ABSTRACT

A BIOINFORMATICS FRAMEWORK FOR RNA STRUCTURE MINING, MOTIF DISCOVERY AND POLYADENYLATION ANALYSIS

by

Mugdha Khaladkar

The RNA molecules play various important roles in the cell and their functionality depends not only on the sequence information but to a large extent on their structure. The development of computational and predictive approaches to study RNA molecules is extremely valuable. In this research, a tool named RADAR was developed that provides a multitude of functionality for RNA data analysis and research. It aligns structure annotated RNA sequences so that both the sequence as well as structure information is taken into consideration. This tool is capable of performing pair-wise structure alignment, multiple structure alignment, database search and clustering. In addition, it provides two salient features: (i) constrained alignment of RNA secondary structures, and (ii) prediction of consensus structure for a set of RNA sequences. This tool is also hosted on the web and can be freely accessed and the software can be downloaded from http://datalab.njit.edu/biodata/rna/RSmatch/server.htm . The RADAR software has been applied to various datasets (genomes of various mammals, viruses and parasites) and our experimental results show that this approach is capable of detecting functionally important regions.

As an application of RADAR, a systematic data mining approach was developed, termed GLEAN-UTR, to identify small stem loop RNA structure elements in the Untranslated regions (UTRs) that are conserved between human and mouse orthologs and exist in multiple genes with common Gene Ontology terms. This study resulted in 90

distinct RNA structure groups containing 748 structures, with 3' Histone stem loop (HSL3) and Iron Response element (IRE) among the top hits.

Further, the role played by structure in mRNA polyadenylation was investigated. Polyadenylation is an important step towards the maturation of almost all cellular mRNAs in eukaryotes. Studies have identified several cis-elements besides the widely known polyadenylation signal (PAS) element (AATAAA or ATTAAA or a close variant) which may have a role to play in polyA site identification. In this study the differences in structural stability of sequences surrounding poly(A) sites was investigated and it was found that for the genes containing single poly(A) site, the surrounding sequence is most stable as compared with the surrounding sequences for alternative poly(A) sites. This indicates that structure may be providing a evolutionary advantage for single poly(A) sites that prevents multiple poly(A) sites from arising. In addition the study found that the structural stability of the region surrounding a polyadenylation site correlates with its distance from the next gene. The shortest distance corresponding to a greater structural stability.

A BIOINFORMATICS FRAMEWORK FOR RNA STRUCTURE MINING, MOTIF DISCOVERY AND POLYADENYLATION ANALYSIS

by
Mugdha Khaladkar

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2009

APPROVAL PAGE

A BIOINFORMATICS FRAMEWORK FOR RNA STRUCTURE MINING, MOTIF DISCOVERY AND POLYADENYLATION ANALYSIS

Mugdha Khaladkar

4/22/09

Dr. Jason T.L. Wang, Dissertation Co-Advisor — Date
Professor, Department of Computer Science, NJIT

4/22/09

Dr. Bin Tian, Dissertation Co-Advisor — Date
Assistant Professor, Department of Biochemistry and Molecular Biology, UMDNJ

4/22/09

Dr. Narain Gehani, Committee Member — Date
Professor, Department of Computer Science, NJIT
Dean, College of Computing Sciences, NJIT

4/22/09

Dr. James McHugh, Committee Member — Date
Professor, Department of Computer Science, NJIT

4-22-09

Dr. Marvin K. Nakayama, Committee Member — Date
Associate Professor, Department of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Mugdha Khaladkar

Degree: Doctor of Philosophy

Date: May 2009

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2009

- Bachelor of Engineering in Computer Engineering,
  Pune University, Pune, Maharashtra, India, 2004

Major: Computer Science

Presentations and Publications:

1. Khaladkar, M., Liu, J., Wen, D., Wang, J.T. and Tian, B. (2008) Mining small RNA structure elements in untranslated regions of human and mouse mRNAs using structure-based alignment. *BMC Genomics*, 9, 189.

2. Khaladkar, M., Patel, V., Bellofatto, V., Wilusz, J. and Wang, J.T. (2008) Detecting conserved secondary structures in RNA molecules using constrained structural alignment. *Comput Biol Chem.*

3. Khaladkar, M., Bellofatto, V., Wang, J.T., Tian, B. and Shapiro, B.A. (2007) RADAR: a web server for RNA data analysis and research. *Nucleic Acids Res*, 35, W300-304.

4. Khaladkar, M., Bellofatto, V., Wang, J.T., Patel, V. and Nakayama, M.K. (2007) Constrained RNA Structural Alignment: Algorithms and Application to Motif Detection in the Untranslated Regions of Trypanosoma brucei mRNAs. Proceedings of the *7th IEEE International Conference on Bioinformatics and Bioengineering*, Boston, Massachusetts, pp. 334-341.

5. Khaladkar, M. and Wang, J.T. (2007) Detecting Conserved RNA Secondary Structures in Viral Genomes: The RADAR Approach. Proceedings of the *2007 NSF Workshop on Biosurveillance Systems and Case Studies*, New Brunswick, New Jersey, pp. 222-227.

6. Khaladkar, M., Bellofatto, V., Wang, J.T., Tian, B. and Zhang, K. (2006) RADAR: An Interactive Web-Based Toolkit for RNA Data Analysis and Research. *BIBE 2006*, pp. 209-212.

To my parents who are my greatest idols,
To my husband who is the wind beneath my wings.

"The woods are lovely, dark and deep.
But I have promises to keep,
And miles to go before I sleep."
~Robert Frost

ACKNOWLEDGMENT

As I approach the light at the end of the tunnel, I wish to extend my heartfelt gratitude towards all those who made this journey possible. Firstly, I thank my dissertation advisors: Prof. Jason T.L. Wang and Prof. Bin Tian, for their guidance and counsel, and for having the faith in me. I thank Prof. Wang for providing me all the resources and opportunities that have helped me achieve what I have. His exuberant passion for research has been contagious. Besides being an excellent scientific advisor he has been a great moral support. His constant encouragement and kind words have pulled me through many frustrating times. Prof. Tian has significantly bolstered my research foundation and enabled me to proceed with confidence towards making a career in research. I am extremely thankful to him for making me part of his Bioinformatics group. I also thank him for his patience when I just couldn't get it, his honest criticism, for BOIL meetings, and much more, but above all, for the inspiration that I have got from him.

I am thankful to my committee members for taking out time to be part of this. I thank Prof. Marvin Nakayama for investing the efforts to provide me with his expert knowledge in statistics. I would also like him to know that his course was the best and I enjoyed being his Teaching Assistant. Sincere thanks to Prof. James McHugh for his interest in my research and for agreeing to be on my committee. Prof. Narain Gehani has been a guardian angel for me at NJIT. From time to time I have barged into his office and he has showed nothing but great kindness and concern. I really appreciate all his support and assistance.

I thank Prof. Barry Cohen for lending a kind ear to my concerns and for the counsel he has provided to me many a times.

I am deeply grateful to my colleagues, in particular, Yang Song, Dongrong Wen, Lei Hua, Ju Youn Lee, Ji Yeon Park, Zhe Ji, Komal Jain, Michael Tsai, Zhenhua Pan. It has been a great pleasure to work with them.

This dissertation would not have been possible without the love and support of my family and friends. I am blessed to have wonderful parents, Mohan and Vinita Khaladkar, who spared no means to provide us with the very best in life and it is because of them that I am here, my brother Manish, who is truly something else. I thank Malti Aji, Pushpa Aji and Ajoba for being the world's best grandparents; Tai Maushi and Srikant Kaka for being the pillars of strength; Amita Mami-Bhaimama for being there for me and moreover to Bhaimama for the inspiration he provides by being such an outstanding researcher; Vaishu Maushi-Raju Kaka for their utmost care and concern for me, their support through these years has been a blessing for me. Special thanks to my fabulous in-laws who have showered me with so much love and blessings. I also thank all my dearest friends who hold a very special place in my heart.

Last but certainly not the least; I don't know what I would have done without my husband Amit. He made all this a lot easier by standing besides me. He is my strength, my joy and my peace and words cannot express what he means to me. Thanks for being in my life.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Objective

The objective of this dissertation is to present a bioinformatics framework conceived for the study of RNA molecules. This research was aimed specifically towards RNA secondary structures and the functionality conferred upon the RNA molecule due to this structure.

To achieve this goal, a tool named "RADAR" which stands for RNA Data Analysis and Research was developed. It comprises of novel algorithms for the detection of conserved secondary structures present in RNA sequences which in turn provides valuable indication of an associated functionality of the molecule. RADAR was tested on several biological datasets and the results exemplify that the method is successful in achieving its purpose.

The framework also comprises of a computational approach termed GLEAN-UTR for the discovery of hitherto unknown RNA structure elements that may be playing important roles in the cell. It was applied to un-translated regions (UTRs) of human and mouse genome and yielded several unique results.

Further, the regulatory role of RNA structure in the process of mRNA polyadenylation was investigated. Polyadenylation is a crucial step in the post-transcriptional gene regulation of most mammalian mRNAs. This study found some correlation between polyadenylation strength and structural stability, and attempted to identify other factors that may increase the efficiency of this process.

1.2 Overview

Ribonucleic acid (RNA) plays various roles in the cell. Many functions of RNA are attributable to their structural particularities (herein called RNA motifs). RNA motifs have been extensively studied for noncoding RNAs (ncRNAs), such as transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), etc. (1). More recently, small interfering RNA (siRNA) and microRNA (miRNA) have been under intensive studies (2,3). Less well characterized are the structures in the un-translated regions (UTRs) of messenger RNAs (mRNAs). However, biochemical and genetic studies have demonstrated a myriad of functions associated with the UTRs in mRNA metabolism, including RNA translocation, translation, and RNA stability. Chapter 2 describes an approach developed to mine novel structures from this region that are conserved between human and mouse orthologs and exist in multiple genes with common Gene Ontology term. This methodology was termed GLEAN-UTR (4) and it uses an RNA structure alignment tool called RSmatch (3) which can efficiently align RNA secondary structures for motif detection. RSmatch can find the optimal global and local alignment between two RNA secondary structures using two different scoring matrices, one for single stranded region and one for the double stranded region. It follows a dynamic programming algorithm of time complexity $O(mn)$, where m is the size of the query structure and n is the size of the subject structure.

RSmatch algorithm suffers from a drawback in that it does not allow the users to specify characteristics specific to the input RNA structures which could enhance the alignment and thus, improve the results. To tackle this program, the framework includes a novel method named constrained structural alignment that is capable of performing a

dynamic alignment based on user-specified constraints, if provided with the input. This algorithm is described in Chapter 3 along with the experimental results. The algorithm is part of the web server and standalone tool RADAR (5) which provides a platform consisting of multitude of functionality for RNA structure analysis. Chapter 4 describes this web server. RADAR was tested on several different biological datasets. Chapter 5 describes its application to find conserved structures from viral genome.

The final focus of this work was on polyadenylation which is a very important process for post-transcriptional regulation of most mRNAs in mammals. Post-transcriptional regulation is the mechanism that controls/regulates the synthesis of protein by genes after the RNA synthesis has begun (6). This field of study has become hugely important since the several discoveries which show that it is a key mechanism that can rapidly change the expression of genes. Chapter 6 describes the work which investigates into the factors that may affect the strength of polyadenylation, with an emphasis on the role played by structure in this process.

This dissertation concludes with a summary of the results obtained, implications of this work and the future research that would go towards further strengthening it.

CHAPTER 2

MINING SMALL RNA STRUCTURE ELEMENTS IN UNTRANSLATED REGIONS OF HUMAN AND MOUSE mRNAs USING STRUCTURE-BASED ALIGNMENT

UnTranslated Regions (UTRs) of mRNAs are involved in various steps of mRNA metabolism, including mRNA localization, translation, and mRNA stability. Regulation of gene expression through UTRs occurs at various developmental stages and is involved in diverse cellular pathways. Several RNA stem-loop structures in UTRs have been experimentally identified, including the histone 3' -UTR stem-loop structure (HSL3) and iron response element (IRE). These stem-loop structures are conserved among mammalian orthologs, and exist in several genes with similar functions. It is not known, however, to what extent RNA structures like these exist in all mammalian UTRs. This chapter describes a systematic approach using the tool RSmatch (3), named GLEAN-UTR, to identify structural elements in human and mouse UTRs that are conserved between human and mouse orthologs and exist in multiple genes with common Gene Ontology terms. This approach resulted in 90 distinct RNA structure groups containing 748 structures, with HSL3 and IRE among the top hits based on conservation of structure. The result indicates that there may exist many conserved stem-loop structures in mammalian UTRs that are involved in coordinate post-transcriptional regulation of biological pathways.

2.1 Background

RNA *cis* elements residing in the UnTranslated Regions (UTRs) of mRNAs have been shown to play various roles in post-transcriptional gene regulation, including mRNA localization, translation, and mRNA stability (7-10). The function of a *cis* element can be attributable to its primary sequence or structure. For simplicity, they are called sequence elements and structural elements, respectively. Well-known sequence elements include AU-rich elements (ARE), some of which contain one or several tandem AUUUA sequences and are involved in modulation of mRNA stability (11,12), and miRNA target sites, which base pair with their cognate miRNA molecules and are involved in the regulation of translation or mRNA stability (13,14). Well-characterized structural elements include Internal Ribosome Entry Site (IRES) (15) and Iron Response Element (IRE) (16) in the 5' UTR, Selenocysteine Insertion Sequence (SECIS) (17), IRE, and histone 3' UTR stem-loop structure (HSL3) (18) in the 3' UTR. Each element type exists in multiple genes, and thus can be considered as an RNA motif (similar to the concept of protein motif). IRE and HSL3 elements are highly similar to one another within each type; some divergence has been reported for SECIS (17) and there is no extensive similarity in primary sequence or secondary structure among IRES elements (15). These characteristics may reflect the ways that the RNA structures function. In addition, various gene-specific structure elements in 5' or 3'UTRs have been shown to play roles in RNA metabolism (7).

Functional RNA sequence elements in the human genome have been heavily studied in recent years, including elements responsible for pre-mRNA splicing, polyadenylation, and miRNA target sites (19-23). In contrast, RNA structure elements

have been investigated to a much lesser extent, partly due to the difficulties in accurately predicting and aligning RNA structures, and assessing false discovery rate (FDR). Recent developments of genome-wide prediction of RNA structures based on aligned genomes (24,25) or unalignable regions (26) have resulted in large numbers of conserved RNA structures. On one hand, all methods reported high potential FDR. On the other hand, these results vary from one another in coverage, indicating that there may exist even more structures to be discovered. Here, the approach described is not based on genome alignments, and is dubbed GLEAN-UTR (grouping by structural distance and ontology for RNA elements in UTRs) to uncover conserved RNA structures in UTRs. The focus was on detecting small stem-loop structures. The folded RNA structures in UTR sequences for orthologous genes were compared by using RNA structure alignment tool RSmatch (3). Similar orthologous structures were then compared in an all-against-all fashion to derive RNA structure groups. Using cluster analysis and Gene Ontology (GO) information, the RNA structures that exist in multiple genes that share common biological pathways were identified. For 10,448 human genes which were analyzed, 90 RNA structure groups, containing 748 distinct RNA structures in 3' or 5' UTRs from 698 genes were obtained. HSL3 and IRE are among the top hits based on conservation of structure. Using a randomized data set, estimated FDR of 15% for all the structures was determined. About 12% of the structures overlap genomic regions identified by other whole-genome wide studies for RNA structures. This bioinformatics study lays groundwork for future wet lab examination of putative conserved RNA structure elements in human and mouse UTRs.

2.2 Results

2.2.1 Mining RNA Structural Elements in UTRs

The aim was to identify functional structure elements in human UTRs. Previous studies have used aligned vertebrate genomes to predict conserved structures in the whole genome (24,25). However, a recent report indicated that many human genome regions containing RNA structures cannot be aligned with the mouse genome (26). This suggests that reliance on genome alignments containing divergent species, such as human and fish, may result in many false negatives. This situation can be exacerbated for UTRs, which typically do not exhibit large rates of sequence conservation. To explore approaches other than using aligned genomes, this method was designed and named GLEAN-UTR, which is based on the rationale that there exist structure elements in 5' and 3' UTRs that are encoded by a group of genes involved in the same biological pathways, similar to IRE and HSL3 structures (see Figure A.1). This method was applied to human and mouse UTRs. Figure 2.1 shows the overall design and procedure of this method.

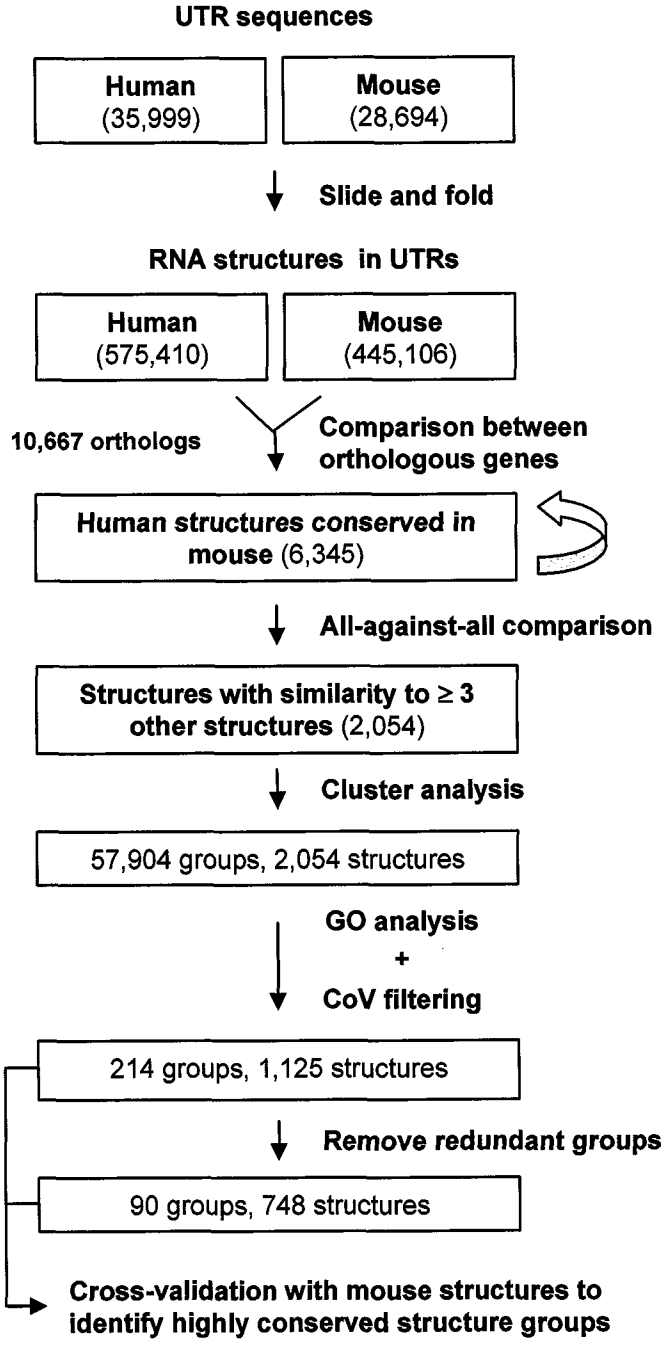

Figure 2.1 The flowchart represents the overall methodology termed "GLEAN-UTR". The number of RNA structures and structure groups are indicated in each step.

First, the UTR sequences were extracted from NCBI RefSeq sequences. Then a "slide and fold" method was used to construct RNA structures in 5' and 3' UTRs (Section 2.4.1). With this method, subsequences in UTRs, 100 nucleotides (nt) long or less, were folded according to thermodynamic properties using the Vienna RNA package (27). Adjacent subsequences were overlapped by 50 nt. This method can derive RNA structures accurately and efficiently for two reasons: (1) Predicting small structures is more accurate and efficient than for large ones; (2) Structures with size less than 50 nt were folded twice as subsequences of two different larger structures, further increasing the chance of getting accurate RNA structures. Further, the setting in the Vienna package that yields multiple RNA structures with the same minimum energy for a given sequence was used to further improve the folding accuracy. On the other hand, since only RNA structures derived from 100 nt subsequences of UTRs was used, the discovery is limited to small structures, such as short stem-loops. Thus, large RNA structures, such as IRES and SECIS, are not analyzed in this study. This step resulted in 575,410 RNA structures from human UTRs and 445,106 RNA structures from mouse UTRs.

Next, the RNA structures from human and mouse orthologs (10,667 pairs in total) were compared. For each orthologous gene pair, the set of RNA structures from the human gene were compared with the set of structures from the mouse gene using RSmatch (3), which aligns RNA structures by taking into account both sequence and structure information. Alignments with a positive score were kept and the rest were discarded. In order to assess the significance of the alignments, three values of a structure alignment were used: size of the alignment, size of the double stranded region of the

alignment, and RSmatch score of the alignment. The distributions of the values for all alignments are shown in Figure 2.2.

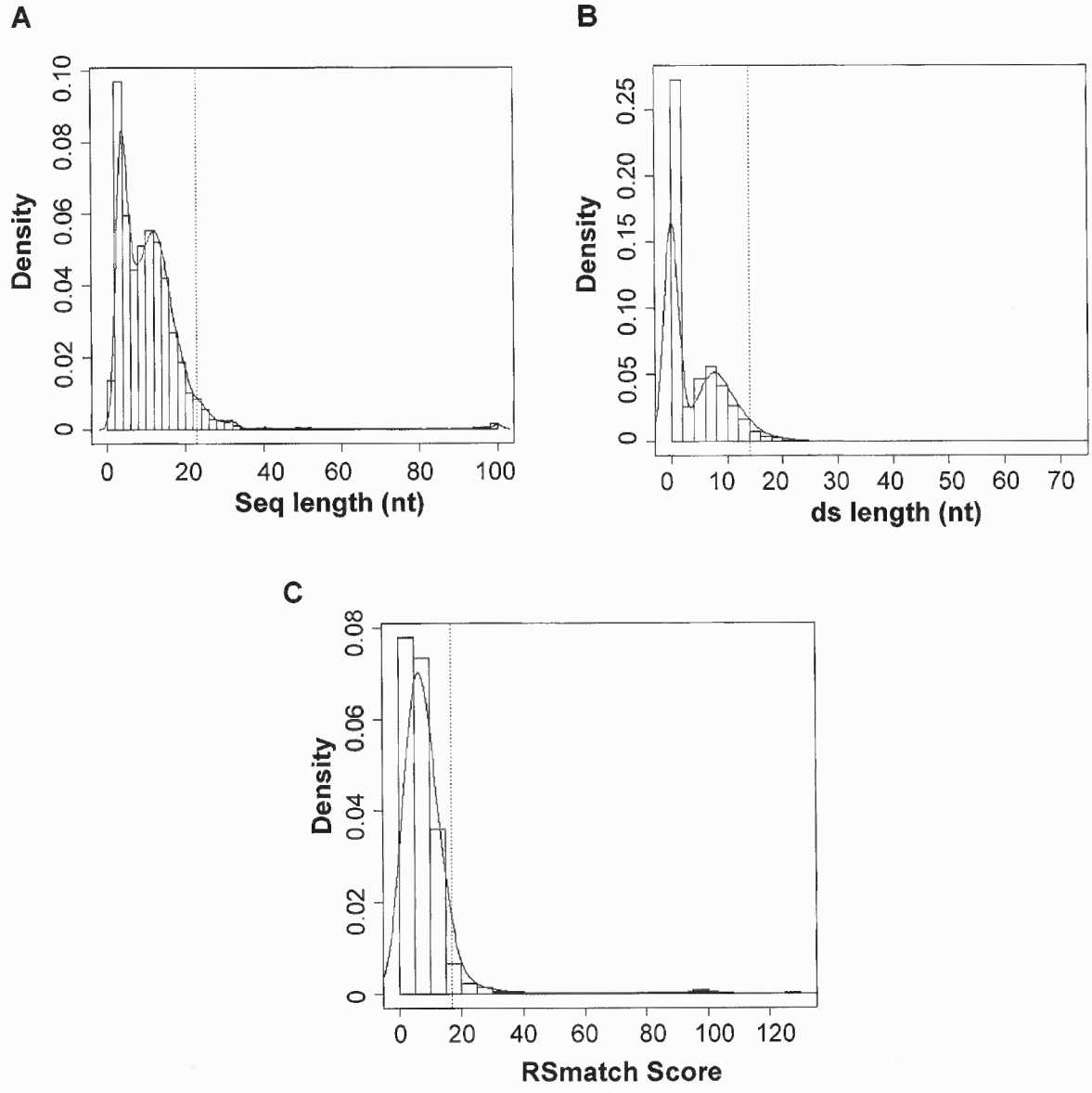

Figure 2.2 Characteristics of aligned RNA structures in human and mouse UTRs. Structures in human UTRs were aligned with those in mouse UTRs from orthologous genes. **(A)** Distribution of overall structure length. **(B)** Distribution of ds region length. **(C)** Distribution of RSmatch alignment score. Dotted vertical lines are cutoff values derived from randomized structures.

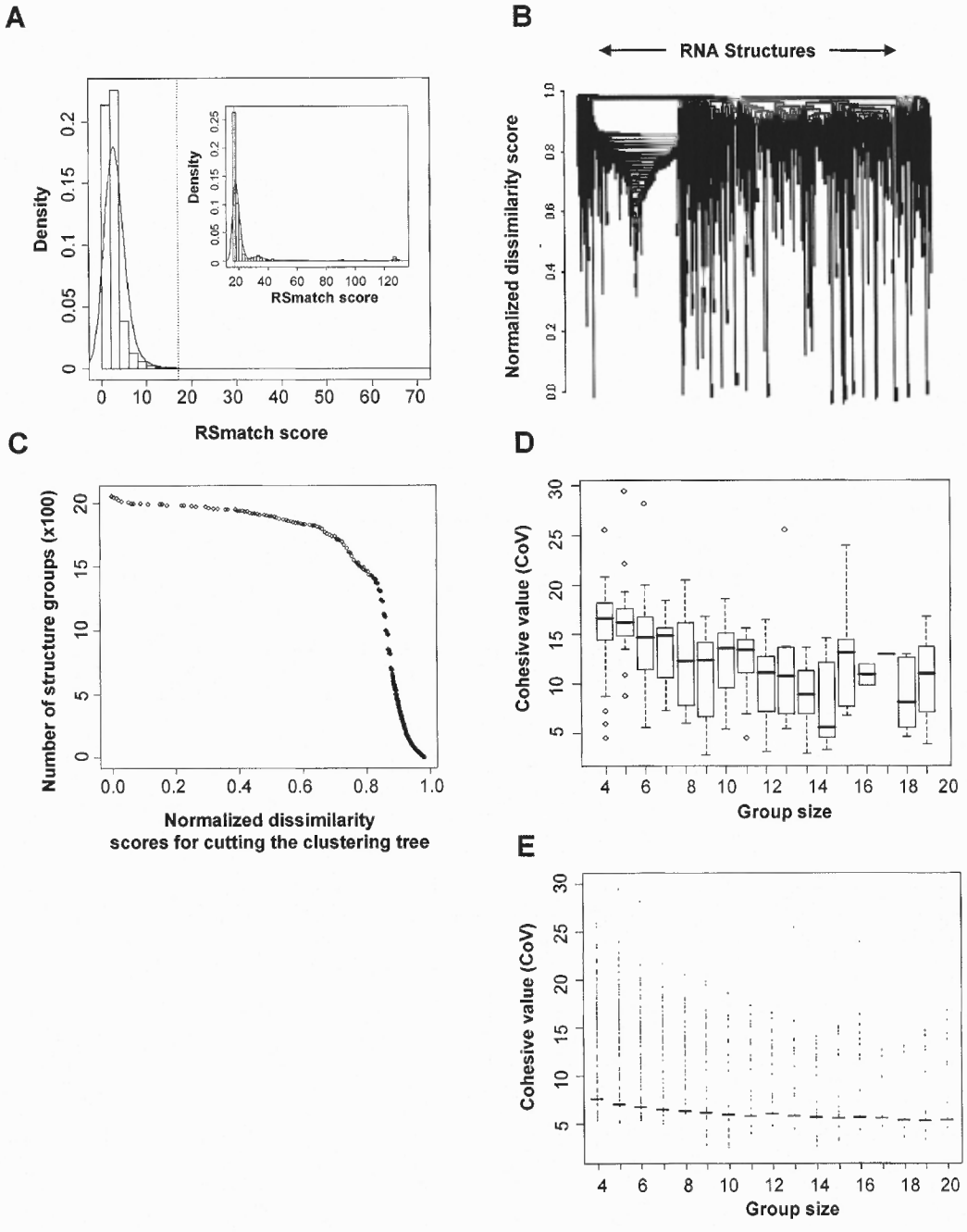

Figure 2.3 (A) Distribution of RSmatch scores for all-against-all pair-wise comparisons of 6,345 human RNA structures. The cutoff value=17, as indicated by a dotted vertical line. The distribution of scores for the selected structures (2,054 in total) is shown in the inset. **(B)** Hierarchical clustering of all 2,054 human RNA structures by the normalized dissimilarity score. **(C)** One hundred normalized dissimilarity scores were used to cut the hierarchical clustering tree to obtain structure groups. Distribution of CoV vs. group size using **(D)** real data and **(E)** randomized data. Horizontal lines in **(E)** are mean values for different groups, which were used as cutoff values for selecting structure groups for the real data.

In order to select significant structures, a randomization method was applied to obtain expected values. Since most known RNA sequence elements in UTRs have the length around 6 nt, the sequences were randomized by shuffling hexamers in UTRs with the goal of separating sequence conservation from structure conservation. For each aforementioned value type, the cutoff value was the 95th percentile of all values from the randomized set. They were found to be 23 nt, 14 nt, and 17 for the size of an aligned structure, the size of a ds region, and the RSmatch score, respectively. To balance selectivity and sensitivity, the structure alignments that had at least two of three values higher than the respective cutoff values were retained. The structure alignments in which two matching structures had identical sequences were eliminated, as the focus of this study was to find elements conserved on the structure level, and it was not possible to differentiate structure conservation from sequence conservation for those alignments. The reasoning was that the ~100 million years since the split of human and mouse ancestors should have given functional RNA structures enough time to have random mutations in insignificant parts of the structure and compensatory mutations in the structure, and the sequences are not expected to be identical unless sequence constraint is also in play. This step resulted in 6,345 alignments.

Then all-against-all pairwise comparisons of all 6,345 RNA structures were carried out. To make the approach computationally efficient, this was done only on human RNA structures obtained from the alignments. Each comparison yielded an alignment score. The structures that were similar to at least two other structures with the alignment score > 17 were selected. This step resulted in 2,054 RNA structures (see Figure 3A for distribution of scores). Both alignments in the single-stranded (ss) and

double stranded (ds) regions can contribute to the final RSmatch score. To assess the contribution of sequence to the selection of these structures, the RNA structures were randomized by swapping nucleotides in both ss and ds regions, while keeping the overall secondary structure intact. With the same selection criteria, 851 structures from the randomized set were selected. Thus, about 40% of the selected structures are primarily due to their structure information, and the remaining 60% are due to both sequence and structure information.

To group similar RNA structures together, hierarchical clustering was applied to the data. First, using pair-wise structure alignment scores, normalized dissimilarity scores were derived to represent distances among the structures (Section 2.4.3). Then a hierarchical tree was constructed containing all 2,054 structures based on their mutual dissimilarities (Figure 2.3(B) and Figure B.1). The hierarchical tree can be "cut" to yield sub trees that represent RNA groups. Figure 2.3 (C) gives the distribution of the number of structure groups obtained by cutting the tree at every value of normalized dissimilarity score. Values at every percentile of this distribution were selected to derive 100 cut heights, i.e. 1st percentile, 2nd percentile, etc. Using these 100 values to cut the tree, 57, 904 groups of structures were obtained, each containing several RNA structures.

In order to find structures that exist in multiple genes involved in the same pathways, the RNA structure groups were further examined by their Gene Ontology (GO) information for the biological process category. The hypergeometric test was applied to measure the significance of association between the genes for a structure group and GO terms (Section 2.4.4). A structure group was selected for further analysis if the group was significantly associated with a GO term (p-value < 0.05), and there were at least two

genes in the group that were annotated with the significant GO term. To measure how member structures in each selected group are similar to one another, a measurement called Cohesive Value (CoV) was used, which is the average of all pair-wise similarity scores among structures in the same group. Figure 2.3(D) shows the distribution of CoVs against group size for all groups. To assess the significance of the CoVs, the same numbers of structures from 2,054 structures were randomly selected to form groups and their CoV values were calculated. For a given group size, this process was repeated 100 times and then the mean value is used as the expected CoV for groups of the given size. Since the numbers of structures in a group ranged from 4 to 20,the expected values were derived for groups with 4–20 structures (Figure 2.3(E)). Groups which had a CoV below the expected values were eliminated. After GO analysis and CoV filtering, 214 structure groups, corresponding to 1,125 distinct structures were obtained.

Since one structure may exist in several groups due to the 100 height values used in cutting the hierarchical tree, the groups that overlapped with other groups with a greater number of structures and lower p-values for the associated GO terms were eliminated while giving preference to groups that were highly conserved between human and mouse based on a cross-validation method (Section 2.4.5). This resulted in 90 structure groups in all, corresponding to 748 distinct structures from 698 genes. Of the structures, 74 are from 5' UTRs and 674 are from the 3' UTRs. Of the groups, 58 groups contain only 3' UTR structures, 30 groups contain structures from both 5' and 3' UTR and 2 groups contain only 5' UTR structures. The top 10 groups based on CoV are shown in Table 2.1. All the structure groups identified by this study, including the ones that are

overlapping with other groups, have been provided in an online database named GLEAN-UTR. It can be accessed freely at http://datalab.njit.edu/biodata/GLEAN-UTR-DB/.

Table 2.1 Top 10 Structures from the "Highly Conserved Set" based on Structure Conservation

Group ID[1] (CoV[2])	Structure[3]		
GO Entries[4]			
3 (HSL3) (28.13)	NM_005321:721-785	AACC-C-AAAGGCTCTTTTCAGAGCCACCCA	HIST1H1E: histone cluster 1, H1e
	NM_021062:401-431	AACC-C-AAAGGCTCTTTTCAGAGCCACCTA	HIST1H2BB: histone cluster 1, H2bb
	NM_005319:704-732	AACC-CAAAAGGCTCTTTTCAGAGCCACC-A	HIST1H1C: histone cluster 1, H1c
	NM_003526:412-438	--CC-C-AAAGGCTCTTTTAAGAGCCACCCA	HIST1H2BC: histone cluster 1, H2bc
	NM_002105:545-578	A-CCAC-AAAGGCCCTTTTAAGGGCCACC-A	H2AFX: H2A histone family, member X
	NM_003516:510-534	-A-----AAAGGCTCTTTTCAGAGCCACCCA	HIST2H2AA3: histone cluster 2, H2aa3
	#=GC SS_cons	..........((((((....))))))....	
GO:0006334 (0) nucleosome assembly GO:0007001 (0) chromosome organization and biogenesis (sensu Eukarya)			
9 (IRE) (19.93)	NM_003234:3430-3460	TTTATCAGTGACAGAGTTCACTATAAA	TFRC: transferrin receptor (p90, CD71)
	NM_014585:197-237	AACTTCAGCTACAGTGTTAGCTAAGTT	SLC40A1: solute carrier family 40 (iron-regulated...
	NM_003234:3884-3912	ATTATCGGGAGCAGTGTCTTCCATAAT	TFRC: transferrin receptor (p90, CD71)
	NM_003234:3481-3509	ATTATCGGAAGCAGTGCCTTCCATAAT	TFRC: transferrin receptor (p90, CD71)
	NM_000032:13-36	GT--TCGTCCTCAGTGCAGGGCA--AC	ALAS2: aminolevulinate, delta-, synthase 2 (side...
	NM_000146:20-40	TG---CTTCAACAGTGTTTGGA---CG	FTL: ferritin, light polypeptide
	#=GC SS_cons	(((((.(((((.....)))))))))))	
GO:0006826 (0) iron ion transport GO:0006879 (0) iron ion homeostasis			
15 (17.40)	NM_015556:203-223	TCATTTAACCTTTTAAATGA	SIPA1L1: signal-induced proliferation-associated 1 like...
	NM_018947:5349-5370	AAATTTAACATTTTAAATTT	CYCS: cytochrome c, somatic , nuclear gene encoding
	NM_000617:2349-2372	TAATTTCTCAGTGGAAGTTA	SLC11A2: solute carrier family 11 (proton-coupled dival...
	NM_018970:469-543	TATATTTTCAGTAAAATGTA	GPR85: G protein-coupled receptor 85
	NM_173494:843-866	TATTGTGACCATTTACAGTA	CXorf41: chromosome X open reading frame 41
	#=GC SS_cons	(((((((((....)))))))))	
GO:0006810 (0.012265) transport			

Table 2.1 Top 10 Structures from the "Highly Conserved Set" based on Structure Conservation (Continued)

Group ID[1] (CoV[2])	Structure[3]		
GO Entries[4]			
17 (17.33)	NM_004441:3717-3813 NM_004443:3616-3640 NM_005398:2077-2107 NM_032827:2394-2416 #=GC SS_cons	TCTTCATATTGAAGA TCTTCATATTGAAGA CCTTCATATTGAAGG GCTTCAAATTGAAGT (((((.....)))))	EPHB1: EPH receptor B1 EPHB3: EPH receptor B3 PPP1R3C: protein phosphatase 1, regulatory (inhibitor) subu… ATOH8: atonal homolog 8 (Drosophila)
GO:0007169 (0.00033) transmembrane receptor protein tyrosine kinase signaling pathway GO:0007165 (0.031793) signal transduction GO:0006468 (0.00927) protein amino acid phosphorylation			
19 (17.17)	NM_000314:502-530 NM_032564:144-170 NM_014751:110-164 NM_016233:2056-2074 #=GC SS_cons	CCTCCCGCTCCTGGAGCGGGGGG GCCCTGGCCCCGGGGGCCGGGGC -CGCTGGC-CCCGG-GTCAGCG- -CCTGTCC-CCCTG-GGGCGGG- ((((((((((...))))))))))	PTEN: phosphatase and tensin homolog (mutated in multi… DGAT2: diacylglycerol O-acyltransferase homolog 2 (mou… MTSS1: metastasis suppressor 1 PADI3: peptidyl arginine deiminase, type III
GO:0045786 (0.00108) negative regulation of cell cycle GO:0007049 (0.009836) cell cycle GO:0006629 (0.001806) lipid metabolism			
21 (17.00)	NM_000899:1060-1087 NM_015355:3606-3643 NM_003081:1331-1430 NM_002893:1613-1645 #=GC SS_cons	TTGCTTCATAAATGAAGCAG ATTCTTTATTTATAAAGGAT -TTATGCATTTATGCATGA- --GCTTGATTTATCAAGC-- (((((((.....)))))))	KITLG: KIT ligand SUZ12: suppressor of zeste 12 homolog (Drosophila) SNAP25: synaptosomal-associated protein, 25kDa RBBP7: retinoblastoma binding protein 7
GO:0016568 (0.000785) chromatin modification GO:0008283 (0.002712) cell proliferation			

Table 2.1 Top 10 Structures from the "Highly Conserved Set" based on Structure Conservation (Continued)

Group ID[1] (CoV[2])	Structure[3]		
GO Entries[4]			
23 (16.90)	NM_001546:1287-1309	CATCTATTGTTTAAAATAGATG	ID4: inhibitor of DNA binding 4, dominant negative …
	NM_020834:2941-2962	CAGGTTTGGTTTTACAAACCTG	KIAA1443: KIAA1443
	NM_005643:1316-1339	CTTTAATGGTTTCACATTGAAG	TAF11: TAF11 RNA polymerase II, TATA box binding pr…
	NM_017617:8938-8965	G-GATTTTGTTTAAAAAATC-T	NOTCH1: Notch homolog 1, translocation-associated (…
	NM_016120:2737-2778	C--ATTT-GTTTAA-AAAT--G	RNF12: ring finger protein 12
	#=GC SS_cons	((((((((.....))))))))	
GO:0016568 (0.000785) chromatin modification GO:0008283 (0.002712) cell proliferation			
25 (16.80)	NM_004625:1678-1700	ATATTAATTTATTTAATTAAA	WNT7A: wingless-type MMTV integration site family, me…
	NM_015508:3801-3828	ATATTTATTTTTTTAATAAAA	TIPARP: TCDD-inducible poly(ADP-ribose) polymerase
	NM_031371:4664-4722	ATATTAAAGATTCTCTTTAAA	ARID4B: AT rich interactive domain 4B (RBP1-like), tr…
	NM_016513:2522-2555	---TTAAAGTTTTTTTTTAA-	ICK: intestinal cell (MAK-like) kinase, transcript va…
	NM_004744:2276-2294	---TTAATTTTTCAAATTAA-	LRAT: lecithin retinol acyltransferase (phosphatidylc…
	NM_138290:1864-1884	---GTAAATGTTTAATTTAC-	RPIB9: Rap2-binding protein 9
	#=GC SS_cons	...((((((.....)))))).	
GO:0007275 (0.036763) development			
27 (16.57)	NM_000252:3053-3080	TTTTACAATGATTTGTAAAG	MTM1: myotubular myopathy 1
	NM_003582:2041-2075	TTTTTATATGATTATAAAAG	DYRK3: dual-specificity tyrosine-(Y)-phosphorylation r…
	NM_001635:2828-2849	GTTTTGCCTAATGGCAAAAC	AMPH: amphiphysin (Stiff-Man syndrome with breast canc…
	NM_001338:2060-2081	ATTTTTCTTATTAGAAAAAT	CXADR: coxsackie virus and adenovirus receptor
	NM_152267:3108-3127	ATTTTCACTGTTGTGAAAGT	FLJ38628: hypothetical protein FLJ38628
	NM_006329:2418-2435	-TTTTGAGTATTTTTAAAA-	FBLN5: fibulin 5
	NM_005627:1871-1929	-TCTTCCATATTTGGAAGA-	SGK: serum/glucocorticoid regulated kinase
	NM_000170:3730-3747	-TTATTAGTATTCTAATAA-	GLDC: glycine dehydrogenase (decarboxylating; glycine …
	#=GC SS_cons	((((((((....))))))))	
GO:0007155 (0.027609) cell adhesion			

Table 2.1 Top 10 Structures from the "Highly Conserved Set" based on Structure Conservation (Continued)

Group ID[1] (CoV[2])	Structure[3]		
GO Entries[4]			
29 (16.30)	NM_002025:8958-8980	GCTGATGCTTTCAGC	AFF2: AF4/FMR2 family, member 2
	NM_014506:1434-1458	GCTGTTCTTTGCAGC	TOR1B: torsin family 1, member B (torsin B)
	NM_014417:1285-1350	-CTCCTCCTGGGAG-	BBC3: BCL2 binding component 3
	NM_007011:2104-2126	-CTCTTCCTGGGAG-	ABHD2: abhydrolase domain containing 2
	NM_004215:1327-1350	-CTAGTGTTTCTAG-	EBAG9: estrogen receptor binding site associated, antigen, 9…
	#=GC SS_cons	(((((.....)))))	
GO:0006915 (0.011186) apoptosis			

[1] Group ID is a serial number, which can be used to query the GLEAN-UTR database.
[2] CoV, cohesive value, which reflects the conservation of structure.
[3] Structures are aligned, and a consensus structure is given for each group.
[4] Significant GO terms associated with each structure group are shown and p-values from hypergeometric tests are indicated in parenthesis.

HSL3 and IRE are ranked among the top hits with respect to CoV values (1st and 2nd) as can be seen in Table 2.1. This result not only validated the approach, but also indicated that other groups of RNA structures may also exist, though probably not as well conserved as HSL3 or IRE. Using the multiple alignment function of RSmatch (3), a consensus structure was generated for each structure group. In a sense, each structure group represents a putative RNA structure element type. The sizes of the consensus structures ranged from 15 to 31. All groups and structures can be searched, retrieved and viewed through an on-line database named GLEAN-UTR DB (4).

To assess the false data rate (FDR) for this method, all the above steps were repeated using randomized human and mouse UTR sequences maintaining overall dimer frequencies, and the number of selected entries at each step was calculated (Figure C.1). In the last step, this randomized set resulted in 17 groups consisting of 110 human structures. Thus, the FDR is ~18.89% for the groups and ~14.71% for the structures. Of these groups, 3 groups with 14 structures also passed the cross-validation with mouse orthologs, giving FDR ~8.82% for the groups and ~5.96% for the structures.

2.2.2 Comparison with other Genome-wide RNA Structure Studies

Three recently carried out studies for finding conserved RNA structure regions in the human genome (24-26) were selected. Their results were examined for structures that differed from and overlapped with the results obtained in this study. Using 8-way human-referenced vertebrate genome alignments, Washietl et al. (24) detected 91,676 conserved RNA structures (at $P > 0.5$) using the RNAz program, which identifies RNA structures with similar thermodynamic stabilities across species. Pederson et al. (25) developed

phylogenetics stochastic context-free grammar (phylo-SCFG), and identified 48,479 candidate RNA structures using the same genome alignments. Torarinsson et al. (26) focused on human and mouse genomic sequences that could not be aligned on the sequence level, and identified conserved structures by FOLDALIGN, a tool that simultaneously predicts and aligns RNA structures.

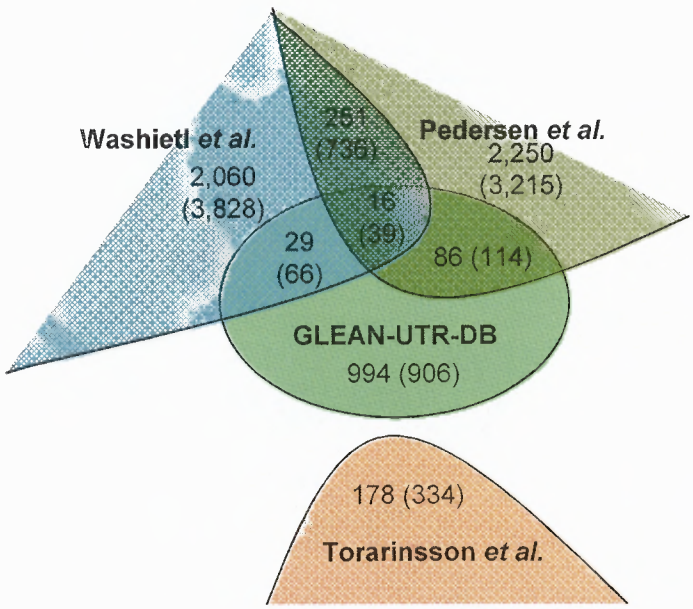

Figure 2.4 The Venn diagram shows overlapping structures in UTRs among the results reported by Washietl et al. (24), Pedersen et al. (25), Torarinsson et al. (26), and this study. The number in the paranthesis indicates the number of overlapped structures if only the genomic regions are considered, i.e. without consideration of the strand.

First the structures reported by these studies that are located in UTRs were identified, and these were compared with structures found by GLEAN-UTR approach. Of the 1,125 structures that were identified prior to removal of redundant groups (see above), 131 (12%) structures overlapped with those reported by Washietl et al. (24) and Pedersen et al. (25) (Figure 2.4 and Table D.1). If only the genomic region is examined (without consideration of the strand), 219 (19%) structures were found to be overlapping

with those in these two studies. Of the 178 structures predicted by Torarinsson et al. (26) that overlapped with UTR regions, none overlapped with the results on this study. A detailed analysis found that this was caused by differences in human and mouse UTR coverage (127 cases), gene ortholog information (27 cases), or structure alignment (24 cases).

2.3 Discussion

This chapter describes a systematic approach that was designed to identify RNA structure elements conserved in human and mouse UTRs which may function coordinately in post-transcriptional regulation of biological pathways. The approach contains three major steps: (1) compare RNA structures between orthologous genes; (2) compare RNA structures among all genes; and (3) select RNA structure groups significantly associated with certain GO terms. Presumably, mRNAs containing RNA structure elements from the same group can be coordinately regulated via *trans*-acting protein factors, like those having HSL3 and IRE, leading to concerted modulation of a biological pathway. This method was applied to mining small RNA structures in this study, primarily because those structures can be more accurately predicted by RNA prediction programs using only thermodynamic parameters. As more powerful RNA structure prediction programs become available, particularly those reliant on phylogenetic information for structure prediction, this approach can be extended to larger RNA structures. The major strength of this approach is the ability to assign functions to candidate RNA structures in the genome. In addition, it may help improve the accuracy in RNA structure identification, as

structures shared by multiple genes can be more reliable than those encoded by a single gene.

The assessment of FDR is critical in RNA structure analysis (28). Using randomized sequences, FDR of 15% was estimated for the structures identified in this study. False negative rate or sensitivity is another important issue, particularly in this study in which stringent cutoff values were applied at multiple steps. However, it is difficult to address due to lack of knowledge on true positive structure groups.Two well-known RNA structure elements, HSL3 and IRE, were examined for sensitivity. For HSL3 and IRE genes that have orthologous gene information, about 35% (6 out of 17) HSL3 elements and 60% (6 out of 9) IRE elements are included in the final result. Thus the sensitivity can be low for some structure groups and high for others. Several steps can result in exclusion of conserved functional RNA structures in our method. First, the current coverage of orthologous genes and UTRs is not complete. In fact, most of the human HSL3 true positive structures (44 in total) were not even analyzed in this study due to lack of orthologous gene or UTR information. This will improve as more comprehensive gene annotations, and more accurate transcription start sites and polyadenylation sites are available. Second, it is known that RNA structure prediction by thermodynamic parameters has limitation in accuracy (29). Third, some structures may reside in genes for which GO information is not adequately annotated.

One potential approach to improve sensitivity is to search the genome with consensus RNA structures derived from the groups. This idea was tested by first generating RNA structure patterns for the groups and using them to search human UTRs by PatSearch (30). Candidate elements were further analyzed for GO terms to ensure

consistency in their association with biological pathways as the original groups. As expected, the group size increased exponentially (Figure E.1). While this approach seems promising in reducing the false negative rate, the control for false positive rate needs to be further developed. This work is left for future exploration.

About 12% of the structures identified in this work overlap those reported by other studies (Figure 2.4). Interestingly, each genome-wide approach resulted in a large fraction of unique structures, suggesting that RNA structure identification is largely influenced by the chosen method. Many structures in UTRs identified by other studies are not in our final result (Figure 2.4). This is attributable to several aspects of the design of our study, in addition to the technical difference and false negative issues described above. First, this analysis is based on RNA structure groups, and functional structures located in individual genes are not included. It was found that this is the case for several recently reported RNA structures in UTRs (31,32). Second, RNA structures with similar functions but different secondary structures, like IRES, cannot be identified. Third, large structures, like SECIS, are not examined. Notwithstanding these issues, the structures that overlap between this study and others are of higher importance for further wet lab validations (Table D.1).

In summary, the result indicates that there may be present many conserved stem-loop structures in human UTRs that are involved in coordinate post-transcriptional gene regulation of biological pathways, similar to HSL3 and IRE structures. This bioinformatics study lays a ground work for future wet lab validations of putative RNA stem-loop groups and represents a framework which can be used to analyze RNA structures identified by other approaches and in other species.

2.4 Materials and Methods

2.4.1 UTR Sequence and Structure Databases

28,926 human and 26,243 mouse RefSeq mRNA sequences were downloaded from NCBI. UTRs of RefSeq sequences were extracted according to RefSeq's GenBank annotation. The information regarding human and mouse orthologs was obtained from the HomoloGene database (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/). RNA structures in the UTRs were prepared by a method called "slide and fold" as described in (3). Briefly, for each UTR sequence, 100 nt subsequences were taken at every 50 nt nucleotide position from 5' to 3' resulting in consecutive subsequences overlapped with one another on a 50 nt segment. Subsequences shorter than 100 nt, e.g. at the 5' or 3' ends, were also kept. Then all of the subsequences were folded using the RNAsubopt function from the Vienna RNA package (27), with the setting "-e 0". With this setting, multiple structures with the same minimum energy can be generated. Using this method, 575,410 structures from human UTRs, and 445,106 structures from mouse UTRs were obtained.

2.4.2 RNA Structure Comparison

Pairwise comparisons of RNA structures (human vs. mouse and human vs. human) were carried out by RSmatch (3), with the "dsearch" function and default scoring matrices for ss and ds regions. Specifically, nucleotide match scores were 1 and 3 in ss and ds regions, respectively; and mismatch scores were -1 and 1, in ss and ds regions, respectively. Gap penalty was -6 for both ss and ds regions. This scoring scheme in effect gave more

weight on matches in ds regions than those in ss regions. Randomization of RNA structure was carried out by a PERL script.

2.4.3 Cluster Analysis of RNA Structures

To cluster RNA structures, the normalized dissimilarity scores $D_{i,j}$ were calculated between all structures: $D_{i,j}=(S_{max}-S_{i,j})/S_{max}$, where $S_{i,j}$ was the similarity score derived from RSmatch using the local structure alignment function between structures i and j, and S_{max} was the maximum similarity score obtained from all structure comparisons. For cluster analysis, the hierarchical clustering function in R was used (33) with the "average linkage" method for joining nodes. To select groups of RNA structures, the "cutree" function was applied to cut the hierarchical tree obtained from R into groups using the normalized dissimilarity scores, which were also called heights in the tree. Structures in each group were aligned by the multiple structure alignment function of RSmatch (3) with default scoring matrices. Structures in the same group were also compared in a pairwise manner; the average of all pair-wise similarity scores for the group was called the Cohesive Value (CoV) of that group, which indicated the degree of similarity among structures in the group.

2.4.4 Gene Ontology Analysis

The biological process (BP) category of Gene Ontology (GO) was downloaded from the Gene Ontology database (34). The mapping between genes and GO entries was obtained from NCBI Gene database (35). Hypergeometric analysis was used to assess whether an RNA structure group was significantly associated with some GO entries.

$$f(x,m,n,k)=\frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \tag{2.1}$$

Briefly, in the hypergeometric test, there are four parameters: (1) m, the number of white balls in an urn, (2) n, the number of black balls in the urn, (3) k, the number of balls drawn from the urn, and (4) x, the number of white balls drawn from the urn. The probability that x out of the k balls drawn are white from the urn containing $m + n$ balls is For each RNA structure group M containing multiple genes, all GO entries are examined to evaluate their associations with M. Through the mapping information between M and a GO entry G in a GO category C, we are able to calculate four numbers: (1) $N1$, the number of genes associated with any GO entry in C, (2) $N2$, the number of genes associated with G in C, (3) $N3$, the number of genes in M associated with any GO entry in C, and (4) $N4$, the number of genes in M associated with G in C, where $N1 \geq N2$ and $N3 \geq N4$. The p-value of the GO entry G is calculated by $p(G) = f(N4, N2, N1 - N2, N3)$, where the function f is defined in Equation 2.1.

2.4.5 Cross-validation with Mouse UTR Structures

After performing the GO analysis and CoV filtering, the selected human RNA structure groups were cross-validated with their orthologous mouse structures. For each group, mouse UTR structures corresponding to each human structure in the group were retrieved. Then the mouse UTR structure which is most similar to the human structure is selected. All these selected mouse structures are compared by the multiple structure alignment function of RSmatch which gives the consensus structure. The consensus structure of human RNA structures was then compared to that of mouse ones. An RNA

structure group is considered to be highly significant if: (1) the human consensus was identical to the mouse one or (2) the human consensus was contained within the mouse one or vice verse. In case (2), a consensus of human and mouse structures was built to represent the structure group.

2.4.6 Comparison with Structural Elements from other Studies

The datasets for Pedersen et al. and for Washietl et al. were downloaded from their respective web sites (24,25). The dataset from Torarinsson et al. (26) was obtained from the authors. BLAT was used to find genomic locations for all structure elements, including for the ones predicted by this study, and overlapped ones were identified by their locations.

CHAPTER 3

DETECTING CONSERVED SECONDARY STRUCTURES IN RNA MOLECULES USING CONSTRAINED STRUCTURAL ALIGNMENT

Constrained sequence alignment has been studied extensively in the past. Different forms of constraints have been investigated, where a constraint can be a subsequence, a regular expression, or a probability matrix of symbols and positions. However, constrained structural alignment has been investigated to a much lesser extent. Here, described is an efficient method for constrained structural alignment which is applied to detecting conserved secondary structures, or structural motifs, in a set of RNA molecules. The proposed method combines both sequence and structural information of RNAs to find an optimal local alignment between two RNA secondary structures, one of which is a query and the other is a subject structure in the given set. This allows a biologist to annotate conserved regions, or constraints, in the query RNA structure and incorporate these regions into the alignment process to obtain biologically more meaningful alignment scores. A statistical measure is developed to assess the significance of the scores. Experimental results based on detecting internal ribosome entry sites in the RNA molecules of *hepatitis C virus* and *Trypanosoma brucei* demonstrate the effectiveness of the proposed method and its superiority over existing techniques.

3.1 Introduction

In recent years, it is becoming clear that post transcriptional processes at the RNA level play a major role in determining the complexity of the proteome along with a significant amount of regulation of gene expression (36,37). Numerous examples of co-regulation of sets of transcripts in RNA regulons have also been described (38). The identification and characterization of RNA sequence and structural regulatory elements, therefore, is of fundamental importance to molecular biology (1,2).

Inspired by the success of proteomics using sequence-based techniques, researchers anticipated achieving the same level of success in RNA study. Unfortunately, till now the accomplishment is far from what had been expected. A typical example is with RNA motif exploration: unlike protein motif searching which can be accomplished through the development of sophisticated amino acid substitution matrices and sequence alignment tools, detecting RNA motifs is still at a primitive stage without broadly accepted methods in the literature. One important reason for the failure of substitution matrices-based alignment methods in analyzing RNA sequences is that nucleotide bases do not carry as much functional information as amino acid residues do (39). To properly characterize an RNA motif, information concerning both distant base interactions and sequential nucleotide composition is required to define its structure, and hence its function.

At the sequence level, one important topic is to measure the similarity of two biosequences (40,41). The next step is to find an alignment between two sequences or among several sequences. Tools capable of performing sequence alignments include

BLAST (42), FASTA (43), ClustalW (44), with their primary goal of detecting homologs from sequence databases.

However, biological activities of many molecules, such as non coding functional RNAs, are largely dependent on their secondary or tertiary structures. Furthermore, it has been observed that myriad functions involved in post-transcriptional gene regulation are accomplished by RNA protein binding mechanisms, which require conserved structural RNA motifs to be present at the binding sites. Thus, it is biologically justifiable that conserved RNA motifs in the form of secondary or tertiary structure could be more important and informative than those in the primary sequence format (45).

This research proposes a new approach to RNA secondary structure alignment and also applies it to the search for conserved secondary structures, or structural motifs, in RNAs. The problem tackled here is defined as follows: given a query structure Q and a set of RNA subject structures, find the subject structures that are most similar to the query structure where the similarity between the query structure Q and a subject structures S is measured by the score of local matches between Q and S. When the query structure is a structural motif or a conserved secondary structure, the problem becomes finding those subject structures containing the conserved secondary structure and displaying the locations of the conserved secondary structure in those subject structures.

Central to the approach is an efficient constrained structural alignment (CSA) method for comparing two RNA secondary structures with quadratic time and space complexities. The CSA method allows the user to annotate a portion of the query structure, or the entire query structure, as conserved, and then uses this information, or *constraint*, to align the query structure Q with each subject structure S in the given set.

The constraint guides the alignment process, which dynamically varies the alignment scores between portions of Q and S to obtain a more accurate alignment between the two structures.

The RNA structures are obtained by folding RNA sequences using either mfold (46) or RNAfold (27). In (47) a general edit distance was considered for comparing RNA secondary structures. RNAforester (48) extended the tree model to a forest model.

Corpet and Michot (49) designed RNAlign to provide more rigorous RNA structural comparisons at the cost of computing efficiency: $O(n^4)$ in space and $O(n^5)$ in time where n is the length of the RNA structures to be compared. Several other tools are available that carry out RNA folding and alignment at the same time, such as Dynalign (50) and FOLDALIGN (51). These tools can achieve better structure prediction and alignment at the expense of computing time. In addition, algorithms using derivative-free optimization techniques, such as genetic algorithms and simulated annealing (52,53) have been proposed to increase the accuracy in structure-based RNA alignment. Most of these methods suffer from high time complexities, making the structure-based RNA tools much less efficient than sequence-based tools.

There are pattern-matching methods for RNA analysis (39,54,55). In (55) a sequence-scanning technique was proposed, called PatSearch. The pattern present in an RNA secondary structure is depicted by a series of pattern description units. The sequences in a dataset are scanned one by one to decide whether the given pattern can match these sequences. In another related study (39), a profile-based sequence-scanning algorithm was proposed and implemented under the name ERPIN. Like most statistical model based methods, ERPIN requires a multiple alignment of sequences with secondary

structure annotation and infers a statistical secondary structure profile (SSP). This SSP is then matched with the sequences in the dataset by using a dynamic programming algorithm to calculate scores of the best matches.

Some probabilistic models, such as stochastic context-free grammars (SCFGs) (56) and covariance models (CMs) (57), have been applied to RNA structural alignment. A model is first trained by a set of manually curated sequences with known structural similarities. The trained model is then used to compare with other related RNA structures. Since a prior multiple sequence alignment (with structural annotation) is needed to train the model, its applicability is limited to RNA types for which structures of a large number of sequences are available, such as snoRNA and tRNA (56,58). In (59) SCFGs were extended to find homologs of structured RNA sequences using RIBOSUM substitution matrices derived from ribosomal RNAs to score the matches in single-stranded (ss) and double-stranded (ds) regions. The pairwise SCFG method requires computing time as high as $O(n^3)$ (59). More recently, better algorithms based on the probabilistic models have been developed (60,61). However these methods do not deal with constrained alignments as described in the next section.

3.2 Methods

Constrained structural alignment (CSA) constructs the alignment between a query RNA structure and a subject RNA structure based upon the knowledge of the conserved region in the query structure. This method has been implemented as part of the web server RADAR which is described in Chapter 4.

Figure 3.1 shows the input interface of RADAR for aligning a query structure with a set of subject structures. The structures are represented in the Vienna style Dot Bracket format (27). Each position of the conserved region in the query RNA structure is marked using a special character "*". Figure 3.2 shows the output obtained from the input data in Figure 3.1, where RADAR compares the query structure with each subject structure using the proposed CSA method and ranks the subject structures in the dataset based upon their similarities to the query structure. The top ranked subject structure is most similar to the query structure, with the maximum alignment score. The score diminishes as the quality of the alignment decreases. A statistical measure, namely a *p*-value, is associated with each alignment score, which indicates the significance of the score (Section 3.2.5).

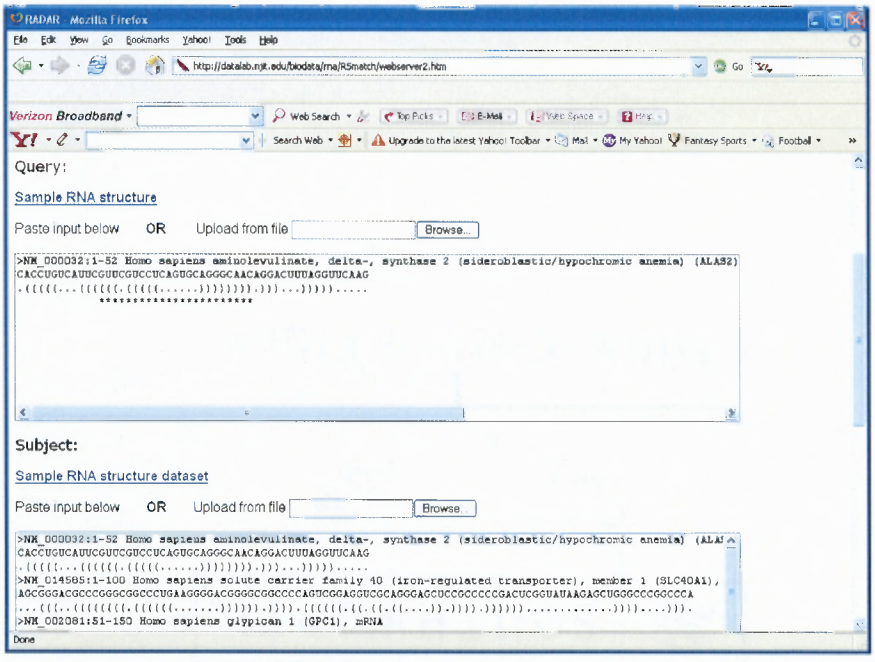

Figure 3.1 The input interface of RADAR for the constrained structural alignment. The first text box contains the query structure. Constrained region of the query is marked using "*". The second text box lists the subject RNA structures forming the dataset.

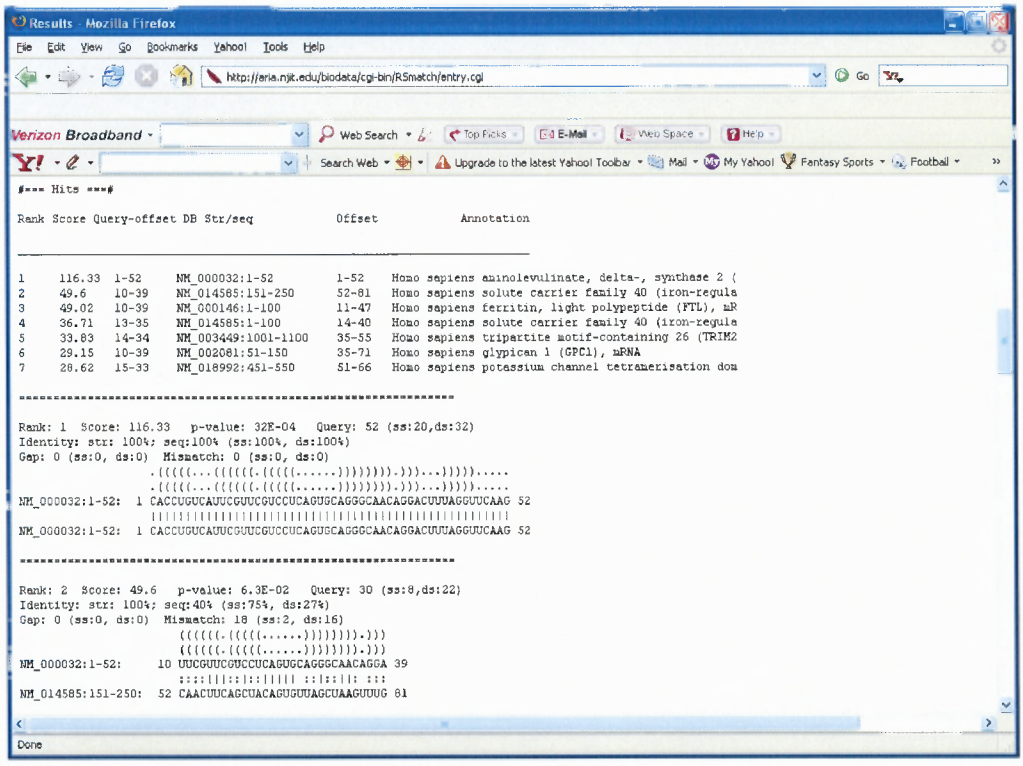

Figure 3.2 Output obtained after performing constrained structural alignment between the query structure and the subject structures in Figure 3.1. Output shows a summary of the top ranked alignments including the score, subject structure name and aligned region for each alignment. Then each alignment is shown one after the other starting with the top-ranked one.

3.2.1 Extended Loop and Structural Component

The proposed CSA method is built on previously developed RSmatch algorithm for RNA structural alignment (3). RNAs are modeled using a structural decomposition scheme similar to the loop-decomposition method commonly used in RNA structure prediction algorithms (46). Thus pseudoknots are not allowed. An RNA secondary structure is completely decomposed into units called *extended loops* (Figure 3.3(A)). An extended loop, or simply a loop when the context is clear, is a set of structural components (single bases or base pairs), which are reachable from one another by traversing within the loop

without crossing any bond. The extended loops considered here differ from the commonly used loops described by (46) in that the extended loops can be part of a stem in an RNA secondary structure.

The above obtained extended loops can be organized into a hierarchical tree according to their relative positions in the secondary structure, where each node corresponds to an extended loop (Figure 3.3(B)). The tree construction is as follows. The root node is established as the extended loop containing the 5' most and 3' most bases. Within the root loop, each base-pair r is used to form a subtree (or child tree) whose root corresponds to another extended loop containing r. This process is iteratively performed until no further extended loop can be found and the tree is completely constructed. Furthermore, we require that the nucleotide pairs be processed from 5' to 3' within the extended loops. Consequently, the final tree is an ordered tree in which the order among sibling nodes is important.

In describing the relative positions between two structural components (single base or base pairs), the precedence and hierarchical relationships between them are taken into consideration. Let c_1 and c_2 be two structural components in an RNA sequence and its secondary structure. It is said that c_1 precedes c_2 if at the sequence level the 3'-base of c_1 is closer to the sequence's 5'-end than the 3'-base of c_2. To specify the hierarchical relationship of c_1 and c_2, a mapping from the structural components to extended loops in the tree needs to be established that will represent the RNA secondary structure. It is obvious that each single base component can be mapped to a unique loop. However, a base pair component can be mapped to up to two alternate loops where one is an ancestor of the other. To resolve this ambiguity, the ancestor loop is chosen as the base pair's

mapping target. Suppose c_1 is mapped to loop e_1 and c_2 is mapped to loop e_2. The hierarchical relationship between c_1 and c_2 is one of the following: (1) c_1 is hierarchically identical to c_2 if e_1 and e_2 are the same; (2) c_1 is an ancestor (descendant, respectively) of c_2 if e_1 is an ancestor (descendant, respectively) of e_2; or (3) c_1 and c_2 are cousins if e_1 and e_2 are cousins or siblings in the tree.

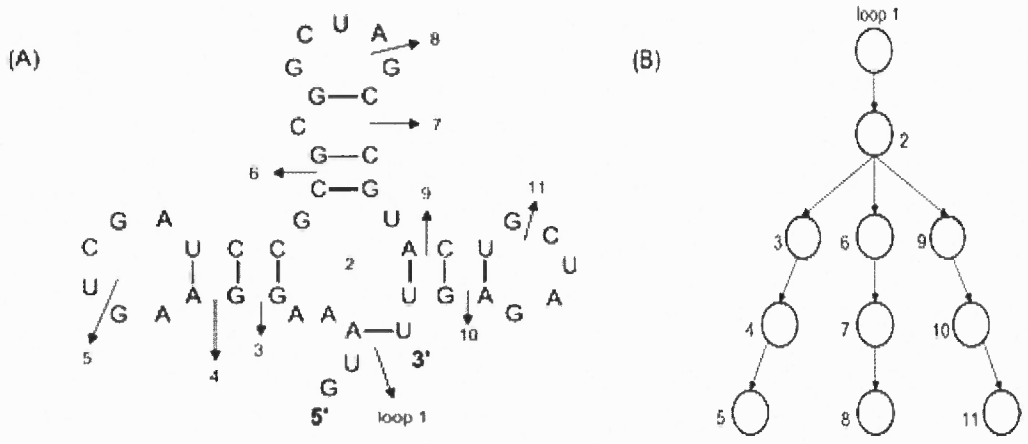

Figure 3.3 (A) A hypothetical RNA secondary structure is decomposed into extended loops. **(B)** The hierarchical tree comprising of the extended loops for the RNA secondary structure in (A).

3.2.2 Partial Structure

A structural component is either a single base or a base pair. The *partial structure* induced by a structural component α, is a set of structural components S_α such that for any structural component $c \in S_\alpha$ the following three conditions are satisfied: (1) c precedes α; (2) c is not an ancestor of α; and (3) α itself belongs to S_α. Furthermore, since

a base pair could appear in two extended loops, the partial structure induced by a base pair could be divided into two smaller substructures: *parent structure* and *child structure* (Figure 3.4). Formally, if the structural component α is a base-pair, its parent structure is a set of components $P_\alpha \subset S_\alpha$ (excluding α) such that for any component $c \in P_\alpha$, c's 3'-base is always 5' upstream of α's 5'-base; its child structure contains a set of components $C_\alpha \subseteq S_\alpha$ (including α itself) such that for any component $c \in C_\alpha$, c's 5'-base is always 3' downstream of α's 5'-base. It can be verified that $P_\alpha \cup C_\alpha = S_\alpha$ and $P_\alpha \cap C_\alpha = \varphi$.

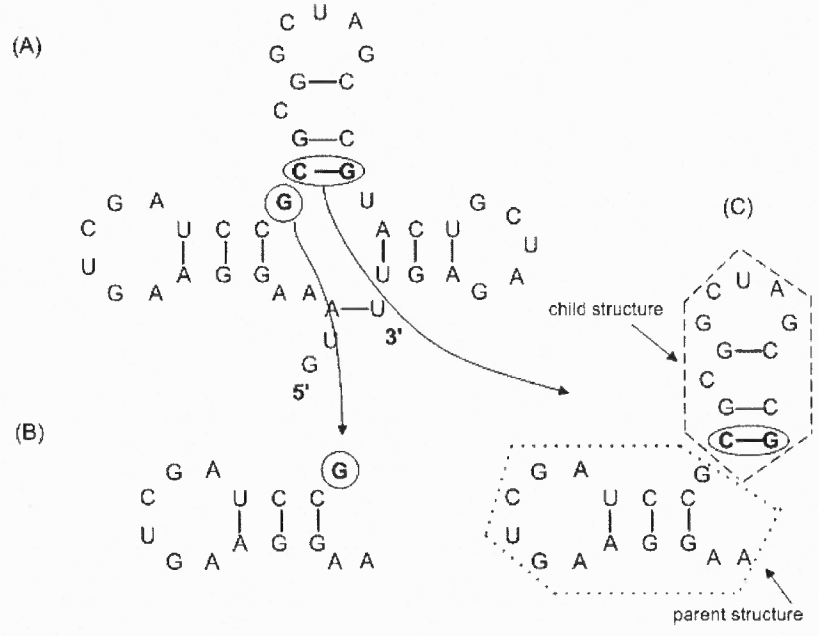

Figure 3.4 (A) A hypothetical RNA secondary structure is used to illustrate how partial structures are determined. **(B)** The partial structure induced by the single base G is shown. **(C)** The partial structure induced by the base-pair C-G consists of 2 parts, a parent structure and a child structure. The base-pair is included in the child structure.

Using the concept of partial structures, the two given RNA secondary structures are progressively aligned using a dynamic programming (DP) algorithm by initially aligning smaller partial structures and expanding each partial structure one structural component at a time. Ultimately, the two partial structures will become the two overall

structures, and the DP scoring table will be fully filled with alignment scores from which we can find the optimal local alignment between the two given RNA secondary structures.

3.2.3 Scoring Scheme

To measure the quality of an alignment, a scoring scheme must be provided. The proposed CSA method leaves great latitude for the choice of various scoring schemes. One important aspect of a scoring scheme is to define an alignment function of two structural components to measure the quality of matching one component to the other. The other important aspect is a penalty parameter, which punishes the action of aligning structural component(s) to gap(s). During the course of computation, one structural component (single base or base pair) could be matched to a gap; or one parent substructure or child substructure could also be matched to a big gap. Intuitively, the bigger the gap, the heavier the penalty is. In this implementation, a basic penalty was used for the smallest gap involving only one base. Then the larger gap is punished proportionally to the number of bases involved in the gap. Let μ denote the basic penalty in the following discussions. Let x be a structural component in the query structure and let y be a structural component in the subject structure. Let $h(x, y)$ denote the alignment score between x and y. This function can be extended to represent the alignment score between two substructures D_Q, D_S from the query structure Q and the subject structure S, respectively, as follows:

$$\varphi(D_Q, D_S) = \sum_{\substack{i \in D_Q \\ j \in D_S}} h(i,j) + \mu.G \tag{3.1}$$

where G represents the total number of gaps in aligning D_Q and D_S.

In calculating the alignment function h, the constraint, or conserved region, annotated in the query structure needs to be considered. Refer to Figure 3.1. Each position of the conserved region in the query RNA structure is marked using a special character '*' underneath the position. This is termed *binary 0/1 conservation* since any position in the query RNA structure is treated to be either 100% conserved (if it is marked with '*') or not conserved at all. If it has been found, from wet lab experiments or other sources, that a particular RNA structure contains a motif that needs to be searched for in other RNA structures from a data set, then that particular RNA structure can be used as a query structure and that motif region can be marked by '*' to indicate that it is conserved in the query structure.

Let $g(\alpha, \beta)$ be the alignment score between two structural components α, β where no constraint is involved. In our implementation presented here, $g(\alpha, \beta)$ is similar to that defined in (3), as shown below:

$$g(\alpha,\beta) = \begin{cases} 1 & \text{if } \alpha, \beta \text{ are single bases and } \alpha = \beta \\ -1 & \text{if } \alpha, \beta \text{ are single bases and } \alpha \neq \beta \\ -2 & \text{if } \alpha \text{ is a single base and } \beta \text{ is a gap, or vice} \\ 3 & \text{if } \alpha, \beta \text{ are base pairs and } \alpha = \beta \\ 1 & \text{if } \alpha, \beta \text{ are base pairs and } \alpha \neq \beta \\ -4 & \text{if } \alpha \text{ is a base pair and } \beta \text{ is a gap, or vice versa} \end{cases} \tag{3.2}$$

The alignment function h in Equation (3.1) is calculated by:

$$h(x,y) = \begin{cases} \lambda g(x',y) & \text{if } x \text{ is constrained} \\ g(x,y) & \text{otherwise} \end{cases} \quad (3.3)$$

where x (y, respectively) is a structural component in the query RNA structure (subject RNA structure, respectively), and λ is used to increase or diminish the score to take into account the conserved region in the query structure. When x is constrained, we use x' to represent the corresponding structural component without the constraint.

With binary 0/1 conservation, λ is defined as

$$\lambda = 1 + \frac{L}{N} \quad (3.4)$$

where L is the length of the conserved region and N is the total length of the query RNA structure.

3.2.4 Recurrence Formulas

This subsection presents scoring formulas for aligning partial structures induced by structural components from the query structure Q and the subject structure S respectively. The recurrence formulas in the proposed dynamic programming algorithm take into account the constraint occurring in the query structure. When a structural component involved in an alignment is a base pair, only the child and partial structures induced by the base pair need to be considered (3). The reason is that the parent structure induced by a base pair can always be derived as a partial structure induced by another structural component and hence is considered when the alignment score of that structural component is calculated (3).

Given the query RNA structure Q and the subject structure S, the proposed CSA method is a dynamic programming (DP) algorithm that matches partial structures from Q and S, respectively. Let x be a single base in Q and let y be a single base in S. Let x^p denote the structural component that precedes x. In matching the partial structure S_x with the partial structure S_y there are three cases: (i) x is aligned with y; (ii) x is aligned with a gap; and (iii) y is aligned with a gap. Thus the score of matching S_x with S_y can be calculated by the following equation:

$$\varphi(S_x, S_y) = \max \begin{cases} \varphi(S_{x^p}, S_{y^p}) + h(x, y) \\ \varphi(S_{x^p}, S_y) + \mu \\ \varphi(S_x, S_{y^p}) + \mu \end{cases} \tag{3.5}$$

where $h(x,y)$ is defined in Equation (3.3) and μ = -2 is the basic penalty for aligning a base with a gap, cf. Equation (3.2).

Next, consider the situation where x is a base pair and y is a single base. (The situation where x is a single base and y is a base pair is similar and hence omitted.) As discussed before, besides the partial structure S_x the child structure C_x for the base pair x also needs to be compared. First the structural alignment score between the child structure C_x and the partial structure S_y is calculated. There are two cases: (i) the single base component y is aligned with a gap; and (ii) the base pair x is aligned with a gap. Therefore,

$$\varphi(C_x, S_y) = \max \begin{cases} \varphi(C_x, S_{y^p}) + \mu \\ \varphi(S_{x^p}, S_y) + 2.\mu \end{cases} \tag{3.6}$$

In aligning the partial structure S_x with the partial structure S_y, there are three cases: (i) the single base y matches with a gap; (ii) the partial structure S_y matches with the child structure C_x; (iii) the partial structure S_y matches with the parent structure P_x.

Thus,

$$\varphi(S_x, S_y) = \max \begin{cases} \varphi(S_x, S_{y^p}) + \mu \\ \varphi(C_x, S_y) + |P_x| . \mu \\ \varphi(P_x, S_y) + |C_x| . \mu \end{cases} \tag{3.7}$$

Then, consider the situation where x is a base pair and y is also a base pair. This requires the computation of four alignment scores because each base pair corresponds to two structures: one child structure and one partial structure. While aligning the child structure C_x with the child structure C_y, it is clear that

$$\varphi(C_x, C_y) = \max \begin{cases} \varphi(S_{x^p}, S_{y^p}) + h(x, y) \\ \varphi(S_{x^p}, C_y) + 2.\mu \\ \varphi(C_x, S_{y^p}) + 2.\mu \end{cases} \tag{3.8}$$

since both x and y are the last components in the respective child structures.

Equation (3.9) gives alignment score between the partial structure S_x and the child structure C_y:

$$\varphi(S_x, C_y) = \max \begin{cases} \varphi(S_x, S_{y^p}) + 2.\mu \\ \varphi(P_x, C_y) + |C_x| . \mu \\ \varphi(C_x, C_y) + |P_x| . \mu \end{cases} \tag{3.9}$$

The first case corresponds to that y is aligned with a gap. If y does not match with a gap, it can be shown that, the second and third cases in Equation (3.9) cover all possible

situations. Similarly, we can calculate the score of aligning the child structure C_x and the partial structure S_y as shown in Equation (3.10):

$$\varphi(C_x, S_y) = \max \begin{cases} \varphi(S_{x^p}, S_y) + 2.\mu \\ \varphi(C_x, P_y) + |C_y|.\mu \\ \varphi(C_x, C_y) + |P_y|.\mu \end{cases} \tag{3.10}$$

In aligning the partial structure *Sx* with the partial structure *Sy*, there are five cases: (i) the parent structure *Px* is matched with the parent structure *Py* and the child structure *Cx* is matched with the child structure *Cy*; (ii) the child structure *Cx* is matched with gaps; (iii) the child structure C_y is matched with gaps; (iv) the parent structure *Px* is matched with gaps; and (v) the parent structure P_y is matched with gaps. Therefore,

$$\varphi(S_x, S_y) = \max \begin{cases} \varphi(P_x, P_y) + \varphi(C_x, C_y) \\ \varphi(P_x, S_y) + |C_x|.\mu \\ \varphi(S_x, P_y) + |C_y|.\mu \\ \varphi(C_x, S_y) + |P_x|.\mu \\ \varphi(S_x, C_y) + |P_y|.\mu \end{cases} \tag{3.11}$$

It can be shown that this CSA method for aligning the query structure Q and the subject structure S allowing constraints to exist in Q has a polynomial time complexity of $O(mn)$ where m is the length of the query structure and n is the length of the subject structure.

3.2.5 Computation of p-Value

To determine what match is likely or unlikely to occur by chance, the computation of a statistical measure, namely a *p*-value, is incorporated into the CSA method (Figure 3.2). In (62) it was showed that in the case of a gapless alignment, the distribution of alignment scores of random sequences is the Gumbel or extreme value distribution (63). However for a gapped alignment, there is no theory that predicts the distribution of alignment scores for random sequences. It has been conjectured based on numerical evidence that the score distribution is still of the Gumbel form (64-66). This assumption is adopted while computing the statistical measure. For the comparison of random sequences of sufficient lengths m and n, the number of distinct local alignments with score at least x is approximately Poisson distributed, with mean

$$E(x) = Kmne^{-\lambda x} \tag{3.12}$$

where λ and K can easily be calculated (62). The optimal alignment score S' follows an extreme-value distribution with

$$\Pr ob(S' \geq x) = 1 - e^{-E(x)} \tag{3.13}$$

Accurate estimation of λ and K is essential to using these equations. The Island method (67,68) has been used to do the estimation. As suggested by this method, first the constrained structural alignments of biologically occurring RNA secondary structures chosen randomly from Rfam (1) is computed. While performing the alignment between

two RNA secondary structures, one of the structures are annotated to be constrained. Thus the scores obtained from the alignments are consistent with the proposed constrained structural alignment scoring scheme. The local alignment results are several locally optimal matches, each being comparable to an island in the large sequence. All the scores that are greater than a threshold c are selected. In this study, the c value is set to 10. The threshold value is chosen such that it is a reasonable score obtained when aligning short RNA motifs of the commonly occurring length. Let the set I_c of such local alignment islands have cardinality R_c and the mean score in excess of c for these islands be S_c:

$$S_c = \frac{\sum_{i \in I_c} [S(i) - c]}{R_c} \tag{3.14}$$

where $S(i)$ is the score of island i. Then the maximum-likelihood estimator for λ is

$$\lambda_c = \ln(1 + \frac{1}{S_c}) \tag{3.15}$$

The maximum likelihood estimator for K is

$$K_c = R_c e^{\lambda_c \cdot \frac{c}{A}} \tag{3.16}$$

where A is the aggregate "area" of the search space from where the local alignments are taken. If a single pair of structures of length m and n is used, then $A = mn$. If B such comparisons are performed, then $A = Bmn$. Once λ_c and K_c are determined, these values are used to calculate the p-value for an alignment score x by plugging λ_c and K_c in Equations (3.12) and (3.13). The p-value is the probability, by chance, that there is another alignment with a similarity score greater than or equal to the score x. The p-value

is a measure of the reliability of the score x. The smaller the p-value, the more reliable x is.

3.4 Experiments and Results

The proposed constrained structural alignment method was tested by detecting internal ribosome entry sites in the RNA sequences of *T. brucei* and hepatitis C virus respectively. An internal ribosome entry site (IRES) is a nucleotide sequence which functions to allow for translation initiation in the non-coding region of an mRNA sequence (69). An IRES element is able to attract the eukaryotic ribosome to close vicinity of a start codon and thus to initiate its translation. The secondary structure of an internal ribosome entry site in *T. brucei* mRNA sequences is portrayed in Figure 3.5.

Two different datasets were used in these experiments. For the first dataset $D1$, 20 non-redundant untranslated regions (UTRs) of *T. brucei* mRNA sequences that contain internal ribosome entry sites were extracted from UTRdb (70). These IRES containing mRNA sequences, listed in Table 3.1, formed the positive data for the dataset $D1$. Their lengths are in the range 85-993 nt. The presence of IRESs in these sequences was suggested by UTRscan (70) which is a sequence analysis tool provided by UTResource. UTRscan analyzes user-submitted sequences for the functional elements defined in the UTRsite database of UTResource. Notice that even though the 20 UTRs of *T. brucei* mRNA sequences contain internal ribosome entry sites, there are no known conserved secondary structures, or structural motifs, in the IRES-containing UTRs. Also, 30 other sequences were added from UTRdb that were not known to contain internal ribosome entry sites. These 30 sequences formed the negative data for the dataset $D1$. All these 50

sequences were folded using RNAfold (27). Finally, 5 of the 20 IRES-containing *T. brucei* sequences were randomly selected and the IRES-containing region in each of the 5 sequences was extracted. These IRES-containing regions were separately folded using RNAfold and that formed the query structures in our experiment involving *D*1.

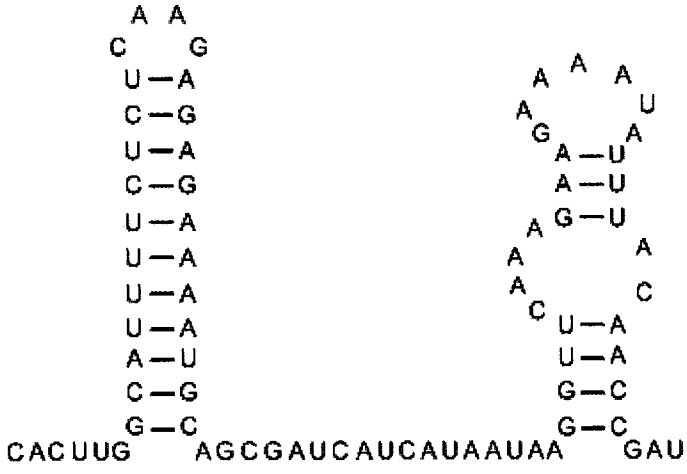

Figure 3.5 The secondary structure of an internal ribosome entry site in T. brucei mRNA sequences.

Table 3.1 The 20 IRES-containing *T. brucei* UTR Sequences used as Positive Data in D_1

EMBL accession number	Description	IRES start	IRES end
AB033824	5'UTR in *T. brucei* GPI10 mRNA for GPI anchor biosynthesis protein, complete cds.	5	92
AF007547	5'UTR in *T. brucei* Trab5B mRNA, complete cds.	73	158
AF049901	5'UTR in *T. brucei* rhodesiense prohibitin mRNA, complete cds.	72	166
AF068705	5'UTR in *T. brucei* rhodesiense transferring-binding protein (ESAG 6-d) mRNA, complete cds.	475	558
AF101480	5'UTR in *T. brucei* pf20 homolog (TWD1) mRNA, complete cds.	1	101
AF189284	5'UTR in *T. brucei* nucleolar G-protein NOG1 (NOG1) mRNA, complete cds.	168	254
AF226674	5'UTR in *T. brucei* 20S proteasome beta 5 subunit (PSB5) mRNA, complete cds.	267	346
AF301417	5'UTR in *T. brucei* procyclin-associated gene 2 polypeptide (PAG2), procyclin-associated gene 4 polypeptide (PAG4), GU2 (GU2), and GU1 (GU1) genes, complete cds.	9178	9265
AF404116	5'UTR in *T. brucei* proteasome regulatory non-ATP-ase subunit 8 (Rpn8) mRNA, complete cds.	135	235
AJ242519	5'UTR in *T. brucei* mRNA for cyclin 2 (CYC2 gene)	6	103
AM159084	5'UTR in *T. brucei* mRNA for RNA polymerase I subunit RPA12 (RPA12 gene)	3	97
AM159570	5'UTR in *T. brucei* mRNA for RNA polymerase I subunit RPC40 (RPC40 gene)	213	308
AY157028	5'UTR in *T. brucei* putative G1 cyclin CycE2 mRNA, complete cds.	124	217
AY157032	5'UTR in *T. brucei* putative mitotic B-type cyclin CycB3 mRNA, complete cds.	142	239
AY370775	5'UTR in *T. brucei* strain Lister 427 Rab23 mRNA, complete cds.	22	116
K02198	5'UTR in *T. brucei* spliced leader mRNA (pSLc4) from procyclic stage.	11	109
K02945	5'UTR in *T. brucei* gambiense calmodulin mRNA 2 with a spliced leader sequence.	15	104
L03777	5'UTR in *T. brucei* protein kinase (nrkB) allele nrkB-2 mRNA, complete non-functional cds and alleles nrkB-1 and nrkB-3.	901	993
U18329	5'UTR in *T. brucei* small GTP-binding protein mRNA, clone rtb9, complete cds.	75	157

Table 3.1 The 20 IRES-containing *T. brucei* UTR Sequences used as Positive Data in D_1 (Continued)

EMBL accession number	Description	IRES start	IRES end
U80910	5'UTR in T. brucei ribonucleotide reductase large subunit (RNR1) mRNA, complete cds.	8	85

Table 3.2 The 20 IRES-containing HCV Sequences used as Positive Data in D_2

EMBL accession number	Description	IRES start	IRES end
AF021888	HCV strain GE 174 5' non-coding region type 1a	1	190
AF021898	HCV strain GE 56 5' non-coding region type 4	1	190
AF021904	HCV strain SL 34 5' non-coding region type 1a	1	190
AF034628	HCV type 3 5' noncoding region, partial sequence	2	253
AF041264	HCV isolate 498 5' untranslated region	1	191
AF041266	HCV isolate 611 5' untranslated region	1	191
AF041267	HCV isolate 614 5' untranslated region	1	191
AF041300	HCV isolate 966 5' untranslated region	1	191
AF055303	HCV type 1a strain CHCH3 5' untranslated region, partial sequence	1	240
AF055305	HCV type 1a strain CHCH5 5' untranslated region, partial sequence	1	239
AF041309	HCV isolate 982 5' untranslated region	1	191
AF041329	HCV type 2c isolate 760 5' noncoding sequence and core protein gene, partial cds	1	267
AF056005	HCV type 1b strain CHCH6 5' untranslated region, partial sequence.	1	237
AF055301	HCV type 1a 5' untranslated region, partial sequence.	1	238
AF057147	HCV type 2b strain CHCH13 5' untranslated region, partial sequence.	1	240
AF057150	HCV type 3a strain CHCH16 5' untranslated region, partial sequence.	1	237
AF077228	HCV isolate patient 20 5' non-coding region, partial sequence	1	250
AF141989	HCV isolate 8-63 polyprotein mRNA, 5' untranslated region, partial sequence	1	195
AF216795	HCV isolate SOM1 5'UTR, partial sequence	3	205
AF217298	HCV clone Sot10 5'UTR sequence	1	256

The second dataset *D*2 was made up of 20 non-redundant hepatitis C virus (HCV) sequences, which contained internal ribosome entry sites, from Rfam (1). These sequences belong to the IRES_HCV family in Rfam. Table 3.2 lists these sequences, which formed the positive data for the dataset D2. Their lengths are in the range 190-267 nt. In Rfam, these 20 HCV sequences share a consensus or conserved secondary structure. Another 30 sequences were taken from UTRdb and added to the dataset *D*2. These 30 sequences did not belong to the hepatitis C virus, and were not known to contain internal ribosome entry sites. These 30 sequences formed the negative data for the dataset *D*2. All these sequences were folded using RNAfold (27). Separately, 5 of the 20 IRES containing HCV sequences were randomly selected from the dataset *D*2 just the IRES region from each of the 5 sequences was extracted. These IRES containing regions were folded using RNAfold (27). The resulting 5 structures formed the query structures in the experiment involving *D*2.

On these two datasets *D*1 and *D*2 , constrained structural alignment (CSA) method was applied, first using the binary conservation option (0/1 constraints), and then using sequence logos, by aligning each of the 5 selected query structures one by one with the RNA secondary structures in *D*1 and *D*2, respectively. (For the binary conservation option, every base in a query structure was marked with "*"). For comparison purposes, two other methods were also applied to the same datasets. They were the regular pair-wise structural alignment method without constraints offered in RSmatch (3) and the RNAforester structural alignment method (48). Thus, a database search was carried out with each of these alignment methods by aligning the corresponding query structures one by one with the subject structures in *D*1 and *D*2, respectively. Then, from the top 20 hits,

i.e. the top 20 RNA subject structures with the largest alignment scores, in a search result, the true positives and false positives were computed. True positives are those hits in which an internal ribosome entry site is actually present. False positives are those hits that appear in the search result as containing internal ribosome entry sites, though in reality they are not known to contain internal ribosome entry sites. The error rate *(e),* defined below, is used to evaluate the effectiveness of an alignment method:

$$e = \frac{FP}{TP+FP} \tag{3.17}$$

where *TP* is the number of true positives, *FP* is the number of false positives, and *TP* + *FP* = 20 in these experiments.

Table 3.3 shows the results and presents the average error rate obtained from using the 5 different *T. brucei* query structures for each alignment method. Table 3.4 presents this data for the 5 different queries belonging to the HCV dataset. As can be seen from the tables, the proposed CSA method with 0/1 constraints gives the lowest average error rate, outperforming the other three alignment techniques. These results were obtained by using the optimal structure for each sequence. The alignment algorithms were also compared by using twenty percent suboptimal structures for each sequence, and the qualitative conclusion remains the same.

It was observed that there is little similarity shared by the IRES-containing *T. brucei* sequences. The average pairwise sequence identity for the 20 IRES-containing *T. brucei* sequences is 29%. This explains why the three alignment algorithms have high error rates for the *T. brucei* dataset (Table 3.3). On the other hand, the 20 IRES-containing HCV sequences are conserved at both the sequence and the secondary

structure level. The average pairwise sequence identity for the 20 IRES-containing HCV sequences is 88%. Under this circumstance, all the three alignment algorithms have good performance; the algorithms have much lower error rates for the HCV dataset (Table 3.4) than for the *T. brucei* dataset (Table 3.3).

Table 3.3 The Average Error Rate Calculated by using 5 *T. brucei* Queries against the Dataset D_1

Query	CSA with 0/1 Constraints		RSmatch		RNAforester	
	TP*	FP*	TP	FP	TP	FP
Q1	14	6	15	5	11	9
Q2	11	9	10	10	11	9
Q3	11	9	10	10	12	8
Q4	12	8	11	9	10	10
Q5	12	8	10	10	12	8
Average error rate	0.40		0.44		0.44	

* TP = True positive, FP = False positive that occur in the top 20 hits of a search.

Table 3.4 The Average Error Rate Calculated by using 5 HCV Queries against the Dataset D_2.

Query	CSA with 0/1 Constraints		RSmatch		RNAforester	
	TP*	FP*	TP	FP	TP	FP
Q1	18	2	18	2	16	4
Q2	20	0	20	0	17	3
Q3	19	1	17	3	18	2
Q4	20	0	20	0	18	2
Q5	19	1	19	1	17	3
Average error rate	0.04		0.06		0.14	

* TP = True positive, FP = False positive that occur in the top 20 hits of a search.

From Table 3.3, it can be seen that among the top 20 hits, the CSA method with 0/1 constraints found 11-14 positive structures and 6-9 negative structures. The consensus of the found positive structures may suggest a conserved secondary structure or structural motif in the *T. brucei* UTRs. Figure 3.6 shows the consensus secondary structure together with its Vienna style Dot Bracket representation of the top 10 positive structures most similar to query Q1 in Table 3.3 according to the CSA method with 0/1 constraints. The consensus secondary structure is computed by the multiple structural alignment (MSA) function of the RADAR tool (5). For the HCV data in Table 3.4, the consensus secondary structure found by the proposed constrained structural alignment method in combination with RADAR's MSA function is consistent with that documented in Rfam (1).

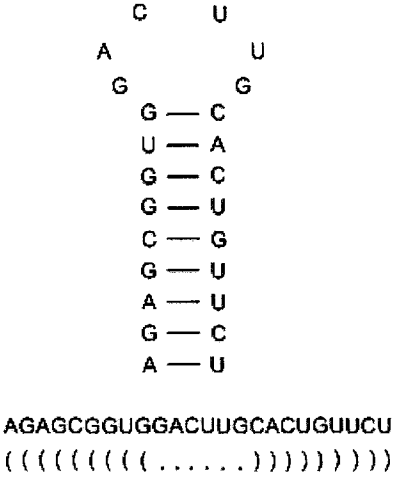

Figure 3.6 A putative structural motif in *T. brucei* UTRs obtained from the multiple structural alignment of the top 10 positive structures that occurred in the search result of query Q1 in Table 2.3 using the proposed CSA method with 0/1 constraint.

3.5 Conclusions

Here a constrained structural alignment algorithm for matching two RNA secondary structures was introduced. A statistical measure was developed for assessing the significance of alignment scores. The proposed techniques are applied to searching for internal ribosome entry sites in RNA sequences of *T. brucei* and hepatitis C virus, respectively. For the HCV sequences, there is a known consensus secondary structure, as documented in Rfam (1), and our method accurately detected the consensus secondary structure in the HCV sequences. For the *T. brucei* sequences, there is little similarity shared by their IRES containing sequences, and our experimental results suggested the possible existence of a conserved secondary structure in the IRES containing *T. brucei* sequences. The results also showed the superiority of the proposed techniques over existing methods.

CHAPTER 4

RADAR: A WEB SERVER FOR RNA DATA ANALYSIS AND RESEARCH

RADAR is a web server that provides a multitude of functionality for RNA data analysis and research. It can align structure-annotated RNA sequences so that both sequence and structure information are taken into consideration during the alignment process. This server is capable of performing pairwise structure alignment, multiple structure alignment, database search and clustering. In addition, RADAR provides two salient features: (i) constrained alignment of RNA secondary structures, and (ii) prediction of the consensus structure for a set of RNA sequences. RADAR will be able to assist scientists in performing many important RNA mining operations, including the understanding of the functionality of RNA sequences, the detection of RNA structural motifs and the clustering of RNA molecules, among others.

The web server together with a software package for download is freely accessible at http://datalab.njit.edu/biodata/rna/RSmatch/server.htm.

4.1 Introduction

The web server, RADAR (acronym for RNA Data Analysis and Research), performs a multitude of functions related to RNA structure comparison, including pair-wise structure alignment, constrained structural alignment, multiple structure alignment, database search, clustering and consensus structure prediction. The aim behind developing this web server was to have a versatile tool that provides a computationally efficient platform

for performing several tasks related to RNA structure. RADAR has been developed using Perl-CGI and Java. In each run, the server can accept at most 50 RNA sequences or secondary structures for pair-wise structure alignment and constrained structural alignment and at most 10 RNA sequences or secondary structures for the other functions where each sequence or structure has at most 300 bases, though the downloadable version does not have this restriction. For the sample data provided by the server, it takes a few seconds for most of the server's functions to complete and display results on the web. It takes about one minute to produce a multiple structure alignment when RNA sequences are fed as input. The database search function needs several minutes to search the Rfam database (1); the results of this function are returned to the user via email, rather than on the web.

4.2 Method

RADAR employs the RSmatch algorithm (3) for computing the alignment of two RNA secondary structures. Briefly, it decomposes each RNA secondary structure into a set of basic structure components that are further organized by a tree model. With this model, pseudoknots are not allowed. A dynamic programming algorithm is employed to align the two RNA secondary structures. RSmatch is capable of performing both global and local alignment of two RNA secondary structures. The time complexity of the algorithm is O(mn), where m and n are the sizes of the two structures, respectively. This method is an efficient solution to the problem of RNA structure alignment. By using this structure comparison algorithm, other functionalities were developed such as pair-wise structure alignment, multiple structure alignment, database search, clustering, constrained

structural alignment and consensus structure prediction, and incorporated into RADAR. Pair-wise structure alignment involves the alignment of a query structure with each of the subject structures in a set. Multiple structure alignment uses the same alignment algorithm along with a position specific scoring matrix to build up an alignment by including one structure at a time until no appropriate structure can be included in the alignment (3). Database search is done by aligning a query structure one by one with the consensus structures of the non-coding RNA families stored in the release 8.0 of Rfam (1) to find the consensus structures similar to the query structure. This function returns the top *k* hits as the search result, where *k* is an adjustable parameter. Clustering is done to compute and display a similarity matrix for a set of RNA secondary structures. A constrained version of RNA structure alignment has been developed to improve the sensitivity of the alignment (as described in Chapter 3). This allows the user to annotate a region of an input RNA structure as conserved. The conserved region, or constraint, is incorporated into the alignment process to produce biologically more meaningful alignment results. RADAR also includes a novel method for computing the structure of a group of closely related RNA sequences. This method is explained below.

4.2.1 Consensus Structure Prediction

This method works in four steps, as described below:

i. Determine individual RNA structures: For the input RNA sequences, compute their structures having energies that fall within a particular range of the minimum energy using the Vienna RNA package's RNAsubopt function (27). Therefore, for

each sequence there can be more than one possible structure. The result consists of the predicted RNA structures for all the RNA sequences in the input file.

ii. Compute a pair-wise scoring matrix: In this step, the pair-wise alignment scores between all structures except for the structures that represent the same RNA sequence are computed. The result is a matrix that gives the score of alignment for every pair of structures. The score of comparison between RNA structures of the same sequence is set to 0, since these structures are for the same RNA sequence and so they are treated as being very close to each other.

iii. Select one structure for each RNA sequence: From the matrix produced in step ii, select the pair of structures which have the best score. These structures are then said to be the chosen structures for the RNA sequences they correspond to. The pair-wise scoring matrix is modified to eliminate all the other structures of these RNA sequences. Once again the same process of selecting the best pair of structures and then eliminating the other structures of the sequences they belong to is carried out. This is repeated until we a structure is selected for each of the input sequences.

iv. Predict the common RNA substructure: This step deals with predicting the consensus RNA substructure that is common to as many RNA sequences in the input file as possible. This is obtained by computing a multiple structure alignment of the RNA structures selected in step iii.

4.3 Web Server

The RADAR web server together with a standalone downloadable version is freely available at http://datalab.njit.edu/biodata/rna/RSmatch/server.htm.

4.3.1 Input

RADAR accepts, as input data, either RNA sequences in the standard FASTA format or RNA secondary structures in the Vienna style Dot Bracket format (27). The input data can be stored in a file to be uploaded to the server or entered directly into the text boxes provided by the server. Figure 4.1 shows the input interface of RADAR for aligning an RNA secondary structure with a set of subject structures. When RNA sequences are fed as input, RADAR invokes Vienna RNA v1.4 (27) to fold the sequences into RNA secondary structures. Based upon the function chosen, there are different alignment parameters such as gap penalty, scoring matrix, alignment type (global or local) or folding parameters such as minimum free energy, sliding window size, etc. that can be customized by the user. For performing constrained structural alignment, it is required that users annotate the query RNA structure to indicate which region is conserved by marking the region with '*'.

4.3.2 Output

Upon completion of a structure alignment job, RADAR presents the alignment result on a web page where the alignment result can be downloaded to a file on a local machine. In Figure 4.2, the common region of two RNA secondary structures given in an alignment

result is portrayed using RnaViz (71), where the local matches in the two structures are highlighted with the green color. For pair-wise structure alignment and constrained structural alignment, RADAR ranks the subject structures in the given set based upon their similarities to the query structure. The top ranked subject structure is most similar to the query structure, with the maximum alignment score. The score diminishes as the quality of the alignment decreases.

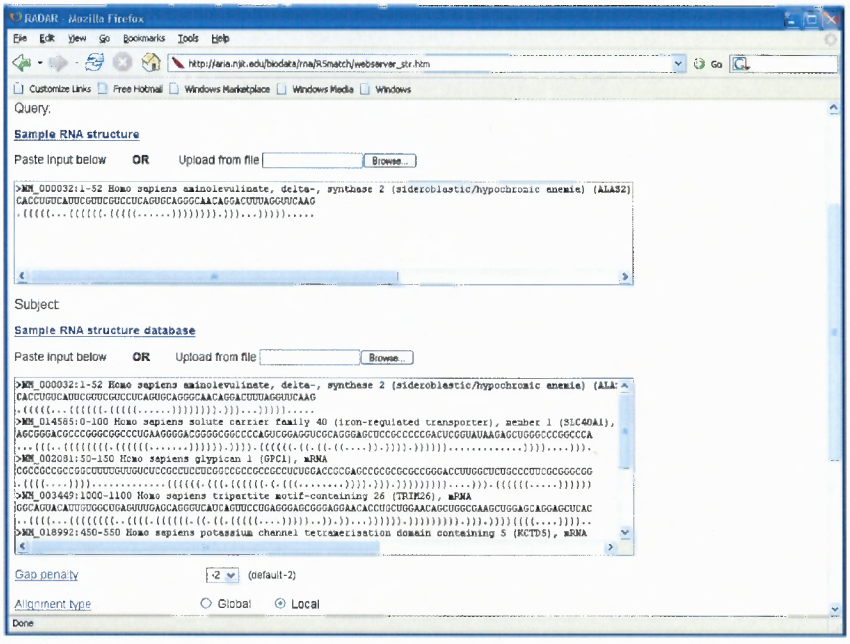

Figure 4.1 The input interface of RADAR for aligning an RNA secondary structure with a set of subject structures.

In the case of a multiple structure alignment, RADAR displays the set of RNA structures followed by the best possible multiple structure alignment obtained from them. Inference can be made from looking at this alignment about the presence of a common motif. Such knowledge greatly helps in studying and discovering the functionalities of

RNA molecules. Figure 4.3 shows an example of the output from the RADAR function to predict the consensus structure for a set of RNA sequences.

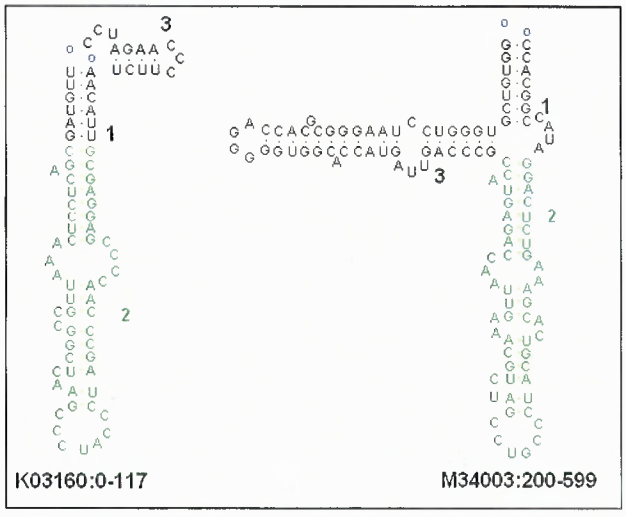

Figure 4.2 Figure illustrates a common region between two RNA secondary structures with green color.

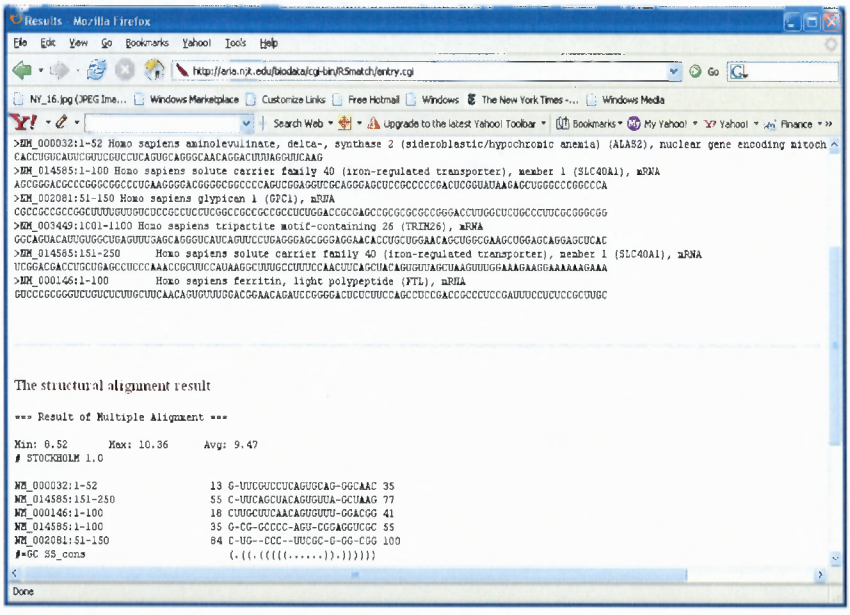

Figure 4.3 Sample output from RADAR's consensus-structure prediction function for a set of RNA sequences. The result shows a group of subsequences from the input that share a common structure. Here the common structure is that of the IRE motif (72).

The input sequences are shown on the top of the figure and the consensus structure is shown at the bottom of the figure. The consensus structure is that of an iron response element (IRE) (72) and all the input sequences are known IRE-containing sequences. The IRE motif is displayed as a multiple structure alignment where the alignment shows the positions at which the motif occurs in each input sequence. These positions indicate the offsets within a sequence. For example, in NM_014585:151-250, the motif begins at the 55th position and ends at the 77th position of the sequence.

4.4 Conclusions

The RADAR web server provides multiple capabilities for RNA structure alignment data analysis, which includes pair-wise structure alignment, multiple structure alignment, constrained structural alignment, database search, clustering and the prediction of a consensus RNA structure from structure alignments for a set of RNA sequences. The web server is implemented in Perl-CGI, rather than SOAP, and hence it requires human–computer interaction.

CHAPTER 5

DETECTING CONSERVED RNA SECONDARY STRUCTURES IN VIRAL GENOMES: THE RADAR APPROACH

5.1 Introduction

Conserved regions, or motifs, present among RNA secondary structures serve as a useful indicator for predicting the functionality of the RNA molecules. Automated detection or discovery of these conserved regions is emerging as an important research topic in health and disease informatics. In practice, biologists favor integrating their knowledge about conserved regions into the alignment process to obtain biologically more meaningful similarity scores between RNAs. Constrained alignment method (described in chapter 3) was used for detecting conserved regions in RNA secondary structures of some viral genomes. The experimental results show that the proposed approach is capable of efficiently detecting conserved regions in the viral genomes and is comparable to existing methods.

5.2 Implementation and Experimental Results

Several experiments have been conducted to evaluate the performance of the proposed constrained structural alignment algorithm by applying this method to finding structural motifs in viral genomes. Study of viral genomes has shown that they often contain functionally active RNA structural motifs that play an important role in the different stages of the life cycle of the virus (73). Detection of such motifs or conserved regions would greatly assist the study of these viruses.

One of experiments designed was to search for a short GC–rich hairpin (tetraloop) which follows an unpaired GGG element, shown in Figure 5.1, present at the 5' end of the *Levivirus* genome (73). Constrained structural alignment algorithm, with the binary conservation option, was applied to a dataset comprising 6838 RNA structures each with length 200 nt formed from ten *Levivirus* genomes and four other randomly selected viral genomes. The query structure used was the GC–rich hairpin. There were ten structures in this dataset containing the region of interest. The algorithm was able to correctly identify 8 out of the 10 structures. The same experiment was repeated using the non-constrained alignment method of RSmatch (3), and it could identify only 6 out of the 10 structures. These six structures were part of the eight structures found by the constrained structural alignment (CSA) algorithm. This shows that the CSA method improves upon the performance of the existing RSmatch method and has a better sensitivity. The Infernal tool (45) was also applied to this same viral genome dataset. Infernal also detected only 6 out of the 10 structures. Again, these six structures were part of the eight structures found by the CSA method.

Figure 5.1 The secondary structure of a GC–rich hairpin that is found to be conserved within the *Leviviridae* family (73).

5.3 Conclusions

The application of the constrained structural alignment algorithm to viral genomes demonstrates the use of the algorithm in RNA informatics research and its ability to detect conserved regions in RNA secondary structures. The work described here is part of a large project aiming to build a cyber infrastructure (http://datalab.njit.edu/bioinfo) for RNA structural motif discovery in human, virus and trypanosome mRNAs. Human immunodeficiency virus type 1 is the causative agent of AIDS and is related to many cancers in humans. Hepatitis C virus is related to hepatocellular cancer in humans. *Trypanosoma brucei* causes African trypanosomaiasis, or sleeping sickness, in humans and animals in Africa. RNA motifs or conserved structures have been shown to play various roles in post-transcriptional control including mRNA translation, mRNA stability, and gene regulation, among others. This cyber infrastructure will contribute to integrated genomic and epidemiological data analysis, by enabling access, retrieval, comparison, analysis, and discovery of biologically significant RNA motifs through the Internet as well as the integration of these motifs with online biomedical ontologies.

CHAPTER 6

THE STRENGTH OF A POLYADENYLATION SITE IS INFLUENCED BY THE STRUCTURAL STABILITY OF THE SURROUNDING REGION AND ITS DISTANCE FROM THE NEIGHBORING GENE

Polyadenylation is a crucial step towards the maturation of almost all cellular mRNAs in eukaryotes. Studies have identified several cis-elements besides the widely known polyadenylation signal (PAS) element (AATAAA or ATTAAA or a close variant) which may have a role to play in polyA site identification. This study investigated the differences in structural stability of sequences surrounding poly(A) sites. It was found that for the genes containing single poly(A) site, the surrounding sequence is most stable as compared with the surrounding sequences for genes with alternative poly(A) sites. This suggests that structure may be providing some evolutionary advantage for genes containing a single poly(A) sites that prevents other poly(A) sites from arising. In addition this research shows that the structural stability of the region surrounding a polyadenylation site correlates with its distance from the closest neighboring gene. The shorter the distance, higher was the structural stability.

6.1 Background

Polyadenylation is an important post-transcriptional regulation step towards the generation of mature mRNA transcripts that can be translated to proteins (74). This is a two step process that includes a specific cleavage at the 3' end of nascent mRNA and then the addition of poly(A) tail (75). The poly(A) tail is located at the 3'-end of all mature mRNAs except some histone genes (18,74), and is critical for many aspects of mRNA metabolism, including mRNA stability, translation, and transport (76,77).

The polyadenylation process involves the use of two major components: the cis-elements or poly(A) signals of the pre-mRNA, and the trans-acting factors that carry out the cleavage and the addition of the poly(A) tail at the 3'-end (78). Sequences flanking the poly(A) site is called the poly(A) region. Several cis-elements residing near to poly(A) sites have been found to promote polyadenylation. A hexamer AAUAAA or AUUAAA or a close variant, usually referred to as the polyadenylation signal (PAS), is located 10-35 nt upstream of most human poly(A) sites (79). In addition, TGTA, TATA, G-rich and C-rich elements in upstream or downstream regions have been implicated in regulation of polyadenylation by different experimental and/or bioinformatics studies (19,80,81). Some studies have also identified RNA structure to be a critical determinant of poly(A) site definition (82,83). Here, the primary goal was to further investigate the role played by RNA secondary structure in polyadenylation and to study the different types of poly(A) sites for factors that affect their strength.

More than half of all human genes have been found to contain multiple poly(A) sites (79,84), which leads to alternative gene products, while others have only a single poly(A) site. The multiple poly(A) sites can be located downstream of the stop codon in the 3'-most exon, leading to transcripts with variable 3'-untranslated regions (UTRs), or in internal exons, leading to transcripts with variable protein products and 3'-UTRs (85). In this study, the analysis deals with the genes that contain only one poly(A) site, referred henceforth as S-type poly(A) sites, and with genes that have multiple poly(A) sites downstream of the stop codon in the 3'-most exon. The alternative poly(A) sites are further classified into three types as follows: 5' most poly(A) site is referred as F-type, 3'

most poly(A) site is referred as L-type and all other sites between these two are referred as M-type (Figure 6.1).

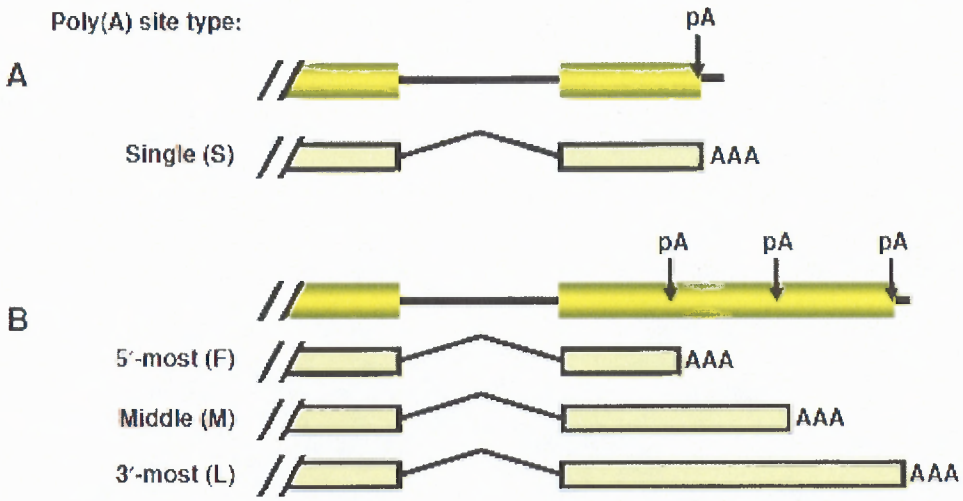

Figure 6.1 Different types of poly(A) sites classified according to their location in the gene. **(A)** Single poly(A) sites (S). **(B)** Sites located in the 3'-most exon are classified into 5'-most site (F), middle site (M) and 3'-most site (L).

Source: Lee, J.Y., Ji, Z. and Tian, B. (2008) Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Res*, **36**, 5581-5590.

6.2 Results

All poly(A) sites used in this analysis have been obtained from the PolyA_DB 2 database (86). This study only deals with the poly(A) sites of type: Single (S), First (F), Middle (M) and Last (L) (Section 6.3.1).

6.2.1 Structural Stability of the Poly(A) Region for the Different Types of Poly(A) Sites

Sequences flanking the poly(A) sites (-200 to +200 and -100 to +100) of type: S, F, M and L are extracted and their minimum folding free energy (mfe) is computed using RNAfold (27). Each set of poly(A) sites are divided into conserved and non-conserved sites (Section 6.3.2) and the density plots of mfe were obtained separately for the conserved and non-conserved set of each type. From the distribution of the mfe for all the conserved polyA sites of different types it was observed that the S-type polyA site sequences showed the maximum stability (Figure 6.2(A-C)).The mfe for the conserved S-type sequences is significantly less than that of the other types (Wilcoxon test p-value < 2.2e-16 when compared with conserved L-type). The same pattern was also observed for the poly(A) region of mouse poly(A) sites (Figure 6.3). It also indicated that the S-type poly(A) region is most stable.

Furthermore the stability of conserved poly(A) regions is more than that of non-conserved (Figure 6.4). Comparison of the observed stability of S-type sequences with the expected stability (using 1-order Markov randomized S-type sequences) shows a clear bias of the observed data towards lower energy (Figure 6.2(D)). Both KS test and

Wilcoxon test showed that the observed stability of S-type sequences is significantly more than the expected value (KS test E-value=0, Wilcoxon test p-value = 0.0018).

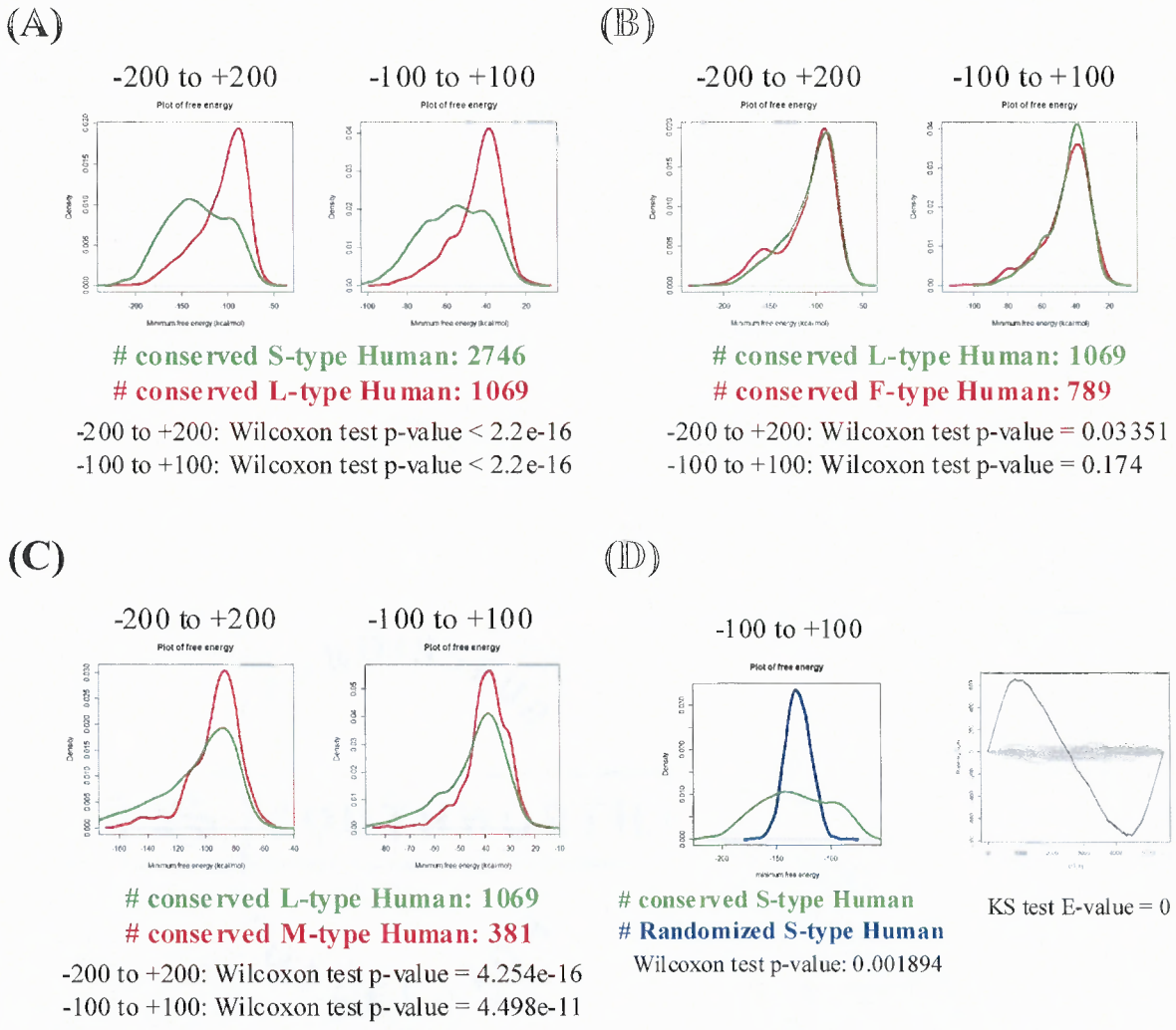

Figure 6.2 Comparison between the minimum free energy distribution of the poly(A) region surrounding conserved Human poly(A) sites: **(A)** between S and L-type **(B)** between L and F-type **(C)** between L and M-type **(D)** between observed and expected distribution for S-type. Wilcoxon and mKS tests are used to provide the significance of the difference.

Note: This work was performed in collaboration with members of Prof. Bin Tian's research group at UMDNJ.

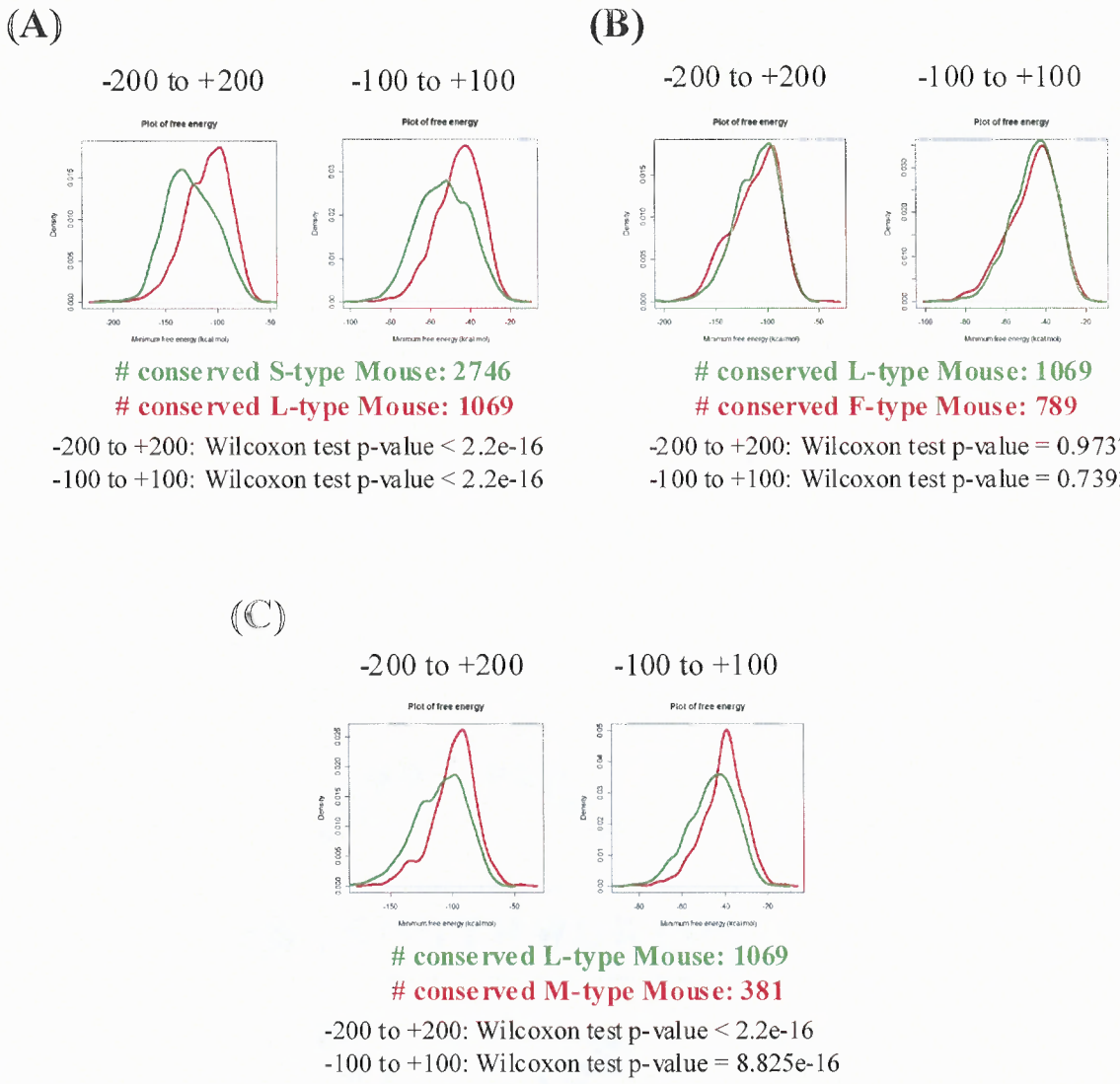

Figure 6.3 Comparison between the minimum free energy distribution of the poly(A) region surrounding conserved Mouse poly(A) sites: **(A)** between S and L-type **(B)** between L and F-type **(C)** between L and M-type.

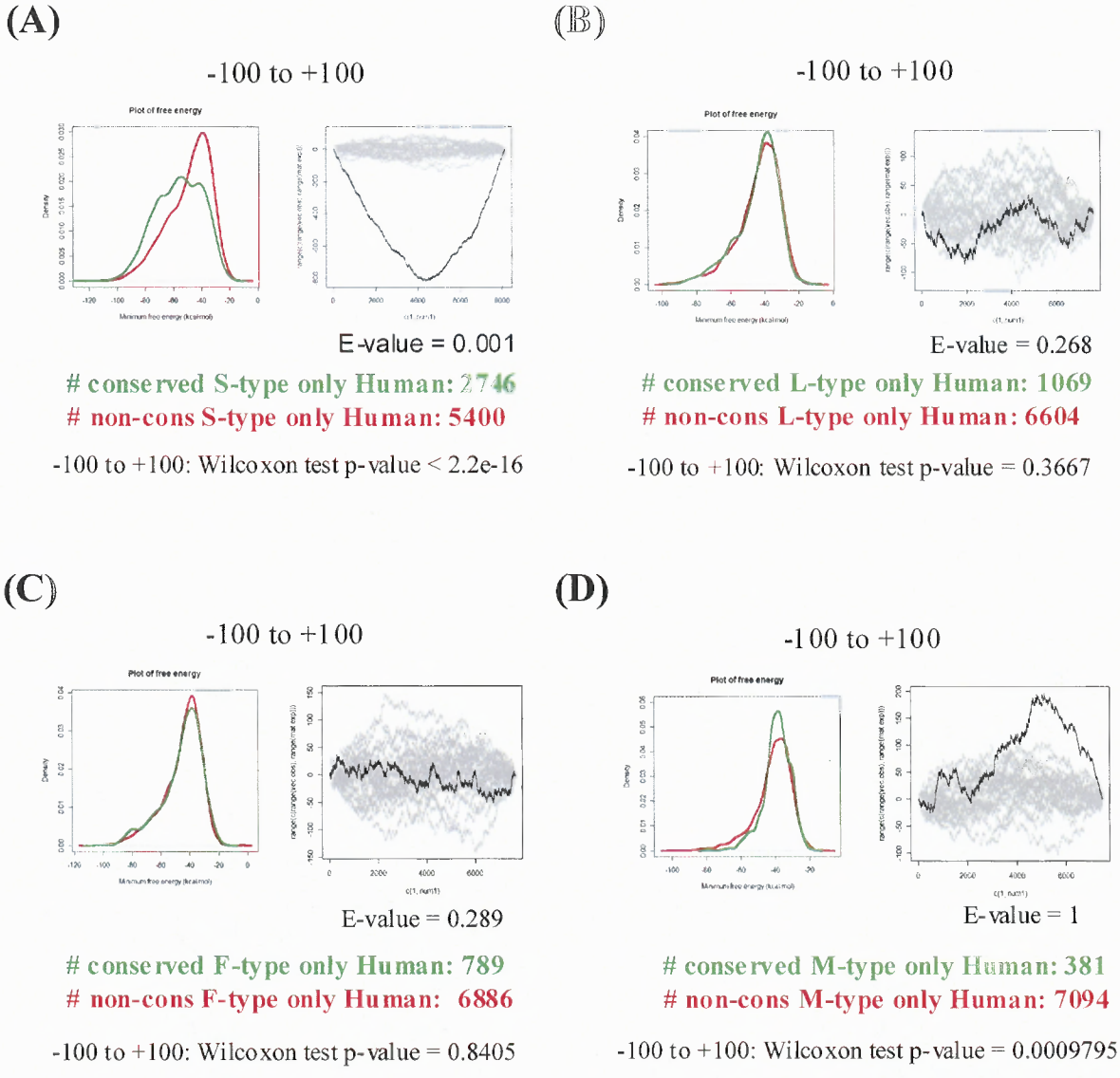

Figure 6.4 Comparison between the minimum free energy distribution of the poly(A) region surrounding conserved sites with that of the non-conserved sites. Wilcoxon and mKS test were used to provide the significance of the difference between these distributions.

6.2.2 Structural Differences between different Regions surrounding the Poly(A) Sites

Next, the poly(A) region was divided into subsequences 50 nt long upstream and downstream of the polyA site (-200 to +200). The resulting 8 regions are labeled as a, b, c, d, e, f, g and h as can be seen in Figure 6.5. Then the number of base pairs between each of these regions is found. Base pairing is an indicator for judging the capability of the region to form stable structures and so it was used to find out how the stability varies for the regions surrounding the poly(A) site. Figures 6.5(A-D) show heat maps of the ratio of observed to expected (using 1-order Markov randomized sequences) average base pairing among the different regions for each type of conserved poly(A) site. An interesting observation from this is that the "d" region seems to have a general avoidance for structure with other regions except with itself and with region "e". This avoidance is most pronounced for the S-type poly(A) site sequences.

To further verify the above results, the free energy contributed by each region towards the free energy of the structure for the entire sequence was calculated using RNAeval (27). Here again it was observed that there was higher energy (hence less structural stability) when the region "d" is involved (Figure 6.5(E-H)). It's also seen that the conserved S-type sequences have overall lower energies than the others which again reiterates the previous result that showed S-type poly(A) regions to have the least minimum free energy.

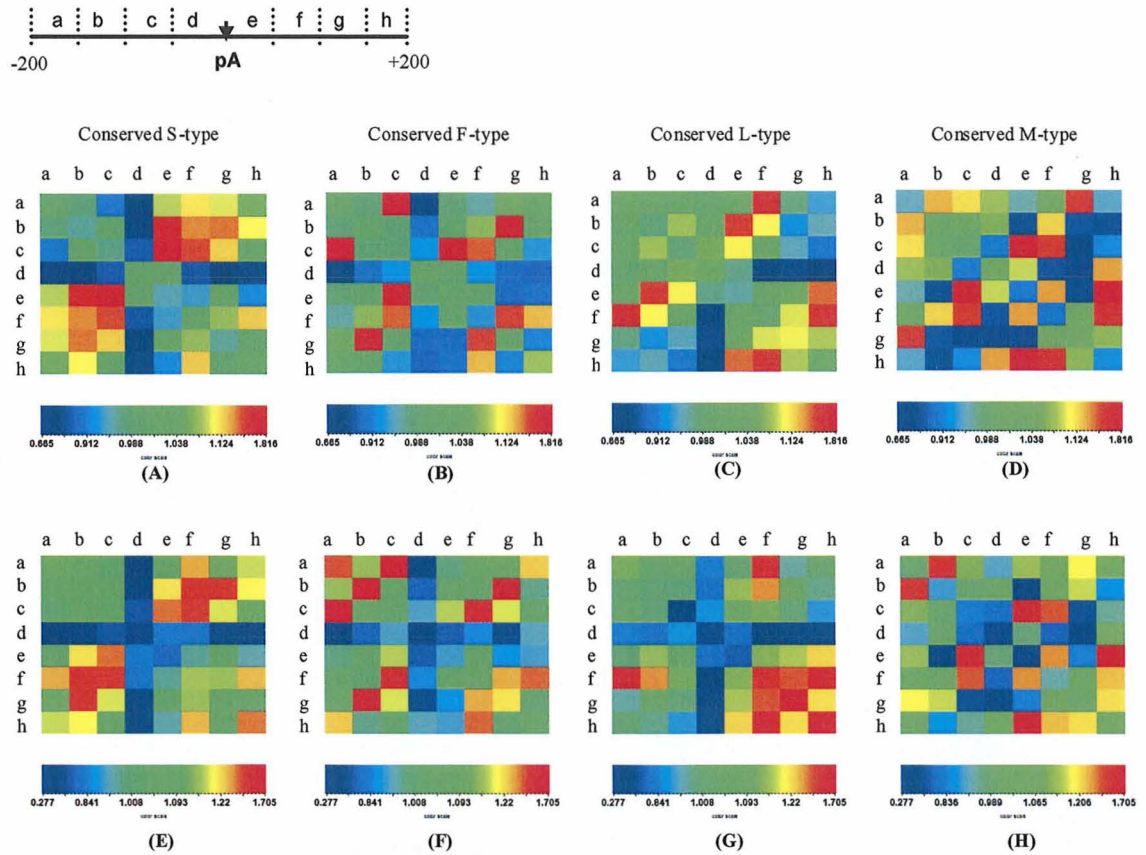

Figure 6.5 The sequence 200 nt upstream and downstream of the poly(A) site is divided into regions 50 nt long and labeled from a-h. **(A-D)** Ratio of the observed vs. expected base pairing amongst the different regions upstream and downstream of the poly(A) site of conserved S, F, L and M-type respectively. **(E-H)** Ratio of the observed vs. expected free energy contributed by the different pairs of regions upstream and downstream of the poly(A) site.

6.2.3 Differences in the Co-occurrence of cis-regulatory Elements Surrounding Poly(A) Sites

A previous study had identified several cis-elements that were over-represented in frequently used poly(A) sites as compared to the weaker poly(A) sites (19). The goal here was to find a network of co-occurring interactions between these cis-elements (and possibly other unidentified cis-elements) existing uniquely in the stable poly(A) regions. This information will further lead to the discovery of cis-elements that co-occur structurally to provide some functionality.

This analysis was done on the conserved S-type poly(A) regions as they were found to be the most stable group from previous results. The region 100 nt upstream and 100 nt downstream of the poly(A) site was selected and divided into regions 50 nt long as before and labeled c, d, e and f (Figure 6.6). Between every pair of regions, the *Z-scores* for the co-occurrence frequency of every existing tetramer pair in that region (Section 6.3.3) were computed. The pairs with a Z-score >= 2.5 were selected for further analysis. These significant interactions were visualized using the program Cytoscape (87). The network between the different pairs of regions is shown in Figure 6.6. Each edge is color-coded based on the Z-score of the interaction. The co-occurring tetramers between regions d-e are largely A-T rich and most of them involve the poly(A) signal (PAS) element whereas the extreme upstream and downstream regions (cf) contain more GC-rich tetramers as was also seen previously (19). Some of these pairs are also complementary to one another suggesting that they may be base-pairing together such as AAAA-TTTT (de), CCCA-TGGG (cf), CCCA-TGGG (ce).

Several of the tetramers could be selected to extend this network to involve more regions such as c-d-e, c-d-f so on (Figure 6.6), which shows that there may exist a

complex circuitry of interactions between cis-elements which occurs during the process of polyadenylation.

For more detailed analysis, the conserved S-type set was divided into two groups: 1) containing the sites with minimum free energy of their poly(A) region in the first quartile (minimum free energy < 25^{th} percentile i.e., most stable), and 2) containing the sites with minimum free energy of their poly(A) region in the third quartile (minimum free energy > 75^{th} percentile i.e., least stable). For each of these two groups all the significant tetramer pairs (*Z-score* >= 2.5) between every pair of regions were obtained (Figure 6.7). It was observed that the co-occurrence of tetramers differs significantly for these two groups. For the structurally more stable group (I quartile), the number of tetramer pairs found significantly co-occurring is the highest between two extreme upstream (*c*) and downstream (*d*) regions whereas this is much less for the structurally least stable group (III quartile). On the other hand the first group has lower number of significant interactions between the *d* and *e* regions (immediate upstream and downstream of the poly(A) site) as compared to the second group, for which this number is very high. It also shows that there are fewer interactions amongst the upstream regions and more interactions amongst the downstream regions for the first group and this is the opposite for the second group. This difference between these two structurally extreme groups suggests that the variation in the nucleotide composition in the different regions may lead to formation or avoidance of structures which might affect poly(A) site recognition.

6.2.4 Separation of the Poly(A) Site from the Neighboring Gene is Correlated with its Structural Stability

This analysis has so far indicated that the S-type poly(A) regions are most stable. Further investigating into the reasons, it was found that the distance separating the poly(A) site from its neighboring gene on the same strand (head to tail) as well as the distance from the closest poly(A) site on the opposite strand (tail to tail) is the least for the poly(A) regions having the least energy and it is found to be higher for regions with higher folding energy (Figure 6.8(A)).

Next, the conserved S-type poly(A) sites were divided into two parts: 1) all the sites for which the closest poly(A) site on the opposite strand is also S-type, and 2) all remaining S-type poly(A) sites. For the first group, the energy of the poly(A) region and the distance from the neighboring poly(A) site is lower than that of the second (Figure 6.8(B) and Figure 6.8(C)). Further dividing each group based on their energy, we find that the distance is closest in both cases for the sequences with least energies and it increases as the energy increases (Figure 6.8(D)). This suggests that the structure might be playing a role in strengthening the poly(A) site especially in situations where it becomes crucial for the transcription termination to occur in a timely manner to avoid interference with surrounding genes.

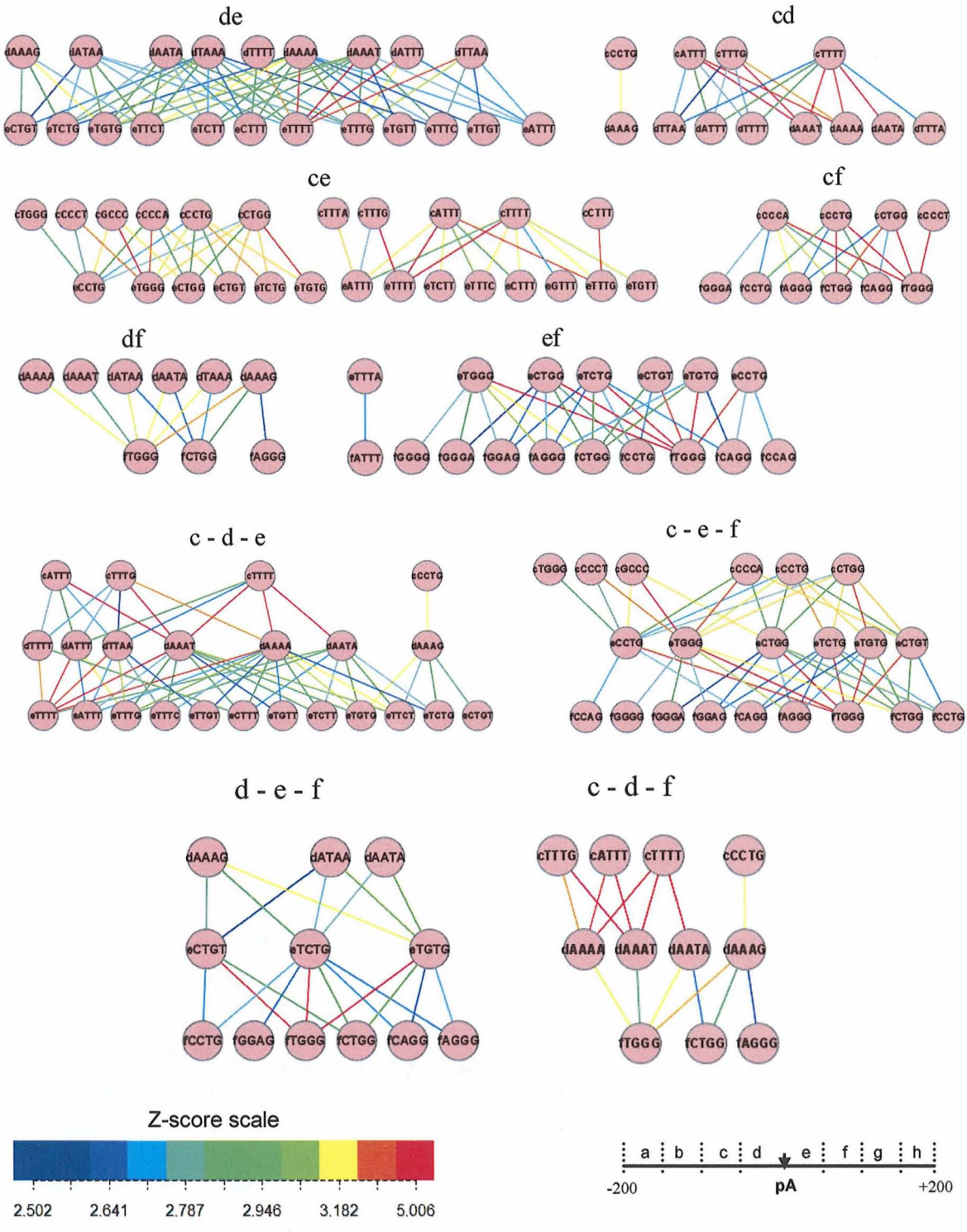

Figure 6.6 The network showing significant (Z-score >= 2.5) co-occurrences of tetramers between the different upstream and downstream regions (-100 to +100) of conserved S-type poly(A) sequences. The Z-score of each interaction is shown by using a color-coded scale.

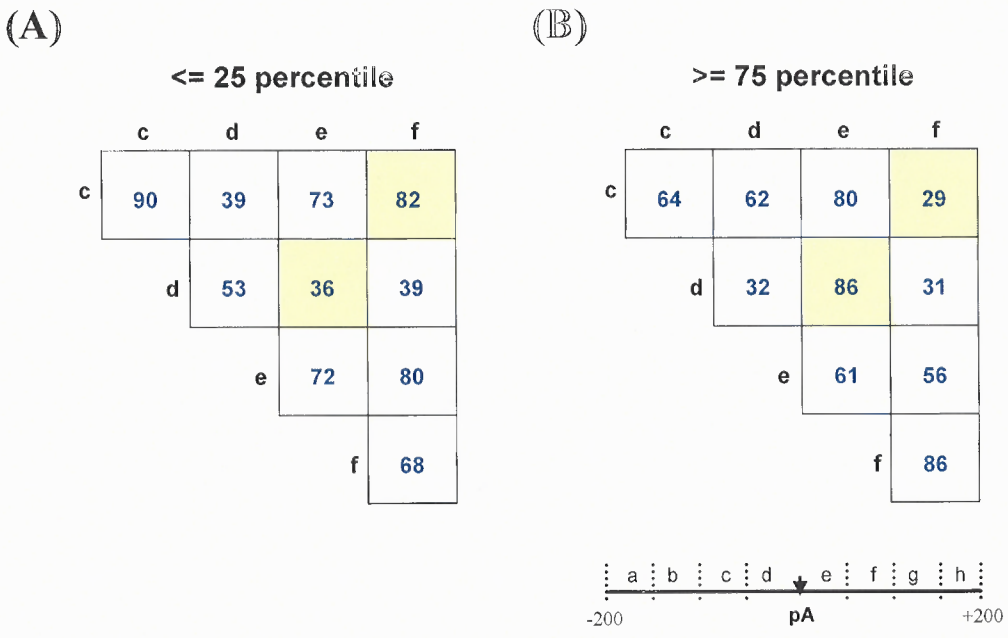

Figure 6.7 Number of significant tetramer pairs (Z-score >= 2.5) found between different upstream and downstream region pairs (-100 to +100) for the conserved S-type poly(A) sequences having **(A)** minimum free energy <= 25 percentile (most stable), and **(B)** minimum free energy >= 75 percentile (least stable), of the energy distribution.

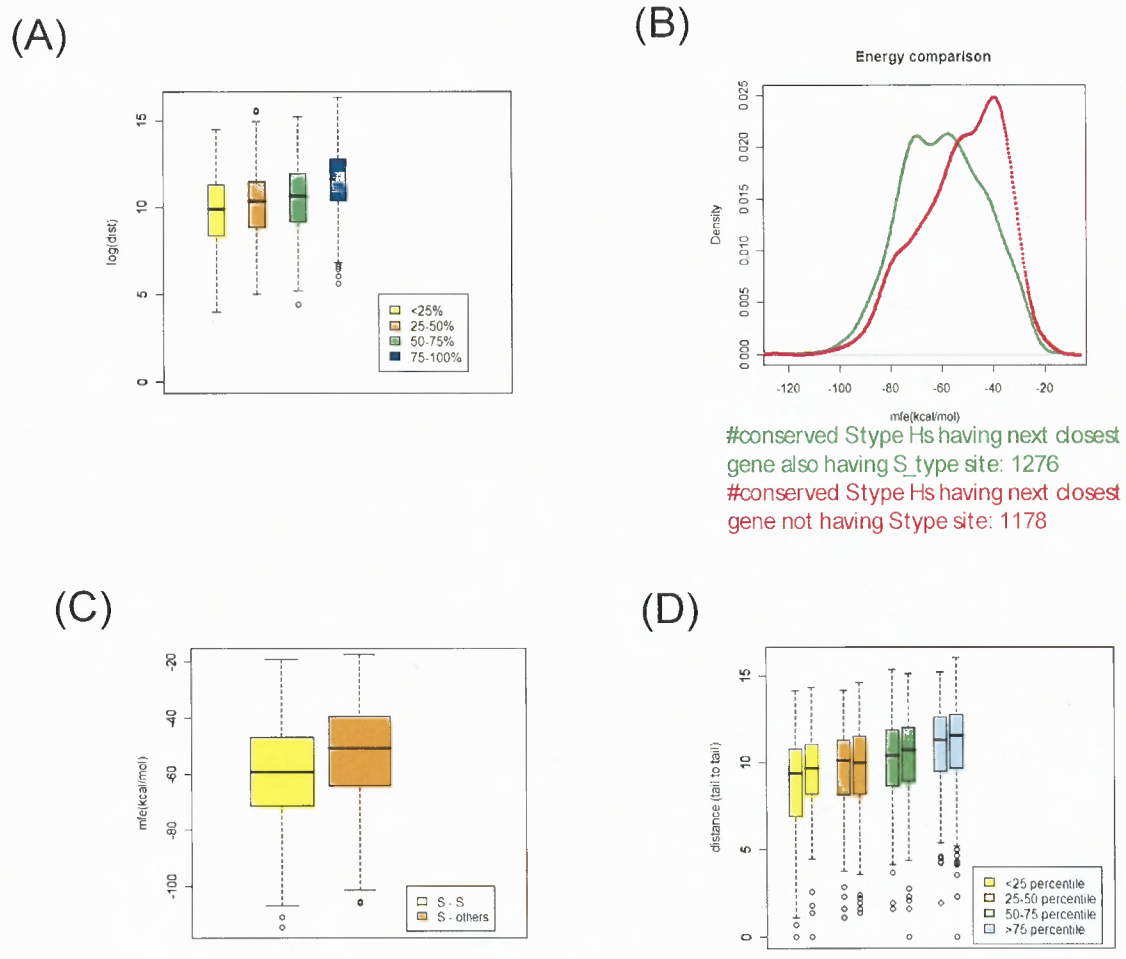

Figure 6.8 (A) Conserved S-type poly(A) sites are divided into 4 groups based on the minimum free energy of its surrounding region. For each group the distance of the poly(A) site from the transcription start site of the closest neighboring gene on the same strand i.e. head to tail or from the poly(A) site on the opposite strand i.e. tail to tail (whichever is smaller) is obtained and plotted. **(B)** Minimum free energy distribution for the conserved S-type sequences (-100 to +100) for which the other nearest poly(A) site on the opposite strand is also S-type (S-S) vs. the conserved S-type sequences for which the nearest poly(A) site is not S-type (S-others). **(C)** Box-plot of the energy for the two groups in (B). **(D)** Each of the two groups from (B) is further divided into 4 parts based on minimum free energy distribution and then the box plot of the tail to tail distance for each of these sets is shown, first for the S-S group and next for S-others group).

6.3 Materials and Methods

6.3.1 Poly(A) Site Dataset

The information about the poly(A) sites was obtained from the PolyA_DB 2 database (86). All the 54, 686 Human poly(A) sites and 30, 235 Mouse poly(A) sites were downloaded. The poly(A) sites which are of types: Single (S) and the exonic alternative polyadenylation sites : First (F), Middle (M) and Last (L) were selected. These poly(A) sites were identified as described in (79). Briefly, human, mouse and rat cDNA/EST (NCBI, August 2005 versions) sequences were aligned with their respective genomes (UCSC; hg17 for human, mm5 for mouse and rn3 for rat) by BLAT (88). Dangling poly(A) tails (> 8nt) of the aligned cDNA/ESTs were used to find the poly(A) sites. Sites located in A-rich regions, i.e., six or more consecutive As or seven or more As in 10-nt window in the -10 to +10 nt region surrounding the site were considered as internal priming candidates and were not used in this study. cDNA/ESTs without poly(A) tails were also used if their 3′ ends were located within 24 nt from a site supported by poly(A/T)-tailed cDNA/ESTs. The orientation of a cDNA/EST on the genome was inferred by its splicing sites as previously described (79).

6.3.2 Identification of Conserved Orthologous Poly(A) Sites

Orthologous poly(A) sites were identified as described in (89) by using UCSC human versus mouse (hg17 vs. mm5), mouse versus human (mm5 vs. hg17), human versus rat (hg17 vs. rn3), and rat versus human (rn3 vs. hg17) whole genome alignments (axtNet files) (90). A pair of human and mouse/rat poly(A) sites were considered orthologous when (a) the human and mouse/rat sites are located within 24 nt in the human and

mouse/rat genome alignment; and (b) they are nearest to one another in a reciprocal manner, i.e., the mouse/rat poly(A) site is the nearest one to the human poly(A) site using hg17 versus mm5 or hg17 versus rn3, and the human one is the nearest to the mouse/rat one using mm5 versus hg17 or rn3 versus hg17. Further, if these orthologous poly(A) sites are of the same type then we select the poly(A) site pair as being conserved orthologous poly(A) sites.

6.3.3 Network of Co-occurring Tetramers Using Z-score Calculation

For any given tetramer pair: k1~k2 where k1 falls in region "r1" and k2 falls is region "r2" surrounding the poly(A) site, first we find the frequency of their co-occurrence, say its $F_{k1,r1:\ k2,r2}$. We then find the frequency of co-occurrence of k1 in region r1 with all other tetramers in region r2. Next, the mean (m1) and standard deviation (sd1) of these frequencies is calculated using which the first Z-score is obtained as follows:

$$Z1 = (F_{k1,r1:\ k2,r2} - m1) / sd1 \qquad (6.1)$$

Using a similar procedure, we then obtain the frequency of co-occurrence of k2 with all the tetramers in region r1. Again, we obtain the mean (m2) and standard deviation (sd2) for this distribution. This gives the second Z-score:

$$Z2 = (F_{k1,r1:\ k2,r2} - m2) / sd2 \qquad (6.2)$$

These Z-scores are a measure of the significance of co-occurrence of the pair k1~k2 between regions r1 and r2. We select the pairs where both the z-scores are >= 2.5 to be significant in this study.

6.3.4 Statistical analysis

Wilcoxon rank sum tests and mKS tests were carried out using the statistical analysis software R (http://www.r-project.org). For mKS test, we followed the method described in (91). Briefly, given a set of values N containing n entries and another set M containing m entries, the following method was used to assess whether values in M were significantly higher or lower than those in N. N and M are joined together, and the combined set (M+N) is ordered from high value to low value. A running sum is computed across all entries starting at the highest value. A value of $v1$ was added to the running sum if the entry is from the set N, and otherwise $v2$ is added, where $v1 = \sqrt{(m/n)}$, and $v2 = -\sqrt{(n/m)}$. Thus, the overall sum comes out to be zero. The maximum and minimum values, *Omax* and *Omin* respectively, of the running sum were used as empirical statistics and can be considered as observed values. To obtain their significance, we randomly selected m entries from (M+N), and calculated the maximum and minimum values, *Emax* and *Emin* respectively, which are considered to be the expected values. The process was repeated 1000 times. The probability for rejecting the null hypothesis that M contains larger values than N was the fraction of 1000 *Emax* that were higher than *Omax*. The probability for rejecting the null hypothesis that M contains smaller values than N was the fraction of 1000 *Emin* that were smaller than *Omin*. These probabilities were called E-values in this study.

CHAPTER 7

CONCLUSIONS AND FUTURE RESEARCH

The primary objective of this research was to undertake a bioinformatics approach to the study of RNA secondary structures and discover novel biological results. This chapter reviews the findings of this study, implications of the results and the additional research in future that will further consolidate the work.

7.1 GLEAN-UTR

GLEAN-UTR approach was developed to discover novel putative conserved ncRNAs from the untranslated regions (UTRs) of orthologous Human and Mouse genes. This approach resulted in 90 distinct RNA structure groups containing 748 structures (Chapter 2). These groups were formed of RNA sequences that have a similar structure and also share Gene Ontology annotations of the Biological Process category which indicates a possibility that the structures in these groups may have some common function in the biological pathway. The approach also discovered the well known Histone 3' UTR stem loop structures and the Iron Response element structures as the top two groups in the results. This provides some validation for the approach that it does group structurally and functionally similar structures together. However, for the other groups it is hitherto unknown what function these structures carry out in the cell and if they do so at all. So, the next step will be to design wet-lab experiments that can find out whether any of these structures are functional in the cell.

The GLEAN-UTR approach is generic and can be applied to other species as well to analyze and identify conserved RNA structures. Recently large numbers of species have been sequenced and this data is publicly available. Taking advantage of this fact, in future GLEAN-UTR approach will be used to study other organisms.

This method was applied to mining small RNA structures in this study, primarily because those structures can be more accurately predicted by RNA prediction programs using only thermodynamic parameters. With the development of more sophisticated RNA prediction algorithms, the accuracy will increase and it will also be possible to identify large conserved RNA structures.

In summary, this study indicates that many more conserved stem-loop structures are present in human UTRs and they might be involved in coordinate post-transcriptional gene regulation of biological pathways, similar to HSL3 and IRE structures. This bioinformatics study lays a ground work for future wet lab validations of putative RNA stem-loop groups and represents a framework which can be used to analyze RNA structures identified by other approaches and in other species.

7.2 Method Development

Computational analysis of biological data has opened a great deal of avenues for ground-breaking discoveries. Development of various software tools, databases and efficient algorithms in conjunction with statistical analysis has wielded the path towards an exciting exploration of the complex cellular machinery.

In order to aid this research of RNA secondary structures, a powerful software framework was developed termed as RADAR (Chapter 4). This is an online web-based as well as standalone tool that provides wide range of functions such as database search, multiple structure alignment, consensus structure prediction, clustering and so on, which aid in detecting conserved RNA secondary structures. By using this predictive approach, biologists will be able to reduce the expensive wet-lab experiments by rejecting data that may not seem interesting while being able to find promising results very quickly. This tool is based on alignment of RNA structures using a dynamic programming algorithm (*O(mn)*) RSmatch (3). It also incorporates a novel algorithm that improves the structure alignment function of RSmatch termed as Constrained Structural alignment (Chapter 3) which significantly increases the specificity of the results and provides more flexibility for the user to provide special characteristics of the input data as per their requirements.

Several applications of this framework are possible and have been described in this dissertation (Chapter2, Chapter 3, Chapter 5) resulting in good findings. Since this tool depends on the accuracy of the RNA secondary structures provided as input, the performance can be greatly enhanced as newer more powerful methods for prediction are developed especially the ones based on phylogeny. Currently it also incorporates a p-value for each alignment as a statistical indicator for the reliability of the results (Chapter 3). This p-value depends on the score of alignment, which is computed by RADAR using very basic scoring matrices. Development of more complex and biologically obtained matrices will lead to a better outcome. Finally, through the application of these methods to various different RNA sequences coupled with biological experiments will lead to a stronger validation.

7.3 Polyadenylation Analysis

In Chapter 6 of this dissertation, focus was on the post-transciptional gene regulation process: Polyadenylation, which is a crucial step towards the maturation of almost all cellular mRNAs in eukaryotes. The process involves cleavage at 3' end of mRNAs and addition of poly(A) tail. This study set out to inquire into the questions pertaining to the strength and usage of poly(A) sites by focusing on Human poly(A) sites. Several genes have multiple poly(A) sites leading to alternate gene products (Alternative polyadenylation) and others have only a single poly(A) site. Presumably genes that have a single (S) poly(A) site would depend more on it to be efficiently detected and cleaved to have proper functioning. This gives rise to an interesting question as to whether these poly(A) sites have some evolutionary advantage that gives them a higher strength. Research undertaken attempts to answer this by investigating into the structural differences between the S-type poly(A) region as compared to that of multiple poly(A) sites present in 3' UTR : First (F), Middle (M) and Last (L). It was found that the S-type region is significantly more stable than the others and further the S-type site which are also conserved in Mouse have the highest stability as compared to those that are not conserved. It is known from previous studies that RNA structure is a critical determinant of poly(A) site definition.

Another factor that might influence the selection of poly(A) site is the distance that separates it from the neighboring gene on the same strand and distance from the closest poly(A) site of opposite strand. A correlation was seen between this distance and the structural stability: shorter the distance, lower is the minimum free energy of the poly(A) region hence higher stability. It can be hypothesized that a short distance would

mean that the poly(A) site would be stronger to prevent transcription interference. This also then means that there is a correlation between structural stability and poly(A) site strength. Future work involves designing wet-lab experiments that would prove this theory.

Ths study also found a network of co-occurring interactions between tetramers in different regions surrounding the poly(A) sites and it was observed that these interactions differ based on the structural stability of the sequence. Further research need to be done that would unequivocally model these interactions for different types of poly(A) sites and tie it to the strength of polyadenylation.

APPENDIX A

HSL3 AND IRE MOTIFS

Figure A.1 shows the secondary structure of the Histone 3' UTR stem loop (HSL3) and that of the Iron Response Element (IRE).

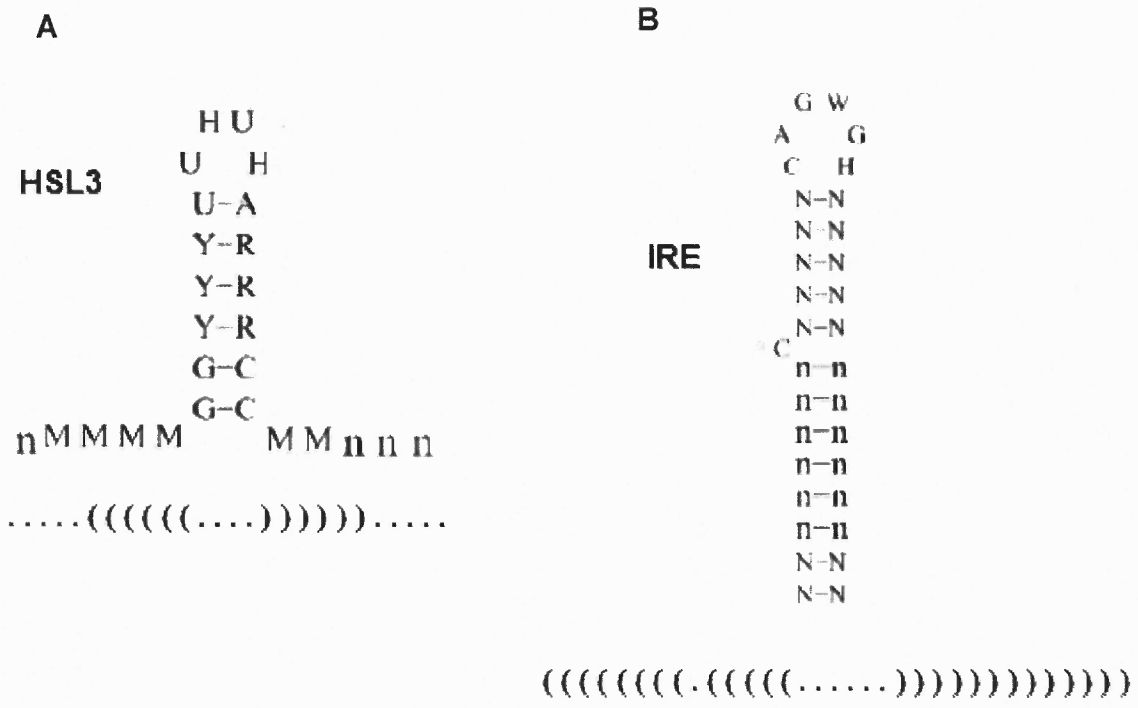

Figure A.1 The graphical representation of **(A)** HSL3 motif and **(B)** IRE motif. The structures are also represented in the dot-bracket format.

APPENDIX B

HIERARCHICAL CLUSTERING RESULTS

The GLEAN-UTR approach found 2,054 structures that were similar to alteast two other structures and satisfied the alignment score cutoff. In order to group the similar structures hierarchical clustering was applied. Figure B.1 shows the heatmap for the outcome of clustering. The results show that several structures are similar to one another and they have been clustered together.

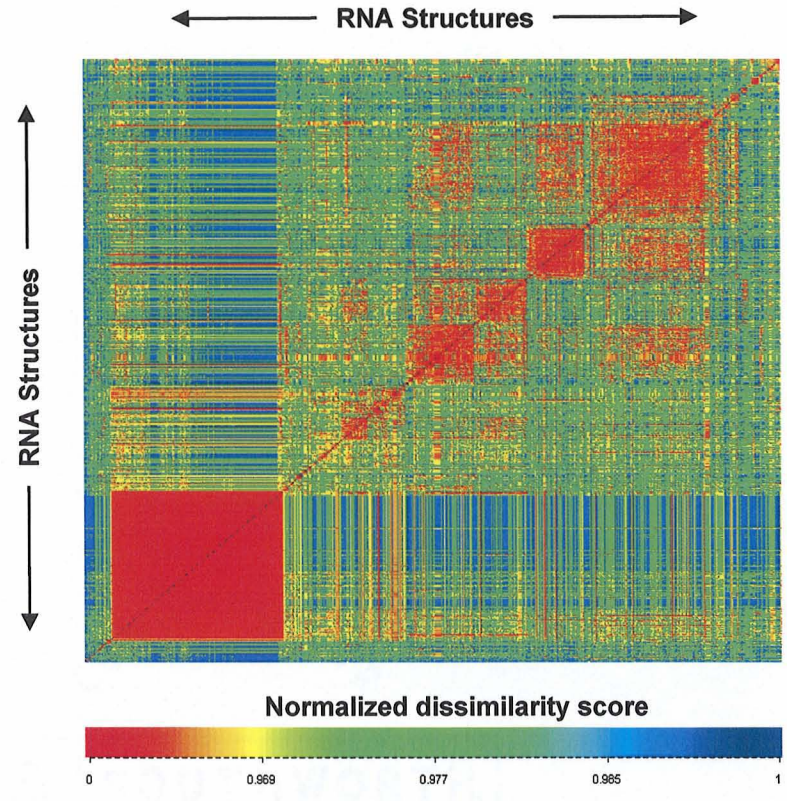

Figure B.1 Heat map for all-against-all comparisons of 2,054 human RNA structures. The normalized dissimilarity score is represented by color based on the scale shown at the bottom. The structures are in the same order as those shown in the hierarchical clustering tree in Figure 5.3(B).

APPENDIX C

GLEAN-UTR FOR RANDOMIZED UTR SEQUENCES

Figure C.1 shows the result GLEAN-UTR approach to randomized (using 1-order Markov model) human and mouse UTR sequences.

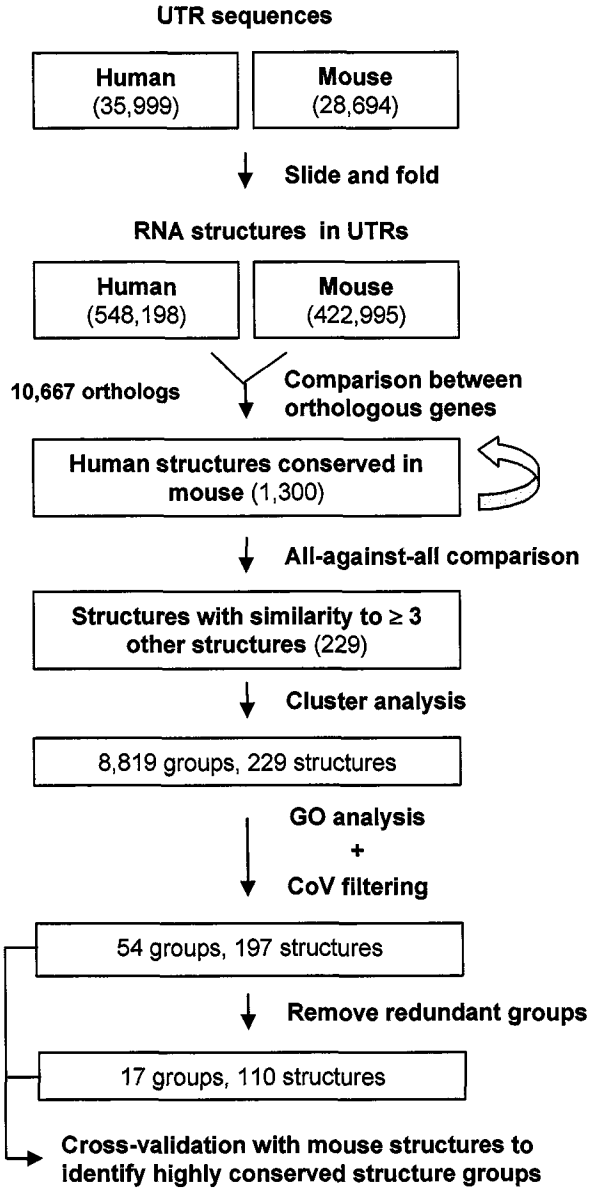

Figure C.1 UTR sequences randomized by 1-order Markov chain were subject to the same GLEAN-UTR approach as shown in Figure 5.1. The number of structures and structure groups are shown at each step.

APPENDIX D

GLEAN-UTR RESULTS OVERLAPPING WITH STRUCTURES FROM OTHER STUDIES

The human conserved structures obtained from the GLEAN-UTR approach were compared with structures obtained by other similar studies (Washietl et al., 2005, Pedersen et al., 2006 and Torarinsson et al., 2006). 131 structures were found to overlap with Washietl et al., 2005 and Pedersen et al., 2006 whereas no overlap was found with Torarinsson et al., 2006. Following table shows the overlapping structures.

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
115	NM_005922:5007-5032	Homo sapiens mitogen-activated protein kinase kinase kinase 4 (MAP3K4), transcript variant 1, mRNA	TATGTAATATTTACATA (((((((...)))))))	P
156	NM_006265:3592-3636	Homo sapiens RAD21 homolog (S. pombe) (RAD21), mRNA	TATAACTTTTCTAATAAAAGTTGTG ((((((((((.....))))))))))	P
193	NM_007203:3560-3639	Homo sapiens A kinase (PRKA) anchor protein 2 (AKAP2), transcript variant 1, mRNA	ATTTTAATGGACTATTTATTAAAGT ((((((((((...))))))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
179	NM_139168:3644-3678	Homo sapiens splicing factor, arginine/serine-rich 12 (SFRS12), mRNA	GCTTTACTATGTAAAGT (((((((...)))))))	P
224	NM_018959:1740-1782	Homo sapiens DAZ associated protein 1 (DAZAP1), transcript variant 2, mRNA	TATGTTAAAGAAAAAATATA ((((((........))))))	P
180	NM_005249:2317-2397	Homo sapiens forkhead box G1B (FOXG1B), mRNA	TGTATATTTTGATGTATG (((((((....)))))))	P, W
128	NM_173469:2838-2914	Homo sapiens hypothetical protein LOC92912 (LOC92912), mRNA	TAAACTTGCATCAAGTTTA (((((((...)))))))	P
215	NM_004093:4035-4084	Homo sapiens ephrin-B2 (EFNB2), mRNA	ATTGCTGCATATTTGTGCCGTAAT (((((.((((....)))).)))))	P
124	NM_020245:10602-10629	Homo sapiens tubby like protein 4 (TULP4), mRNA	TTTGCATTTGTTTATAAATGCATTATTT ..(((((((....))))))).....	P, W
186	NM_004396:2183-2242	Homo sapiens DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 (DDX5), mRNA	CCTGAAACAATTTTTAGGT (((((((....)))))))).	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
11	NM_001546:1287-1309	Homo sapiens inhibitor of DNA binding 4, dominant negative helix-loop-helix protein (ID4), mRNA	CATCTATTGTTTAAAATAGATG (((((((.....)))))))	P, W
13	NM_005627:1871-1929	Homo sapiens serum/glucocorticoid regulated kinase (SGK), mRNA	TCTTCCATATTTGGAAGA (((((((....)))))))	P
115	NM_022900:3428-3449	Homo sapiens O-acetyltransferase (CAS1), mRNA	TTTCCAATATTTGGAAA (((((((...)))))))	P
159	NM_004235:2401-2458	Homo sapiens Kruppel-like factor 4 (gut) (KLF4), mRNA	TGTGCAATAATTTGTACA (((((((....)))))))	P
67	NM_005204:2736-2767	Homo sapiens mitogen-activated protein kinase kinase kinase 8 (MAP3K8), mRNA	ATTCAAAACGTGATGTTTTGAAT (((((((((.....)))))))))	P, W
87	NM_014795:5073-5162	Homo sapiens zinc finger homeobox 1b (ZFHX1B), mRNA	AAATAACATTTTATTT ((((((....))))))	P
277	NM_016131:1409-1508	Homo sapiens RAB10, member RAS oncogene family (RAB10), mRNA	TAAAGTTAGAATTAACAATTTTA (((((((.(.....)))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
117	NM_014757:5412-5436	Homo sapiens mastermind-like 1 (Drosophila) (MAML1), mRNA	TGTAAATAAATGTTTACA (((((((....)))))))	P, W
9	NM_014585:197-237	Homo sapiens solute carrier family 40 (iron-regulated transporter), member 1 (SLC40A1), mRNA	AACTTCAGCTACAGTGTTAGCTAAGTT (((((.(((((.....)))))))))))	P
193	NM_015397:2051-2087	Homo sapiens KIAA1892 (KIAA1892), mRNA	CTCAGACTTTCTGTGAAAGTTTGGG ((((((((((...))))))))))	P
128	NM_001905:2511-2531	Homo sapiens CTP synthase (CTPS), mRNA	ACTCCTTGCATCAAGGGGT (((((((...)))))))	P
226	NM_006471:1171-1202	Homo sapiens myosin regulatory light chain MRCL3 (MRCL3), mRNA	AGAAAGTTATTCGCTCGATTTTTT (((((((.......)))))))	P
248	NM_003701:1654-1706	Homo sapiens tumor necrosis factor (ligand) superfamily, member 11 (TNFSF11), transcript variant 1, mRNA	AAATGTCTTGCTGTTGACATAT (.(((((.......))))))	P
193	NM_024045:2432-2468	Homo sapiens DEAD (Asp-Glu-Ala-Asp) box polypeptide 50 (DDX50), mRNA	GTATTTTTTTAAAAAGTAT (((((((...)))))))	P
209	NM_000214:4674-4767	Homo sapiens jagged 1 (Alagille syndrome) (JAG1), mRNA	TTTGATTTATTAACTTAATAATCAA .((((.((((((...))))))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
200	NM_004463:4228-4250	Homo sapiens faciogenital dysplasia (Aarskog-Scott syndrome) (FGD1), mRNA	TTTTTTTTTTTTTAAGAAAAA (((((((.....)))))))	P
55	NM_002973:4142-4178	Homo sapiens spinocerebellar ataxia 2 (olivopontocerebellar ataxia 2, autosomal dominant, ataxin 2) (SCA2), mRNA	TGCTTCTACCAACTGGAAGCA (((((((.....)))))))	P
126	NM_182789:1643-1664	Homo sapiens poly(A) binding protein interacting protein 1 (PAIP1), transcript variant 2, mRNA	TATATAATAGTTTATTATGTA (((((((((...)))))))))	P
118	NM_005776:597-666	Homo sapiens cornichon homolog (Drosophila) (CNIH), mRNA	TTTAAAAAATGACTCCTTATTTTTTAAA ((((((((((((....))))))))))))	P
244	NM_182700:2993-3066	Homo sapiens Sp8 transcription factor (SP8), transcript variant 1, mRNA	TGTATAGTATTTTTCTTGTACA (((((((........)))))))	P
186	NM_001292:1621-1646	Homo sapiens CDC-like kinase 3 (CLK3), transcript variant phclk3/152, mRNA	TGTTATAAAGTTATAATA (((((((....)))))))	P
83	NM_172316:2894-2949	Homo sapiens Meis1, myeloid ecotropic viral integration site 1 homolog 2 (mouse) (MEIS2), transcript variant h, mRNA	TATCAGATCTGCTGTGGAATTGGTA ((((((.(((.....))).))))))	P
61	NM_014497:6409-6441	Homo sapiens NP220 nuclear protein (NP220), mRNA	GGTTTGATTTTTATATCAAATC (((((((((......)))))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
85	NM_004089:656-686	Homo sapiens delta sleep inducing peptide, immunoreactor (DSIPI), transcript variant 2, mRNA	TCTGTCCTCAGGGTGGGCAGA (((((((.....)))))))	P
209	NM_001827:387-441	Homo sapiens CDC28 protein kinase regulatory subunit 2 (CKS2), mRNA	GTATTCAGTGAATAC ((((((...))))))	P, W
109	NM_020432:3107-3145	Homo sapiens putative homeodomain transcription factor 2 (PHTF2), mRNA	GCACAGCTCCAACTGTGC ((((((.....))))))	P, W
150	NM_015024:3682-3707	Homo sapiens exportin 7 (XPO7), mRNA	TGTATAAACATGTACA ((((((....))))))	P
139	NM_018287:3611-3649	Homo sapiens Rho GTPase activating protein 12 (ARHGAP12), mRNA	TATATAATTGTGTTGTATAG (((((((...)))))))).	P
180	NM_013352:3919-3946	Homo sapiens squamous cell carcinoma antigen recognized by T cells 2 (SART2), mRNA	TTTATTTTCCTAAATAAA (((((((....)))))))	P
9	NM_003234:3481-3509	Homo sapiens transferrin receptor (p90, CD71) (TFRC), mRNA	ATTATCGGAAGCAGTGCCTTCCATAAT (((((.(((((.....)))))))))))	P
224	NM_005487:4142-4194	Homo sapiens high-mobility group protein 2-like 1 (HMG2L1), mRNA	GTATAAAGAAATAAAATTTTGTAC (((((((.......)))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
126	NM_170731:1145-1229	Homo sapiens brain-derived neurotrophic factor (BDNF), transcript variant 3, mRNA	TGTATAAATGAAGTTTATACA (((((((((...)))))))))	P
17	NM_004441:3717-3813	Homo sapiens EphB1 (EPHB1), mRNA	TCTTCATATTGAAGA (((((.....)))))	P, W
13	NM_001635:2828-2849	Homo sapiens amphiphysin (Stiff-Man syndrome with breast cancer 128kDa autoantigen) (AMPH), transcript variant 1, mRNA	GTTTTGCCTAATGGCAAAAC (((((((((....)))))))))	P
266	NM_002207:3233-3256	Homo sapiens integrin, alpha 9 (ITGA9), mRNA	AAAAATCTTCTCCAGATTTTT (((((((((.....)))))))))	P
117	NM_005153:2858-2934	Homo sapiens ubiquitin specific protease 10 (USP10), mRNA	TAAAAAAGAAATTTTTTA (((((((....)))))))	P
157	NM_005985:1632-1657	Homo sapiens snail homolog 1 (Drosophila) (SNAI1), mRNA	TTTGTATAGTTATATGTACAGTT .(((((((......)))))))..	P
128	NM_004098:2788-2879	Homo sapiens empty spiracles homolog 2 (Drosophila) (EMX2), mRNA	TATACTTCCAAGAAGTATG (((((((((...)))))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
151	NM_022138:1582-1605	Homo sapiens SPARC related modular calcium binding 2 (SMOC2), mRNA	ACATACAATGTATGT ((((((...))))))	P
186	NM_000266:138-190	Homo sapiens Norrie disease (pseudoglioma) (NDP), mRNA	TCTCAGAAAAGTCTGAGA (((((((....)))))))	P
41	NM_015032:5209-5257	Homo sapiens androgen-induced proliferation inhibitor (APRIN), mRNA	TTTAAAGTATTTTAATTTTAAA (((((((.....)))))))	P
198	NM_004308:3310-3336	Homo sapiens Rho GTPase activating protein 1 (ARHGAP1), mRNA	TTTTTGTATTTCAATAAAAA (((((((.....)))))))	P
136	NM_003564:693-742	Homo sapiens transgelin 2 (TAGLN2), mRNA	AATATATATGTAGATATATATT (((((((((....)))))))))	P
222	NM_000304:839-869	Homo sapiens peripheral myelin protein 22 (PMP22), transcript variant 1, mRNA	TTGAAGATGTATATAATATCTCCGG (((.((((((.....)))))).)))	P
246	NM_025213:8264-8332	Homo sapiens spectrin, beta, non-erythrocytic 4 (SPTBN4), mRNA	GGAGGGGACACCCCCTCC (((((((....)))))))	P, W
73	NM_181552:4943-4969	Homo sapiens cut-like 1, CCAAT displacement protein (Drosophila) (CUTL1), transcript variant 1, mRNA	TTTTCAAGGAAGAAAA (((((.....)))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
117	NM_148976:1421-1445	Homo sapiens proteasome (prosome, macropain) subunit, alpha type, 1 (PSMA1), transcript variant 1, mRNA	GTGAAATAAAATTGTTTCAC (((((((((....)))))))))	P
232	NM_001149:4428-4451	Homo sapiens ankyrin 3, node of Ranvier (ankyrin G) (ANK3), transcript variant 2, mRNA	CAGAATACTAATATTTTG (((((((....)))))))	P
127	NM_002268:3309-3369	Homo sapiens karyopherin alpha 4 (importin alpha 3) (KPNA4), mRNA	TTGCATAAAAGTTATGCAA (((((((((...)))))))))	P
181	NM_005281:1740-1824	Homo sapiens G protein-coupled receptor 3 (GPR3), mRNA	TGGTTTTTTATTTTTTTAAGAAGCCA ((((((((((.....))))))))))	P, W
148	NM_025263:2056-2088	Homo sapiens proline-rich polypeptide 3 (PRR3), mRNA	TTGTTTTCTGTGAAAACAG (((((((((...)))))))))	P, W
85	NM_020360:1706-1732	Homo sapiens phospholipid scramblase 3 (PLSCR3), mRNA	TTTATATG--T-TACATT-CATATAAA (((((((((.........)))))))))	P
193	NM_182898:2027-2094	Homo sapiens cAMP responsive element binding protein 5 (CREB5), mRNA	CATTTTTATAGTTATTATGGAAATG ((((((((((...))))))))))	P
145	NM_173560:3263-3295	Homo sapiens regulatory factor X domain containing 1 (RFXDC1), mRNA	TTATTTGACAAAAAGTCAAATGTG .(((((((.....))))))))..	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
9	NM_003234:3884-3912	Homo sapiens transferrin receptor (p90, CD71) (TFRC), mRNA	ATTATCGGGAGCAGTGTCTTCCATAAT (((((.(((((......)))))))))))	P
214	NM_004593:1186-1279	Homo sapiens splicing factor, arginine/serine-rich 10 (transformer 2 homolog, Drosophila) (SFRS10), mRNA	AAAAGTATGTTTTGCATGTATTTTTTT ((((((((((.....))))))))))..	P
148	NM_002293:7875-7908	Homo sapiens laminin, gamma 1 (formerly LAMB2) (LAMC1), mRNA	ATTTTATTTATAATAAAAT (((((((...)))))))	P
85	NM_198077:880-905	Homo sapiens gm117 (gm117), mRNA	ATATATTTTTAAAGTAAATATATT (((((((((.......)))))))))	P
53	NM_182697:951-1050	Homo sapiens ubiquitin-conjugating enzyme E2H (UBC8 homolog, yeast) (UBE2H), transcript variant 2, mRNA	ATATATATATATTATATAT (((((((.....)))))))	P
214	NM_018948:2571-2593	Homo sapiens mitogen-inducible gene 6 (MIG-6), mRNA	CAACACAAGCTGGCCTTGTGTTG (((((((((.....)))))))))	P
131	NM_032291:3259-3291	Homo sapiens hypothetical protein DKFZp761D221 (DKFZp761D221), mRNA	TTTTTATAACTTGTGTAAAAA (((((((.....)))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
170	NM_015461:4825-4865	Homo sapiens early hematopoietic zinc finger (EHZF), mRNA	GTTTTCCAAGAGGAAAT ((((((.....))))))	P
162	NM_004730:1700-1796	Homo sapiens eukaryotic translation termination factor 1 (ETF1), mRNA	TGAAAAAAAATGATTTTTTTAA (((((((((.....)))))))).	P
85	NM_019028:2342-2383	Homo sapiens HIP14-related protein (HIP14L), mRNA	TAAATATGTAAAAAATATTTA (((((((.......)))))))	P
99	NM_002167:1143-1169	Homo sapiens inhibitor of DNA binding 3, dominant negative helix-loop-helix protein (ID3), mRNA	ACAGGAAGGTGACTTTCTGT (((((((((....)))))))))	P
85	NM_182763:1617-1650	Homo sapiens myeloid cell leukemia sequence 1 (BCL2-related) (MCL1), transcript variant 2, mRNA	TGTAAAAAT-TGTATA-TATTTTTACA (((((((((.........)))))))))	P
21	NM_003081:1331-1430	Homo sapiens synaptosomal-associated protein, 25kDa (SNAP25), transcript variant 1, mRNA	TTATGCATTTATGCATGA (((((((....)))))))	P
121	NM_003927:1600-1648	Homo sapiens methyl-CpG binding domain protein 2 (MBD2), transcript variant 1, mRNA	AGATGTATTTTTGATGTATATATCT (((((((((.........)))))))))	P
200	NM_022763:6852-6894	Homo sapiens FAD104 (FAD104), mRNA	ATATTTATGCCCAATAAATGT (((((((((.....)))))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
150	NM_015578:2037-2116	Homo sapiens chromosome 19 open reading frame 13 (C19orf13), mRNA	TTTTATATAGTTGTAAAA (((((((....)))))))	P, W
190	NM_145175:2274-2301	Homo sapiens NSE1 (NSE1), mRNA	AAAATTTCAAATTGAAATTTT (((((((.....)))))))	P, W
227	NM_017893:4247-4275	Homo sapiens sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4G (SEMA4G), mRNA	ACAATGAATGTATTTATTGT (((((((....)))))))	P, W
210	NM_005487:3043-3125	Homo sapiens high-mobility group protein 2-like 1 (HMG2L1), mRNA	AAATCTCTTAGATTT (((((.....)))))	P, W
211	NM_014901:3848-3886	Homo sapiens ring finger protein 44 (RNF44), mRNA	ATGTATGTATTTGAGAAAATGCTAATATAT ((((((((((((....))))))).)))))	P
224	NM_020177:2578-2628	Homo sapiens fem-1 homolog c (C.elegans) (FEM1C), mRNA	AATATACCATATAATATATT ((((((........))))))	P
203	NM_002657:5587-5626	Homo sapiens pleiomorphic adenoma gene-like 2 (PLAGL2), mRNA	AATGAAGTTGTTTATT ((((((....))))))	P
35	NM_173822:2158-2178	Homo sapiens hypothetical protein MGC39518 (MGC39518), mRNA	TTTTGTTTAAAAACAAAA (((((((....)))))))	P
197	NM_004703:3243-3262	Homo sapiens rabaptin, RAB GTPase binding effector protein 1 (RABEP1), mRNA	TTTATATTAAAATATGAA (((((((....)))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
27	NM_152267:3108-3127	Homo sapiens hypothetical protein FLJ38628 (FLJ38628), mRNA	ATTTTCACTGTTGTGAAAGT (((((((....)))))))	P
239	NM_030797:1250-1337	Homo sapiens hypothetical protein DKFZp566A1524 (DKFZP566A1524), mRNA	ATGTTAATACTTGTGTATTTACAT ((((.(((((....)))))).))))	P
262	NM_002819:1859-1920	Homo sapiens polypyrimidine tract binding protein 1 (PTBP1), transcript variant 1, mRNA	AAAGAGAAATCAGTTTACCTGTTTTT (((((.....(((.....)))))))))	P
256	NM_170677:977-1037	Homo sapiens Meis1, myeloid ecotropic viral integration site 1 homolog 2 (mouse) (MEIS2), transcript variant a, mRNA	TATCCGGACTGGGATA (((((.....)))))	P
200	NM_001677:1898-1994	Homo sapiens ATPase, Na+/K+ transporting, beta 1 polypeptide (ATP1B1), mRNA	TTTTTTCTGCAAGAAAAAG (((((((.....)))))))	P
224	NM_014153:3602-3631	Homo sapiens zinc-finger protein AY163807 (HSPC055), mRNA	TTTAACACTAGTATTTGTTAAA (((((((........)))))))	P
166	NM_005595:2467-2565	Homo sapiens nuclear factor I/A (NFIA), mRNA	TATCTTTGTAAGATA ((((((...))))))	P
246	NM_139135:7428-7485	Homo sapiens AT rich interactive domain 1A (SWI- like) (ARID1A), transcript variant 2, mRNA	GCAGCGGCTACGCTGC ((((((....))))))	P
232	NM_015952:914-976	Homo sapiens RWD domain containing 1 (RWDD1), mRNA	TCAGGAGAATATTCTTCTGA (((((((....)))))))	P

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
3	NM_005321:721-785	Homo sapiens histone 1, H1e (HIST1H1E), mRNA	AACCCAAAGGCTCTTTTCAGAGCCACCCA .....((((((....)))))).....	P
61	NM_020993:3639-3660	Homo sapiens B-cell CLL/lymphoma 7A (BCL7A), mRNA	AGATGAATTTGGATATTTATTT (((((((.....)))))))	W
178	NM_006480:2193-2215	Homo sapiens regulator of G-protein signalling 14 (RGS14), mRNA	GAGGAGGGGCCGGCCCCTCCTC (((((((((....)))))))))	W
278	NM_007040:2857-2881	Homo sapiens E1B-55kDa-associated protein 5 (E1B-AP5), transcript variant 1, mRNA	GGGCTGCCTCCCTCCAGCCC ((((((.......))))))	W
17	NM_004443:3616-3640	Homo sapiens EphB3 (EPHB3), mRNA	TCTTCATATTGAAGA (((((.....)))))	W
221	NM_007373:2827-2873	Homo sapiens soc-2 suppressor of clear homolog (C. elegans) (SHOC2), mRNA	T-ATATATGTATATACAATGCTATATA (.(((((.((((.....))))))))))	W
214	NM_021190:2873-2928	Homo sapiens polypyrimidine tract binding protein 2 (PTBP2), mRNA	TTTTGAAATTGATGTACTTAGTTTCAAGATT ((((((((((((.....)))))))))))))..	W

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (Continued)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
193	NM_016353:2379-2408	Homo sapiens zinc finger, DHHC domain containing 2 (ZDHHC2), mRNA	TTTAGTTTGAGATAAACTAAA (((((((((...)))))))))	W
136	NM_005610:2031-2071	Homo sapiens retinoblastoma binding protein 4 (RBBP4), mRNA	GTAAAGATGTATGTTTTTAC (((((((....)))))))	W
160	NM_001219:3240-3267	Homo sapiens calumenin (CALU), mRNA	TAGAGTGTAAACCAAGTTTTATATTCTG ((((((((((......))))))))))	W
61	NM_006558:1749-1799	Homo sapiens KH domain containing, RNA binding, signal transduction associated 3 (KHDRBS3), mRNA	ATAGAATTTAGTTATTTTAT (((((((......)))))))	W
186	NM_014583:1428-1451	Homo sapiens LIM and cysteine-rich domains 1 (LMCD1), mRNA	TTCTAAGAAGTCTTAGGA (((((((....)))))))	W
240	NM_002569:3320-3346	Homo sapiens furin (paired basic amino acid cleaving enzyme) (FURIN), mRNA	AGCCCGGGCTGCCTGGGCT (((((((...)))))))	W
180	NM_005249:2317-2397	Homo sapiens forkhead box G1B (FOXG1B), mRNA	TGTATATTTTGATGTATG (((((((....)))))))	W
11	NM_001546:1319-1388	Homo sapiens inhibitor of DNA binding 4, dominant negative helix-loop-helix protein (ID4), mRNA	CATCTATTGTTTAAAATAGATG (((((((......)))))))	W

Table D.1 Structures Identified by the GLEAN-UTR Approach as well as by Washietl et al., 2005 or Pedersen et al., 2006 (*Continued*)

Group ID[1]	RefSeq ID	Annotation	Structure	Overlap with[2]
240	NM_005479:2438-2521	Homo sapiens frequently rearranged in advanced T-cell lymphomas (FRAT1), transcript variant 1, mRNA	ACACTTCGCACCGGAGTGT (((((((...)))))))	W
232	NM_002265:3275-3300	Homo sapiens karyopherin (importin) beta 1 (KPNB1), mRNA	AGGCTAGAAGTAGCTT ((((((....))))))	W
65	NM_004429:2279-2308	Homo sapiens ephrin-B1 (EFNB1), mRNA	GTCGCGCCTCGTGCGGCA (((((((...))))))).	W
182	NM_004343:1845-1899	Homo sapiens calreticulin (CALR), mRNA	CAAAATTTCTATTAAATTAAATTTTG (((((((((.........)))))))))	W
214	NM_021190:2247-2290	Homo sapiens polypyrimidine tract binding protein 2 (PTBP2), mRNA	TTTTGAAATTGATGTACTTAGTTTCAAGATT ((((((((((((.....)))))))))))))..	W
245	NM_130470:249-271	Homo sapiens MAP-kinase activating death domain (MADD), transcript variant 1, mRNA	CAGAATTCCTCCTGGGAATGCTG (((.(((((.....))))).)))	W
225	NM_013381:3780-3801	Homo sapiens thyrotropin-releasing hormone degrading ectoenzyme (TRHDE), mRNA	AACTCATTTTCTTTGAGTT ((((((.....))))))	W
89	NM_006599:7764-7837	Homo sapiens nuclear factor of activated T-cells 5, tonicity-responsive (NFAT5), transcript variant 3, mRNA	GGAAATGGTATACTATTTT .(((((((....)))))))	W
97	NM_032208:4055-4083	Homo sapiens anthrax toxin receptor 1 (ANTXR1), transcript variant 1, mRNA	TTGACTGCTGGCAGTCTAA (((((((....)))))).))	W

[1] Group ID is a serial number, which can be used to query the GLEAN-UTR database.

[2] "W" refers to the study by Washietl et al., 2005 and "P" to the one by Pedersen et al., 2006.

APPENDIX E

EXTENDING RNA STRUCTURE GROUPS FOUND BY GLEAN-UTR

The 90 structure groups identified by GLEAN-UTR were used to further select additional similar structures from the human UTRs using PatSearch and then GO analysis. Figure E.1 shows the increase in group size using this approach.

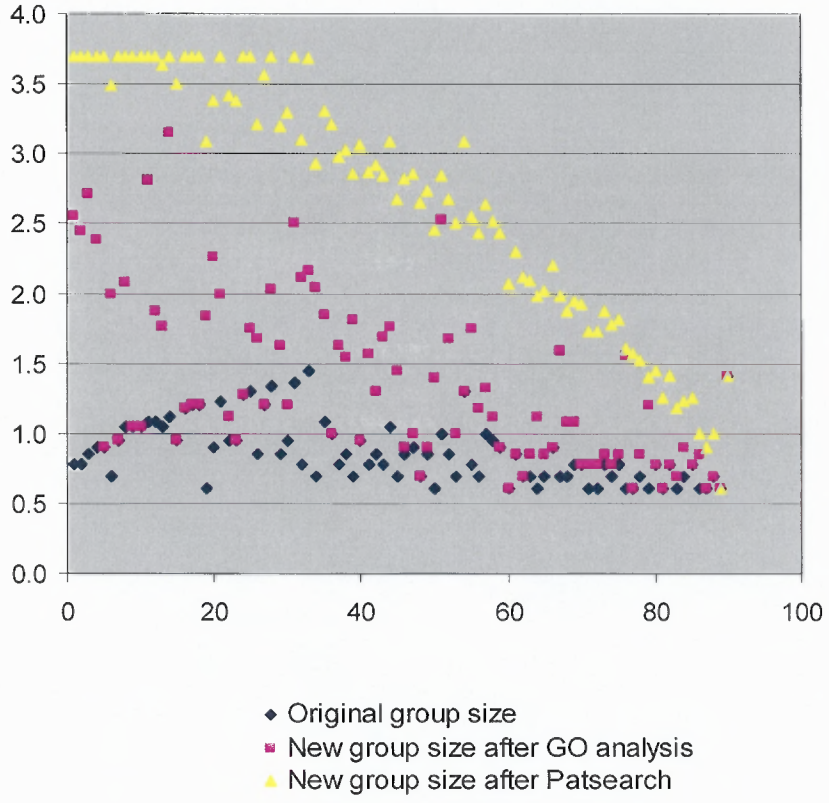

Figure E.1 The 90 structure groups found by GLEAN-UTR approach were used to search human UTRs to obtain additional group members using PatSearch. GO analysis refers to filtering out hits without the same GO term annotation as the original group. The structure groups are ordered according to the difference between the original group size and the group size after PatSearch.

REFERENCES

1. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, **33**, D121-124.
2. Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277-279.
3. Liu, J., Wang, J.T., Hu, J. and Tian, B. (2005) A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics*, **6**, 89.
4. Khaladkar, M., Liu, J., Wen, D., Wang, J.T. and Tian, B. (2008) Mining small RNA structure elements in untranslated regions of human and mouse mRNAs using structure-based alignment. *BMC Genomics*, **9**, 189.
5. Khaladkar, M., Bellofatto, V., Wang, J.T., Tian, B. and Shapiro, B.A. (2007) RADAR: a web server for RNA data analysis and research. *Nucleic Acids Res*, **35**, W300-304.
6. Alberts, B., Wilson, J.H. and Hunt, T. (2008) *Molecular biology of the cell*. 5th ed. Garland Science, New York.
7. Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002) Untranslated regions of mRNAs. *Genome Biol*, **3**, REVIEWS0004.
8. Wilkie, G.S., Dickson, K.S. and Gray, N.K. (2003) Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. *Trends Biochem Sci*, **28**, 182-188.
9. Kuersten, S. and Goodwin, E.B. (2003) The power of the 3' UTR: translational control and development. *Nat Rev Genet*, **4**, 626-637.
10. Keene, J.D. and Tenenbaum, S.A. (2002) Eukaryotic mRNPs may represent posttranscriptional operons. *Mol Cell*, **9**, 1161-1167.
11. Bakheet, T., Frevel, M., Williams, B.R., Greer, W. and Khabar, K.S. (2001) ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Res*, **29**, 246-254.
12. Wilusz, C.J. and Wilusz, J. (2004) Bringing the role of mRNA decay in the control of gene expression into focus. *Trends Genet*, **20**, 491-497.
13. Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281-297.
14. Filipowicz, W., Bhattacharyya, S.N. and Sonenberg, N. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*, **9**, 102-114.
15. Baird, S.D., Turcotte, M., Korneluk, R.G. and Holcik, M. (2006) Searching for IRES. *RNA*, **12**, 1755-1785.
16. Rouault, T.A. (2006) The role of iron regulatory proteins in mammalian iron homeostasis and disease. *Nat Chem Biol*, **2**, 406-414.
17. Grundner-Culemann, E., Martin, G.W., 3rd, Harney, J.W. and Berry, M.J. (1999) Two distinct SECIS structures capable of directing selenocysteine incorporation in eukaryotes. *RNA*, **5**, 625-635.

18. Marzluff, W.F. (2005) Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr Opin Cell Biol*, **17**, 274-280.
19. Hu, J., Lutz, C.S., Wilusz, J. and Tian, B. (2005) Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA*, **11**, 1485-1493.
20. Rajewsky, N. (2006) microRNA target predictions in animals. *Nat Genet*, **38 Suppl**, S8-13.
21. Matlin, A.J., Clark, F. and Smith, C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, **6**, 386-398.
22. Ladd, A.N. and Cooper, T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol*, **3**, reviews0008.
23. John, B., Sander, C. and Marks, D.S. (2006) Prediction of human microRNA targets. *Methods Mol Biol*, **342**, 101-113.
24. Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A. and Stadler, P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol*, **23**, 1383-1390.
25. Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W. and Haussler, D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, **2**, e33.
26. Torarinsson, E., Sawera, M., Havgaard, J.H., Fredholm, M. and Gorodkin, J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res*, **16**, 885-889.
27. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*, **31**, 3429-3431.
28. Babak, T., Blencowe, B.J. and Hughes, T.R. (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, **8**, 33.
29. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288**, 911-940.
30. Grillo, G., Licciulli, F., Liuni, S., Sbisa, E. and Pesole, G. (2003) PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res*, **31**, 3608-3612.
31. Sarnowska, E., Grzybowska, E.A., Sobczak, K., Konopinski, R., Wilczynska, A., Szwarc, M., Sarnowski, T.J., Krzyzosiak, W.J. and Siedlecki, J.A. (2007) Hairpin structure within the 3'UTR of DNA polymerase beta mRNA acts as a post-transcriptional regulatory element and interacts with Hax-1. *Nucleic Acids Res*, **35**, 5499-5510.
32. Brenet, F., Dussault, N., Delfino, C., Boudouresque, F., Chinot, O., Martin, P.M. and Ouafik, L.H. (2006) Identification of secondary structure in the 5'-untranslated region of the human adrenomedullin mRNA with implications for the regulation of mRNA translation. *Oncogene*, **25**, 6510-6519.

33. Venables W.N. , R.B.D. (2002) Modern Applied Statistics with S. *In Statistics and Computing Edited by: Chambers J, Eddy W, Hardle W, Sheather S, Tierney L. Springer.*
34. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.
35. Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, **29**, 137-140.
36. McKee, A.E. and Silver, P.A. (2007) Systems perspectives on mRNA processing. *Cell Res*, **17**, 581-590.
37. Sanchez-Diaz, P. and Penalva, L.O. (2006) Post-transcription meets post-genomic: the saga of RNA binding proteins in a new era. *RNA Biol*, **3**, 101-109.
38. Keene, J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet*, **8**, 533-543.
39. Gautheret, D. and Lambert, A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol*, **313**, 1003-1011.
40. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992) Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Sci*, **1**, 1677-1690.
41. Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. and Claverie, J.M. (1993) Ancient conserved regions in new gene sequences and the protein databases. *Science*, **259**, 1711-1716.
42. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
43. Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444-2448.
44. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680.
45. Eddy, S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
46. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406-3415.
47. Jiang, T., Lin, G., Ma, B. and Zhang, K. (2002) A general edit distance between RNA structures. *J Comput Biol*, **9**, 371-388.
48. Hochsmann, M., Toller, T., Giegerich, R. and Kurtz, S. (2003) Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf*, **2**, 159-168.
49. Corpet, F. and Michot, B. (1994) RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Comput Appl Biosci*, **10**, 389-399.
50. Mathews, D.H. and Turner, D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, **317**, 191-203.

51. Gorodkin, J., Stricklin, S.L. and Stormo, G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res*, **29**, 2135-2144.
52. Kim, J., Cole, J.R. and Pramanik, S. (1996) Alignment of possible secondary structures in multiple RNA sequences using simulated annealing. *Comput Appl Biosci*, **12**, 259-267.
53. Notredame, C., O'Brien, E.A. and Higgins, D.G. (1997) RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Res*, **25**, 4570-4580.
54. Laferriere, A., Gautheret, D. and Cedergren, R. (1994) An RNA pattern matching program with enhanced performance and portability. *Comput Appl Biosci*, **10**, 211-212.
55. Pesole, G., Liuni, S. and D'Souza, M. (2000) PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, **16**, 439-450.
56. Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C. and Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res*, **22**, 5112-5120.
57. Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*, **22**, 2079-2088.
58. Lowe, T.M. and Eddy, S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168-1171.
59. Klein, R.J. and Eddy, S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.
60. Holmes, I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, **6**, 73.
61. Yao, Z., Weinberg, Z. and Ruzzo, W.L. (2006) CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445-452.
62. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, **87**, 2264-2268.
63. Gumbel, E.J. (1958) *Statistics of extremes*. Columbia University Press, New York,.
64. Altschul, S.F. and Gish, W. (1996) Local alignment statistics. *Methods Enzymol*, **266**, 460-480.
65. Collins, J.F., Coulson, A.F. and Lyall, A. (1988) The significance of protein sequence similarities. *Comput Appl Biosci*, **4**, 67-71.
66. Smith, T.F., Waterman, M.S. and Burks, C. (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Res*, **13**, 645-656.
67. Altschul, S.F., Bundschuh, R., Olsen, R. and Hwa, T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res*, **29**, 351-361.
68. Olsen, R., Bundschuh, R. and Hwa, T. (1999) Rapid assessment of extremal statistics for gapped local alignment. *Proc Int Conf Intell Syst Mol Biol*, 211-222.
69. Hellen, C.U. and Sarnow, P. (2001) Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes Dev*, **15**, 1593-1612.

70. Pesole, G., Liuni, S., Grillo, G., Ippedico, M., Larizza, A., Makalowski, W. and Saccone, C. (1999) UTRdb: a specialized database of 5' and 3' untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res*, **27**, 188-191.
71. Rijk, P.D., Wuyts, J. and Wachter, R.D. (2003) RnaViz2: an improved representation of RNA secondary structure. *Bioinformatics*, **19**, 299-300.
72. Theil, E.C. (1993) The IRE (iron regulatory element) family: structures which regulate mRNA translation or stability. *Biofactors*, **4**, 87-93.
73. Hofacker, I.L., Stadler, P.F. and Stocsits, R.R. (2004) Conserved RNA secondary structures in viral genomes: a survey. *Bioinformatics*, **20**, 1495-1499.
74. Edmonds, M. (2002) A history of poly A sequences: from formation to factors to function. *Prog Nucleic Acid Res Mol Biol*, **71**, 285-389.
75. Colgan, D.F. and Manley, J.L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev*, **11**, 2755-2766.
76. Mangus, D.A., Evans, M.C. and Jacobson, A. (2003) Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol*, **4**, 223.
77. Wickens, M., Anderson, P. and Jackson, R.J. (1997) Life and death in the cytoplasm: messages from the 3' end. *Curr Opin Genet Dev*, **7**, 220-232.
78. Loke, J.C., Stahlberg, E.A., Strenski, D.G., Haas, B.J., Wood, P.C. and Li, Q.Q. (2005) Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol*, **138**, 1457-1468.
79. Tian, B., Hu, J., Zhang, H. and Lutz, C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res*, **33**, 201-212.
80. Gilmartin, G.M. (2005) Eukaryotic mRNA 3' processing: a common means to different ends. *Genes Dev*, **19**, 2517-2521.
81. Zhao, J., Hyman, L. and Moore, C. (1999) Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev*, **63**, 405-445.
82. Graveley, B.R., Fleming, E.S. and Gilmartin, G.M. (1996) RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor. *Mol Cell Biol*, **16**, 4942-4951.
83. Das, A.T., Klaver, B. and Berkhout, B. (1999) A hairpin structure in the R region of the human immunodeficiency virus type 1 RNA genome is instrumental in polyadenylation site selection. *J Virol*, **73**, 81-91.
84. Yan, J. and Marr, T.G. (2005) Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res*, **15**, 369-375.
85. Lee, J.Y., Ji, Z. and Tian, B. (2008) Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Res*, **36**, 5581-5590.
86. Lee, J.Y., Yeh, I., Park, J.Y. and Tian, B. (2007) PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res*, **35**, D165-168.
87. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for

integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498-2504.

88. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664.
89. Tian, B., Pan, Z. and Lee, J.Y. (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res*, **17**, 156-165.
90. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res*, **13**, 103-107.
91. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267-273.