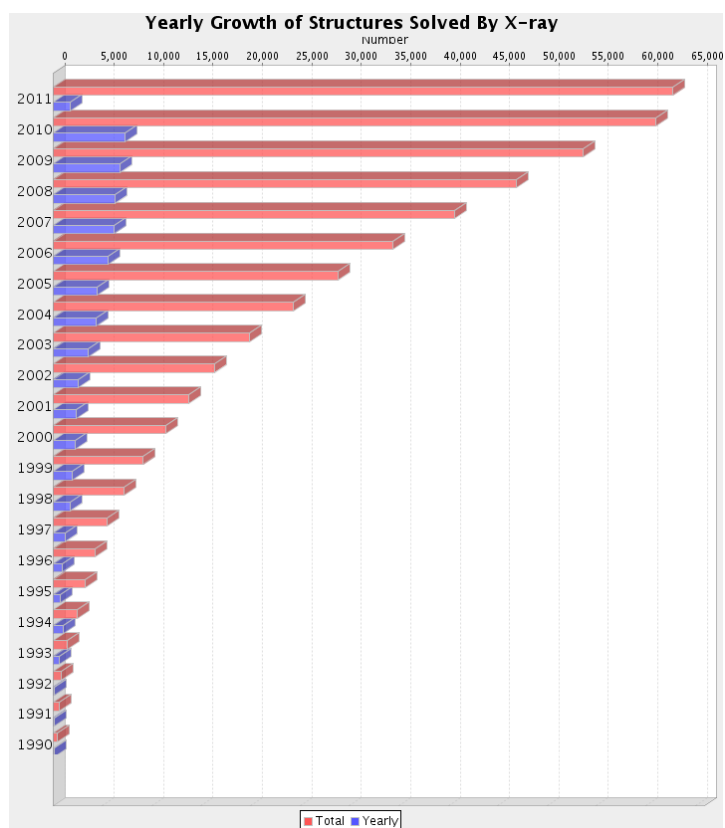
2.3.1(b) NMR Spectroscopy Experimental Method

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 resonance of the RNA nuclei. This response is then exploited to identify and build the tertiary structure of RNA molecules by recording the resonance of the nuclei (Kolk et al., 1998).

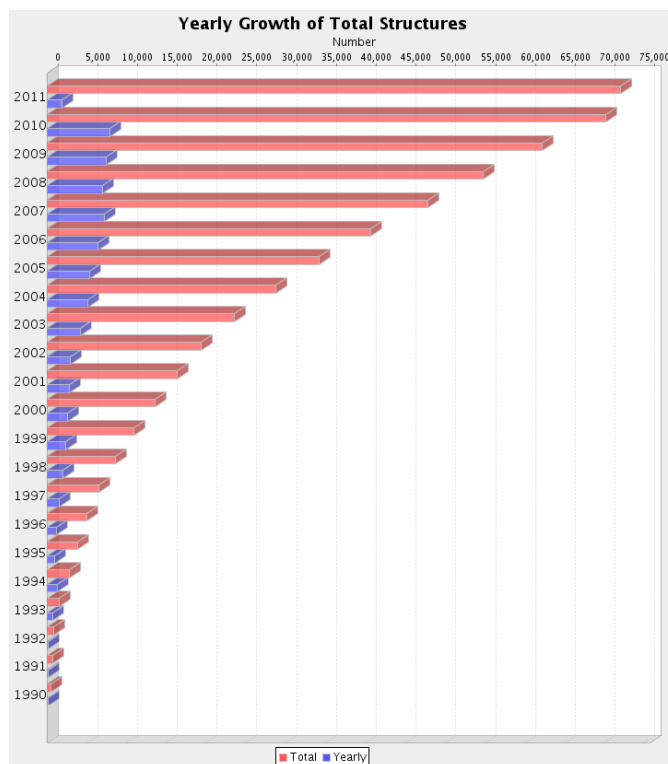
²adapted from <http://www.rcsb.org/pdb/statistics/contentGrowthChart.explMethod-xray&seqid=100>

³adapted from <http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100>

⁴adapted from <http://www.ddbj.nig.ac.jp/images/breakdownstats/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).