Annual Progress Report

Research Activity

by

Shu Yang

Thesis Supervisor
Dr. Irmtraud M. Meyer

Committee Members
Dr. Alexandre Bouchard-Côté
Dr. Anne Codon
Dr. Gregg Morin

Chair
Dr. William S. Evans

A THESIS PROPOSAL SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Department of Computer Science

(The Faculty of Graduate Studies)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver, British Columbia, Canada)

June 2013

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In order to understand how living cells in organisms are organized, Bioinformatics research has so far primarily focused on studying how genes in the genome are activated by proteins, i.e. protein-DNA interactions, and how proteins interact, i.e. protein-protein interactions. Something that has only become fully apparent due to the advances in experimental technologies in the last few years is that protein-RNA interactions are indeed as numerous as protein-DNA interactions and play key role in post-transcriptional regulation of gene expression. Recently, several large scale experiments have identified a number of protein-mRNA interactions [1, 2, 3, 4, 5], which makes it possible to systematically study the features of protein binding sites in mRNA sequences.

In eukaryotic cells, all of post-transcriptional regulation of mRNA stability, translation, localization and splicing, involve the targeting of transcripts by various RBPs that recognize the functional elements in the transcript sequence. However, the sequence preference of such motif does not provide sufficient specificity (unlike TF-DNA). mRNA-protein interaction differs from DNA-protein interaction mostly due to the single stranded structure of mRNA. Protein-DNA interactions are dominated by primary sequence in canonical double stranded DNA, while protein-RNA interactions usually involve secondary structure features of the single stranded mRNA, which may affect protein binding events.

Therefore, mRNA secondary structures, in most cases provide additional accessibility information. Here the term accessibility is used to refer to the binding sites on mRNA: whether they are structurally accessible to RBPs or not. It is a remarkable feature for a number of computational methods, and is also the focus of my current project. I will describe and discuss it in the following chapters.

The rest of the report is orgnized as below: I will first go through a brief review of previous studies in this field, and introduce our proposed project. Next, I will present the main part of the project, the protein binding *motif-wise accessibility*, i.e. the accessibility of a specific sequence fraction. Then I will also show a pilot study on the *base-wise accessibility*, i.e. the accessibility

on single base level. Since a lot of my previous work has been described in the RPE report, here I will only present the updated work that has been done after RPE, including the updated literature review, data collection and project progress.

1.1 Brief Review

In this section, a second round of literature review for RNA-protein interaction is presented (the first round is done in RPE). Briefly, the focuses for this time are mRNA binding sites in vivo and most recent experimental techniques.

An overview of the RNA-protein interaction research including the experimental methods and computational methods is shown in Figure 1.1.

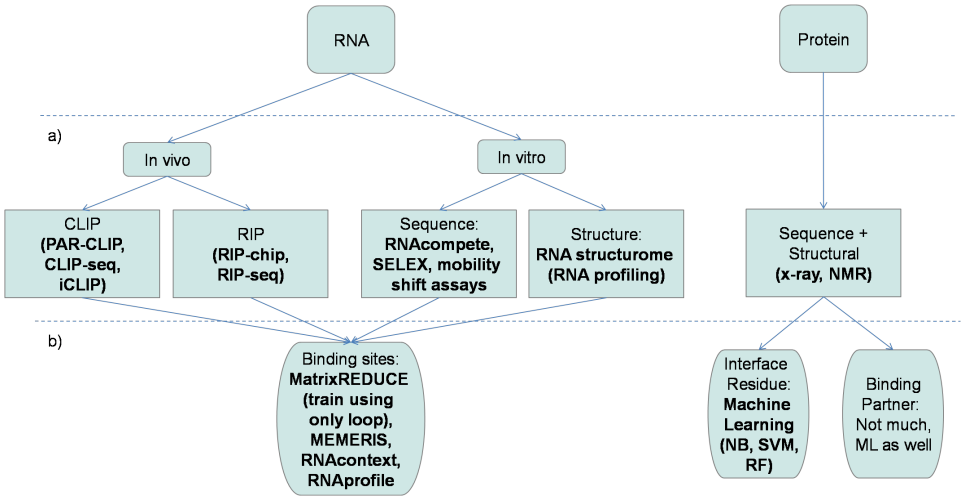

Figure 1.1: Overview of RNA-protein interaction research. a) experimental methods for RNA (both in vivo and in vitro) and protein binding sites detection. b) computational methods (in vitro RNA binding site is not of interest)

Figure 1.1 is basically drawn with guidance of Peter Stadler's recent review paper [6]. The paper discussed both RNA-RNA and RNA-protein interactions. And both experimental and computational methods are reviewed. It also provides pointers to promising databases. Besides, I am very happy that, the "RNA-protein" and "database" sections of this paper are

very similar to the literature review part of my RPE report, which shows my work is on the right track.

1.1.1 Computational Methods

For RNA-protein interactions, this review paper summarized that previous studies have shown that sequence-specific RBPs either recognize and bind to unstructured single-stranded RNA [7] or require at least some of their mRNA binding sites to be unpaired [8]. Based on this point, several representative computational methods, RNAcontext, MEMERIS, Vienna+P and a work from Morris group are introduced. Besides, HuR protein is mentioned as a "model" to assess the binding sites accessibility since HuR binds to unpaired U-rich motifs in 3'UTR.

Peter Stadler's recent review paper covers a wide range of this field and provides a very good guideline for my research. Besides of the works mentioned in his review, there are several other studies which are very relevant to my project.

Particularly, in a study [8] from Morris group at UToronto, the accessibility of a target site or region on mRNA is calculated as the probability that the site or region is unpaired. RNAplfold [9], as minimum free energy (MFE) folding algorithm from Vienna package [10] is used to estimate such probability. MFE folding algorithm predicts the secondary structure that minimizes the overall free energy of a RNA in thermodynamic equilibrium. In another word, such algorithm predicts the most stable structure. By taking accessibility into consideration, the accuracy of predicting RBP bindings has been significantly improved on benchmarked data [11].

There is one recent paper uses a set of structural discovery algorithms followed by SVM training [12]. The authors trained a novel classification model (CisRNA-SVM) on a set of known structured cis-regulatory elements from 3'UTRs. Their data is selected from TargetScan. The interesting part is that the four programs chosen for the training process are RNAalifold, LocARNA, Foldalign and Cove. They represented four different categories of folding strategies: align-then-fold, fold-then-align, simultaneous fold-and-align and covariance-based methods.

There are also several recent papers [13] which focus on protein side and its computational methods (mainly machine learning). Since this is not my focus and my RPE report shares large overlaps with these papers, no further descriptions here.

1.1.2 Experimental datasets

Besides of the computational methods, experimental datasets are also of great importance to my study. As shown in Figure 1.1, from RNA side, there are in vivo experimental techniques RIP and CLIP, and in vitro techniques like SELEX and RNA profiling; from protein side, there are mostly structural techniques like x-ray and NMR. Here in this project, I will primarily focus on in vivo techniques for mRNA, mainly RIP and CLIP.

RIP

RIP is the conventional method for studying RNA-protein interactions. One benchmarked dataset that has been used in many computational papers is generated from Hogan *et al.* [11]. It contains Yeast RIP-chip (RNA-binding protein immunoprecipitation coupled to reverse transcription and a microarray) data of 12,000 pairs of RBP-RNA interactions. It serves as the proof-of-principle data set for several previous studies [14, 8].

Lange *et al.* have recently published a paper on predicting secondary structure and accessibility in mRNAs [15]. The paper provides two benchmarked datasets for secondary structures of mRNA functional elements: 2,500 structured cis-regulatory elements in 95 Rfam families, and 3,196 *in vivo* Saccharomyces cervisiae mRNAs with secondary structure available.

CLIP

The recent cutting-edge technology for RNA-protein interactions is CLIP, i.e. high-throughput (UV) cross-linking immunoprecipitation technique. There is a nice review paper on all kinds of high-throughput CLIP methods [16]. It also introduced downstream procedures to identify binding sites from sequencing data of CLIP. There are two major databases CLIPZ and doRiNA which collect CLIP data generated from several recent large scale experiments. Each data set from these two database contains transcriptome-wide binding site data for both RBPs and miRNAs. DoRiNA hosts 17 such CLIP sets for Human, Mouse, C.elegans. And CLIPZ hosts 999,350 CLIP-determined AGO protein binding sequences for the same species.

Especially for doRiNA database, it is based on the thought that the combinatorial action of RBPs and miRNAs on target mRNAs form a post-transcriptional regulatory code. The authors [4] provide this database that supports the quest for deciphering this code. The database is built based on Hafner's PAR-CLIP data, supplement with other CLIP datasets. They

assign two quality scores based on the characteristics of the PAR-CLIP protocol: # of conversions and entropy of reads.

Since CLIP data is really new and of large scale, it may be very useful for the future studies of Protein-binding RNAs. While, it has many different variants like HITS-CLIP (normal form of CLIP), PAR-CLIP, iCLIP etc., and they are sometimes confusing to researchers in this field. So several representative papers are reviewed here as a comparison of different CLIP techniques.

CLIP vs PAR-CLIP In this paper, the authors found only small differences in accuracies of these methods in identifying binding sites of HuR [1]. To determine whether differences between protocols are reflected in the set of identified target sites, they started from PAR-CLIP and individually modified the two steps that are most likely to bias the identification of binding sites: cross-linking and RNase digestion.

CLIP vs PAR-CLIP vs iCLIP The authors [2] found that reverse transcriptase used in CLIP frequently skips the crosslinked amino-acid-RNA adduct, resulting in a nucleotide deletion, which is more precise of mapping protein-RNA interactions than currently available PAR-CLIP and iCLIP. The used Nova and Argonaut (Ago) HITS-CLIP data. Especially, the cross-linking induced mutation frequency in standard CLIP is lower than that observed from PAR-CLIP, but more meaningful comparisons have to consider signal-to-noise, which is $\tilde{1}550$ fold for CIMS analysis ($\tilde{8}20\%$ crosslinking mutation rate vs. $\tilde{0}.40.5\%$ background deletion rate due to sequencing or alignment errors), and 45 fold in PAR-CLIP (5080% cross-linking induced TC transition vs. 1020% spontaneous transitions).

iPAR-CLIP In vivo PAR-CLIP (iPAR-CLIP) is introduced and performed in this paper [3]. They use KH domain-containing RBP GLD-1 protein which is known to recognize a relatively well-defined primary sequence motif. Note that, according to the paper, Lebedeva2011 and Mukherjee2011 datasets (collected in doRiNA as well) are NOT in vivo.

ENCODE

In the recent published ENCODE project [17], There is a RIP dataset for several human RNA-binding proteins, but not been focused in ENCODE papers.

The tracks in this supertrack contain two forms of information: genes whose transcripts were bound by the given RBP (such as **SUNY RIP GeneSt**) and approximate location of the RBP binding site in the mRNA sequence (such **SUNY RIP Tiling** and **SUNY RIP-seq**). An overview of the data is shown in Table 1.1

	RIP Tiling		RIP-seq	
	GM12878	K562	GM12878	K562
ELAVL1	Y	Y	Y	Y
PABPC1	Y	Y	Y	Y
Negative Control	Y	Y		
Input(Total)	Y	Y		

Table 1.1: The ENCODE datasets for RBP binding sites of *Human*.

And the files are mainly encoded in broadPeak formats. This format is used to provide called regions of signal enrichment based on pooled, normalized (interpreted) data. It is a BED 6+3 format. So besides of normal BED 6 format which has 6 columns to denote: chrom, chromStart, chromEnd, name, score, and strand, the broadPeak format has 3 additional columns: signalValue, pValue and qValue, which are measurements of enrichment for the region.

All the three tier 1 cell lines of ENCODE project are included:

GM12878 is a lymphoblastoid cell line produced from the blood of a female donor with northern and western European ancestry by EBV transformation. It was one of the original HapMap cell lines and has been selected by the International HapMap Project for deep sequencing using the Solexa/Illumina platform. This cell line has a relatively normal karyotype and grows well. Choice of this cell line offers potential synergy with the International HapMap Project and genetic variation studies. It represents the mesoderm cell lineage. Cells will be obtained from the Coriell Institute for Medical Research [coriell.org] (Catalog ID GM12878).

K562 is an immortalized cell line produced from a female patient with chronic myelogenous leukemia (CML). It is a widely used model for cell biology, biochemistry, and erythropoiesis. It grows well, is transfectable, and represents the mesoderm linage. Cells will be obtained from the America Type Culture Collection (ATCC) [atcc.org] (ATCC Number CCL-243).

H1 human embryonic stem cells will be obtained from Cellular Dynamics International [cellulardynamics.com].

1.2 The proposed study

In this project, we propose to assess the secondary structural features of *in vivo* protein binding sites on mRNA transcripts. mRNA sequences have the remarkable ability to form structures which define the functional roles they play in the cell. Proteins, in turn, are known to recognize and bind to certain RNA transcripts with specific structures in a sequence-specific or unspecific way. Currently, there already exist a number of known protein binding sites in mRNA sequences. While, there are only a few computational methods have considered the structural features of these binding sites previously[8, 18], and they all assume that the RNA transcript folds into its thermodynamically most stable structure.

We know, however, that this assumption generally does not hold *in vivo*. Previous research has shown that the optimized thermodynamic structure does not necessarily correspond to the structure that is functional in the cell, especially for molecules such as pre-mRNAs and long mRNAs [19]. The reasons may be the effect of co-transcriptional folding [20] (when the mRNA molecule is being transcribed, it folds at the same time; during this process, the mRNA forms a series of structures which are not the same as the thermodynamic structure) and the influence of other molecules binding the mRNA molecule.

We therefore propose to employ most widely used comparative methods instead to detect evolutionarily conserved RNA structures. These methods include Pfold [21], RNAalifold [10], and RNA-Decoder [22], which do not make the above thermodynamic assumption. Comparative methods in general tend to outperform thermodynamic methods in terms of prediction accuracy [23]. They make use of the evolutionary information from homologous RNA sequences by assuming that homologous RNAs which have similar functions may also have similar structures for carrying out their biological functions *in vivo*. Besides, as a key part of the project, the data set for mRNA-protein interactions is of crucial importance. Recently, several large scale experiments have identified a number of protein-RNA interactions, which makes it possible to systematically study the features of protein binding sites in RNA sequences.

The work has the potential to:

1. contribute a systematic study of mRNA-protein interactions,
2. increase the prediction accuracy of protein binding sites and,
3. detect new types of RNA-protein interactions.

Chapter 2

Results

In this project, we assess the accessibility of mRNAs on both motif-wise and base-wise. Here **motif-wise accessibility** means that we compute the accessibility values for each protein binding motif by taking all the bases in the motif into consideration. And **base-wise accessibility** is based on comparisons of accessibilities between all binding motif positions and all UTR positions at individual base/position level. A motif is a sub-string that contains several consecutive bases on the transcript, normally on 3'UTR region. In general, we want to see if there is any difference of accessibilities between motif region and the background whole UTR region on each transcript.

Here, we apply three comparative methods:

- Pfold
- RNAalifold
- RNA-Decoder

And we use two measurements to define the **accessibility**:

A. the rate of the base-pairing cases in binding sites according to the consensus structures ("fold" mode);

B. the base-pairing probability across binding sites ("scan" mode).

So for each of the three programs, we calculate the values of (A) and (B) for every RNA-binding protein which has binding sites information available in our dataset.

Besides, when calculating base paring probabilities (B) and consensus structures (A), they are calculated according to different regions: N (non-structured region), P (paired region), L (loop/bulge region) for both binding sites and (3'UTR) background sites. In below, I will use NPL to denote such classification.

2.1 RNA-binding proteins and binding sites

In this study, RNA-binding proteins and binding sites are derived from [8]. The data is originally generated by RNP immunoprecipitation-microarray (RIP-chip) experiment. All together, it comprises 18 previously defined RBPs that bind to 3'UTR region of Yeast, Human and Fly mRNA sequences, and has consensus binding sequences (motif) available (Table 2.1). We map back these binding sequences to the corresponding transcripts (and alignments) using a sliding-window based approach.

We download from UCSC Genome Browser [24] the genomes and corresponding multiple genomic sequence alignments for Yeast, Human and Fly species. In order to compare with the work in [8], we choose the same versions as in that paper, namely Yeast (sacCer2, i.e. SGD1.01), Human (hg18) and Fly (dm3, i.e. BDGP Release 5). We also downloaded the annotation for each genome from the Ensembl database [25]. We use a tool, mafsInRegion from UCSC to fetch all of the mRNA transcripts (whole transcript and the three prime untranslated region (3'UTR)) from each genomes' multiple species alignments. We then utilize a widely used perl library Bioperl [26] (version 1.006001) to convert the raw UCSC alignments from maf format to different formats as required by the different programs.

2.1.1 Overview of RBPs

For each of these proteins, the primary sources of its binding information are NCBI, InterPro and UniProt. Besides, for the binding type, the information is directly got from these databases (very rare), or from PDB and GO database (structure from PDB; ss or ds term associated in GO), or from individual sources including species database, like SGD (Saccharomyces Genome Database) or other protein database, like SMART or Pfam, or individual papers. The source of the binding type is annotated in the parentheses in the table below.

Table 2.1: Overview of Yeast RBPs.

RBPs	**Motif**	**Binding type**	**Database ID**	**Remark**

Msl5	UACUAAC	single (inferred from its KH domain)	PDB: 3FMA	pre-mRNA branch point binding; AU-rich elements binding (see AREsite); cooperatively recognize a tetra-loop structure [27]
Puf4	UGUAHMNUA	single (PDB)	PDB: 3BX3 3BWT 3BX2 4DZS	Puf3, Puf4 and Puf5 in Yeast, and Pum1 in humans, are known to bind UGUR tetranucleotide motif [28]
Puf3	CNUGUAHAUA	single (PDB)	PDB: 3K49 3K4E	PUF protein family
Khd1	CNNCNN	single (inferred from its KH domain)	InterPro: P38199	aka HEK2, YBL032W; AU-rich elements binding (see AREsite)
Nab2	DRARAMGMD		PDB: 2JPS 2V75 2LHN 3LCN 4H1K	form complex with Gfd1; poly(A) RNA binding
Yll032c	AAUACCY	single (inferred from its KH domain)		Protein of unknown function that may interact with ribosomes; KH domain contained
Vts1	CNGG	single (NCBI, GO)	PDB: 2D3D 2FE9 2B6G 2F8K 2ES6 2ESE	bind an RNA hairpin termed the Smaug recognition element (SRE); flap-structured DNA and RNA binding (GO)
Pub1	HUUUUUUHW	single (InterPro)	PDB: 2LA4 3MD1 3MD3	poly(A)+ mRNA binding; RRM domain (single-stranded RNA binding) contained

Puf2	UAAUAAW	single (inferred from other PUF proteins)	UniProt: Q12221	Member of the PUF protein family, like Puf4 etc., Pumilio homology domain contained; RRM domain contained
Puf5	WUUGUAWUWU	single (inferred from other PUF proteins)	UniProt: P39016	see Puf4, Pumilio domain contained;
Ssd1	AKUCAUUCCUU	single (PDB Homolog)	UniProt: P24276; PDB: 2VNU 2WP8 4IFD (homolog according to PROSITE and SGD)	RNR ribonuclease family
PAB1	WUAUAUAW	single (inferred from its RRM domain)	PDB: 1IFW	contains PABC domain which is poly(A) binding; RRM domain contained
Nsr1	GGGWAACGGW	single (inferred from its RRM domain)	UniProt: P27476	RRM domain contained; nucleolar protein that binds nuclear localization sequences; required for pre-rRNA processing; single-stranded DNA binding

Nrd1	UUCUUGUW	single (inferred from its RRM domain)	PDB: 2LO6 3CLJ	By recruiting it to RNA Pol II, Nrd1 could co-operate with Nab3 and Sen1 to terminate small nucleolar RNAs and other short RNAs; 3'-end processing of mRNA, snoRNA, snRNA and tRNA; RRM domain contained
Pum1	UGUAHAUA	single (inferred from other PUF proteins)	PDB: 1m8z (homolog)	Pumilio homolog; PUF protein family

Note that, for those RBPs which have different binding motifs, combine by picking the most general one (eg. for CNGG and CNGGN, pick CNGG).

Detailed description of RBPs

Msl5

1. *Species* Yeast
2. *Function* nuclear mRNA splicing, via spliceosome
3. *Domain* KH, 2ZnF_C2H2
4. *Description* Component of the commitment complex, which defines the first step in the splicing pathway; essential protein that interacts with Mud2p and Prp40p, forming a bridge between the intron ends; also involved in nuclear retention of pre-mRNA.

Puf4

1. *Species* Yeast
2. *Function* Loss of chromatin silencing during replicative cell aging, mRNA catabolic process, nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay, protein localization.

3. *Domain* 8Pumilio repeat

4. *Description* Member of the PUF protein family, which is defined by the presence of Pumilio homology domains that confer RNA binding activity; preferentially binds mRNAs encoding nucleolar ribosomal RNA-processing factors.

Puf3

1. *Species* Yeast

2. *Function* aerobic respiration, intracellular mRNA localization, mitochondrion localization, mitochondrion organization, nuclear-transcribed mRNA catabolic process, nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay.

3. *Domain* 8Pumilio repeat

4. *Description* Protein of the mitochondrial outer surface, links the Arp2/3 complex with the mitochore during anterograde mitochondrial movement; also binds to and promotes degradation of mRNAs for select nuclear-encoded mitochondrial proteins.

Khd1

1. *Species* Yeast

2. *Function* Cytoplasm, nuclear chromosome, telomeric region.

3. *Domain* 3KH

4. *Description* RNA binding protein involved in the asymmetric localization of ASH1 mRNA; represses translation of ASH1 mRNA, an effect reversed by Yck1p-dependent phosphoryation; regulates telomere position effect and length; similarity to hnRNP-K.

Nab2

1. *Species* Yeast

2. *Function* mRNA polyadenylation, poly(A+) mRNA export from nucleus, regulation of mRNA stability.

3. *Domain* coiled coil

4. *Description* Nuclear polyadenylated RNA-binding protein required for nuclear mRNA export and poly(A) tail length control; binds nuclear pore protein Mlp1p; autoregulates mRNA levels; related to human hnRNPs; nuclear localization sequence binds Kap104p.

Yll032c

1. *Species* Yeast
2. *Domain* KH
3. *Description* Protein of unknown function that may interact with ribosomes, based on co-purification experiments; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm; YLL032C is not an essential gene. Not use.

Vts1

1. *Species* Yeast
2. *Function* Nuclear-transcribed mRNA catabolic process, Nuclear-transcribed mRNA poly(A) tail shortening.
3. *Domain* Vts1
4. *Description* Post-transcriptional gene regulator, RNA-binding protein containing a SAM domain; shows genetic interactions with Vti1p, which is a v-SNARE involved in cis-Golgi membrane traffic.

Pub1

1. *Species* Yeast
2. *Function* Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, regulation of mRNA stability, stress granule assembly.
3. *Domain* 3RRM
4. *Description* Poly (A)+ RNA-binding protein, abundant mRNP-component protein that binds mRNA and is required for stability of many mRNAs; component of glucose deprivation induced stress granules, involved in P-body-dependent granule assembly.

Puf2

1. *Species* Yeast
2. *Function* Nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay.
3. *Domain* 6Pumilio, RRM
4. *Description* Member of the PUF protein family, which is defined by the presence of Pumilio homology domains that confer RNA binding activity; preferentially binds mRNAs encoding membrane-associated proteins.

Puf5

1. *Species* Yeast
2. *Function* Cell wall organization, loss of chromatin silencing during replicative cell agin, nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay, protein localization, re-entry into mitotic cell cycle after pheromone arrest, replicative cell aging.
3. *Domain* 8Pumilio repeat
4. *Description* Member of the Puf family of RNA-binding proteins; binds to mRNAs encoding chromatin modifiers and spindle pole body components; involved in longevity, maintenance of cell wall integrity, and sensitivity to and recovery from pheromone arrest

Ssd1

1. *Species* Yeast
2. *Function* RNA binding, cell wall organization, chronological cell aging, replicative cell aging, response to drug.
3. *Domain* Coiled coil
4. *Description* Protein with a role in maintenance of cellular integrity, interacts with components of the TOR pathway; ssd1 mutant of a clinical S. cerevisiae strain displays elevated virulence.

PAB1

1. *Species* Yeast
2. *Function* Regulation of translational initiation.
3. *Domain* 4RRM, PolyA
4. *Description* Poly(A) binding protein, part of the 3'-end RNA-processing complex, mediates interactions between the 5' cap structure and the 3' mRNA poly(A) tail, involved in control of poly(A) tail length, interacts with translation factor eIF-4G.

Nsr1

1. *Species* Yeast
2. *Function* Ribosomal small subunit assembly, rRNA processing.
3. *Domain* 2RRM, coiled coil
4. *Description* Nucleolar protein that binds nuclear localization sequences, required for pre-rRNA processing and ribosome biogenesis.

Nrd1

1. *Species* Yeast
2. *Function* Termination of RNA polymerase II transcription, poly (A)-independent.
3. *Domain* RPR, RRM
4. *Description* RNA-binding protein that interacts with the C-terminal domain of the RNA polymerase II large subunit (Rpo21p), preferentially at phosphorylated Ser5; required for transcription termination and 3' end maturation of nonpolyadenylated RNAs.

Pumilio

1. *Species* Fly
2. *Function* mRNA 3'-UTR binding; protein binding; mRNA binding; translation repressor activity, nucleic acid binding.

3. *Domain* 8Pumilio repeat, coiled coil

4. *Description* The gene pumilio is referred to in FlyBase by the symbol Dmel
pum (CG9755, FBgn0003165). The phenotypes of these alleles are annotated with 46 unique terms, many of which group under: organ system; embryonic abdomen; embryonic segment; anatomical structure; female germline cyst; nervous system; embryonic tagma; peripheral nervous system; germarium; embryonic neuron; multi-cell-component structure; cephalopharyngeal skeleton. It has 5 annotated transcripts and 5 annotated polypeptides.

HuR

1. *Species* Human

2. *Function* multicellular organismal development, mRNA stabilization.

3. *Domain* 3RRM

4. *Description* The protein encoded by this gene is a member of the ELAVL protein family. This encoded protein contains 3 RNA-binding domains and binds cis-acting AU-rich elements. It destabilizes mRNAs and thereby regulates gene expression.

Pum1

1. *Species* Human

2. *Function* Membrane organization, post-Golgi vesicle-mediated transport, regulation of translation.

3. *Domain* Cytoplasm, cytosol

4. *Description* A member of the PUF family, evolutionarily conserved RNA-binding proteins related to the Pumilio proteins of Drosophila and the fem-3 mRNA binding factor proteins of C. elegans. The encoded protein contains a sequence-specific RNA binding domain comprised of eight repeats and N- and C-terminal flanking regions, and serves as a translational regulator of specific mRNAs by binding to their 3' untranslated regions. The evolutionarily conserved function of the encoded protein in invertebrates and lower vertebrates suggests that the human protein may be involved in translational regulation

of embryogenesis, and cell development and differentiation. Alternatively spliced transcript variants encoding different isoforms have been described.

PTB

1. *Species* Human
2. *Function* RNA binding, Nucleotide binding, poly-pyrimidine tract binding, protein binding.
3. *Domain* 4RRM
4. *Description* This protein belongs to the subfamily of ubiquitously expressed heterogeneous nuclear ribonucleoproteins (hnRNPs). The hnRNPs are RNA-binding proteins and they complex with heterogeneous nuclear RNA (hnRNA). These proteins are associated with pre-mRNAs in the nucleus and appear to influence pre-mRNA processing and other aspects of mRNA metabolism and transport. While all of the hnRNPs are present in the nucleus, some seem to shuttle between the nucleus and the cytoplasm. The hnRNP proteins have distinct nucleic acid binding properties. The protein encoded by this gene has four repeats of quasi-RNA recognition motif (RRM) domains that bind RNAs. This protein binds to the intronic polypyrimidine tracts that requires pre-mRNA splicing and acts via the protein degradation ubiquitin-proteasome pathway. It may also promote the binding of U2 snRNP to pre-mRNAs. This protein is localized in the nucleoplasm and it is also detected in the perinucleolar structure. Alternatively spliced transcript variants encoding different isoforms have been described.

2.1.2 Base-wise accessibility

As described at the beginning of this chapter, the study of **base-wise accessibility** here is based on comparisons between all binding motif positions and all UTR positions at single base/position level. We would like to detect if there is any difference between the accessibilities at motif region and background region.

And as a pilot study, here we focus on the Pumilio protein of Fly species. We first look at only those transcripts with single UTR (around 7000 such

transcripts). So for each of the three programs Pfold, RNAalifold and RNA-Decoder, we run it in both "scan" and "fold" modes respectively with NPL classification, as described at the beginning of this chapter.

"Scan" mode

Definition: A binding motif is a sub-string on the UTR sequence, which matches a particular protein binding consensus profile as shown in Table 2.1. A binding site is a base within a binding motif. "Scan" mode is to assign a pairing probability to each single base. And the $base\,wise\,accessibility = 1 - P(Pairing)$.

As we can see from Figure2.1 and 2.2, for PPfold and RNAalifold, the averaged base-wise pairing probabilities at background paired regions (3'UTR, P) are higher than those at binding motif paired regions (binding sites, P). But for RNAdecoder2.3, it is not the case. Besides at loop regions and non-structured regions (L and N), neither the averaged pairing probabilities of background nor of binding motif is consistently stronger than the other.

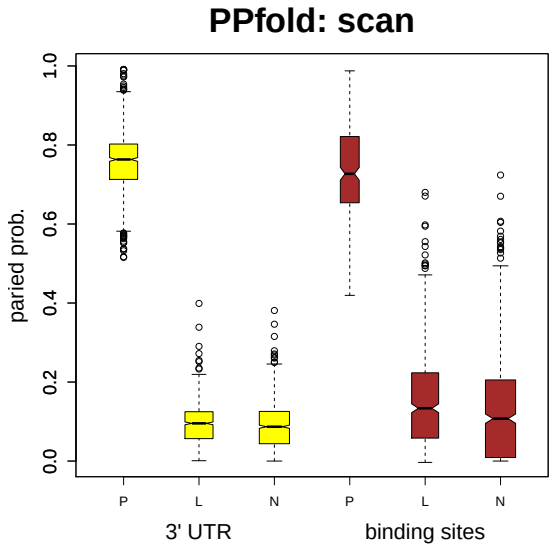

Figure 2.1: Boxplot of base-pairing probabilities for Pumilio binding sites positions and 3'UTR positions. The base-pairing probabilities are generated by PPfold for each alignment position. For each transcript, the probabilities of positions within binding sites region and those in background region (3'UTR) are averaged respectively, and then plotted for comparison.

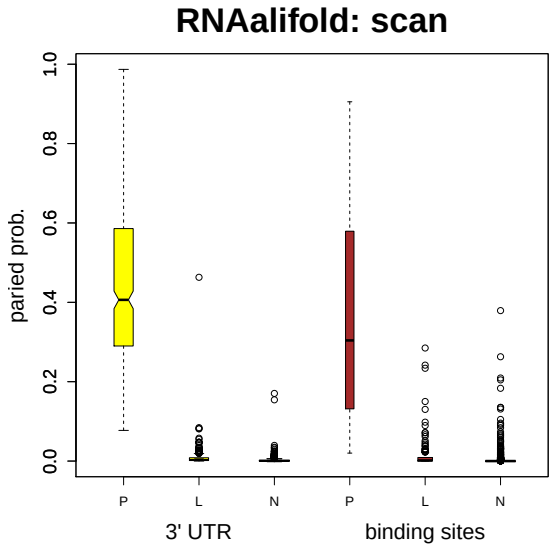

Figure 2.2: Boxplot of base-pairing probabilities for Pumilio binding sites positions and 3'UTR positions. The base-pairing probabilities are generated by RNAalifold for each alignment position. For each transcript, the probabilities of positions within binding sites region and those in background region (3'UTR) are averaged respectively, and then plotted for comparison.

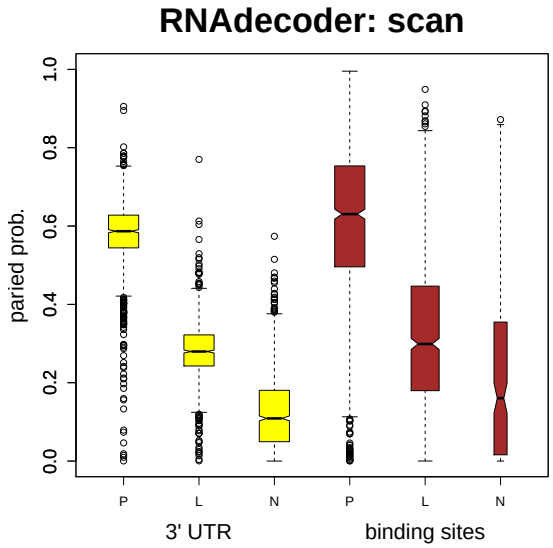

Figure 2.3: Boxplot of base-pairing probabilities for Pumilio binding sites positions and 3'UTR positions. The base-pairing probabilities are generated by RNAdecoder for each alignment position. For each transcript, the probabilities of positions within binding sites region and those in background region (3'UTR) are averaged respectively, and then plotted for comparison.

"Fold" mode

Definition: "Fold" mode is the labeling process, it is aimed to assign a consensus label (one out of N,L,P) to a given base. Here the probability computed is the frequency of a label x ϵ {N,L,P}, P(x) = (count of sites labeled with x)/(count of the total sites).

As we can see from Figure2.4, 2.5 and 2.6, different programs have obvious different strategies when predicting consensus structures. RNAdecoder tends to predict more base pairs (Figure2.4); while RNAalifold tends to predict less base pairs (i.e. less structures) (Figure2.5); PPfold is in between (Figure2.6). In general, the background regions have relatively more base pairing rates (P) and loop rates (L) than binding sites regions and less non-structured rates (N), which is as expected.

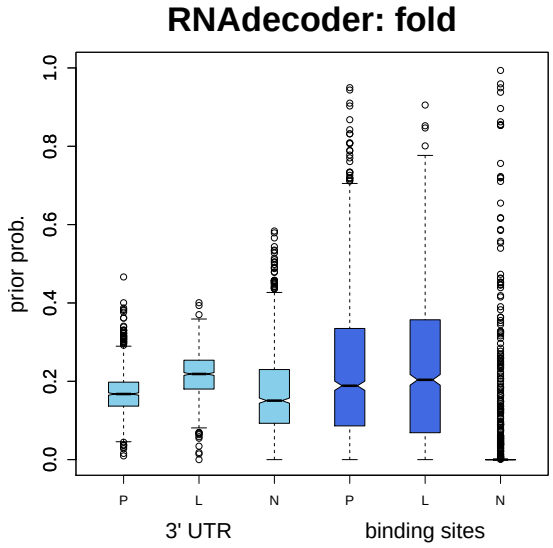

Figure 2.4: Boxplot of structural annotation probabilities for Pumilio binding sites positions and 3'UTR positions. The structural annotation probabilities are generated by RNAdecoder for each alignment position. For each transcript, the probabilities of positions within binding sites region and those in background region (3'UTR) are averaged respectively, and then plotted for comparison.

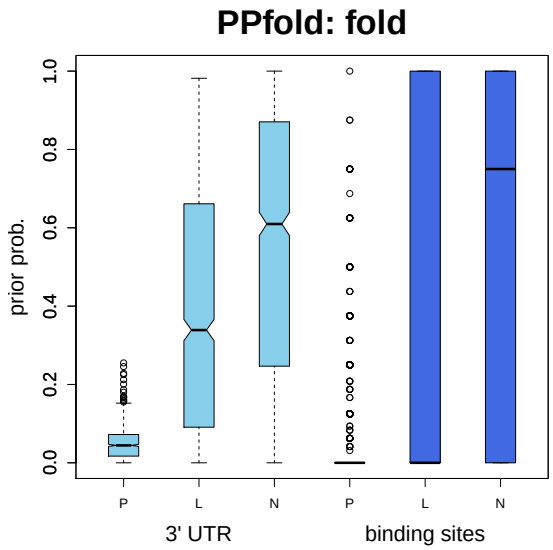

Figure 2.5: Boxplot of structural annotation probabilities for Pumilio binding sites positions and 3'UTR positions. The structural annotations are generated by PPfold for each alignment position. Next, for positions within binding sites region and those in background region (3'UTR) of a transcript, the frequencies for N,P,L categories are used as probabilities. For each transcript, the probabilities are then plotted for comparison.

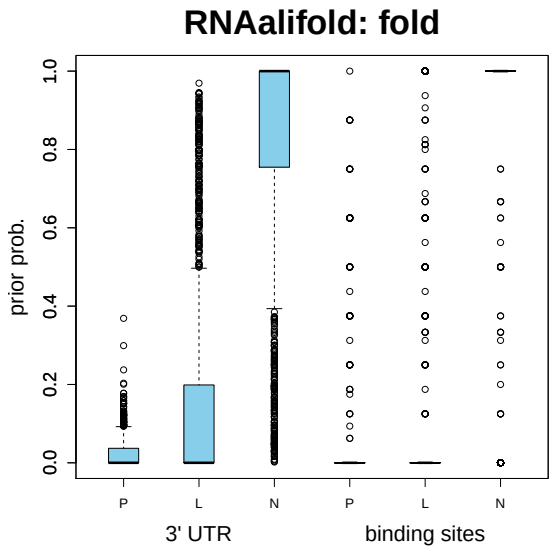

Figure 2.6: Boxplot of structural annotation probabilities for Pumilio binding sites positions and 3'UTR positions. The structural annotations are generated by RNAalifold for each alignment position. Next, for positions within binding sites region and those in background region (3'UTR) of a transcript, the frequencies for N,P,L categories are used as probabilities. For each transcript, the probabilities are then plotted for comparison.

2.1.3 Motif-wise accessibility

As described at the beginning of this chapter, for **motif-wise accessibility**, we look at each motif as a whole rather than consider single base. Similar to base-wise accessibility, we compare the motifs and background regions in "scan" and "fold" modes with N,P,L classification. Compared to base-wise accessibility, the motif-wise accessibility provides more detailed information of the pairing probability distribution on binding motif and background 3'UTRs, which is more important.

In the study of [8], the motif-wise accessibility is computed naturally by RNAplfold: consider the probability of the motif being single stranded in all Boltzmann distributed structures. As we would like to compare with RNAplfold using our comparative methods, we compute such motif-wise accessibility by taking average of the accessibilities of all single bases on a motif. We consider RBPs in all three species. Since there are too many of RBPs, only a subset of representative proteins in Yeast is shown below.

"Scan" mode

Similar to the definition in previous section, the "Scan" mode here is to assign a pairing probability to a sub-string that contains multiple consecutive bases. So $P(Pairing) = Prob^b_{avg}$ = (sum of pairing probabilities for all binding sites in a motif)/(number of the binding sites in that motif), and $Accessibility = 1 - P(Pairing)$.

First we look at the motif level accessibility on a simple case. As shown in Table 2.1, Msl5 motif does not occur very frequent in the Yeast UTR regions, so each of the transcript's UTR contains at most one hit of the motif Figure 2.7.

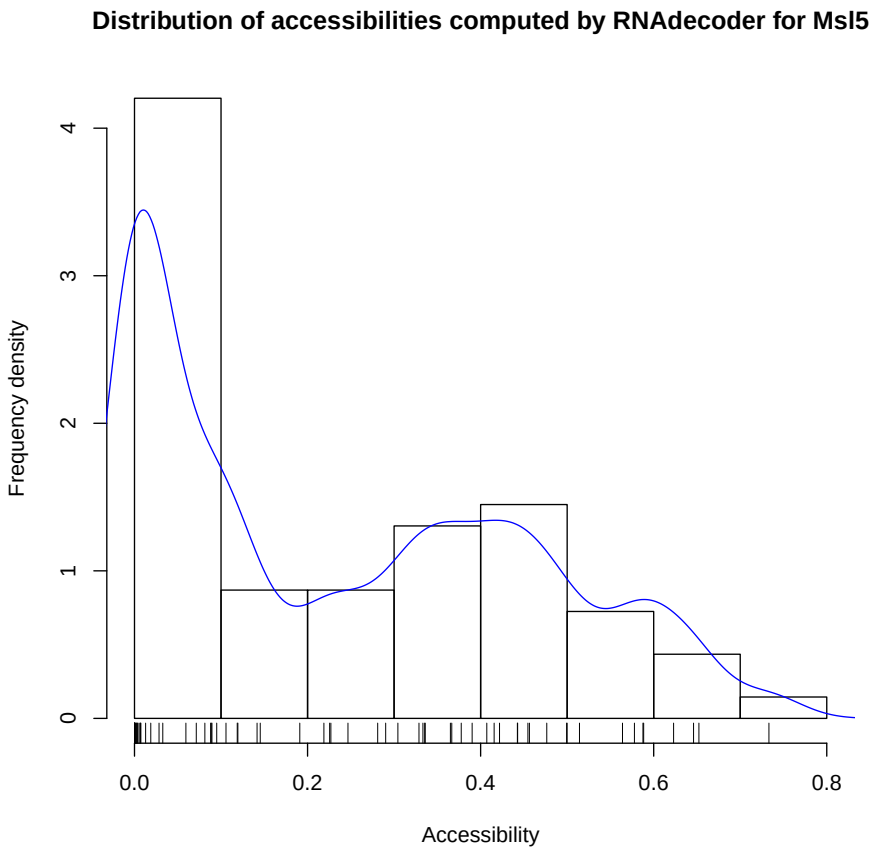

Figure 2.7: **Distribution of accessibilities computed by RNAdecoder for Msl5 protein.** Histogram plot of Msl5 motif-wise accessibility. $Accessibility = 1 - Pairing = 1 - Prob^{b}_{avg}$. The base pairing probabilities are generated by RNA-decoder for each alignment position.

Puf2 motif occurs more frequent than Msl5 in the Yeast UTR regions, so there are transcripts containing more than one motif. So we have two plots here: one for those UTRs with single motif Figure 2.8, one for UTRs with multiple motifs Figure 2.9. In the second case, we further plot that: for any given such UTR, the accessibility value for the best/worst accessible binding motif.

As we can see, for those transcripts with single Puf2 motif in Figure 2.8, the distribution is quite similar to Msl5 case in Figure 2.7. And for those UTRs with multiple Puf2 binding motifs, the best accessible motifs are better than the worst motifs at a wide range, with a weak concentration at $worst = (0.25, 0.5)$ region as shown in Figure 2.9.

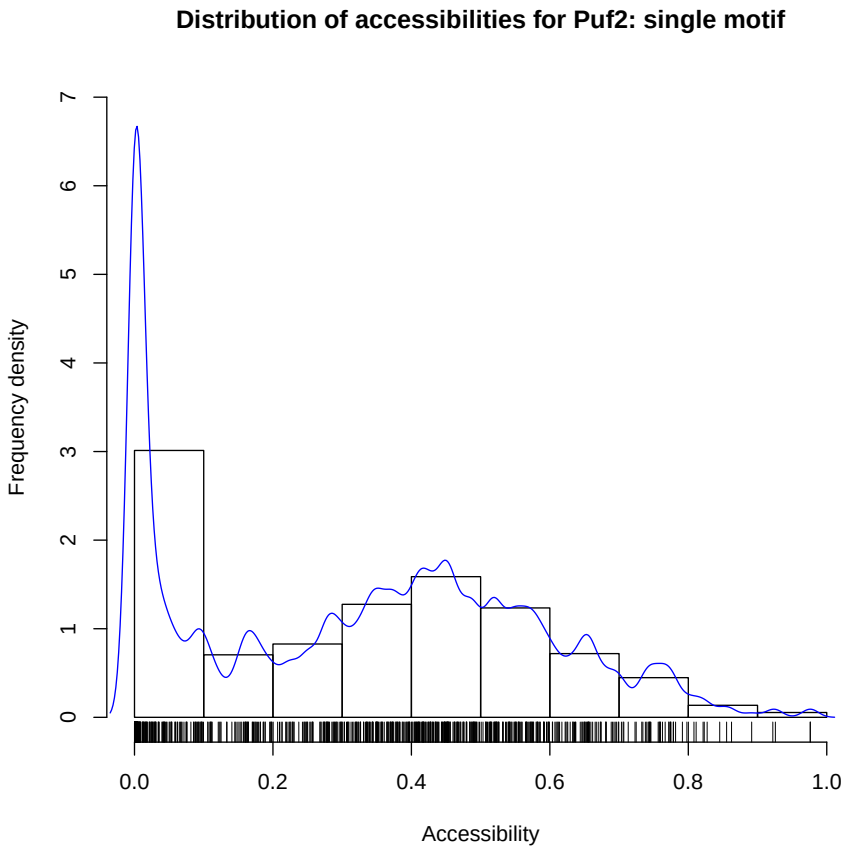

Figure 2.8: **RNAdecoder: Puf2 single motif accessibility distribution.** Histogram plot of Puf2 single motif accessibility. The figure is for those UTRs with single motif. $Accessibility = 1 - Pairing = 1 - Prob^{b}_{avg}$. The base pairing probabilities are generated by RNA-decoder for each alignment position.

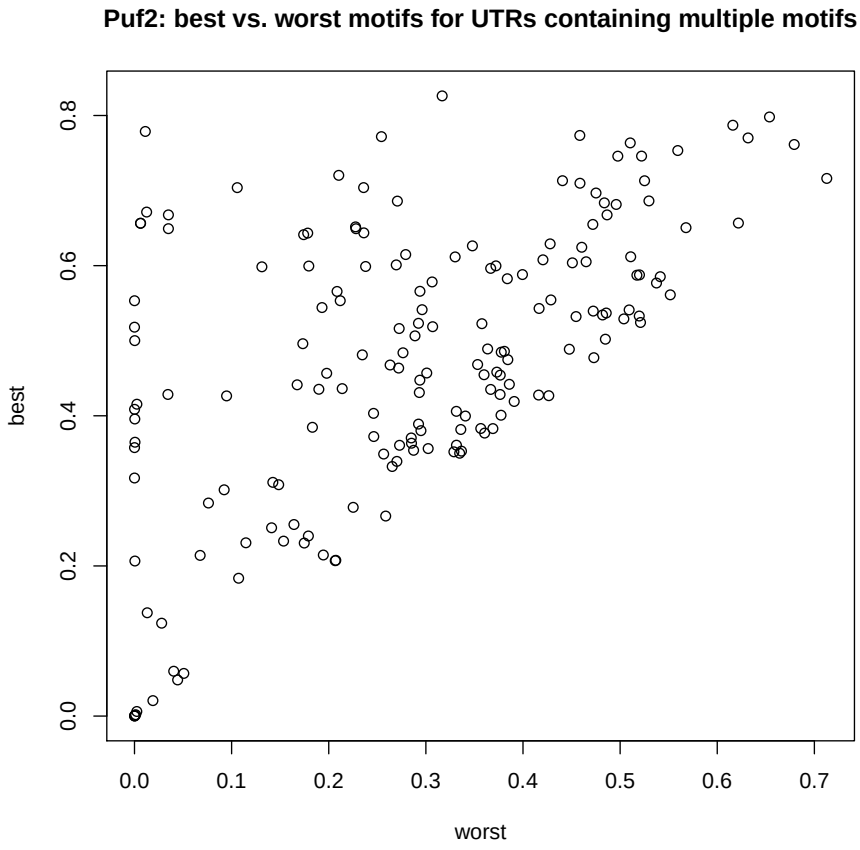

Figure 2.9: **RNAdecoder: best vs. worst Puf2 motifs for UTRs containing multiple binding motifs.** Scatter plot of Puf2 multiple motifs accessibility. Each dot in the plot represents a UTR that contains multiple Puf2 binding motifs. The coordinates of any dot are $(worst, best)$, where $worst = Accessibility^{b}_{min}$, and similarly $best = Accessibility^{b}_{max}$. The structural annotations are generated by RNA-decoder for each alignment position.

"Fold" mode

As defined in previous section, the "Fold" mode is to assign a sequence of consensus label (one out of N,L,P) to a base. While here, we consider the N,L,P compositions in a binding motif. Also, we look at the differences between individual binding motif and its background UTR. Unlike in "Scan" mode, the detailed computation of the probabilities are different among the three programs. The below figures (Fig 2.10, Fig 2.11 and Fig 2.12) show such differences on an example protein Puf3-1. We could observe most of the dots are in quadrant 3 (bottom left) in PPfold and RNAalifold cases while for RNAdecoder the dots are more equally scattered. Still, this could be due to the fact that PPfold and RNAalifold tend to predict less structures (so less pairing and loop) than RNAdecoder.

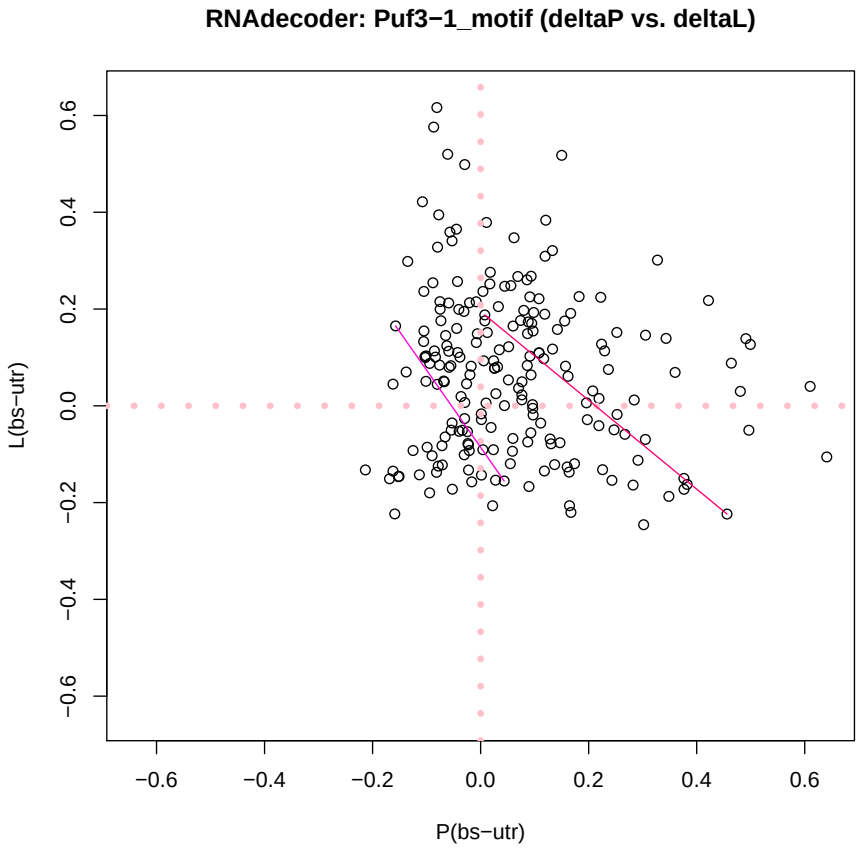

Figure 2.10: Scatter plot of structural annotation probabilities for L vs P, in RNAdecoder case. Each dot in the plot represents the difference between a Puf3-1 protein binding motif and its UTR. So the coordinates of any dot are $(\Delta P, \Delta L)$, where $\Delta P = P^b_{avg} - P^U_{avg}$ with P^b_{avg} = (sum of likelihoods for binding sites labeled with P)/(length of the binding motif), and similarly $\Delta L = L^b_{avg} - L^U_{avg}$. The structural annotations are generated by RNAdecoder for each alignment position. Since one UTR may contain more than one binding motif, we use lines with the same colour to connect dots from the same UTR.

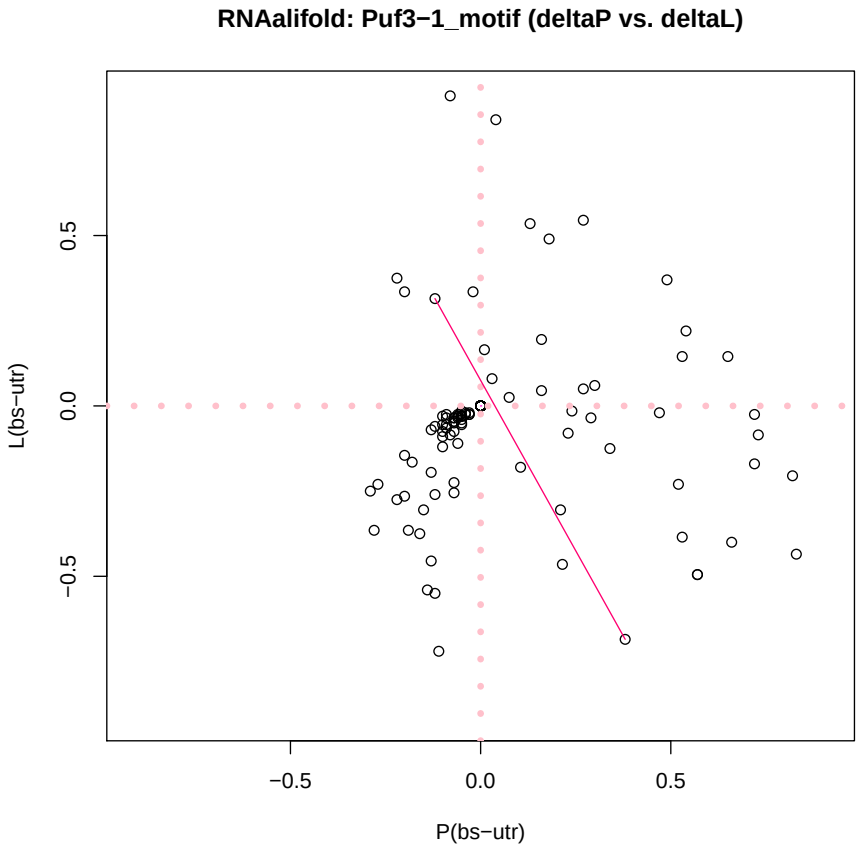

Figure 2.11: Scatter plot of structural annotation probabilities for L vs P. Each dot in the plot represents the difference between a Puf3-1 protein binding motif and its UTR. So the coordinates of any dot are $(\Delta P, \Delta L)$, where $\Delta P = P^b_{avg} - P^U_{avg}$ with P^b_{avg} = (count of binding sites labeled with P)/(length of the binding motif), in RNAalifold case, and similarly $\Delta L = L^b_{avg} - L^U_{avg}$. The structural annotations are generated by RNAalifold for each alignment position. Since one UTR may contain more than one binding motif, we use lines with the same colour to connect dots from the same UTR. Note that the other line is not shown in this figure as that UTR is too long for RNAalifold to compute.

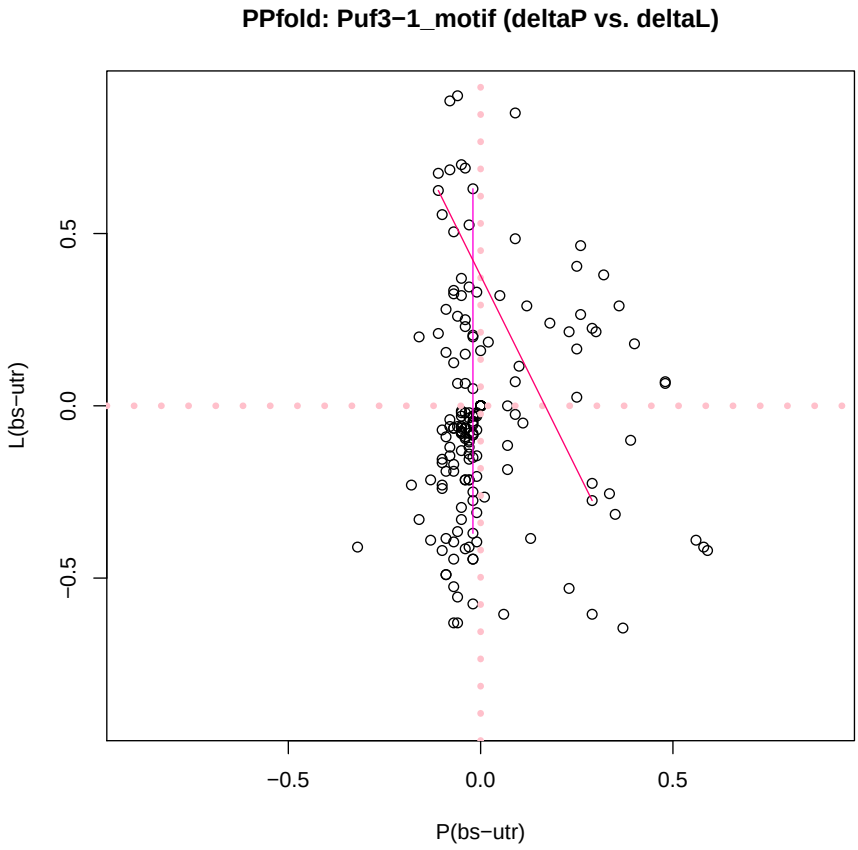

Figure 2.12: Scatter plot of structural annotation probabilities for L vs P, in PPfold case. Each dot in the plot represents the difference between a Puf3-1 protein binding motif and its UTR. So the coordinates of any dot are $(\Delta P, \Delta L)$, where $\Delta P = P^b_{avg} - P^U_{avg}$ with P^b_{avg} = (count of binding sites labeled with P)/(length of the binding motif), and similarly $\Delta L = L^b_{avg} - L^U_{avg}$. The structural annotations are generated by PPfold for each alignment position. Since one UTR may contain more than one binding motif, we use lines with the same colour to connect dots from the same UTR.

Besides, since RNAdecoder is the only one program among the three that can calculate the likelihood for each assigned structural label, we further look at the N,P,L distribution within binding motifs, as shown in figures Fig 2.15, Fig 2.13 and Fig 2.14. As we could see from Figure 2.13 and Fig 2.14, the probability of either L or N alone is relatively lower than P. But when considering L+N, the probability is comparable to P as shown in Fig 2.15.

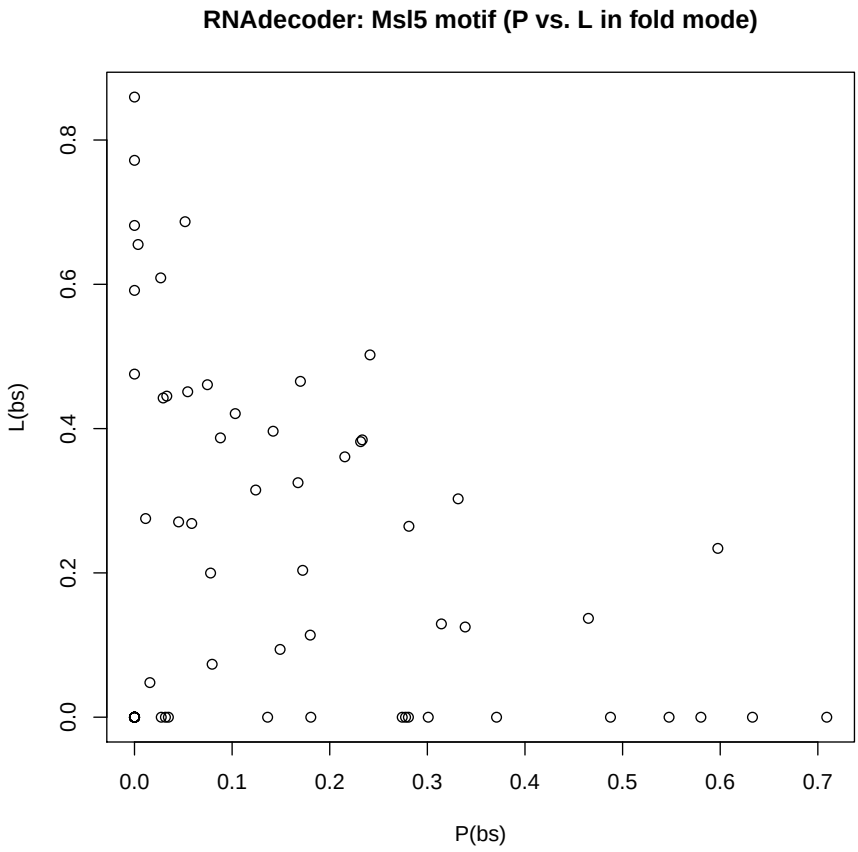

Figure 2.13: **RNAdecoder: Msl5 motif (P vs. L) on absolute scale.** Scatter plot of structural annotation probabilities for L vs P, in RNAdecoder case. Each dot in the plot represents a Msl5 protein binding motif. So the coordinates of any dot are (P, L), where $P = P^b_{avg}$ = (sum of likelihoods for binding sites labeled with P)/(length of the binding motif), and similarly $L = L^b_{avg}$. The structural annotations are generated by RNA-decoder for each alignment position.

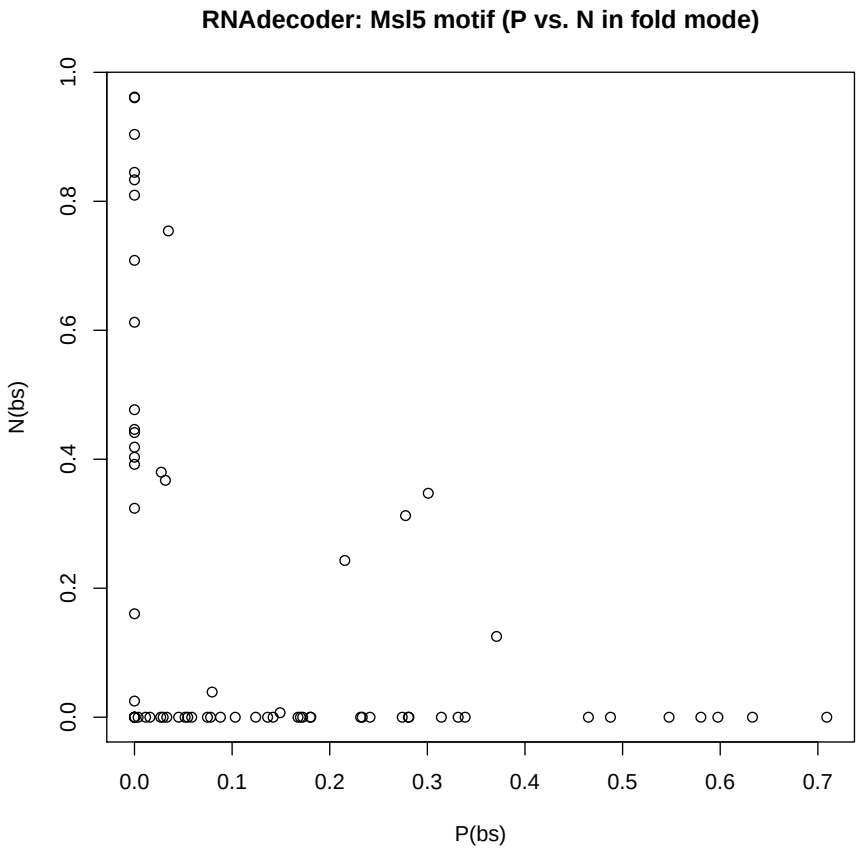

Figure 2.14: **RNAdecoder: Msl5 motif (P vs. N) on absolute scale.** Scatter plot of structural annotation probabilities for N vs P, in RNAdecoder case. Each dot in the plot represents a Msl5 protein binding motif. So the coordinates of any dot are (P, N), where $P = P^b_{avg} =$ (sum of likelihoods for binding sites labeled with P)/(length of the binding motif), and similarly $N = N^b_{avg}$. The structural annotations are generated by RNA-decoder for each alignment position.

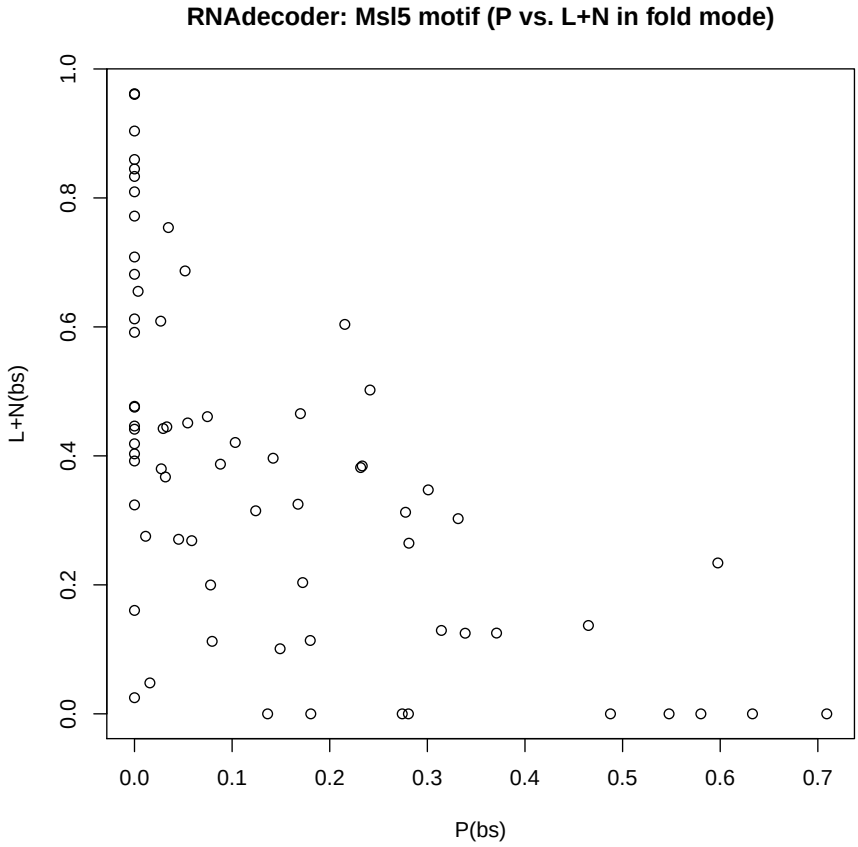

Figure 2.15: **RNAdecoder: Msl5 motif (P vs. L+N) on absolute scale.** Scatter plot of structural annotation probabilities for L+N vs P, in RNAdecoder case. Each dot in the plot represents a Msl5 protein binding motif. So the coordinates of any dot are $(P, L+N)$, where $P = P^b_{avg} =$ (sum of likelihoods for binding sites labeled with P)/(length of the binding motif), and similarly $L+N = L^b_{avg} + N^b_{avg}$. The structural annotations are generated by RNA-decoder for each alignment position.

2.1.4 Visualization of the alignment, structural annotation and accessibility

By all means, it would be very useful if we could visualize the structural context around the binding motif on 3'UTR region. This section presents our progress on this task.

Motif logo

Motif-wise accessibility with regarding to NLP distribution could be shown independently as motif logo in Figure 2.16. However, such figure does not give much useful information as "N" is the dominant case for every position. It would be better to put the binding sites motif back to the sequence context to get an idea of how different it is compared to other non-binding region.

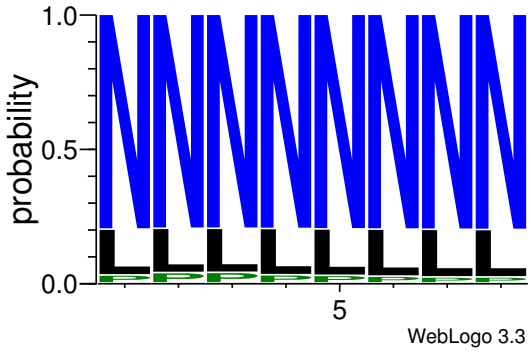

Figure 2.16: This is the motif logo for Pumilio (UGUAHAUA) protein of Fly, predicted by RNAalifold.

More advanced visualization

In order to visualize predictions from "fold" and "scan" modes for individual binding sites in individual UTRs, a combination of berrylogo [29] and our lab's R4RNA package is built (a perl script called ccPBS.pl).

A berryLogo is a seqLogo alternative developed by Charles C. Berry for [29]. Instead of "information content", the y-axis is the log relative frequency with respect to the background frequency, generated originally from the gc_content parameter. Based on the original version I re-coded my version to visualize predictions from "fold". Currently I have not used the background frequency yet.

R4RNA is used to draw the pairing probability (1-accessibility) from "scan". For berrylogo.r, the script uses ggplot2 which is using lattice and based on 'grid' graphics subsystem, while the original R programs including R4RNA use 'base' graphics system. Noted that it is fairly difficult to combine figures from different systems.

Moreover, in order to connect the plots codes from R and computation codes from Perl more efficiently, a perl module Statistics::R is set up so that one can not only pass R commands as inline script in Perl but also directly pass and retrieve scalars or arrays variables between these two languages.

Given the transcripts ID, protein name and program name (like RNAdecoder), ccPBS.pl could generate a pdf with plot for "scan" on top and plot for "fold" at bottom. Two examples are shown below in Figure 2.17 and Figure 2.18

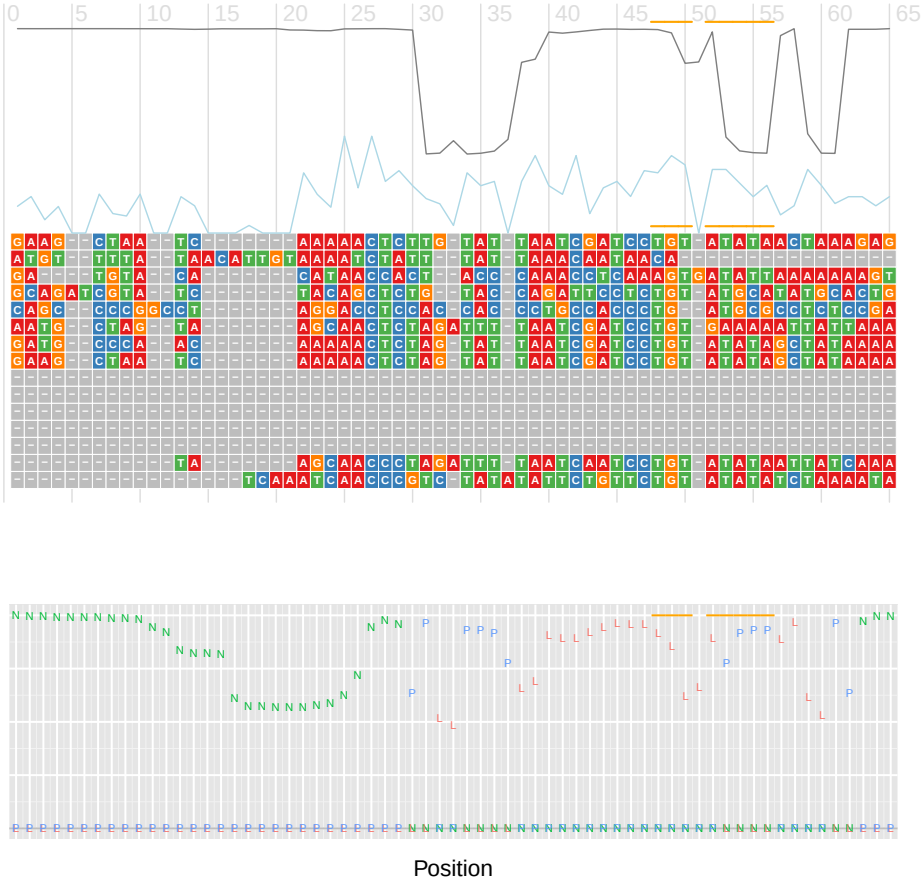

Figure 2.17: This is a figure based on the prediction results from RNAdecoder. RNA: the 3'UTR of Fly FBtr0078004 transcript. Protein: Pumilio. 1). The plot on top is for "scan". Above the alignment, the grey line shows the pairing probability for each position. The binding sites regions are highlighted in yellow (on the grey line). The blue line shows the sequence conservation of the alignment. 2). The plot at bottom is for "fold". For each position, the N,L,P classifications are plotted according to their prior probabilities. Similarly, binding sites regions are highlighted as yellow fragments.

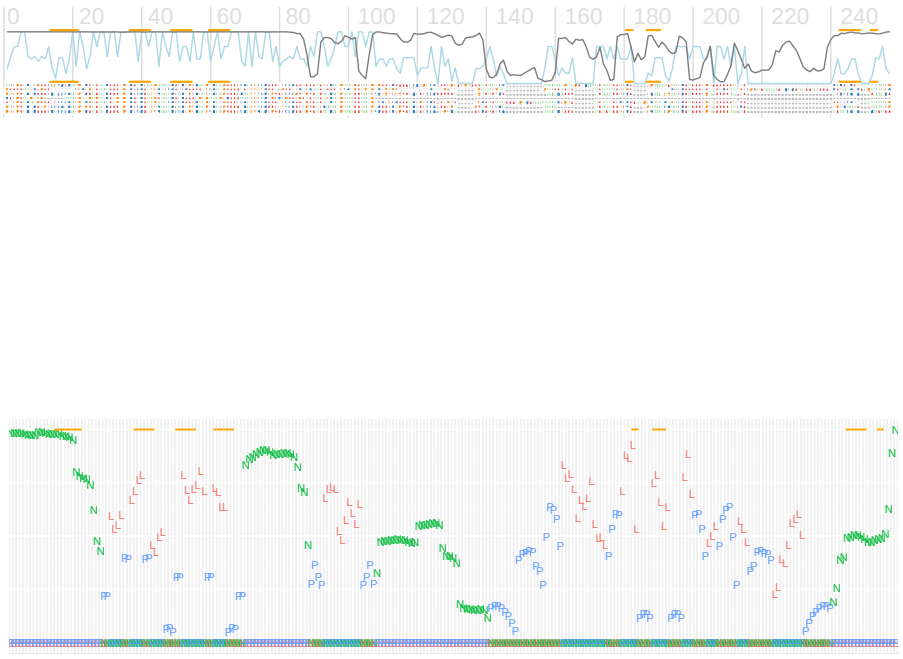

Figure 2.18: This is another example based on the prediction results from RNAdecoder. RNA: the 3'UTR of Yeast YAL001C transcript. Protein: Khd1. It shows the case of a longer alignment with multiple binding motif hits.

Probably we could cluster the UTRs/binding sites according to their accessibility so that we may be able to visualize and see if there is any pattern for the binding sites. For each potential binding motif (i.e. a sequence matching), we could get its pairing probability curve centered at the binding motif with flanking regions. And I found in several ENCODE papers, they have introduced how to cluster curves for features around TSS, like the paper [30]. Those features are mostly CHIP-seq signals or histone modification signals. In our case, we can easily change these to base pairing probabilities generated from those programs. And we could observe if there is any pattern surrounding binding motif by clustering these pairing signals.

2.2 Undergoing work

2.2.1 Realign the original alignment from UCSC

Structurally realign

As shown in the previous section, the performances of the three programs we are using are highly depending on the quality of the input alignments. We propose to realign the raw 3'UTR alignments "cut" from UCSC whole genome alignments so that we would get more convincing results from those comparative methods. The original plan is to structurally realign them then compute the new accessibility.

There are programs like mlocarna and simulfold which can do realign and fold simultaneously. According to my labmate, Alborz's unpublished results, both programs are very slow, especially when the sequence length>1000; mlocarna is a bit faster in general. So I used mlocarna to realign Fly alignments at first place.

However, when I tested on a small set of alignments of Fly 3'UTRs, the program took hours to get the results. The main reason might be that the core part is a program called locarna which is used for pair-wise alignment, and mlocarna is a large perl script calls locarna for progressive alignment. Besides, the original authors of mlocarna suggests to use sequence-based realignment in their paper [31]. In that paper, they use mlocarna for sRNA realign and MAFFT for mRNA realign. So I take their suggestions here in my project.

Sequence-based realign

For sequence-based realignment, since Human species has the longest 3'UTR sequences and the largest alignment (44way) among all three species, I start

from Human to build the computational pipeline.

Pick the best set of species from UCSC MSAs

Binding proteins' homologs In order to control the quality of the MSA, the first thing is to make sure there are corresponding RBPs exist in all the species of the MSA as shown in Figure 2.19. So for each of those human RBPs listed in Table 2.2, we have searched for the homologs of its gene in NCBI.

Table 2.2: Human RBPs in our dataset.

RBPs	**Database ID**	**Database Info**
PTB	AAC99798 (NCBI)	polypyrimidine tract binding
HuR	AAB41913	Elav-like family; binds specifically to AU rich elements (AREs) in 3'UTRs
Pum1	NP_001018494	Puf family

For PTB, the human PTBP1 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, C.elegans, A.thaliana, and rice.

For HuR, since it is a member of Elav-like family, the human ELAVL1 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and C.elegans.

For Pum1, its human gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, and zebrafish. Moreover, this gene encodes a member of the PUF family, evolutionarily conserved RNA-binding proteins related to the Pumilio proteins of Drosophila and the fem-3 mRNA binding factor proteins of C. elegans.

For all the species in the human 44way MSA, the most distant one (outgroup) from human is Petromyzon marinus. So we can conclude that it is safe to include all the 44 species in the MSA, from proteins' point of view.

We have also searched the three proteins in Treefam (using their NCBI ID to search against the DB). The results are basically the same. Besides, Treefam provides the orthologs family trees for each of the protein gene (not found it very useful since the species coverage is small), which are also

consistent.

Sequence similarity The phylogenetics tree for hg18 44way vertebrate MSA is downloaded from UCSC. This tree as shown in Figure 2.19 is used as the starting point. And the total tree length is: 12.117.

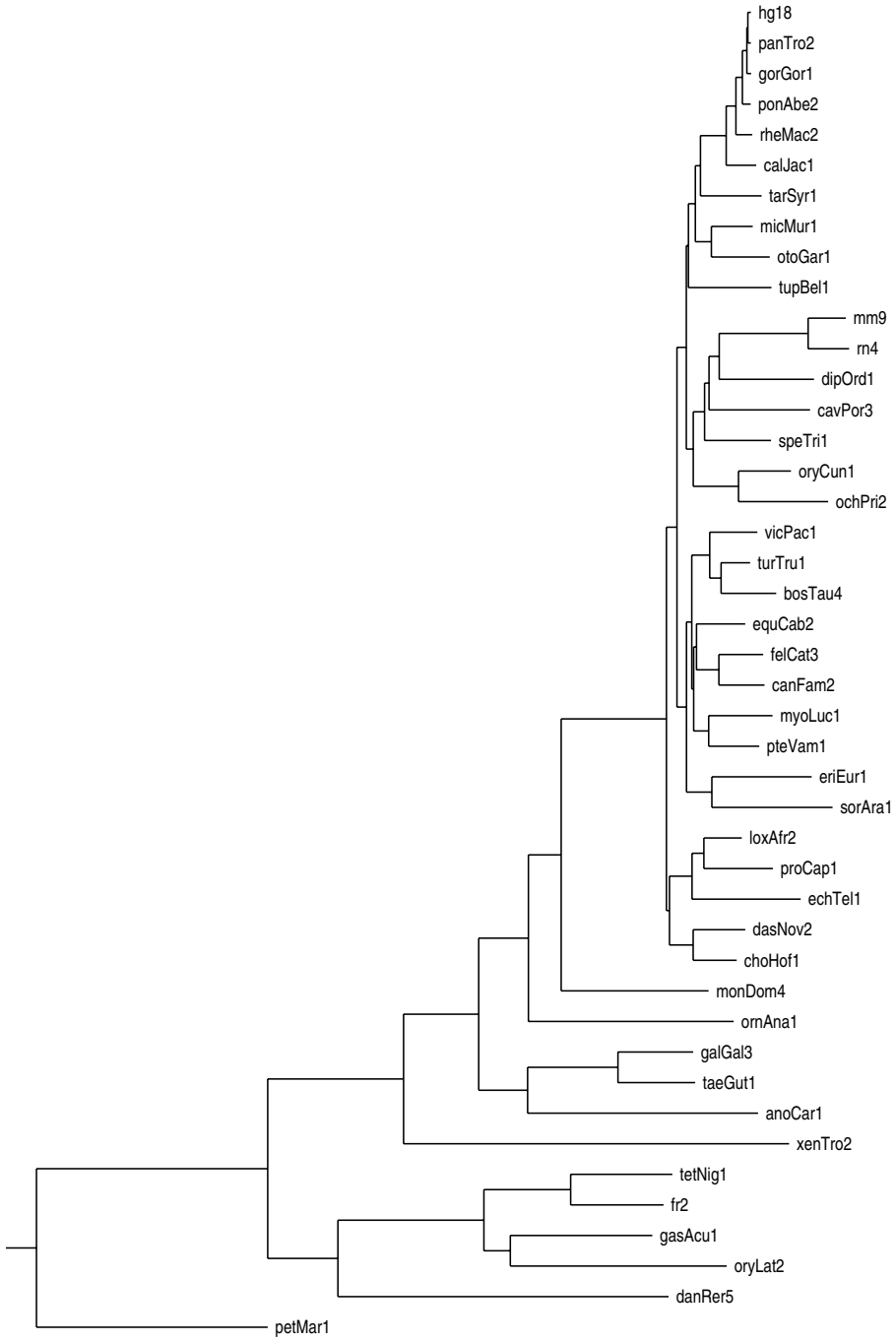

Figure 2.19: Hg18 44way alignment: total tree. This figure is made based on UCSC whole genome MSAs.

Next, we would like to get an idea of the distribution of tree length for human 3'UTR MSAs from UCSC. For any given MSA, we pick the overlapped species between it and the 44way tree. Then based on the 44way tree, the missing species are trimmed. Figure 2.20 shows an example of the 3'UTR MSAs for transcript ENST00000334314. Its total tree length is: 4.847.

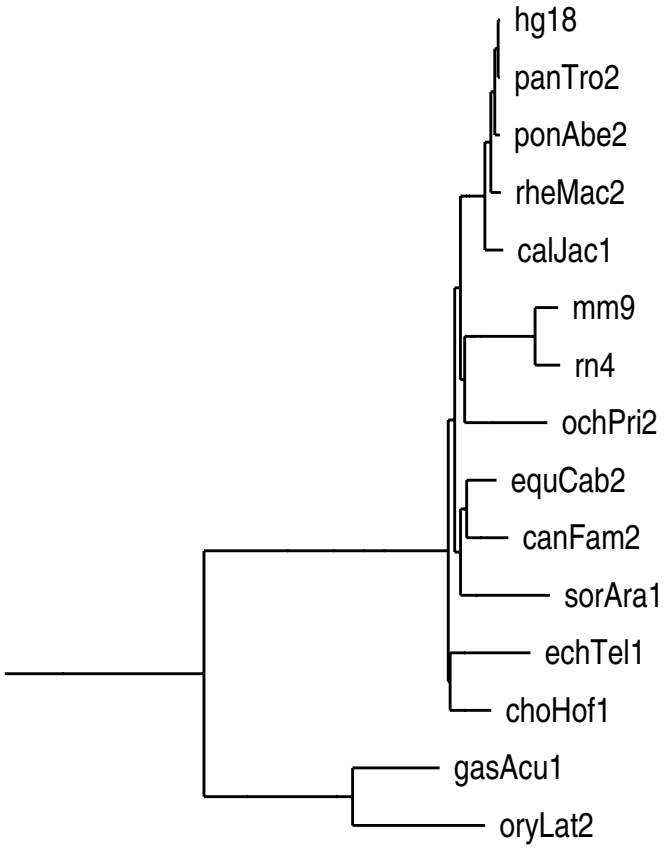

Figure 2.20: ENST00000334314 alignment: 15way tree. This tree is pruned based on the 44way tree. Since for ENST00000334314 MSA, only 15 species out of 44 appear, the rest part of the 44way tree is pruned.

A randomly sampled subset (100 samples per chromosome) of all human 3'UTR MSAs is used to estimate the total tree length distribution of human 3'UTR alignments from UCSC. The results are shown as Figure 2.21. In

summary, the mean length is 7.864 and median is 8.235.

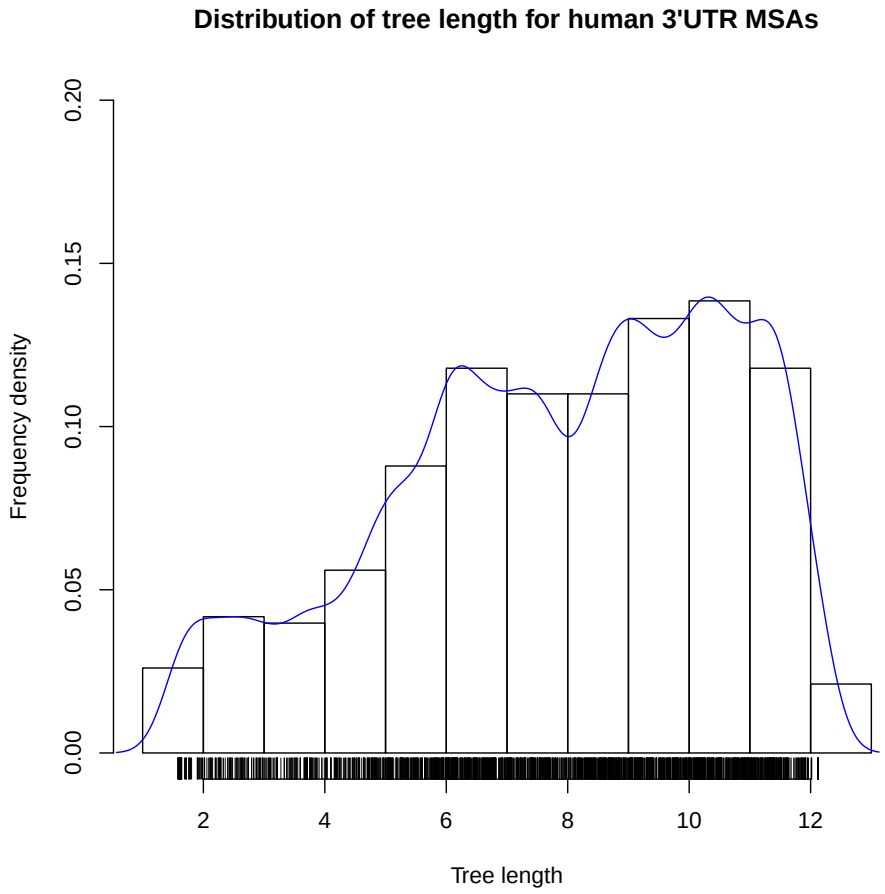

Figure 2.21: Distribution of total tree length for human 3'UTR MSAs. The normalized frequency densities (y-xais) of different total tree lengths (x-xais) are shown. The bars at the bottom indicate the actual tree lengths.

Thus we would aim for a total tree length between 4 and 10 to get an appropriate MSA with regarding to evolutionary distance (sequence similarity).

Actually, the information of hg18 44way alignment is also listed as table at 1 and 2 . From the "% of hg18 matched" column, we could see the matched percentage (i.e. alignment coverage) between human genome and

each of the other 43 species. The rows have been sorted according to the percentage, as shown in Figure 2.22.

Sheet2

ucsc db name	common name	total size	% masked	% of hg18 matched (chainLink table)
hg18	Human	3107677273	% 48.85	100.00%
panTro2	Chimp	3350447512	% 48.90	94.888%
ponAbe2	Orangutan	3446771396	% 50.89	92.892%
rheMac2	Rhesus	2864106071	% 48.28	85.552%
calJac1	Marmoset	3029401840	% 47.50	78.351%
gorGor1	Gorilla	2323645895	% 47.27	61.731%
equCab2	Horse	2484532062	% 40.97	57.162%
canFam2	Dog	2531673953	% 40.63	52.915%
turTru1	Dolphin	2519048486	% 43.86	48.537%
tarSyr1	Tarsier	3179905132	% 41.75	47.999%
bosTau4	Cow	2917974530	% 46.89	46.689%
micMur1	Mouse lemur	2902270736	% 37.31	46.445%
pteVam1	Megabat	1996076410	% 31.75	45.502%
otoGar1	Bushbaby	3420058864	% 34.89	44.638%
cavPor3	Guinea Pig	2723219641	% 27.53	43.971%
vicPac1	Alpaca	2962253608	% 32.24	39.531%
tupBel1	TreeShrew	3660774957	% 20.43	37.348%
felCat3	Cat	4045535322	% 37.66	35.888%
speTri1	Squirrel	3488768592	% 26.90	35.828%
loxAfr2	Elephant	4170414852	% 44.13	35.204%
mm9	Mouse	2725765481	% 44.09	35.201%
choHof1	Sloth	2458927620	% 35.67	34.463%
oryCun1	Rabbit	3464410039	% 37.44	34.015%
dasNov2	Armadillo	4813823562	% 37.23	33.663%
myoLuc1	Microbat	2850051559	% 24.90	33.044%
rn4	Rat	2834127293	% 44.29	32.893%
proCap1	Rock hyrax	2985258999	% 28.11	30.935%
ochPri2	Pika	3445784354	% 11.44	27.974%
dipOrd1	Kangaroo rat	2158502098	% 28.28	27.282%
echTel1	Tenrec	3823724728	% 13.84	23.645%
sorAra1	Shrew	2936119008	% 39.07	20.056%
eriEur1	Hedgehog	3367787358	% 50.89	19.622%
monDom4	Opossum	3605614649	% 55.69	12.385%
ornAna1	Platypus	1996811212	% 47.89	7.870%
anoCar1	Lizard	1781602899	% 42.56	4.774%
galGal3	Chicken	1100480441	% 9.85	3.589%
taeGut1	Zebra finch	1233186341	% 20.42	3.503%
xenTro2	X. tropicalis	1513925492	% 19.65	2.623%
danRer5	Zebrafish	1440582308	% 49.53	2.565%
tetNig1	Tetraodon	402240326	% 4.34	2.001%
gasAcu1	Stickleback	463354448	% 2.58	1.923%
oryLat2	Medaka	869000216	% 33.09	1.829%
fr2	Fugu	400525790	% 19.01	1.766%
petMar1	Lamprey	1027258967	% 40.67	1.251%

Page 9

Figure 2.22: Hg18 44way alignment: matched percentage. This table is built based on the information from 1 and 2. Note for the last column, the percentage is the alignment coverage rather than the evolutionary distance. Four widely studied model organisms (Human, mouse, rat, zebrafish) are highlighted.

So we can see that the rows below Opossum are all species with less than 10% coverage to human. They are basically species which are the most distant from human. If we remove these species, the total tree length reduces to: 6.463. Even with the model organism zebrafish added, we have the tree length: 7.349.

However, as described above, for all the sampled human 3'UTRs the median tree length is 8.235. This means that the majority of the 3'UTRs have a relatively large coverage of the 44 species, including distant ones. So removing all of these distant species seems not to be a good idea.

Another way is to use K-means cluster algorithm to get the "rational" k representative species out of Figure 2.22. A program has been implemented to parse table data using Algorithm::KMeans module. But still, the percentage values here could only be used to measure the distance to human. Even when two species have very close percentages to human, like elephant and mouse, it does not mean they should be clustered together.

Thus, the best way to control the quality for each 3'UTR MSA is really to do cluster on aligned sequences themselves, based on sequence similarity. A program called *usearch* which is developed by the same author of MUSCLE is exactly for this job. That's what I plan to do in the next step.

But here, in order to get a general trimmed tree from the 44way whole genome tree, direct tree manipulating is needed. As described above, for all the three RBPs, all the 44 species are supposed to have homologs. So it is safe to select any one of the 44 species. Besides, there are some obvious clusters from Figure 2.19. We could sample from each of the clusters according to the distances of its members. And this sampling may generate different trees (but their distance should be similar). At the moment, this process is done in a heuristic way: first generate the pair-wise distance matrix (attached), then fix hg18 and select tupBel1 which is the most distant one in that cluster, use the distance (D) between hg18 and tupBel1 as a criteria to select the following ones. It means the third one to join {hg18, tupBel1} is the one which is the most close to hg18 and tupBel1 among those nodes having distances to both hg18 and tupBel1 greater than D. And so is the fourth one.. In practice, this can be done using monophyly test which is implemented in Bio::Tree::Tree: given a set of nodes and a outgroup node, use is_monophyletic() to test whether the common ancestor for the members of the internal_nodes group is more recent than the common ancestor that any of them share with the outgroup node. So for each candidate node, the script loops over each of the rest nodes (excluding the candidate and those already selected) as the outgroup node. Based on this procedure, we could get the trimmed tree as Figure 2.23. And the total tree length is:

9.475. Note that setting the heuristic distance differently at the beginning would generate very different trees. Here, we basically reduce the number of nodes by half and keep almost the same diversity. So we would expect a less conserved alignment generated based on the trimmed tree.

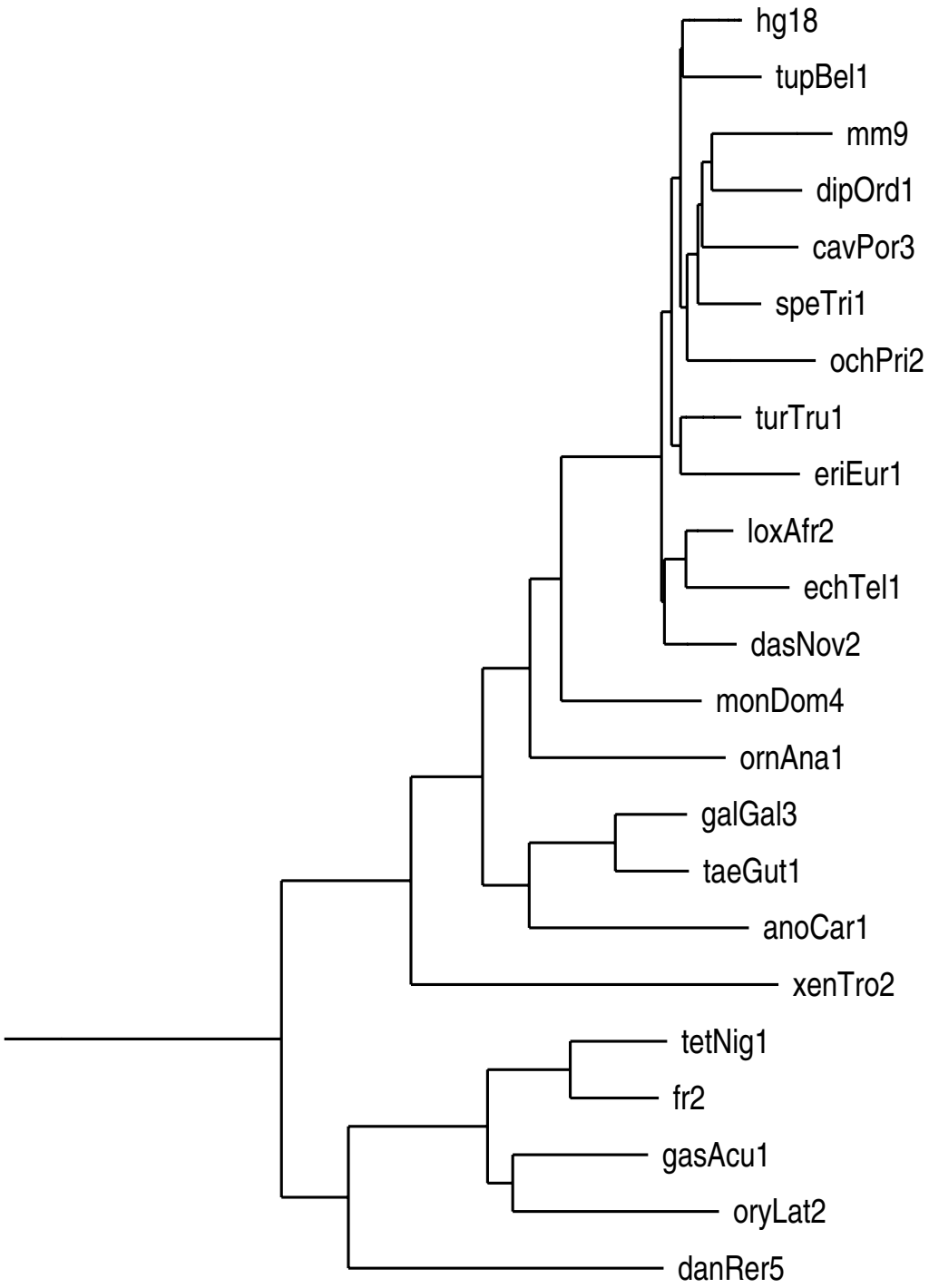

Figure 2.23: Hg18 23way alignment: trimmed tree. This tree is built from the 44way tree from UCSC with quality control.

Actually if we could have more RBPs with diverse homologs ranges in the future, the best way might be to make a set of RBP-specific pruned trees.

Realign MSAs for human 3' UTRs After getting the new tree, we would trim the MSA based on the tree. Given an MSA of some 3'UTR, a pipeline is built to get its overlapped entities with a designated tree, output the overlapped MSA and overlapped entities' tree (length, newick file, eps figure). As a result, any trimmed MSA with total tree length smaller than 4 or greater than 10 is removed.

As suggested in [31], MAFFT is chosen to do the realignment based on sequence conservation. More specifically, as suggested by other labmates, MAFFT-einsi which is one of the three options in MAFFT (the other two are ginsi (for global aln) and linsi (for local aln). einsi is like the compromise between local and global) is used here.

The table below Table 2.3 shows the comparison between trimmed MSAs and raw UCSC MSAs re-alignments. The value in each cell is the average value across all 3'UTRs. It seems the trimmed 23way MSAs have worse conservation and more gaps. However, it highly depends on how to select the trimmed tree nodes. According to the method described above, the tree selected is more diverse compared to the raw one rather than conserve. We could see from the Figure 2.24. The distribution of trimmed tree length is more concentrated in the 4 to 10 region. But the total tree length is still as high as 7.077 (Median, compared to 8.235 for the original), which indicates the variation is increasing.

	Conservation		Gappiness	
	UCSC	**Realigned**	UCSC	**Realigned**
23way	0.397	**0.411**	0.521	**0.582**
33way	0.625	**0.603**	0.335	**0.397**
raw	0.528	**0.525**	0.528	**0.443**

Table 2.3: The alignment quality scores for the **trimmed** UCSC alignments and **raw** UCSC alignments (untrimmed) of *Human*.

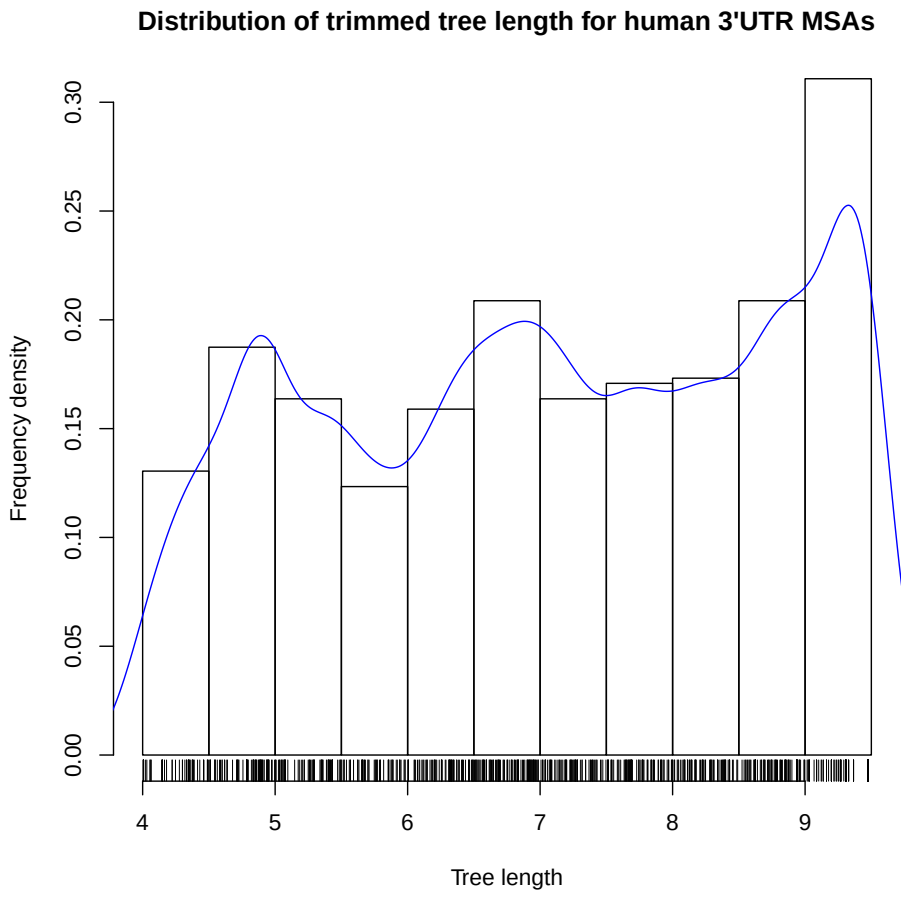

Figure 2.24: Distribution of trimmed tree length for human 3'UTR MSAs. The figure is similar to that Figure 2.21. However, MSAs are trimmed according to the new tree and those with total tree length beyond 4 to 10 region are removed.

As a contrast, if we trim the tree in a conserved manner, we would get good alignment but bad variation. For example, like said in Figure 2.22, if we remove those species with less than 10% coverage, we get a 33way tree with length 6.463.

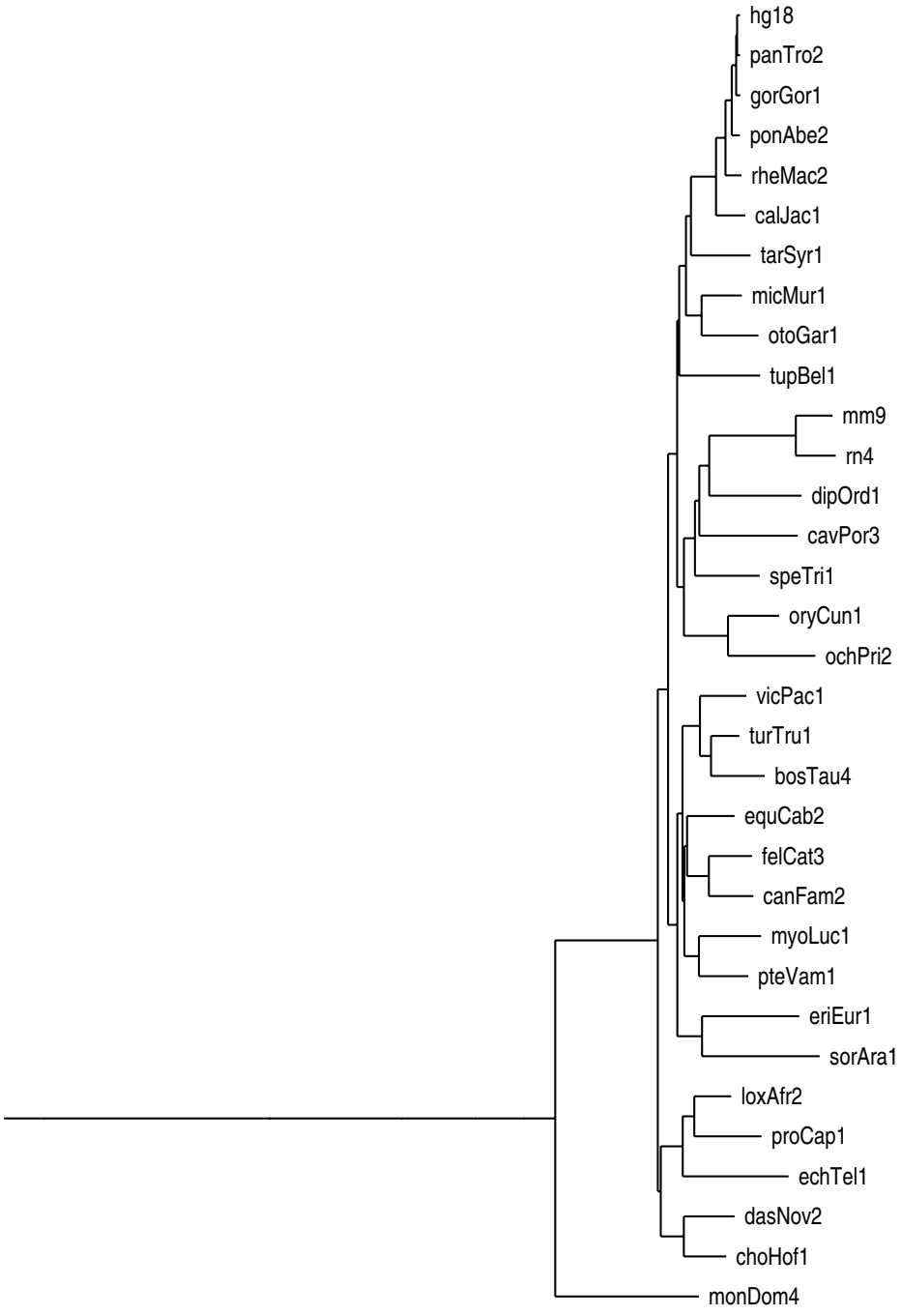

Figure 2.25: Hg18 33way alignment: trimmed tree. This tree is built from the 44way tree from UCSC with removing low coverage species.

Since the final goal is on RNA-secondary structure level, probably we need criteria other than conservation and gappiness which obviously bias to conserved alignments, such as Mutual Information or Transat scores or those methods mentioned at the beginning of this section (mlocarna and simulfold).

2.2.2 Incoperating the RIP-Chip enrichment data

Briefly, the in vivo RIP-Chip data from [11] have been collected for all the proteins (human, yeast and fly) on Table 2.1, so that for most of the transcripts we now have their enrichment values in the copurifying experiments with proteins. These transcripts are then classified into "bound", "unbound" and "unsure" sets using similar approach as [8] did.

Computing the accessibilities based on "bound" and "unbound" sets are much more reasonable than simply considering the matched motif sequence hits. This is because of the obvious fact that sequence matching has a very high false positive rate for protein binding motifs though the proteins are known to be sequence-specific binding. We could further compare the difference between false positive matchings and true positive matchings, which would potentially improve binding motif discovery.

By far, I have performed a pilot study on Msl5 and Khd1 proteins in yeast using "bound" set and observed significant difference (p-values smaller than 0.05) on accessibilities between motif hits and the whole 3'UTR (background) region; also for "unbound" set there is no significant difference. However, I still need to test this for all the other proteins before any conclusion could be drawn. Also, the results are based on raw alignment from UCSC without realignment. The alignments may need to be tuned and the results may differ.

Bibliography

[1] Shivendra Kishore and Lukasz Jaskiewicz1 et al. A quantitative analysis of clip methods for identifying binding sites of rna-binding proteins. *Nature methods*, 8:559–564, 2011.

[2] Chaolin Zhang and Robert B. Darnell. Mapping in vivo protein-rna interactions at single-nucleotide resolution from hits-clip data. *Nature Biotechnology*, 29:604–614, 2012.

[3] Anna-Carina Jungkamp and Marlon Stoeckius et al. In vivo and transcriptome-wide identification of rna binding protein target sites. *Cell*, 44:828–840, 2011.

[4] Gerd Anders, Sebastian D. Mackowiak, Marvin Jens, Jonas Maaskola, Andreas Kuntzagk, Nikolaus Rajewsky, Markus Landthaler, and Christoph Dieterich. dorina: a database of rna interactions in post-transcriptional regulation. *Nucleic Acids Research*, 40(D1):D180–D186, 2012.

[5] Mohsen Khorshid, Christoph Rodak, and Mihaela Zavolan. Clipz: a database and analysis environment for experimentally determined binding sites of rna-binding proteins. *Nucleic Acids Research*, 39(suppl 1):D245–D252, 2011.

[6] Christoph Dieterich1 and Peter F. Stadler. Computational biology of rna interactions. *Wiley Interdisciplinary Reviews: RNA*, 4:107–120, 2012.

[7] Jonathan J. Ellis, Mark Broom, and Susan Jones. Proteinrna interactions: Structural analysis and functional classes. *Proteins: Structure, Function, and Bioinformatics*, 66(4):903–911, 2007.

[8] Xiao Li, Gerald Quon, Howard D. Lipshitz, and Quaid Morris. Predicting in vivo binding sites of rna-binding proteins using mrna secondary structure. *RNA*, 16(6):1096–1107, 2010.

[9] Stephan H. Bernhart, Ivo L. Hofacker, and Peter F. Stadler. Local rna base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, 2006.

[10] Stephan Bernhart, Ivo Hofacker, Sebastian Will, Andreas Gruber, and Peter Stadler. Rnaalifold: improved consensus structure prediction for rna alignments. *BMC Bioinformatics*, 9(1):474, 2008.

[11] Daniel J Hogan, Daniel P Riordan, Andr P Gerber, Daniel Herschlag, and Patrick O Brown. Diverse rna-binding proteins interact with functionally related sets of rnas, suggesting an extensive regulatory system. *PLoS Biology*, 6(10):e255, 10 2008.

[12] Xiaowei Sylvia Chen and Chris M. Brown. Computational identification of new structured cis-regulatory elements in the 3'-untranslated region of human protein coding genes. *Nucleic Acids Research*, 40:261–268, 2012.

[13] Tomasz Puton and Lukasz Kozlowski et al. Computational methods for prediction of proteinrna interactions. *Journal of Structural Biology*, 179:261–268, 2012.

[14] Usha Muppirala, Vasant Honavar, and Drena Dobbs. Predicting rna-protein interactions using only sequence information. *BMC Bioinformatics*, 12(1):489, 2011.

[15] Sita J. Lange, Daniel Maticzka, Mathias Mhl, Joshua N. Gagnon, Chris M. Brown, and Rolf Backofen. Global or local? predicting secondary structure and accessibility in mrnas. *Nucleic Acids Research*, 2012.

[16] Julian Knig, Kathi Zarnack, Nicholas M. Luscombe, and Jernej Ule. Proteinrna interactions: new genomic technologies and perspectives. *Nature Review Genetics*, 13:77–83, 2012.

[17] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 2012.

[18] Vera Pancaldi and Jrg Bhler. In silico characterization and prediction of global proteinmrna interactions in yeast. *Nucleic Acids Research*, 2011.

[19] Steven R. Morgan. Evidence for kinetic effects in the folding of large rna molecules. *The Journal of chemical physics*, 105(1):7152–7157, 1996.

[20] Irmtraud Meyer and Istvan Miklos. Co-transcriptional folding is encoded within rna genes. *BMC Molecular Biology*, 5(1):10, 2004.

[21] Bjarne Knudsen and Jotun Hein. Pfold: Rna secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, 2003.

[22] Jakob Skou Pedersen, Irmtraud Margret Meyer, Roald Forsberg, Peter Simmonds, and Jotun Hein. A comparative method for finding and folding rna secondary structures within protein-coding regions. *Nucleic Acids Research*, 32(16):4925–4936, 2004.

[23] Irmtraud M. Meyer. A practical guide to the art of rna gene prediction. *Briefings in Bioinformatics*, 8(6):396–414, 2007.

[24] Pauline A. Fujita, Brooke Rhead, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Melissa S. Cline, Mary Goldman, Galt P. Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R. Dreszer, Belinda M. Giardine, Rachel A. Harte, Jennifer Hillman-Jackson, Fan Hsu, Vanessa Kirkup, Robert M. Kuhn, Katrina Learned, Chin H. Li, Laurence R. Meyer, Andy Pohl, Brian J. Raney, Kate R. Rosenbloom, Kayla E. Smith, David Haussler, and W. James Kent. The ucsc genome browser database: update 2011. *Nucleic Acids Research*, 2010.

[25] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas Khri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Bert Overduin, Bethan Pritchard, Harpreet Singh Riat, Daniel Rios, Graham R. S. Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Jana Vandrovcova, Albert J. Vilella, Simon White, Steven P. Wilder, Amonida Zadissa, Jorge Zamora, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xos M. Fernndez-Suarez, Javier Herrero, Tim J. P. Hubbard, Anne Parker, Glenn Proctor, Jan Vogel, and Stephen M. J. Searle. Ensembl 2011. *Nucleic Acids Research*, 39(suppl 1):D800–D806, 2011.

[26] Jason E. Stajich, David Block, Kris Boulez, Steven E. Brenner, Stephen A. Chervitz, Chris Dagdigian, Georg Fuellen, James G.R. Gilbert, Ian Korf, Hilmar Lapp, Heikki Lehvslaiho, Chad Matsalla,

Chris J. Mungall, Brian I. Osborne, Matthew R. Pocock, Peter Schattner, Martin Senger, Lincoln D. Stein, Elia Stupka, Mark D. Wilkinson, and Ewan Birney. The bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618, 2002.

[27] Mara Flor Garca-Mayoral, David Hollingworth, Laura Masino, Irene Daz-Moreno, Geoff Kelly, Roberto Gherzi, Chu-Fang Chou, Ching-Yi Chen, and Andres Ramos. The structure of the c-terminal {KH} domains of {KSRP} reveals a noncanonical motif important for mrna degradation. *Structure*, 15(4):485 – 498, 2007.

[28] Limor Leibovich and Zohar Yakhini. Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Research*, 2012.

[29] Charles Berry and Sridhar Hannenhalli et al. Selection of target sites for mobile dna integration in the human genome. *PLoS Computational Biology*, 2, 2006.

[30] Anshul Kundaje, Sofia Kyriazopoulou-Panagiotopoulou, Max Libbrecht, Cheryl L. Smith, Debasish Raha, Elliott E. Winters, Steven M. Johnson, Michael Snyder, Serafim Batzoglou, and Arend Sidow. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research*, 22(9):1735–1747, 2012.

[31] Andreas S. Richter and Rolf Backofen. Accessibility and conservation: General features of bacterial small rnamrna interactions? *RNA Biology*, 9:954–965, 2012.