SYSTEMATICALLY LEARNING OF INTERNAL RIBOSOME ENTRY SITE AND PREDICTION BY MACHINE LEARNING

by

Junhui Wang

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biological Sciences

West Lafayette, Indiana

May 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Michael Gribskov, Chair
Department of Biological Sciences
Dr. Richard J. Kuhn
Department of Biological Sciences
Dr. Daisuke Kihara
Department of Biological Sciences
Dr. Barbara L. Golden
Department of Biochemistry

Approved by:

Dr. Daniel Suter
Head of the Graduate Program

ACKNOWLEDGMENTS

I want to thank my Ph.D. advisor Michael Gribskov, for his encouragement, patience, and advices over the years. I want to thank him for the communication, feedback and unlimited support to my dissertation preparation. I want to thank my committee members: Daisuke Kihara, Babrara Golden, and Richard Kuhn for their time, comments and support all the time. Most of all, I owe my deepest gratitude to my wife Mingyu and my son Will. Without your endless love, I cannot earnf this achievement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IRES	Internal ribosome entry sites
eIF	Eukaryotic initiation factor
UTR	Untranslated region
ITAF	IRES trans-acting factor
ORF	Open reading frame
PV	Poliovirus
EMCV	Encephalomyocarditis virus
HCV	Hepatitis C virus
CrPV	Cricket paralysis virus
TSV	Taura syndrome virus
GBDT	Gradient boosting decision tree
XGBoost	Extreme Gradient Boosting

ABSTRACT

Author: Wang, Junhui. PhD
Institution: Purdue University
Degree Received: May 2019
Title: Systematically Learning of Internal Ribosome Entry Site and Prediction by Machine Learning
Committee Chair: Michael Gribskov

Internal ribosome entry sites (IRES) are segments of the mRNA found in untranslated regions, which can recruit the ribosome and initiate translation independently of the more widely used 5' cap dependent translation initiation mechanism. IRES play an important role in conditions where has been 5' cap dependent translation initiation blocked or repressed. They have been found to play important roles in viral infection, cellular apoptosis, and response to other external stimuli. It has been suggested that about 10% of mRNAs, both viral and cellular, can utilize IRES. But due to the limitations of IRES bicistronic assay, which is a gold standard for identifying IRES, relatively few IRES have been definitively described and functionally validated compared to the potential overall population. Viral and cellular IRES may be mechanistically different, but this is difficult to analyze because the mechanistic differences are still not very clearly defined. Identifying additional IRES is an important step towards better understanding IRES mechanisms. Development of a new bioinformatics tool that can accurately predict IRES from sequence would be a significant step forward in identifying IRES-based regulation, and in elucidating IRES mechanism. This dissertation systematically studies the features which can distinguish IRES from nonIRES sequences. Sequence features such as kmer words, and structural features such as predicted MFE of folding, Q_{MFE}, and sequence/structure triplets are evaluated as possible discriminative features. Those potential features incorporated into an IRES classifier based on XGBboost, a machine learning model, to classify novel sequences as belong to IRES or nonIRES groups. The XGBoost model performs better than previous predictors, with higher accuracy and lower computational time. The number of features in the model has been greatly reduced, compared to previous predictors, by adding global kmer and structural features. The trained XGBoost model has been implemented as the first high-throughput bioinformatics tool for IRES prediction, IRESpy. This website provides a public tool for all IRES researchers and can be used in other genomics applications such as gene annotation and analysis of differential gene expression.

CHAPTER 1. INTRODUCTION OF INTERNAL RIBOSOME ENTRY SITE

The central dogma of molecular biology contains three important steps, replication, transcription and translation. Information passes from DNA to RNA, and then to protein. Most of the biological molecular functions are executed by proteins. Translation initiation is therefore an important regulatory step. There are two main ways to initiate protein translation, the canonical 5' cap dependent method, and initiation at internal sites called internal ribosome entry sites (IRES) method. In contrast to the 5' cap dependent mechanism, the mechanism of initiation at IRES is not completely clear. This chapter discusses the definition of internal ribosome entry sites, the characteristics of IRES, possible initiation mechanisms, the importance and history of computationally prediction of IRES, the importance of learning them, IRES databases and methods for identifying novel IRES.

1.1 What are internal ribosome entry sites?

Internal ribosome entry sites (IRES) are a segment of mRNA in the untranslated region (UTR) of mRNAs which can directly recruit the ribosome and initiate translation without the help of a 5' cap. Typically, the initiation of translation is a key part of protein expression, and requires a specific cap structure, m7G(5')ppp(5')N, to be placed at the 5' end of the mRNA. The cap structure is recognized by eukaryotic initiation factor 4F (eIF4F), which is composed of eIF4A, eIF4E, and eIF4G. The 43S complex is formed by the 40S ribosomal complex (ribosome small subunit), factor eIF3, and another complex composed of eIF1, eIF1A, eIF2, eIF5B, and the initiator Met-tRNAi. The 43S complex is recruited by eIF4F, and scans the mRNA in the 5' to 3' direction until it meets the first initiation codon (normally AUG). The initiation factors of the 40S ribosomal

complex are then unloaded, and the 60S ribosomal complex (large subunit) added. The AUG start codon occupies a specific site on the ribosome, by base-pairing to the Met-tRNAi, and protein translation is initiated (Hinnebusch, 2014).

Unlike the canonical translation initiation mechanism, IRES do not need the 5' cap mRNA structure, nor the help of many of eukaryotic initiation factors (eIFs). They can directly recruit the 40S and 60S ribosomal complexes and initiate protein translation (Sharathchandra, Katoch, & Das, 2014). The mechanisms of IRES-dependent translation initiation have not been fully explained. Initiation has been shown to function through secondary structures, in some cases using specific sequences which can bind the ribosome, and with or without the help of eukaryotic initiation factors (eIFs) or IRES-specific trans-acting factors (ITAFs) (Mailliot & Martin, 2018; Plank & Kieft, 2012). The eIFs may be the same as those used in the canonical mechanism, while ITAFs can be a variety of small molecules or proteins that function to modify the structure of IRES or eIFs and promote interaction with the ribosome. The first discovered ITAF was polypyrimidine tract-binding protein (PTB) which acts with the poliovirus (PV) IRES (Meerovitch, Pelletier, & Sonenberg, 1989). Later, researchers found that ITAF contributions to IRES function could be positive or negative (Stoneley & Willis, 2004).

IRES were first discovered in the poliovirus (PV) and encephalomyocarditis virus (EMCV) RNA genomes in 1988 (Jang et al., 1988; Pelletier & Sonenberg, 1988). A cap-independent translation mechanism must exist in these viruses because both PV and EMCV are naturally uncapped. A specific region of the viral RNA was found to play an important role in viral translation and those regions were named internal ribosome entry sites (Pelletier & Sonenberg, 1988). Many other viral IRES were later found (Weingarten-Gabbay et al., 2016). Many positive-strand RNA viruses, or [++] ssRNA viruses, which make up more than one-third of known virus

genera, are naturally uncapped and rely heavily on IRES-dependent translation for expressing their genome.

Additional IRES have been found in cellular genes, especially those whose function depends on extreme conditions. The first cellular IRES-containing mRNA was discovered in 1991 in the gene encoding immunoglobulin heavy-chain binding protein (BiP), and this started the search for cellular IRES (Macejak & Sarnow, 1991). Cellular IRES have been widely found in genes involved in proliferation, growth and apoptosis (Baird, Turcotte, Korneluk, & Holcik, 2006).

Extensive investigation of IRES function has found it is likely that more than one mechanism exists. Viral IRES can be divided into four types depending on their sequence, secondary or tertiary structure, and mechanism (Mailliot & Martin, 2018). Class I and II, such as PV and EMCV, need almost the full set of eukaryotic initiation factors (eIFs) to function. The complex of IRES and eIFs acts as a complex to recruit the ribosome. In the absence of eIFs, their IRES function disappears (Mailliot & Martin, 2018). In contrast, Class III IRES, such as that of hepatitis C virus (HCV), only need the help of a subset of eIFs. Class IV viral IRES, such as that of cricket paralysis virus (CrPV), are the most different. They do not need eIFs for efficient IRES function (Mailliot & Martin, 2018). The different dependence on cellular factors observed in these four groups may be because the functions of some eIFs have been replaced by part of the structure of the Class IV IRES RNA (Mailliot & Martin, 2018).

The mechanism of cellular IRES differs from viral IRES (Baird et al., 2006). Cellular IRES are more likely to need the help of eIFs or ITAFs than viral IRES (Kozak, 2005). Many kinds of ITAFs have been identified or proposed, making the mechanism of cellular IRES more difficult to clearly define. Important sequence motifs have been identified by point mutation or deletion analysis. Partial deletions have been found that barely eliminate the cellular IRES function. The

minimal size 9-nt cellular IRES has been reported which shows that the structure of cellular IRES may not be as important as in viral IRES (Chappell, Edelman, & Mauro, 2000; Stoneley & Willis, 2004). It has been proposed that the structure of the complex of cellular IRES, the eIFs, and the ITAFs enables their function, rather than structure of the cellular IRES itself (Stoneley & Willis, 2004). Not all IRES work efficiently under the same physiological conditions, and a single IRES may have different responses to different cellular environments such as cell-types or tissues (Bonnal et al., 2003; Stoneley & Willis, 2004).

The currently known IRES are diverse in length, position, sequence and even structure. Their length varies from 19 nt to hundreds of bases. Their positions relative to the translation start site differ too. They were first found near the 5' end of mRNA untranslated regions (UTRs), but this relative position varies between different IRES. More recently, researchers found that 3' UTRs and even protein coding regions may contain IRES (Weingarten-Gabbay et al., 2016). There is no universally conserved short sequence or secondary structure that has been reported to be present in all IRES. All-in-all, the detailed mechanism of IRES remains a mystery.

1.2 Why are internal ribosome entry sites important?

Initiation of translation at internal ribosome entry sites (IRES) provides an alternative translation initiation mechanism compared to the typical 5' cap dependent initiation. It supplies a back-up plan for protein translation when the canonical mechanism is repressed or inactivated. IRES have been found to play important roles in viral infection, cellular apoptosis, cellular differentiation and response to external stimuli such as hypoxia, serum deprivation and heat shock (Baird et al., 2006; Stoneley & Willis, 2004).

IRES are found both in virus and cellular genes, and almost 10% of cellular mRNAs are believed to have the potential to initiate translation using IRES (Baird et al., 2006), which means

this mechanism is widely used. There are many known viral IRES because many viruses do not have capped 5' mRNA structures, and heavily rely on IRES function. A significant number of cellular IRES exist because the regulation of protein translation is necessary, especially when various stimuli repress the function of cap-dependent translation initiation (Schwanhausser et al., 2011).

Viral IRES play a vital role in viral infection, so they can act as therapeutic targets. For instance, HCV infects 130–150 million people in the world (www.who.int/mediacentre/factsheets/fs164/en/), and it is a major cause of hepatocellular carcinoma. The HCV IRES is important for viral propagation and virulence, because it is essential for the expression of viral proteins. Attempts at preventing IRES function, IRES directed therapies, include identifying antagonists that can interrupt IRES function and control the expression of viral proteins (A. A. Komar & Hatzoglou, 2015). Such drugs could be small-molecule inhibitors such as peptide nucleic acids (PNAs), short hairpin RNAs (shRNAs), small interfering RNAs, antisense oligonucleotides, and ribozymes (A. A. Komar & Hatzoglou, 2015; Martinand-Mari, Lebleu, & Robbins, 2003; Nulf & Corey, 2004).

Cellular IRES work as a backup plan when the cells are exposed to extreme or unusual physiological conditions, and the canonical 5' cap mechanism is repressed. For example, in apoptosis, cap-dependent protein translation is inhibited by modification of several eIFs (Clemens, Bushell, Jeffrey, Pain, & Morley, 2000). However, some genes, such as c-*myc*, DAP5, XIAP and PKCδ, have been found to be expressed using IRES (Henis-Korenblit, Strumpf, Goldstaub, & Kimchi, 2000; Holcik, Yeh, Korneluk, & Chow, 2000; Morrish & Rumsby, 2002; Stoneley et al., 2000). Amino-acid starvation leads to silencing of global protein synthesis by significantly increasing eIF2α phosphorylation. But the expression of the cationic amino-acid transporter gene

CAT1 is activated under IRES control (Fernandez et al., 2001), even during amino-acid starvation. IRES can also stimulate $p58^{PITSLRE}$ translation during G2/M phase of the cell cycle, when most protein translation stops. Evidence shows that an IRES exists in the UTR of the $p58^{PITSLRE}$ mRNA (Cornelis et al., 2000). The above examples suggest that IRES may play a role in tumorigenesis. A significant increase in the expression of c-*myc* proteins in the human neoplasia multiple myeloma (MM) has been attributed to an IRES mediated mechanism (A. A. Komar, & Hatzoglou, M., 2005). Improved understanding of cellular IRES function under different physiological conditions will help us to understand the response of cells in proliferation, apoptosis and tumorigenesis.

1.3 How many internal ribosome entry sites have been found?

About 10% of protein translation initiation events in eukaryotic cells depend on IRES mechanisms (A. A. Komar, & Hatzoglou, M., 2005; Stoneley & Willis, 2004). However, since the discovery of the first viral IRES to 2016, 30 years, fewer than 300 IRES have been reported. The limited number of known IRES shows they are not easily defined. This is because traditional approaches to identifying IRES are time consuming and have a high false positive rate.

The bicistronic assay is the best experimental method for confirming the presence of IRES (Baird et al., 2006). Basically, the potential IRES segment is placed between two reporter genes in a reporter construct. The expression level of the downstream gene can be compared with that of the upstream gene to see whether the segment of interest stimulates translation initiation, or simply acts to increase transcription.

The bicistronic assay may also have high false positive rates. When the putative IRES are examined more carefully, they are often found to harbor cryptic promoters or splice sites (Kozak, 2005). The reason that splicing activity produces false positives is because of the widely use of

pRF vector in most reported bicistronic assays. The pRF vector, which contains the *Renilla* luciferase reporter gene near 5' end, and *firefly* luciferase reporter gene near the 3' end, has been found to generate spliced transcripts (Van Eden, Byrd, Sherrill, & Lloyd, 2004). For example, the poliovirus and c-*myc* IRES have been found to exhibit more than 20-fold less induction of gene expression when the order of the two reporter genes is reversed in the pRF vectors (Hennecke et al., 2001; Nevins, Harder, Korneluk, & Holcik, 2003). The identification of a splice-donor segment within the *Renilla* luciferase gene explains why splicing activity might cause the false positives in the bicistronic assay (Van Eden et al., 2004). The existence of cryptic promoter activity may cause false positives in the bicistronic assay, as well (Han & Zhang, 2002). Additional experiments typically need to be done confirm the absence of cryptic promoters and splicing sites in the IRES bicistronic assay.

IRESite was the first database to systematically summarize all reported IRES (Mokrejs et al., 2010; Mokrejs et al., 2006). IRESite includes, in total, 52 viral IRES and 64 cellular IRES. The predicted secondary structures of the IRES are included, as well as their corresponding ITAFs (when known). The IRES are annotated, including the positions of their boundaries, the confidence level of IRES function, the organism, the ORF absolute position, the protein annotation of the ORF, and the papers identifying them. IRESite has not been updated since 2009.

In 2016, a high-throughput IRES activity detection assay was developed to find additional IRES in human and viral genomes (Weingarten-Gabbay et al., 2016). It represents the first time that researchers have tried to find IRES on a large scale. This work has increased the number of sequences with known IRES activity by more than 10-fold. This assay has selected 55,000 sequences with defined lengths of 173 nt from viral 5'UTRs and segments of complete viral genomes, 5'UTRs of human genes, and segments of complete transcripts, all the reported IRES,

and some mutated IRES segments. This dataset is available online (https://bitbucket.org/alexeyg-com/irespredictor/src). IRES activity for of the library of sequences has been tested by inserting them into a lentiviral bicistronic plasmid, between mRFP and eGFP reporter genes, and infecting H1299 cells, which results in integration of a single oligonucleotide construct in each cell. The cells are sorted with FACS and assigned to 16 bins on the basis of eGFP expression. IRES activity is defined by those expression levels. False positives due to cryptic splicing and promoter activity have been eliminated by further experiments. For promoter activity measurements, the library was cloned into a plasmid that lacks a promoter and the eGFP+ population sequenced. Sequences for which >20% of their reads were obtained in the eGFP+ population were assumed to contain active promoters, and were not considered to be IRES. For cryptic splicing elimination, deep-sequencing reads were obtained from both cDNA and genomic DNA samples, and the ratio between the two was compared. Any sequences whose expression showed significant reduction, indicating cryptic splice sites, were not considered to be IRES.

1.4 How do we find more internal ribosome entry sites?

Although the high-throughput IRES detection assay has increased the total number of found IRES to more than 2000, there are still more novel IRES waiting to be discovered. For example, MrTV, which is associated with moralities, has been isolated, characterized (Pan, Xiaoyi, et al. 2016) and reported to contain a Dicistrovirus-type IGR IRES. Traditional wet lab experiments, such as the bicistronic assay, are always time consuming and labor-costing. If there is a way to select a group of sequences more likely to be IRES sequences from a large candidate sequence pool, it will save much time and effort. A high-accuracy bioinformatics tools might help.

An advantage of bioinformatics compared with traditional biological bench work is that it is faster and less costly. Instead of doing bicistronic assays directly, which might take months,

bioinformatics tools such as sequence alignment, secondary structure prediction and others can be carried out in advance to make a preliminary prediction of whether a sequence is an IRES. This can serve as a guide to prioritize the testing IRES candidate pool, or even as a tool to make novel discoveries. For example, the designed viral IRES prediction system (VIPS) uses the RNA Align program to predict viral IRES (Hong, Wu, Chang, & Chen, 2013; Wu, Hsieh, Hong, Chen, & Tsai, 2009). However, this comparative approach tends to miss candidates with low sequence similarity. In addition, the VIPS can only predict viral and not cellular IRES, and the reported high predictive accuracy appears to be an artifact of testing on their training data. When the recently discovered viral IRES have been tested in this program, the accuracy is extremely low.

Based on the discussion in Chapter 1.1, IRES, including viral and cellular IRES, are mechanistically diverse, and they lack common motifs in either sequence or structure. Some short motifs have been reported to be conserved in certain subgroups of IRES, but not in all of them. Considering this, machine learning tools might help to predict IRES. Machine learning is an approach which can extract informative knowledge from a mass of data. It enables computers to assist humans in the analysis of large, complex data sets. Usually, it works by dividing the whole dataset into training and testing parts, building a model based on features that reflect important characteristics of the training dataset, and predicting the results on the testing dataset. The prediction error on the test dataset, which was not included in the predictive model, indicates the quality of the model.

Many machine learning algorithms have been successfully applied in genetics and genomics in applications such as annotation of sequence elements, classification of functional RNA, and so on (Degroeve, De Baets, Van de Peer, & Rouze, 2002; Heintzman et al., 2007; Libbrecht & Noble, 2015; Ohler, Liao, Niemann, & Rubin, 2002). Support vector machines are

one kind of machine learning algorithm, which apply the statistics of support vectors to classify unlabeled data by maximizing the distance (margin) between two labeled groups. IRESpred is an example of using support vector machine to predict IRES (Kolekar, Pataskar, Kulkarni-Kale, Pal, & Kulkarni, 2016). The model incorporates 35 features, which are sequence and structure related features of the UTRs, and the probabilities of interactions between the UTRs and small subunit ribosomal proteins (SSRPs). It shows improved the prediction performance compared to VIPS, but some defects in IRESPred still exist. The model performance comparison with VIPS is based on the IRESPred training dataset which is not convincing. It shows that the IRESPred model works better only on the IRESPred training dataset, rather than the VIPS training datasets. To test the accuracy of IRESPred, I have generated 50 random sequences greater than 250 nt long and submitted them as queries to the IRESPred website. 48 of these random queries were predicted to be potential IRES, which suggests an extremely high false positive rate. Many features that are included in IRESPred, such as UTR length and number of upstream AUGs, are not particularly relevant to IRES function, probably increasing the prediction noise. Furthermore, the positive training dataset is too small (only 192 samples), probably leading to overfitting. And only 10 sequences can be tested at a time on the IRESPred website which makes large-scale testing impossible.

Considering the defects of VIPS and IRESPred, we can see the importance of the training dataset and choice of features in predicting IRES by machine learning. Optimally, one should use a large positive IRES training dataset, and the model features should be representative of the nature of IRES themselves. The high-throughput IRES activity detection assay developed in 2016 has increased the number of positive IRES by more than 10-fold and makes the prediction of IRES by machine learning more practical (Weingarten-Gabbay et al., 2016). Based on that larger positive

IRES dataset, a stochastic gradient boosted tree algorithm has been implemented to predict IRES (Gritsenko et al., 2017). In that model, 5814 different sequence kmers have been used as features. The structural features, such as the number of unpaired nucleotides, was also examined in this work, but they did not improve model performance. The problem with this gradient boosting approach is likely to be the inclusion of too many features. Compared with the training dataset of 23,000 examples, 5814 features, many of which are highly correlated, is too many to use to fit the model.

IRESfinder uses a logit model with framed kmer features to find cellular IRES based on the same dataset (Zhao et al., 2018). The disadvantage of IRESfinder is that it is designed only for cellular IRES. And the logit model, as a transformed linear model, may not work well for non-linear relationships. In addition, the independent dataset is very small (only 13 sequences), which will cause the reported AUC to be overestimated. 9 structural features, the number of predicted hairpin-, bulge-, internal-, and multi-loops, the total number of loops, the maximum loop length, the maximum hairpin-loop length, the maximum hairpin-stem length, and the number of unpaired bases were included in this model, but showed no importance.

1.5 Objective of this study.

Understanding the mechanism of IRES is important for us to better understand viral disease therapeutic treatments, as well as the response of cells in proliferation, apoptosis and tumorigenesis. However, due to the limited number of reported IRES and the diversity of their sequence and folded structures, the mechanism of IRES are still not very clear. Systematically studying the existing IRES database and identifying features that can distinguish IRES and nonIRES sequences is an important step in predicting IRES. Such discriminative features can then be incorporated into machine learning models to achieve high prediction accuracy.

The objective of this study is to systematically study the pattern of all the known Internal Ribosome Entry Sites from point of view of sequence, structure, and any other possible shared common motifs, and accurately predict novel IRES using machine learning methods.

Table 1-1 List of Advantages of disadvantages of different IRES prediction tools

	Dataset	Features	Methods	Disadvantages
VIPS	4 Types of viral IRES as positive, coding sequence as negative.	N/A	Align predicted secondary structure	1. Coding sequence cannot set to be negative training control. 2. Low true positive for novel sequence. 3. Training dataset is too small.
IRESPred	Known IRES as positive, coding sequence and housekeeping UTR as negative.	UTR length, # of AUGs, hairpin-loops, MFE, predicted interaction probabilities between UTR and SSRP	Support Vector Machine	1. UTR is not a good feature because of the diversity of IRES sequence length. 2. # of AUGs is not a good feature because many IRES are reported not to have AUG. 3. Coding sequence can not set to be negative training control. 4. Training dataset is too small.
IRES-intepreter	20872 Synthetic sequences	6120 Kmer features	Gradient Boosting Decision Tree (GBDT)	1. Too many features for training and the global, local kmers have high correlation. 2. GBDT training time is way too slow which takes several days.
IRESfinder	Human IRES and non IRES in the 55000 Synthetic sequences	19 Kmer features	Logit Model	1. Logit model which not work well on non-linear relationships. 2. 13 Independent testing dataset is too small and they are not randomly generated.

CHAPTER 2. DESCRIPTIVE ANALYSIS: ROLE OF INTERNAL RIBOSOME ENTRY SITES IN PROTEIN TRANSLATION INITIATION

2.1 Introduction

Internal ribosome entry sites are segments of mRNA in the untranslated regions of mRNA which can initiate protein translation without the presence of a 5' cap structure. (A. A. Komar, & Hatzoglou, M., 2005; Mailliot & Martin, 2018). They have been found in many cellular and viral genomes. Almost 10% of mRNAs are believed to potentially function as IRES (Hershey, Sonenberg, & Mathews, 2012; Stoneley & Willis, 2004). However, due to the limitations of the IRES bicistronic assay, which is the gold standard for experimental identification of IRES relatively few IRES have been definitively described and functionally validated compared to the potential overall population (Baird et al., 2006). The mechanism IRES function is different between viral IRES and cellular IRES (Baird et al., 2006). Cellular IRES usually require eukaryotic initiation factors (eIFs) and IRES trans-acting factors (ITAFs) to function (A. A. Komar, & Hatzoglou, M., 2005). Viral IRES have been divided into four groups based on sequence and functional differences (Mailliot & Martin, 2018). Some sequence motifs or structural regions have been found to be important for the IRES function, but such features are not consistently found existing in all reported IRES. In a word, the mechanism of IRES function remains unclear. Researchers know which factors, such as the eIFs or ITAFs, are involved in the IRES activity, but they are not sure what the determined reason for the IRES function. The important IRES motifs have been identified by investigating the effects of point mutations or small deletions on IRES function. Two main mechanisms of IRES have been proposed (Weingarten-Gabbay et al., 2016). One class of IRES have been called global-sensitive IRES. The function of these IRES is affected

by mutations in any of several sequence segments. This implies that the mechanism of global-sensitive IRES is more likely to depend either on folded structures in the mRNA, or by structures stabilized by eIFs or ITAFs (Weingarten-Gabbay et al., 2016). The AEV IRES is one such global-sensitive IRES, and there are multiple regions in its sequence in which mutation affect its function (Weingarten-Gabbay et al., 2016). The other IRES functional class are called local-sensitive IRES. Locally-sensitive IRES lose their IRES function by only when mutations occur in a single specific region. Mutations in other regions will not alter their IRES function. Such IRES are more likely to require eIFs or ITAFs, and it is believed that the specific regions might be important for protein binding. Any change in these binding regions might prevent the formation of the necessary ribonucleoprotein complex and thereby repress or eliminate IRES function. Other mutations, those that do not affect the binding, should not affect their IRES function. XIAP is an example of a local-sensitive IRES (Stoneley & Willis, 2004).

Identifying a large set of IRES is an important step towards understanding IRES mechanism. However, the bicistronic assay is time consuming and labor intensive. Development of a new bioinformatic tool to classify IRES and nonIRES sequences can work as a tool to pre-filter all the tested sequences and is important to find more potential IRES. To achieve this goal one must systematically study the sequence and structural features of known IRES in order to define features that can be used to consistently identify IRES from their sequence.

In this dissertation, I study the sequence as well as the structure of all known IRES and explore their roles in the protein translation. Two existing databases that have been created to systematically study the IRES, provide useful background information for this study. The first database, referred to as Dataset 1 in this dissertation, is composed of IRESite and some selected 5’UTRs of housekeeping genes. 52 viral IRES and 64 cellular IRES from IRESite are labeled as

IRES in Dataset 1. Housekeeping genes principally utilize the 5' cap dependent mechanism for initiation and 51 of them have been selected as the nonIRES group used for comparison in Dataset 1 (A. A. Komar, Mazumder, & Merrick, 2012). Dataset 2 results from a high-throughput bicistronic assay developed in 2016, and its application has increased the number of known IRES by more than 10-fold (Weingarten-Gabbay et al., 2016). This large increase in the number of examples of IRES provides an opportunity to better learn the relationship between sequence and structural features and IRES mechanism. In this chapter, I have made a descriptive analysis of all the reliable reported IRES and identified distinctive characteristics of IRES and nonIRES sequences.

These distinctive characteristics can be used as features by machine learning models trained to predict IRES. A feature is a measurable property or characteristic of an observed phenomenon (Nasrabadi, 2007). Classification models are one kind of machine learning methodology which can use those features to distinguish classes of examples, in this case, IRES and nonIRES. Many groups of functional RNAs can be predicted by machine learning methods. For example, miRAlign uses features based on the predicted secondary structure of pre-miRNAs to detect miRNAs (Wang et al., 2005). MicroRNA precursors have been predicted by a combination of sequence and structure features which have been called triplet features (Xue et al., 2005). IRESPred uses both primary sequence features and predicted secondary structure features to classify IRES (Kolekar et al., 2016).

Many different features could be explored. Short sequence motifs have been reported to be conserved in some sub-groups of viral IRES, and the kmer features have been used to predict IRES in a stochastic tree model (Fernandez-Miragall & Martinez-Salas, 2003; Gritsenko et al., 2017). The primary sequence of all reported IRES may contain some features which can be used for the

prediction, for instance sites for sequence-specific protein binding or iteration between the IRES and the ribosomal RNA. The structure of RNA is believed to be correlated with its function, and plays a very important role in many RNA groups such as tRNA and microRNA (Mattick, 2018; Mattick & Makunin, 2006). The pseudoknoted structures have been reported to be important in Dicistrovirus IRES (Mailliot & Martin, 2018). This chapter will explore possible IRES features from both sequence and structure perspectives, and Chapter 3 will use those features to build machine learning models.

2.2 Datasets and Methods

2.2.1 Datasets:

There are two datasets used in this research. The first dataset is derived from IRESite, which contains 52 viral IRES and 64 cellular IRES (Mokrejs et al., 2010; Mokrejs et al., 2006). The total number of IRES observations is 52+64 =116. The identification of an appropriate nonIRES control group is important for identifying IRES features as well as for fitting of classification models in Chapter 3. IRES can be located in the 5'UTR regions of the mRNA, protein coding regions, or even in 3'UTR regions (Weingarten-Gabbay et al., 2016). The nonIRES regions in the 5'UTR of known viral and cellular IRES can be treated as nonIRES because they have been experimentally tested and shown to lack IRES function when the IRES regions have been deleted. For each IRES in Dataset 1, an identical length nonIRES has been randomly selected to serve as a nonIRES, so the number and lengths of the nonIRES sequence is the same as that of the positive IRES (116). Housekeeping genes principally use the cap dependent initiation pathway (A. A. Komar et al., 2012). UTRS of 51 housekeeping genes were selected from the recent publications, and their UTRs included as nonIRES sequences (Eisenberg & Levanon, 2013; Kolekar et al., 2016). The lengths of housekeeping UTRs lie within the range of those sequences

in the positive dataset. The total number of nonIRES sequences is thus 116+51=167. The length distribution and GC content of the positive and negative datasets are similar. The whole dataset will be divided into training and testing partitions at a later stage.

The second dataset, Dataset 2, is derived from a work of Weingarten-Gabbay et al.,(Weingarten-Gabbay et al., 2016). In this work, Segal's group developed the first high-throughput assay for IRES activity. This experimental dataset has increased the number of known IRES by more than 10 times. They tested 55,000 sequences which were from several different resources - reported IRES, 5'UTRs of human genes, 5'UTRs of viral genes, sequences complementary to 18S rRNA. Human transcripts and viral genome sequence fragment were screened using a consistent 173 nt insert size, removing any length effects. This dataset is available online (https://bitbucket.org/alexeyg-com/irespredictor/src). From the 55,000 tested sequences, 28,669 native fragments from human and viral genomes have been used in this dissertation. The rest are synthetic sequences and they haven't been included.

Based on the reported replicate measurements of IRES activity, promotor activity, and splicing activity, we further filter the selected set of sequences to obtain a reliable set for model training. All sequences with splicing scores below -2.5 or promoter activity above 0.2 were removed because of the possible artifacts of cryptic promoters and splicing sites. Finally, 20872 native sequences are included in this dataset. 2129 sequences with IRES activity scores above 600 are defined as IRES, and the other 18743 sequences with IRES activity score below 600 are defined as nonIRES. The ratio of IRES to nonIRES is about 1:9 in dataset 2, as shown in Figure 2.8. The distribution of IRES across different groups is shown in Figure 2.9. Most detected IRES are in viral segments especially in the 5'UTR regions.

Those two datasets can't be combined because the data source of those two datasets is different. The IRES in Dataset 1is from IRESite collected from various reported IRES. Their experimental conditions and proving methods are different for each IRES. The IRES from Dataset 2 is from one experimental condition, a high-throughput bicistronic assay. Having two datasets is important because they include both experimentally confirmed results for IRES sequences and nonIRES sequences. The number of overlapping sequences is low, and there are only 43 shared sequences which are all IRES in the two datasets (Figure 2.1). Dataset 2 does not include known IRES whose sequence is longer than the 173 nt fragment size used in the sequence construct. It is important to explore shared features among all known IRES with different sequence lengths. So having Dataset 1 is necessary. On the other hand, some features such as predicted MFE, are highly dependent on the sequence length, and it is meaningless to test them on Dataset 1 where the sequence lengths vary widely, but they can be tested on Dataset 2, in which all sequences share the same length. What's more, there is rapid increasing number of IRES in Dataset 2 and those more IRES examples will benefit the machine learning model training. In chapters 2 and 3, both Dataset 1 and Dataset 2 have been used for selection and model training.

The idea that there may be two distinct kinds of IRES, globally sensitive IRES and locally sensitive IRES, was raised by Weingarten-Gabbay et al. in 2016 (Weingarten-Gabbay et al., 2016). Globally sensitive IRES were proposed to be a group of IRES whose function is abrogated by mutations in any of several segments, because globally sensitive IRES require the entire secondary structure to maintain IRES function. Locally sensitive IRES differ in that their IRES activity is not affected by mutations in the bulk of their sequences, but are affected by mutations in specific areas important for the binding of ITAFs or eIFs. The proposed list of representatives of those two different IRES groups is shown in Table 2.4.

2.2.2 Methods

The goal of this section is to describe features that might be used to distinguish IRES and nonIRES sequences. To explore which features can be used, we must understand what features may be important for IRES function. In another words, what features are highly correlated with IRES function.

There is no unique location of IRES with respect to the translation start site. Most IRES are found in the 5' UTR region of mRNAs. More recently, IRES have been found in the 3'UTR, and even in protein coding regions (Weingarten-Gabbay et al., 2016). The genome wide landscape was first examined in 2016, and figure 2.8 shows the distribution of detected IRES in humans and positive-strand RNA viruses.

RNA structure determines the function of IRES, and groups of RNAs which shares similar functions may also share similar structures. RNA structure can be divided into primary structure, which is the sequence of the RNA, and secondary structure, which includes the base-paired stem regions, loops and so on. Tertiary structure is the three-dimensional conformation of the RNA polymer. The primary sequences of IRES have been reported to be diverse, although, some small common sequence motifs are seen in specific viral IRES groups (Fernandez-Miragall & Martinez-Salas, 2003). Parts of the secondary structure are conserved in some groups, but there is considerable variability even in viral IRES. Viral IRES can be divided into four types (Fig 2.2). Some of the structural motifs, shown in boxes in Fig 2.2, have been reported to be conserved and important for the IRES function (Plank & Kieft, 2012). However, even these short motifs are restricted to certain species, and there are no universal motifs that are in common in both viral and cellular IRES. Thus, distinct features shared by all IRES are not easy to define. Different potential features that can be used to separate IRES and nonIRES groups are discussed below.

In the following section, features are discussed and their ability to distinguish between IRES and nonIRES groups evaluated. The two sample Student's t-test has been applied to test whether the difference in the mean values of the features is significant. The differences between IRES and nonIRES groups are considered to be significant when their test p values are less than 0.05 (α=0.05).

2.2.2.1 The primary sequence features

There is no clear evidence showing a difference between IRES and nonIRES in GC content or location (Baird et al., 2006). However, some groups of viral IRES have common primary sequence motifs. Viral IRES have been grouped into four groups based on their potentially different mechanisms and different shared motifs (Mailliot & Martin, 2018; Plank & Kieft, 2012). For example, a GNRA motif has been reported to be important for IRES activity in the central domain of the foot-and-mouth disease virus (FMDV) and Poliovirus (PV) (Fernandez-Miragall & Martinez-Salas, 2003) IRES. It is also possible that some weak features or complex features may have been missed, so primary sequence may still be considered as a useful source of features. Since there are no shared short sequences universally present in viral or cellular IRES, kmers (subsequences of length k) are good candidate features which can be used to represent IRES similarity from a sequence perspective.

As for mRNA, there are four possible choices, adenine (A), cytosine (C), guanine (G), or uracil (U) in each nucleotide position. The number of different kmers with length k is 4^k. For example, there are four 1mer features, sixteen 2mer, 4^3=64 3mer and 4^4=256 4mer features respectively. The frequency of each kmer has been used as features and they are calculated as the count of kmers divided by the sequence length. I consider two types of kmer feature: global kmers and local kmers. Global kmers are counted over the entire length of the sequence, whereas local

kmers are counted in certain regions that are defined with respect to other features (for instance, the translation initiation codon).

2.2.2.2 The secondary structure features

Secondary structure features include all the possible statistics that can be used to describe the secondary structure of IRES. The secondary structure of RNA is the base-pairing interaction of the single stranded nucleic acid polymer with itself (Dirks, Lin, Winfree, & Pierce, 2004). Usually, it can be decomposed into stems (multiple base-pairings at successive positions) and loops (nonpaired bases at successive positions). Stem-loop structures (also called "hairpin loops") contains a base-paired helical stem ending with a short unpaired loop. Hairpin loops are extremely common in RNA structure and are building blocks of larger structural motifs. Stems contribute more to thermal stability of the RNA secondary structure than loops (Yakovchuk, Protozanova, & Frank-Kamenetskii, 2006). the presence of more stems usually indicates a more stable (lower deltaG of folding) secondary structure. The predicted minimum free energy (MFE) is the most popular secondary structure feature. It is calculated using a nearest-neighbor thermodynamic model (Zuker, 1981). The predicted MFE is increased by the presence of non-paired bases and decreased by the stacking energy of paired bases. A lower predicted MFE indicates a higher degree of folding and greater stability.

Several secondary structure features such as MFE and the number of loops and hairpins have been incorporated in previous IRES prediction approaches (Gritsenko et al., 2017; Kolekar et al., 2016; Zhao et al., 2018). None of these features were found to have significant predictive value. However, this does not mean that structural features are not important for the IRES function. It just means the structural features that they used do not work well on their training dataset and for the models they were training. If the model or dataset is different, structural features might still

have predictive value. Other structural features are considered in this chapter to see whether they can contribute to the model. (Zhu et al., 2017)

Secondary structure features are important because they have been demonstrated to be correlated with the function of many RNA groups. What's more, IRES sequences are usually located in highly regulated regions to which both ribosomes and other regulatory proteins bind, and secondary structure has been reported to be important for the binding of ribosome or ITAFs (Baird et al., 2006; Lozano, Fernandez, & Martinez-Salas, 2016). IRES regions are likely to form some RNA secondary structures and the predicted MFE is thus expected to be lower than that of random sequences with the same length. In contrast, The MFE of protein coding regions and UTRs of housekeeping genes are more likely to have higher predicted MFE than random sequences because they are thought to be less extensively folded and have less regulation at the translational level.

2.2.2.3 The predicted Q_{MFE} value

The predicted minimal free energy (MFE) is highly correlated with sequence length (Trotta, 2014). In this work we try to find features that reflect the degree of base-pairing without being explicitly dependent on the sequence length. The Q_{MFE} value of the predicted MFE, which is based on the ration of the predicted MFE and the predicted MFE of randomized is such a feature (Bonnet, 2004). Q_{MFE} is the quantile of the original sequence MFE, divided by the MFE of random sequences. Q_{MFE} is calculated as follows:

(1) Calculate the predicted minimum freedom energy of the secondary structure from the original sequence by RNAfold.

(2) The original sequence has been randomized by permuting the dinucleotide ratios. Then the MFE of the randomized sequence has been generated.

(3) Repeat step 2 many times (for example 2000) in order to obtain a distribution of the predicted MFE values.

(4) If N is the number of iterations and n is the number of randomized sequences which MFE value are less or equal to the original value, then Q_{MFE} is calculated as:

$$Q_{MFE} = \frac{n}{N+1}$$

It is believed that the stability of an RNA secondary structure depends crucially on the stacking of adjacent base pairs (Yakovchuk et al., 2006). Therefore, the frequency of dinucleotides in the random sequences is an important consideration in calculating the MFE of randomized sequences (Clote, Ferre, Kranakis, & Krizanc, 2005). In calculating Q_{MFE} a dinucleotide preserving randomization method has been used to generate random sequences. The Ushuffle program (Jiang, Anderson, Gillespie, & Mayne, 2008), which is based on the Euler algorithm, can generate such qualified negative controls to calculate Q_{MFE}. It can randomize RNA sequences with a more biological meaning by maintaining the dinucleotide counts (Jiang et al., 2008).

Q_{MFE} can be used to compare the degree of predicted secondary structure in different sequences regardless of length. This length independent feature is a great tool to see whether a sequence is more likely to form a more complex secondary structure. For example, viral IRES have been found to have highly folded secondary structure that is critical to their functions. The structures of Dicistrovirus IRES are conserved and comprise highly folded structure with three pseudoknots. Cellular IRES usually need ITAFs to initiate translation, and the binding between ITAFs and cellular IRES may change thee IRES structure from a relaxed status to a rigid status (Filbin & Kieft, 2009). Cellular IRES are therefore likely to have a less folded secondary structure. The 5'UTRs of housekeeping genes do not require highly folded structures because they use a cap-dependent translation initiation process. The ribosome complex, as well as some eIFs are recruited to the end of 5' cap, and in the most popular model, the ribosome scans UTR regions until it finds

an initiation codon. Highly structured UTRs might stop the movement of the ribosome, leading to the assumption that there should be no highly stable structures in these regions. The Q_{MFE} values are expected to be different in viral IRES, cellular IRES and the UTRs of housekeeping genes.

2.2.2.4 Triplet features

MFE and Q_{MFE} score reflect the folding status of the RNA. Features that represent both the primary sequence and the base-paired structure are also considered. Triplet features combine contiguous paired or unpaired predicted structures with sequence information (Vitsios et al., 2017). The first successful application of this kind of features was in the implementation of a support vector machine algorithm for classifying pre-miRNAs (Xue et al., 2005). The definition of triplet features is shown in Figure 2.5.

2.3 Results and Discussions

2.3.1 Sequence features like kmer are selected to predict IRES and nonIRES

Kmer features are semi-dependent of the length of the sequence. The count of kmer is obviously dependent of sequence length. The frequency of kmer is independent of length for most cases. But it is length dependent when a relative short sequence is compared with a long sequence if longer length of k is considered. Since the length of IRES in Dataset 1 varies a lot, exploring the predictive ability of kmer features makes more sense when they are tested on Dataset 2 in which all sequences are the same length. Four different lengths of kmer features were introduced in 2.2.2.1, and their total number is 4+16+64+256 = 340. Global kmers features are the counts of those four types over the entire length of the sequence. For each kmer features, local kmer features are counted within each 20 nt moving window, with an offset of 10 nt between windows, across the whole sequence. In Dataset 2, the length of every sequence is fixed at 173 nt, so there are 17

local kmer features for each global kmer feature. In total, 340*(1+17) = 6120 global and local kmers were examined. The Student's two sample t-test found significant differences in the frequencies of most kmers. 5842 out of 6120 kmer features including both global kmer and local kmer showed significant differences and these features are potential feature candidates (Figure 2.9) for building a classifier. Part of the significant global kmers have been listed in table 2.5.

If all 6120 kmers were used as features to train a model, over-training is possible because of correlation and redundancy between the kmer features. In addition, the 6120 kmer features are not independent, and global kmer features and their corresponding local kmer features are highly correlated. Figure 2.10 and figure 2.11 show a comparison of the global kmer 'T' and "CGT" and their 17 local kmer features. Whenever global kmer 'T' differs significantly between the IRES and nonIRES group, its local kmers are more likely to show significant differences as well. When global kmer 'CGT' is not significantly different, its corresponding local kmer features are also unlikely to show significant differences. Correlation plots of something and something, figure 2.12), show the highly correlated relationship as well. Whether it is a good idea to include both global and local kmers in the same model will be further discussed in Chapter 3.

2.3.2 Secondary structure features are selected to predict IRES and nonIRES

Secondary structure features such as MFE, $Q_{MFE,}$ and triplets have been tested on both Dataset 1 and Dataset 2. It is important to test on these two datasets because they are a good complementary. Predicted MFE is usually dependent on the sequence length. It is meaningless to test MFE on Dataset 1 because the IRES group and nonIRES groups have different sequence lengths. MFE can be tested in Dataset 2 because the lengths of all the sequences are the same. Dataset 1 is still necessary because Dataset 2 does not include some known IRES whose sequences

are longer than the predefined 173 bases insert used in the sequence construct. Q_{MFE} and triplet features have been tested on both Dataset 1 and Dataset 2.

In Dataset 1, a large difference in Q_{MFE} of viral IRES, cellular IRES and the 5'UTRs of housekeeping genes is observed (Figure 2.4). The Q_{MFE} values of viral IRES are the lowest. The cellular IRES Q_{MFE} score is usually around 0.5, which indicates an intermediate degree folding of secondary structure. The 5'UTRs of housekeeping genes have the highest Q_{MFE}. Figure 2.3 shows examples of calculating Q_{MFE} by plotting the frequency of predicted MFE between tested sequence and their randomized sequence. Those examples including CrPV (viral IRES, Q_{MFE}=0.001), ERH (housekeeping gene UTR, Q_{MFE}=0.99), Apaf-1 (cellular IRES, Q_{MFE}=0.66) and CrPV nonIRES region (Q_{MFE}=0.94) show their Q_{MFE} value when they are compared with the randomized sequences from Ushuffle. These results confirm that the Q_{MFE} value can represent the degree of predicted secondary structure in various sequence classes and may be useful in distinguishing IRES and nonIRES.

To better explore the predictive ability of Q_{MFE}, the whole CrPV mRNA sequence has been divided into segments with the same length, 200 nt. The Q_{MFE} value of each segments was calculated and we found three locations with higher Q_{MFE} than the other regions (Figure 2.5). Two of them are exactly the regions where the known the 5'UTR IRES (bases 1-708) and intergenic IRES (6000-6200 bases) are found. It indicates that Q_{MFE} may be a powerful discriminatory feature that can locate IRES positions in a mRNA.

Q_{MFE} and MFE have been tested on Dataset 2. The Student's two sample t-test shows significant differences in both Q_{MFE} and MFE between IRES and nonIRES groups (Figure 2.13). It is notable that the IRES group has lower Q_{MFE} and higher MFE compared with the nonIRES group, which indicates there may be more highly folded secondary structure in the nonIRES group.

This may be because there are many cellular IRES which decrease the average Q_{MFE} and increase the average MFE. To further investigate the relationship between Q_{MFE} and the folded secondary structure, the value of Q_{MFE} and MFE in global-sensitive IRES and local-sensitive IRES was compared and we can see that the global-sensitive IRES, which are believed to rely on extensively folded structures for activity, have higher Q_{MFE} and lower MFE (Figure 2.14). Student's t-test shows the differences between MFE and Q_{MFE} are significant, making them potential features for predicting/classifying IRES.

Student's two sample t-test on Dataset 1 shows that 23 out of 32 triplet features are significantly different between viral IRES and 5'UTRs of housekeeping genes (Figure 2.15). The triplet features have also been compared on Dataset 2 where 30 out of 32 were found to be significant (Figure 2.16).

2.4 Conclusions

This chapter has established two well defined datasets that include experimentally confirmed IRES as well as several defined sets of nonIRES. It provides the sufficient examples to explore which features can separate the IRES and nonIRES groups. Features with significant differences between IRES and nonIRES include both sequence features and structural features that may explain the role of IRES in protein translation initiation. From this study, specific kmer features, MFE, Q_{MFE}, and certain triplet features show significant difference between IRES group and nonIRES group and can serve as potential features to predict IRES. This provides a rich set of sequence and structural features to be used in constructing machine learning models for distinguishing IRES and nonIRES. Furthermore, some features reveal significant differences between viral and cellular IRES, and between global-sensitive and local sensitive IRES.

Table 2-1 List of Housekeeping Genes

Gene Name	UCSC ID	Start	End
ENSA	uc001eve.3	150601947	150602142
ERH	uc001xlc.2	69864951	69864991
FH	uc001hyx.3	241683023	241683183
FPGS	uc004bsh.1	34646586	34646726
GPI	uc002nvi.2	34884172	34884364
H1FX	uc003elx.3	129034746	129034901
LSG1	uc003fui.3	194392892	194393311
LSS	uc002zik.2	47647545	47648294
MAEA	uc011bvd.2	1303599	1304051
MAVS	uc002cvv.3	3929918	3930220
MCU	uc001jtd.3	74452377	74452757
AKR7A2	uc001bbw.3	19638619	19638983
AKIRIN1	uc001ccw.3	39456916	39457881
ARNT	uc001evr.2	150849044	150849313
C1orf43	uc001fei.2	154192884	154193801
APOA1BP	uc001fpk.3	156561558	156561750
ARV1	uc001huh.3	231114823	231116019
PDHX	uc001mvt.3	34937677	34937843
AIP	uc001olv.3	67250505	67250939
ARCN1	uc001ptq.3	118443102	118443404
APEX1	uc001vxg.3	20923290	20923505
ARIH1	uc002aut.4	72766667	72766969
ATP5D	uc002lrn.3	1241749	1241990
AES	uc002lwy.1	3062199	3062685
CHMP2A	uc002qti.3	59065580	59066051
COA5	uc002syz.3	99224869	99225064
AGFG1	uc002vpd.2	228336888	228337166
AF055024	uc002vyh.3	239359013	239359195
CSTB	uc002zdr.4	45196151	45196344
C21orf33	uc002zed.4	45553494	45553923
AP1B1	uc003afh.3	29727806	29728205
AX747758	uc003apb.1	36633473	36633790
ADSL	uc003ayp.4	40742504	40743022
BTF3	uc003kcr.1	72794250	72795189
AGGF1	uc003kes.3	76326210	76326628
COX7C	uc003kir.3	85913784	85914320
FSCN3	uc003vmc.1	127231463	127231775
APOOL	uc004eem.3	84258898	84259395
ATP6AP1	uc004flh.1	153657191	153657292
APH1A	uc010pbz.2	150240126	150240519
AKIP1	uc010rbs.2	8932739	8933080
API5	uc010rfh.1	43333505	43333962

Table 2.1 continued

AMBRA1	uc010rgt.2	46564265	46564741
AHSA1	uc010tvk.1	77924373	77924560
KLC1	uc010tyd.1	104029299	104029382
ANP32A-IT1	uc010uka.2	69098985	69099243
ASXL1	uc021wbw.1	30946147	30946299
AAAS	uc001scr.4	53715250	53715529
DPH1	uc031qxv.1	1943966	1944264
ECI1	uc002cps.3	2301568	2301715
FAM178A	uc001krq.4	102672326	102672831

Table 2-2 List of viral IRES in IRESite.

Virus Name	Accession No.	Start	End
Feline leukemia virus	AB818696.1	582	819
Drosophila C virus strain EB	AF014388.1	6078	6266
Rhopalosiphum padi virus	AF022937.1	1	579
Rhopalosiphum padi virus	AF022937.1	6875	7106
Triatoma virus	AF178440.1	1	694
Triatoma virus	AF178440.1	5929	6149
White spot syndrome virus	AF227911.1	303	482
Porcine teschovirus 1	AF231769.1	1	432
Gallid herpesvirus 2	AF243438.1	131117	131361
Taura syndrome virus	AF277675.1	6741	6990
Foot-and-mouth disease virus	AJ133357.1	578	1038
Hepatitis GB virus B	AJ277947.1	30	445
Simian sapelovirus 1	AY064708.1	253	746
Youcai mosaic virus	AY318866.1	4649	4876
Youcai mosaic virus	AY318866.1	5456	5601
Swine vesicular disease virus	AY429470.1	69	635
Human coxsackievirus B3	AY752946.1	1	750
Human enterovirus 71	DQ060149.1	1	748
Homalodisca coagulata virus-1	DQ288865.1	5802	5989
Reticuloendotheliosis virus	DQ387450.1	363	939
Human parechovirus 1	EF051629.2	298	538
Human herpesvirus 1	FJ655111.1	535	573
Moloney murine leukemia virus	J02255.1	495	621
Poliovirus	K01392.1	1	742
Equine rhinovirus 1	L43052.1	245	956
Drosophila melanogaster gypsy transposable element	M12927.1	1	330
Drosophila melanogaster gypsy transposable element	M12927.1	530	790
Hepatitis A virus	M14707.1	151	734
Theiler's murine encephalomyelitis virus	M16020.1	1	1040
Hepatitis C virus subtype 1a	M67463.1	1	383
Bovine viral diarrhea virus 1	NC_001461.1	1	385
Encephalomyocarditis virus	NC_001479.1	257	832
Human immunodeficiency virus 1	NC_001802.1	104	336
Murid herpesvirus 4	NC_001826.2	25330	25715
Plautia stali intestine virus	NC_003779.1	6002	6146
Cricket paralysis virus	NC_003924.1	1	708
Cricket paralysis virus	NC_003924.1	6025	6216
Avian encephalomyelitis virus	NC_003990.1	1	494
Ectropis obliqua picorna-like virus	NC_005092.1	1	390

Table 2.2 continued

Epstein-Barr virus	S45894.1	465	608
Hepatitis GB virus A	U22303.1	15	707
Hepatitis GB virus C	U36380.1	13	642
Human herpesvirus 8	U75698.1	122973	123206
Human poliovirus 1	V01149.1	320	631
Tobacco mosaic virus	V01408.1	4670	4900
Human rhinovirus 2	X02316.1	11	614
Friend murine leukemia virus	X02794.1	1	621
Mouse DNA for virus-like (VL30) retrotransposon BVL-1	X51336.1	462	1144
Equine Rhinovirus type 2	X96871.1	162	920
Turnip vein-clearing virus	Z29370.1	26	173
Turnip vein-clearing virus	Z29370.1	655	795
Hog cholera virus (Classical swine fever virus)	Z46258.1	1	373

Table 2-3 List of cellular IRES in IRESite

Organism and gene Name	Accession No.	Start	End
Drosophila melanogaster, antennapedia (Antp)	NM_206445.1	1323	1574
Mus musculus, apoptotic protease activating factor 1 (Apaf-1)	AF064071.1	1	583
Homo sapiens, apoptotic protease activating factor 1 (Apaf-1)	AK307509.1	504	734
Homo sapiens, mercurial-insensitive water channel (AQP4)	U34845.1	10	293
Homo sapiens, bcl-2-alpha protein (bcl-2)	M13994.1	322	1458
Homo sapiens, v-myb avian myeloblastosis viral oncogene homolog (C-MYB)	NM_005375.3	2	151
Homo sapiens, c-myc oncogene (C-MYC)	V00568.1	1	393
Rattus norvegicus, cationinc amino acid transporter 1 (Cat1)	AF245000.1	47	270
Homo sapiens cyclin D1 (CCND1)	NM_053056.2	1	209
Homo sapiens, eukaryotic translation initiation factor 4 gamma (DAP5)	NM_001418.3	112	416
Homo sapiens, eukaryotic initiation factor 4 gamma (eIF4G)	D12686.1	1	357
Homo sapiens, cDNA FLJ43058 fis (ELG1)	AK125048.1	755	1214
Homo sapiens, clone UGL16c06 (FGF1)	DQ655917.2	50	483
Mus musculus, (GTX)	L08074.1	1	196
Drosophila melanogaster, mRNA for hairless serine rich protein (hairless)	X67239.1	308	742
Zea mays, heat shock protein HSP101 (HSP101)	AF133840.1	1	161
Homo sapiens, laminin (LamB1)	NM_002291.2	1	335
Homo sapiens, MAX network transcriptional repressor (MNT)	NM_020310.2	75	267
Homo sapiens, myelin transcription factor 2 (MYT2)	AF006822.1	997	1152
Mus musculus, N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1 (Ndst1)	NM_008306.4	48	467
Mus musculus, N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 2 (Ndst2)	NM_010811.2	1	750
Nicotiana tabacum, heat shock factor (NtHSF1)	AB014483.1	1	453
Rattus norvegicus, ornithine decarboxylase (ODC1)	M16982.1	1	303
Homo sapiens, protein kinase PITSLRE alpha 2-2 (P58PITSLRE)	U04816.1	745	1125
Canis familiaris, scamper (scamper)	AF263546.2	1	365
Drosophila melanogaster, Ultrabithorax (Ubx)	BT010241.1	1	966
Saccharomyces cerevisiae, chromosome XI reading frame ORF (YKL109w)	Z28109.1	8	277
Homo sapiens, KIAA0086 (hSNM1)	D42045.1	1	918

Table 2.3 continued

Homo sapiens, heat shock 70kDa protein 1A (Hsp70, HSPA1A)	NM_005345.5	51	243
Mus musculus, potassium voltage-gated channel, shaker-related subfamily, member 4 (Kcna4)	NM_021275.4	3	1199
Saccharomyces cerevisiae strain CBS5112 Ure2p (URE2)	AF525191.1	418	584
Rattus norvegicus calcium/calmodulin-dependent protein kinase II alpha (Camk2a)	NM_012920.1	1	41
Drosophila melanogaster Adh-related (Adhr), transcript variant B	NM_001032101.2	844	1146
Rattus norvegicus activity-regulated cytoskeleton-associated protein (Arc)	NM_019361.1	1	216
Mus musculus betaPix-b mRNA	AF247654.1	1	303
Mus musculus Bcl-xL	L35049.1	1	242
Saccharomyces cerevisiae Bem1p-interacting protein (BOI1)	L31406.1	1	487
Mus musculus Cx32 gene for connexion (Cx32)	AJ271753.1	7081	7552
Rattus norvegicus gap junction protein, alpha 1 (Gja1)	NM_012567.2	1	196
Aplysia californica egg-laying hormone (ELH)	NM_001204741.1	1	279
Saccharomyces cerevisiae (FLO8)	U51431.1	1	183
Saccharomyces cerevisiae (GIC1)	BK006934.2	222479	222672
Rattus norvegicus, glutamate receptor (Gria2)	NM_001083811.1	1	430
Saccharomyces cerevisiae, G protein coupled receptor (GPR1)	BK006938.2	392058	392457
Drosophila melanogaster grim (grim)	NM_079413.3	1	318
Drosophila melanogaster (hid)	NM_079412.4	1	519
Rattus norvegicus, insulin-like growth factor I (IGFI-R) receptor	M37807.1	416	1355
Drosophila melanogaster Insulin-like receptor (InR)	NM_001144622.2	1	419
Homo sapiens insulin receptor (INSR)	M76592.1	39	575
Gallus gallus jun proto-oncogene (JUN)	NM_001031289.1	1	313
Homo sapiens Sjogren syndrome antigen B (autoantigen La1)	NM_001294145.1	1	498
Rattus norvegicus microtubule-associated protein 2 (Map2)	NM_013066.1	1	102
Homo sapiens methionine synthase (MS)	U73338.1	1	394
Saccharomyces cerevisiae (MSN1)	BK006948.2	99467	99808
Saccharomyces cerevisiae (NCE102)	CP006243.1	806383	806840
Mus musculus NK6 homeobox 1 (Nkx6-1)	NM_144955.2	1	477

Table 2.3 continued

Rattus norvegicus protein kinase C, delta (PKCD)	BC076505.1	1	188
Arabidopsis thaliana 40S ribosomal protein S18 (RPS18C)	NM_117048.4	20	103
Rattus norvegicus neurogranin/RC3 protein (RC3)	U22062.1	4217	4475
Mus musculus ring finger protein 2 (Rnf2, Ring1b)	XM_006529269.3	53	205
Homo sapiens soluble guanylyl cyclase subunit beta 2 (GUCY1B2)	AF038499.2	1	280
Saccharomyces cerevisiae, (YMR181c)	CP005424.2	595521	595819
Rattus norvegicus dendrin (Ddn)	NM_030993.1	1	148
Mus musculus utrophin (Utrn)	NM_011682.4	1	506

Table 2-4 List of global-sensitive and local-sensitive IRES (Weingarten-Gabbay et al., 2016)

Global-sensitive IRES	**Local-sensitive IRES**
NM_001242486:EIF1AD:145:642:1:ER_stress	NCBI_human_RNA_viruses:NC_001722:Human immunodeficiency virus 2 (HIV-2):ss-RNA:6239:6502:Fw
Virus_Human:PV_type_1_Mahoney:Picornaviridae:139	NM_182691:SRPK2:188:2254:1:Rapamycin
NCBI_human_RNA_viruses:NC_011510:Rotavirus A: RNA:10:2340:Fw	Rattus_norvegicus:ODC1::130
NM_032966:CXCR5:289:1272:1:Apoptosis	Human:ELG1:C17orf85:287
NM_001122634:CPS1:510:3659:3:polio,Rapamycin,Rapamycin	Mus_Musculus:NDST2::490
NM_001198625:RUNX1T1:761:2494:3:ER_stress,Apoptosis,IRESite	NM_001193317:VIPAS39:324:1805:1:ER_stress
NM_001195684:TGFBR3:352:2904:2:polio,UVB	NM_153742:CTH:199:1284:2:ER_stress,Rapamycin
Virus_Invertebrate:BQCV_IGRpred:Dicistroviridae:17	Human:AML1/RUNX1:RUNX1:1388
Human:DAP5:EIF4G2:1	Human:DAP5:EIF4G2:132
one_rRNA_element:NM_001008387:Poliovirus_t_2:127:140:CCACACTTCCTTTA:3	NCBI_human_RNA_viruses:NC_006577:Human coronavirus HKU1 (HCoV-HKU1):ss-RNA:206:21753:Fw
NCBI_human_RNA_viruses:NC_001542:Rabies virus:ss-RNA:5418:11846:Fw	Virus_Human:REV-A:Retroviridae:317
NM_001005619:ITGB4:9:5426:1:hypoxia_2007	NCBI_human_RNA_viruses:NC_002728:Nipah virus:ss-RNA:5108:6166:Fw
one_rRNA_element:NM_001004750:Rbm3:109:119:TTCTTGGCAAT:1	Human:UNR:CSDE1:256
NM_001256571:RXRG:307:1329:1:ER_stress	Mus_Musculus:NDST2::577
NM_006022:TSC22D1:297:731:3:ER_stress,ER_stress,Apoptosis	NM_001242927:ZNF410:451:1668:1:ER_stress
Human:FGF1A:FGF1:87	two_rRNA_elements:NM_001118886:2:Rbm3:19:29:TTCTCAGCAAA:2,Rbm3:73:83:TTCCTGCCAAA:2
NM_001267061:SNX17:144:1496:1:ER_stress	Virus_Human:PV_type_1_Leon:Picornaviridae:569
NM_001040110:NRF1:101:1612:2:ER_stress,IRESite	NCBI_human_RNA_viruses:NC_001488:Human T-lymphotropic virus 2:ss-RNA:6:119:Fw
NM_001202404:ALDH7A1:112:1620:1:ER_stress	NCBI_human_RNA_viruses:NC_001802:Human immunodeficiency virus 1 (HIV-1):ss-RNA:5377:7970:Fw

Table 2.4 continued

NM_001114309:ELF3:268:1383:3:polio ,Apoptosis,hypoxia_2007	NM_202001:ERCC1:189:1160:3:ER_stress ,UVB,YFP_lib
Virus_Human:HCV_type_1a:Flaviviridae:210	IRESite_SV40_661-830:NC_001669:Simian virus 40: DNA:562:1620:Fw
Virus_Human:GBV-C:Flaviviridae:370	IRESite_SV40_661-830:NC_001669:Simian virus 40: DNA:1499:2593:Fw
IRESite_REV-A:DQ387450:Reticuloendotheliosis virus: DNA:935:2434:Fw	NCBI_human_RNA_viruses:NC_001906:Hendra virus:ss-RNA:6618:8258:Fw
NM_183422:TSC22D1:492:3713:3:ER_ stress,ER_stress,Apoptosis	Virus_Invertebrate:DCV_IGR:Dicistroviridae:1
Human:UNR:CSDE1:169	Human:NRF:NKRF:377
NM_019101:APOM:74:640:2:YFP_lib, Rapamycin	one_rRNA_element:NM_000742:Random_ ICS1_23:81:94:CACAGAATCCAGCA:3
Virus_Invertebrate:SINV1_IGRpred:Dicistroviridae:1	IRESite_SV40_661-830:NC_001669:Simian virus 40: DNA:916:1620:Fw
Human:eIF4G1:EIF4G1:97	two_rRNA_elements:NM_001005484:2:TEV:132:138:GACTCCC:1,TEV:156:162:TACTTCC:1
NM_001194995:C12orf65:473:973:1:ER_stress	two_rRNA_elements:NM_001005240:2:TEV:132:138:GACTCCC:1,TEV:156:162:TACTTCC:1
Virus_Vertebrates:ERAV_245-961:Picornaviridae:365	Human:Hsnm1:DCLRE1A:658
Mus_Musculus:NDST1::247	one_rRNA_element:NM_001004312:TEV:54:60:TACTCCC:0
one_rRNA_element:NM_001005193:Random_ICS1_23:62:75:CATGGAAGCGAGAA:2	IRESite_LINE-1:AF016099:Mus musculus (house mouse): DNA:1:9278:Rv
NM_001013251:SLC3A2:153:1742:3:ER_stress,UVB,YFP_lib	NCBI_human_RNA_viruses:NC_001906:Hendra virus:ss-RNA:11400:18134:Fw
NM_080918:DGUOK:86:655:2:UVB,Rapamycin	NCBI_human_RNA_viruses:NC_005831:Human coronavirus NL63:ss-RNA:24542:25219:Fw
NCBI_human_RNA_viruses:NC_005079:Machupo virus:ss-RNA:465:7094:Rv	Human:eIF4G1:EIF4G1:184
Rattus_norvegicus:cat_1_224::1	Human:XIAP:XIAP:287
Human:Bcl2:BCL2:790	NCBI_human_RNA_viruses:NC_006577:Human coronavirus HKU1 (HCoV-HKU1):ss-RNA:27373:27621:Fw

Table 2.4 continued

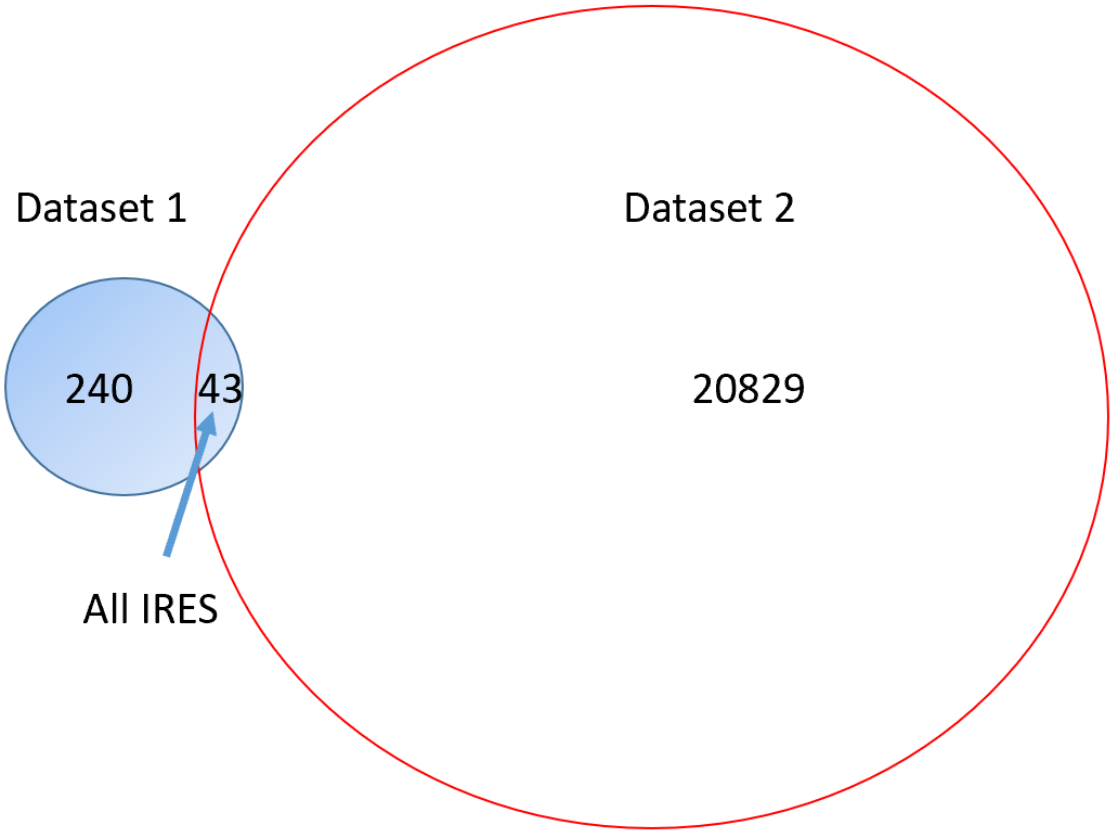
IRESite_ERBV_162-920:X96871:Equine rhinitis B virus 1: RNA:895:8664:Fw	
NCBI_human_RNA_viruses:NC_001653:Hepatitis delta virus:ss-RNA:1632:363:Rv	
NCBI_human_RNA_viruses:NC_001475:Dengue virus 3: RNA:95:10267:Fw	
NM_001127206:HMOX2:172:1122:1:ER_stress	
Virus_Vertebrates:ERBV_162-920:Picornaviridae:499	
NM_001199723:CRABP2:209:625:2:hypoxia_2004,hypoxia_2007	
two_rRNA_elements:NM_001145465:3:Random_ICS1_23:142:155:CAGGAAATCGAGAC:3,TEV:98:104:TAATCCC:1,Rbm3:157:167:ATCTTGGCTAA:2	
NCBI_human_RNA_viruses:NC_002645:Human coronavirus 229E: RNA:24750:24983:Fw	
NCBI_human_RNA_viruses:NC_010810:Human TMEV-like cardiovirus:ss-RNA:956:7837:Fw	
NM_181358:HIPK1:64:2514:3:Apoptosis,Rapamycin,Rapamycin	
NM_001128425:MUTYH:217:1866:1:ER_stress	
Virus_Human:GBV-B:Flaviviridae:156	
NM_001206885:CTNND1:572:3391:1:ER_stress	
NM_013314:BLNK:179:1549:2:ER_stress,Rapamycin	
two_rRNA_elements:NM_001256932:2:TEV:108:114:AACTCCC:1,TEV:129:135:TACTGCC:1	
NM_001007245:IFRD1:471:1826:1:Apoptosis	
Virus_Invertebrate:TrVIGR:Dicistroviridae:1	
Mus_Musculus:Utr4::159	

Table 2.4 continued

two_rRNA_elements:NM_005598:2:TEV:77:83:TCCTCCC:1,TEV:87:93:TCCTCCC:1	
NCBI_human_RNA_viruses:NC_001722:Human immunodeficiency virus 2 (HIV-2):ss-RNA:5898:6239:Fw	
Mus_Musculus:NDST4L::325	
NCBI_human_RNA_viruses:NC_006577:Human coronavirus HKU1 (HCoV-HKU1):ss-RNA:21773:22933:Fw	
NM_001206797:PKM:285:1658:2:ER_stress,hypoxia_2007	
NCBI_human_RNA_viruses:NC_005147:Human coronavirus OC43 (HCoV-OC43): RNA:211:21497:Fw	
Virus_Vertebrates:AEV:áPicornaviridae:147	
NCBI_human_RNA_viruses:NC_005831:Human coronavirus NL63:ss-RNA:20472:24542:Fw	

Table 2-5 Global kmer features with significant differences (Student's t-test) between IRES and nonIRES groups at a significant level 0.05.

"T","TT","TTT","TTTT","CT","CCT","TCT","CTT","TC","TTC","CTC","TCC","TTTC","CTCC","CTTT", "TTCT","TCCT","CTCT","TCTT","CCCT","CCTC","TCTC","CTTC","TTCC","CCTT","TGT","GTT","TGTT","TTGT","TTA","GCC","TATT","CGT","TTAT","TTG","ATT","CTGT","TAT","CCG","CGC","ATTT", "TTTG","TAC","GCTT","TTCA","TTTA","GTTT","ACTT","TCA", "A","AA","AAA","AAAA","G","GG","GA","AG","AAG","GGA","AGA","GAG","GAA","AGG","GGG","AGGA","AAAG","GGAA","AGAA","GGAG","AAGA","GAAG","GAGA","AGAG","GCG","CGG","GAC","CGA","AGT","GGC","TAA","ACG","TGG","TGAG","AGC","AT","TA"



	Dataset1	Dataset2
number of IRES	116	2129
number of non-IRES	167	18743
shared sequence	43	

Figure 2-1 Comparison of sequences in Datasets 1 and 2.

Dataset 2 has increased the number of IRES by more than 10 times. The number of experimental proved nonIRES has been enlarged too. Little overlap between these two datasets.

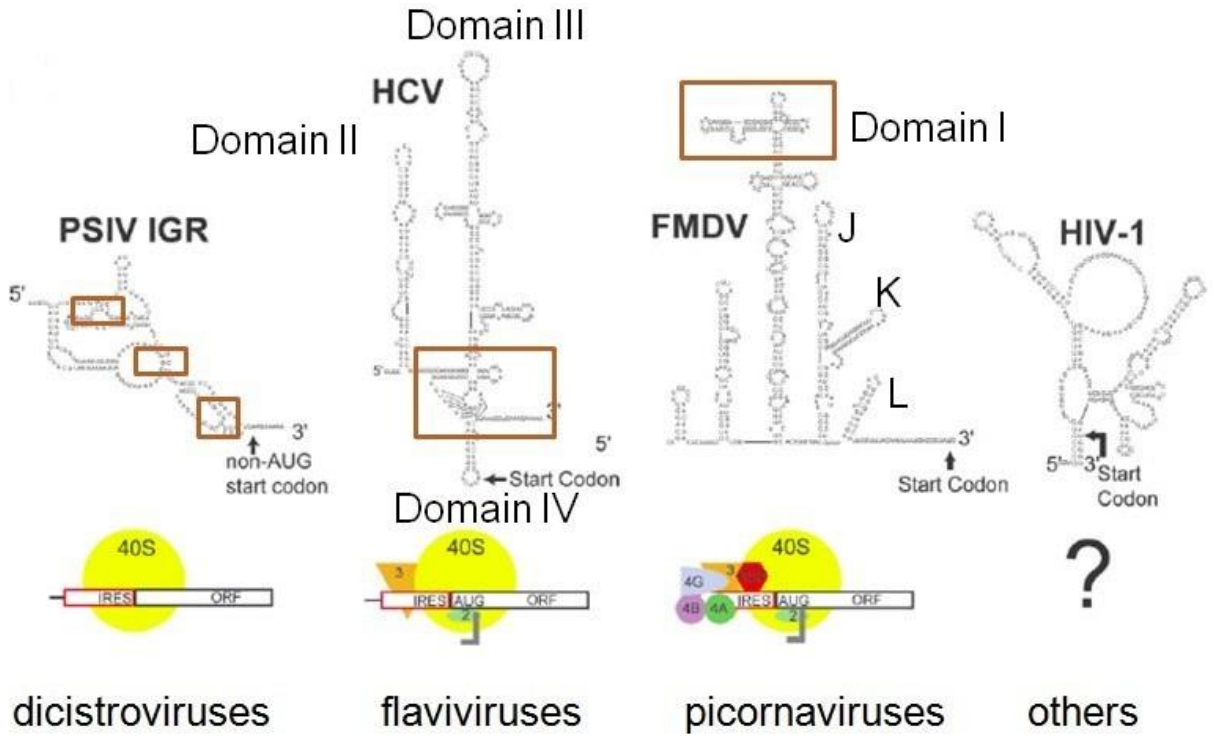


Figure 2-2 Diversity of viral IRES secondary structure. (Plank & Kieft, 2012)

Diversity of viral IRES secondary structure indicates diversity of their mechanism. Conserved regions within each sub-group are shown in brown boxes. Different eIFs and ITAFs play different role in each sub-group.

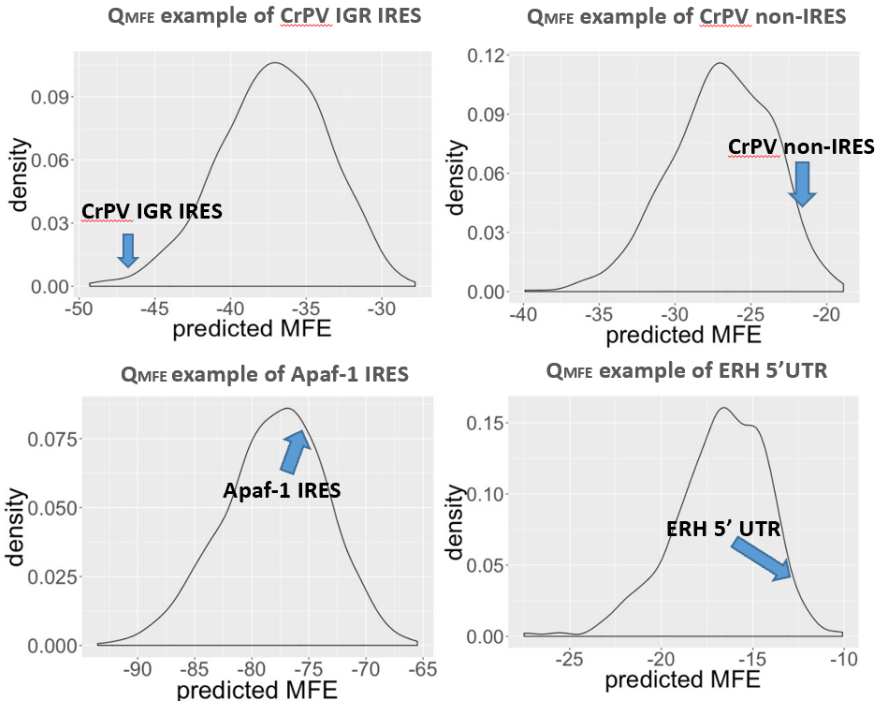


Figure 2-3 Examples of QMFE in four different sequences.

Q_{MFE} is calculated by finding the quantile of the predicted MFE from sequences of interest compared with the distribution of predicted MFE for randomized sequences. Those sequences of interest include CrPV IRES, Apaf-1 IRES, nonIRES segment UTR of CrPV, and ERH housekeeping gene UTR. Sequences were independently randomized 1000 times using Ushuffle the all the predicted MFE was predicted by UNAfold. The distribution plot of those predicted MFE has been showed.

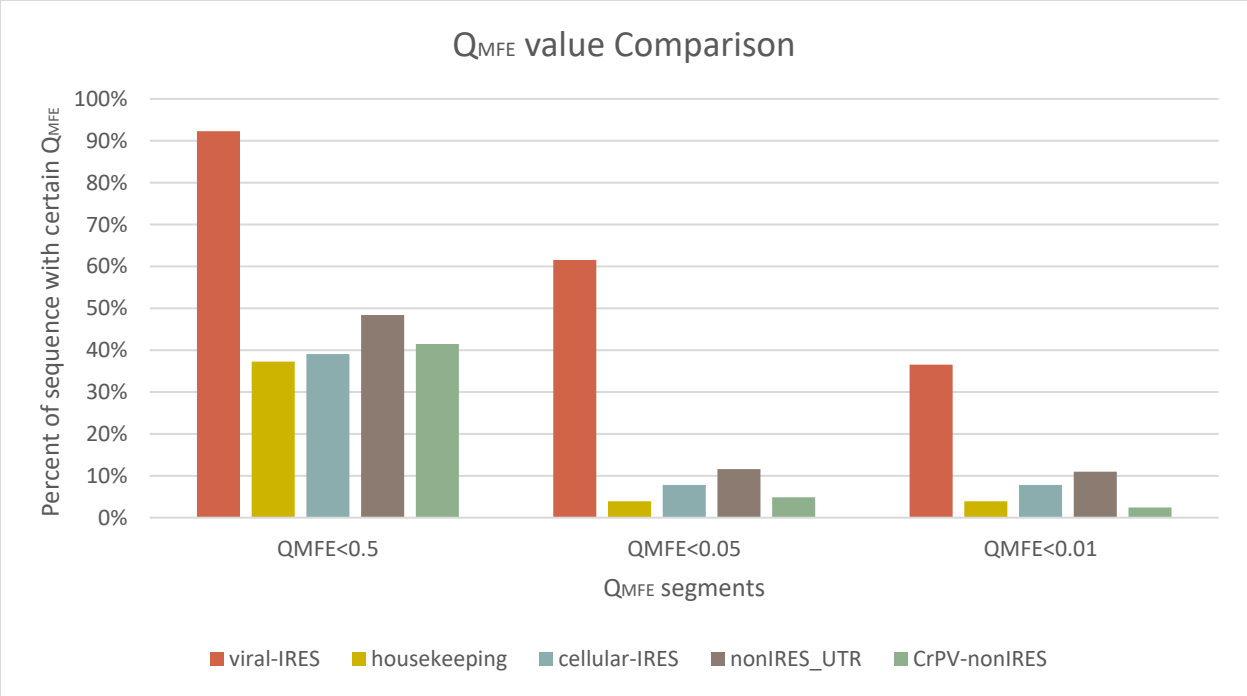


Figure 2-4 Comparison of QMFE in viral IRES, housekeeping genes, cellular IRES, nonIRES UTR and CrPV nonIRES.

The range of QMFE is between 0 and 1 and we have clustered them into three groups: <0.5, <0.05 and <0.01 as X-axis. The Y-axis shows the percent of sequences fall into each group.

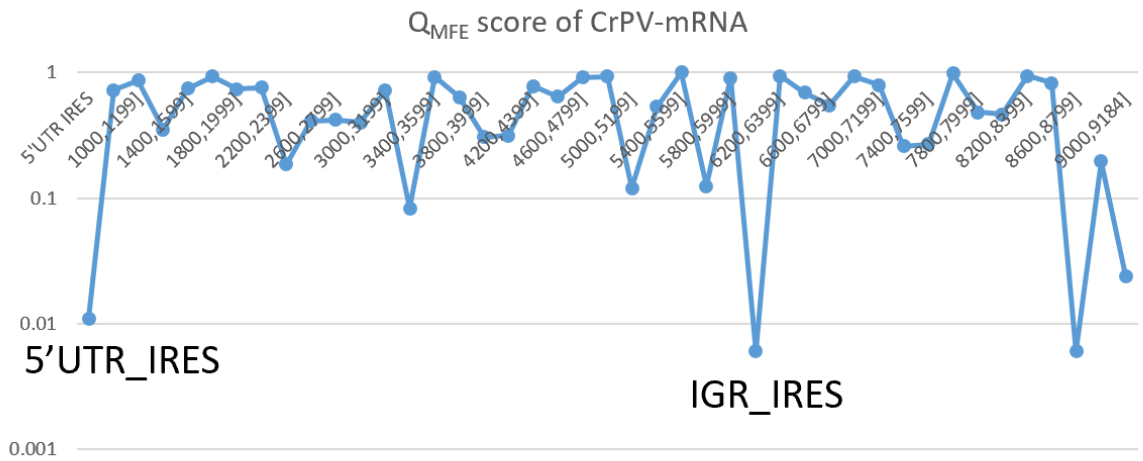


Figure 2-5 Positional analysis of CrPV IRES by QMFE

The Q_{MFE} of every adjacent 200 nt segment is calculated across the mRNA of CrPV. Two of them are exactly the regions where the known the 5'UTR IRES (bases 1-708) and intergenic IRES (6000-6200 bases) are found

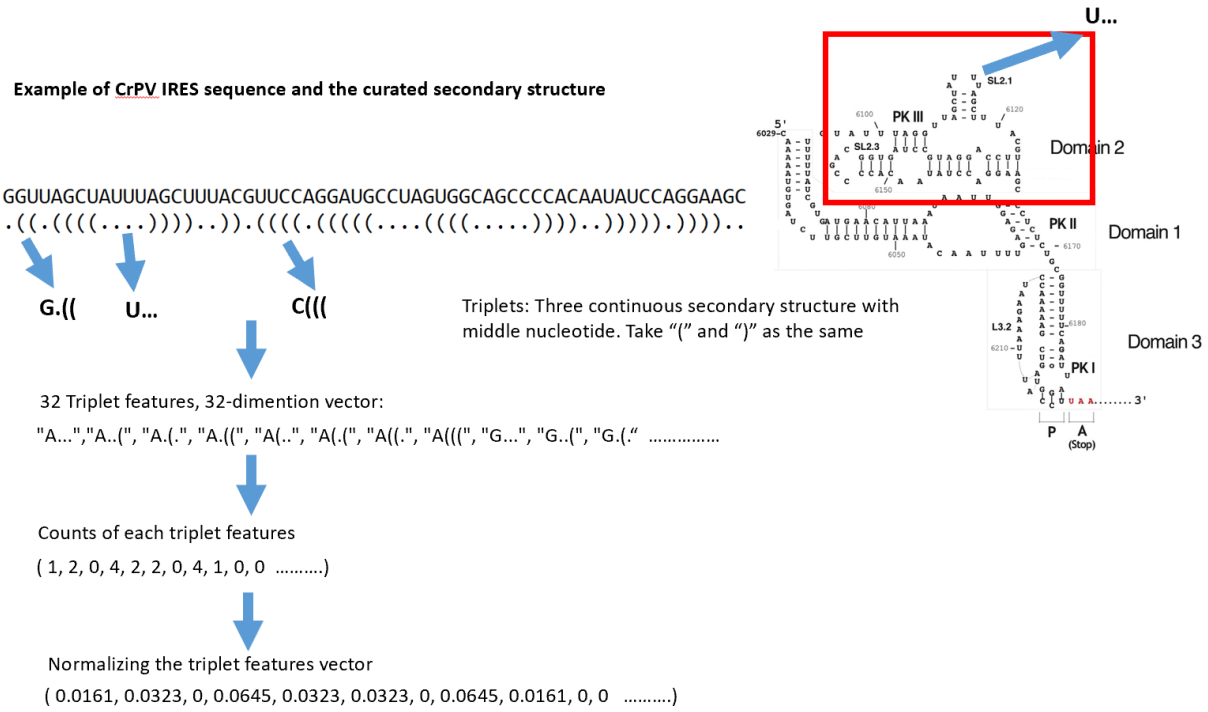


Figure 2-6 Triplet feature calculation.

(1) Calculate the secondary structure of the candidate sequence using UNAfold (Markham & Zuker, 2008). For each nucleotide, only two states are possible, paired or unpaired. Brackets "(" or dots "." represent the and unpaired nucleiotides in the predicted secondary structure, respectively.

(2) For any 3 adjacent nucleotides, there are 8 possible structural states: "(((", "((.", "(..","(.(",".((",".(.","..(", and "…". Triplet features consider the cases where the identity of the central base is known so there are 32 (8*4=32) triplet features in total.

(3) Triplet features are normalized by dividing the observed numbers of each triplet by the total number of all the triplet features.

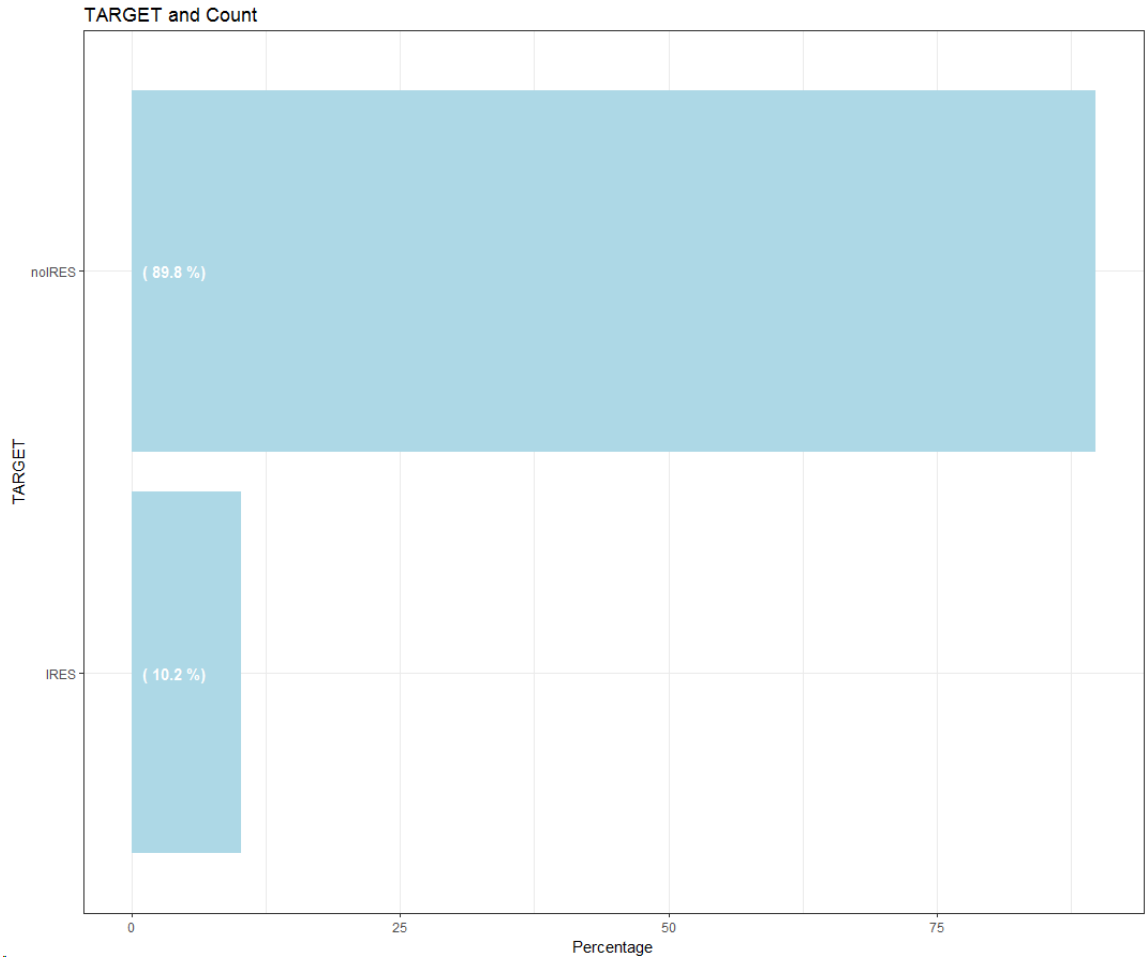


Figure 2-7 The ratio of IRES and nonIRES in Dataset 2.

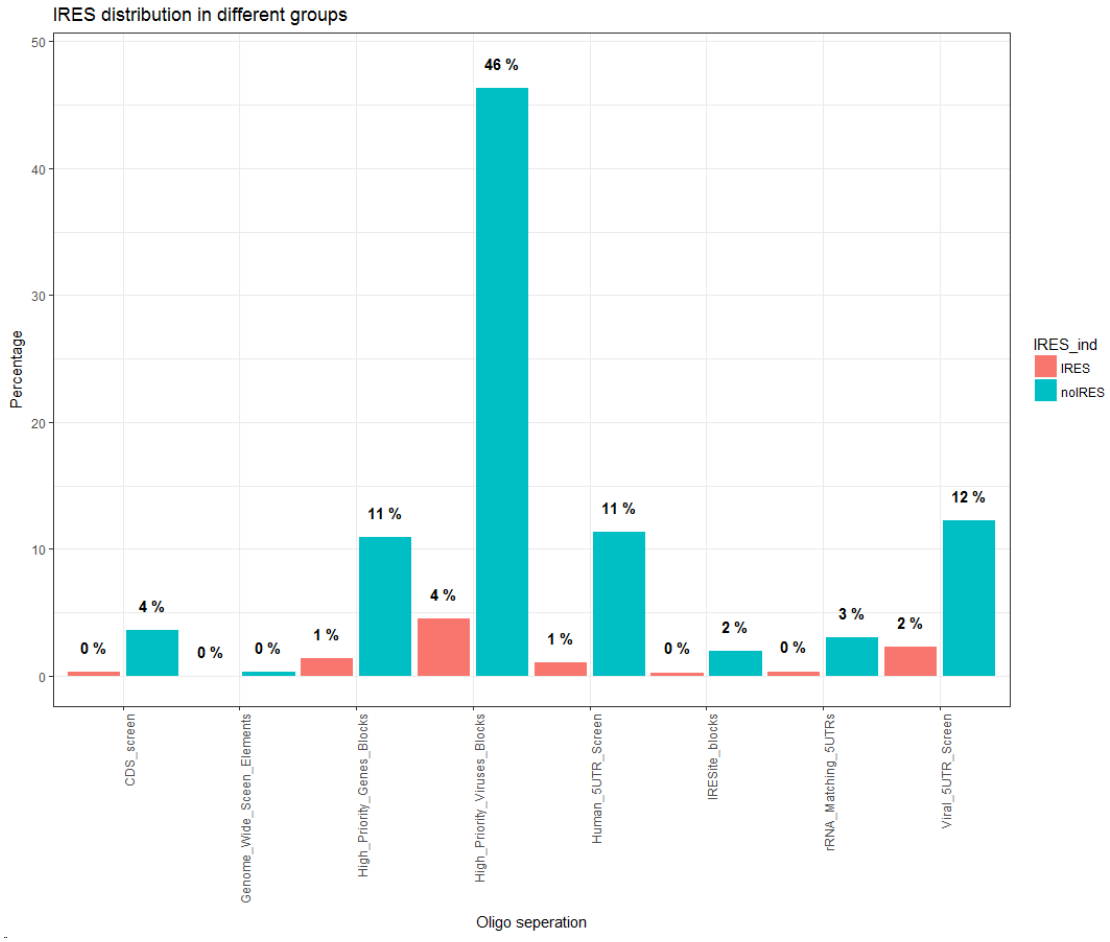


Figure 2-8 IRES distribution in different groups in Dataset 2

The distribution of two labels IRES (red) and nonIRES (blue) is showed among different sequence groups which include: CDS_screen, Genome_Wide_screen_elements, High_Priority_Genes_Blocks, High_Priority_Viruses_Blocks, Human_5UTR_Screen, IREite_blocks, rRNA_Matching_5UTRs, Viral_5UTR_Screen

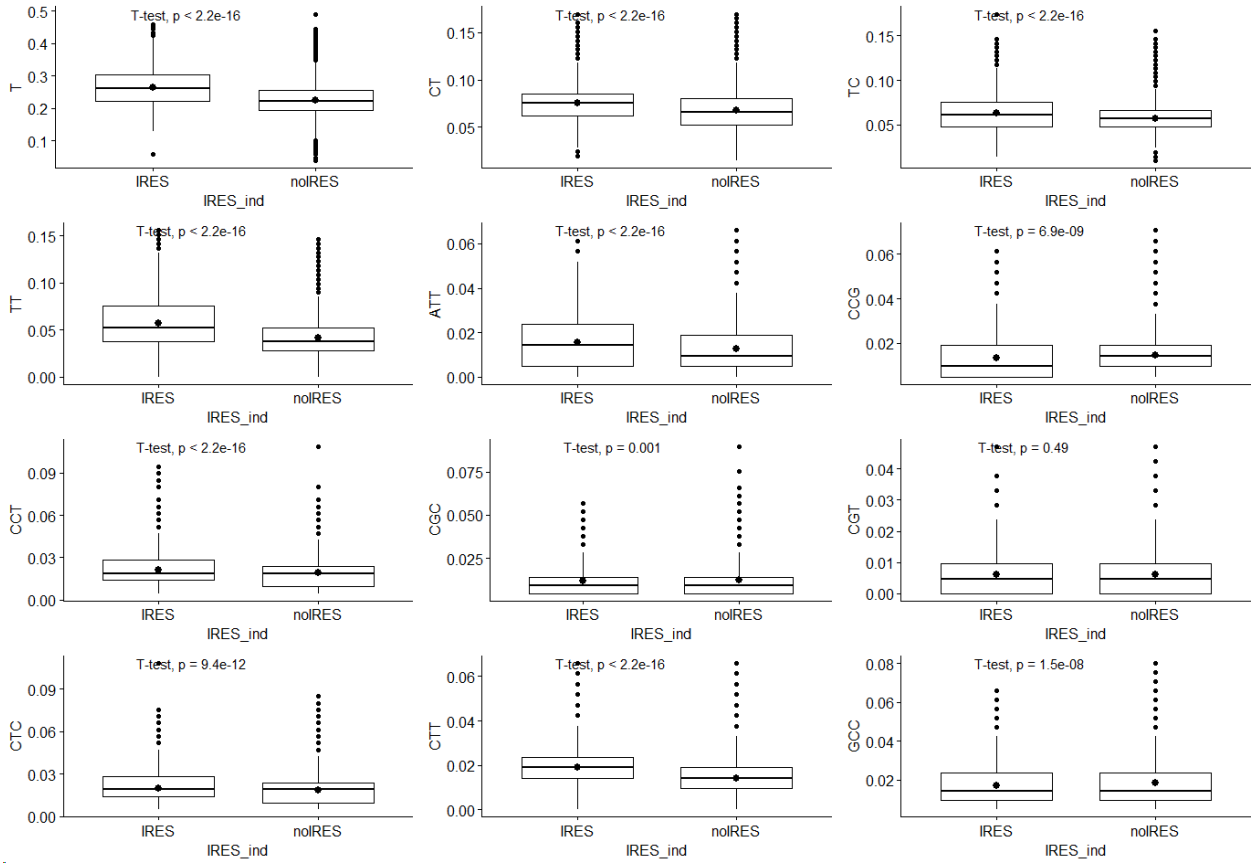


Figure 2-9 Kmer features box plot in IRES and nonIRES groups in Dataset 2 (partial plot)

Box plot of kmer features has been drawn with median, upper and lower quartiles (lines), and outliers (dots). The Student's two sample t-test show the significance with a p-value.

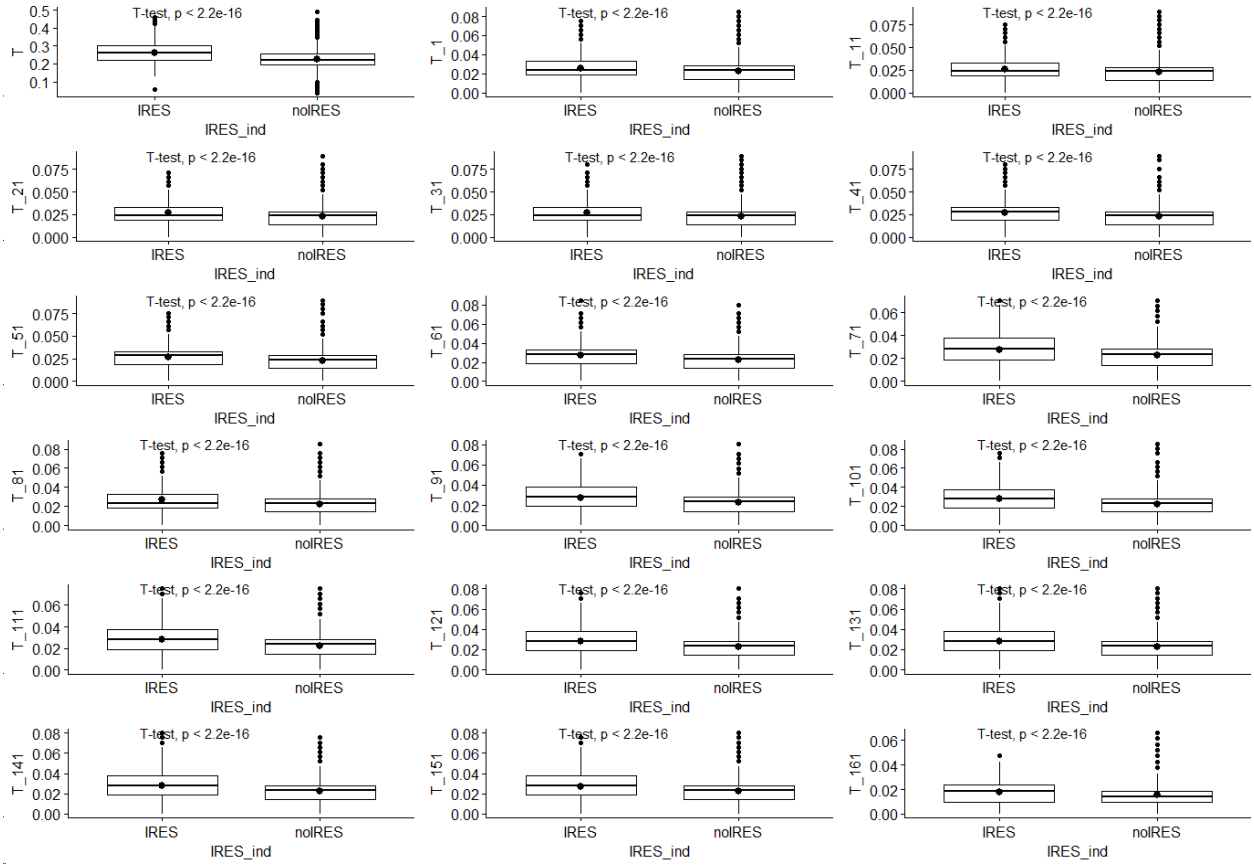


Figure 2-10 Comparison of global and local "T" features in IRES and nonIRES groups in Dataset2

Box plot of "T" global kmer feature as well as its 17 local kmer features have been drawn with median, upper and lower quartiles (lines), and outliers (dots). The Student's two sample t-test show the significance with a p-value. When the global "T" feature is significant, its 17 local kmer features are more likely to be significant.

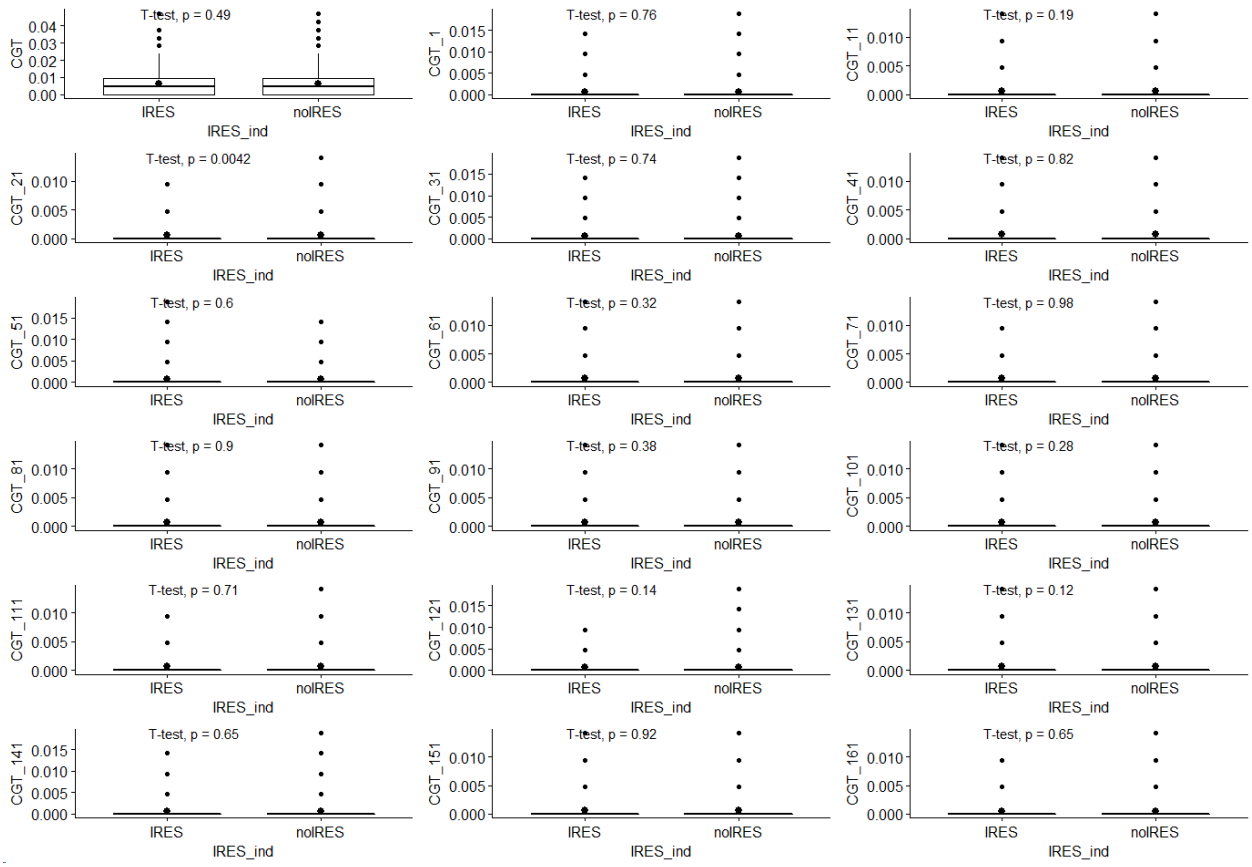


Figure 2-11 Comparison of global and local "CGT" features in IRES and nonIRES groups in Dataset 2

Box plot of "CGT" global kmer feature as well as its 17 local kmer features have been drawn with median, upper and lower quartiles (lines), and outliers (dots). The Student's two sample t-test show the significance with a p-value. When the global "CGT" feature is insignificant, its 17 local kmer features are more likely to be insignificant.

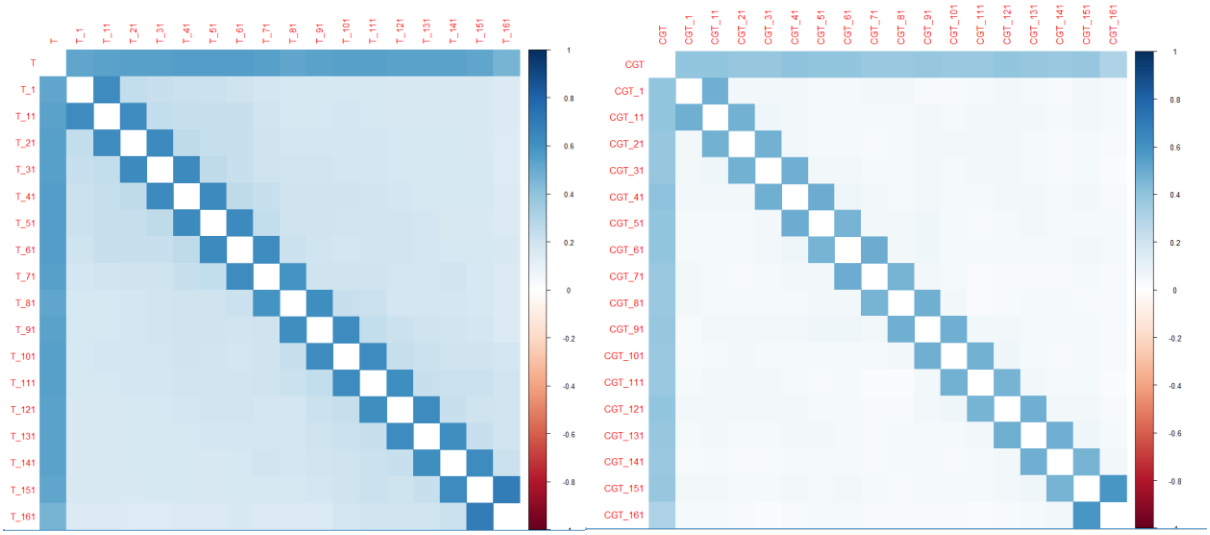


Figure 2-12 The correlation of global kmer features “T”, “CGT” and their local kmer features in Dataset 2.

The correlation of global kmer features “T”, “CGT” and their local kmer features is high (above 0.7). The correlation among local kmer features is low (below 0.4).

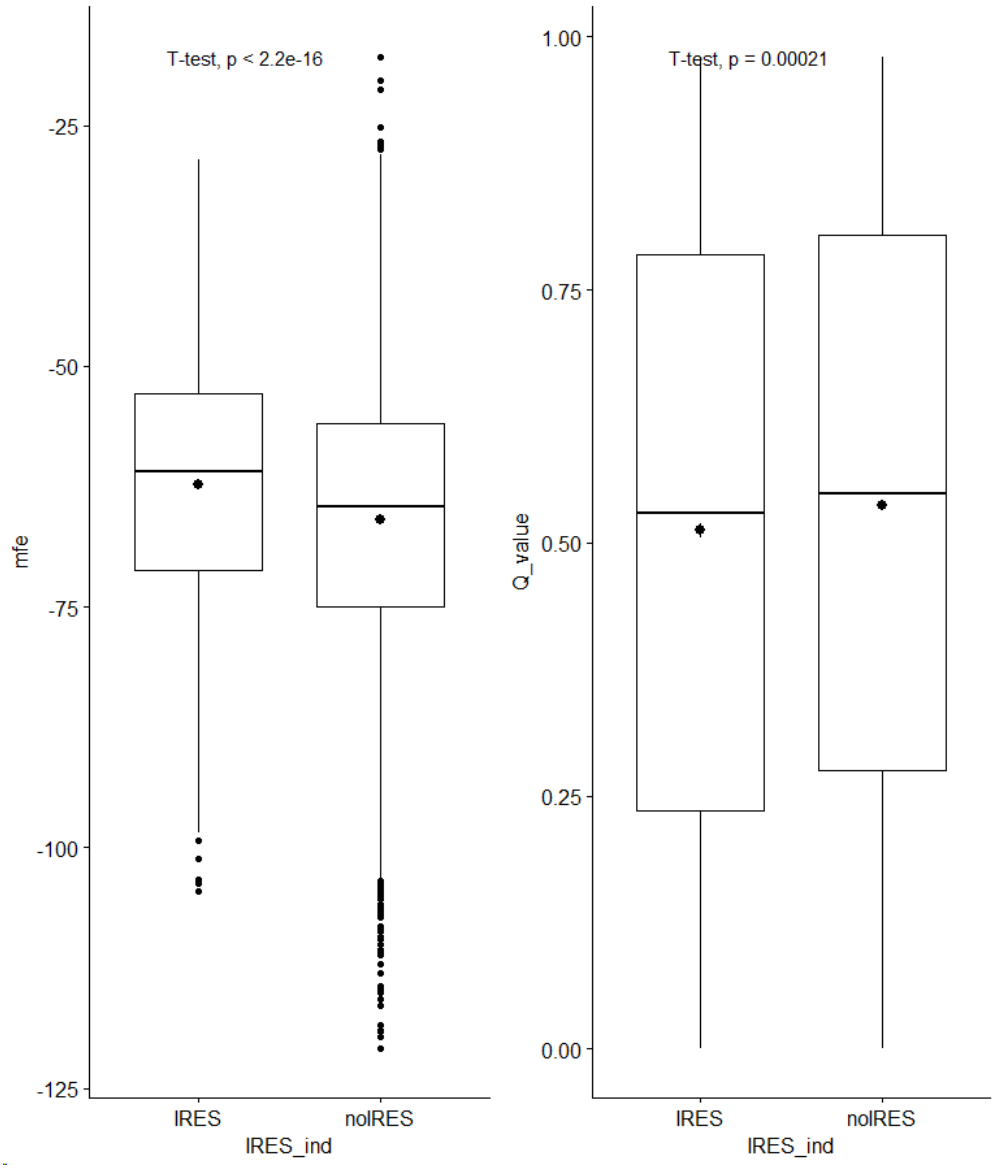


Figure 2-13 MFE, QMFE features in IRES and nonIRES groups in Dataset 2

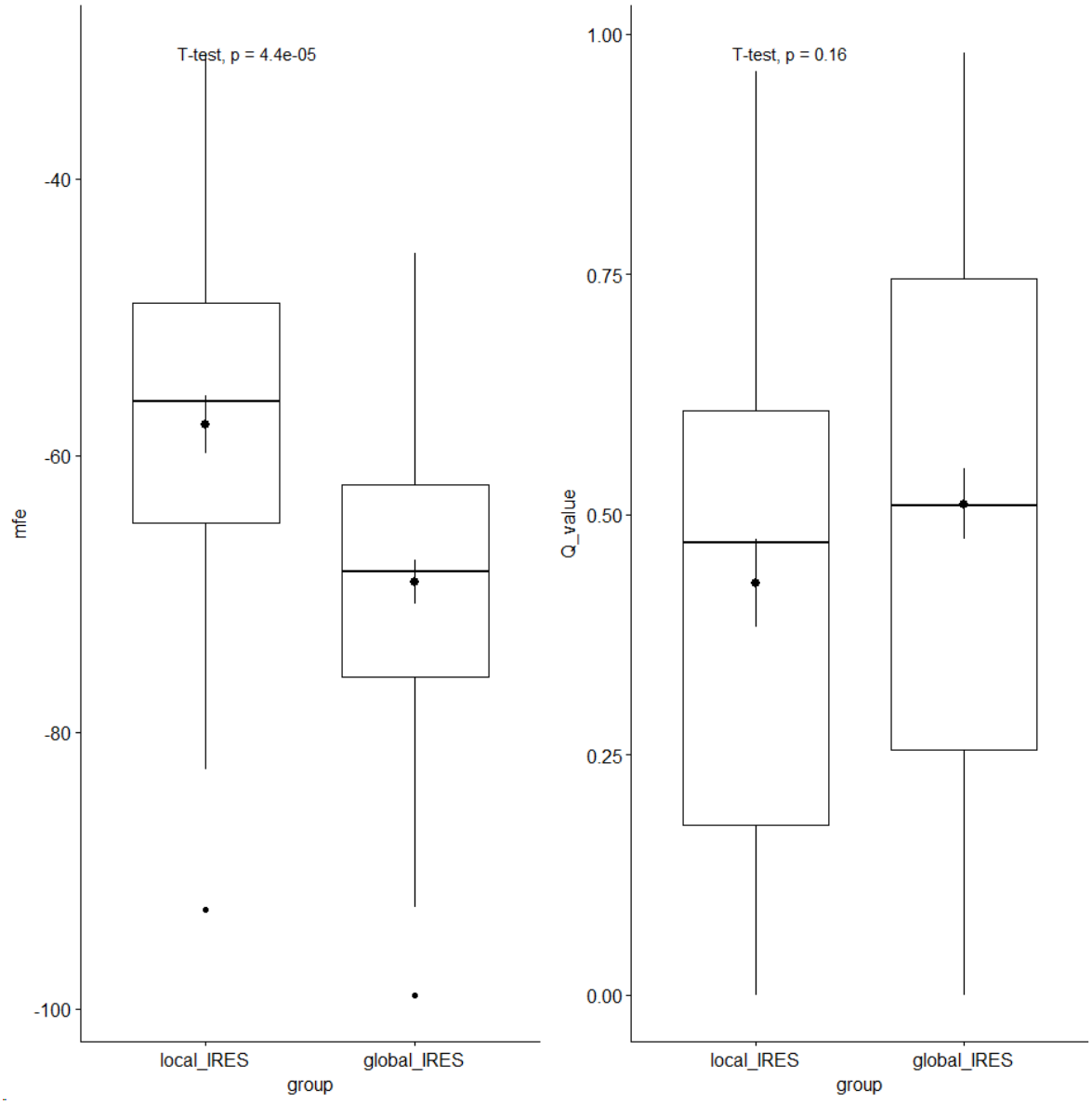


Figure 2-14 MFE, QMFE in global-sensitive IRES and local-sensitive IRES in Dataset 2

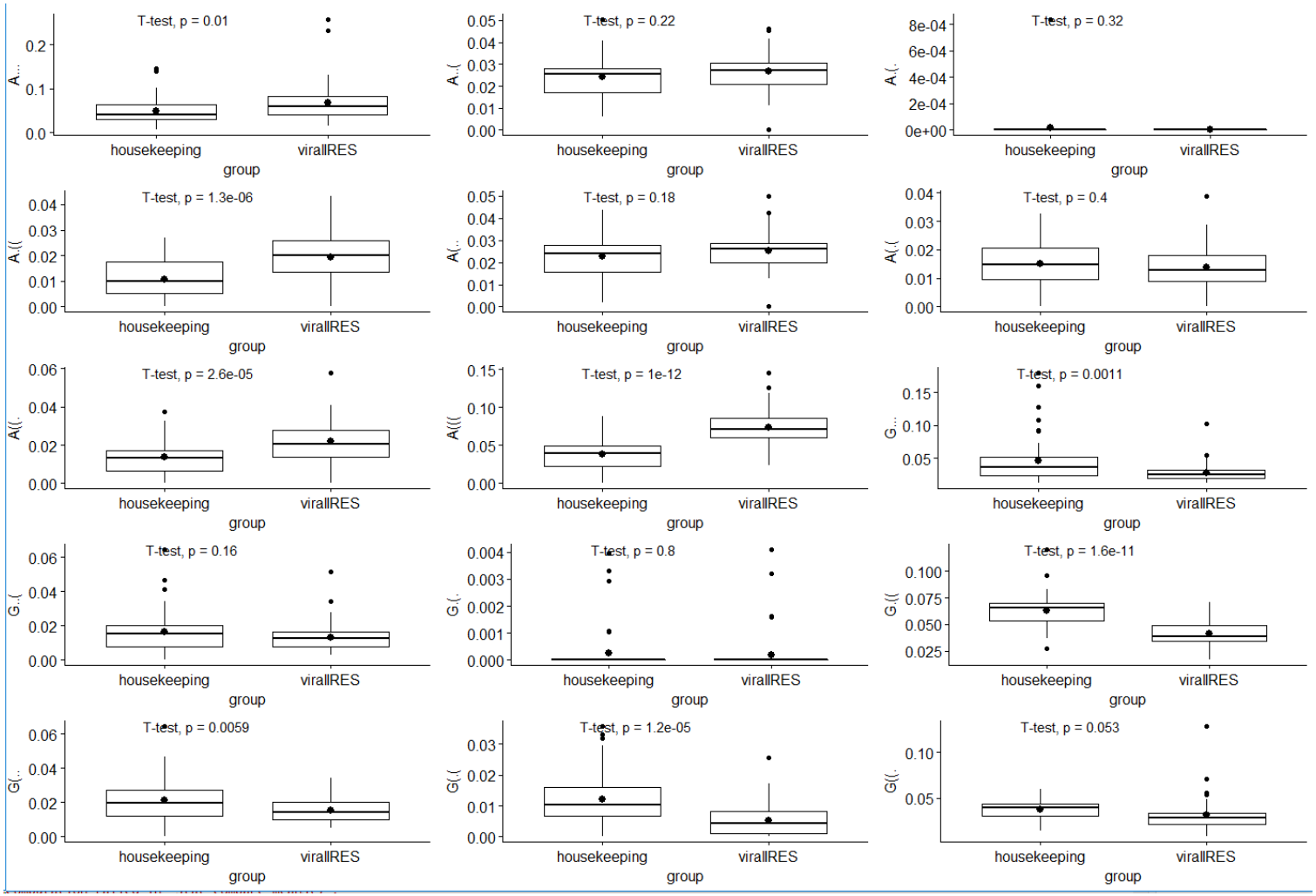
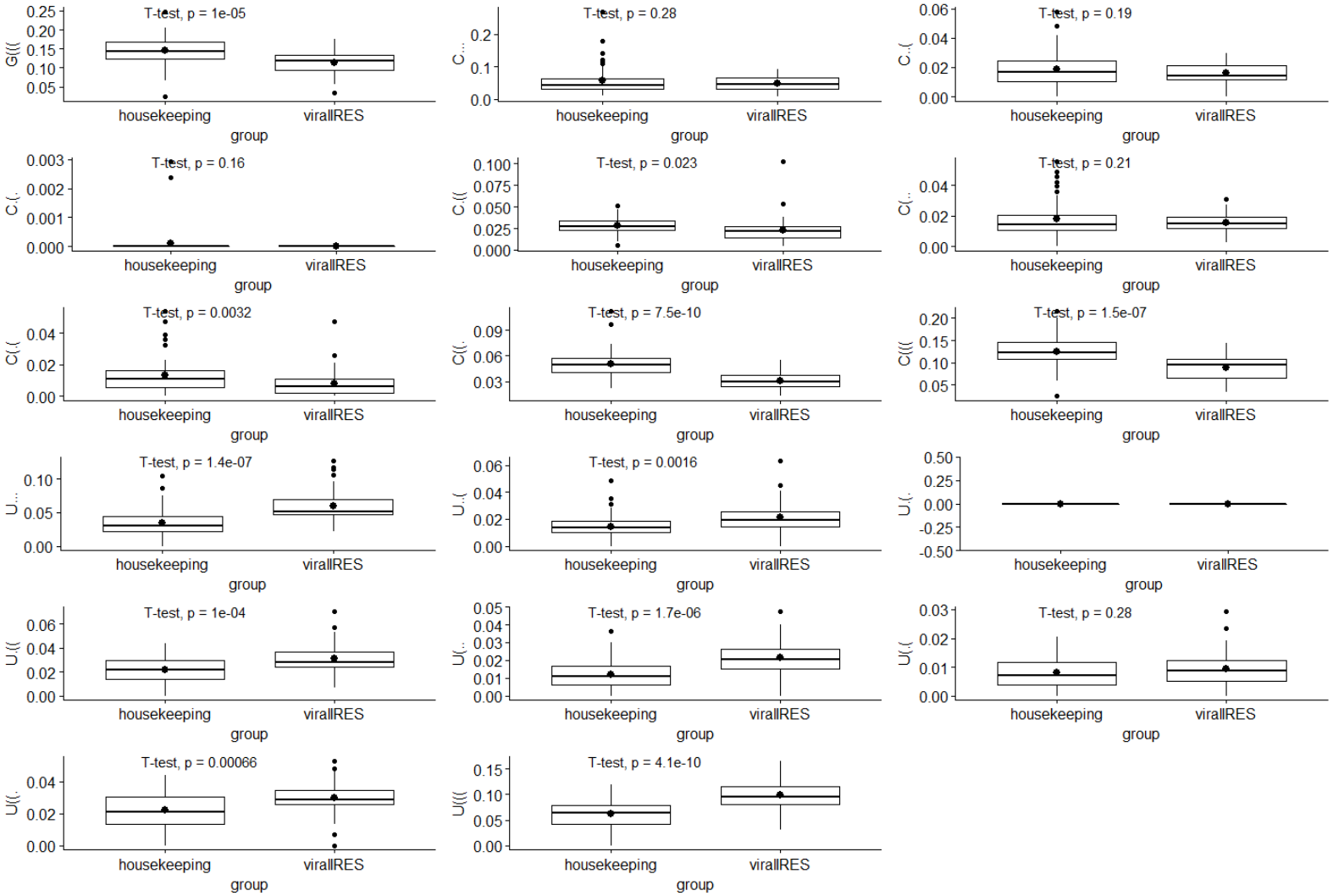


Figure 2-15 Differences in Triplet Features in IRES and housekeeping UTR groups in Dataset 1

Box plot of triplet features in Dataset 1 have been drawn with median, upper and lower quartiles (lines), and outliers (dots). The Student's two sample t-test show the significance with a p-value. 21 out of 32 triplet features show significance under a 0.05 confidence level.

Figure 2.15 continued



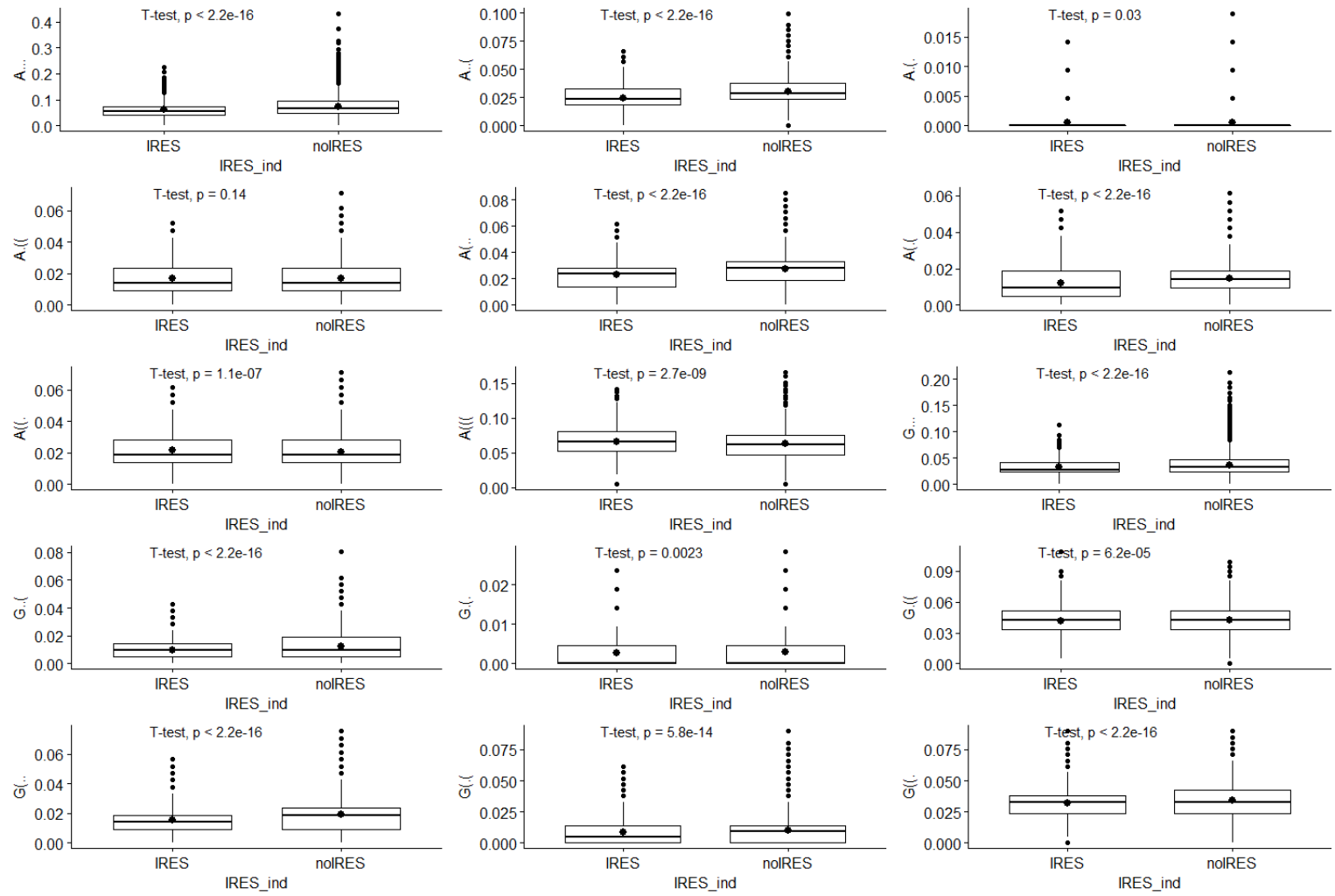
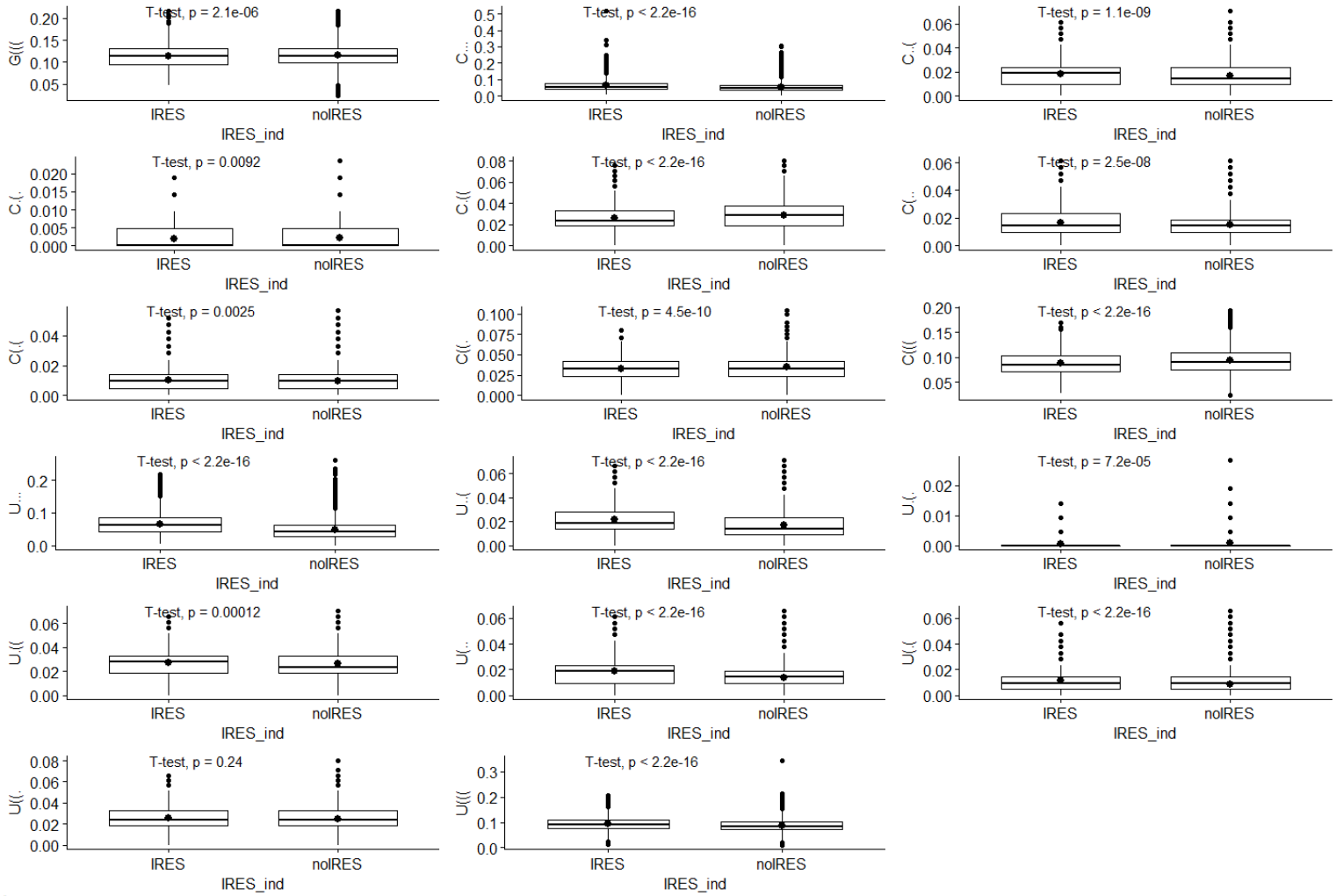


Figure 2-16 Comparison of Triplet Features in IRES and nonIRES in Dataset 2

Box plot of triplet features in Dataset 2 have been drawn with median, upper and lower quartiles (lines), and outliers (dots). The Student's two sample t-test show the significance with a p-value. 30 out of 32 triplet features show significance under a 0.05 confidence level.

Figure 2.16. continued



CHAPTER 3. FINDING INTERNAL RIBOSOME ENTRY SITE BY MACHINE LEARNING

3.1 Introduction

Internal ribosome entry sites (IRES) are segments of the mRNA in untranslated regions which can recruit the ribosome and initiate translation, especially when the conventional translation mechanism has been blocked or repressed. They have been found to play important roles in viral infection, cellular apoptosis, and response to many other external stimuli (Hung et al., 2014; Jo et al., 2008; Sharathchandra et al., 2014; Spriggs, Bushell, Mitchell, & Willis, 2005). The mechanism IRES function not clear, but they are known to function with or without the help of IRES trans-acting factors (ITAFs), which can be small molecules or proteins depending on the different types of IRES.

IRES are widely found in both viral and cellular mRNA. They were first discovered in the poliovirus (PV) and encephalomyocarditis virus (EMCV) RNA genomes in 1988 using a constructed bicistronic assay (Pelletier & Sonenberg, 1988). The assay design by place potential IRES sequence segments between two reporter genes and measures the expression of the reporter gene in comparison to a nonIRES control construct. The bicistronic assay is considered to the best experimental method to confirm the presence of IRES, because the upstream reporter gene's expression can act as a control which is not included in any monotronic assay. However, the defect of this method is that it is very time consuming and labor intensive. In the past 30 years, fewer than 200 IRES have been reported. The most widely used compendium of known IRES is IRESite, which provides a summary of all IRES reported up to 2009 (Mokrejs et al., 2010).

It has been estimated that about 10% of mRNA in both virus and cell can utilize IRES to initiate protein translation (Stoneley & Willis, 2004). The limited number of confirmed IRES

prevents better study and understanding of their function. Researchers have continued trying to find faster and more efficient ways to identify IRES than the bicistronic assay, but with limited success. Comparative analysis of sequences, secondary structures, and tertiary structures of reported IRES has been tried but little commonality has been found across all IRES. Some small motifs have been reported to be shared within specific viral IRES groups, for instance, a GNRA segment has been reported to be shared in picornavirus IRES (Fernandez-Miragall & Martinez-Salas, 2003). However, the absence of universally conserved features across all IRES makes their prediction difficult from the bioinformatics perspective. Instead of using features common to all IRES to determine whether or not a segment of mRNA contains an IRES or not, some machine learning methods, such as support vector machine and random forest models, can use multiple features which are shared by IRES sub-groups to predict the existence of IRES and potentially increase prediction accuracy.

For example, the tool called VIPS predicts the secondary structure of an RNA from its sequence, and uses the RNA Align program to align the predicted structure to forecast IRES (Hong et al., 2013). VIPS predictions are limited to only viral IRES. Although the accuracy rate of VIPS was assessed as 98.53%, 90.80%, 82.36% and 80.41% for four different viral IRES sub-groups, this tool seems to achieve a high prediction accuracy by validating their existing training dataset. Its prediction ability on new finding viral IRES is extremely low. IRESPred, a more recent method, uses 35 features that are based on sequence and structural properties of UTRs, and the probabilities of interactions between UTRs and small subunit ribosomal proteins (SSRPs) to predict IRES (Kolekar et al., 2016). However, due to the limited number of positive IRES, and misleading features such as UTRs' length, number of upstream AUGs, which do not represent the true characteristics of IRES, the performance of their models is not convincing.

Clearly, both the selected features as well as the models are important for predicting the existence of IRES. The main drawbacks of VIPS and IRESPred are the misuse of length dependent features such as the length of UTRs, and the number of upstream AUGs. To overcome these issues, I focus more on the use of an *ab initio* classification approach for IRES prediction. *By ab initio* classification, I mean the use of only the primary sequence and the predicted structure to predict whether or not a sequence contains an IRES. All the features considered here are sequence length independent. The idea of using an *ab initio* classification model to predict the existence of functional RNA is not new. It has previously been applied to predict microRNA precursors (Xue et al., 2005). Chapter 2 discussed the available IRES training datasets and potential sequence length independent features. This chapter will build an *ab initio* classification model to predict IRES based on the training dataset and the sequence length independent features.

In 2016, Eran Segal's group developed a high-throughput IRES activity detection assay, and employed it to identify thousands of novel IRES in human and viral genomes (Weingarten-Gabbay et al., 2016). The identification of many new IRES improves the likelihood that a machine learning model can be successfully implemented. Based on the Segal's dataset, Alexey Gritsenko built a stochastic gradient-boosting random-forest model to predict IRES using 6120 kmer features (Gritsenko et al., 2017). However, their feature set does not consider any structural features, and all features are sequence related. The large feature set leads to the model overfitting and increased computation time. Incorporating of structure related features and better models may achieve higher accuracy and decreased computational time.

Potential features such as kmer words, MFE, $Q_{MFE,}$ and triplet features have been discussed in Chapter 2. In this chapter, the relationship between different features are discussed, and a machine learning model is built based on those features. Machine learning is a tool that can extract

informative knowledge from large scale data. It enables computers to assist humans in the analysis of large, complex datasets. Usually it works by dividing the whole dataset into training and testing parts, referred to as the training and testing datasets, building a model incorporating different features reflecting specific characteristics of the data, and predicting the results on the testing dataset. It is important that the testing and training dataset be independents so that the performance on the testing dataset reflects the probable performance on novel data. The prediction error on the testing dataset thus measures the quality of the model. Different machine learning models including support vector machine, random forest, gradient boosting, and extreme gradient boosting (XGBoost) have been tested on Dataset2 to find a model with improved performance.

The objective of this chapter is to use Dataset2 and all of the previously discussed potential features to correctly classify IRES and nonIRES sequences using machine learning models. The performance matrix of the models will be compared with other available tools to show improvements in accuracy and performance. To implement an improved IRES classifier, I examine dimensionality reduction, feature selection, cross-validation and machine learning model training. Finally, a R-Shiny website toolbox has been created to share the model with the public.

3.2 Materials and Methods

3.2.1 Datasets

Two datasets previously described in Chapter 2, Dataset 1 and Dataset 2, have been used to train the IRES classifiers. The IRES positive group of Dataset 1 mainly comprises sequences extracted from IRESite (Mokrejs et al., 2010), which is the most widely used IRES database. And the negative IRES group is built from nonIRES regions of IRES UTRs, and 5’ UTRs of housekeeping genes. However, the total number of IRES group in Dataset 1 is too small for training many kinds of models. That doesn’t mean one cannot fit a two-class classification model from

only a few positive observations. Some information can always be abstracted from small datasets. A linear model might work better than a more complicated model in such a case. But for IRES classification, using a small positive dataset has several serious drawbacks.

1. Overfitting is more likely with fewer positive training examples. If the ratio of positive data to negative data in the training dataset is too small, the trained model must be more complicated to learn those few positive data, which might cause overfitting.
2. Outliers will be more significant and skewed compared with the limited number of positive observations.

I have chosen to use Dataset2 to develop machine learning methods for the classification of IRES. Dataset2 is the first high-throughput dataset based on a bicistronic assay to detect a large group of sequences with IRES activity. This assay has increased the number of known IRES number by more than 20 times, as well as increasing size of the training dataset, which includes both experimentally confirmed positive IRES and negative IRES, by more than 50 times. The larger dataset makes the use of more complicated machine learning models possible. In total, 28,669 native sequence fragments from Dataset 2 have been used to build the models.

3.2.2 Features

The original feature list included all the potential features discussed in Chapter 2. These features include 340 global kmers and their corresponding local kmer features, structural features such as the predicted MFE, Q_{MFE} and triplet features. The total number of potential features is 6120+32+1+1 = 6154.

Usually, features need to be treated by some selection or transformation to make them better represent the pattern of the data or to meet assumptions of different models. The features

discussed in Chapter 2 have been included in the feature selection process, because 6154 features is too many for efficient model fitting. This high dimensionality can cause overfitting in models, so dimensionality reduction is necessary. There are two major advantages to removing redundant features. One is that reducing the number of features decreases computational time and complexity. The second that it reduces the danger of overfitting. I use the following four approaches to select features:

1. Remove variables which have close to zero variance, because such variables provide almost no positive information for classification.
2. Remove highly correlated features. Multicollinearity is a severe and common problem when there are many features that have high correlations with each other. The correlation matrix between all features has been calculated and any variable pair with correlations greater than 0.75 removed.
3. Variables with high skewness (right skewed or left skewed) are transformed. Transformation methods such as centering, scaling, and box-Cox were considered. Centering and scaling are the most common ways to remove skewness. Typically, skewed variables are transformed into standard normal deviates by subtracting the mean and dividing by the standard deviation. The box-Cox approach uses maximum likelihood estimation to determine the best transformation, including log, power, square root, inverse, and other transformations.
4. Principal component analysis (PCA) is often used to reduce the number of predictors, by linearly combining the variables to form principal components

(PCs). Typically, a limited number of PCs can capture the bulk of the variability of the predictors.

3.2.3 Dataset splitting

The total available dataset must be split into training and testing datasets. In that way, that model training can be monitored in order to achieve the high accuracy without overfitting. An overfitt refersto a model performs well on the training data, but cannot successfully classify novel testing datasets (Olson, Cava, Mustahsan, Varik, & Moore, 2018). Overfitting is usually a problem with nonparametric and nonlinear models that comprise many parameters and are, in effect, able to learn the noise in the training dataset. The noise obscures the true pattern of the training data and degrades the performance of the model on novel datasets.

There are many ways to reduce overfitting. Holding out a testing dataset is one way to tune the parameters trained in the training dataset while making sure that the final trained model yields the best and most realistic results. Because the testing data is not included in the training dataset, overtraining does not have any influence on prediction performance on the testing dataset. Penalty parameters (regularizes) that act to reduce overfitting are also included in some models. For example, the gamma value, which is used as a Lagrangian multiplier in the XGBoost model, is one such regularization parameters used to reduce overfitting. Similarly, parameters such as maximum depth, minimum child weight parameters in any other all tree-based models.

In general, it is desirable to keep the training and testing dataset as homogeneous as possible. Usually, random splitting is the most straightforward way to divide the total dataset into testing and training sets. But if the whole dataset is small, it is easy to obtain a random but heterogeneous split. Or if the ratio of the classes is very skewed, the imbalanced dataset can lead to imbalanced separation. The strategy used to split the whole dataset into training and testing

partitions is therefore a serious issue. Stratified random sampling, in which random sampling is applied within each class of the labeled response is one approach that provides homogeneous testing and training datasets. The whole dataset can also be split by maximum dissimilarity sampling applied to each feature.

Resampling methods and repetition are necessary in dataset splitting to achieve more accurate learning. K-Fold Cross-validation is a resampling method in which the samples are equally divided into k sets; one set is used as the testing partition, and the remainder used for training in each run. In successive runs, different partitions are held out for testing. The best fitt parameters are summarized in the end to generate the final model. K-fold cross validation guarantees that each data point will be used as training K-1 times, and the total data population will be used. Leave-one-out cross-validation (LOOCV) is a special case of K-Fold where K is the number of training examples.

Repetition is another way to fully exploit the dataset. Basically, the dataset is randomly split into training and testing datasets multiple times. Bias in the fit parameters is reduced by increasing the number of subsets. Bootstrapping is a third method, in which random sampling with replacement is used to generate testing and training datasets.

Different sampling methods might be chosen for different sample sizes and different training objectives. If the goal is to compare the performance of different models on the same dataset, bootstrapping might be used because it typically has lower variance than k-fold cross validation. If the sample size is small, repeated k-fold cross validation should be considered. However, if the total dataset is too large, computational time will become a priority rather than model performance. In this case, a small k and repetitions might be tried to achieve good

computational efficiency, because the difference between different resampling trials is small. K-fold cross validation with repeat has been used for the data in this research.

3.2.4 Machine Learning Methods

Machine learning methods can be grouped into supervised and unsupervised approaches depending on whether the target classes are known in advance (labeled). They can also be divided into regression models and classification models based on whether the target classes are numerical or categorical. From this point of view, IRES prediction is a supervised classification problem. There are many well studied supervised regression and classification models available for this situation such as logistic regression, support vector machine, artificial neural network, and decision tree-based models.

The gradient-boosting decision-tree model (GBDT) is a derivative of the random forest algorithm. Unlike random forest, which assigns equal weights to individual trees, boosting grows trees using the information from previous training rounds. It slowly learns from data and improved its prediction rate by assigning different weights to the trees that have more mismatching. A summary of the GBDT method is shown in Figure 3. The Gradient-boosting decision-tree model (GBDT), which has been used in previous research (Gritsenko et al., 2017), was selected as the base model in order to see whether the model performance could be improved by incorporating additional features, feature engineering, and employing a more efficient training algorithm..

Extreme Gradient Boosting (XGBoost) is a tree-based boosting model that improves on standard stochastic-gradient boosting models. In the application of XGBoost to the Higgs-1M data, XGBoost runs more than 10X faster than a gradient-boosting decision-tree model (Torlay, Perrone-Bertolotti, Thomas, & Baciu, 2017). XGBoost can be more efficiently parallelized, and incorporates regularization and tree pruning. First, and maybe the most important, is that the

parallel implementation of XGBoost provides much higher training speed. Second, XGBoost incorporates regularization, which reduces overfitting by setting up the gamma parameters. Third, XGBoost allows for tree pruning, which prevent missing a positive loss following a negative loss. Gradient boosting is a greedier algorithm, because it stops splitting nodes when a negative loss happens. In contrast, XGBoost continues splitting until a predefined max_depth is reached, even though a negative loss occurs. After splitting is completed, XGBoost goes back and prunes the trees. Fourth, the objective function of XGBoost is estimated using a Taylor expansion which contains both first and second derivative terms, and it supports user defined objective functions. However, GBDT only uses of first derivative.

3.2.5 Model evaluation

For a classification problem, the most straightforward evaluation is the confusion matrix, which lists the numbers of the target examples assigned to the possible classes by the model. Statistical measures such as recall (sensitivity), precision (specificity), MCC, and ACC, which can be calculated from the confusion matrix are commonly provided. I define the following measures:

$$\text{Recall}=\frac{TP}{TP+FN} \text{ ; Precision}=\frac{TP}{TP+FP}$$

$$\text{MCC}=\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \text{ ;ACC}=\frac{TP+TN}{TP+FP+FN+TN}$$

Where TP: number of true positives; TN: number of true negatives

FP: number of false positives; FN: number of false negatives

In the best scenario, the final model would have both high precision and high recall. However, there is always a variance-bias trade off. More complex models generally have a higher variance due to overfitting, while simpler models typically have a higher bias because of

underfitting. Simultaneously maximizing both recall and precision is difficult. A balance needs to be achieved, and a statistic that can evaluate this trade-off must be used. Receiver operating characteristic (Rocchi et al.) curves are an example of such a statistic. ROC curve plots the false-positive rate as the x-axis and the true-positive rate (recall) as the y-axis across a continuum of thresholds of a proxy statistic. In this case, the proxy statistic is the prediction made by the trained model. The area under the ROC curve, or AUC, is the probability that a randomly selected positive will have a higher score than a negative – equivalent to a Wilcoxon test. The AUC of a random guess model is 0.5. So the range of AUC is usually 0 to 1. Within that range, a higher AUC indicates a better model.

3.2.6 R-Shiny Toolbox Website Design

The website is available at: https://irespy.shinyapps.io/IRESpy/

And the interface is shown in figure 3.9.

3.3 Results and Discussions

3.3.1 Classification of IRES by kmer features

Gritsenko et al. (2017) implemented a gradient boosting decision tree model based on the Dataset 2, with a testing AUC = 0.77, and a training AUC close to 1.0 (Gritsenko et al., 2017). This model has been taken as the base model and we attempt to improve the performance of this base model. Efforts such as changing the model, adding more features, and feature selection which have been discussed in Chapter 3.2 have been tried. In this work, I focus on the XGBoost model because of its incorporation of regularization functions, and fast fitting speed. From Figure 3.2, XGBoost requires 75% less training time, but improved AUC by 5% compared with gradient boosting, without any hyperparameter tuning.

With the same features as the base model, but different model and different parameter tuning, the XGBoost model can reach a testing AUC of 0.793 and training AUC 0.947. This is better than the base model AUC of 0.77, and training AUC of 1.0 (Figure 3.3). Parameter tuning using grid searching for the best combination of maximum number of threads, step size shrinkage, minimum loss reduction, maximum depth of the decision trees, minimum sum of instance weight, maximum delta step allowed for each leaf, and subsample ratio of columns. This improvement shows the XGBoost model works better than the gradient boosting decision tree (BGDT) model, with all kmer features, on Dataset 2.

Important features ranking (Figure 3.4). shows that "T" kmer group is very important for accurate classification, which is consistent with the previous finding (Gritsenko et al., 2017). Global kmers comprise of 50% of the top 20 most important features and 60% of the top 10 most important kmer features.

To test the importance of global kmer features, the model was run with the same parameter settings, but incorporating only global kmer features. The testing AUC is still 0.771 and a training AUC 0.911. This model achieves the same performance as the base model but requires many fewer features. The total number of features has been decreased by 94.12% and their importance ranking plot is in figure 3.5.

3.3.2 Classification of IRES by sequence features and structural features

Structural features, including MFE, Q_{MFE}, and sequence-structure triplets, which have been discussed in Chapter 2, have been tested to see whether they contribute to model improvement. The combination of global kmer features and structural features increases the testing AUC to 0.775. The importance ranking plot, figure 3.6, shows that Triplet 25 "U…", the MFE, Q_{MFE} value, and Triplet 01 "A…" are important.

In another test we examined a model incorporating only structural features without any kmer features. In this case, the testing AUC is 0.741 and the feature importance plot is shown in figure 3.7. The high AUC indicates that structural features alone can still capture most of the information contained in the kmer features, and the number of features can be decreased from 340 to 35.

To better test the importance of structural features, the labeled target classes have been changed to global sensitive IRES and nonIRES. In this case, the testing AUC is 0.790, which is very high considering only 35 features have been incorporated. The feature importance ranking plot is shown in figure 3.8.

To sum up, I have successfully explored some the addition of structural features such as MFE, Q_{MF}, and triplet features, to random forest models to classify the positive and negative IRES samples. The models built by XGBoost can achieve a higher performance compared with the previous report by 5% increase.

3.3.3 Building R-Shiny Toolbox Website

The website is available at: https://irespy.shinyapps.io/IRESpy/

The interface is as figure 3.9.

3.4 Conclusions

This chapter has discussed the work flow of machine learning and algorithms such as gradient boosting decision tree and XGBoost. It also discussed the utilization of all the potential features which were discussed in chapter 2 to predict the existence of IRES. I have reached several conclusions:

1. XGBoost performs better than gradient boosting decision tree model on Dataset 2. Using the same datasets and features but switching from gradient boosting decision tree model to XGBoost, increases the testing AUC by 5% and the decreases the computational time by 75%.
2. Global kmer and local kmer features are highly correlated. Using only global kmer features in the XGBoost model achieves the same model performance as the gradient boosting decision tree base model. But the total number of features has been rapidly decreased by 94.12% which means the huge computational time decreasing.
3. Incorporation of structural features such as MFE, Q_{MFE}, and triplets improve model performance. The combination of global kmer features and structural features increases the testing AUC compared with that of global kmer features alone. Using structural features alone can still achieve a relatively high model performance. At the same time, the number of features can be greatly decreased, from 340 to 35.
4. Structural features achieve better separation of global-sensitive IRES and nonIRES than the separation of IRES and nonIRES. This is because global-sensitive IRES contain more folded structure than nonIRES sequences, and structural features such as MFE, Q_{MFE}, and triplets better represent this difference. Those significant structural features might better explain the relationship of IRES structure and function.
5. The first high-throughput testing bioinformatics online tool IRESpy has been released to predict IRES. This website provides a public tool for all the IRES researchers in the world.

Table 3-1 Glossary of machine learning terminology.

Terminology	**Explanation**
Example	One data example used in a machine learning model.
Feature	A single measurement or descriptor of an example used in a machine learning model.
Label	The target of a prediction task. In classification, the label is discrete (e.g., “IRES” or “nonIRES”).
Supervised algorithm	Machine learning algorithm that is trained on labeled examples and used to predict the label of unlabeled examples.
Unsupervised algorithm	Machine learning algorithm that does not require labels, such as a clustering algorithm.
Overfitting	A common pitfall in machine learning analysis where a complex model is trained specifically to the training data and resulting in poor performance on new data.
Feature selection	The process of choosing a subset of features from the total available features
Cross validation	A resampling method in which the samples are equally divided into k sets; one set is used as the testing partition, and the remainder used for training in each run.

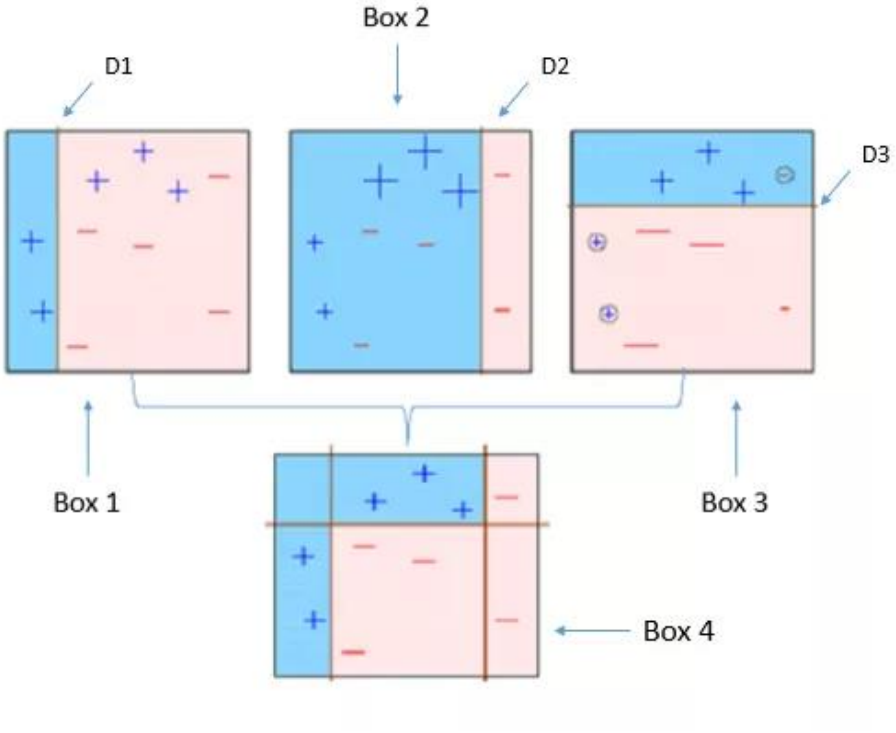


Figure 3-1 Difference of Triplet features between viral IRES and housekeeping genes. (adapted from Chen et al., 2016)

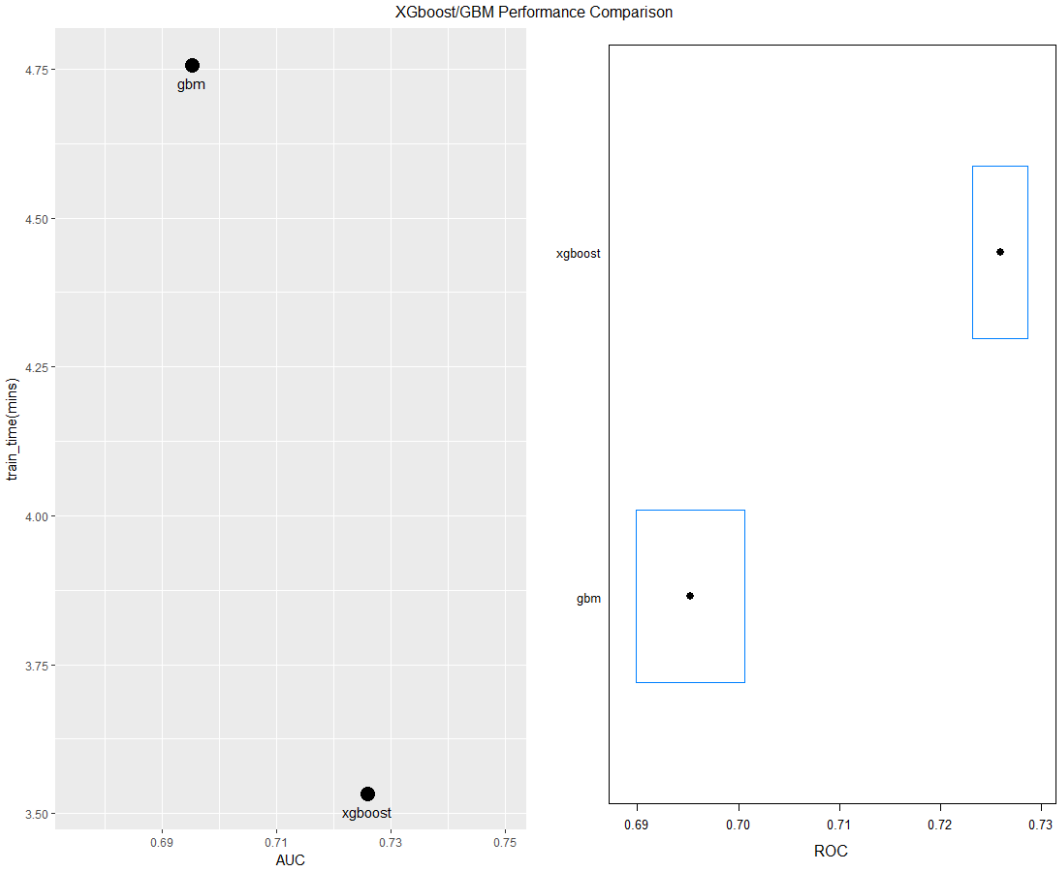


Figure 3-2 Model performance comparison between XGBoost model and GBDT model on Dataset 2

XGBoost requires 75% less training time, but improved AUC by 5% compared with gradient boosting, without any hyperparameter tuning (left panel). The right panel shows the box plot of AUC values with a 10-fold cross validation. We can see XGBoost can achieve better model performance with less computational time than GBDT.

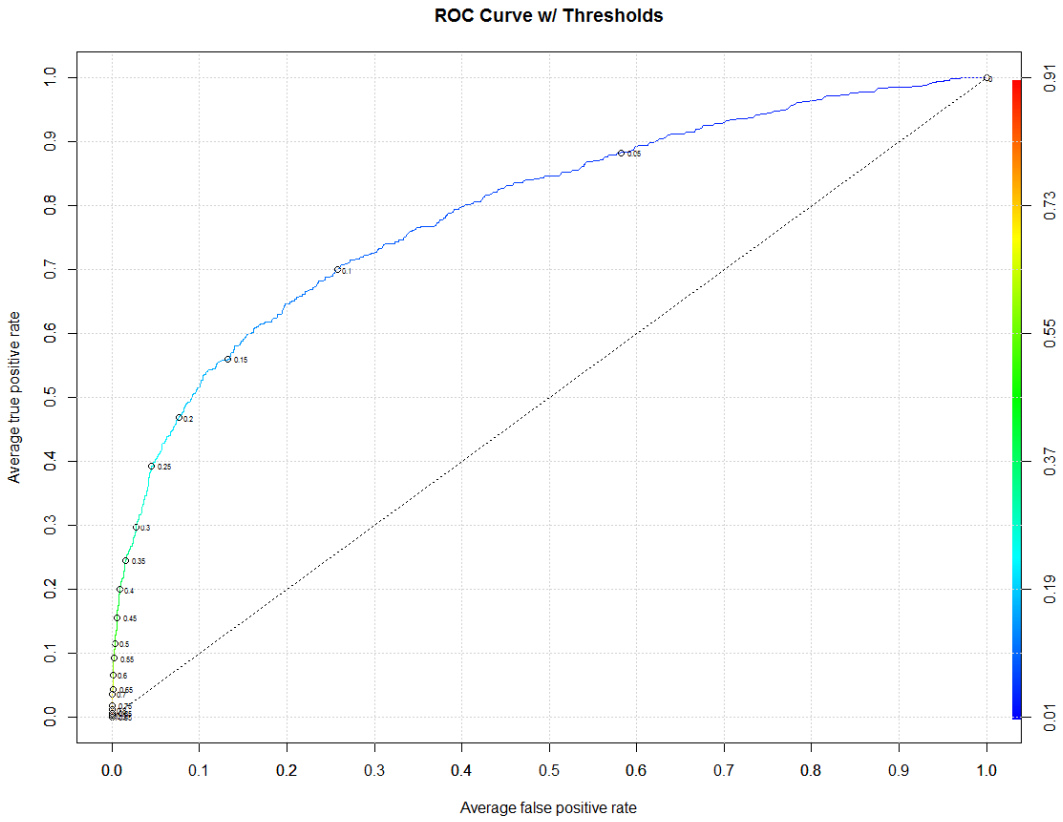


Figure 3-3 ROC curve using the combination of all global and local kmer features by XGBoost in Dataset 2.

It is a ROC curve plot by XGBoost model in Dataset 2 with the combination of all global and local kmer features. The X-axis is false positive rate and Y-axis is true positive rate by different threshold. The threshold is indicated on the curve with a color scale. The AUC is the area under the ROC curve and is 0.793.

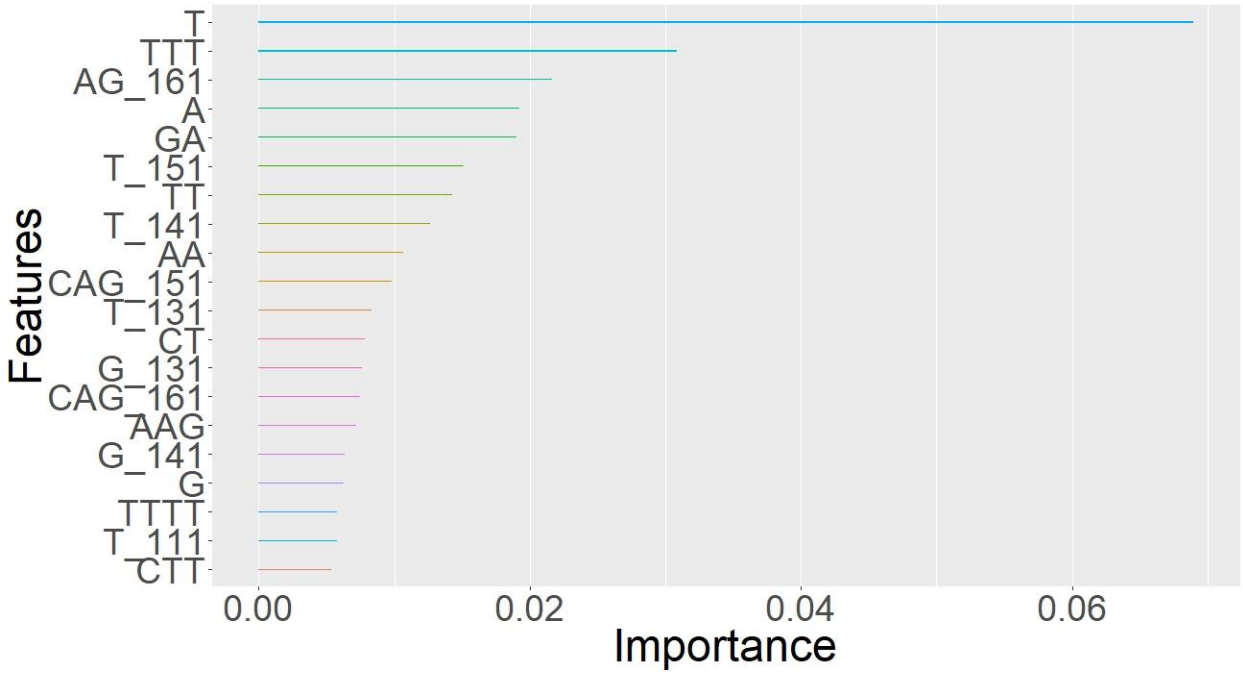


Figure 3-4 Feature importance ranking of global and local kmer in XGBoost model

The feature importance ranking plot shows magnitude of each global and local kmer feature relationship with the response as compared to other features used in the XGBoost. The model identifies “T”, “TTT”, “GA” kmer features are the top 3 variables impacting the sequence to be an IRES.

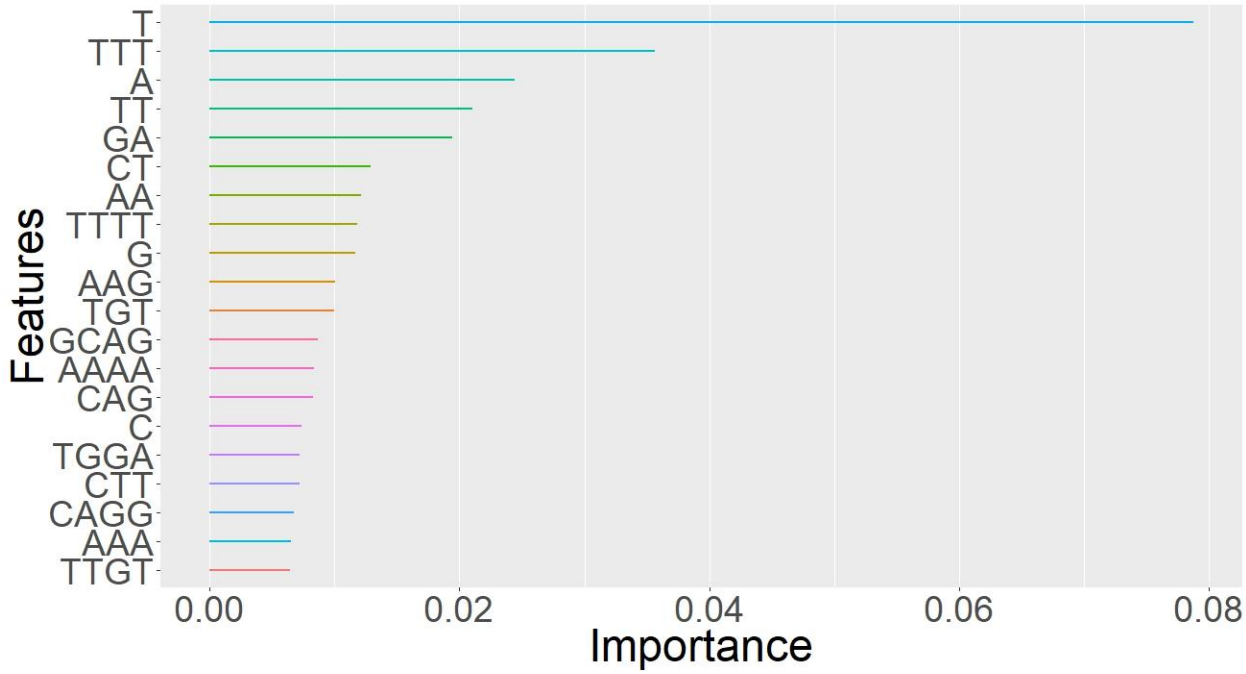


Figure 3-5 Features importance ranking of global kmer features in the XGBoost model

The feature importance ranking plot shows magnitude of each global kmer feature relationship with the response as compared to other features used in the XGBoost. The model identifies “T”, “TTT”, “GA” kmer features are the top 3 variables impacting the sequence to be an IRES.

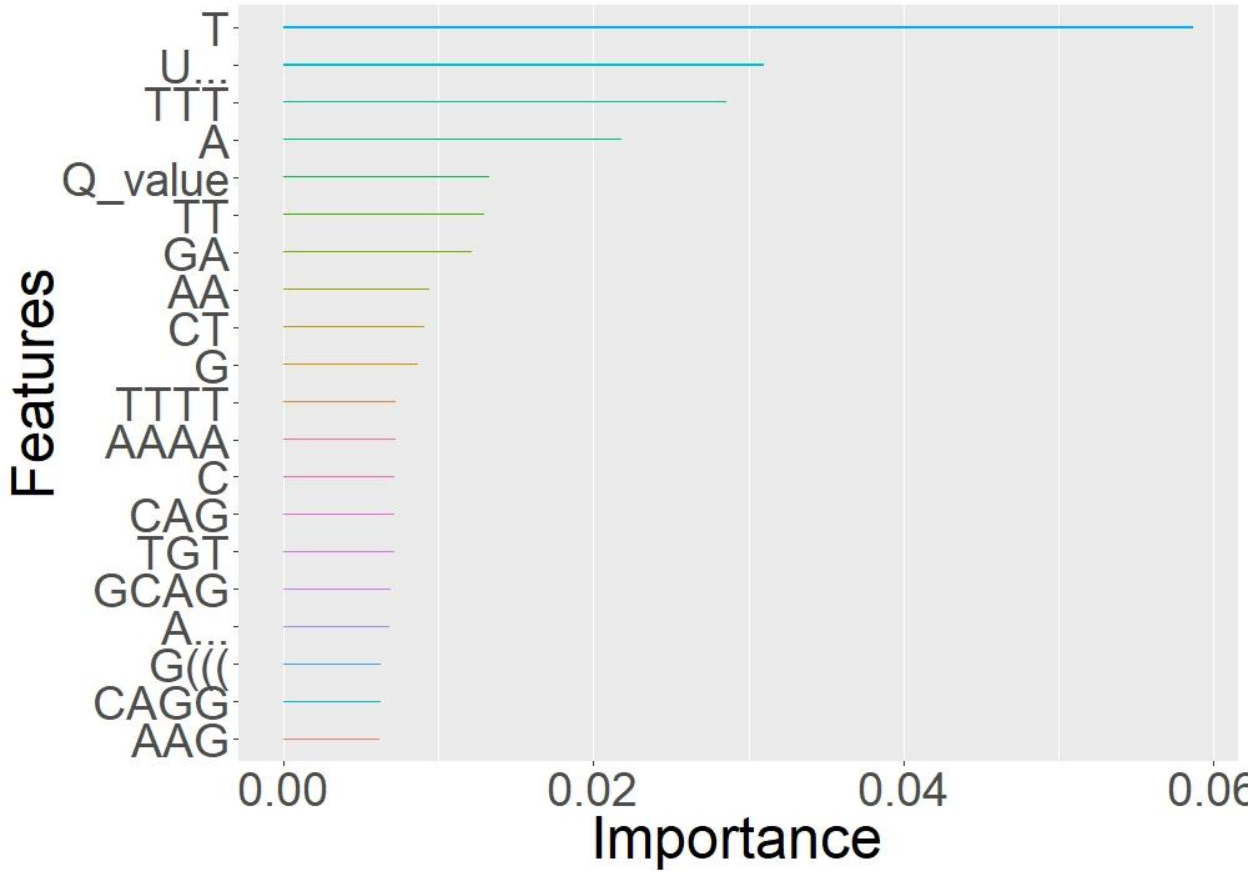


Figure 3-6 Feature importance ranking of global kmer and structural features in XGBoost model

The feature importance ranking plot shows magnitude of each global and structural feature relationship with the response as compared to other features used in the XGBoost. The model identifies “T”, “TTT” kmer features and Triplet25 which is “U…” are the top 3 variables impacting the sequence to be an IRES

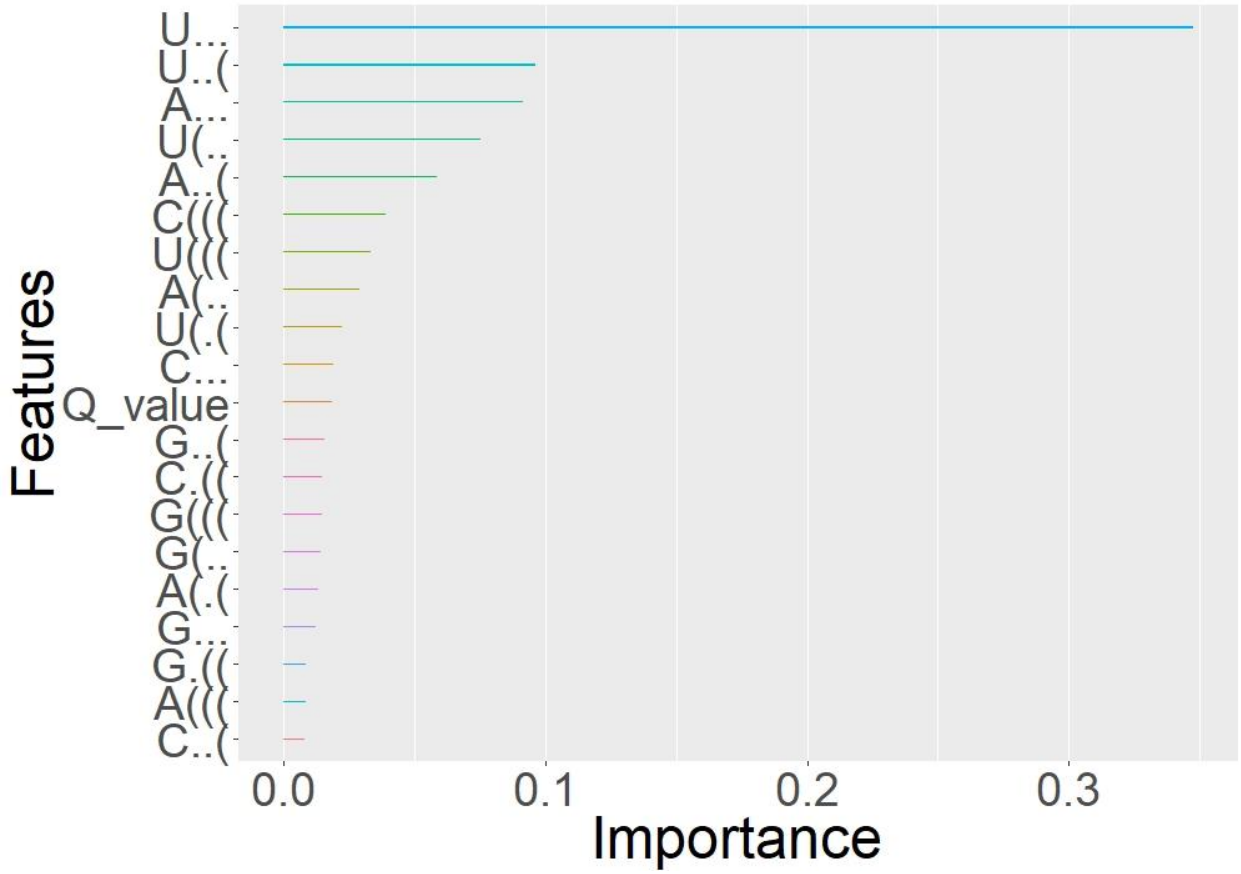


Figure 3-7 Feature importance ranking of structural features in the XGBoost model

The feature importance ranking plot shows magnitude of each structural feature relationship with the response as compared to other features used in the XGBoost. The model identifies Triplet "U…", Triplet "U..(" and Triplet "A…" are the top 3 variables impacting the sequence to be an IRES.

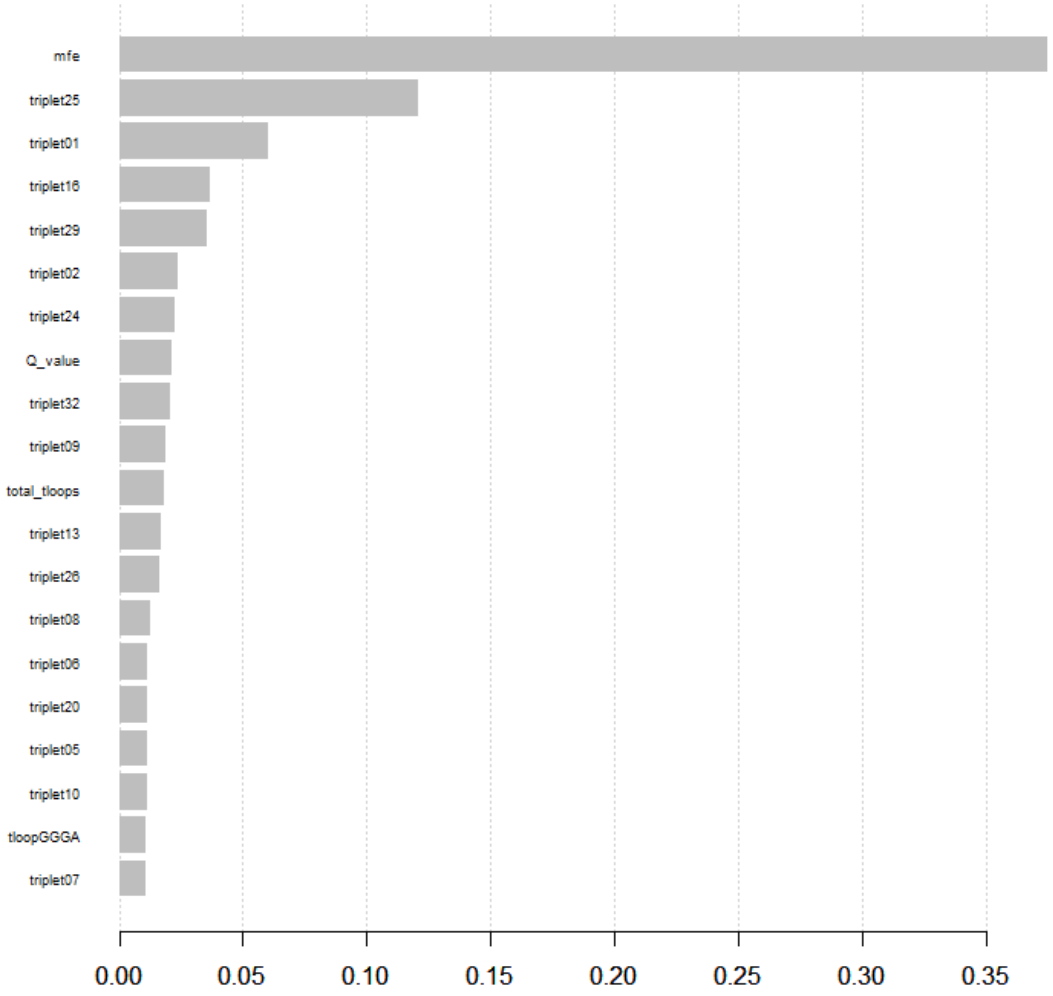


Figure 3-8 Feature importance ranking of structural features in the XGBoost model for classification of global-sensitive IRES and local-sensitive IRES

The feature importance ranking plot shows magnitude of each structural feature relationship with the response as compared to other features used in the XGBoost. The model identifies MFE, Triplet25 which is "U…" and Triplet01 which is "A…" features are the top 3 variables impacting the sequence to be a global-sensitive IRES rather than a local-sensitive IRES.

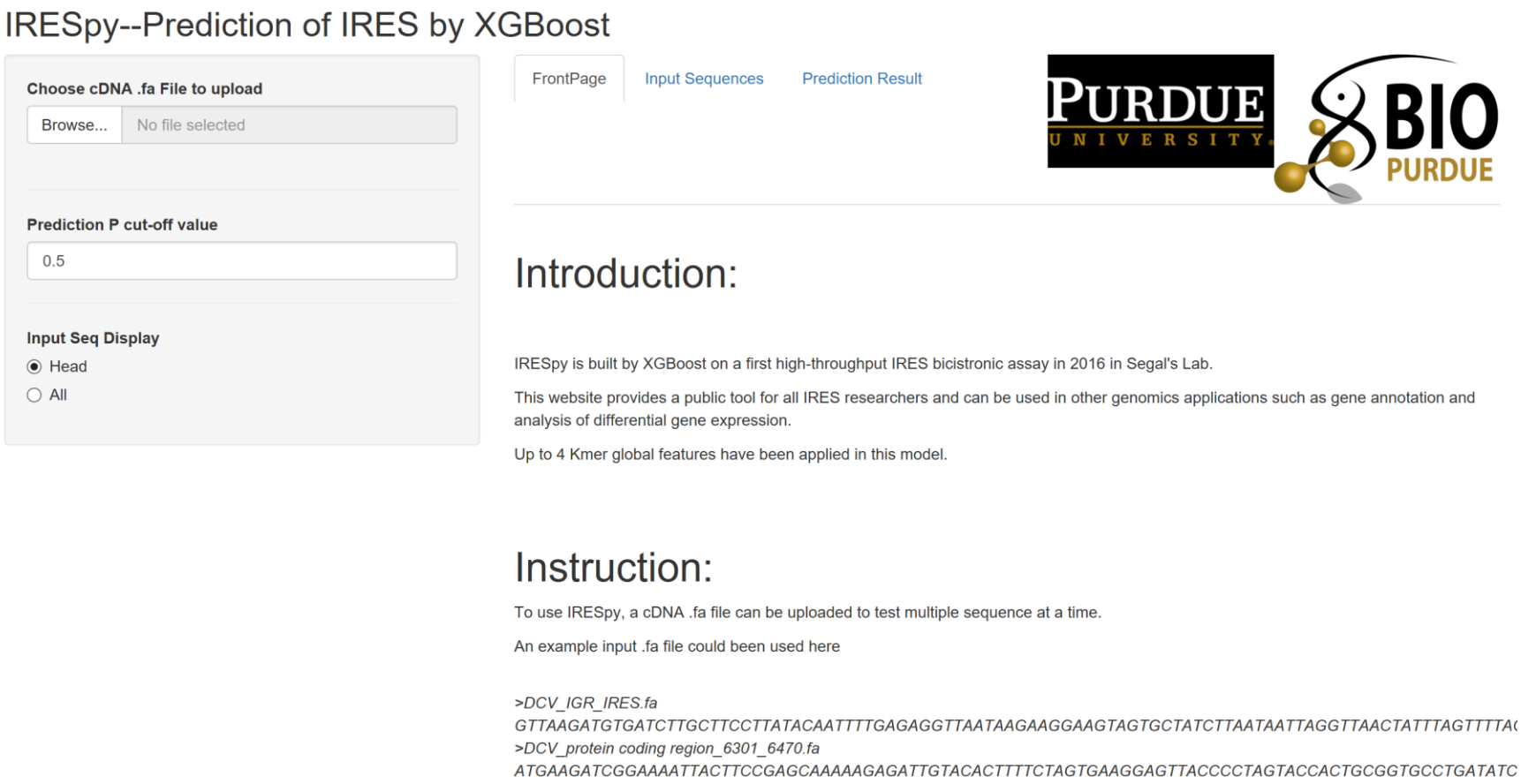


Figure 3-9 IRESpy website

CHAPTER 4. IRESPY: AN XGBOOST MODEL FOR PREDICTION OF INTERNAL RIBOSOME ENTRY SITES

4.1 Abstract

Internal ribosome entry sites (IRES) are segments of the mRNA found in untranslated regions that can recruit the ribosome and initiate translation independently of the more widely used 5' cap-dependent translation initiation mechanism. IRES usually function when 5' cap-dependent translation initiation has been blocked or repressed. This paper systematically studies the features that can distinguish IRES from non-IRES sequences. Sequence features such as kmer words, structural features such as QMFE, and sequence/structure hybrid features are evaluated as possible discriminators. They are incorporated into an IRES classifier based on XGBoost. The XGBoost model performs better than previous classifiers, with higher accuracy and much shorter computational time. The number of features in the model has been greatly reduced, compared to previous predictors, by including global kmer and structural features. The contributions of model features can be explained from both global and local perspectives. The trained XGBoost model has been implemented as a bioinformatics tool for IRES prediction, IRESpy (https://irespy.shinyapps.io/IRESpy/). This website provides a public tool for all IRES researchers and can be used in other genomics applications such as gene annotation and analysis of differential gene expression.

4.2 Introduction

Internal ribosome entry sites (IRES) are segments of the mRNA in untranslated regions that can recruit the ribosome and initiate translation, especially when the conventional cap-dependent

translation initiation mechanism has been blocked or repressed. They have been found to play important roles in viral infection, cellular apoptosis, cellular differentiation and response to external stimuli such as hypoxia, serum deprivation and heat shock (Hung et al., 2014; Jo et al., 2008; Sharathchandra, Katoch, & Das, 2014; Spriggs, Bushell, Mitchell, & Willis, 2005). IRES have been identified as potential therapeutic targets for antagonists that can interrupt IRES function and control the expression of viral proteins (A. A. Komar & Hatzoglou, 2015). Such drugs could be small-molecule inhibitors such as peptide nucleic acids (PNAs), short hairpin RNAs (shRNAs), small interfering RNAs, antisense oligonucleotides, and ribozymes (A. A. Komar & Hatzoglou, 2015; Martinand-Mari, Lebleu, & Robbins, 2003; Nulf & Corey, 2004). An improved understanding of cellular IRES function under different physiological conditions will increase understanding of the response of cells in proliferation, apoptosis and tumorigenesis.

IRES are widely found in both viral and cellular mRNA. They were first discovered in the Poliovirus (PV) and Encephalomyocarditis virus (EMCV) RNA genomes in 1988 using a synthetic bicistronic assay (Pelletier & Sonenberg, 1988). The assay places potential IRES sequence segments between two reporter genes, and measures the expression of the reporter genes in comparison to a non-IRES control construct. The bicistronic assay is considered to be the best experimental method to confirm the presence of IRES. However, this method is time consuming and labor intensive, and in the past 30 years, only a few hundred IRES have been confirmed. The difficulty in identifying IRES is complicated by incomplete understanding of the mechanism(s) of IRES function. In the simplest case, that of Dicistroviruses, IRES function without the help of eukaryotic initiation factors (eIFs) or IRES trans-acting factors (ITAFs), but in other viruses, and in most cellular IRES, eIFs and ITAFs are required. Various lines of evidence implicate RNA structure in IRES function (Filbin & Kieft, 2009; Lozano, Fernandez, & Martinez-Salas, 2016;

Martinez-Salas, Lopez de Quinto, Ramos, & Fernandez-Miragall, 2002; Plank & Kieft, 2012), especially in IRES that do not require additional protein factors, but the relative importance of RNA structure, ITAFs, and (possibly unidentified) RNA binding proteins remains unclear. Whether all IRES share a common mechanism, and therefore common sequence and structural features, has not been determined, and universal features shared by all IRES have yet to be identified (A. A. Komar, & Hatzoglou, M., 2005; Mailliot & Martin, 2018). This substantial gap in our knowledge can be largely attributed to the relatively small number of confirmed IRES, which, until recently, has made identification of common features difficult.

It has been estimated that about 10% of cellular and viral mRNA may use IRES to initiate translation (Stoneley & Willis, 2004), but the limited number of confirmed IRES has prevented study and understanding of IRES function. Alternative approaches to IRES identification, such as comparative analysis of IRES primary/secondary/tertiary structure, have been tried, but little commonality has been found across all IRES (Filbin & Kieft, 2009; Hong, Wu, Chang, & Chen, 2013). Small sequence motifs have been reported to be conserved within specific viral IRES groups, for instance, a GNRA sequence is shared in picornavirus IRES (Fernandez-Miragall & Martinez-Salas, 2003). The SL2.1 stem/loop contains a U rich motif that has been found to be important for ribosome binding in the Dicistrovirus IGR IRES(Costantino & Kieft, 2005; Schuler et al., 2006).

The absence of universally conserved features across all IRES makes their prediction difficult from a bioinformatics perspective, but several systems have been implemented. For example, the Viral IRES Prediction System (VIPS) predicts the secondary structure of an RNA from its sequence, and uses the RNA Align program to align the predicted structure to known IRES to predict whether the sequence contains an IRES (Hong et al., 2013). However, VIPS

predictions are limited to viral IRES, and although the accuracy rate of VIPS was assessed as over 80% for four viral IRES sub-groups, the prediction accuracy was assessed only on the training dataset and is substantially overestimated. The ability of VIPS to find novel viral IRES is low in our hands (note that the VIPS server is no longer available). A more recent method, IRESPred, uses 35 sequence and structural features and the probabilities of interactions between RNA and small subunit ribosomal proteins to predict IRES (Kolekar, Pataskar, Kulkarni-Kale, Pal, & Kulkarni, 2016). IRESpred was trained using a non-IRES negative training set that included viral protein coding and cellular protein coding mRNA sequences; unfortunately some of these sequences were later found to contain IRES (Weingarten-Gabbay et al., 2016). In addition, IRESpred incorporates features such as UTR length and the number of upstream AUGs. Such features are dependent on the length of the query sequence, and most of the positive training set is substantially longer than the negative training set. The overall false positive rate for IRES prediction is high: in a test of 100 random 400 base sequences, 98 were predicted to be IRES (results not shown). This high false positive rate has been confirmed by other investigators, as well (Zhao et al., 2018).

Instead of using features common to all IRES to determine for prediction, recent results suggest that machine learning approaches that combine multiple weak learners to predict IRES may be effective (Libbrecht & Noble, 2015; Valentini, Tagliaferri, & Masulli, 2009). In 2016, Weingarten-Gabbay et al. developed a high-throughput IRES activity assay and employed it to identify thousands of novel IRES in human and viral genomes (Weingarten-Gabbay et al., 2016). The identification of many new IRES improves the likelihood that a machine learning model can be successfully implemented. Based on the Weingarten-Gabbay et al. dataset, Gritsenko et al. built a stochastic gradient-boosting decision tree model (GBDT) to predict IRES using 6120 kmer

features (Gritsenko et al., 2017). However, the large feature set leads to possible model overfitting and slow model fitting time.

IRESfinder, the most recent method, uses only the human genome part of the Weingarten-Gabbay et al. dataset and implements a logit model with framed kmer features to predict cellular IRES (Zhao et al., 2018). The IRESfinder logit model is trained only on cellular IRES, and, as a transformed linear model, may not work well for non-linear relationships. In addition, the independent testing dataset is very small (only 13 sequences), possibly leading to overestimation of the AUC.

In this manuscript, we describe a machine learning model that combines sequence and structural features to predict both viral and cellular IRES, with better performance previous models. In order to make the predictive model widely available, it has been implemented as a simple to execute R/Shiny app. The optimized model, IRESpy, is very fast, and can be used to make genome scale predictions.

4.3 Results

In a typical scenario, one has only the sequence of the RNA available and does not have additional information (such as experimentally determined secondary and tertiary structure). In this work we focus on features that can be obtained from the sequence alone, rather than on comparative information, which requires a curated comparative database. We consider three kinds of features: sequence features, structural features, and sequence-feature hybrid features.

4.3.1 Sequence Features

Sequence features are the tabulated frequencies of kmer words in the target sequences. Given the four base RNA alphabets, there are 4k words of length k, yielding four 1mer, sixteen 2mer, sixty-four 3mer, and two hundred and fifty-six 4mer features (total=340). It is possible that sequence features, which might correspond to protein binding sites, could be localized with respect to other features in the IRES. To incorporate this possibility, we consider both global kmers, the word frequency counted over the entire length of the sequence, and local kmers, which are counted in 20 base windows with a 10-base overlap, beginning at the 5' end of the sequence of interest. In all cases, the kmer count is divided by the sequence length to give the kmer frequency. An example of kmer calculation in CrPV IGR IRES is shown in Fig.1.

4.3.2 Structural Features

The predicted minimum free energy (PMFE) is highly correlated with sequence length (Trotta, 2014). This is undesirable as could lead to false positive predictions based on the length of the query sequence. While this effect is reduced using Dataset 2, in which all training sequences are the same length, sequence length is clearly a conflating variable that should be excluded.

QMFE, the ratio of the PMFE and the PMFE of the randomized sequence (Bonnet, 2004), is much less dependent on sequence length (see materials and methods). It is believed that the stability of RNA secondary structure depends crucially on the stacking of adjacent base pairs (Jaeger, Turner, & Zuker, 1989; Turner, Sugimoto, & Freier, 1988). Therefore, the frequencies of dinucleotides in the randomized sequences are an important consideration in calculating the PMFE of randomized sequences (Clote, Ferre, Kranakis, & Krizanc, 2005). In calculating QMFE, a dinucleotide preserving randomization method has been used to generate randomized sequences.

QMFE can be used to compare the degree of predicted secondary structure in different sequences regardless of length. This length independent statistic indicates whether the degree of secondary structure is relatively lower or higher than that of randomized sequences, respectively. Viral IRES have been found to have highly folded secondary structures that are critical for their function. The structures of Dicistrovirus IRES, in particular, are conserved and comprise folded structures with three pseudoknots. Cellular IRES typically need ITAFs to initiate translation, and the binding between ITAFs and cellular IRES has been proposed to activate the IRES structure by changing it from a relaxed status to a rigid status (Filbin & Kieft, 2009). Cellular IRES are therefore likely to have a less extensively base-paired secondary structure. The 5' UTRs of housekeeping genes, in general, do not require highly folded structures because they use the cap-dependent translation initiation process.

Average QMFE values clearly differ in viral IRES, cellular IRES and the UTRs of housekeeping genes (Fig 2). We expect that QMFE should be also different in IRES and non-IRES regions of the same mRNA. Figure 2A shows the observed differences in QMFE of selected viral IRES, cellular IRES, and a housekeeping gene 5'UTR. The QMFE of the viral IRES is the lowest. The cellular IRES QMFE is about 0.5, which indicates this sequence an intermediate degree of secondary structure, but still more than would be expected for randomized sequences, and the 5'UTR of the ERH housekeeping genes has the highest QMFE, indicating a relatively low degree of secondary structure. These results suggest that the QMFE can indicate the degree of base-paired secondary structure in various sequence classes, and may be useful in distinguishing IRES and non-IRES sequences. Fig 2B shows the QMFE of 200 base segments of CrPV. Two of the low QMFE regions exactly match the regions of the known the 5'UTR IRES (bases 1-708) and

intergenic IRES (bases 6000-6200), again indicating that QMFE may be a powerful discriminatory feature that can be used to identify IRES positions mRNA sequences.

4.3.3 Hybrid Features

Triplet features, which combine the primary sequence and predicted base-paired structure, have been used in miRNA prediction (Vitsios et al., 2017). The first successful application of this kind of feature was in a support vector machine algorithm for classifying pre-miRNAs (Xue et al., 2005). The definition and calculation of triplet features are shown in Figure 3. Triplet features encode the local predicted secondary structure as a series of characters indicating the predicted structure (where the symbols '(' and '.' indicate base-paired and unpaired bases, respectively) and the base at the center of the triplet. The triplet feature "A(((" thus indicates a sequence where three bases are base-paired, and the center base is an 'A'.

4.3.4 Approach

In this work, we focus on an ab initio classification approach for IRES prediction. All the features considered here are sequence length independent - kmer words, QMFE, and triplets, and thus should be equally appropriate for scanning long (genomic) or short (specific target) sequences.

Two existing databases that have been created to systematically study the IRES provide useful background information for this study. The first database, referred to as Dataset 1 in this work comprises confirmed IRES drawn from IRESite and includes selected 5'UTRs of housekeeping genes. 52 viral IRES and 64 cellular IRES from IRESite are labeled as IRES in Dataset 1. Housekeeping genes principally utilize the 5' cap-dependent mechanism for initiation, and 51 of them were randomly selected as the non-IRES group used for comparison in Dataset 1 (A. A. Komar, Mazumder, & Merrick, 2012). Dataset 2 results from a high-throughput bicistronic

assay developed in 2016, and its application has increased the number of known IRES by more than 10-fold (Weingarten-Gabbay et al., 2016). This large increase in the number of examples of IRES provides an opportunity to better learn the relationship between sequence and structural features and IRES mechanism.

We primarily rely on the Dataset 2 to build the machine learning model due to its large size and semi-quantitative measure of IRES activity. The entire Dataset 2 has been randomly divided into a training partition (80%) and a validation partition (20%). The training dataset was used in a grid search to optimize the learning rate, maximum tree depth, subsample ratio of the training instances, and subsample ratio of the features, used when constructing each tree. Each combination of parameters was evaluated using a 10-fold cross validation approach, in which the training partition was equally divided into 10 sets; one set is used for testing, and the remainder used for training in each run. In successive runs, different partitions are held out for testing. In the end, the best fit parameters are summarized to generate the final model. The data in the validation is not included in either hyperparameter or parameter training and thus provides an unbiased evaluation of the final trained model.

XGBoost stands for eXtreme Gradient Boosting. It combines weak learners (decision trees) to achieve stronger overall class discrimination. XGBoost learns a series of decision trees to classify the labelled training data. Each decision comprises a series of rules that semi-optimally split the training data. Successive trees that "correct" the errors in the initial tree are then learned to improve the classification of positive and negative training examples. Compared with gradient boosting, XGBoost can be more efficiently parallelized computed, incorporates regularization and tree pruning that prevent over-fitting. A variety of hyperparameters must be optimized in the

XGBoost method, including the learning rate, maximum tree depth, subsample ratio of the training instances, and subsample ratio of the features.

A succession of decision trees is generated where each tree, metaphorically, corrects the errors made in the previous trees. Due to the nature of this process, it is often difficult to map the importance of the features directly onto biological importance since each individual “rule” in the decision tree is likely to be noisy.

4.3.5 Training on kmer features

Machine learning models, including GBDT, and extreme gradient boosting (XGBoost), have been compared for IRES prediction. The approach used here, XGBoost exhibits higher AUC performance, and substantially lower training time than the GBDT model. As shown in Fig 4A, XGBoost requires 75% less training time, but improves AUC by 5% compared with GBDT, without any hyperparameter tuning. With the same features, but different model and parameter tuning, the XGBoost model can reach a testing AUC of 0.793 and training AUC 0.947. This is substantially better than the GBDT which showed a testing AUC of 0.77, and training AUC of 1.0 (Figure 4B). To investigate the relative importance of global and local kmer features, the XGBoost model was run with the same parameter settings, but incorporating only global kmer features. The testing AUC is 0.771 and training AUC is 0.911 (Figure 4B); this model achieves the same performance as GBDT, but requires many fewer features. The final model includes 1281 individual trees and each tree is built by 340 features. The depth of each tree is set to be 6.

4.3.6 Training on kmer + structural features

Structural features such as the number of predicted hairpin-, bulge-, and internal- loops; maximum loop length, maximum hairpin-loop length, maximum hairpin-stem length, and the

number of unpaired bases have been previously studied (Gritsenko et al., 2017; Kolekar et al., 2016; Zhao et al., 2018), but none were found to have significant predictive value. We hypothesized that QMFE, and triplet features, because they are length independent and combine sequence and structural information, might act as better features to classify IRES and non-IRES sequences. In particular, triplet features have the potential to reveal locally conserved sequence motifs that appear in a specific structural context. These features have been combined with the previously examined global kmer features in a sequence-structural model that is better than a simple sequence-based model. The testing AUC of the combined model increases slightly, from 0.771 to 0.775 (Fig. 5). The small magnitude of the increase probably indicates the presence of correlation between the global kmer and structural features. When using the structural features alone, the testing AUC is 0.741, which means that the structural features can still capture most of the variance of the dataset with only 33 features.

The high AUC of the structural feature-based model indicates that structural features alone can capture most of the information contained in the kmer features, while decreasing the number of features from 340 to 33. The structural features therefore have a relatively high information content. However the lack of improvement in the combined model compared to either the global kmer or structural model suggests that the information in kmer words and the structural features may be largely redundant.

4.3.7 Biological significance of discriminative features

As mentioned previously, it is not usually straightforward to understand the biological relevance of the selected features. Machine learning (ML) models are often considered “black boxes” due to their complex inner mechanism. Understanding the contribution of each feature to the model has been recognized as a very difficult aspect of machine learning. The SHAP (SHapley

Additive exPlanations) method assigns values that measure the marginal contribution of each feature in the model (Lundberg, 2017). It combines game theory with local explanations and is well suited for machine learning explanation. Unlike feature importance measures based on weight, cover, or information gain, the SHAP value is the only consistent and locally accurate additive method, and it can be interpreted as indicating which features are the most globally important for classification. Figure 6A shows the top 20 most important features in models trained with both global and local kmers. Red indicates higher feature values and blue indicates lower feature values. Higher frequencies of U rich kmers, such as “U”, “UU”, “UUU”, “UUUU”, “CU”, and “UGU”, are associated with higher predicted probability of being IRES. This is consistent with the previous reports that pyrimidine-rich kmers, especially U rich kmers are important for IRES function. (Weingarten-Gabbay et al., 2016). Importance of global kmer and local kmer features follow similar patterns, for instance, the local kmer features U_121, U_131, U_141, U_151, and U_161 all support classification of sequences as IRES, as do the global kmer features. The importance of the local region from base 121-161 may be important as an ITAF binding site (perhaps pyrimidine tract binding protein), as suggested by Weingarten-Gabbay et al. Whether the CU feature is related to the poly U feature is difficult to tell. It is worth noting that in picornoviral IRES, one of the most conserved features is the SL3A “hexaloop” in which a CU dinucleotide is highly conserved (Fernandez, Buddrus, Pineiro, & Martinez-Salas, 2013). Figure 6B lists the SHAP values of the top important features for the global kmer only model. The similar importance of features in different models suggests that the models are detecting essentially the same features. Figure 6C shows the SHAP values for both the global kmer and structural features model. Some structural features, such as ‘U..’, ‘G(((’, and the QMFE , are more important than most global kmers. Figure

6D lists the structural features which serves as a potential structural motif list much like a differentially expressed genes list in the RNA-seq analysis.

In order to understand the biological meaning of the trained model we can examine how the response variable, in this case classification as IRES vs non-IRES, changes with respect to the values of the features. SHAP values show the change in the predicted value as a specified feature varies over its marginal distribution, for each important feature. Figure 7A shows examples of two highly ranked features. An increase in the frequency of the UUU 3mer, from 0.01 to 0.03, increases the probability that a sequence is an IRES, while an increase in the frequency of the GA 2mer from 0.04 to 0.08 decreases the probability that the sequence is IRES.

For novel sequences, instead of simply predicting the probability that a sequence is an IRES, we want to know which features can explain the prediction. Local Interpretable Model-agnostic Explanations (LIME) analysis explains the contribution of individual features to the overall prediction (Kemp, MacAulay, & Palcic, 1997; Zhang et al., 2018). The assumption of LIME is that every complex model has a linear or explainable relationship in the local space of the dataset. It is possible to fit a simple model around a sequence by slightly permuting its feature matrix. In LIME, a similarity matrix that measures the distance between a query sequence and a certain number of permutations is constructed. Each permutation is classified by the XGBoost model, and the predicted class, IRES or non-IRES, is further classified by a simple model. The simple model uses the same features as the XGBoost model, and mimics how the XGBoost model behaves in the local space defined by the permutations. Figure 7B shows, for instance, why the predicted probability of CrPV IGR IRES is high (p=0.861), but the predicted probability of an IRES in the CrPV protein coding sequence is very low (p=0.067). There are more green bars,

which represent the positively weighted features in the CrPV IGR IRES, than in the CrPV protein coding sequences (non-IRES).

4.3.8 Structural Features

We use importance ranking plots to analyze the importance of triplet features in IRES prediction. Figure 6B shows that triplets “U…”, “A…”, “A..(” are important in the model including both global kmers and structural features, as well as in the model including only structural features. In particular, the triplet “U…”, a loop with a central U base, can be seen to be important. This feature may correspond to the conserved U rich loop motif found in the SL2.1 region of Dicistrovirus IGR IRES. The SL2.1 stem/loop has been found to be important for ribosome binding (Costantino & Kieft, 2005; Schuler et al., 2006), and in the Cryo-EM structure of the CrPV IRES, it is complexed with the ribosome, with the SL2.1 region positioned at the interface of the IRES and the ribosome (Jan & Sarnow, 2002; Schuler et al., 2006), in direct contact with the ribosome. Mutations in the SL2.1 region result in loss of IRES function (Hatakeyama, Shibuya, Nishiyama, & Nakashima, 2004; Jang & Jan, 2010; Mailliot & Martin, 2018).

4.3.9 Prediction probability VS IRES activity

The IRES activity of the sequences in Dataset 2 was measured by inserting them into a lentiviral bicistronic plasmid, between mRFP and eGFP reporter genes, and transfecting H1299 cells, which results in integration of a single oligonucleotide construct in each cell (Weingarten-Gabbay et al., 2016). The cells are sorted with FACS and assigned to 16 fluorescence intensity bins on the basis of eGFP expression. IRES activity, in the range 206 to 50000, is defined by those expression levels. The correlation between the IRES probability predicted by our XGBoost model and the quantitative IRES experimental activities has been explored, and the result shows that the

predicted IRES probability is significantly higher for high-activity (>1000) IRES, than for those where the IRES activity is close to the base level (<1000) in Fig 8. This suggests that the predictive accuracy of the XGBoost model for high activity IRES is higher than for marginally active sites, and implies that, when high precision is a priority, precision can be increased at the expense of recall.

4.3.10 Scan of human UTRs

IRESpy has been applied to scan human 5'UTRs (124315 UTR sequences listed in UTRdb). Fig 8 shows the distribution of IRES prediction probability for the positive and negative training sets in Dataset 2, and all human UTRs. The distribution of probabilities in the human UTR dataset strongly resembles the Dataset 2 negative class, but has a larger tail. This suggests that IRESpy is successfully distinguishing IRES from non-IRES in the uncharacterized human UTRs. When a prediction threshold of 0.1 is used for both datasets, 13.47% of the human IRES are predicted to contain IRES which is close to the 10% value cited in previous reports (Stoneley & Willis, 2004).

4.3.11 IRESpy prediction tool

The XGBoost model based on global kmer features, has been implemented as a shiny application, IRESpy. It is available online: https://irespy.shinyapps.io/IRESpy/. Compared with IRESpred (Table 1), IRESpy shows better predictive performance, with both higher sensitivity (recall) and higher precision on the validation dataset (not included in parameter or hyperparameter training).

To further test the predictive ability of IRESpy, it has been applied to 202 highly structured non-IRES RNAs (see methods) (J. Huang, Li, & Gribskov, 2016), Dataset 1, which includes the reported sequences of IRES from IRESite (positives) (Mokrejs et al., 2010), and of housekeeping

genes 5'UTRs (presumed negatives). IRESpy clearly distinguishes IRES and non-IRES sequences in Dataset 1. The low predicted IRES probability for all highly structured RNA groups suggests that IRESpy is not simply detecting relatively structured RNA. Since relatively high secondary structure is widely considered to be a hallmark of IRES, the test against highly structured RNAS represents an especially difficult test.

4.4 Discussion

Clearly, both the selected features and the models are important for predicting the existence of IRES. A limitation of VIPS and IRESPred are the inclusion of length dependent features such as the length of UTRs, and the number of upstream AUGs. This is a serious drawback when predicting IRES in UTRs, which vary greatly in length. IRESpy performs better than the GBDT method, using a smaller number of features. Using the same datasets and features (global and local kmer features), but switching from the GBDT model to XGBoost, increases the validation AUC by 5%, and the decreases the training time by 75%.

Global kmer and local kmer features are highly correlated. The XGBoost model achieves the same model performance as the GBDT model incorporating only global kmer features. The modest increase in classification performance, accompanied by a 94% decrease in the number of features, suggests that the IRESpy model shows better generalization. The reduced number of model features results in a decrease in both training time and classification time (making the XGBoost model more appropriate for genome wide scanning).

Surprisingly, incorporation of structural features such as QMFE and triplet features, has relatively little effect on model performance, although some of the highly ranked features such as "U…" can be directly related to known mechanistic features of some IRES. The reason for this

lack of improvement are not obvious. Several explanations seem possible. The extensive nature of the QMFE, while it provides an overall measure of the degree of secondary structure, may not be sensitive enough to particular structural and topological features that are important to IRES function. Alternatively, while the prediction MFE RNA structures is relatively good, generally estimated to be about 80% accurate (Mathews, 2006; Zuker, 2003) at the base pair level, it may not be good enough to reliably detect structural motifs. Furthermore, the RNA structure prediction approach used here does not predict pseudoknots which, based our knowledge of viral IRES, may be highly important to IRES function. On the other hand, triplet features take a very local view of structure and sequence and may be too detailed to capture larger important structural motifs. Another explanation may be that, in fact, IRES function involves many different mechanisms (Plank & Kieft, 2012) – the XGBoost decision tree models can capture the fact that different features are important for different IRES, but unfortunately, teasing this information out of the trained model is difficult – the interpretation of the importance of features in machine learning models is a topic of high interest in the machine learning community. The SHAP feature importance plots shown in figure 6 can serve as a potential motif list for researchers to test by lab experiment. In particular, the triplet "U…" may work importantly for IRES like a conserved U rich loop motif found in the SL2.1 region of Dicistrovirus IGR IRES. The CU kmer is part of a known tetraloop motif (CUYG) which might be important in stabilizing the IRES structure (Moore, 1999). The combination of global kmer features and structural features increases the validation AUC compared with that of the model incorporating global kmer features alone, but only modestly. Using structural features alone achieves relatively high classification performance, and at the same time, reduces the number of features from 340 to 33. From one point of view, this indicates that

the structural features are relatively powerful, providing higher performance per feature, but why these features cannot greatly increase predictive performance remains unclear.

In summary, IRESpy is a high-throughput online tool for IRES prediction. Its prediction quality is better than previous tools, and it is able to predict both viral and cellular IRES with good performance. IRESpy uses only length-independent features in its prediction making in appropriate for analyzing RNAs of different lengths. The computational time is low making IRESpy appropriate for genome wide comparisons and for use in genome annotation. The IRESpy application is freely available as a R/shiny app making it easily available to both computationally sophisticated and more computationally naïve users.

4.5 Materials and Methods

4.5.1 Training Data (Dataset 2)

The dataset used to train the XGBoost model is derived from Weingarten-Gabbay et al. (Weingarten-Gabbay et al., 2016). The original dataset includes 55,000 sequences – selected from reported IRES, 5'UTRs of human genes, 5'UTRs of viral genes, and sequences complementary to 18S rRNA. Sequence fragments were screened in a high-throughput bicistronic assay using a consistent 173 base insert size, removing any length effects. This dataset is available online (https://bitbucket.org/alexeyg-com/irespredictor/src). From the 55,000 tested sequences, 28,669 native subsequences originating from human and viral genomes were selected as dataset for use in this work. The remaining sequences are synthetic sequences introduced to test the effect of specific mutations on IRES activity. Based on the reported replicate measurements of IRES activity, promotor activity, and splicing activity, we further filtered the dataset retained only sequences with splicing scores greater than -2.5 and promoter activity less than 0.2. The final training dataset,

referred to as Dataset 2, comprises 20872 subsequences: 2129 sequences with IRES activity scores above 600 are defined as IRES, and the other 18743 as nonIRES. The ratio of IRES to nonIRES is about 1:8.6.

20872 native sequences in Dataset 2 have been checked identity by Blastn. The results show 7.56% sequences have more than 80% identity, 15.3% sequences have more than 50% identity, and 17.02% sequences have more than 30% identity. There is no any sequence holding 100% identity. When constructing the database, those more than 80% identity sequences are introduced by scanning the similar homologs of viral UTRs, and different sequences sources when scanning the cellular and viral genome. Since the ratio of highly identity sequences is low, the XGBoost model has been tested again by excluding those similar sequences. We found the model performance is similar.

4.5.2 Highly structured RNA data

Highly structured RNA group includes 202 examples of 16S RNA, 23S RNA, 5S RNA, g1, g2, rnasep, tmRNA and tRNA (J. Huang et al., 2016). The sequences have been carefully screened to remove any sequences with greater than 40% sequence identity.

4.5.3 Dataset 1

Dataset 1 is composed of sequences from IRESite (Mokrejs et al., 2010) and selected 5’UTRs of housekeeping genes. 52 viral IRES and 64 cellular IRES from IRESite are labeled as IRES in Dataset 1. Housekeeping genes principally utilize the 5’ cap dependent mechanism for initiation and 51 of were selected as the non-IRES group in Dataset 1 (A. A. Komar et al., 2012).

4.5.4 Human UTRs

124315 of human 5'UTR sequences have been collected from UTRdb (Grillo et al., 2010).

4.5.5 Kmer features

The frequency of each kmer is calculated as the count of the kmer divided by the sequence length. Global kmer features are counted over the entire length of the sequence. Local kmer features are counted in 20 base windows, with a ten-base overlap between adjacent windows (ref to figure1).

4.5.6 Predicted minimum free energy (PMFE) and Q_{MFE}

The predicted minimum free energy is calculated by UNAfold-3.9 (Markham & Zuker, 2008). Q_{MFE} is calculated as follows:

(1) Calculate the predicted minimum freedom energy of the secondary structure from the original sequence by RNAfold.

(2) The original sequence has been randomized by permuting the dinucleotide ratios. Then the MFE of the randomized sequence has been generated.

(3) Repeat step 2 many times (for example 2000) in order to obtain a distribution of the predicted MFE values.

(4) If N is the number of iterations and n is the number of randomized sequences which MFE value are less or equal to the original value, then QMFE is calculated as:

QMFE = n/(N+1)

The Ushuffle program (Jiang et al., 2008), which is based on the Euler algorithm, is used to randomize the sequences used in calculating the Q_{MFE}. Ushuffle uses an exact method that

produces randomized sequences with exactly the same dinucleotide composition as the original sequences.

4.5.7 XGBoost Software and parameters

The XGBoost model is fitted under R (Version 3.5.0) with the xgboost package (Version 0.71.2). The parameters have been used in the XGBoost model include: eta=0.01, gamma=0, lamda=1, alpha=0, max_depth=5, min_child_weight=19, subsample=0.8, colsample_bytree=0.65). IRESpy is deployed online with the shiny package (Version 1.2.0). It is available on line: https://irespy.shinyapps.io/IRESpy/.

4.6 Supplemental Materials

4.6.1 Nested Cross-Validation

The model was trained using a nested cross-validation approach, as shown in Fig S1. In the inner loop, 10-fold cross-validation was used to search for the best model with the best hyper-parameters. Since the validation dataset in the outer loop has never been used in the model training, it can be used to test the generalization ability of the model. In the inner loop, ten-fold cross-validation is used to determine the best model with the best hyper-parameters. This model is then applied to predict the validation AUC in the outer loop. The average validation AUC measures the model generalization ability, and the model with the highest validation AUC was picked as the final model.

4.6.2 Hyper-parameter tuning

The goal of tuning the hyper-parameters is to obtain the best testing ROC-AUC without over-fitting the training data. There are several important hyper-parameters in XGBoost.

- number of trees
- eta (η): The learning (or shrinkage) parameter, which determines how fast the model converges. The range is 0 to 1.
- max_depth: controls the maximum depth of each tree.
- min_child_weight: controls the minimum number of observations in a leaf.
- colsample_bytree: controls the portion of variables to grow a new node in each tree.
- sub_sample: controls the ratio of the training samples in each tree.
- Gamma (γ): controls the minimum reduction in the loss function required to add a new node to a tree.
- Alpha (α), lambda (λ): L1 and L2 regularization terms on weights.

There are eight hyper-parameters that can be optimized. If all possible combinations of them are tried in a fully grid search, for example, three possible values of eight hyper-parameters, then there are 3^8=6561 different possible combinations. Instead, fixing all parameters except one and optimizing that one is much less time consuming. We first tune the most important hyper-parameters – those which affect the model performance the most. This stepwise optimized search is especially useful and applicable for models with more hyper-parameters to be tuned and more efficient to find a relatively good

combination of them. To eliminate some possible bias introduced by that one-by-one tuning, the grid search will be applied on eta, max_depth and min_child_weight in the end.

The suggested parameter ranges for XGBoost models are shown in Table S1 (Chen, 2016). The ranges of the parameters used in the earlier gradient boosting model (Gritsenko et al., 2017), learning rate r=[0.001,0.002,0.004,0.008], minimum leaf samples m=[5,25,125] and subsampling fraction f=[0.9,0.7] have been considered as well. The following approach uses the suggested parameter ranges in Table S1 to tune the hyper-parameters.

My approach for hyper-parameter tuning:

1. Eta, the learning rate, determines how fast the model fits. Eta has a huge effect on model performance. Higher eta means faster fitting. Initially, the learning rate is set to a relatively high value (eta=0.01). Then a combination of hyper-parameters within the ranges shown in Table S1 were randomly selected as initial values (showed later). Under these conditions, the optimum number of trees was determined, and the training AUC and testing AUC (Figure S2) calculated.
2. A grid search over the tree-specific hyper-parameters, max_depth, min_child_weight, colsample_bytree, sub_sample and gamma were then performed using the ranges shown in Table S1. The exact tuning process is described below.
3. Next, the regularization parameters, lambda and alpha, were tuned with all the other parameters fixed.

4. The learning rate was tuned with the best combinations of all other hyper-parameters.
5. Grid search of eta, max_depth and min_child_weight to eliminate some possible bias introduced by that one-by-one tuning. The values picked for those parameters are: eta= [0.1,0.01,0.001], max_depth= [1,5,9], min_child_weight= [19,29,39]. So there are 3^3=27 runs.

I started with eta=0.01, max_depth=3, min_child_weight=29, colsample_bytree=0.8, sub-sample= 0.8, gamma=0, alpha=0 and lambda=1 as initial values. The training process plot is shown in figure S2-A. It is a good start because the testing AUC approaches a plateau as more trees are trained. The best number of tress is 1661, which is within the suggested range (Table S1). The number of trees is an important hyper-parameter for XGBoost. It usually depends on the size of the training dataset and the range is typical between 100-2000.

After fixing the number of trees at 1661, the other tree-related parameters were tuned in stepwise grid searches. The result is shown in Figure S3. The max_depth and min_child_weight parameters were tuned first because they have higher impact on model performance. The best max_depth is 5 and the best min_child_weight is 19. Then other parameters were tuned one by one later, obtaining gamma=0, subsample=0.8, colsample_bytree=0.7 as the best values.

Alpha is the L1 regularization term on weights and lambda is the L2 regularization term on weights. Increasing either value gives models a higher penalty score for more

complexed structures. The grid search shows the default values are the best options for those two hyper-parameters. So alpha=0 and lambda=1.

Finally, eta, the learning rate, was tuned. The results show eta = 0.01 is the best choice. In the end, a grid search of eta, max_depth and min_child_weight has been tried to eliminate some possible bias introduced by that one-by-one tuning. The values picked for those parameters are: eta= [0.1,0.01,0.001], max_depth= [1,5,9], min_child_weight= [19,29,39]. There are 3^3=27 runs in total and the final results show that eta=0.01, max_depth=5 and min_child_weight=19 is the best combination by getting the highest validation AUC.

Figure S2 shows the change of the model parameters before and after tuning the hyper-parameters. The test AUC has a slightly increase from 0.752 to 0.756. And the number of trees has been decreased from 1661 to 901. It shows the goal of tuning the hyper-parameters which is improving the test AUC but at the same time reduce model complexity.

4.6.3 Sequence similarity

20872 native sequences in Dataset 2 have been checked identity by CD-hit program. CD-hit is a tool for clustering biological sequences on a large scale (Y. Huang, Niu, Gao, Fu, & Li, 2010). It is fast, scalable and flexible based on short word filtering and a greedy incremental clustering algorithm (Li, Jaroszewski, & Godzik, 2002). The results show 7.56% sequences have more than 80% identity, 15.3% sequences have more than 50% identity, and 17.02% sequences have more than 30% identity. There is no any sequence holding 100% identity. When constructing the database, those more than 80% identity sequences are introduced by scanning the similar homologs of viral UTRs, and different sequences sources when scanning the cellular and viral genome. Since the ratio of highly identity

sequences is low, the XGBoost model has been tested again by excluding those similar sequences. We found the model performance is similar.

4.6.4 Model performance comparison

There are four tools focusing on prediction of IRES before our method IRESpy. Their training dataset, methods, training features, and pros & cons have been listed in Table S2.

The summary in Table S2 tells that there are significant defects existing in VIPS and IRESPred. The inclusion of length dependent features such as the length of UTRs, and the number of upstream AUGs in those two methods introduces a serious drawback when predicting IRES in UTRs, which vary greatly in length. VIPS predictions are limited to viral IRES, and although the accuracy rate of VIPS was assessed as over 80% for four viral IRES sub-groups, the prediction accuracy was assessed only on the training dataset and is substantially overestimated. The ability of VIPS to find novel viral IRES is low in our hands (note that the VIPS server is no longer available). IRESpred was trained using a non-IRES negative training set that included viral protein coding and cellular protein coding mRNA sequences; unfortunately some of these sequences were later found to contain IRES (Weingarten-Gabbay et al., 2016). IRESfinder, the most recent method, uses only the human genome part of the Weingarten-Gabbay et al. dataset and implements a logit model with framed kmer features to predict cellular IRES (Zhao et al., 2018). The IRESfinder logit model is trained only on cellular IRES, and, as a transformed linear model, may not work well for non-linear relationships. In addition, the independent testing dataset is very small (only 13 sequences), possibly leading to overestimation of the AUC.

After all, I compare the performance of IRESpy, IRESpred and IRESfinder based on the data from the IRESfinder paper and the result has been shown in Table S3 (Zhao et al., 2018).

Even though the defects of IRESpred and IRESfinder exist, IRESpy works better in accuracy, sensitivity and precision in Table S3. The more straightforward comparison is between IRES-interpreter and IRESpy. Because they are working on the same dataset and the algorism of gradient boosting decision tree (GBDT) and the model XGBoost are similar. There is slightly increase of validation AUC from GBDT to XGBoost (Figure 4). But considering the much faster running time to fit the model (Figure 4), IRESpy is a faster, more efficient, more reliable tool to predict IRES compared to IRES-interpreter.

IRESpy provides the first, fast and high-throughput online testing tool for IRES screen. Its website can be used in other genomics applications such as gene annotation and analysis of differential gene expression. The number of features in the model has been greatly reduced, compared to previous predictors, by including global kmer features. It is the first time that structural features such as triplets and QMFE have been explored in the research of IRES. Unlike other structural features like the number of stem-loops and MFE, triplets and QMFE are length independent features and show significance in predicting IRES.

4.6.5 Identification of human 5'UTR IRES

IRESpy has been applied to scan human 5'UTRs (124315 UTR sequences listed in UTRdb). Fig 9 shows the distribution of IRES prediction probability for the positive and negative training sets in Dataset 2, and all human UTRs. The distribution of probabilities in the human UTR dataset strongly resembles the Dataset 2 negative class, but has a larger

tail. This suggests that IRESpy is successfully distinguishing IRES from non-IRES in the uncharacterized human UTRs.

The top 20 predicted human UTRs by IRESpy has been listed in Table S4. The gene ontology analysis (David 6.8) in Figure S5 shows that IRES might be a widely existing mechanism shared with many biological processes. Spleen tyrosine kinase (SYK), participating in intracellular signal transduction, protein autophosphorylation, Immunity, and protein complex, might potentially utilize IRES. It's 5'UTR is 201 BP long with a high CT ratio. To future demonstrate the IRES activity, the top 20 predicted UTRs have been aligned by the high-throughput bi-cistronic assay in Weingarten-Gabbay lab (Weingarten-Gabbay et al., 2016). If there is a match, the results have been showed in IRES_activity column in Table S4. The IRES_activity of SYK is 1378.92 compared with a 206 background level. So IRES mechanism might be used by SYK.

Table 4-1 Comparison between IRESpy and IRESpred model performance. IRESpy performs better than IRESpred in accuracy, sensitivity (recall), specificity, precision and MCC.

	Validation Dataset		Equation
	IRESpred	IRESpy	
Accuracy (%)	52.5%	77.8%	ACC = (TP + TN) / (P + N)
Sensitivity (%)	62.5%	79.6%	TPR = TP / (TP + FN)
Specificity (%)	42.5%	61.8%	SPC = TN / (FP + TN)
Precision (%)	52.1%	94.8%	PPV = TP / (TP + FP)
MCC	0.0510	0.2900	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

Table 4-2 Hyper-parameters tune guide.

Hyper Parameter	Tuning Approach	Range	
Number of trees	Fixed value	100-2000	Depending on data size and the averaging testing error by cross validation
Eta	Stepwise Optimized/ Grid search	[0.001,0.01,0.1]	
Max_depth	Stepwise Optimized/ Grid search	[1,3,5,7,9]	
min_child_weight	Stepwise Optimized/ Grid search	[9,19,29,39,49]	3/(% of rare events), rule of thumb
colsample_bytree	Stepwise Optimized/ Grid search	[0.6,0.7,0.8]	
sub_sample	Stepwise Optimized/ Grid search	[0.7,0.8,0.9]	
Gamma	Stepwise Optimized/ Grid search	[0,0.1, 0.3]	
Alpha, Lambda	Stepwise Optimized/ Grid search	Alpha=[0,0.1,0.2], Lambda=[1,2,3]	

Table 4-3 Previous IRES prediction model summary.

	Dataset	Features	Methods	Pros & Cons
VIPS	4 Types of viral IRES as positive, coding sequence as negative	N/A	Align predicted secondary structure	1. Can only predict viral IRES but not cellular IRES. 2. Wrong negative training dataset. 3. Low true positive for novel sequence.
IRESPred	Known IRES as positive, coding sequence and housekeeping UTR as negative	UTR length, # of AUGs, hairpin-loops, MFE, predicted interaction probabilities between UTR and SSRP	Support Vector Machine	1. Improper use of features leads to high false positive rate in longer sequence. 2. Training dataset is too small. 3. Wrong negative training dataset.
IRES-intepreter	55000 Synthetic sequences	6120 Kmer features	Gradient Boosting Decision Tree (GBDT)	1. Too many features for training: the global, local kmers have high correlation. 2. Model training time is way too slow which takes several days.
IRESfinder	Human IRES and non IRES in the 55000 Synthetic sequences	19 Kmer features	Logit Model	1. Designed only for cellular IRES. 2. Independent testing dataset is too small. 3. Logit model which not work well on non-linear relationships.

Table 4-4 Top 20 predicted human UTRs by IRESpy.

UTRdb_id	Reference Sequence	Gene symbol	Predicted_P	IRES_activity
5HSAA053317	NM_023015	integrator complex subunit 3 (INTS3)	0.8011613	N/A
5HSAA088788	NM_012294	Rap guanine nucleotide exchange factor (GEF) 5 (RAPGEF5)	0.7756774	N/A
5HSAA049422	NM_002121	major histocompatibility complex, class II, DP beta 1(HLA-DPB1)	0.7600551	N/A
5HSAA106226	NM_001135052(NM_001174167)	spleen tyrosine kinase (SYK)	0.7551038	1378.92
5HSAA061003	NM_000627	latent transforming growth factor beta binding protein 1 (LTBP1)	0.7488986	206.29
5HSAA090748	NM_000324	Rh-associated glycoprotein (RHAG)	0.7478946	N/A
5HSAA103586	NM_001024209	small proline-rich protein 2E (SPRR2E)	0.7406377	N/A
5HSAA089416	NM_006743	RNA binding motif (RNP1, RRM) protein 3 (RBM3)	0.7368885	N/A
5HSAA086413	NM_007039	protein tyrosine phosphatase, non-receptor type 21 (PTPN21)	0.7365369	N/A
5HSAA115306	NM_153235	taxilin beta (TXLNB)	0.7360649	2109.43
5HSAA003064	NM_031900	alanine-glyoxylate aminotransferase 2 (AGXT2), nuclear gene encoding mitochondrial protein	0.733718	N/A
5HSAA019728	NM_001040031	CD37 molecule (CD37)	0.7291359	N/A
5HSAA028545	NM_000555	doublecortin (DCX)	0.7285784	N/A
5HSAA108294	NM_153046	tudor domain containing 9 (TDRD9)	0.7167501	N/A
5HSAA058893	NM_001014434	LIM homeobox 9 (LHX9)	0.7088233	N/A
5HSAA019725	NM_001040031	CD37 molecule (CD37)	0.707556	N/A
5HSAA082471	NM_007055	polymerase (RNA) III (DNA directed) polypeptide A, 155kDa (POLR3A)	0.7065421	N/A
5HSAA098269	NM_173354	salt-inducible kinase 1 (SIK1)	0.7060764	N/A
5HSAA084551	NM_002734	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1) (PRKAR1A)	0.7036859	N/A
5HSAA107274	NM_001143964	TBC1 domain family, member 7 (TBC1D7)	0.700797	N/A
5HSAA071005	NM_016250	NDRG family member 2 (NDRG2)	0.7001775	N/A

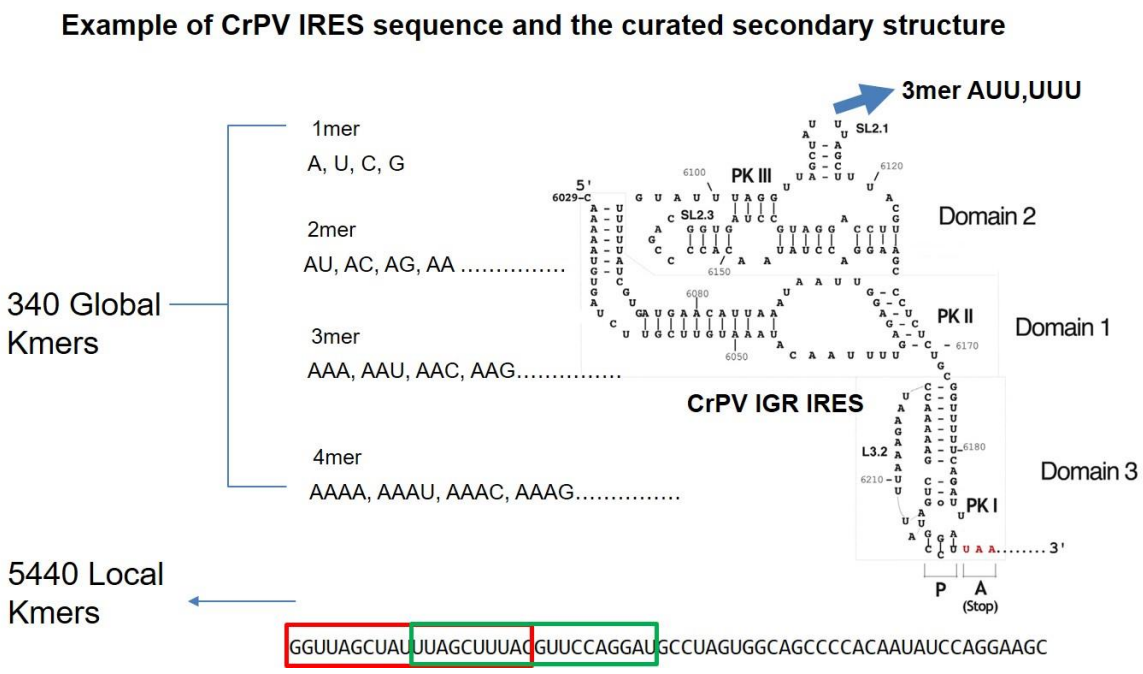


Figure 4-1 Calculation of Kmer features. An example of kmer features in the Cricket paralysis virus (CrPV) intergenic region (IGR) are shown. From 1mer to 4mer examples are shown. The red and green boxes show examples of the observation window used to calculate local kmers. 340 global kmers and 5440 local kmers have been tested in this research.

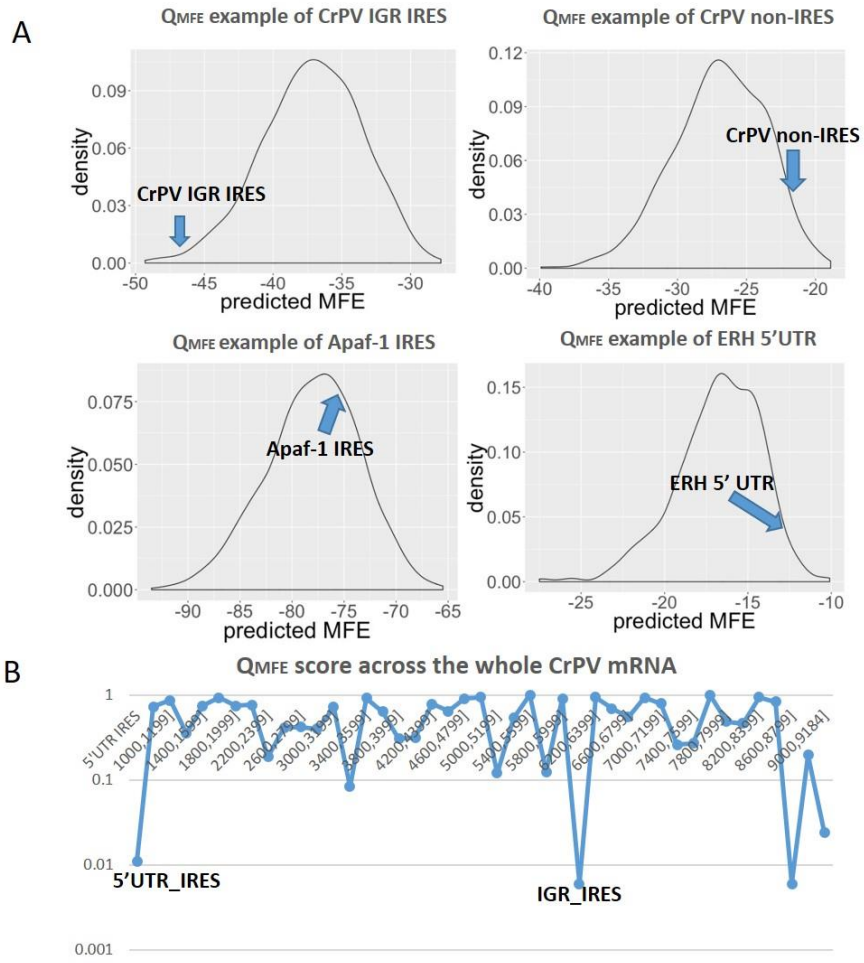


Figure 4-2 QMFE calculation examples of IRES and non-IRES sequences. A. PMFE of randomized sequences (density plot) and PMFE of the CrPV IGR IRES (viral IRES, PMFE=-47.5, QMFE=0.001), the ERH 5' UTR (housekeeping gene, PMFE=-12.7, QMFE=0.99), Apaf-1 cellular IRES (PMFE=-76, QMFE=0.66), and CrPV non-IRES regions (position: 6200-6399, PMFE=-22.2, QMFE=0.94). B. QMFE of 200 base segments across the whole genomic CrPV mRNA. The QMFE shows minimal values in the regions of the known the 5'UTR IRES (bases 1-708) and IGR IRES (bases 6000-6200).

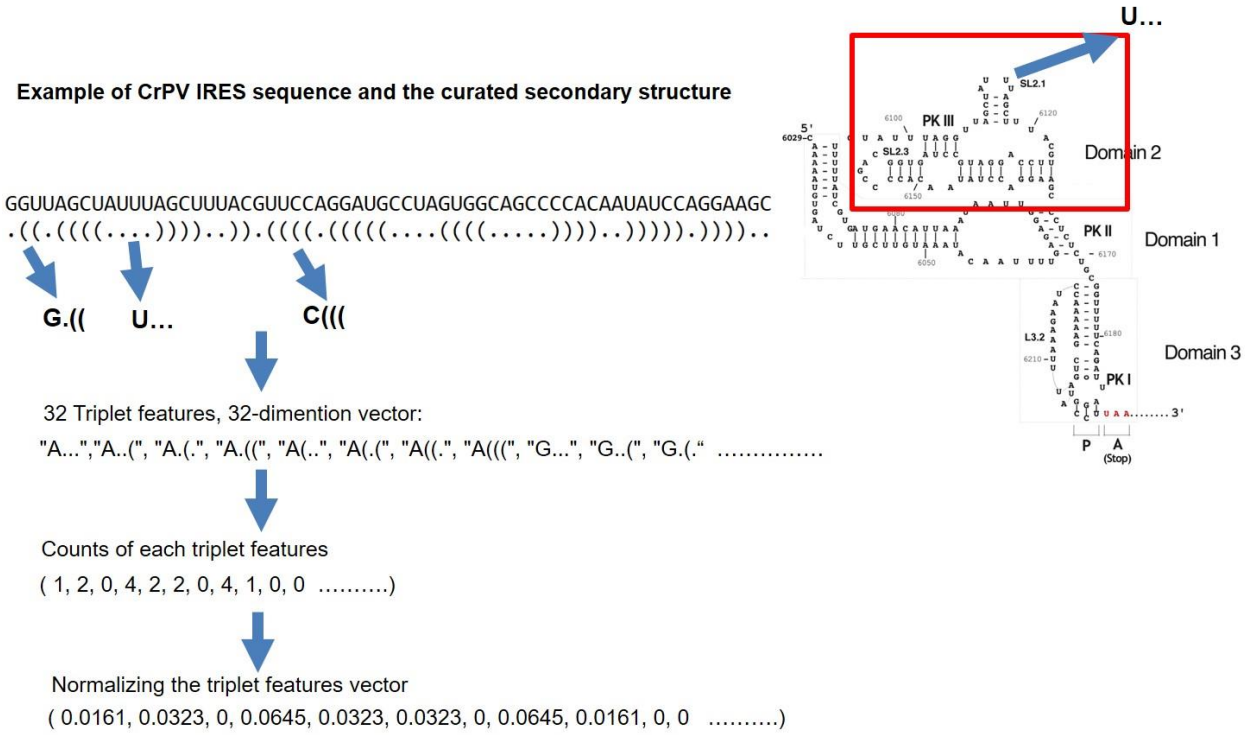


Figure 4-3 Calculation of triplet features. An example of triplet features in the Cricket paralysis virus (CrPV) intergenic region (IGR) are shown. The secondary structure of the candidate sequence was predicted using UNAfold (Markham & Zuker, 2008). For each nucleotide, only two states are possible, paired or unpaired. Parenthesess "( )" or dots "." represent the paired and unpaired nucleotides in the predicted secondary structure, respectively. For any 3 adjacent bases, there are 8 possible structural states: "(((", "((.", "(..","(.(","((.",".(.","..(", and "…". Triplet features comprise the structural states plus the identity of the central base, A, C, G, or U, so there are 32 (8*4=32) triplet features in total. Triplet features are normalized by dividing the observed number of each triplet by the total number of all the triplet features.

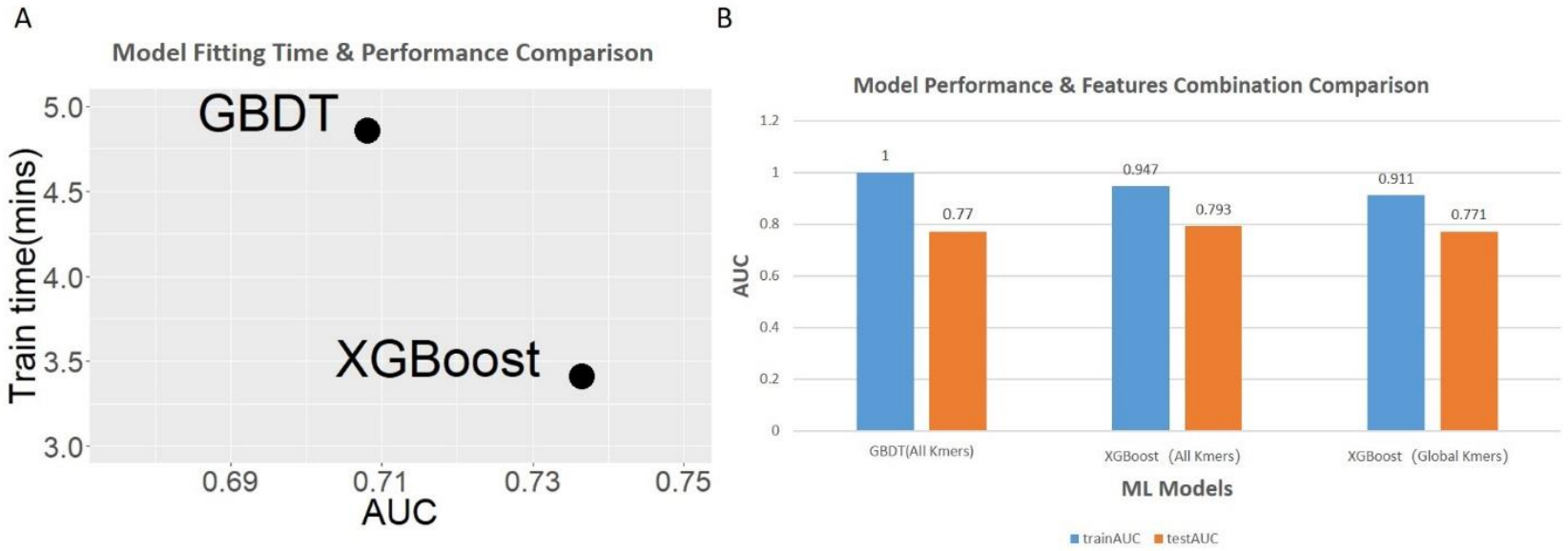


Figure 4-4 Model performance of XGBoost and GBDT. A. The model performance of XGBoost and GBDT for only the global kmer features, without any hyperparameter tuning. B. Model performance comparison using area under the ROC curve (AUC). The XGBoost model has lower training AUC but higher testing AUC than the GBDT model. The XGBoost model trained with only local mers performs the same as the GBDT model, but the number of features is reduced from 5780 to 340.

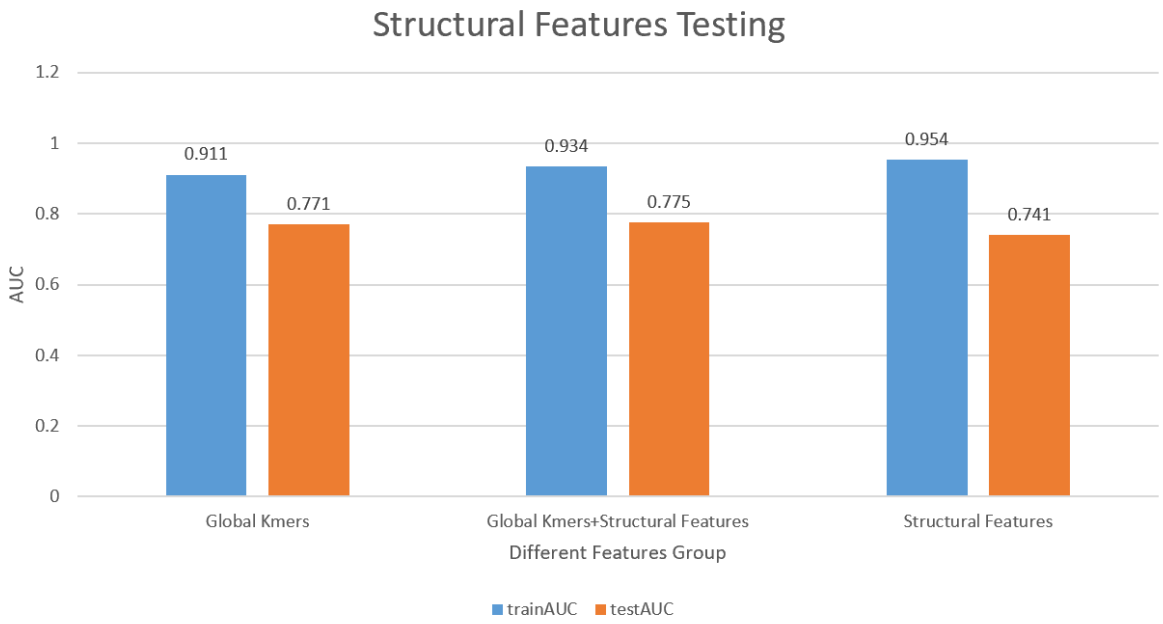


Figure 4-5 Effect of incorporating structural features. QMFE and triplet features were included in a combined model with global kmer features. We examined models incorporating only global kmer features, only structural features, and a combination of global kmer and structural features.

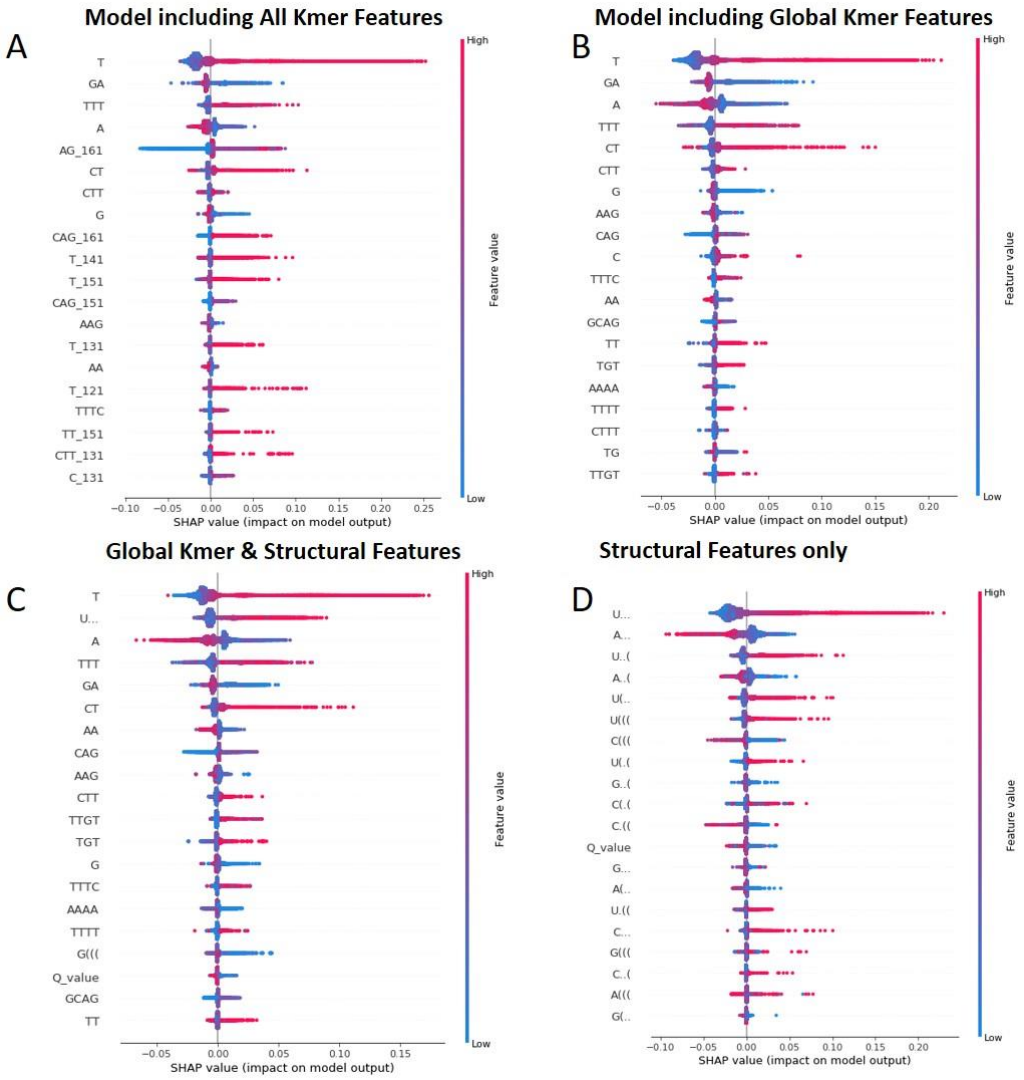


Figure 4-6 XGBoost model feature importance explained by SHAP values at the global scale. A. The summary of SHAP values of the top 20 important features for model including both global kmers and local kmers. B. The summary of SHAP values of the top 20 important features for models including only global kmers. C. The summary of SHAP values of the top 20 important features for models including both global kmers and structural features. D. The summary of SHAP value of the top 20 important features for model including only structural features.

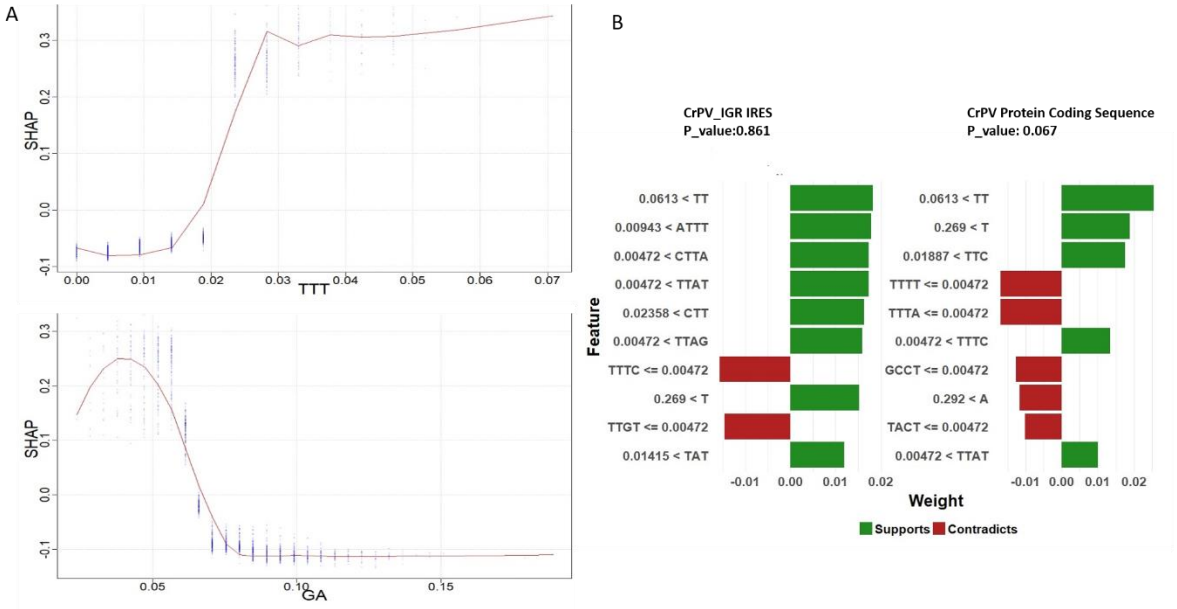


Figure 4-7 XGBoost model feature importance explained by SHAP and LIME at a local scale. A. SHAP (SHapley Additive exPlanation) dependence plots of the importance of the UUU and GA kmers in the XGBoost model. B. Local Interpretable Model-agnostic Explanations (LIME) for the CrPV IGR IRES and CrPV protein coding sequence. The green bar shows the weighted features that support classification as IRES and red bars are the weighted features that oppose

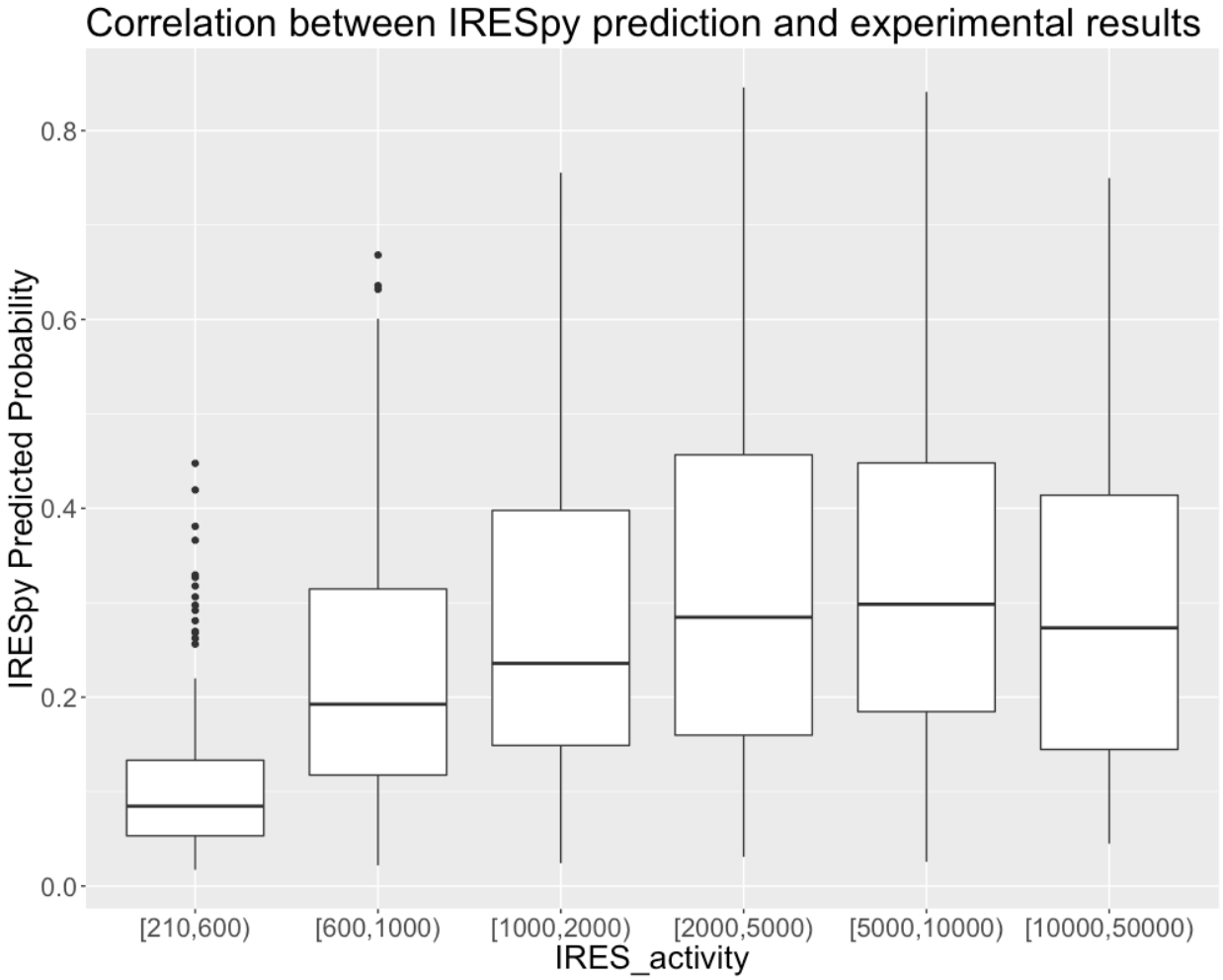


Figure 4-8 Correlation between IRESpy prediction and experimental results.

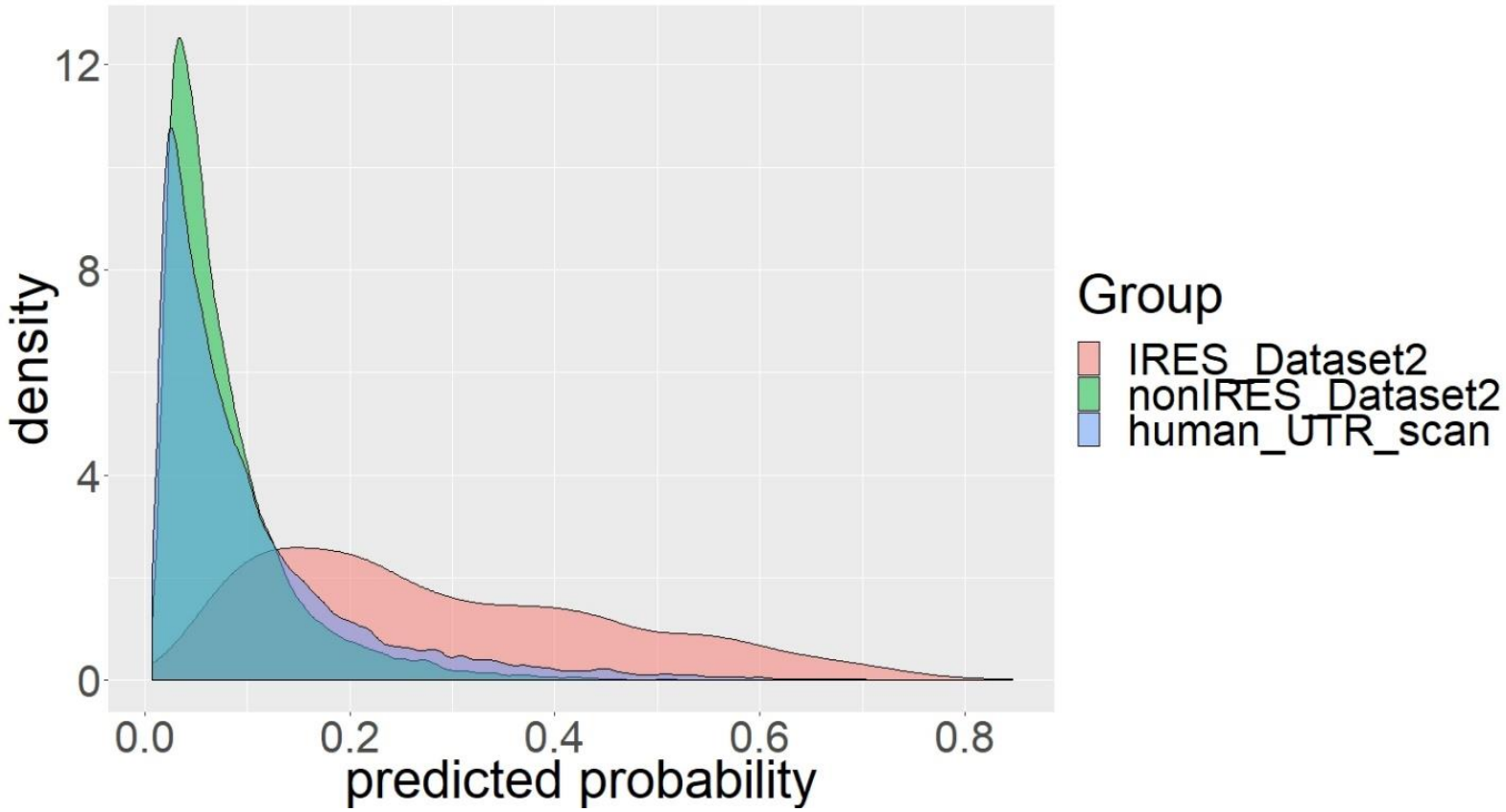


Figure 4-9 The density distribution of predicted IRES probability in Dataset 2 and human UTR scan.

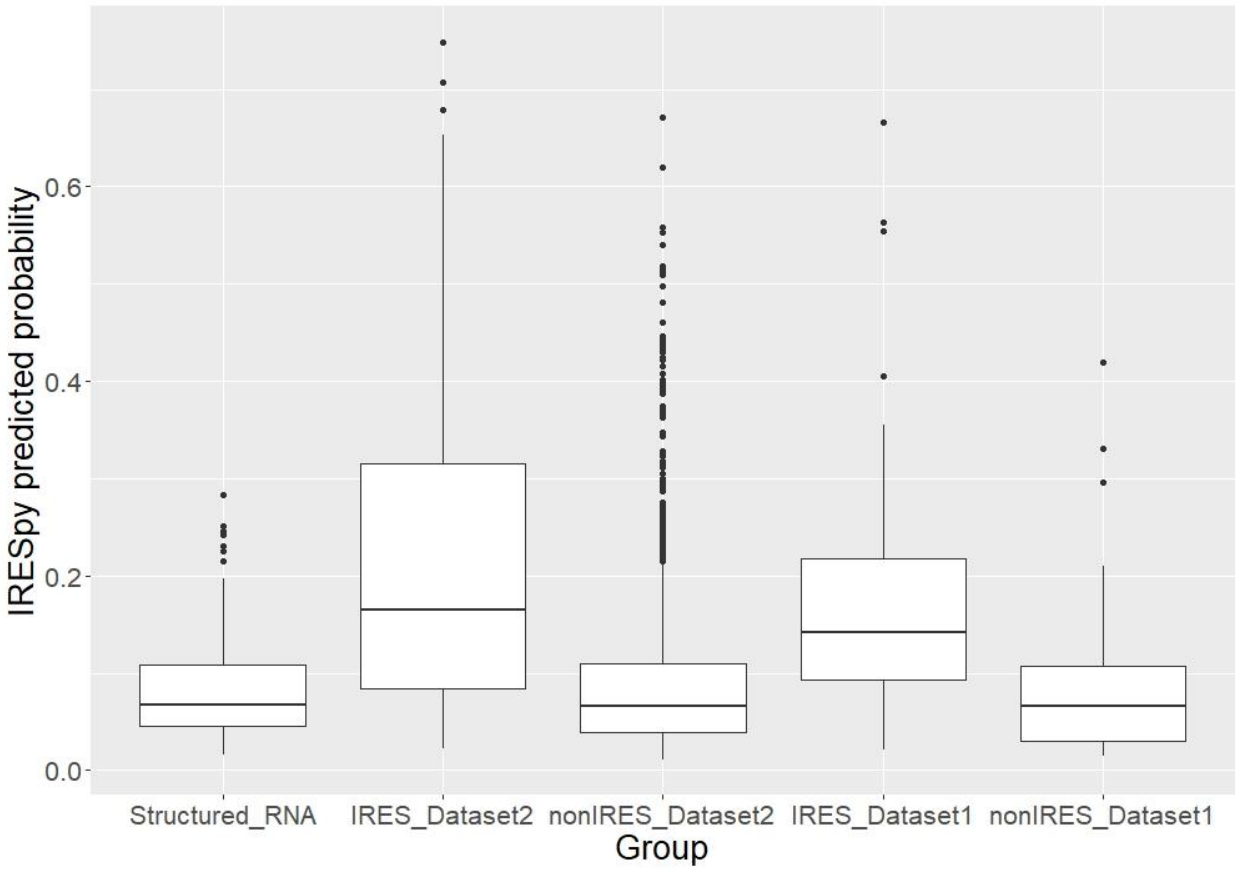


Figure 4-10 Predicted probability of IRES for highly structured RNA families, and IRES and non-IRES classes in Datasets 1 and 2.

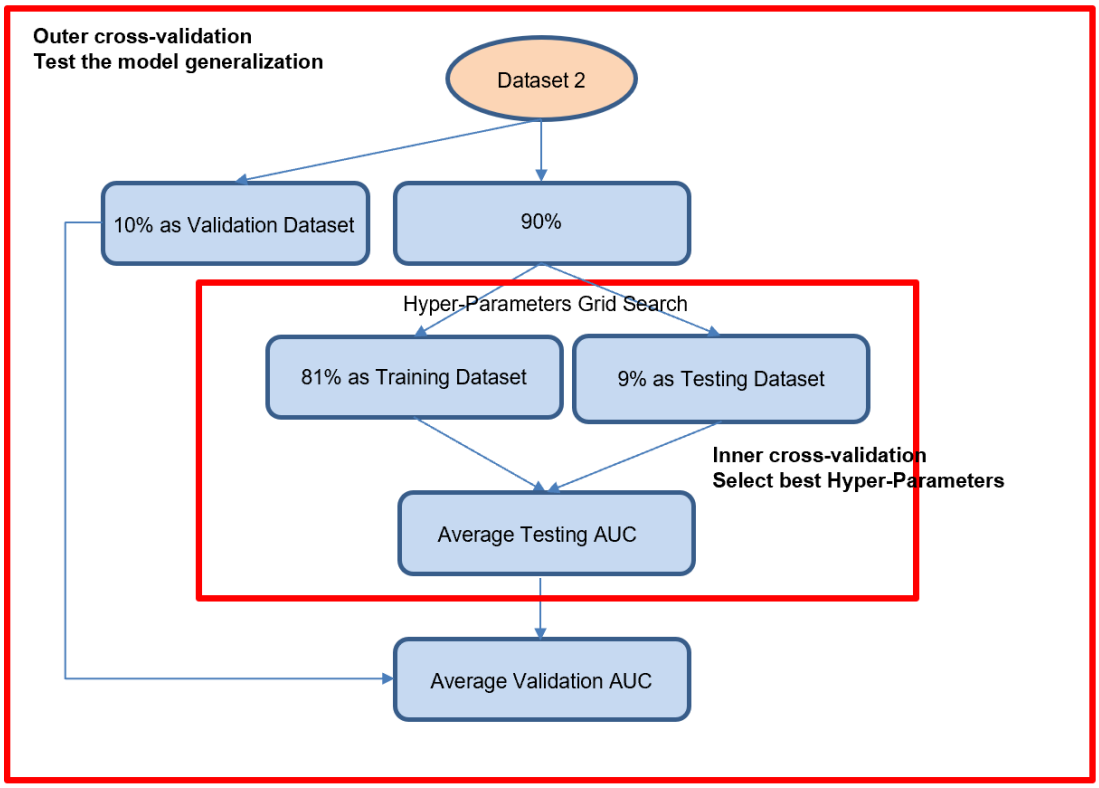


Figure 4-11 Nested cross validation design map.

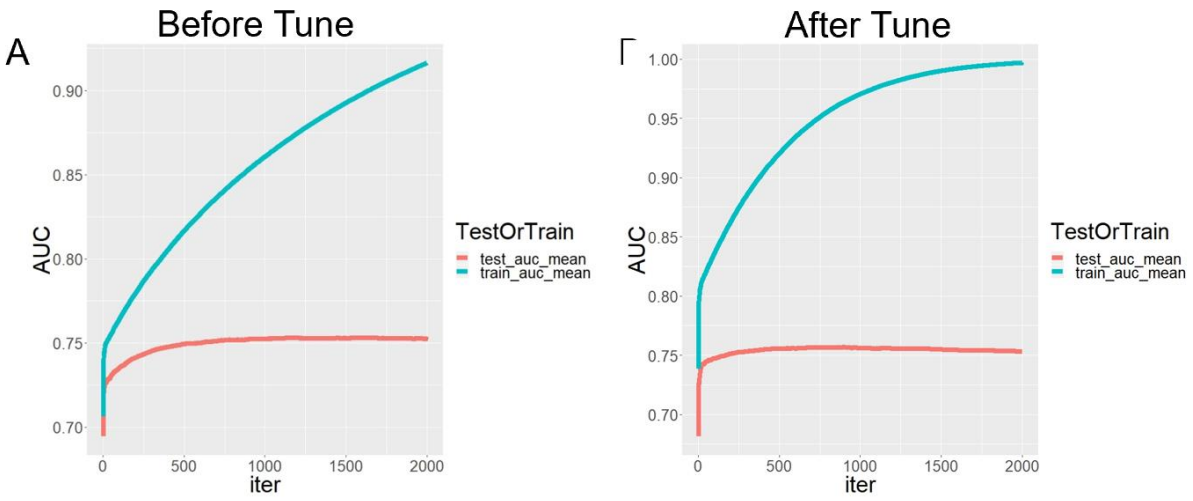


Figure 4-12 The comparison of Inner loop cross validation performance as trees get bigger before hyper-parameters tuning and after hyper-parameters tuning. Iter, the X-axis, indicates the number of trees. (A). The initial parameters before tuning were: eta=0.01, max_depth=3, min_child_weight=29, colsample_bytree=0.8, sub-sample= 0.8, gamma=0, alpha=0 and lambda=1. The largest test AUC was 0.752, which is obtained when the number of trees is 1661. (B). The parameters after tuning are: eta = 0.01, max_depth=5, min_child_weight=19, subsample=0.8, colsample_bytree=0.7, gamma=0, alpha=0 and lambda=1. The largest test AUC is 0.756 when the number of trees is 901.

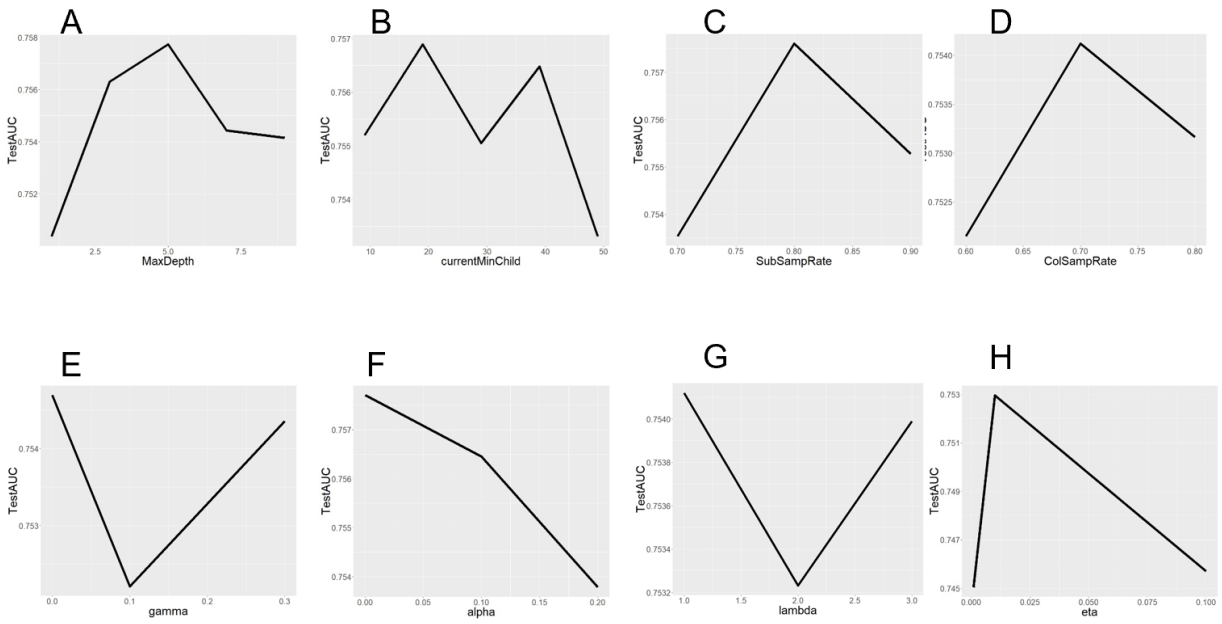


Figure 4-13 Tree-related hyper-parameter tune results. The effect of varying each parameter separately with the final tuned parameters is shown. (A). Max_depth=[1,3,5,7,9] (B). min_child_weight=[9,19,29,39,49] (C). sub_sample= [0.7,0.8,0.9] (D). colsample_bytree=[0.6,0.7,0.8] (E). gamma=[0,0.1, 0.3] (F). Alpha=[0,0.1,0.2] (G). Lambda=[1,2,3] (A). eta=[0.001,0.01,0.1]

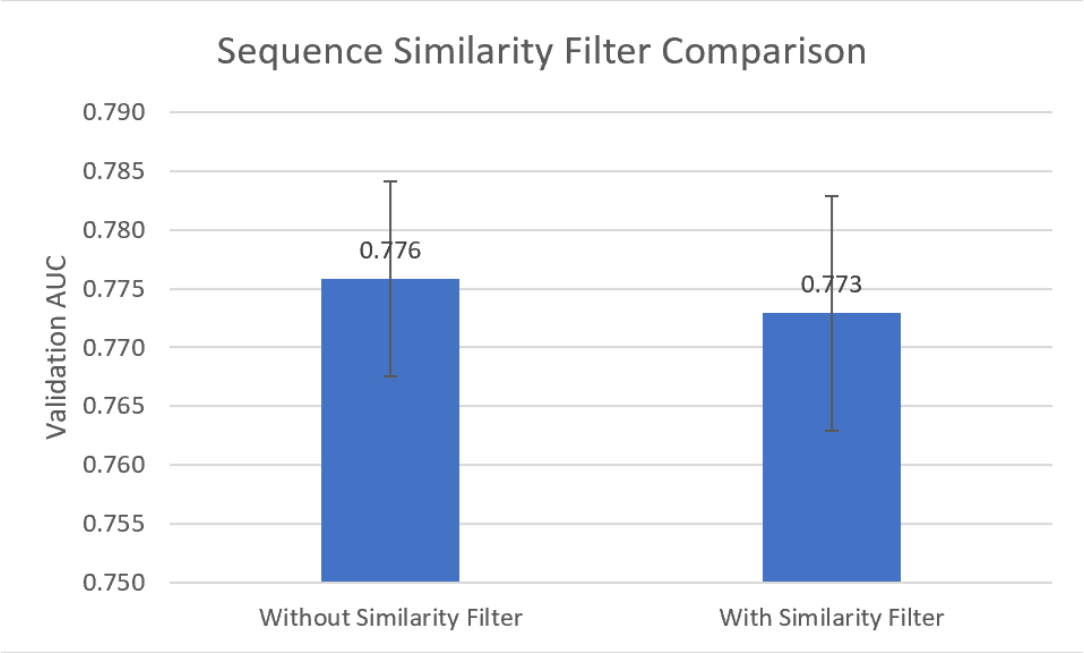


Figure 4-14 The validation AUC comparison between filtering the 80% sequence similarity and no-filtering the sequence at all.

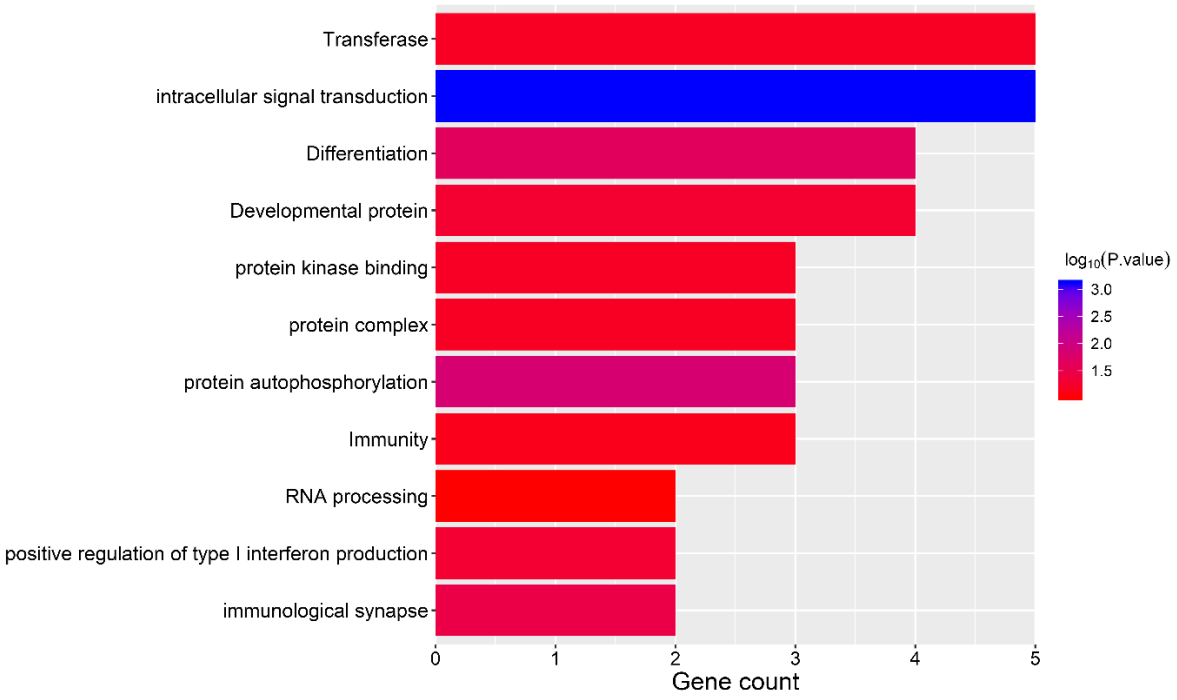


Figure 4-15 The Gene ontology analysis of the top 20 predicted human UTRs by David 6.8.

REFERENCES

Baird, S. D., Turcotte, M., Korneluk, R. G., & Holcik, M. (2006). Searching for IRES. *RNA, 12*(10), 1755-1785. doi:10.1261/rna.157806

Bonnal, S., Schaeffer, C., Creancier, L., Clamens, S., Moine, H., Prats, A. C., & Vagner, S. (2003). A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons. *J Biol Chem, 278*(41), 39330-39336. doi:10.1074/jbc.M305580200

Bonnet, E., Wuyts, J., Rouzé, P., & Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. . *Bioinformatics, 20(17)*, 2911-2917.

Chappell, S. A., Edelman, G. M., & Mauro, V. P. (2000). A 9-nt segment of a cellular mRNA can function as an internal ribosome entry site (IRES) and when present in linked multiple copies greatly enhances IRES activity. *Proc Natl Acad Sci U S A, 97*(4), 1536-1541.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794*.

Clemens, M. J., Bushell, M., Jeffrey, I. W., Pain, V. M., & Morley, S. J. (2000). Translation initiation factor modifications and the regulation of protein synthesis in apoptotic cells. *Cell Death Differ, 7*(7), 603-615. doi:10.1038/sj.cdd.4400695

Clote, P., Ferre, F., Kranakis, E., & Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA, 11*(5), 578-591. doi:10.1261/rna.7220505

Cornelis, S., Bruynooghe, Y., Denecker, G., Van Huffel, S., Tinton, S., & Beyaert, R. (2000). Identification and characterization of a novel cell cycle-regulated internal ribosome entry site. *Mol Cell, 5*(4), 597-605.

Degroeve, S., De Baets, B., Van de Peer, Y., & Rouze, P. (2002). Feature subset selection for splice site prediction. *Bioinformatics, 18 Suppl 2*, S75-83.

Dirks, R. M., Lin, M., Winfree, E., & Pierce, N. A. (2004). Paradigms for computational nucleic acid design. *Nucleic Acids Res, 32*(4), 1392-1403. doi:10.1093/nar/gkh291

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends Genet, 29*(10), 569-574. doi:10.1016/j.tig.2013.05.010

Fernandez-Miragall, O., & Martinez-Salas, E. (2003). Structural organization of a viral IRES depends on the integrity of the GNRA motif. *RNA, 9*(11), 1333-1344.

Fernandez, J., Yaman, I., Mishra, R., Merrick, W. C., Snider, M. D., Lamers, W. H., & Hatzoglou, M. (2001). Internal ribosome entry site-mediated translation of a mammalian mRNA is regulated by amino acid availability. *J Biol Chem, 276*(15), 12285-12291. doi:10.1074/jbc.M009714200

Filbin, M. E., & Kieft, J. S. (2009). Toward a structural understanding of IRES RNA function. *Curr Opin Struct Biol, 19*(3), 267-276. doi:10.1016/j.sbi.2009.03.005

Grillo, G., Turi, A., Licciulli, F., Mignone, F., Liuni, S., Banfi, S., . . . Pesole, G. (2010). UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res, 38*(Database issue), D75-80. doi:10.1093/nar/gkp902

Gritsenko, A. A., Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., de Ridder, D., & Segal, E. (2017). Sequence features of viral and human Internal Ribosome Entry Sites predictive of their activity. *PLoS Comput Biol, 13*(9), e1005734. doi:10.1371/journal.pcbi.1005734

Han, B., & Zhang, J. T. (2002). Regulation of gene expression by internal ribosome entry sites or cryptic promoters: the eIF4G story. *Mol Cell Biol, 22*(21), 7372-7384.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., . . . Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet, 39*(3), 311-318. doi:10.1038/ng1966

Henis-Korenblit, S., Strumpf, N. L., Goldstaub, D., & Kimchi, A. (2000). A novel form of DAP5 protein accumulates in apoptotic cells as a result of caspase cleavage and internal ribosome entry site-mediated translation. *Mol Cell Biol, 20*(2), 496-506.

Hennecke, M., Kwissa, M., Metzger, K., Oumard, A., Kroger, A., Schirmbeck, R., . . . Hauser, H. (2001). Composition and arrangement of genes define the strength of IRES-driven translation in bicistronic mRNAs. *Nucleic Acids Res, 29*(16), 3327-3334.

Hershey, J. W., Sonenberg, N., & Mathews, M. B. (2012). Principles of translational control: an overview. *Cold Spring Harb Perspect Biol, 4*(12). doi:10.1101/cshperspect.a011528

Hinnebusch, A. G. (2014). The scanning mechanism of eukaryotic translation initiation. *Annu Rev Biochem, 83*, 779-812. doi:10.1146/annurev-biochem-060713-035802

Holcik, M., Yeh, C., Korneluk, R. G., & Chow, T. (2000). Translational upregulation of X-linked inhibitor of apoptosis (XIAP) increases resistance to radiation induced cell death. *Oncogene, 19*(36), 4174-4177. doi:10.1038/sj.onc.1203765

Hong, J. J., Wu, T. Y., Chang, T. Y., & Chen, C. Y. (2013). Viral IRES prediction system - a web server for prediction of the IRES secondary structure in silico. *PLoS One, 8*(11), e79288. doi:10.1371/journal.pone.0079288

Huang, J., Li, K., & Gribskov, M. (2016). Accurate Classification of RNA Structures Using Topological Fingerprints. *PLoS One, 11*(10), e0164726. doi:10.1371/journal.pone.0164726

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics, 26*(5), 680-682. doi:10.1093/bioinformatics/btq003

Hung, C. Y., Yang, W. B., Wang, S. A., Hsu, T. I., Chang, W. C., & Hung, J. J. (2014). Nucleolin enhances internal ribosomal entry site (IRES)-mediated translation of Sp1 in tumorigenesis. *Biochim Biophys Acta, 1843*(12), 2843-2854. doi:10.1016/j.bbamcr.2014.08.009

Jang, S. K., Krausslich, H. G., Nicklin, M. J., Duke, G. M., Palmenberg, A. C., & Wimmer, E. (1988). A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol, 62*(8), 2636-2643.

Jiang, M., Anderson, J., Gillespie, J., & Mayne, M. (2008). uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics, 9*, 192. doi:10.1186/1471-2105-9-192

Jo, O. D., Martin, J., Bernath, A., Masri, J., Lichtenstein, A., & Gera, J. (2008). Heterogeneous nuclear ribonucleoprotein A1 regulates cyclin D1 and c-myc internal ribosome entry site function through Akt signaling. *J Biol Chem, 283*(34), 23274-23287. doi:10.1074/jbc.M801185200

Kolekar, P., Pataskar, A., Kulkarni-Kale, U., Pal, J., & Kulkarni, A. (2016). IRESPred: Web Server for Prediction of Cellular and Viral Internal Ribosome Entry Site (IRES). *Sci Rep, 6*, 27436. doi:10.1038/srep27436

Komar, A. A., & Hatzoglou, M. (2005). Internal ribosome entry sites in cellular mRNAs: The mystery of their existence. *Journal of Biological Chemistry*.

Komar, A. A., & Hatzoglou, M. (2015). Exploring Internal Ribosome Entry Sites as Therapeutic Targets. *Front Oncol, 5*, 233. doi:10.3389/fonc.2015.00233

Komar, A. A., Mazumder, B., & Merrick, W. C. (2012). A new framework for understanding IRES-mediated translation. *Gene, 502*(2), 75-86. doi:10.1016/j.gene.2012.04.039

Kozak, M. (2005). A second look at cellular mRNA sequences said to function as internal ribosome entry sites. *Nucleic Acids Res, 33*(20), 6593-6602. doi:10.1093/nar/gki958

Li, W., Jaroszewski, L., & Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics, 18*(1), 77-82.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat Rev Genet, 16*(6), 321-332. doi:10.1038/nrg3920

Lozano, G., Fernandez, N., & Martinez-Salas, E. (2016). Modeling Three-Dimensional Structural Motifs of Viral IRES. *J Mol Biol, 428*(5 Pt A), 767-776. doi:10.1016/j.jmb.2016.01.005

Macejak, D. G., & Sarnow, P. (1991). Internal initiation of translation mediated by the 5' leader of a cellular mRNA. *Nature, 353*(6339), 90-94. doi:10.1038/353090a0

Mailliot, J., & Martin, F. (2018). Viral internal ribosomal entry sites: four classes for one goal. *Wiley Interdiscip Rev RNA, 9*(2). doi:10.1002/wrna.1458

Markham, N. R., & Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol, 453*, 3-31. doi:10.1007/978-1-60327-429-6_1

Martinand-Mari, C., Lebleu, B., & Robbins, I. (2003). Oligonucleotide-based strategies to inhibit human hepatitis C virus. *Oligonucleotides, 13*(6), 539-548. doi:10.1089/154545703322860834

Mattick, J. S. (2018). The State of Long Non-Coding RNA Biology. *Noncoding RNA, 4*(3). doi:10.3390/ncrna4030017

Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. *Hum Mol Genet, 15 Spec No 1*, R17-29. doi:10.1093/hmg/ddl046

Meerovitch, K., Pelletier, J., & Sonenberg, N. (1989). A cellular protein that binds to the 5'-noncoding region of poliovirus RNA: implications for internal translation initiation. *Genes Dev, 3*(7), 1026-1034.

Mokrejs, M., Masek, T., Vopalensky, V., Hlubucek, P., Delbos, P., & Pospisek, M. (2010). IRESite--a tool for the examination of viral and cellular internal ribosome entry sites. *Nucleic Acids Res, 38*(Database issue), D131-136. doi:10.1093/nar/gkp981

Mokrejs, M., Vopalensky, V., Kolenaty, O., Masek, T., Feketova, Z., Sekyrova, P., . . . Pospisek, M. (2006). IRESite: the database of experimentally verified IRES structures (www.iresite.org). *Nucleic Acids Res, 34*(Database issue), D125-130. doi:10.1093/nar/gkj081

Morrish, B. C., & Rumsby, M. G. (2002). The 5' untranslated region of protein kinase Cdelta directs translation by an internal ribosome entry segment that is most active in densely growing cells and during apoptosis. *Mol Cell Biol, 22*(17), 6089-6099.

Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of electronic imaging, 16(4)*(049901).

Nevins, T. A., Harder, Z. M., Korneluk, R. G., & Holcik, M. (2003). Distinct regulation of internal ribosome entry site-mediated translation following cellular stress is mediated by apoptotic fragments of eIF4G translation initiation factor family members eIF4GI and p97/DAP5/NAT1. *J Biol Chem, 278*(6), 3572-3579. doi:10.1074/jbc.M206781200

Nulf, C. J., & Corey, D. (2004). Intracellular inhibition of hepatitis C virus (HCV) internal ribosomal entry site (IRES)-dependent translation by peptide nucleic acids (PNAs) and locked nucleic acids (LNAs). *Nucleic Acids Res, 32*(13), 3792-3798. doi:10.1093/nar/gkh706

Ohler, U., Liao, G. C., Niemann, H., & Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. *Genome Biol, 3*(12), RESEARCH0087.

Olson, R. S., Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput, 23*, 192-203.

Pelletier, J., & Sonenberg, N. (1988). Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature, 334*(6180), 320-325. doi:10.1038/334320a0

Plank, T. D., & Kieft, J. S. (2012). The structures of nonprotein-coding RNAs that drive internal ribosome entry site function. *Wiley Interdiscip Rev RNA, 3*(2), 195-212. doi:10.1002/wrna.1105

Rocchi, L., Pacilli, A., Sethi, R., Penzo, M., Schneider, R. J., Trere, D., . . . Montanaro, L. (2013). Dyskerin depletion increases VEGF mRNA internal ribosome entry site-mediated translation. *Nucleic Acids Res, 41*(17), 8308-8318. doi:10.1093/nar/gkt587

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., . . . Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature, 473*(7347), 337-342. doi:10.1038/nature10098

Sharathchandra, A., Katoch, A., & Das, S. (2014). IRES mediated translational regulation of p53 isoforms. *Wiley Interdiscip Rev RNA, 5*(1), 131-139. doi:10.1002/wrna.1202

Spriggs, K. A., Bushell, M., Mitchell, S. A., & Willis, A. E. (2005). Internal ribosome entry segment-mediated translation during apoptosis: the role of IRES-trans-acting factors. *Cell Death Differ, 12*(6), 585-591. doi:10.1038/sj.cdd.4401642

Stoneley, M., Chappell, S. A., Jopling, C. L., Dickens, M., MacFarlane, M., & Willis, A. E. (2000). c-Myc protein synthesis is initiated from the internal ribosome entry segment during apoptosis. *Mol Cell Biol, 20*(4), 1162-1169.

Stoneley, M., & Willis, A. E. (2004). Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression. *Oncogene, 23*(18), 3200-3207. doi:10.1038/sj.onc.1207551

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform, 4*(3), 159-169. doi:10.1007/s40708-017-0065-7

Trotta, E. (2014). On the normalization of the minimum free energy of RNAs by sequence length. *PLoS One, 9*(11), e113380. doi:10.1371/journal.pone.0113380

Van Eden, M. E., Byrd, M. P., Sherrill, K. W., & Lloyd, R. E. (2004). Demonstrating internal ribosome entry sites in eukaryotic mRNAs using stringent RNA test procedures. *RNA, 10*(4), 720-730.

Vitsios, D. M., Kentepozidou, E., Quintais, L., Benito-Gutierrez, E., van Dongen, S., Davis, M. P., & Enright, A. J. (2017). Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Res, 45*(21), e177. doi:10.1093/nar/gkx836

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., & Li, Y. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics, 21*(18), 3610-3614. doi:10.1093/bioinformatics/bti562

Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., Gritsenko, A. A., Stern-Ginossar, N., Yakhini, Z., . . . Segal, E. (2016). Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science, 351*(6270). doi:10.1126/science.aad4939

Wu, T. Y., Hsieh, C. C., Hong, J. J., Chen, C. Y., & Tsai, Y. S. (2009). IRSS: a web-based tool for automatic layout and analysis of IRES secondary structure prediction and searching system in silico. *BMC Bioinformatics, 10*, 160. doi:10.1186/1471-2105-10-160

Xue, C., Li, F., He, T., Liu, G. P., Li, Y., & Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics, 6*, 310. doi:10.1186/1471-2105-6-310

Yakovchuk, P., Protozanova, E., & Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res, 34*(2), 564-574. doi:10.1093/nar/gkj454

Zhao, J., Wu, J., Xu, T., Yang, Q., He, J., & Song, X. (2018). IRESfinder: Identifying RNA internal ribosome entry site in eukaryotic cell using framed k-mer features. *J Genet Genomics*. doi:10.1016/j.jgg.2018.07.006

Zhu, Y., Huang, P., Yang, N., Liu, R., Liu, X., Dai, H., . . . Sun, C. (2017). Establishment and Application of a High Throughput Screening System Targeting the Interaction between HCV Internal Ribosome Entry Site and Human Eukaryotic Translation Initiation Factor 3. *Front Microbiol, 8*, 977. doi:10.3389/fmicb.2017.00977

Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research, 9(1)*, 133-148.

PUBLICATIONS

Chen, Y., Wang, J., Yang, S., Utturkar, S., Crodian, J., Cummings, S., . . . Casey, T. (2017). Effect of high-fat diet on secreted milk transcriptome in midlactation mice. *Physiol Genomics, 49*(12), 747-762. doi:10.1152/physiolgenomics.00080.2017

APS select award certificate (http://apsselect.physiology.org/), a collection from the APS that showcases some of the best recently published articles in physiological research.

Poster presented in the 39TH ANNUAL MIDWEST BIOPHARMACEUTICAL STATISTICS WORKSHOP, “Identification, Classification, and Prediction of Functional RNA by machine learning methods”

Wang, J., Gribskov, M.: IRESpy: An XGBoost model for prediction of Internal Ribosome Entry Sites. **(Submitted)**