SEARCHING THE EDGES OF THE PROTEIN UNIVERSE USING DATA SCIENCE

by

Mengmeng Zhu

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 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Michael Gribskov, Chair

Department of Biological Sciences

Dr. Dan Goldwasser

Department of Computer Science

Dr. Faming Liang

Department of Statistics

Dr. Cynthia Stauffacher

Department of Biological Sciences

Approved by:

Dr. Jason R. Cannon and Dr. Janice P. Evans

I dedicate this work to the past, through which I learned upekkha.

ACKNOWLEDGMENTS

I would like to express my deepest and sincerest gratitude to my PhD advisor, Dr. Michael Gribskov. Dr. Gribskov has provided me tremendous support over the past years. He always respected my research ideas, provided resources whenever needed, and guided me whenever I felt lost. His open, cheerful, and witty attitudes toward life also inspired me. To me, he is an invaluable mentor and also a warm friend.

I would also like to thank Dr. Dan Goldwasser, Dr. Markus Lill, Dr. Faming Liang, and Dr. Cynthia Stauffacher for being in my committee and offering invaluable suggestions and guidance on my research. In addition, I would like to thank Dr. Daisuke Kihara for his generous mentorship and support in my earlier PhD research career. I would also like to thank my collaborator, Dr. Woong-Hee Shin for his generous and patient guidance in my earlier work.

Last but not least, I would like to thank anyone who have helped me, guided me, or whom I’ve learned from during my PhD career.

TABLE OF CONTENTS

LIST OF TABLES

Supplemental Tables

LIST OF FIGURES

Supplemental Figures

LIST OF ABBREVIATIONS

DAF	derived allele frequency
DCNN	deep convolutional neural network
FDA	fraction of rare derived alleles
FMA	raction of rare minor alleles
kb	1,000 nucleotide bases
MAF	minor allele frequency
miRNA	microRNA
MS	Mass spectrometry
nt	nucleotides
PDB	Protein Data Bank
piRNA	piwiRNA
rRNA	ribosomal RNA
scaRNA	small Cajal body RNA
SI	sequence identity
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
tRNA	transfer RNA
uORF	upstream ORF

ABSTRACT

Data science uses the latest techniques in statistics and machine learning to extract insights from data. With the increasing amount of protein data, a number of novel research approaches have become feasible.

Micropeptides are an emerging field in the protein universe. They are small proteins with <= 100 amino acid residues (aa) and are translated from small open reading frames (sORFs) of <= 303 base pairs (bp). Traditionally, their existence was ignored because of the technical difficulties in isolating them. With technological advances, a growing number of micropeptides have been characterized and shown to play vital roles in many biological processes. Yet, we lack bioinformatics methods for predicting them directly from DNA sequences, which could substantially facilitate research in this field with minimal cost. With the increasing amount of data, developing new methods to address this need becomes possible. We therefore developed MiPepid, a machine-learning-based method specifically designed for predicting micropeptides from DNA sequences by curating a high-quality dataset and by training MiPepid using logistic regression with 4-mer features. MiPepid performed exceptionally well on holdout test sets and performed much better than existing methods. MiPepid is available for downloading, easy to use, and runs sufficiently fast.

Long noncoding RNAs (LncRNAs) are transcripts of > 200 bp and does not encode a protein. Contrary to their “noncoding” definition, an increasing number of lncRNAs have been found to be translated into functional micropeptides. Therefore, whether most lncRNAs are translated is an open question of great significance. To address this question, by harnessing the availability of large-scale human variation data, we have explored the relationships between lncRNAs, micropeptides, and canonical regular proteins (> 100 aa) from the perspective of genetic variation, which has long been used to study natural selection to infer functional relevance. Through rigorous statistical analyses, we find that lncRNAs share a similar genetic variation profile with proteins regarding single nucleotide polymorphism (SNP) density, SNP spectrum, enrichment of rare SNPs, etc., suggesting lncRNAs are under similar negative selection strength with proteins. Our study revealed similarities between micropeptides, lncRNAs, and canonical proteins and is the first attempt to explore the relationships between the three groups from a genetic variation perspective.

Deep learning has been tremendously successful in 2D image recognition. Protein binding ligand prediction is fundamental topic in protein research as most proteins bind ligands to function. Proteins are 3D structures and can be considered as 3D images. Prediction of binding ligands of proteins can then be converted to a 3D image classification problem. In addition, a large number of protein structure data are available now. We therefore utilized deep learning to predict protein binding ligands by designing a 3D convolutional neural network from scratch and by building a large 3D image dataset of protein structures. The trained model achieved an average F1 score of over 0.8 across 151 classes on the holdout test set. Compared to existing methods, our model performed better. In summary, we showed the feasibility of deploying deep learning in protein structure research.

In conclusion, by exploring various edges of the protein universe from the perspective of data science, we showed that the increasing amount of data and the advancement of data science methods made it possible to address a wide variety of pressing biological questions. We showed that for a successful data science study, the three components – goal, data, method – all of them are indispensable. We provided three successful data science studies: the careful data cleaning and selection of machine learning algorithm lead to the development of MiPepid that fits the urgent need of a micropeptide prediction method; identifying the question and exploring it from a different angle lead to the key insight that lncRNAs resemble micropeptides; applying deep learning to protein structure data lead to a new approach to the long-standing question of protein-ligand binding. The three studies serve as excellent examples in solving a wide range of data science problems with a variety of issues.

CHAPTER 1. INTRODUCTION

1.1 Data science

1.1.1 What is data science?

As a buzz word in today's society, data science is gaining more and more attention. Although it lacks a clear definition, data science can be simply understood as a discipline emerging from big data. As the amount of data is growing exponentially in many fields, there is an urgent need to generate value from these huge amounts of data. And to generate value, the most direct way is to extract knowledge and insights from data (Dhar, 2013). Data science simultaneously employs statistics, data analysis, machine learning, etc. to maximize the value obtained from the data (Dhar, 2013).

1.1.2 The three components of a successful data science study

Data science is extracting knowledge and insight from data. Based on my experience, three components are indispensable to conduct a successful data science study. 1. An understanding of the type of knowledge or insight we would like to extract, i.e., before we employ data science, we need to have a clear goal in mind. 2. The data that we can obtain. Many methods require large amounts of data and, furthermore, the data must be crefully cleaned and curated. 3. An understanding of, and careful selection of data science methods / techniques that can successfully produce insights using the data we can obtain. These three components intertwine with each other. The lack of any one of them will make the effort in vain.

1. The end goal of a data science project

This depends on the need of the data scientists that work on the study. For data scientists that work on biological research, the goal is often to achieve a certain biological significance.

2. The data

A high-quality dataset is indispensable for a successful data science study. By high-quality it means the dataset (1) is representive of the population we are assessing; (2) is of sufficient size for the problem to be addressed, and the computational method to be employed (3) has minimal noise, i.e., has few mislabeled data or irrelevant data.

3. Data science methods

Broadly speaking, any method that could be to used to work with data can be called a data science method. Most data science methods come from stastistics, machine learning, or computer science.

1.1.3 Machine learning

Machine learning (ML) is a collection of algorithms for learning hidden patterns from a set of data in order to classify, cluster, etc. Machine learning is broadly categorized into supervised learning and unsupervised learning. In supervised learning, the labels of the data are known, and the goal is generally to learn the hidden mapping patterns between the data and the labels so that the learned patterns can be utilized to predict labels of future data. Many ML algorithms belong to supervised learning, e.g., regression (linear regression, logistic regression, generalized linear regression, etc.), support vector machine (SVM), tree-based methods (decision tree, random forest, etc.). In unsupervised learning, the labels of the data are generally unknown, and the goal is to better understand the data, usually by clustering the data. (Bishop, 2006)

1.1.4 Deep learning

Deep learning is a special type of machine learning. Among deep learning algorithms, artificial neural networks are the most common (Schmidhuber, 2015). Deep learning is data-hungry, i.e., it generally requires a lot of training data to address overfitting issues since it usually has considerably more learnable parameters than simpler ML algorithms, such as linear models. In addition, deep learning is usually deployed for learning special types of data, e.g., image data. For learning image data, a special type of deep learning algorithm – convolutional neural networks, are often used (LeCun et al., 2015). Since convolutional neural networks use a special type of feature – convolutional units, which capture local features of an image that could help distinguish an image from one of a different category (e.g., a dog's paw helps identify a dog), they are better at classifying images than other algorithms.

1.2 Micropeptides – an emerging edge of the protein universe

1.2.1 What are micropeptides?

Micropeptides are a special type of proteins. Proteins, or strictly speaking, single-chain proteins or polypeptides, are composed of a single chain of amino acid residues. A protein is translated from a messenger RNA (mRNA) transcript, which is a linear chain of nucleotides that generally contains three tandem regions: a 5'-untranslated region (5'-UTR), the coding sequence (CDS), and a 3'-untranslated region (3'-UTR). The CDS is responsible for coding the peptide, and as their names suggest, the 5'-UTR and 3'-UTR are generally not translated and are involved in the translation process directed by ribosomes.

A CDS is also an open reading frame (ORF) and contains the genetic information for directing the translation of the peptide. It is composed of codons. A codon is a 3-nucleotide code encoding an amino acid. As there are 4 different DNA nucleotides in total: A, T, C, G, and a codon contains 3 nucleotides, there are $4^3 = 64$ different codons. Among the 64, ATG is the canonical start codon and encodes methionine, and TAA/TAG/TGA are the stop codons and do not encode any amino acid. An ORF starts with a start codon and ends with a stop codon.

A typical protein, what I call here a regular protein/peptide, generally has more than 100 amino acids, so the ORF that encodes it generally has over 101 codons, i.e., it is > 303 base pairs (bp). On the other hand, a micropeptide, as the name suggests, is much shorter than a typical protein. By definition, it is a small protein that is encoded by a CDS region that is <= 303 bp, i.e., a micropeptide is <= 100 amino acids (aa) (Chugunova et al., 2018; Couso & Patraquim, 2017; Makarewich & Olson, 2017). Since the CDS region is short compared to that of a regular protein, it is also referred to a short/small open reading frame (sORF).

1.2.2 The history of micropeptides

Due to the short sequence, micropeptides were traditionally ignored.

1. Because of their small size, micropeptides are easily lost during sample preparation. In addition, micropeptides may be of low abundance. Together this makes them hard to identify in proteomics studies. (Olexiouk et al., 2016)
2. Because of their short length, the sORF that is responsible for encoding a micropeptide has a higher probability of occurring by chance compared to the ORF of a regular

> protein (Olexiouk et al., 2016). Consequently, to reduce noise, sORFs were excluded in many *in silico* methods for predicting the coding potential of ORFs (Kang et al., 2017; Kong et al., 2007; L. Wang et al., 2013). This annotation procedure, while successful in filtering out spurious noncoding ORFs, does leave out the sORFs that indeed encode micropeptides, and annotates them as noncoding. (Makarewich & Olson, 2017)

Despite the biases of proteomics studies and computational methods in unintentionally or intentionally filtering out sORFs, examples of micropeptides were still slowly identified by serendipity.

Among the earliest examples of a micropeptide, (Rohrig et al., 2002) identified two overlapping sORFs that encode a 12 amino acid (aa) and a 24 aa micropeptide, respectively, in the *ENOD40* gene of legumes, a gene that was previously considered to transcribe non-translatable noncoding RNAs (ncRNAs), as this gene contains only sORFs (Asad et al., 1994; Crespi et al., 1994). Both micropeptides were found to function in binding to sucrose synthase in the nitrogen-fixing nodules of legumes. From that point, growing attention has been paid to sORFs and their potential for coding micropeptides (Yeasmin et al., 2018).

Interest in micropeptides increased exponentially starting with the advent of ribosome profiling. Ribosome profiling (Ribo-Seq) is a transcriptomic technique that sequences the RNA fragments bound to ribosomes (also known as ribosome footprints/ ribosome-protected fragments), thereby revealing the transcripts that are potentially being translated by ribosomes (Ingolia et al., 2009). The employment of Ribo-Seq enabled large-scale discovery of many potential coding sORFs in various species, ranging from yeast (Smith et al., 2014a), to zebrafish (Bazzini et al., 2014a; Chew et al., 2013), to mouse (Ingolia et al., 2011a) and human (Bazzini et al., 2014a). These studies also revealed that transcriptomes are pervasively translated (Ingolia et al., 2014).

1.2.3 Functions of micropeptides

Ribo-Seq studies showed that many sORFs have clear signatures of translation, and a growing number of micropeptides were experimentally validated. Yet for the majority of translated sORFs, the functions of the translation products are still unknown (Yeasmin et al., 2018).

Nevertheless, the well characterized cases to date provide a peek into the functional universe of micropeptides.

In plants, micropeptides have been found to be involved in nodule organogenesis (Rohrig et al., 2002), leaf morphogenesis (Casson et al., 2002; Chilley et al., 2006; Frank & Smith, 2002; Narita et al., 2004), plant organogenesis (P. Guo et al., 2015; Wen et al., 2004), programmed cell death (Blanvillain et al., 2011), and pollen development (Dong et al., 2013; Jinxia Ma et al., 2008; D. Wang et al., 2009).

In animals, micropeptides have been found to play roles in embryogenesis (Galindo et al., 2007; T Kondo et al., 2010; Takefumi Kondo et al., 2007), cell migration (Pauli et al., 2014), stem cell differentiation (Kikuchi et al., 2009), calcium homeostasis (D M Anderson et al., 2015; Magny et al., 2013a), regulation of muscle performance (Bi et al., 2017; Nelson et al., 2016), DNA repair (Sarah A Slavoff et al., 2014), mRNA recycling (D'Lima et al., 2017), and programmed cell death (B. Guo et al., 2003; Hashimoto et al., 2001).

1.2.4 Current methods for large-scale identification of micropeptides

Ribosome profiling (Ribo-Seq)

As stated in "The history of micropeptides" section, Ribo-Seq has enabled discoveries of numerous sORFs that show translation signatures (Bazzini et al., 2014a; Chew et al., 2013; Ingolia et al., 2011a, 2014; Smith et al., 2014a). To date, Ribo-Seq remains one of the major tools for detecting translating sORFs in large scale.

While Ribo-Seq is powerful in identifying many putative translated sORFs, ribosome occupancy itself does not necessarily indicate a message is actually translated or establish the functional relevance of a translated peptide (Olexiouk et al., 2016). The sORF may just be associated with ribosomes (Chugunova et al., 2018). A number of technical improvements have been made to address the issues of Ribo-Seq. For instance, in Poly-Ribo-Seq, only polysomes, which represent active translation (Aspden et al., 2014; Galindo et al., 2007), are used to isolate the RNA. In addition, computational procedures have been developed to differentiate coding regions from noncoding fragments in Ribo-Seq data, including Ribo taper (Calviello et al., 2016), FLOSS (Ingolia et al., 2014), RSS (Guttman et al., 2013), ORF-RATER (Fields et al., 2015).

Despite those efforts, detection by Ribo-Seq still cannot be automatically interpreted as identifying true coding RNAs and still requires experimental verification.

Mass spectrometry

Mass spectrometry (MS) is a method for directly detecting peptides and determining their amino acid sequence. Protein samples are digested into (fragments) of peptides and are detected by MS. Identification of those peptides is then achieved by matching the resulting MS spectra against theoretic spectra of all possible candidate peptides retrieved from a reference protein database (Chugunova et al., 2018).

A number of novel micropeptides have been discovered using MS. (Banfai et al., 2012) identified 85 unique peptides, 65 of which were mapped to known lncRNAs, using tandem mass spectrometry (MS/MS). By incorporating MS into a proteomic data pipeline, (S A Slavoff et al., 2013) discovered 86 uncharacterized micropeptides in K562 cells. Later on, with a modified pipeline, (Jiao Ma et al., 2014) discovered an additional 195 micropeptides in K562 cells.

Despite of advances in MS-based micropeptide identification, there are drawbacks with this method. A prominent one is that MS favors abundant proteins, which means that micropeptides, which may be of low abundance in cells, are hard to detect using MS. Consequently, the number of micropeptides identified by MS to date is limited. (Yeasmin et al., 2018)

Bioinformatics methods

Bioinformatics methods can generally be divided into conservation-based methods and nucleotide-composition-based methods.

Conservation-based methods find conserved sORFs by multiple alignments of sequences from multiple species. Similar to regular proteins, conserved sORFs are more likely to be functional since conservation implies selection for retaining the sequence(Sousa & Farkas, 2018; Yeasmin et al., 2018). A well-known conservation-based method is PhyloCSF (Lin et al., 2011). Despite being useful in identifying conserved sORFs, conservation-based methods are not applicable to all cases (Sousa & Farkas, 2018; Yeasmin et al., 2018). Not all sORFs are conserved; some sORFs may evolve rapidly and therefore lack conservation. In addition, for shorter conserved sequences (such as sORFs) more species are needed for multiple genome alignments to achieve a reliable statistical power (Eddy, 2005). And for many species, there are not a sufficient number of genomes of close species for the required multiple alignments.

For nucleotide-composition-based methods, the underlining principle is that a coding sequence is different from a randomly occurred sequence with respect to nucleotide composition and arrangement. There are a number of nucleotide composition features that could differentiate coding sequences from noncoding ones. A simple feature could be single nucleotide proportions. For instance, human coding sequences are more GC-enriched than noncoding (Louie et al., 2003). Codon composition and preference can also be utilized.

A number of state-of-the-art methods employed nucleotide composition features for predicting the coding potential of an ORF, including CPC (Kong et al., 2007), CPC2 (Kang et al., 2017), CPAT (L. Wang et al., 2013), etc. However, they were all designed for predicting regular proteins, and sORFs were generally penalized and classified as noncoding during their training process. To our knowledge, there are few bioinformatics methods that specifically address the prediction of sORFs.

uPEPperoni (Skarshewski et al., 2014) is a conservation-based method for detecting conserved sORFs in the 5'-UTR of mRNAs (i.e., in the upstream region of regular proteins). Although a number of studies indicate that 5' upstream ORFs (uORFs) can be involved in regulating translation of the downstream peptide, many coding sORFs are located elsewhere.

sORF finder (Hanada et al., 2010) is a nucleotide-composition-based method and predicts the coding capability of a sORF using a pre-trained model of nucleotide frequency conditional probabilities. However, the method was developed a decade ago, when little training data was available, and the web server is no longer available.

1.2.5 The need for new bioinformatics methods of predicting coding sORFs

Although prediction of micropeptides by bioinformatics methods still entails experimental verification, as required for Ribo-Seq and MS as well, a well-established bioinformatics method can greatly facilitate research with minimal cost. In contrast to Ribo-Seq and MS, which still have substantial costs for sample preparation and sequencing/ mass spectrometry, utilization of bioinformatics methods costs almost nothing.

Prior to the work described in this dissertation, there were no good available bioinformatics method for predicting micropeptides from DNA sequences. Conservation-based methods cannot be applied to micropeptides that are not conserved or are differently conserved from typical proteins. State-of-the-art coding potential prediction methods generally consider sORFs as noncoding. The few methods designed for sORF prediction all have significant drawbacks. Therefore, there is the need for building a new bioinformatics method specifically for predicting the coding potentials of sORFs.

1.2.6 Available data for building bioinformatics methods

Today, there are sufficient data that may be utilized for developing new methods for micropeptide prediction. sORFs.org has collected examples of millions of sORFs that derived from translation signatures from various Ribo-Seq studies (Olexiouk et al., 2016, 2018). SmProt curated sORFs from Ribo-Seq, MS, literature mining, and other large-scale studies, and grouped the collected sORFs into high-confidence or non-high-confidence based on the available evidence for each sORF (Hao et al., 2018).

1.3 Long noncoding RNAs (LncRNAs) – a potential edge of the protein universe?

1.3.1 What are lncRNAs?

Long noncoding RNAs (lncRNAs), by definition, are noncoding RNA transcripts that have a sequence length of > 200 nucleotides (nt) and that are not translated into proteins (Kapranov et al., 2007; Mercer et al., 2009; Wilusz et al., 2009). Since it is usually unknown whether an RNA is translated, this definition is quite arbitrary and includes no criteria related to the structure or function of the RNA.

Contrary to the earlier belief that only a small portion of a genome is transcribed and that mostly into mRNAs, studies revealed that mammalian genomes are pervasively transcribed, and a large number of the transcription products are long RNAs that seem not possess a long ORF of coding potential (Carninci et al., 2005; Derrien et al., 2012; Johnson et al., 2005; Okazaki et al., 2002). These RNAs are termed lncRNAs, and the reason why the threshold of "200 nt" was chosen was simply for distinguishing them from traditional "small" noncoding RNAs (ncRNAs), including microRNAs (miRNAs), small interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs), piwiRNAs (piRNAs), etc., which are generally less than 200 nt (L. Ma et al., 2013).

1.3.2 Functions of lncRNAs

81,022, curated lncRNA transcripts have been identified in the current release of the human reference genome in Ensembl (release 98) (Cunningham et al., 2018), which exceeds the number of protein coding transcripts. Despite the enormous number, the functions of the majority of lncRNA transcripts remain to be answered. Nevertheless, there are a number of lncRNAs functions of which have been experimentally identified. According to LncBook (L. Ma et al., 2019), 1,867 lncRNAs have been documented with experimental validation backed up by 2,632 publications. Based on functional mechanisms, those lncRNAs are grouped into transcriptional regulation, ceRNA, splicing regulation, translational control, protein localization and RNAi.

1.3.3 Connections between lncRNAs and proteins

Similar to protein-coding transcripts, lncRNAs are also transcribed by RNA polymerase II. Many lncRNAs are 5’ capped, 3’ polyadenylated, have multiple exons, etc., just like protein-coding mRNAs (Cabili et al., 2011; Carninci et al., 2005; Chen, 2016; Derrien et al., 2012; Hartford & Lal, 2020; L. Ma et al., 2013). One major difference between lncRNAs and protein-coding transcripts is that lncRNAs typically have much lower sequence conservation (Makarewich & Olson, 2017), which contributes to the belief that they are “noncoding”.

With the advent of Ribo-Seq, many studies have shown that, contrary to earlier beliefs, many noncoding transcripts, including lncRNAs, exhibited clear translation signatures in a number of species, including human (Calviello et al., 2016; Ingolia et al., 2014; Ji et al., 2015; Raj et al., 2016; Ruiz-Orera et al., 2014), mouse (Ingolia et al., 2011a; Ruiz-Orera et al., 2018), zebrafish (Chew et al., 2013), fruit fly (Ruiz-Orera et al., 2014), plants (Bazin et al., 2017), and yeast (Ingolia et al., 2009). Consistent with results from Ribo-Seq studies, in recent years a large number of lncRNAs have been experimentally verified to encode proteins, especially micropeptides. The sORFs present in them that once were considered to have low coding capability are in fact translated. Some well-studied cases are listed in Table 1.1.

Table 1.1. Examples of micropeptides previously annotated as lncRNAs

name	**length (aa)**	**species**	**function**	**reference**
Bacteria				
Sda	46	*B. subtilis*	inhibition of sporulation	(Burkholder et al., 2001)
MciZ	40	*B. subtilis*	regulation of cell division	(Handler et al., 2008)
Plant				
ENOD40	12 or 24	legumes	binding to sucrose synthase	(Rohrig et al., 2002)
Zm401	89	maize	pollen development	(Jinxia Ma et al., 2008)
Animals				
polished rice (pri)	11 or 32	Drosophila	regulation of actin-based cell morphogenesis	(Takefumi Kondo et al., 2007)
Sarcolamban (scl)	28 or 29	Drosophila	regulation of cardiac calcium uptake	(Magny et al., 2013a)
Toddler	54	zebrafish	an embryonic signal promoting cell movement	(Pauli et al., 2014)
Myoregulin (MLN)	46	human and mouse	regulation of muscle performance	(D M Anderson et al., 2015)
SPAR	75	human and mouse	regulation of muscle regeneration	(Matsumoto et al., 2016)

aa: number of amino acid residues

Table 1.1. continued

DWORF	34	human and mouse	enhancing SERCA activity in muscle	(Nelson et al., 2016)
Myomixer	84	human and mouse	control of muscle formation	(Bi et al., 2017)
NoBody	68	human	interacting with the mRNA decapping complex	(D’Lima et al., 2017)
CASIMO1	83	human and mouse	breast-cancer-associated; control of cell proliferation	(Polycarpou-Schwarz et al., 2018)
MOXI / mitoregulin	56	human and mouse	Enhancement of mitochondrial β-oxidation	(Makarewich et al., 2018; Stein et al., 2018)

aa: number of amino acid residues

1.3.4 Are lncRNAs indeed translated?

As shown in the previous section, more and more lncRNAs have been re-annotated as protein-coding. This raises the fundamental question: do lncRNAs actually function via firstly being translated into micropeptides? i.e. Maybe they should not be called long *noncoding* RNAs, but instead they may represent a source of micropeptides that are yet to be identified. With a limited number of characterized lncRNAs to date, it is hard to answer this question. Yet it calls for further studies to explore the relationships between lncRNAs and proteins.

1.3.5 Genetic variation data and analysis

If most lncRNAs function by producing translated proteins, then their sequences may exhibit features resembling those in proteins. One of the features could come from genetic variation profile.

Genetic variation has long been used to study natural selection strength and thereby to infer functional relevance (Bhartiya et al., 2014; De Silva et al., 2014; Mu et al., 2011; Ward & Kellis, 2012). For example, the CDS regions of protein coding DNA sequences are generally considered to be under stronger negative selection than UTR regions, and results from several genetic variation studies indeed support this idea, suggesting the functional importance of the CDS regions (Jha et al., 2015; Khurana et al., 2013; Mu et al., 2011). By analogy, we can explore the relationships between lncRNAs and proteins from the perspective of genetic variation.

More importantly, today we have a large amount of genetic variation data. The 1000 Genomes project (Auton et al., 2015) is one of the most influential large-scale population genetics projects. It has produced the largest catalogue of human variation data, including over 84 million single nucleotide polymorphisms (SNPs) collected from 2,504 individuals across the world. The 1000 Genomes project provides a large variation dataset that enables studies with good statistical power.

1.4 Protein ligand binding with deep learning – an updated edge of the protein universe

Protein ligand binding is an age-old (and very important, of course) field of biological research. The reasons are simple - most proteins function via binding to other molecules (which are termed ligands), including other proteins, small molecules, etc. For instance, many proteins can bind small molecules, therefore, drugs, most of which are small molecules, are developed to bind certain disease-related proteins in human body to either promote or inhibit the functions of those proteins. Thus, the search for potential molecules that could bind target proteins is a fundamental aspect of drug discovery.

There are a large number of methods for identifying drug candidates of target proteins from a variety of perspectives, yet each of them has its own advantages and drawbacks (Xie & Hwang, 2015). Advances in deep learning, in particular its huge success in image classification (LeCun et al., 2015), shed new light on this long-standing field of protein ligand binding. The idea behind it is straightforward: proteins are in essence 3D structural molecules, which can be considered as 3D images. The features of their 3D structures are highly associated with the potential molecules they can bind. Therefore, determination of whether a protein can bind a certain ligand can be converted to a classification problem of the 3D image of this protein, employing ideas borrowed from the application of deep learning in 2D image classification.

More importantly, the increasing number of protein structures makes the "deep learning" idea feasible. To date, there are over 100,000 protein structures in the Protein Data Bank (PDB) (Burley et al., 2018). This large dataset enables the attempt to apply deep learning to study protein-ligand interaction.

1.5 Incorporating data science into the search of the edges of the protein universe

The protein universe consists of hundreds of thousand proteins carrying out a variety of functions. New proteins are discovered each day making this universe, while immense, still rapidly growing. What is growing as well is the number of data, which makes the study of this protein universe using data-driven approach more feasible.

As discussed above, the 3 components of a successful study are the goal, the data, and the method.

Micropeptides constitute an emerging edge of the protein universe. Despite the growth of newly identified micropeptides, we still need a bioinformatics method specifically designed for predicting micropeptides directly from DNA sequences, which could help identify potential micropeptides on a large scale with almost no cost. The increasing data in this field makes achieving this goal possible. With sufficient amount of labeled data, we can utilize machine learning algorithms to build an efficient method.

LncRNAs may constitute a potential edge of the protein universe as more are discovered to encode proteins. There are hundreds of thousands of lncRNA transcripts, but functions of the majority are unknown, nor are their coding capabilities. The coding phenomenon of lncRNAs, which contradicts their “noncoding” definition, raises the fundamental question about their functional mechanism – are lncRNAs indeed translated? We do not know the answer to this question, but we can explore it from a data-driven approach, and one feasible way is to infer the functional relevance of lncRNAs via genetic variation analysis. More importantly, we have sufficient data from the large-scale human variation project – the 1000 Genomes project. So, with the method of genetic variation analysis and the data from the 1000 Genomes project, it is now feasible to achieve the goal – explore the key question of whether lncRNAs are translated.

Protein-ligand interaction is a fundamental activity of proteins. Consequently, the computational prediction of protein-ligand binding is a long-standing field with many published methods, each with drawbacks and advantages. Advances in data science methods, in particular the huge success of deep learning in computer vision, sheds new light on this field of protein-ligand binding. Specifically, proteins are essentially 3D macromolecules and can be considered as 3D images. Consequently, the computational study of protein structures can be converted to the study of 3D images, which are analogous to that of 2D images in deep learning. Moreover, the amount of available protein structure makes the learning possible. So, with the new method in data science – deep learning and the large number or protein structure data, it is now possible to achieve the goal - the prediction of protein-ligand binding from a different approach.

CHAPTER 2. MIPEPID: MICROPEPTIDE IDENTIFICATION TOOL USING MACHINE LEARNING[1]

2.1 Abstract

Background: Micropeptides are small proteins with length <= 100 amino acids. Short open reading frames that could produce micropeptides were traditionally ignored due to technical difficulties, as few small peptides had been experimentally confirmed. In the past decade, a growing number of micropeptides have been shown to play significant roles in vital biological activities. Despite the increased amount of data, we still lack bioinformatics tools for specifically identifying micropeptides from DNA sequences. Indeed, most existing tools for classifying coding and noncoding ORFs were built on datasets in which "normal-sized" proteins were considered to be positives and short ORFs were generally considered to be noncoding. Since the functional and biophysical constraints on small peptides are likely to be different from those on "normal" proteins, methods for predicting short translated ORFs must be trained independently from those for longer proteins.

Results: In this study, we have developed MiPepid, a machine-learning tool specifically for the identification of micropeptides. We trained MiPepid using carefully cleaned data from existing databases and used logistic regression with 4-mer features. With only the sequence information of an ORF, MiPepid is able to predict whether it encodes a micropeptide with 96% accuracy on a blind dataset of high-confidence micropeptides, and to correctly classify newly discovered micropeptides not included in either the training or the blind test data. Compared with state-of-the-art coding potential prediction methods, MiPepid performs exceptionally well, as other methods incorrectly classify most *bona fide* micropeptides as noncoding. MiPepid is alignment-free and runs sufficiently fast for genome-scale analyses. It is easy to use and is available at https://github.com/MindAI/MiPepid.

Conclusions: MiPepid was developed to specifically predict micropeptides, a category of proteins with increasing significance, from DNA sequences. It shows evident advantages over

[1] This chapter appeared as MiPepid: MicroPeptide identification tool using machine learning. Zhu M and Gribskov M. BMC Bioinformatics, 20(1), 559. (2019). https://doi.org/10.1186/s12859-019-3033-9

existing coding potential prediction methods on micropeptide identification. It is ready to use and runs fast.

Keywords: micropeptide, small ORF, sORF, smORF, coding, noncoding, lncRNA, machine learning

2.2 Introduction

Micropeptides are generally defined as small proteins of <= 100 amino acid residues that are translated from small open reading frames (sORFs or smORFs, <= 303 base pairs (bp)) (Chugunova et al., 2018; Couso & Patraquim, 2017; Makarewich & Olson, 2017). Their existence was traditionally ignored because few micropeptides had been shown to be functionally important, mostly due to technological limitations in isolating small proteins (Olexiouk et al., 2018). Consequently, sORFs that encode micropeptides are generally ignored in gene annotation and have been considered to be noise (occurring by chance) and to be unlikely to be translated into proteins (Chugunova et al., 2018; Olexiouk et al., 2016, 2018).

With improved technology, an increasing number of micropeptides have been discovered, and have been shown to play important roles in muscle performance (D M Anderson et al., 2015), calcium signaling (Douglas M Anderson et al., 2016), heart contraction (Magny et al., 2013a), insulin regulation (Lee et al., 2015), immune surveillance (Schwab et al., 2003; R. F. Wang et al., 1996), etc. In particular, many micropeptides were shown to be translated from transcripts that were previously annotated as putative long noncoding RNAs (lncRNAs) (Cai et al., 2017; Yeasmin et al., 2018). This fact challenges the “noncoding” definition and raises questions about the functional mechanisms of lncRNAs, *i.e.*, whether they function through their 3D RNA structure, or via the micropeptides translated from encoded sORFs, or both.

With the increasing recognition of the importance of the “once well forgotten” field of micropeptides, it is increasingly important to develop a large-scale method for identifying them in a cost-effective way. Ribosome profiling (Ingolia, 2014; Ingolia et al., 2009) (Ribo-Seq) is a recent high-throughput technique for identifying potentially coding sORFs by sequencing mRNA fragments captured within translating ribosomes. Despite its advantages, there currently is no community consensus on how Ribo-Seq data should be used for gene annotation (Mudge & Harrow, 2016), as some investigators have questioned whether capture of RNAs by ribosomes necessarily implies translation; some capture could be transient or non-specific rather than truly functional (Ingolia, 2016; Raj et al., 2016). Ribo-Seq requires the use of next generation sequencing and thus has significant costs. In addition, depending on the sequencing depth and quality, it may suffer from false positives, and may not reveal all coding sORFs due to differences in sORF expression in different tissues, developmental stages, and conditions. Therefore, the sORFs discovered from Ribo-Seq still require experimental verification of their coding potentials.

It is much less expensive to predict coding sORFs from DNA sequences using bioinformatic tools. Although experimental verification is still required for predicted sORFs, a bioinformatic prediction of the coding potential of any sORFs before experimental verification is valuable since bioinformatics analysis costs almost nothing and could potentially provide useful insights.

There are currently few bioinformatic tools specifically designed for predicting the coding potential of small ORFs. uPEPperoni (Skarshewski et al., 2014) is a web server designed to detect sORFs in the 5' untranslated regions (5'-UTR) of mRNAs. It detects conserved sORFs without explicitly predicting their coding potential. Although 5'-UTR sORFs are an important component of the sORF population, many sORFs are located elsewhere, such as within the coding region of an mRNA, in lncRNAs, etc. The sORF finder (Hanada et al., 2010) program specifically identifies sORFs using the nucleotide frequency conditional probabilities of the sequence, however it was developed nearly a decade ago, and the server is no longer accessible. In addition, because many micropeptides have been discovered in the last decade, a much larger training dataset can now be assembled, and this should greatly improve the prediction quality. Data pipelines have been described (Bazzini et al., 2014b; Crappé et al., 2013; Mackowiak et al., 2015) that calculate the coding abilities of sORFs, especially those identified from Ribo-Seq data; however, these pipelines are not standalone packages readily available for other users. Other well-known coding potential prediction tools such as CPC (Kong et al., 2007), CPC2 (Kang et al., 2017), CPAT (L. Wang et al., 2013), CNCI (Sun et al., 2013), PhyloCSF (Lin et al., 2011), etc. were trained on datasets consisting primarily of normal-sized proteins. Because of the differences between sORF peptides and globular proteins, and because these methods were not trained on large sORF datasets, it is likely they do not perform well in sORF prediction (as shown in the Results section below). In general, most coding potential predictors penalize short ORFs and those that lack significant similarity to known proteins; both of these factors compromise the ability of existing tools to correctly predict sORFs.

With the ongoing development of techniques such as Ribo-Seq and mass spectrometry (MS), an increasing number of micropeptides have been experimentally identified and verified. We have a reasonable amount of data that can be leveraged for the development of bioinformatics tools specifically for micropeptide prediction. sORFs.org (Olexiouk et al., 2016, 2018) is a repository of small ORFs identified specifically from Ribo-Seq and MS data. And SmProt (Hao et al., 2018) is a database of micropeptides collected from literature mining, known databases, ribosome profiling, and MS.

Machine learning (ML) is a set of algorithms for learning hidden patterns within a set of data in order to classify, cluster, etc. The development of a successful ML-based method for a particular problem depends on a good dataset (clean, with sufficient data, etc.), and a good choice of specific ML algorithm. ML has been used in developing numerous bioinformatics tools, and has been used, for instance, in ORF coding potential prediction (Kang et al., 2017; Kong et al., 2007; Sun et al., 2013; L. Wang et al., 2013).

In this study, we present MiPepid, a ML-based tool specifically for identifying micropeptides directly from DNA sequences. It was trained using the well-studied logistic regression model on a high-quality dataset, which was carefully collected and cleaned by ourselves. MiPepid achieves impressive performance on several blind test datasets. Compared with several existing state-of-the-art coding potential prediction tools, MiPepid performs exceptionally well on *bona fide* micropeptide datasets, indicating its superiority in identifying small-sized proteins. It is also a lightweight and alignment-free method that runs sufficiently fast for genome-scale analyses and scales well.

2.3 Datasets

To collect positive as well as negative datasets for micropeptides that are representative yet concise, we selected 2 data sources: SmProt (Hao et al., 2018) and traditional noncoding RNAs.

2.3.1 The positive dataset

SmProt (Hao et al., 2018) is a database of small proteins / micropeptides which includes data from literature mining, known databases (UniProt ("UniProt: a hub for protein information.," 2015), NCBI CCDS (Farrell et al., 2014; Harte et al., 2012; Pruitt et al., 2009)), Ribo-Seq, and MS.

In particular, SmProt contains a high-confidence dataset consisting of micropeptide data that were collected from low-throughput literature mining, known databases, and high-throughput literature mining data or Ribo-Seq data with supporting MS evidence.

The SmProt high-confidence dataset (containing 12,602 human micropeptides in total) is a reliable data source for positive data since many of the peptides have been experimentally verified, and the rest are supported by multiple evidence. Based on this dataset, we cleaned our own positive dataset using the following pipeline:

1. Obtain the nucleotide sequences of the data. In SmProt, only the amino acid sequences rather than the DNA sequences are provided, although for the majority of data points their corresponding transcript IDs (primarily in Ensembl (Zerbino et al., 2018), with others in RefSeq (Zerbino et al., 2018) or NONCODE (Zerbino et al., 2018)) are provided. Since the DNA sequence of a micropeptide contains essential information that the translated sequence cannot provide (such as nucleotide frequency, etc.), we therefore obtained the corresponding DNA sequences by mapping the protein sequences back to their corresponding transcripts using GeneWise (Zerbino et al., 2018). To ensure the quality of the dataset, only micropeptides that gave a perfect match (no substitutions or indels) were retained.
2. Obtain a nonredundant positive dataset. Proteins with similar sequences may share similar functions, and families of related sequences create a bias towards certain sequence features. To ensure that our positive dataset is not biased by subgroups of micropeptides with similar sequences, we selected a nonredundant set with protein sequence identity ≤ 0.6. This serves as our **positive** dataset and it contains 4,017 data points.

2.3.2 The negative dataset

It is hard to define a truly negative dataset for micropeptides as more and more sequences that were formerly considered noncoding have been shown to encode translated proteins, such as 5'-UTRs of mRNAs, lncRNAs, etc. Despite the limitations of our current knowledge, we are still able to collect ORFs that are highly likely to be noncoding.

Traditional noncoding RNAs, such as microRNA (miRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), etc. are highly likely to be truly noncoding. While there is growing evidence that lncRNAs (Ji et al., 2015; Ruiz-Orera et al., 2014) may sometimes encode translated sORFs, the possibility of sORFs in traditional noncoding RNAs has seldom been mentioned or discussed in literature. In addition, some pipelines for predicting coding regions from Ribo-Seq data utilized those ncRNAs to construct their negative datasets (Guttman et al., 2013; Mackowiak et al., 2015). While there are examples of lncRNAs and "noncoding" regions of mRNAs that encode micropeptides in the SmProt high-confidence dataset, there are no examples of micropeptides encoded by traditional ncRNAs.

We therefore chose human miRNA, rRNA, snRNA, snoRNA (small nucleolar RNA), tRNA (transfer RNA), and scaRNA (small Cajal body RNA, a nucleolar RNA) as the data source for our negative dataset. We selected all human transcripts in the Ensembl database (Zerbino et al., 2018) annotated with these 6 biotypes and extracted all possible ORFs from those transcripts, *i.e.*, ORFs with valid start and stop codons from all 3 translation frames. Although there is evidence that non-ATG codons sometimes serve as sORF start codons (Olexiouk et al., 2016), to ensure the validity of our dataset, we consider only ATG start codons in constructing the negative dataset; in the positive dataset, nearly 99% of sORFs begin with ATG start codons.

We finally gathered 5616 negative sORFs. In the same way as for the positive data, we selected a nonredundant **negative** dataset of size 2,936 with pairwise predicted protein sequence identity ≤ 0.6.

2.3.3 The training set and the blind test set

We randomly selected 80% of the examples in the positive and negative datasets to build our training set for the machine learning model training; the remaining 20% were used as a blind test set which was only used for model evaluation (Table 2.1).

Table 2.1. Training and test data sets

dataset	**#positive**	**#negative**	**#total**
training	3,194	2,369	5,563
test	823	567	1390

#positive: number of positive data points
#negative: number of negative data points
#total: total number of data points

2.3.4 The synthetic_negative dataset

To further test the performance of our method, we generated a synthetic dataset that preserves the length distribution as well as the dinucleotide frequencies (H. Zhang et al., 2013) of the negative dataset. Since this dataset mimics the negative data, our method is expected to predict negative on this dataset. This **synthetic_negative** dataset is of the same size as the negative dataset (2,936), and it was generated using the ushuffle software (Jiang et al., 2008) in the MEME suite (Bailey et al., 2009).

2.4 Methods

2.4.1 Feature generation

In machine learning, identifying a set of relevant features is the next important step toward constructing a classifier. A set of well-chosen features greatly facilitates differentiating between different classes.

In our study, we believe the key to determining whether a small ORF is translated lies in the nucleotide patterns in the sequence. A translated sORF should have a DNA sequence that is constrained by the physicochemical properties of the translated peptide, the preference of ribosome occupancy, the codon bias of the organism, etc.

k-mer features have been widely used to effectively capture nucleotide patterns. A k-mer is a subsequence of length k, where k is an integer ranging from 1 to as high as hundreds depending on the requirements of specific questions. For DNA k-mers, there are only 4 types of nucleotides (A, T, C, and G), so the number of distinct k-mers for a specific k is 4^k. The k-mer features are simply encoded as a vector of size 4^k (denoted as $\boldsymbol{v}$), with each value in the vector denoting the frequency of one unique k-mer in the sequence. If we slide a window of length k across the sequence from beginning to end with a step size of s, we obtain $\left\lfloor \frac{|S|-k+1}{s} \right\rfloor$ k-mers in total, where $|S|$ denotes the length of the sequence. Therefore, $|\boldsymbol{v}|_1 = \left\lfloor \frac{|S|-k+1}{s} \right\rfloor$, where $|\boldsymbol{v}|_1$ is the L_1norm of $\boldsymbol{v}$. To exclude the sequence length effects in $\boldsymbol{v}$, we can use the normalized k-mer features, *i.e.*, the *fractional* frequency of each k-mer rather than the frequency itself. In this case, $|\boldsymbol{v}|_1 = 1$.

Regarding the choice of k, a hexamer (*i.e.,* 6-mer) is often used in bioinformatics tools for various biological questions (Chan & Kibler, 2005; Hanada et al., 2010). Yet hexamers would give a feature vector of size $4^6 = 4{,}096$. Compared to $5{,}563$, the size of our training data, a model with as many as $4{,}096$ parameters could potentially overfit the dataset although $5{,}563$ is larger than $4{,}096$. To ensure the generalizability as well as the efficiency of our method, we chose to use 4-mer features. A 4-mer, while short, still captures information about codons, and any dependencies between adjacent amino acid residues since every 4-mer covers parts of 2 adjacent codons / amino acids. A 4-mer feature vector has a reasonable size of 256, much less than 4,096, therefore should produce less model overfitting and have shorter running time. To eliminate the length information of a sORF, we chose to use normalized k-mer features. And to better capture the codon information of the translation frame, we chose a step size of 3 for k-mer extraction.

2.4.2 Logistic regression

From many possible supervised machine learning algorithms, we chose logistic regression for our study. Logistic regression is well-studied and provides easy-to-interpret models that have been shown to be successful in numerous cases and scenarios. The model can be tuned to minimize overfitting by, for instance, including regularization penalties. When used for prediction, the model returns the probability of an instance being in the positive category rather than just a label, which gives more insight into the prediction.

The loss function for logistic regression is:

$$\min_{\boldsymbol{w},b} \sum_{i=1}^{n} \log\left(1 + e^{-(y_i({X_i}^T\boldsymbol{w}+b)}\right) + \lambda\boldsymbol{w}^T\boldsymbol{w}$$

, where $\{X_1, \dots, X_n\}$ are the set of the data points and for each X_i, $y_i \in \{-1, +1\}$ is the label. $\boldsymbol{w}$ is the weight vector and b is the bias term. $\sum_{i=1}^{n} \log\left(1 + e^{-(y_i({X_i}^T\boldsymbol{w}+b)}\right)$ is the negative log likelihood. $\lambda\boldsymbol{w}^T\boldsymbol{w}$ is the regularization term which helps constrain the parameter space of $\boldsymbol{w}$ to reduce overfitting, and λ is a hyperparameter controlling the regularization strength. For a set of $\boldsymbol{w}$ and b, the classifier assigns the label to data point X_i based on the following:

$$f(X_i) = \frac{1}{1 + e^{-(\mathbf{w}^T X_i + b)}} \begin{cases} \geq t, & \hat{y}_i = +1 \\ < t, & \hat{y}_i = -1 \end{cases}$$

, where $\hat{y}_i$ is the predicted label from the classifier and t is the threshold between the positive $(+1)$ and the negative (-1) classes. Although $t = 0.5$ is generally used, $0 \leq t \leq 1$ is also a tunable hyperparameter.

2.4.3 Performance evaluation

To evaluate the performance of MiPepid and existing methods, we used the following metrics.

accuracy

For a dataset S, denote the number of correctly classified cases by a method as c, then the accuracy is $\frac{c}{|S|}$, where $|S|$ is the size of the dataset. This definition applies to any dataset used in this paper.

$\boldsymbol{F_1}$ *score*

For a dataset that contains both positive and negative data, the F_1 score of the performance of a method on this dataset is:

$$F_1 = 2\frac{pr \times rc}{pr + rc}$$

, where pr is the precision and rc is the recall, and

$$pr = \frac{TP}{TP + FP}, \qquad rc = \frac{TP}{TP + FN}$$

, where TP is the number of true positives, *i.e.*, the number of correctly classified cases in the positive subset; FP is the number of false positives, *i.e.*, the number of misclassified cases in the negative subset; FN is the number of false negatives, *i.e.*, the number of misclassified cases in the positive subset. The F_1 score ranges from 0 to 1, with a higher value implying better performance. In this study, the F_1 score is used for the training and the test sets, as both of them consist of both positive and negative data.

2.4.4 10-fold cross validation

N-fold cross validation is commonly used to select good hyperparameters. Here n is an integer ranging from 2 to as high as dozens. In cross validation, the dataset is randomly and evenly divided into n folds. For every set of hyperparameter candidates, and for each fold, a model is trained using the other $n - 1$ fold(s) and is evaluated on the left-out fold. The (weighted) average of the n evaluations is taken as the overall evaluation for that set of hyperparameter candidates.

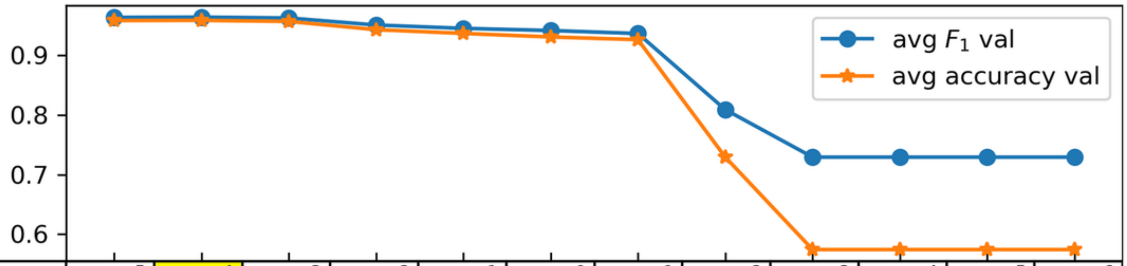
This cross validation is done for every set of hyperparameter candidates in order to select a set that gives the best performance.

As stated in 3.3, there are 2 hyperparameters in logistic regression: the regularization strength λ and the threshold t. We performed 10-fold cross validation to tune these 2 hyperparameters. For $\lambda \in \{1\text{E-}5, 1\text{E-}4, 1\text{E-}3, \ldots, 1\text{E+}5, 1\text{E+}6\}$ and $t \in \{0, 0.05, 0.1, \ldots, 0.95, 1.0\}$, we selected the combination of λ and t that gave the best performance.

2.5 Results

2.5.1 Hyperparameters tuning using 10-fold cross validation

As stated above, in the logistic regression model, the regularization strength λ and the threshold t are tunable hyperparameters. Therefore, before training the model on the training dataset, we first determined the best combination of λ and t using 10-fold cross validation. As shown in Figure 2.1, when $\lambda = 10^{-4}$ and $t = 0.60$, both the average F_1 (0.9639) and accuracy (0.9585) on the 10 validation sets are the highest.



λ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}
best t	.60	.60	.60	.60	.65	.60	.60	.55	.50	.50	.50	.50
avg F_1 val	.9635	.9639	.9625	.9507	.9450	.9414	.9364	.8088	.7292	.7292	.7292	.7292
avg accu val	.9581	.9585	.9567	.9428	.9365	.9306	.9261	.7293	.5741	.5741	.5741	.5741

Figure 2.1. 10-fold cross validation results with different λ and t combinations on the training set. λ: the hyperparameter for regularization strength in logistic regression; t: the hyperparameter for threshold in logistic regression; best t: when λ is fixed, the t from $t \in \{0, 0.05, 0.1, \ldots, 0.95, 1.0\}$ that gives the best performance; avg F_1 val: the average F_1 score on the 10 validation sets when both λ and t are fixed; avg accu val: the average accuracy on the 10 validation sets when both λ and t are fixed.

2.5.2 Training using the tuned hyperparameters

We therefore chose $\lambda = 10^{-4}$ and $t = 0.60$ and re-trained on the complete training dataset to obtain the MiPepid model. This model achieved an F_1 score of 0.9845 and an overall accuracy of 0.9822 on the training set (Table 2.2).

Table 2.2. MiPepid results on the training set

$\boldsymbol{F_1}$	**accuracy**		
	positive	**negative**	**overall**
0.9845	0.9818	0.9827	0.9822

“positive” and “negative” refer to the accuracies of MiPepid on the positive and negative subsets, respectively; “overall” refers to the accuracy on the whole training set (positive + negative).

2.5.3 MiPepid generalizes well on the hold-out blind test set

The blind test set contains 1,390 sequences and was not used during the training stage. As shown in Table 2.3, MiPepid achieved an F_1 score of 0.9640 and an overall accuracy of 0.9576 on this test set. Compared with Table 2.2, although the results are slightly lower, they are still comparably good. In addition, MiPepid performed almost equally well on the positive and negative subsets of the test set as indicated by the corresponding accuracies (0.9587 *vs.* 09559). Therefore, MiPepid generalizes well and has a balanced performance on both positive and negative data.

Table 2.3. MiPepid results on the blind test set

$\boldsymbol{F_1}$	**accuracy**		
	positive	**negative**	**overall**
0.9640	0.9587	0.9559	0.9576

"positive" and "negative" refer to the accuracies of MiPepid on the positive and negative subsets, respectively; "overall" refers to the accuracy on the whole test set (positive + negative).

2.5.4 MiPepid performs well on the synthetic_negative dataset

The synthetic_negative dataset mimics the negative dataset by preserving the dinucleotide frequency as well as the length distribution of the real negative data, but because it has been randomized, should have no true sORFs. MiPepid achieved an accuracy of 0.9659 on the synthetic_negative dataset, a very close result to the one on the negative subset of either the training or test set, indicating the robustness of MiPepid.

2.5.5 MiPepid correctly classifies newly published micropeptides

In the positive dataset, part of the data were collected by low-throughput literature mining in SmProt (Hao et al., 2018), *i.e.*, they were biologically/ experimentally verified on the level of protein, cell, phenotype, etc. SmProt (Hao et al., 2018), which was released in 2016, is based on literature published by December 2015. We searched for new examples of verified micropeptides, supported by extensive experimental evidence, published after Dec 2015, and found 5 new micropeptides in the literature (Table 2.4). Among these 5 cases, 3 are actually already recorded in SmProt (Zerbino et al., 2018), however they were in the non-high-confidence subset, *i.e.*, there was only indirect evidence on the presence of those micropeptides.

Table 2.4. List of micropeptides published after Dec 2015

micropeptide name	protein sequence length	in SmProt non-highConf	reference
MOXI	56	yes	(Makarewich et al., 2018)
DWORF	35	yes	(Nelson et al., 2016)
Myomixer / Minion	84	yes	(Bi et al., 2017)
SPAR	90	no	(Matsumoto et al., 2016)
HOXB-AS3	53	no	(Huang et al., 2017)

in SmProt non-highConf: If this micropeptide was already included in the SmProt (Zerbino et al., 2018) non-high-confidence subset, then the value is "yes", otherwise "no".

These 5 cases were taken as the **new_positive** dataset. They are analogous to “the future cases” if the time boundary were Dec 2015. One of the major purposes of MiPepid is for future prediction. Therefore, its performance on “future cases” matters.

We applied MiPepid on this new_positive dataset, and MiPepid correctly classified all of the 5 micropeptides. And this is another result showing the good generalization of MiPepid.

2.5.6 Comparison with existing methods

Comparison with current ORF coding potential prediction methods

There are several state-of-the-art bioinformatics methods built to predict the coding/noncoding capability of a DNA sequence, including CPC (Kong et al., 2007), CPC2 (Kang et al., 2017), CPAT (L. Wang et al., 2013), CNIT (Sun et al., 2013), PhyloCSF (Lin et al., 2011), etc. However, all of them were designed to work on “average” transcript datasets, *i.e.*, datasets that consist primarily of transcripts of regular-sized proteins and noncoding RNAs. In these methods, sORFs present in either an mRNA encoding a regular protein, or in a noncoding RNA, are generally penalized and are likely to be classified as noncoding; in the former case there is already a longer ORF present so shorter ones are treated as noncoding, and in the latter case the ORFs are automatically considered to be noncoding because they are found in “noncoding” RNAs. Therefore, despite the good performance of these methods in predicting regular-sized proteins, they may not be able to identify micropeptides, which also play critical biological roles.

In contrast, MiPepid is specifically designed to classify small ORFs in order to identify micropeptides. Here we chose CPC (Kong et al., 2007), CPC2 (Kang et al., 2017), and CPAT (L. Wang et al., 2013) as representatives of current methods and evaluated their performances on the hold-out blind test set as well as on the new_positive dataset, both of which the positive data are composed of high-confidence micropeptides.

As shown in Table 2.5, while the 3 methods (CPC (Kong et al., 2007), CPC2 (Kang et al., 2017), CPAT (L. Wang et al., 2013)) performed exceptionally well on negative cases (100% accuracy), they indeed struggled to classify the positive cases.

Table 2.5. Comparison with existing methods on the blind test set and the new_positive dataset

method	**blind test set**						**new_positive**	
	positive		**negative**		**overall**			
	#correct	**accuracy**	**#correct**	**accuracy**	F_1	**accuracy**	**#correct**	**accuracy**
CPC (Kong et al., 2007)	17	0.02	567	1.00	0.04	0.42	0	0.00
CPC2 (Kang et al., 2017)	61	0.07	567	1.00	0.14	0.45	0	0.00
CPAT (L. Wang et al., 2013)	261	0.32	567	1.00	0.48	0.60	3	0.60
MiPepid (our method)	789	0.96	542	0.96	0.96	0.96	5	1.00

positive: the positive subset of the blind test set;

negative: the negative subset of the blind test set;

overall: the overall performance on the blind test set;

#correct: the number of correctly classified cases by a method;

accuracy: #correct divided by the total number of cases in that dataset/subset;

F_1: the F_1 score

The positive cases in the blind test set are sORFs of high-confidence micropeptides supported by at least 2 different types of experimental evidence. CPC (Kong et al., 2007) and CPC2 (Kang et al., 2017) considered over 90% of them as noncoding, while CPAT (L. Wang et al., 2013) did better with 32% accuracy but is still below half. In contrast, while MiPepid performed slightly worse on the negative cases (96%), it correctly classified 96% of the high confidence micropeptides. And regarding sORFs of the newly-published micropeptides, all of which are supported by protein-level and phenotypic evidence, CPC (Kong et al., 2007) and CPC2 (Kang et al., 2017) did not consider any of them to be coding, and CPAT (L. Wang et al., 2013) correctly classified only 3 out of 5. These results are not surprising as all three existing methods were trained on datasets primarily consisting of regular-sized proteins. It is clear from those results that sORFs are a special subpopulation of ORFs and predictions on which entail specially designed methods.

Comparison with sORF finder

As mentioned in the Introduction section, sORF finder predicts sORFs by calculating nucleotide frequency conditional probabilities of hexamers; however, the server is no longer accessible. We located a downloadable version at http://hanadb01.bio.kyutech.ac.jp/sORFfinder/ and ran it locally. sORF finder does not provide a trained model for human sORFs, nor is there any human dataset included in this software. To conduct the comparison, we therefore used sORF finder to train a model using our own training dataset and then evaluated on our test set. It took hours to train the model using sORF finder, as compared to seconds needed for MiPepid.

Table 2.6. Comparison with sORF finder

method	blind test set					
	positive		negative		overall	
	#correct	accuracy	#correct	accuracy	F_1	accuracy
sORFfinder	708	0.86	506	0.89	0.89	0.87
MiPepid (our method)	789	0.96	542	0.96	0.96	0.96

positive: the positive subset of the blind test set;
negative: the negative subset of the blind test set;
overall: the overall performance on the blind test set;
#correct: the number of correctly classified cases by a method;
accuracy: #correct divided by the total number of cases in that dataset/subset;
F_1: the F_1 score

As shown in Table 2.6, sORF finder correctly predicts around 87% of the examples in the test set, which is fairly good. However, it is clear that MiPepid performs significantly better. It is not surprising that sORFfinder achieved a similar performance to MiPepid. sORF finder utilizes hexamer information and a naïve Bayes approach to calculate the posterior coding probability of a sORF given its hexamer composition. MiPepid uses 4-mer information, but rather than naïve Bayes, uses logistic regression to learn patterns from the data automatically. Notably, MiPepid achieves better classification using a much smaller feature vector, and much less computational time for training the model.

2.6 Discussion

2.6.1 MiPepid's predictions on non-high-confidence micropeptides

The SmProt database (Hao et al., 2018) has a high-confidence subset, which contains examples of micropeptides that are supported by multiple kinds of evidence; the rest of the data belong to the non-high-confidence subset. We collected those data and obtained their corresponding DNA sequences using the same pipeline used for the positive dataset (see Methods). We then used MiPepid to predict the coding capabilities of those data. Overall, MiPepid predicted 74% of them as positive. Table 2.7 shows detailed results based on different data sources.

Table 2.7. MiPepid's prediction on the non-high-confidence data in SmProt

data source	#sORFs	avg sORF length (aa)	#predicted positive	proportion
high-throughput literature mining	25663	44	20516	0.80
ribosome profiling	13715	36	8596	0.63
MS data	324	15	233	0.72

high-throughput literature mining: published sORFs that were identified using high-throughput experimental methods;
ribosome profiling: sORFs predicted from Ribo-Seq data;
MS data: sORFs predicted from MS data;
#sORFs: number of sORFs from a particular data source;
avg sORF length (aa): the average length of sORFs measured in number of amino acids;
#predicted positive: number of sORFs that are predicted as positive by MiPepid;
proportion: $\frac{\text{avg sORF length}}{\text{\#predicted positive}}$

As can be seen in Table 2.7, among the over 25 thousand sORFs collected by high-throughput literature mining, MiPepid predicted 80% of them as positive, which is a fairly high proportion. There are only 324 sORFs derived from MS data, and MiPepid labeled 72% of them as positive. Note that, on average, MS sORFs are significantly shorter than those from other sources. In contrast, among the over 13 thousand Ribo-Seq derived sORFs, MiPepid only predicted 63% of them as positive. This is not very surprising as there has been debate on the reliability of predicting peptides from Ribo-Seq data; some investigators have argued that the capture of an RNA transcript by the ribosome does not always lead to translation (Mudge & Harrow, 2016), and that some of the ribosome associated RNAs found in Ribo-Seq may be regulatory or non-specifically associated.

We are interested in looking at the relationship between the length of a sORF and its coding probability predicted by MiPepid.

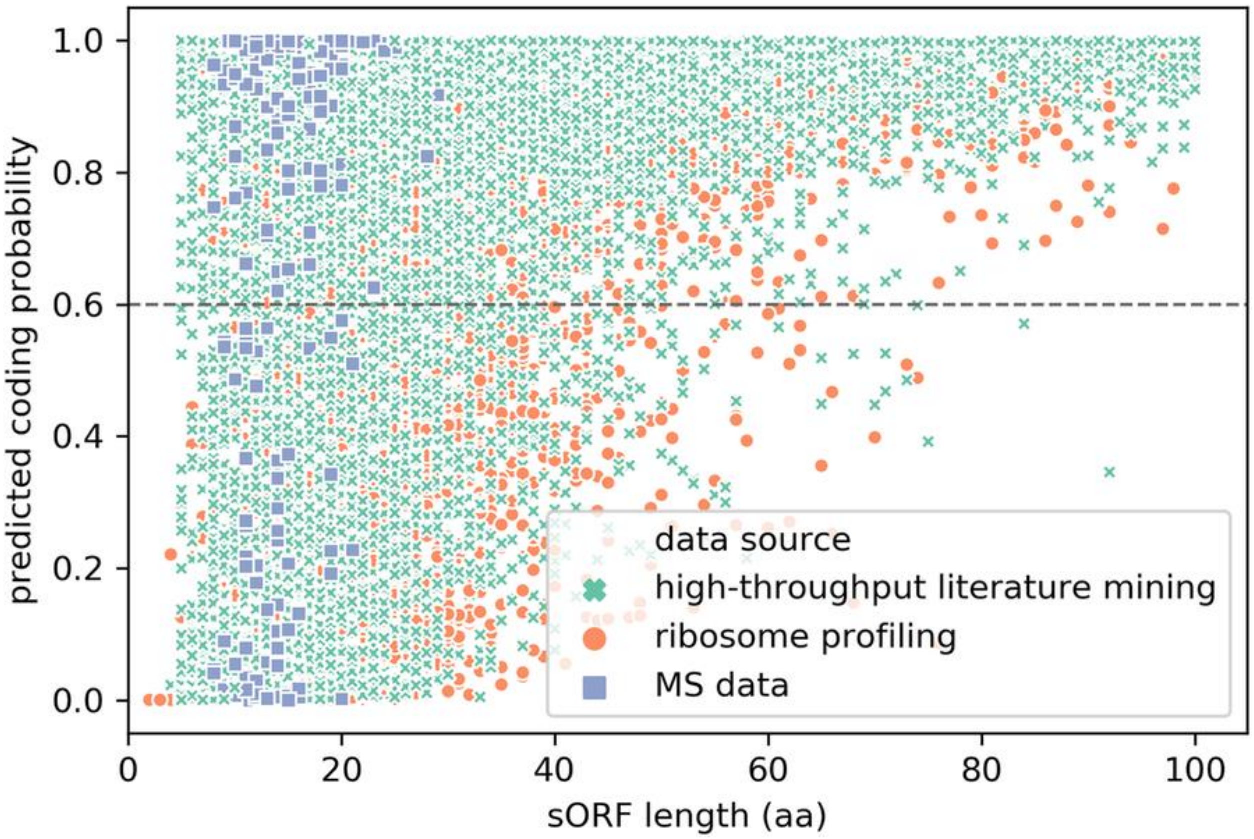


Figure 2.2. Scatterplot of the length of sORF *vs.* predicted coding probability for the non-high-confidence sORFs in SmProt. aa: number of amino acids. The $y = 0.6$ horizontal line separates sORFs that are predicted as positive (predicted coding probability ≥ 0.6) and the rest predicted negative.

Figure 2.2 shows a moderately positive trend between the length of a sORF and its coding probability predicted by MiPepid. This is reasonable considering the following: (1) the longer a sORF, the less likely it occurs by chance; (2) the longer a sORF, the more 4-mer information it contains, which helps MiPepid to better classify it. Yet, we do see that for many very short sORFs (< 20 aa), MiPepid was able to identify the positives, and for long sORFs (> 50 aa), MiPepid was not misled by the length, and was still able to identify some as negatives. In figure 2, one can also see that sORFs derived from the MS data are very short (< 30 aa).

2.6.2 MiPepid's prediction on uORFs of protein-coding transcripts

A uORF (upstream open reading frame) is an ORF (usually short) located in the 5'-UTR (untranslated region) of a protein-coding transcript. A number of uORFs have been discovered to encode micropeptides and to play important roles in biological activities (Plaza et al., 2017), and Ribo-Seq evidence suggests that many uORFs are translated (Skarshewski et al., 2014). uORFs have drawn increasing attention, and there is a great interest in determining the coding potentials of uORFs.

We extracted all possible small uORFs (from all 3 translation frames) of all annotated protein-coding transcripts in the Ensembl (Zerbino et al., 2018) human database. We then used MiPepid to determine the coding potentials of the extracted uORFs.

From 12,221 protein-coding transcripts, we extracted 42,589 small uORFs in total. 34.24% of the uORFs were predicted by MiPepid as coding. Among the 12,221 transcripts, 55.80% of them (6820) contain at least one potential micropeptide-encoding uORF. For the readers' interest, we compiled all the small uORFs together with their coding potential score, location in the corresponding transcript, etc. into a supplemental file. This file is available along with the MiPepid package.

2.6.3 MiPepid's prediction on lncRNAs

Long noncoding RNAs (lncRNAs) are RNA transcripts that lack a long ORF, and therefore were initially considered to be untranslated. Yet a growing number of lncRNAs have been discovered to be actually translated into functional micropeptides (Cohen, 2014; Ji et al., 2015; Matsumoto et al., 2016; Nelson et al., 2016).

We extracted all possible sORFs (from all 3 translation frames) of all human lncRNA transcripts in Ensembl (Zerbino et al., 2018) (those with the following biotypes: non_coding, 3prime_overlapping_ncRNA, antisense, lincRNA, retained_intron, sense_intronic, sense_overlapping, macro_lncRNA, or bidirectional_promoter_lncRNA). From the 26,711 lncRNA transcripts, we extracted 371,123 sORFs, averaging ~ 14 sORFs per transcript. 31.28% of the sORFs were predicted as coding. 86.63% of lncRNA transcripts were predicted to have at least one sORF that could potentially be translated into a micropeptide.

We present MiPepid's prediction results on lncRNAs not for evaluating its performance but to show that the proportion of sORFs in lncRNAs that are "similar" to sORFs of high-confidence micropeptides in our training set is very high. It is impossible to evaluate MiPepid using the lncRNA results as we have very little data on which sORFs in lncRNAs are truly positive, and which are not. The results serve as a reference for researchers interested in further work on any of those lncRNAs. The supplemental file containing MiPepid results on the 26,711 annotated lncRNAs is also available in the MiPepid software package.

2.6.4 MiPepid's prediction on small protein-coding genes in other model organisms

MiPepid was trained on human data, and we expect that it would work well on related mammalian species, such as mouse, rat, etc. Yet, we want to know how well it generalizes to other species, *e.g.*, plants, bacteria, etc. We therefore collected all annotated small protein-coding sequences (<= 303 bp) in *E. coli*, yeast, Arabidopsis, zebrafish, and mouse from the Ensembl database (Zerbino et al., 2018), and examined whether they are predicted to be coding sequences by MiPepid. MiPepid successfully predicts at least 93% of the sequences as coding for these 5 species (Table 2.8). This indicates that MiPepid has been able to successfully learn generalized sequence patterns typical of human sORFs, and in addition, suggests that small protein-coding gene sequences share hidden patterns across biological kingdoms.

Table 2.8. MiPepid's prediction on small protein-coding genes in model organisms

species	#seq	%predicted positive
E. coli	422	96.68%
yeast (*S. cerevisiae*)	502	93.63%
arabidopsis (*A. thaliana*)	2888	98.61%
zebrafish (*D. rerio*)	2481	96.78%
mouse (*M. musculus*)	6451	97.54%

#seq: number of small protein-coding sequences
%predicted positive: percentage of sequences predicted as coding by MiPepid

2.7 Conclusions

MiPepid is designed to take a DNA sequence of a sORF and predict its micropeptide-coding capability. We suggest using sequences with transcriptome-level evidence, *i.e.*, DNA sequences that are indeed transcribed, as MiPepid was trained to determine whether a transcript can be translated, and the training data did not include sORFs from untranslated DNA regions. The potential for an untranslated DNA sequence, such as an intergenic region, to be transcribed and translated was not addressed. MiPepid was specifically developed to predict small ORFs and “regular-sized” ORFs were not included in the training. Therefore, we recommend using MiPepid only on sORFs; MiPepid is not trained to efficiently predict long ORFs such as those found in typical mRNAs. MiPepid was trained on human data, but should work for related mammalian species, such as mouse, rat, etc. Retraining the model on other species requires only a set of known micropeptides and the corresponding genomic sequence.

CHAPTER 3. GENETIC VARIATION EVIDENCE SUGGESTS LONG NONCODING RNAS RESEMBLE MICROPEPTIDES

3.1 Abstract

Micropeptides are small proteins translated from short coding sequences (CDS) with a length <= 303 base pairs (bp). They are drawing increasing attention as more are discovered and are shown to play various vital roles. Yet, as an emerging group, they are still poorly characterized compared to regular proteins (CDS > 303 bp).

Long noncoding RNAs (lncRNAs) are noncoding transcripts of > 200 bp. Studies of lncRNAs are growing dramatically, with tens of thousands of predicted lncRNA transcripts recorded in public databases, yet the functional roles of the majority still remain explored. Moreover, an increasing number of lncRNAs are redefined as encoding proteins, especially micropeptides. Whether lncRNAs function via translated proteins is an open question of great significance.

Genetic variation analysis has long been used to study the natural selection of genomic elements and to infer functional relevance. For instance, previous variation studies have shown protein CDS regions are under stronger purifying selection than a number of other genomic elements, indicating their functional importance.

In this study, to better characterize the relationships between the two growing families – micropeptides and lncRNAs, as well as their relationships with regular proteins, we explored the three categories from the perspective of genetic variation.

We find the three categories share similar single nucleotide polymorphism (SNP) densities, SNP spectra, and enrichments of rare SNPs. The SNP density of lncRNAs is statistically equal to that of micropeptides, suggesting lncRNAs and micropeptides are under the same level of purifying selection strength. Rare SNPs are less enriched in lncRNAs than in micropeptides, suggesting lncRNAs are under weaker purifying selection than micropeptides, yet the difference is very small. The CDS regions of micropeptides and those of regular proteins are under the same purifying selection strength despite of the stark discrepancies in CDS lengths. In addition, CDS show stronger negative selection strength than untranslated regions (UTR) in both regular proteins and micropeptides. We used our published method – MiPepid - to predict potential coding regions in lncRNAs and found the predicted CDS regions are also under stronger purifying selection than UTR.

Overall, our findings show lncRNAs share similar variation profiles with both regular proteins and micropeptides, and therefore may share sizeable functional overlaps with proteins. That is, that many proposed lncRNAs may actually encode translated or translatable peptides. To our knowledge, this study is the first attempt to explore the relationships between regular proteins, micropeptides, and lncRNAs from the angle of genetic variation.

Keywords: micropeptide, lncRNA, small protein, genetic variation, SNP, natural selection

3.2 Introduction

Micropeptides are small proteins that are translated from short open reading frames (sORFs, <= 303 bp, i.e. <= 100 amino acids) (Chugunova et al., 2018; Couso & Patraquim, 2017; Makarewich & Olson, 2017). Traditionally, they were largely ignored because they were difficult to isolate biochemically (Olexiouk et al., 2018). In recent years, a growing number of micropeptides have been identified and shown to participate in a variety of biological roles, including immune surveillance (Schwab et al., 2003; R. F. Wang et al., 1996), calcium signaling (Douglas M Anderson et al., 2016), heart contraction (Magny et al., 2013b), muscle performance (D M Anderson et al., 2015), etc. In addition to the well characterized cases, ribosome profiling (Ribo-Seq) has identified numerous sORFs with clear translation signatures (Chugunova et al., 2018; Hao et al., 2018; Olexiouk et al., 2016, 2018). The function of these translated products awaits experimental verification, but it suggests the micropeptides discovered to date may be just the tip of an iceberg.

Long noncoding RNAs (lncRNAs) have been defined as transcripts that are >= 200 nucleotides (nt) and are not translated into a protein (L. Ma et al., 2013). With the increasing availability of next generation sequencing technology, more and more lncRNAs are being discovered, and it is widely accepted that lncRNAs constitute a significant part of the RNA family. The number of lncRNA transcripts curated in Ensembl (Cunningham et al., 2018) already exceeds the number of protein coding transcripts. LncRNAs have been shown to participate in many processes, including transcriptional regulation, splicing regulation, translational control, protein localization, etc. (L. Ma et al., 2019) Despite the rapidly advancing research in this field, the functions of the majority of lncRNAs still remain to be elucidated.

Contrary to the assumption that lncRNAs exert functions via folded three dimensional structures, several large-scale studies have suggested that a considerable portion of lncRNAs may be translated into proteins (Ingolia et al., 2011b; Ji et al., 2015; Smith et al., 2014b). In fact, a growing number of lncRNAs have been confirmed to actually encode proteins, in particular micropeptides (Yeasmin et al., 2018). Based on our simple analysis, at least 78 human intergenic lncRNA transcripts in Ensembl have been reannotated as protein coding transcripts between 2014 (release 75) (Flicek et al., 2013) and 2019 (release 98) (Cunningham et al., 2018). This growing list contradicts the "noncoding" definition of lncRNAs and also confounds our understanding of their elusive roles. Is it an exception or the rule for a lncRNA to function via its protein product?

Genetic variation, by definition, is the difference in DNA among individuals (Wikipedia contributors, 2020). There are many types of genetic variation, and the most heavily studied are single nucleotide polymorphisms (SNPs) (Collins et al., 1998), which are the substitution of one nucleotide by another at a position in the genome. On the microscopic level, the change of a single nucleotide, an individual SNP, causes a change in the DNA sequence, which may lead to a change in the amino acid sequence if the DNA is translated to a protein. In turn, the amino acid change may alter the function of the protein, and thereby may even cause a disease (Rees et al., 2010) or phenotypic difference. This is one of the mechanisms implicating SNPs in various human diseases and is also a force driving the ongoing studies in SNPs. On the macroscopic level, propagation of a SNP mutation in a population is subject to the force of natural selection depending on the consequence of this mutation. A deleterious SNP mutation is less likely to be inherited and is therefore usually observed at a low frequency, while an advantageous one is more likely to be propagated and can therefore be maintained at a higher frequency in a population. Thus, study of the SNP profile, which includes the SNP density (Bhartiya et al., 2014; Ward & Kellis, 2012), the SNP frequency distribution (De Silva et al., 2014; Mu et al., 2011), etc., of a category of DNA sequences can be used to analyze the strength of natural selection exerted on it, and thus to infer its functional relevance. Negative selection and positive selection are two of the major types of natural selection. Negative selection, or purifying selection, selectively removes adverse alleles; while positive selection, or directional selection, favors beneficial alleles and causes their frequencies to increase in the population over time. (Hamilton, 2011)

SNPs originate in random mutations, which are more likely to be deleterious than beneficial (Vitti et al., 2013). Consequently, SNPs are more likely to be under negative selection than positive selection, most of SNPs are thus observed at low frequencies in the population (Auton et al., 2015). Therefore, enrichment of rare SNPs (low frequency SNPs) is often used to estimate the strength of negative selection (Jha et al., 2015; Khurana et al., 2013). For instance, coding sequence (CDS) are generally considered to be under stronger purifying selection than untranslated regions (UTR) in a transcript, and a number of previous variation studies showed that rare SNPs are indeed enriched in CDS regions compared to UTR (Jha et al., 2015; Khurana et al., 2013; Mu et al., 2011), thus supporting this belief. In addition to enrichment of rare SNPs, SNP density, the number of SNPs within a certain length of DNA, also can be used to assess negative selection strength (Bhartiya et al., 2014; Ward & Kellis, 2012). As negative selection drives the elimination of deleterious mutations, a lower SNP density indicates stronger purifying selection.

SNP variation studies are only possible with data from large-scale human variation projects, and the 1000 Genomes project (Auton et al., 2015) is among the most influential ones. The 1000 Genomes project started in 2008 and has generated the largest public catalogue of human variation data. It includes 3 phases: the pilot phase, concluded in 2010, with genomes of 179 unrelated individuals sequenced (Durbin et al., 2010); phase 1, ending in 2012, in which the genomes of a total of 1,092 individuals were sequenced (McVean et al., 2012); and phase 3, completed in 2015, with the final released data containing variations in 2,504 human genomes (Auton et al., 2015). The final dataset includes whole-genome sequencing plus deep exome sequencing of a large sample size collected from 6 mega-populations across the world - the 1000 Genomes project therefore provides a relatively unbiased dataset with a much higher statistical power and is invaluable for variation studies.

The progress of the 1000 Genomes project propelled a surge of genetic variation studies. Using the pilot phase data, Mu et al. (Mu et al., 2011) found that ncRNAs are generally less selectively constrained than coding sequences by analyzing the SNP spectra of those regions. Also with the pilot phase data, Ward et al. (Ward & Kellis, 2012) analyzed SNP densities and average SNP frequencies and found that in human genome, a variety of biochemically active nonconserved elements (transcribed, regulatory, etc.) are under strong purifying selection, suggesting lineage-specific selection; while conserved regions lacking activity are under less selective constraint, suggesting they recently became nonfunctional. In addition, De Silva et al. (De Silva et al., 2014) found that deeply conserved human enhancers show stronger selective constraints compared to coding sequences as they exhibit a higher enrichment of rare SNPs. With the release of the 1000 Genomes phase 1 data, Khurana et al. (Khurana et al., 2013) identified “ultrasensitive” regions - noncoding regions with strong selective constraint that is comparable to coding regions in the human genome by accessing the enrichment of rare SNPs.

With the increasing focus on lncRNA research, some analyses have characterized the genetic variation profiles of lncRNAs. The first study was conducted by Bhartiya et al. (Bhartiya et al., 2014), who analyzed the pattern of genetic variations in lncRNAs and found that lncRNAs showed a SNP density that is at least 10 times that of protein-coding genes, suggesting lncRNAs are under much weaker purifying selection than proteins. However, the variation data they used was from an earlier project HapMap (Durbin et al., 2010; Gibbs et al., 2003) instead of the 1000 Genomes project. According to (Mu et al., 2011), the HapMap data can be biased since HapMap used known probes to identify SNPs, and SNPs that are adjacent to probes, or novel SNPs, may not be identified; this bias was confirmed by another study (De Silva et al., 2014). Jha et al. (Jha et al., 2015) analyzed the natural selection strength of various coding and noncoding elements in human genome using enrichment of rare SNPs, and found intergenic lncRNAs are under weaker purifying selection than proteins; yet the study was published before the release of 1000 Genomes phase 3 data, and the data they used was from phase 1, which includes only 1,092 individuals, less than half of the sample size of phase 3.

As the open question of whether lncRNAs function via being translated to proteins becomes increasingly controversial, we consider it a great interest to explore this question from the perspective of genetic variation. Our hypothesis is straightforward: if it is typical for lncRNAs to be translated into proteins, then lncRNA transcripts may share similarities with protein coding transcripts regarding genetic variation. In particular, since most lncRNAs do not have a long ORF, it is expected they may share more similarities with micropeptide transcripts. In addition, as a growing number of micropeptides are characterized, it is also of interest to compare them with regular proteins (ORF > 303 bp) regarding the genetic variation profiles since micropeptides have a much shorter CDS.

In this study, with the complete phase 3 data from the 1000 Genomes project (Auton et al., 2015), we analyzed the SNP variation in regular proteins, micropeptides, and lncRNAs regarding SNP density, enrichment of rare SNPs, etc. We find that lncRNAs show a SNP density that is not statistically different from that of micropeptides. Although rare SNPs are less enriched in lncRNAs, suggesting lncRNAs are under less purifying selection strength than proteins, the difference is small. We also found that the CDS regions of regular proteins and micropeptides are under the same purifying selection despite the difference in length. We used MiPepid (M. Zhu & Gribskov, 2019) - our published method specifically designed for predicting sORFs in RNA transcripts to predict potential CDS in lncRNAs. We found the predicted lncRNA CDS are under stronger purifying selection than the rest regions (*i.e.* the predicted UTR regions), which is consistent with the pattern observed in protein coding transcripts. Our results reveal the similarities between regular proteins and micropeptides, and also similarities between lncRNAs and micropeptides. Our findings are consistent with the increasing number of lncRNAs being reannotated as protein-coding.

3.3 Datasets and Methods

3.3.1 Datasets

Protein-coding and lncRNA transcripts

Protein-coding transcripts were downloaded from Ensembl (Cunningham et al., 2018). Only transcripts with a complete CDS, *i.e.*, without annotation as either “CDS 5’ incomplete” or “CDS 3’ incomplete”, were retrieved. The transcripts were further divided into regular proteins (with a CDS > 303 bp) and micropeptides (with a CDS <= 303 bp). LncRNA transcripts were also downloaded from Ensembl (Cunningham et al., 2018).

Single nucleotide polymorphisms (SNPs)

The locations of SNPs in protein-coding and lncRNA transcripts were retrieved from the Ensembl MySQL database (Cunningham et al., 2018). SNP information includes the minor allele sequence, the global frequency of the minor allele from the 1000 Genomes project (Auton et al., 2015), as well as the ancestral allele sequence.

A SNP is substitution of a nucleotide, the corresponding position therefore has as least two alleles, with the one having the second highest frequency in the population defined as the minor allele. The ancestral allele can be identified from multi-species alignments (see the 1000 Genomes project, Auton et al., 2015). Mutations from the ancestral sequence are referred to as derived alleles. A derived allele can either be the minor allele or the major allele of a SNP, the derived allele frequency is therefore either the minor allele frequency or (1 - the minor allele frequency).

A gene can have multiple overlapping transcripts, and multiple genes can be overlapped at a genome location; therefore, a SNP can be located in several transcripts or genes. To avoid counting a SNP multiple times, for a set of sequences, we only record the number of distinct SNPs.

Table 3.1 lists the number of transcripts and the number of distinct SNPs for each dataset used in this paper.

Table 3.1. Basic statistics of the datasets used in this study

dataset	**number of transcripts**	**number of distinct SNPs**	**number of distinct derived alleles**
regular protein	60,520	4,738,251	460,172
micropeptide	4,390	117,890	14,186
lncRNA	76,326	2,404,884	154,846

3.3.2 Methods

Definitions

SNP density: the SNP density of a DNA sequence dataset is the number of SNPs per 1,000 nucleotide bases (kb) in this dataset.

SNP spectrum: for a set of SNPs, the SNP spectrum is the distribution of the allele frequencies of the distinct SNPs in this set. In this study, the SNP spectrum is derived by binning SNPs into ranges of (0, 0.05), [0.05, 0.10), …, [0.95, 1) based on a SNP's frequency.

Rare allele: a rare allele has a frequency below a preset threshold. In this study, 3 thresholds are used: 0.05, 0.01, and 0.005. Depending on whether a rare allele is a derived allele or a minor allele, it is also called a rare derived allele or a rare minor allele.

Fraction of rare derived alleles (FDA): for a set of derived alleles, the FDA is the fraction of rare derived alleles in this set. Let N_d denote the total number of derived alleles and r_d be the number of rare derived alleles in this set, then $FDA = \frac{r_d}{N_d}$.

Fraction of rare minor alleles (FMA): for a set of minor alleles, the FMA is the fraction of rare minor alleles in this set. Let N_m denote the total number of derived alleles and r_m be the number of rare derived alleles in this set, then $FMA = \frac{r_m}{N_m}$.

Procedures

Prediction of potential CDS in lncRNA

MiPepid (M. Zhu & Gribskov, 2019) was used to predict potential coding regions in lncRNA transcripts: for each lncRNA transcript, we used MiPepid to predict the coding capability of each canonical small ORF (<= 303 bp, with ATG as the start codon and TAA/TAG/TGA as the stop codon) in all 3 translating frames, and then among all predicted small ORFs, we assigned the longest one as the final potential coding region for that lncRNA transcript.

Statistical tests

Hypothesis testing for two SNP densities was done using 2-proportion Z test. The SNP densities were treated as proportions, i.e. the proportion of nucleotides that have SNPs in a sequence set of nucleotides.

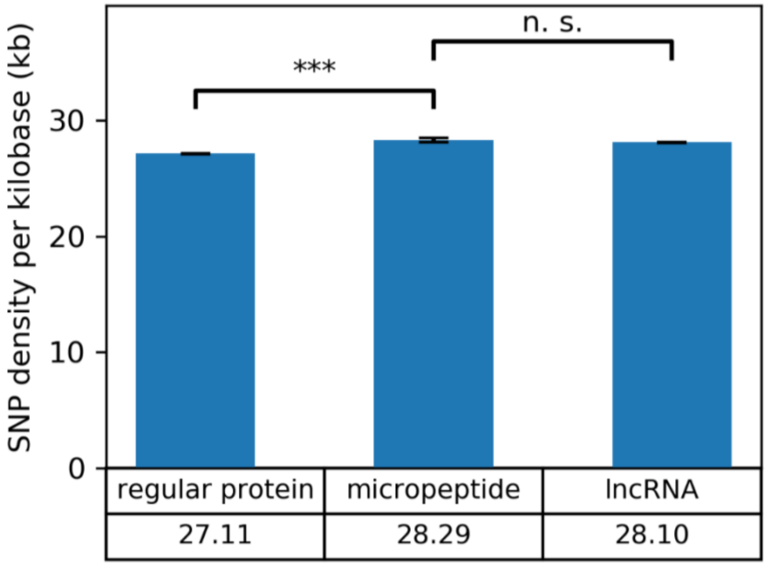
Hypothesis testing for two fractions of rare SNPs, either rare derived alleles or rare minor alleles, were also done using 2-proportion Z test.

Hypothesis testing for two ratios was done by obtaining the confidence interval of each ratio using bootstrapping. If the two confidence intervals do not overlap, then the difference is statistically significant.

3.4 Results

3.4.1 Regular proteins, micropeptides, and lncRNAs have similar SNP densities

We compared the SNP density (number of SNPs per kilobase) across proteins, micropeptides, and lncRNAs. As shown in Figure 3.1, regular proteins have 27.11 SNPs/kb, micropeptides have 28.29 SNPs/kb, while lncRNAs have 28.10 SNPs/kb. The SNP density of lncRNAs is not significantly different from that of micropeptides. While the density in regular proteins is significantly lower than that in either micropeptides or lncRNAs, the difference is small, around one fewer SNP per kilobase. This result suggests regular proteins are under a stronger negative selection strength, while micropeptides and lncRNAs are under the same purifying selection strength. Overall, regular proteins, micropeptides, and lncRNAs have similar variabilities regarding SNP density.



regular protein	micropeptide	lncRNA
27.11	28.29	28.10

Figure 3.1. SNP density of regular proteins, micropeptides, and lncRNAs. Error bars denote 95% confidence intervals of binomial proportions. “***” denotes a p-value < 0.001. “n. s.” denotes “not significant” and a corresponding p-value >= 0.05. Bonferroni correction was used for multiple hypothesis testing.

3.4.2 Regular proteins, micropeptides, and lncRNAs have similar SNP spectra

We generated SNP spectra for regular proteins, micropeptides, and lncRNAs by binning SNPs based on their derived allele frequency (DAF). As shown in Figure 3.2, the 3 categories share nearly identical SNP spectra, with over 90% of SNPs having a DAF < 0.05.

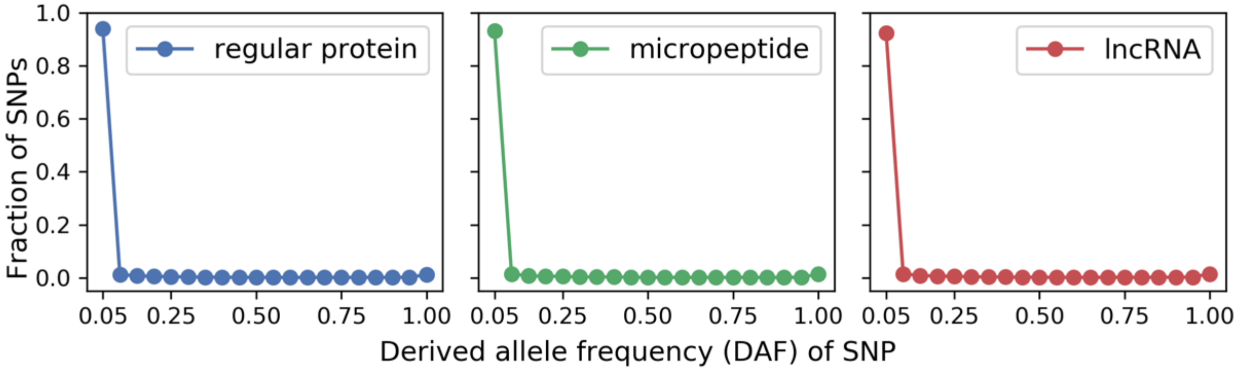


Figure 3.2. The derived allele frequency (DAF) spectra of SNPs of regular proteins, micropeptides, and lncRNAs. Spectra were plotted by binning SNPs into ranges of (0, 0.05), [0.05, 0.10), …, [0.95, 1) based on a SNP's DAF.

3.4.3 Regular proteins and micropeptides are under slightly stronger purifying selections than lncRNAs

As regular proteins, micropeptides, and lncRNAs share a similar SNP spectrum, with the majority of SNPs having a DAF < 0.05, it is more efficient and straightforward to focus only on the low frequency SNPs when comparing the 3 categories. SNPs with DAF < 0.05 occur infrequently in the population, *i.e.*, they are rare SNPs. A number of previous studies have used the fraction of rare SNPs for genetic variation analysis (Jha et al., 2015; Khurana et al., 2013), and besides 0.05, other DAF cut-off / threshold values were also used (Khurana et al., 2013).

In this study, we considered 3 different thresholds - DAF < 0.05, < 0.01, and < 0.005, for defining rare SNPs for a more comprehensive analysis. As shown in Figure 3.3, which is also reflected in Figure 3.2, regular proteins, micropeptides, and lncRNAs have similar fractions of rare derived alleles (FDA) with minor differences. For SNPs of DAF < 0.05, regular proteins and micropeptides have FDAs of 0.940 and 0.932, respectively. Although the difference is statistically significant, it is < 1%. Furthermore, lncRNAs have a FDA of 0.924, which is < 1% lower than that of micropeptides. The same trend is also observed with DAF < 0.01 and DAF < 0.005. These results suggest regular proteins and micropeptides are under slightly stronger purifying selection than lncRNAs.

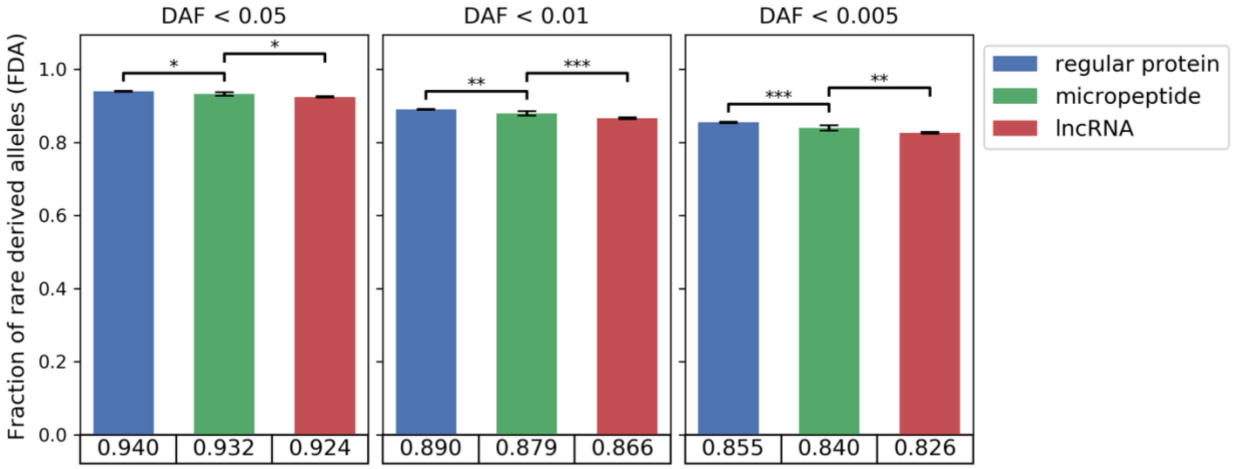


Figure 3.3. Fraction of rare derived alleles (FDA) in regular proteins, micropeptides, and lncRNAs under different derived allele frequency (DAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions. “*”, “**”, and “***” denote a p-value < 0.05, < 0.01, and < 0.001, respectively. Bonferroni correction was used for multiple hypothesis testing.

3.4.4 The coding regions of regular proteins and micropeptides are under the same purifying selection strength

To further examine protein coding transcripts, we analyzed the SNPs in coding regions in particular. As shown in Figure 3.4, considering SNPs with DAF < 0.05, the CDS of regular proteins have a FDA of 0.953, compared to 0.949 in the CDS of micropeptides. Although the former is slightly higher than the latter, the difference is not statistically significant, despite the large data size. The same pattern is observed for alleles with DAF < 0.01 and DAF < 0.005. The CDS of micropeptides are generally much shorter than that of regular proteins. Based on our analysis, regular proteins have a median CDS length of 1698 bp while that of micropeptides is 231 bp. Our results suggest that despite of the wide discrepancy in CDS length, micropeptides and regular proteins are under the same purifying selection strength.

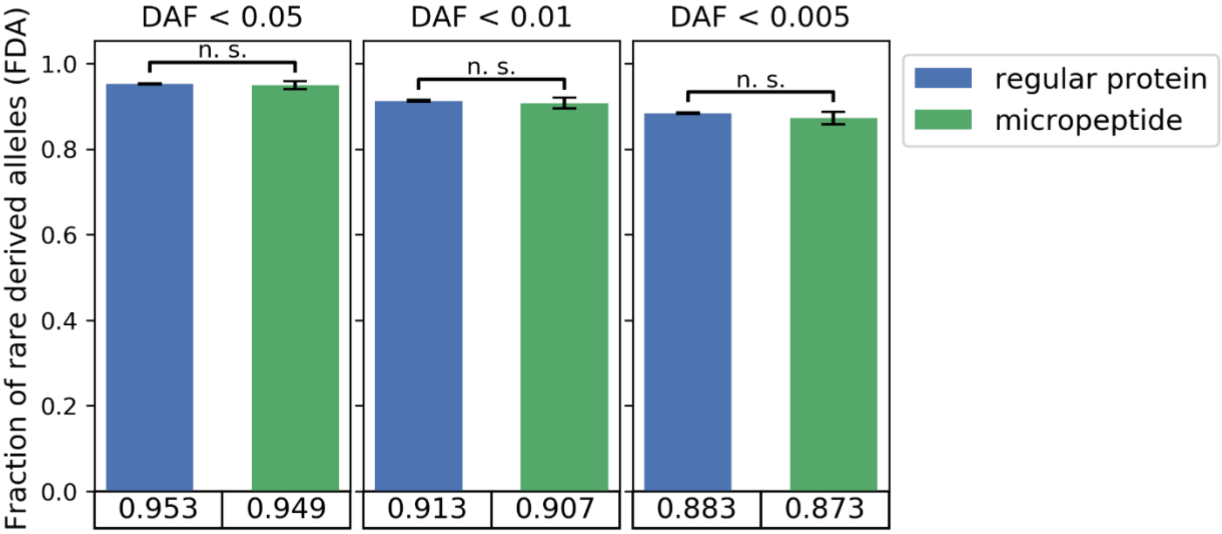


Figure 3.4. Fraction of rare derived alleles (FDA) in the coding sequences (CDS) of regular proteins and micropeptides under different derived allele frequency (DAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions. “n. s.” denotes “not significant” and a corresponding p-value >= 0.05.

In coding regions, nonsynonymous SNPs usually have greater biological (phenotypic) effects than their synonymous counterparts due to the direct change they contribute to the amino acid sequence. Yet, studies have shown that synonymous SNPs also can be associated with protein functional alterations and diseases (Chu & Wei, 2019; Rogozin et al., 2018; Simhadri et al., 2017). Nevertheless, nonsynonymous mutations are more likely to be under strong selective pressure. When we examine nonsynonymous derived alleles in coding regions (Figure 3.5), we see a similar pattern to Figure 3.4 - the CDS-FDAs in regular proteins and micropeptides show no significant differences with DAF < 0.05 and < 0.01. For alleles with DAF < 0.005, the CDS-FDA of regular proteins is 1% higher than that of micropeptides. Overall, these results further suggest that the coding regions in regular proteins and micropeptides can be considered to be under the same selection pressure.

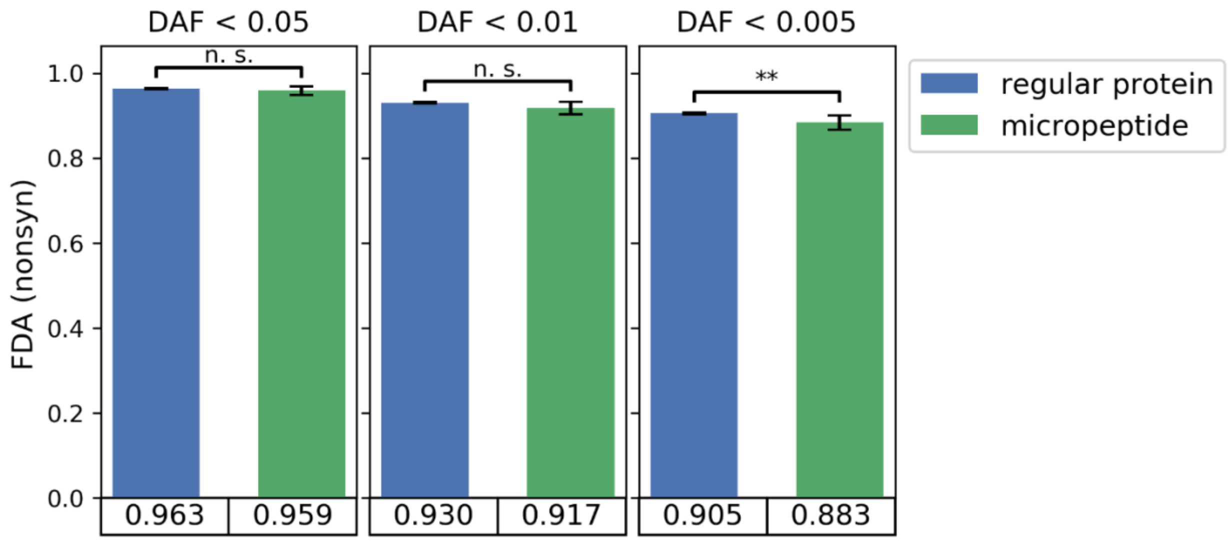


Figure 3.5. Fraction of rare *nonsynonymous* derived alleles (FDA, nonsyn) in the coding sequences (CDS) of regular proteins and micropeptides under different derived allele frequency (DAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions. “n. s.” denotes “not significant” and a corresponding p-value >= 0.05. “**” denotes a p-value < 0.01.

3.4.5 Predicted CDS regions in lncRNAs are under slightly weaker purifying selection than micropeptides

As a growing number of lncRNAs have been reannotated as protein coding, we attempt to assess the genetic variation profile of lncRNAs from the perspective of their potential coding capabilities. We used MiPepid (M. Zhu & Gribskov, 2019) and selected potential coding regions in lncRNA transcripts (see Methods for details). We then examined the SNPs in those MiPepid-predicted coding regions.

As shown in Figure 3.6, when alleles with DAF < 0.05 are used, the CDS-FDA of lncRNAs is lower than that of micropeptides, with a 2% difference. The same trend is seen with alleles with DAF < 0.01 and < 0.005. Similar results were obtained using only nonsynonymous SNPs (Supplemental Figure 3.1).

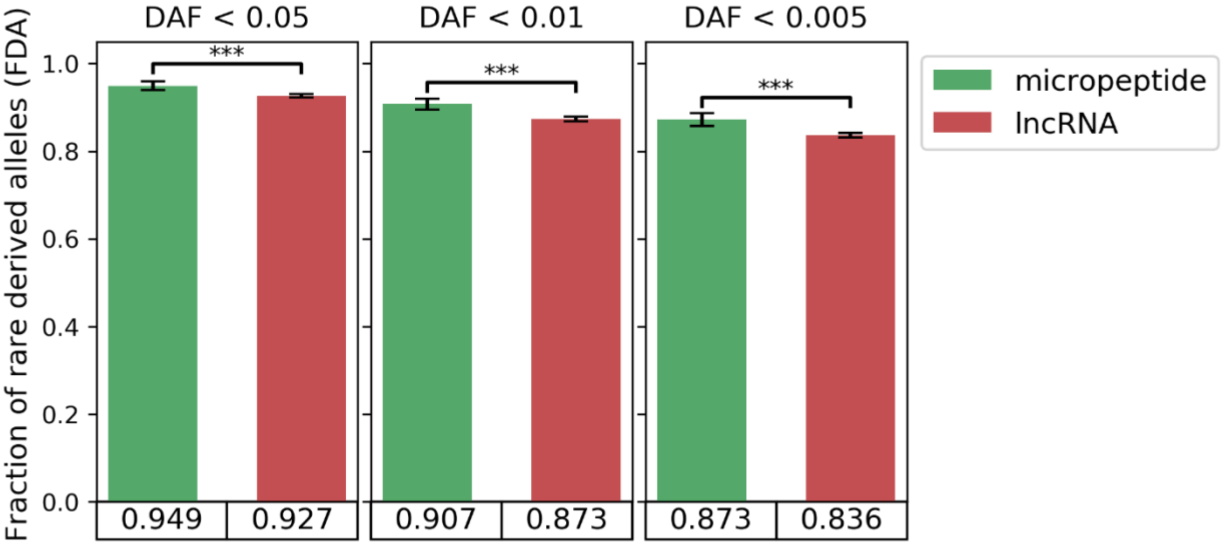


Figure 3.6. Fraction of rare derived alleles (FDA) in the coding sequences (CDS) of micropeptides and the MiPepid-predicted CDS of lncRNAs under different derived allele frequency (DAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions. "***" denotes a p-value < 0.001.

3.4.6 The majority of micropeptides and lncRNAs only have rare derived alleles in their (predicted) coding regions

To further examine the results in Figure 3.6, we looked at the FDA in each individual transcript. We found that in micropeptide transcripts 73% of them have a CDS-FDA of 1, *i.e.* all the SNPs in the coding region are rare derived alleles. And in MiPepid-predicted coding regions of lncRNA transcripts, this fraction is 66%. (Figure 3.7)

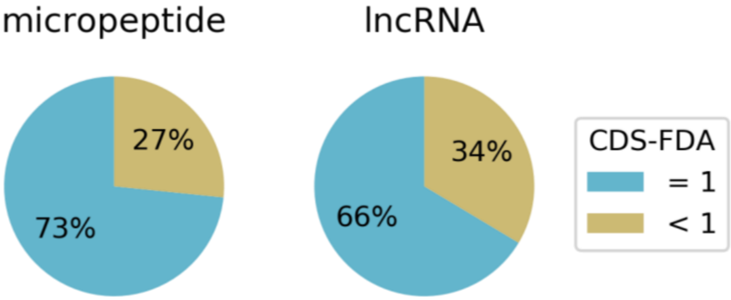


Figure 3.7. Distribution of CDS-FDA in micropeptide and lncRNA transcripts.

A similar pattern is observed when using nonsynonymous SNPs only (Supplemental Figure 3.2).

3.4.7 (Predicted) untranslated regions (UTR) in regular proteins, micropeptides, and lncRNAs are under the same purifying selection strength

We examined SNP profiles in UTR of regular proteins, UTR of micropeptides, and predicted UTR of lncRNAs. As shown in Figure 3.8, the UTR show the same level of FDA across the 3 categories under any of the 3 DAF thresholds. These indicate the (predicted) untranslated regions are under the same level of selective constraint regardless of their origin.

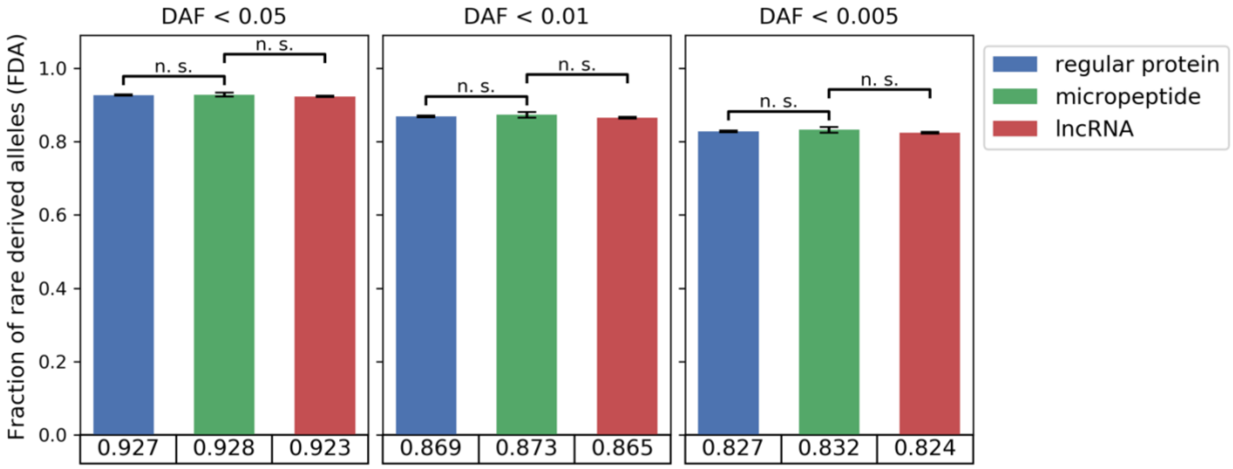


Figure 3.8. Fraction of rare derived alleles (FDA) in the untranslated regions (UTR) of regular proteins, micropeptides, and lncRNAs under different derived allele frequency (DAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions. “n. s.” denotes “not significant” and a corresponding p-value >= 0.05. Bonferroni correction was used for multiple hypothesis testing.

3.4.8 CDS regions are under stronger purifying selection than UTR regions

Many studies support the idea that the coding region is under stronger purifying selection than the untranslated regions of protein coding transcripts (Jha et al., 2015; Khurana et al., 2013; Mu et al., 2011). To analyze the relationship between CDS and UTR with respect to the SNP profile, we calculated the CDS-FDA to UTR-FDA ratios for regular proteins and micropeptides. As shown in Figure 3.9, with any of the three DAF thresholds, the CDS/UTR FDA ratios are significantly above 1 in both regular proteins and micropeptides. These results are consistent with the belief that CDS regions are under stronger purifying selection than UTR regions.

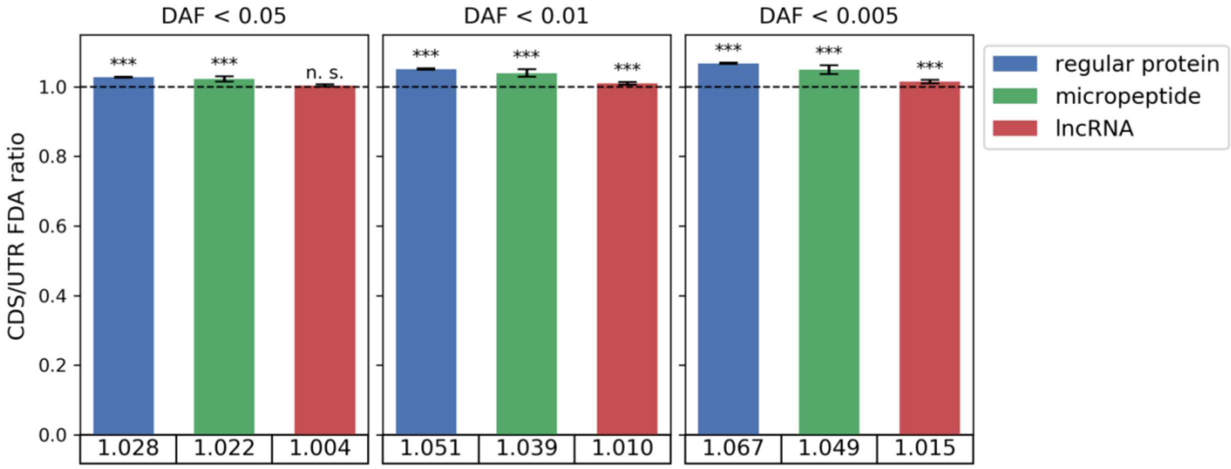


Figure 3.9. The ratio between CDS-FDA and UTR-FDA for regular proteins, micropeptides, and lncRNAs under different derived allele frequency (DAF) thresholds. Error bars denote 95% bootstrapped confidence intervals of ratios (with 5,000 bootstrap samples). "***" denotes the ratio is significantly greater than one with a p-value < 0.001. "n. s." denotes the difference between the ratio and one is not significant, and the corresponding p-value >= 0.05.

We also compared the enrichment of rare derived alleles in MiPepid-predicted CDS of lncRNAs to that of the corresponding predicted UTR in lncRNAs. The pattern is similar to that of proteins - the CDS/UTR FDA ratio of lncRNAs are also > 1 under any of the three DAF thresholds, and the differences for alleles with DAF < 0.01 and < 0.005 are both significant.

3.4.9 Analyses with minor alleles show similar results to those with derived alleles

Derived alleles are generally used in genetic variation analysis compared to minor alleles (Bhartiya et al., 2014; Khurana et al., 2013; Mu et al., 2011; Ward & Kellis, 2012). Yet, as shown in Table 3.1, the number of derived alleles is roughly only 1/10 of that of minor alleles. Derived alleles are relative to ancestral alleles, which are obtained through multiple-species alignments (Auton et al., 2015). That the ancestral alleles of 90% of SNPs cannot be identified reveals the small portion of conserved regions in human genome relative to close species.

To make the most of available variation data, we also used minor alleles for our analysis. As shown in Supplemental Figure 3.3 – 3.12, results from minor alleles analysis in general share the same patterns as observed in results from derived alleles. Regular proteins, micropeptides, and lncRNAs share almost identical SNP spectra of minor allele frequency (MAF) (Supplemental Figure 3.3). LncRNAs show a slightly lower fraction of minor alleles (FMA) than proteins

(Supplemental Figure 3.4). The CDS regions of regular proteins and micropeptides share the same level of FMA using alleles with MAF < 0.05 and < 0.01; although for alleles with MAF < 0.005, the FMA of regular proteins is higher than that of micropeptides, the difference is small (~ 1%) (Supplemental Figure 3.5). We do see a statistically significantly higher FMA in CDS of regular proteins than in micropeptides, but only when considering nonsynonymous SNPs. This difference from what is seen with derived alleles can be explained by the much larger number of minor alleles compared to derived alleles. The MiPepid-predicted CDS regions of lncRNAs show a lower FMA than micropeptide CDS (Supplemental Figure 3.7 – 3.8), which is further explained by the fact that 80% of micropeptide transcripts have only rare minor alleles in the CDS, compared to 75% of lncRNA transcripts (Supplemental Figure 3.9), or 88% vs. 80% when only considering nonsynonymous SNPs (Supplemental Figure 3.10). Although UTRs show slightly different FMAs across the three categories (Supplemental Figure 3.11), again this may be explained by a much larger number of observed minor alleles compared to derived alleles. And finally, the CDS regions are more enriched with rare minor alleles than UTRs across all the three categories (Supplemental Figure 3.12).

The similarities between the results from derived alleles and those from minor alleles strongly suggest that minor alleles can also be used for genetic variation analysis.

3.5 Discussion

3.5.1 SNP densities are similar across proteins and lncRNAs

A number of previous studies used SNP density to evaluate selection strength, including (Bhartiya et al., 2014; Ward & Kellis, 2012), etc. One former study (Bhartiya et al., 2014) found that the SNP density of lncRNAs was at least an order higher than protein-coding transcripts, with lncRNAs having 14.72 SNPs/kb, CDS of protein-coding genes having 0.37 SNPs/kb, and UTR regions having 0.46 SNPs/kb. However, the SNP data they used came from HapMap (Durbin et al., 2010; Gibbs et al., 2003), which used known probes to discover SNPs, therefore can have a confirmation bias (De Silva et al., 2014; Mu et al., 2011). The 1000 Genomes project (Auton et al., 2015) was launched after HapMap. In this study, we used the complete final phase data from 1000 Genomes project, which has over 84 million SNPs, and we showed that the both lncRNAs and protein-coding transcripts have a SNP density close to 30 SNPs/kb, which are both much higher than the findings in the previous study. In addition, we showed that regular proteins, micropeptides, and lncRNAs share very similar SNP densities.

3.5.2 Difference in overall enrichment of rare SNPs between regular proteins and micropeptides can be explained by CDS length discrepancy

Based on the results section, the fraction of rare derived alleles (FDA) in the CDS regions of micropeptides is indistinguishable from the CDS-FDA in regular proteins (Figure 3.4); the FDA in the UTR regions of micropeptides is also similar to that of UTR-FDA in regular proteins (Figure 3.8). Based on this, one might expect that the overall FDA in micropeptides should be similar to that of regular proteins. However, as shown in Figure 3.3, the overall FDA of micropeptides is *lower* than that of regular proteins. This may seem confusing, yet it can easily be explained. As shown in Table 3.2, although a typical regular protein has a longer UTR region (1,336 bp) than a typical micropeptide (984 bp), it has a much longer CDS region (1698 bp) compared to a micropeptide (231 bp). Consequently, compared to the full sequences, the CDS regions take 56% in regular proteins, but only take 12% in micropeptides. Therefore, although both CDS-FDAs and the UTR-FDAs are the same in regular proteins and micropeptides, the overall FDA in micropeptides in lower compared to regular proteins, since they have much shorter CDS regions.

Table 3.2. Sequence length comparisons between regular proteins and micropeptides

	regular protein	**micropeptide**
CDS median length (bp)	1,698	231
UTR median length (bp)	1,336	984
CDS fraction by length	0.56	0.12

3.5.3 lncRNAs have similar SNP profiles to those of proteins

The *ad hoc* definition of lncRNA as a transcript that is > 200 bp and is not translated into proteins (L. Ma et al., 2013) is often considered too generic and does not contain much valuable information regarding the functions of the transcript. The ongoing re-discovery of the coding capacities in this lncRNA family further obscures this definition. In this study, we showed that lncRNAs share a similar SNP density, a similar SNP spectrum, and a similar fraction of rare SNPs with protein coding transcripts, and in particular with micropeptides, indicating lncRNAs are under similar purifying selection strength to protein-coding transcripts. This similarity may also suggest functional overlaps between lncRNAs and proteins.

3.5.4 Predicted CDS regions in lncRNAs are also under stronger purifying selection than UTR regions

CDS regions of regular proteins are under stronger purifying selection than UTR regions; and this has been confirmed by many previous studies (Jha et al., 2015; Khurana et al., 2013; Mu et al., 2011). Our findings once again confirmed this trend. In addition, we attempted to use MiPepid to predict potential CDS regions in lncRNAs, and the predicted CDS are also under stronger purifying selection than the rest UTR regions. Together with the predicted fraction of individual lncRNA CDS that has a FDA of 1, which is close to the fraction in micropeptides, all the numbers suggest MiPepid predictions captured the coding potentials in lncRNAs, which is consistent with the fact that more and more lncRNAs are being reannotated to be protein-coding.

3.5.5 More rare SNPs were identified compared to common SNPs with increased sample size

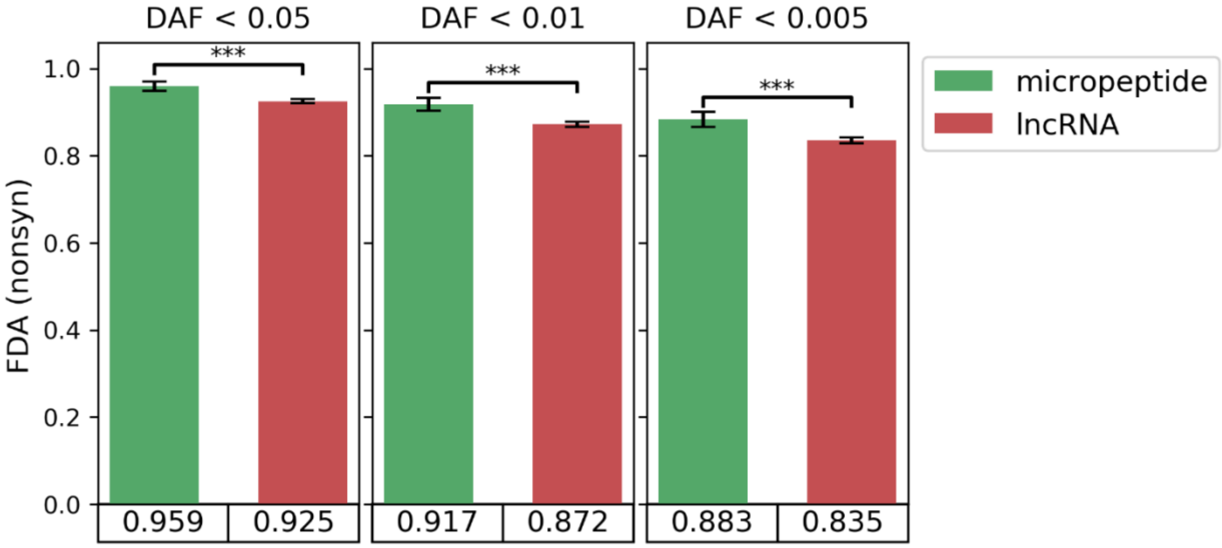
Using data from the pilot phase of the 1000 Genomes project (Durbin et al., 2010), which contains variations from 179 individuals, Mu et al. (Mu et al., 2011) generated derived allele SNP spectra for CDS and UTR regions of proteins and some other noncoding elements (not including lncRNAs). The spectra show exponential decay, and similarly to the spectra we showed in Figure 3.2, the largest fraction of SNPs have frequency < 0.05; however, they only account for around 30% of all SNPs. Later, Jha et al. (Jha et al., 2015) used the phase 1 data (1,092 individuals) (McVean et al., 2012) and showed the fraction of derived alleles with a frequency < 0.05 in CDS regions was around 60%. In our findings, with the phase 3 data (2,504 individuals) of the 1000 Genomes project (Auton et al., 2015), the fractions of rare derived alleles with a frequency < 0.05 are above 90% in both proteins and lncRNAs. From this trend, it is clear that with expanded sample sizes from the pilot phase to phase 1, then to phase 3, many more rare SNPs have been identified.

3.5.6 Minor alleles are suitable for genetic variation analysis

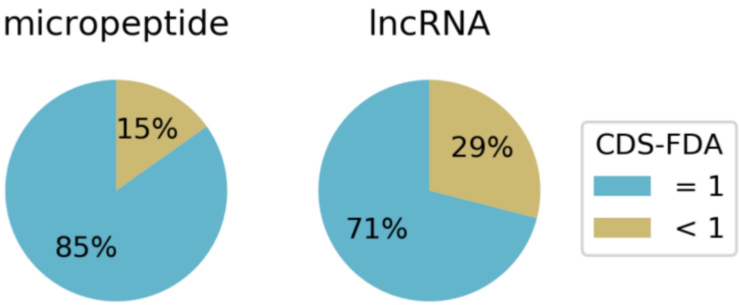
Derived alleles have long been used for SNP variation analysis, but minor alleles have been used only rarely. Yet the ten-fold difference in the number of derived and minor alleles makes us reconsider the usability of minor alleles and their relationships with derived alleles. In this study, we showed that even though minor alleles are much more abundant than derived alleles, SNP profiles obtained using minor alleles are similar to those of derived alleles. Therefore, minor alleles can also be used for genetic variation analysis.

3.6 Supplemental Figures

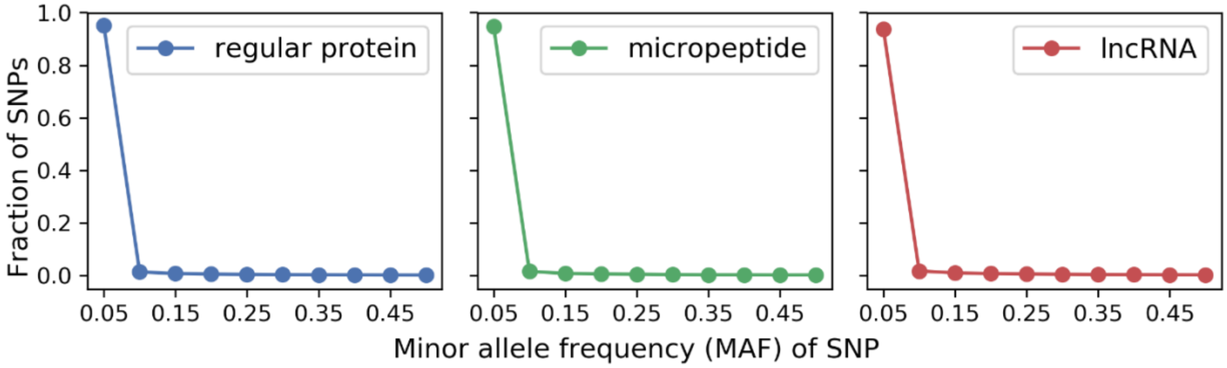
For statistical test results in the following supplemental figures, “n. s.” denotes “not significant” and a corresponding p-value >= 0.05; “*”, “**”, and “***” denote a p-value < 0.05, < 0.01, and < 0.001, respectively; Bonferroni correction was used for multiple hypothesis testing.



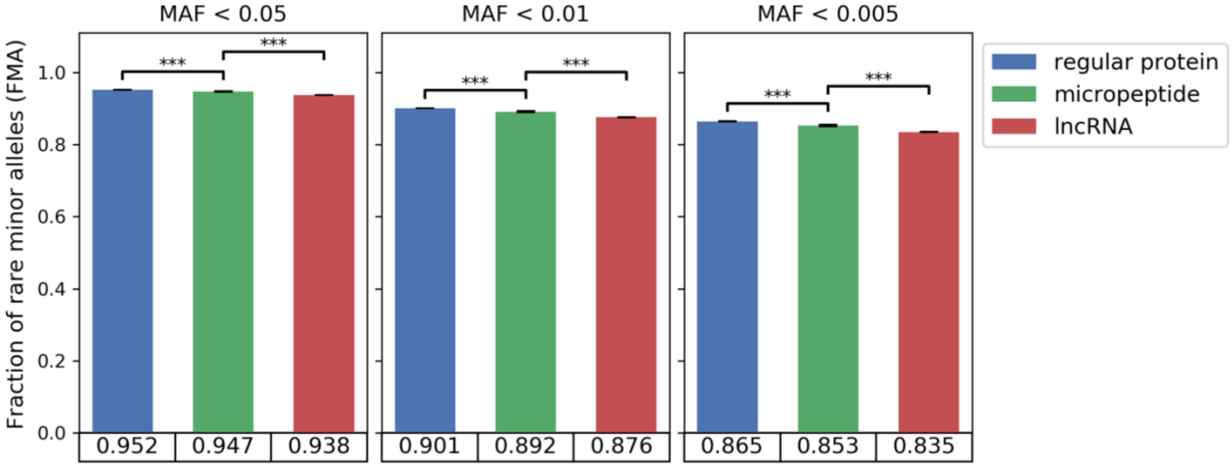
Supplemental Figure 3.1. Fraction of rare *nonsynonymous* derived alleles in the coding sequences (CDS) of micropeptides and the MiPepid-predicted CDS of lncRNAs under different derived allele frequency (DAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions.



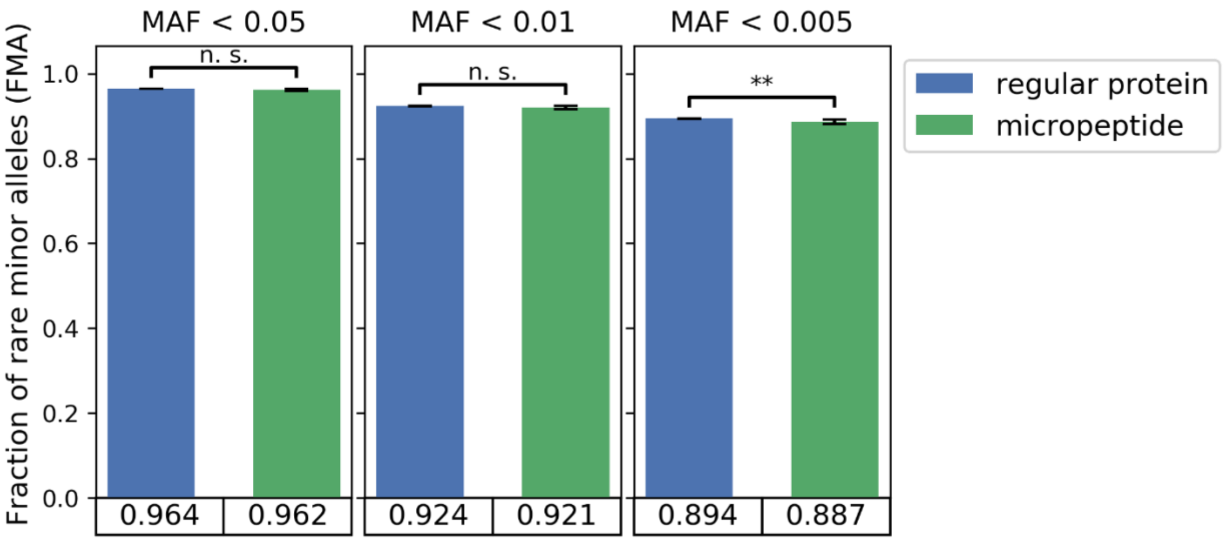
Supplemental Figure 3.2. Distribution of *nonsynonymous* CDS-FDA in micropeptide and lncRNA transcripts.



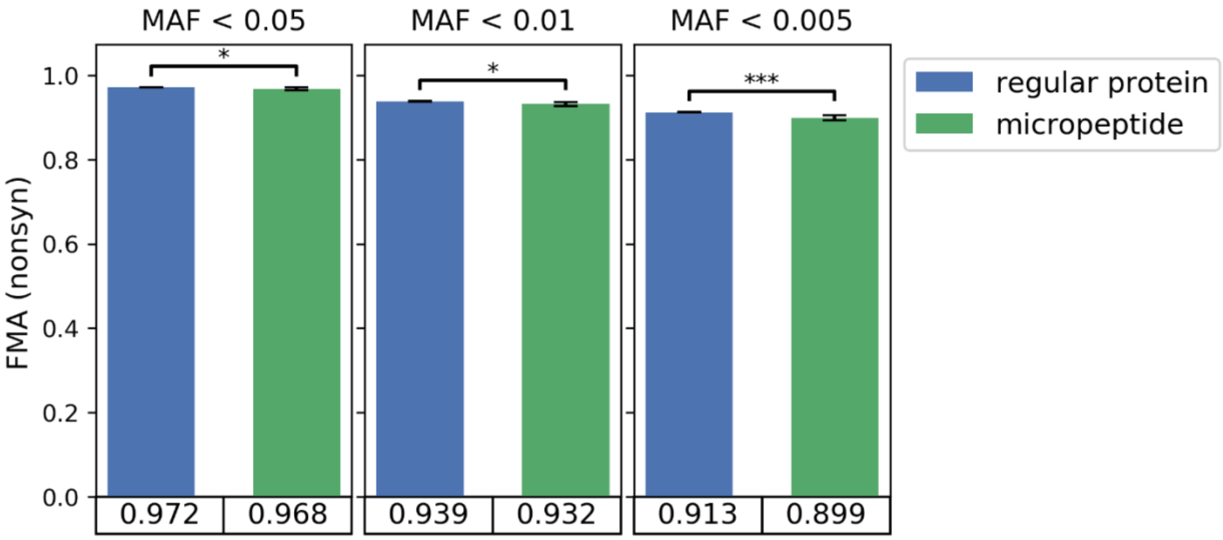
Supplemental Figure 3.3. The minor allele frequency (MAF) spectra of SNPs of regular proteins, micropeptides, and lncRNAs. Spectra were plotted by binning SNPs into ranges of (0, 0.05), [0.05, 0.10], …, [0.45, 0.5) based on a SNP’s MAF.



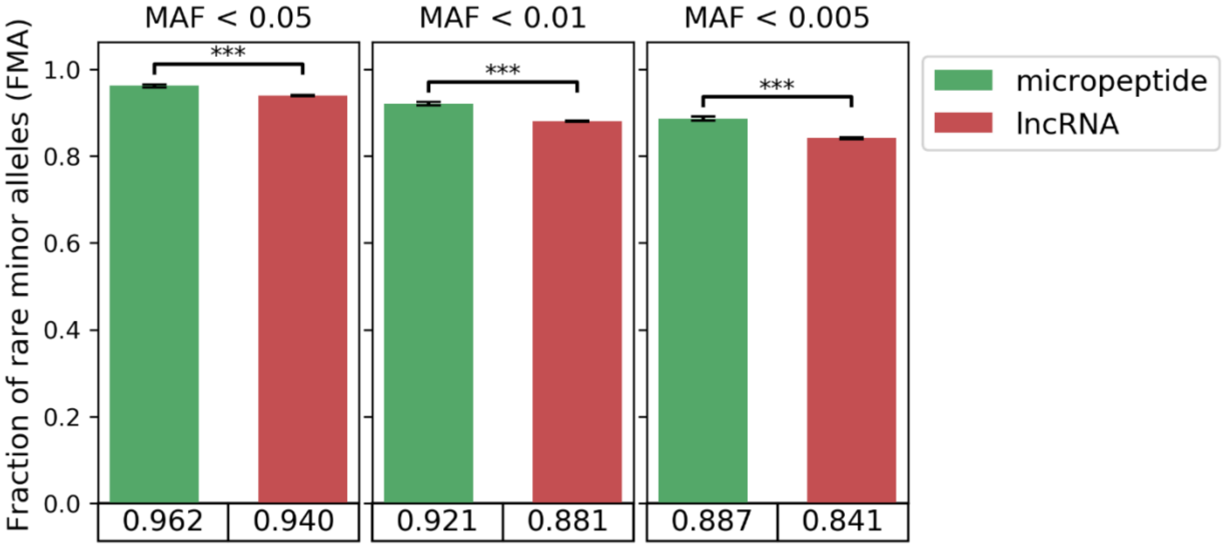
Supplemental Figure 3.4. Fraction of rare minor alleles (FMA) in regular proteins, micropeptides, and lncRNAs under different minor allele frequency (MAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions.



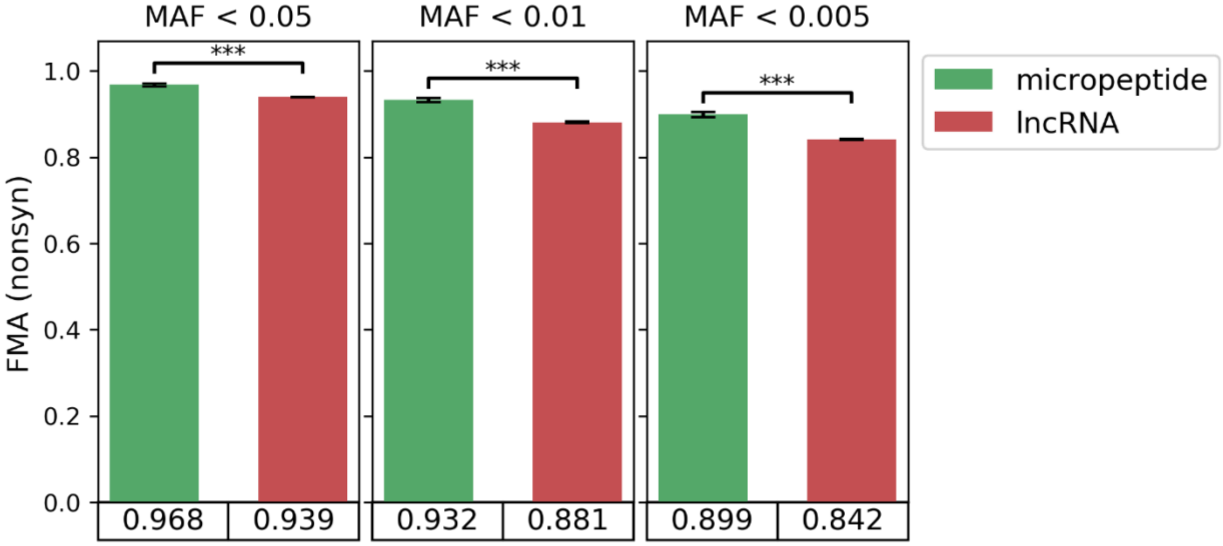
Supplemental Figure 3.5. Fraction of rare minor alleles (FMA) in the coding sequences (CDS) of regular proteins and micropeptides under different minor allele frequency (MAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions.



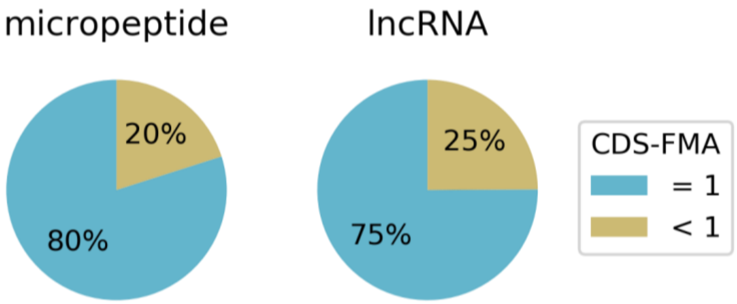
Supplemental Figure 3.6. Fraction of rare *nonsynonymous* minor alleles (FMA, nonsyn) in the coding sequences (CDS) of regular proteins and micropeptides under different minor allele frequency (MAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions.



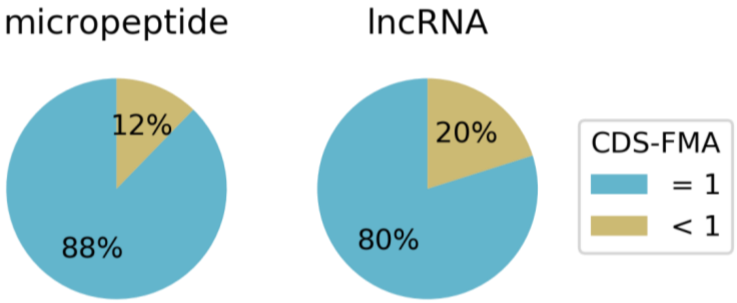
Supplemental Figure 3.7. Fraction of rare minor alleles (FMA) in the coding sequences (CDS) of micropeptides and the MiPepid-predicted CDS of lncRNAs under different minor allele frequency (MAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions.



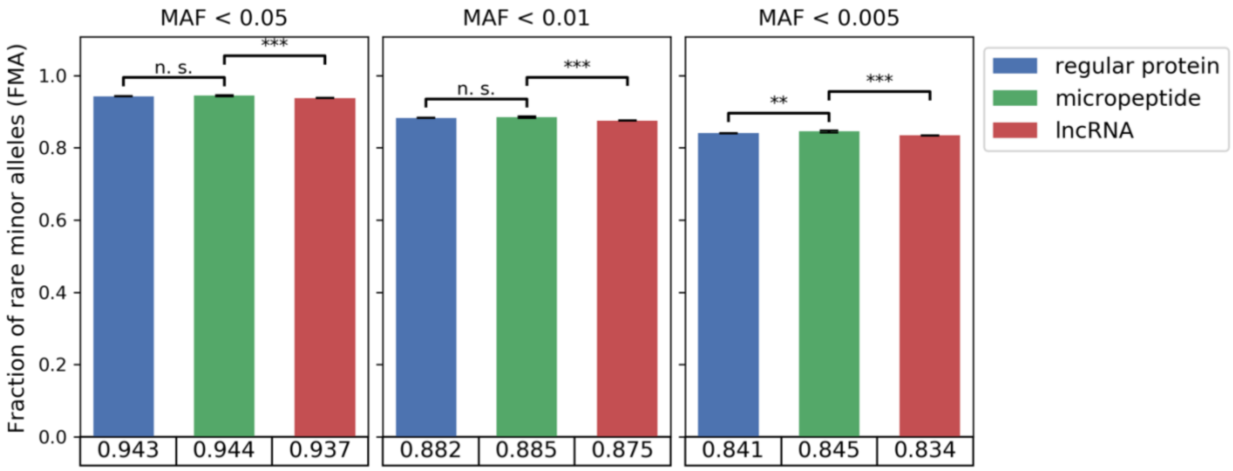
Supplemental Figure 3.8. Fraction of *nonsynonymous* rare minor alleles (FMA) in the coding sequences (CDS) of micropeptides and the MiPepid-predicted CDS of lncRNAs under different minor allele frequency (MAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions.



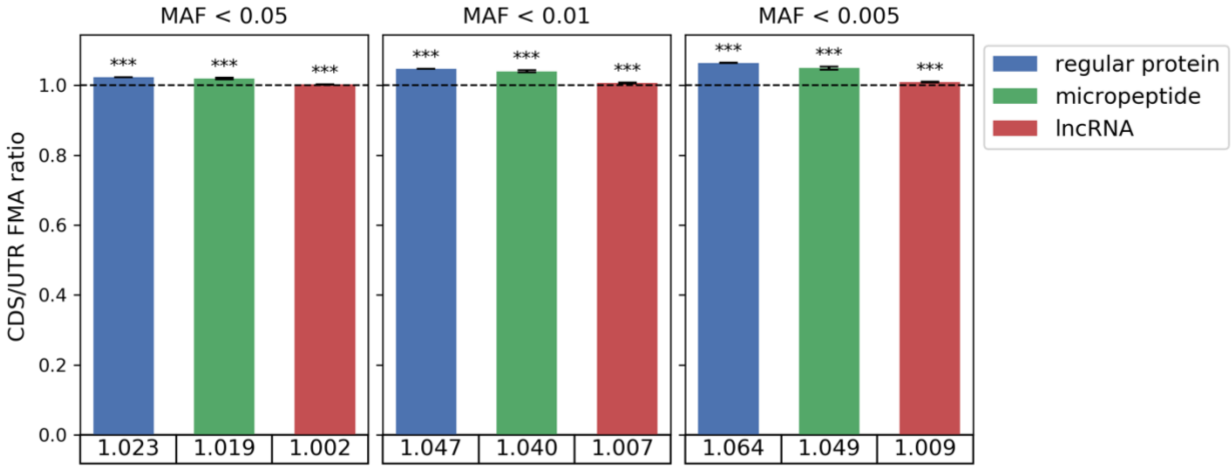
Supplemental Figure 3.9. Distribution of CDS-FMA in micropeptide and lncRNA transcripts.



Supplemental Figure 3.10. Distribution of nonsynonymous CDS-FMA in micropeptide and lncRNA transcripts.



Supplemental Figure 3.11. Fraction of rare minor alleles (FMA) in the untranslated regions (UTR) of regular proteins, micropeptides, and lncRNAs under different minor allele frequency (MAF) thresholds. Error bars denote 95% confidence intervals of binomial proportions.



Supplemental Figure 3.12. The ratio between CDS-FMA and UTR-FMA for regular proteins, micropeptides, and lncRNAs under different minor allele frequency (MAF) thresholds. Error bars denote 95% bootstrapped confidence intervals of ratios (with 5,000 bootstrap samples). "***" denotes the ratio is significantly greater than one with a p-value < 0.001.

CHAPTER 4. PROLIG: PROTEIN BINDING LIGAND PREDICTION USING 3D DEEP CONVOLUTIONAL NEURAL NETWORKS[2]

4.1 Abstract

Most proteins bind ligands to function. Therefore, predicting protein binding ligands is important in that it can be used for large-scale protein function annotation, drug design, predicting drug side effects, etc. By representing protein pockets as 3D images with physicochemical information included, using 3D deep convolutional neural networks, we developed a computational framework named ProLig that can predict the binding ligand for a protein pocket among over 150 different ligand species. ProLig was trained on a large-scale dataset of over 77, 000 pockets and works considerably better than existing non-deep-learning methods. The overall top-1 accuracy is over 71%, and the top-3 and top-10 accuracies are > 83% and > 91%, respectively. The F-score for as many as 37 ligand species are over 0.8, and 99 ligand species have a top-10 accuracy of over 80%. Our framework can still perform very well when only the shape of a pocket is known and does not entail any other information regarding the pocket.

4.2 Introduction

In biological cells, most proteins perform their functions upon binding small molecules called ligands. Most ligands are bound to areas on protein surfaces, and these areas are protein pockets. The correct prediction of the binding ligand to a protein pocket is of great significance since it can be used for protein function annotation, drug design, as well as identifying potential drug side effects.

Existing methods for predicting pocket binding ligands are roughly in two categories: global-structure-based and local-structure-based.

Methods that use the global structure information of the target protein generally retrieve structurally similar protein complexes as templates and use the bound ligands of those templates as candidates for the binding ligand prediction. These methods include FINDSITELHM (Brylinski

[2] The work of this chapter was done during my stay in Dr. Daisuke Kihara's lab. Later on, I joined Dr. Michael Gribskov's lab in May 2018. This work was funded by the National Science Foundation (DMS1614777).

& Skolnick, 2009), FunFOLD (Roche et al., 2011), Fun-FOLD2 (Roche et al., 2013), GalaxySite (Heo et al., 2014). However, proteins that bind the same ligand could possibly have very different global structures, while those that bind different ligands may share similar global structures.

Local-structure-based methods employ only the local structure information of the pocket itself without considering the global structure of the target protein. Among those methods, pockets are either represented by pseudo-atoms (Hoffmann et al., 2010), all atoms (Najmanovich et al., 2008), Cα atoms (Gao & Skolnick, 2013), or by reduced descriptors such as spherical harmonics (Morris et al., 2005), 3D Zernike descriptors (Chikhi et al., 2010; Sael & Kihara, 2012; X. Zhu et al., 2015). Most of these methods are designed for pocket comparison, either by using alignment algorithms or by computing the distances between the reduced descriptors. While the pocket comparison methods can be used for pocket binding ligand prediction, mostly by comparing the target pocket against a database of pockets with known binding ligands, they are subject to the following two drawbacks: 1. Pockets that bind to the same ligand could be different and those that bind to different ligands could be similar; 2. The database of pockets for comparing against needs to be carefully selected and cross validated, and the computation time grows linearly with the number of pockets in the database.

Deep learning is a set of machine learning algorithms primarily for learning hidden representations from raw data (Bengio et al., 2013; LeCun et al., 2015; Schmidhuber, 2015). In recent years, deep learning has achieved state-of-the-art performances in fields such as computer vision (Krizhevsky et al., 2012). The deep convolutional neural networks have brought huge advances in image classification with performances even surpassing humans (Ciregan et al., 2012).

Protein pockets are 3D structures and can be represented as 3D images. The pocket binding ligand prediction problem can therefore be transformed into the 3D image classification problem, which is analogous to 2D image classification. By employing the advantages of deep learning in this field, we developed the computational framework ProLig with deep learning at its core, that can predict the binding ligand given a pocket. ProLig was trained with a large-scale dataset of over 70, 000 pockets binding over 150 ligand types. The hidden representations used for classification were learned from the dataset therefore the prediction of an unknown pocket is fast and does not entail the comparison against any pocket database. ProLig performs significantly better than existing methods and could be implemented to many other computational research that also use

3D biomolecular structure data, such as protein-protein binding, protein-DNA/RNA binding, and EM map analysis.

4.3 Methods

4.3.1 Datasets

Pocket representation by 3D images

We define a pocket as the protein surface area that is within 4 Å of the binding ligand. Each pocket is represented as a set of 3D images, with each image containing information of one of the 4 features of this pocket: shape, electrostatic potential, hydrophobicity, visibility. Each image is a 25 × 25 × 25 voxel grid, with the resolution 1.5 Å for each voxel. The image of shape is binary, with the voxels occupied by the pocket having a value of 1, and the rest having a value of 0. The image of electrostatic potential is computed using the APBS program (Baker et al., 2001). We used the Kyte-Doolittle scale (Baker et al., 2001) to represent hydrophobicity, which ranges from -4.5 (hydrophilic) to 4.5 (hydrophobic). The visibility of a given voxel is defined as the number of visible directions from the voxel within a 10×10×10 Å3 cube. It ranges from 0 to 1, with 0 indicating the voxel is completely buried in a protein while 1 indicating the voxel is not near the protein. A large visibility value suggests that the voxel is located at a concave region, while a small one suggests convexness (Li et al., 2008). Figure 4.1 shows several examples of pocket shape images.

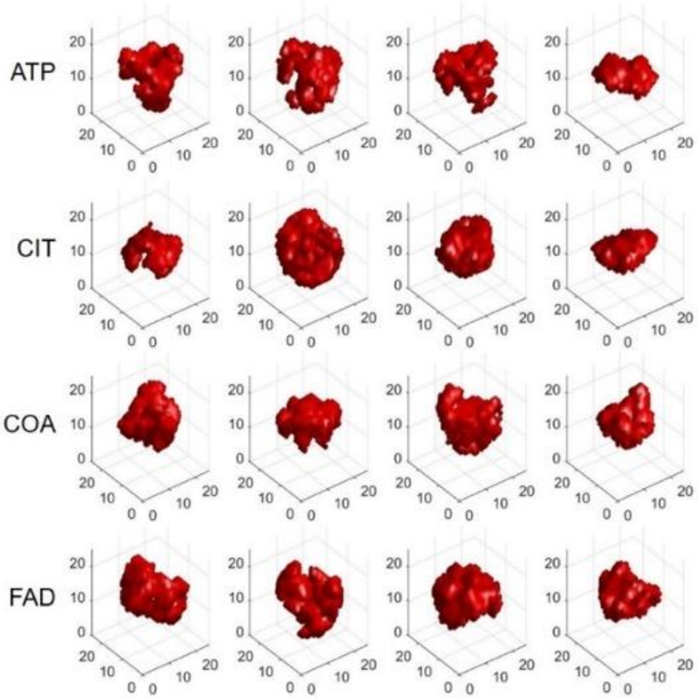


Figure 4.1. Examples of pockets represented by 3D images. Here the shapes of the pockets are rendered in 25 × 25 × 25 3D grids, and names of the corresponding ligands the pockets can bind are shown in the leftmost column. ATP: adeno-sine-5'-triphosphate; CIT: citric acid; COA: coenzyme A; FAD: flavin-adenine dinucleotide.

The whole dataset

We extracted pockets from ligand-bound protein complex structures in the Protein Data Bank (PDB) database (Berman et al., 2000). We selected ligand types that have a certain number of bound protein structures (over 50) in PDB, that are not ions, and that have a molecular weight of over 100 g/mol. This procedure finally generated a dataset consisting of 77,087 pockets that bind to 151 ligand species. Figure 2 shows the number of pockets for each ligand species.

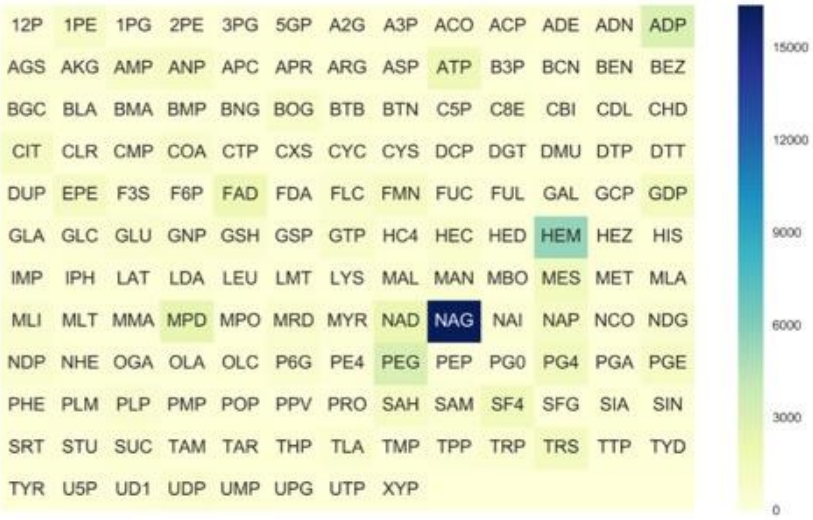


Figure 4.2. Heat map showing the number of pockets for each ligand species in the whole dataset.

The nonredundant datasets

To address the redundancy issue of the PDB data, besides our whole dataset, we also built several nonredundant datasets by culling the whole dataset. For a subset of pockets that bind a same ligand species, if their touching protein chains share pairwise sequence or structure similarity above a set of thresholds, we would then randomly keep only one pocket of the subset. We used TM-align (Y. Zhang & Skolnick, 2005) to determine the sequence similarity (Sequence Identity (SI) score) and structure similarity (TM score) for all the pairs of proteins in our whole dataset. We then collected 4 nonredundant datasets based on the cutoff values of SI-0.5, SI-0.8, TM-0.6, and TM-0.8, respectively. SI-0.5 means for all the pairs of pockets in the dataset, their touching protein chains do not share pairwise sequence identity score of over 0.5. The same rule applies to SI-0.8, TM-0.6, and TM-0.8, where TM represents TM score.

4.3.2 3D deep convolutional neural network (3D DCNN)

Architecture

To build our binding ligand prediction framework, we designed a 3D deep convolutional neural network (3D DCNN) to train our datasets. This 3D DCNN contains 11 layers, with 6 convolutional layers, 2 max pooling layers, and 2 fully-connected layers (Figure 4.3, Table 4.1).

The input layer takes multi-channel 3D images, as in our case each channel corresponds to one of the 4 features (shape, electrostatic potential, hydrophobicity, visibility).

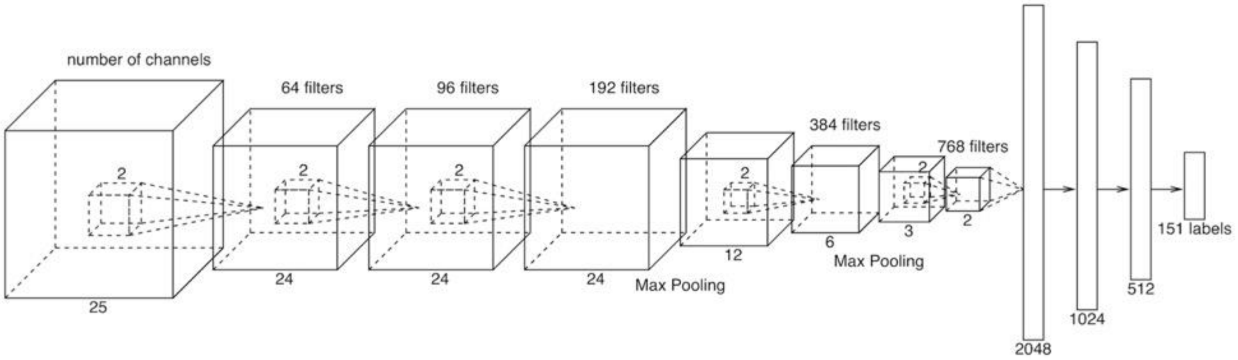


Figure 4.3. Diagram of the 3D deep convolutional neural network (3D DCNN). The input layer takes (multi-channel) 3D images. The convolution kernel size is 2×2×2 for all convolutional layers. For illustration purpose, only one channel of the input and only one filter of each convolutional layer are shown.

Table 4.1. The architecture of our 3D deep convolutional neural network (3D DCNN)

Layer No.	Layer Name	Filter Size	Stride	Layer size
1	Input			25×25×25×#Ch
2	Conv	2×2×2×#Ch×64	1	24×24×24×64
3	Conv	2×2×2×64×96	1	24×24×24×96
4	Conv	2×2×2×96×192	1	24×24×24×192
5	Pool	2×2×2×1	2	12×12×12×192
6	Conv	2×2×2×192×384	2	6×6×6×384
7	Pool	2×2×2×1	2	3×3×3×384
8	Conv	2×2×2×384×768	1	2×2×2×768
9	Conv	2×2×2×768×2048	1	1×1×1×2048
10	FC			1024
11	FC			512
12	Output			151

Conv: Convolutional; Pool: Max pooling; FC: Fully-connected; #Ch: number of channels

The hyperparameter setups in our 3D DCNN are as follows. We used a batch size of 128. We used L2 regularization for every hidden layer, and the regularization strength is 4.08×10^{-6}. The initial learning rate is 4.14×10^{-4}. Both the regularization strength and initial learning rate were tuned on the validation set. The learning rate was decayed by 0.99 every 10000 steps. We used Adam algorithm (Kingma & Ba, 2014) to do the optimization, which computes the adaptive learning rates from estimates of the first and second moments of the gradients, and the 2 hyperparameters in Adam were set to β1 = 0.9, β2 = 0.999, which correspond to the exponential decay rate for the first- and second-moment estimates, respectively. We also used moving average to enhance the performance, i.e. computing the moving averages of parameters such as weights and biases along each training step, and the moving average rate was set to 0.9999. The model was trained 100, 000 steps in total. We used ReLU (Glorot et al., 2011) as the activation function for each hidden layer, and we used the softmax classifier for the output layer. Our code was written using the TensorFlow framework (Abadi et al., 2016).

Training

Data splitting

The original data splitting procedure: for a dataset, we randomly split the dataset into training and test sets, with the ratio of the number of data in the 2 sets as 9-to-1. Since the numbers of data for different ligand types are unbalanced, when splitting the data, we kept the ratios between different classes (ligand types) as the same for both the training and the test sets.

We applied the original data splitting procedure to both the whole dataset and the 4 nonredundant datasets.

Data augmentation

A pocket in the 3D grid can take arbitrary orientation. To augment the dataset as well as to lessen the difficulty of prediction due to the orientation freedom issue, for each pocket in the training set we generate 100 poses by rotating the pocket randomly in the 3D space.

Model training

When training the model, for any of the whole dataset and the 4 nonredundant datasets, we selected a fraction of data from the training set as the validation set and tuned the hyperparameters of the model using the validation set; after hyperparameter tuning was done, we then used the whole training set to do the final model training.

Performance evaluation

There are generally 3 measures used in this paper to evaluate the performances of our method as well as other existing methods: top-k accuracy, mean top-k recall, and mean F-score (k = 1, 2, …). For a given pocket, the 3D DCNN in our method will predict its binding ligand by giving a ranking of 151 potential binding ligands, corresponding to the 151 different ligand species in our dataset. The top-k accuracy is computed by evaluating the overall accuracy on the test set.

For example, the top-3 accuracy is the percentage of pockets in the test set whose true binding ligand (true label) is among the top-3 predicted binding ligands (top-3 predictions). Among the predicted results for a dataset consisting of at least 2 classes, for an example in a particular class, if the predicted label is correct, then this predicted result is called a true positive (TP) for this class; if it is incorrect, then this is a false negative (FN); if an example in another class is predicted to be in this class, then this example if a false positive (FP) for this class. The precision for this class is precision = #TP / (#TP + #FP), where #TP refers to the number of true positives, and the same as #FP. The recall for this class is recall = #TP / (#TP + #FN), so the recall can also be interpreted as the accuracy for this particular class. The top-k recall for a class is then the number of examples predicted correctly within top-3 predictions divided by the total number of examples in this class. The mean top-k recall is the average top-k recall across all the classes in a dataset, which can also be interpreted as the top-k average accuracy. The F-score is computed by F-score = 2 * (precision * recall) / (precision + recall). So, the F-score for a class is computed from the precision and top-1 recall for this class, and the mean F-score is the average F-score across all the classes in a dataset.

4.4 Results

4.4.1 Overall results

Using the whole dataset with 2 feature channels (hydrophobicity and visibility), trained with the neural network shown in Figure 4.3 and Table 4.1 and with the hyperparameter settings specified in the Methods section, with 151 classes, our method ProLig achieved on the test set top-1 accuracy of 0.711, top-5 accuracy of 0.877, and top-10 accuracy of 0.919, and the mean F-score of the 151 classes is 0.615 (Table 4.2).

Table 4.2. Overall results on the whole dataset trained with the neural network in Table 4.1

Accuracy	**top-1**	**top-3**	**top-5**	**top-10**
	.711	.836	.877	.919
Mean Recall	**top-1**	**top-3**	**top-5**	**top-10**
	.594	.719	.766	.815
Mean F-score	.615			

Mean top-k (k = 1, 3, 5, 10) recall is equivalent to the top-k average accuracy across the 151 classes; mean F-score is the average F-score across the 151 classes

To show that we indeed randomly split the dataset into training and test, we constructed 2 follow-up datasets in which part of the training set was used for testing, and the rest of the training set + the original test set for training. The 2 follow-up datasets both gave similar results to those in Table 2 (See Supplemental Table 4.1 for details).

ProLig showed considerably good results on some selected classes, with 37 classes the F-score being over 0.8 (Table 4.3). And figure 4 shows the distributions of the F-scores and selected top-k recalls.

Table 4.3. Classes with the F-score over 0.8

ligand	**precision**	**Recall**				**F-score**
		top-1	**top-3**	**top-5**	**top-10**	
THP	1.000	1.000	1.000	1.000	1.000	1.000
BTN	1.000	.944	1.000	1.000	1.000	.971
MAL	.963	.963	.963	.963	.963	.963
HEM	.951	.971	.989	.991	.993	.961
TRP	.950	.950	.950	.950	.950	.950
PPV	.875	1.000	1.000	1.000	1.000	.933
HIS	.857	1.000	1.000	1.000	1.000	.923
MBO	1.000	.857	.857	.857	.857	.923
SIA	.895	.944	.944	.944	1.000	.919
NAG	.874	.961	.986	.990	.996	.915
BMP	1.000	.833	.833	.833	.833	.909
FAD	.892	.916	.979	.984	.989	.904
GLU	.933	.875	.875	.896	.917	.903
TPP	.917	.880	1.000	1.000	1.000	.898
HC4	1.000	.778	.889	.889	.889	.875
CYC	.868	.868	.974	.974	.974	.868
A3P	.900	.818	.818	.818	.818	.857
CMP	.778	.955	1.000	1.000	1.000	.857
DMU	.900	.818	1.000	1.000	1.000	.857
PEP	.800	.923	.923	1.000	1.000	.857

Table 4.3. continued

ligand	precision	Recall				F-score
		top-1	top-3	top-5	top-10	
UMP	.923	.800	.867	.933	.933	.857
GSH	.886	.780	.780	.860	.880	.830
MPO	.875	.778	.778	.778	.778	.824
APR	.818	.818	1.000	1.000	1.000	.818
LAT	.750	.900	1.000	1.000	1.000	.818
LYS	.818	.818	.909	.909	.909	.818
MYR	.857	.783	.783	.783	.783	.818
NAD	.772	.865	.940	.945	.975	.816
PLP	.867	.765	.863	.922	.961	.813
BEN	.900	.730	.784	.811	.811	.806
CHD	.800	.800	.800	.800	.800	.800
FUL	1.000	.667	.833	.833	.833	.800
HEC	.820	.781	.953	.953	.969	.800
TYR	1.000	.667	.667	1.000	1.000	.800
IMP	.857	.750	.750	.750	.875	.800
OGA	.889	.727	.909	1.000	1.000	.800
PGA	.889	.727	.909	1.000	1.000	.800

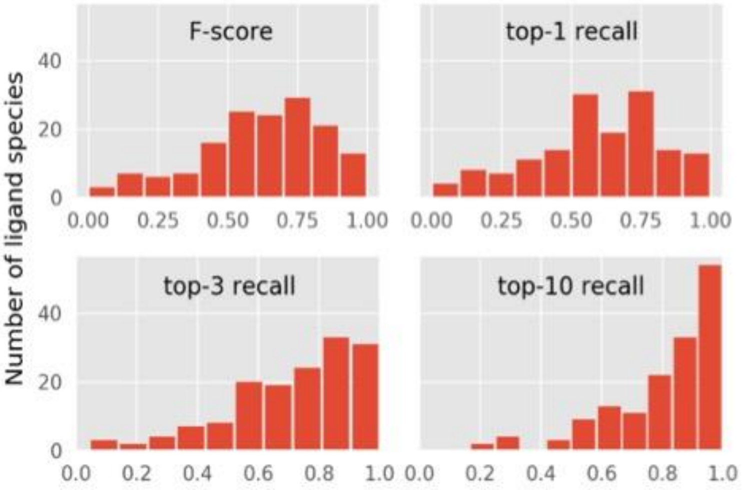


Figure 4.4. Distributions of the F-scores and top-1, 3, 10 recalls of the 151 classes.

4.4.2 Results on different features

We collected four different features for each pocket: shape, electrostatic potential, hydrophobicity, visibility. In the beginning, we thought it would give the best result using all the 4 features. We then tried to remove each one feature at a time to determine which feature contributes the most. However, after excluding the feature of electrostatic potential, trained on the same whole dataset with the same model shown in Figure 4.3, the model achieved a better result than using all 4 features. As shown in the left panel of Figure 4.5, the model trained without electrostatic potential (-E) has higher mean F-score than the one trained using all 4 features (SEHV). And we observed the same pattern for the measures mean top-1 recall (the middle panel) and top-1 accuracy (the right panel). We further tried using combinations of 2 features and using only one feature, and the combination of hydrophobicity and visibility gave the best results regarding both the measure mean F-score and the mean top-1 recall, although its top-1 accuracy is slightly lower than that of using all 4 features. We finally decided to use hydrophobicity and visibility for training other models and datasets.

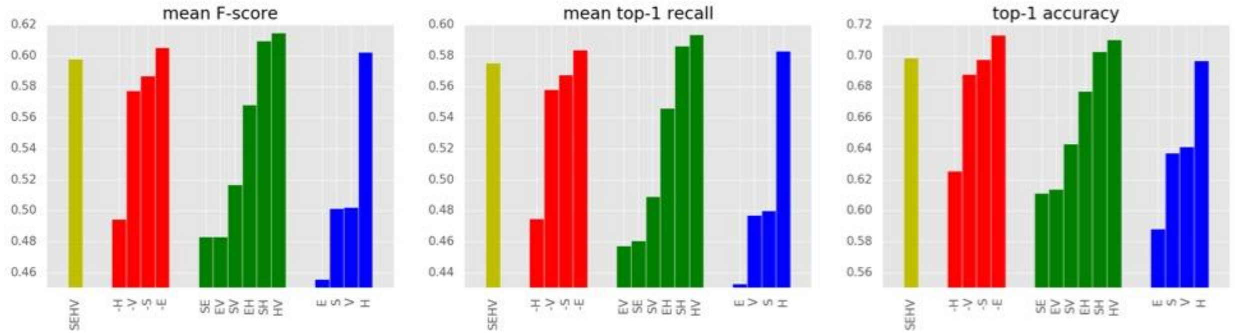


Figure 4.5. Bar plots of results on using different feature combinations. S, E, H, V represent the 4 different feature channels (S: shape; E: electrostatic potential; H: hydrophobicity; V: visibility). The minus sign '-' denotes using all the other 3 features except the feature subtracted, e.g. -H means using S, E, V but not H.

Based on the comparisons among different 3 feature combinations as well as among 1 feature trials, the importance ranking of the 4 features is hydrophobicity > visibility > shape > electrostatic potential.

The force of hydrophobicity / hydrophilicity is a major determinant for protein ligand recognition (Gao & Skolnick, 2013; Nicolau Jr et al., 2014; Scarsi et al., 1999; Snyder et al., 2011), therefore the hydrophobic-hydrophilic pattern on the surface of the protein pocket is important for ligand recognition. Here our results also suggest the critical role hydrophobicity plays in protein ligand binding.

The feature visibility is shown to be more important than the feature shape. This is reasonable. The shape channel is binary, with the grid occupied by the pocket having the value 1, and the others having the value 0. The visibility channel is constructed by mapping all the visibility values (ranging from 0 to 1) onto the grids occupied by the pocket. Therefore, the visibility channel contains most of the information represented by the shape channel, while also containing important information regarding concaveness, convexness, spatial relationship to the protein surface, etc. that the shape channel does not have. Therefore, we expect visibility brings better result than shape.

Electrostatic potential is also an important driving force in protein ligand recognition (Du et al., 2016; Koch et al., 2010). However, our results show that it contributes the least to a good result and combining it with any other feature(s) hinders the performance, as we observe in figure 5 that SE < E, EV < V, EH < H, etc. The reason could probably be that comparing to the 3 other features, electrostatic potential is much noisier. While shape is binary ({0, 1}), hydrophobicity is within -4.5 to 4.5, and visibility is within 0 to 1, the range of electrostatic potential is much broader. While most of the electrostatic potential values are within -10 to 10, some could be as negative as below -300 or as high as above 400. We tried to normalize the electrostatic potential channel, but it did not help much (data not shown), and the reason could be that the standard deviation of the raw data is very high, therefore the normalization makes most of the values very small.

The performance on using only the pocket shape information

Still from Figure 4.5, and from Table 4.4 for detailed comparison, we see that by using only the feature shape we achieved top-1 accuracy of 63.7%, mean F-score of 0.502, and mean top-1 recall of 0.480. Alt-hough these results are lower than those from using hydrophobicity + visibility, they are still good for real application. In particular, since the model only requires the shape information of the pocket, and was trained using lots of poses randomly rotated in the 3D space therefore does not entail any information like the orientation of the pocket open-ing whatsoever, it can be readily used in real-world prediction as long as we know the shape of the pocket and represent it in 3D volumetric data, and it does not require the knowledge of the sequence or the secondary/tertiary structure information of the pocket or its touching protein chain.

Table 4.4. Comparisons of the results from using only the shape feature and using the hydrophobicity + visibility features

Feature	Accuracy				Mean Recall				Mean F-score
	top-1	top-3	top-5	top-10	top-1	top-3	top-5	top-10	
Shape	.637	.786	.837	.894	.480	.621	.678	.755	.502
Hydrophobicity + Visibility	.711	.836	.877	.919	.594	.719	.766	.815	.615

4.4.3 Results on different neural network architectures

The architecture of the neural network (Figure 4.3; Table 4.1) used in this work was carefully crafted to achieve the optimal results after trying other architectures. As shown in Table 4.5, besides the original neural network shown in Table 4.1, we also tried other networks with reduced number of convolutional layers. Among them, network 1 (N1) is not a neural network but instead a multi-class logistic regression model with the linearized input layer directly connected to the softmax output layer; N2 is a fully connected neural net without any convolutional layers; N3, N4, N5 have 1, 3, 4 convolutional layers, respectively. In addition, we tried a network with the same number of convolutional layers as N0 but with a reduced number of filters for each layer (N6). We trained those 6 models with the 2-feature whole dataset (hydrophobicity + visibility) and then evaluated on the test set.

Table 4.5. Comparisons of different network architectures

Layer name	Layer size	N0	N1	N2	N3	N4	N5	N6
Input	25×25×25×#Ch	✓	✓	✓	✓	✓	✓	✓
Conv	24×24×24×64	✓			✓	✓	✓	24×24×24×32
Conv	24×24×24×96	✓			Pool 12×12×12×64	✓	✓	24×24×24×64
Conv	24×24×24×192	✓				✓	✓	24×24×24×128
Pool	12×12×12×192	✓				✓	✓	12×12×12×128
Conv	6×6×6×384	✓					✓	6×6×6×256
Pool	3×3×3×384	✓					✓	3×3×3×256
Conv	2×2×2×768	✓						2×2×2×512
Conv	1×1×1×2048	✓						1×1×1×1024
FC	1024	✓		✓	✓	✓	✓	
FC	512	✓		✓	✓	✓	✓	✓
Output	151	✓	✓	✓	✓	✓	✓	✓

Conv: convolutional; Pool: max pooling; FC: fully-connected; #Ch: number of channels; N0: the original neural network shown in table 1; N1 – N6: network 1-6; ✓denotes a network contains one specific layer

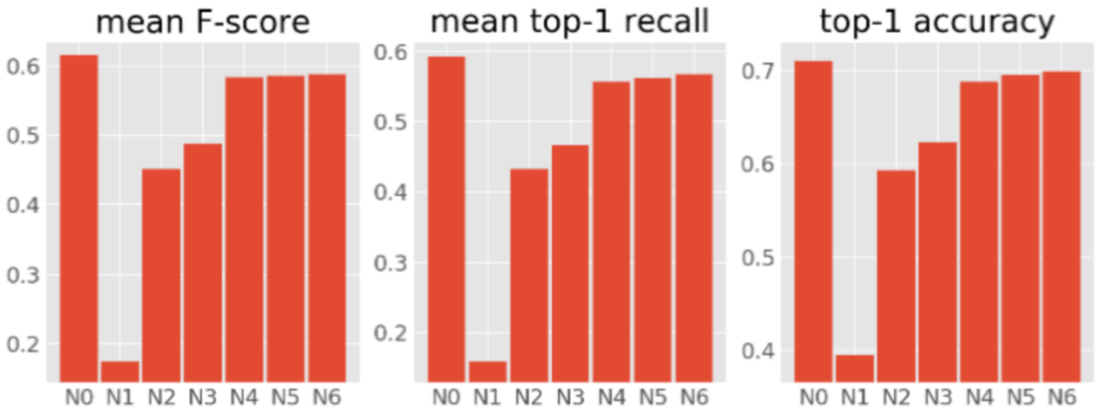


Figure 4.6. Bar plot of the results on network 0 - 6

As shown in Figure 4.6, among the 6 networks N1 – N6, N1 performs the worst. The mean F-score on N1 is < 0.2, as compared to > 0.6 on N0. The same pattern is observed in mean top-1 recall and top-1 accuracy as well. This is reasonable as the input is of very high dimension (25×25×25×2 = 31250, 2 refers to the number of channels, which in this case are hydrophobicity and visibility). With this high dimensional input, a simple classifier like logistic regression is unlikely to perform well. After neural network is applied, even a shallow 2-hidden-layer network without any convolution (N2), we see a giant leap on performance. This clearly shows the advantage of neural network over logistic regression. The performance was boosted even more by adding only one convolutional layer, as is shown in the comparison between N3 and N2, suggesting the advantage of using convolutions for image data. Yet using 3 consecutive convolutional layers, instead of only one layer, gave a much better result. Such a design was carefully chosen for the following reasons. Firstly, we have the assumption that the local small regions on the surface of a pocket is critical for it to recognize the binding ligand, and to extract hidden features of those regions requires small convolutional kernels. Using the 3 consecutive convolutional layers, we can learn features from small local regions of sizes 2×2×2, 3×3×3, and 4×4×4, respectively, therefore sufficiently covering a reasonable range. Secondly, even if regions of sizes 2×2×2 or 3×3×3 are too small to cover meaningful local regions, which we hardly expect, it is still much better to use 3 consecutive convolutional layers with the kernel size of each layer 2×2×2 rather than using only one layer of kernel size 4×4×4. One reason is the former design uses much fewer learnable parameters, therefore making the network more generalizable. With max pooling after the 3 consecutive convolutional layers, adding one more convolutional layer + max pooling still helps (N4 vs. N5), and this is designed to learn large local regions. However, adding 2 more convolutional layers on top helps more as we see the performance difference is noticeable between N5 and N0. These last 2 convolutional layers are designed to learn features on top of the hidden features learned from large local regions, thereby extracting overall global features. We also asked whether we need those amounts of parameters for each layer in N0, therefore we reduced those amounts by around half in each layer and built N6. As we can see the performance gap between N6 and N0 is still noticeable, justifying the parameter amount in N0. By the last convolutional layer in N0, the input size regarding the first 3 dimensions has already been reduced to 1×1×1, therefore we cannot go any deeper in this convolutional neural net design, and finally the selected network is N0, which gave the best performance regarding any of the 3 measures shown in Figure 4.6.

4.4.4 Results on non-redundant datasets

The PDB database contains a large portion of homologous proteins. Our dataset is collected directly from the PDB database, therefore, some subsets of the pockets in our dataset could come from protein complexes that share high sequence and / or structure similarity. For a pair of proteins with considerable sequence and / or structure similarity, even if they bind the same ligand species, the binding pockets could still be quite different. Yet it is undeniable that correlation between protein global similarity and pocket similarity is significantly high. From a ma-chine learning point of view, it is not problematic that the input data share similarities. On the contrary, it is based on the similarities of the input data that the machine learning tools could then extract useful hidden features from the data for future prediction. Our dataset is built on the whole PDB database, therefore shares similar statistical distribution with the PDB database. For a future prediction, the prediction result our method could give will also follow the data distribution of the PDB database. Since newly solved structures usually bias toward existing homologous models, it is therefore reasonable for our method to give predictions following the same pattern.

However, we would still hope to build a model that could largely re-duce this bias. We therefore culled our data to remove potential sequence / structure similarities and collected 4 non-redundant datasets: SI-0.5, SI-0.8, TM-0.6, TM-0.8. We used TM-align (Y. Zhang & Skolnick,

2005) to determine the sequence similarity (Sequence Identity (SI) score) and structure similarity (TM score). SI-0.5 means for all the pairs of pockets in the dataset, their touching protein chains do not share sequence identity score of over 0.5. The same rule applies to SI-0.8, TM-0.6, and TM-0.8, where TM represents TM score.

This data filtering procedure substantially reduced the total amount of data and left the 4 non-redundant datasets with only a fraction of data compared to our original full dataset. Deep learning generally requires large amounts of data to achieve better performance. Indeed, the test results on the 4 non-redundant datasets plummeted compared to the results on the full dataset regarding the top-1 accuracy and the mean top-1 recall (Table 4.6). However, when the criterion is loosened to top-25 accuracy or mean top-25 recall, the performance gap be-tween the whole dataset and all the non-redundant datasets is much smaller. Overall, the performance on the nonredundant sets is still rea-sonably good and could still be used for future prediction.

In addition, to show that the training and test subsets were indeed collected by randomly splitting the nonredundant datasets, we constructed 2 follow-up datasets for each of the 4 nonredundant datasets, and the follow-up datasets indeed gave similar results. (See Supplemental Table 4.2 for details).

Table 4.6. Comparisons of the results between the 4 non-redundant datasets and the full dataset

Dataset	**#examples**	**Accuracy**				**Mean Recall**				**Mean F-score**
		top-1	**top-5**	**top-15**	**top-25**	**top-1**	**top-5**	**top-15**	**top-25**	
TM-0.6	10054	.137	.344	.532	.591	.090	.215	.402	.476	.088
TM-0.8	14335	.184	.448	.666	.767	.120	.260	.465	.576	.121
SI-0.5	16026	.219	.504	.728	.821	.110	.290	.502	.622	.109
SI-0.8	18895	.257	.553	.741	.824	.166	.357	.535	.622	.168
whole	77087	.711	.877	.939	.960	.594	.766	.847	.885	.615

TM-0.6, TM-0.8: datasets created based on TM score cut-off 0.6, 0.8, respectively; SI-0.5, SI-0.8: datasets created based on Sequence Identity score cut-off 0.5, 0.8, respectively; whole: the whole dataset without any redundancy removal

4.4.5 Comparisons with existing methods

Comparison with Patch-Surfer2.0

Patch-Surfer2.0 (PS2.0) (X. Zhu et al., 2015) is a method that predicts the binding ligand for a pocket among 117 candidate ligand species. It represents a protein pocket as a set of small pocket surface local patches characterized by 3D Zernike descriptors, and then predicts the binding ligand given a pocket by comparing its patch representations against a pre-selected pocket database that contains around 3200 pockets.

Since our method ProLig covers 151 ligand species, and the overlap between ProLig and PS2.0 includes 96 different ligand species, to make a fair comparison, we selected a subset of our test set where each pocket binds one of the 96 overlapping ligands. This procedure resulted in a comparison set of 7107 pockets and we predicted their binding ligands using PS2.0. For each pocket in the comparison set, PS2.0 gave a ranking of predicted ligands among 117 ligand species. Since there are only 96 overlapping ligands, for a prediction that is not within the 96 ligands, we simply removed the prediction and moved the following predictions forward in the ranking. We also recalculated our results on the test set to only include the 7107 pockets of the comparison set and their predictions among only the 96 overlapping ligands. In these ways we make the comparison as fair as possible. The results are shown in Table 4.7. Our method ProLig performs considerably better than PS2.0 regarding all the 3 measures. The top-1 accuracy nearly doubled in ProLig, while the mean top-1 and the mean F-score are 5 times more.

Table 4.7. Comparisons with Patch-Surfer2.0 (PS2.0) on our test set

Method	**Accuracy**				**Mean Recall**				**Mean F-score**
	top-1	**top-3**	**top-5**	**top-10**	**top-1**	**top-3**	**top-5**	**top-10**	
PS2.0	.383	.518	.614	.755	.100	. 235	.331	.520	.087
ProLig	.726	.849	.889	.932	.598	.717	.761	.825	.625

Comparison with Apoc

Apoc (Gao & Skolnick, 2013) is an alignment-based pocket comparison method that uses the pocket similarity score PS-score, which ranges from 0 to 1. A large PS-score indicates high similarity between the 2 compared pockets. To compare with Apoc, we did pairwise comparison in our test set (containing 7782 pockets) using Apoc, and then evaluated the performance of Apoc using different pocket selection cutoffs as well as different voting methods. In our dataset, we selected pockets based on the cutoff of 4 Å, i.e. the pocket is within 4 Å of the binding ligand. Here in this comparison, we used 3 different cutoffs (4, 5, 6 Å) to eliminate the potential bias caused by pocket extraction cutoff. We used 2 voting strategies: average and k nearest neighbor (kNN). In average voting, the predicted ligand ranking for a given pocket is based on the ranking of the average PS-scores, with each score calculated by averaging the PS-scores of template pockets that bind one same ligand type. In 1NN voting, the ranking is given by the ranking of the template pockets that have the top-k PS-scores. Comparing with Apoc, our meth-od ProLig performs significantly better regarding any of the 3 measures and regarding any of the 3 pocket extraction cutoffs as well as any of the 2 voting strategies (Figure 4.7).

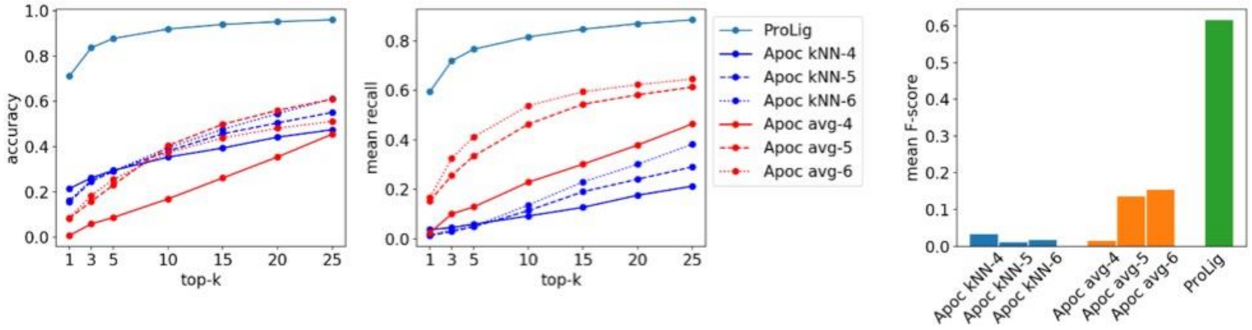


Figure 4.7. Comparisons with Apoc. kNN: k nearest neighbor voting strategy; avg: average voting strategy; kNN-4, kNN-5, kNN6: kNN applied on datasets of pockets extracted within 4 Å, 5 Å, 6 Å, respectively; avg-4, avg-5, avg-6: Avg applied on datasets of pockets extracted within 4 Å, 5 Å, 6 Å, respectively.

4.5 Discussion and conclusion

Proteins are 3D structures. Representing proteins as 3D images can capture important spatial information that 2D or 1D data are unable to represent. Here by representing protein pockets in their original 3D space, and by employing the convolutional neural networks that have shown great performance on image classification, we developed this computational framework ProLig that can predict the binding ligand among over 150 ligand species. Even with only the shape information of the pocket, this framework can still work very well. ProLig demonstrates the edge of employing 3D shape information of biomolecular data, providing a good reference to the research in many other fields of structural biology.

4.6 Supplemental Data

The follow-up data splitting procedure: to further show that our data splitting procedure is indeed random while still taking into consideration that training a model takes a long time, for a split dataset (containing the training set and the test set) we constructed 2 other follow-up datasets using the following procedure: randomly split the training set into 9 folds, each time pick one fold as the test set and combine the rest 8 folds + the original test set as the new training set. We repeated the procedure twice, each time picking a different fold, thus resulting in 2 follow-up datasets.

We applied the follow-up data splitting procedure to both the whole dataset and the 4 nonredundant datasets, resulting in 10 follow-up datasets in total. We then trained those follow-up datasets with the neural network shown in Table 4.1.

Supplemental Table 4.1. Overall results on the whole dataset and the 2 corresponding follow-up datasets trained with the neural network in Table 4.1

	Accuracy				**Mean Recall**				**Mean F-score**
	top-1	**top-3**	**top-5**	**top-10**	**top-1**	**top-3**	**top-5**	**top-10**	
the original whole dataset	.711	.836	.877	.919	.594	.719	.766	.815	.615
follow-up dataset 1	.712	.835	.878	.924	.589	.729	.776	.841	.605
follow-up dataset 2	.712	.837	.879	.926	.589	.717	.762	.834	.607
average	.712	.836	.878	.923	.591	.722	.768	.830	.609
standard deviation	.001	.001	.001	.003	.002	.005	.006	.011	.004

average: the average result across the original whole dataset, follow-up dataset 1, and follow-up dataset 2 results; standard deviation: the standard deviation across the original whole dataset, follow-up dataset 1, and follow-up dataset 2 results

Supplemental Table 4.2. Overall results on the 4 nonredundant datasets and their corresponding follow-up datasets trained with the neural network in Table 4.1

Dataset	Split	Accuracy				Mean Recall				Mean F-score
		top-1	top-5	top-15	top-25	top-1	top-5	top-15	top-25	
TM-0.6	original	.137	.344	.532	.591	.090	.215	.402	.476	.088
	follow-up 1	.134	.403	.637	.733	.075	.192	.429	.530	.073
	follow-up 2	.141	.397	.630	.736	.090	.209	.382	.476	.086
	average	.138	.381	.600	.686	.085	.205	.404	.494	.082
	SD	.003	.027	.048	.068	.007	.010	.019	.025	.007
TM-0.8	original	.184	.448	.666	.767	.120	.260	.465	.576	.121
	follow-up 1	.189	.481	.713	.810	.102	.249	.437	.562	.096
	follow-up 2	.190	.464	.690	.801	.111	.259	.446	.569	.105
	average	.188	.464	.689	.793	.111	.256	.449	.569	.107
	SD	.003	.014	.019	.018	.007	.005	.012	.006	.010
SI-0.5	original	.219	.504	.728	.821	.110	.290	.502	.622	.109
	follow-up 1	.224	.508	.731	.824	.118	.271	.493	.611	.119
	follow-up 2	.215	.473	.693	.787	.115	.261	.453	.559	.113
	average	.219	.495	.718	.811	.114	.274	.483	.597	.114
	SD	.004	.016	.017	.017	.004	.012	.021	.028	.004
SI-0.8	original	.257	.553	.741	.824	.166	.357	.535	.622	.168
	follow-up 1	.282	.559	.759	.835	.167	.347	.561	.662	.166
	follow-up 2	.250	.564	.756	.837	.165	.363	.521	.635	.173
	average	.263	.558	.752	.832	.166	.355	.539	.640	.169
	SD	.014	.004	.008	.006	.001	.007	.016	.016	.003

TM-0.6, TM-0.8: datasets created based on TM score cut-off 0.6, 0.8, respectively;
SI-0.5, SI-0.8: datasets created based on Sequence Identity score cut-off 0.5, 0.8, respectively;
Split: how the dataset was split; original: the dataset was split as described in 2.2.2.1; follow-up 1 & 2: the dataset was split using the follow-up data splitting procedure as described above; average: the average result across the original, follow-up 1, and follow-up 2 results; SD: the standard deviation of the original, follow-up 1, and follow-up 2 results

CHAPTER 5. CONCLUSIONS

Data Science is an exciting and rapidly growing field. With the rapid growth in the amount of data across a variety of disciplines, people are striving to make those data meaningful by deploying data science. The same is happening in biological research.

In this work, by harnessing the growing data in several fields, we explored various edges of the immense protein universe using a range of latest data science techniques.

Firstly, we noticed the emerging field of micropeptides and realized the need for bioinformatics methods to predict them from DNA sequences. We searched the field and found the available data is sufficient for building new methods. We therefore carefully collected and cleaned data from several existing databases to build a high-quality dataset. Based on the fact that the dataset is not large, and that the problem is binary classification, which is not hard to address, we carefully selected a well-studied machine learning algorithm – logistic regression – tailored for this dataset and the problem for optimal results. We trained our model using standard machine learning procedures and validated our model using several blind test sets. Our developed method – MiPepid, performed well on all test sets and also when compared to other methods. Through developing MiPepid, we demonstrated the whole pipeline of developing a new method: define the question, collect and clean data, choose a suitable algorithm, train the model, validate the model, compare with other methods, and finally wrap the model into a standalone package so others can use it with no hassle. Also, by using logistic regression, probably the most common ML algorithm, and obtaining a good performance we showed that when selecting an algorithm, there is no need to blindly pursue “fancy algorithms”, and the key is to find one that suits the data and the problem.

Secondly, when studying micropeptides, we learned that a number of lncRNAs are translated to functional micropeptides. We realized the significance of the open question - are most lncRNAs translated – and we were thinking how to approach this question from a data-driven perspective. We therefore thought of genetic variation analysis, as it has been used to study natural selection for inferring functional relevance. And more importantly, we knew that a large amount of genetic variation data is available, which is essential in making the study possible. Thus, we collected the data in a systematic way and analyzed them using rigorous statistical procedures. Our results showed the similarities between lncRNAs and proteins from various different angles, which were not considered before by this research field. From this study, we demonstrated the possibilities of digging into existing large data reservoirs and extracting insights from a different angle. This work does not look as exciting as building machine learning models. All it relied on were just traditional statistical analyses and a large dataset. Yet data analysis is always the starting point for many data science questions.

Thirdly, we borrowed the ideas from the latest deep learning techniques and applied to the age-old field of protein-ligand binding. We noticed the similarities between 2D images and protein 3D structures, and by rendering the structures into 3D images we also converted the question to a 3D image classification problem. Also, as is always the case, we knew there were enough available data to make our method possible. We designed our 3D deep convolutional neural network from scratch to suit our special dataset. We trained the model using standard deep learning procedures and validated the model on the holdout test set. The model performed well and also achieved better results than other methods. Through this study we showed the possibility of incorporating the latest data science techniques into biological research by treating the questions at hand from a different angle.

The field of data science will continue to grow and so are the biological data. More and more biological questions will be addressed from a data-driven perspective. The biology + data science is an exciting field. Here, with five years' efforts, we addressed several biological questions from different angles of data science. We hope our work could help advance the progress in corresponding specific research fields and also provide examples on deploying data science into biological research.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). Tensorflow: A system for large-scale machine learning. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283.

Anderson, D M, Anderson, K. M., Chang, C.-L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., Kasaragod, P., Shelton, J. M., Liou, J., Bassel-Duby, R., & Olson, E. N. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, *160*(4), 595–606. https://doi.org/10.1016/j.cell.2015.01.009

Anderson, Douglas M, Makarewich, C. A., Anderson, K. M., Shelton, J. M., Bezprozvannaya, S., Bassel-Duby, R., & Olson, E. N. (2016). Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Science Signaling*, *9*(457), ra119 LP-ra119. http://stke.sciencemag.org/content/9/457/ra119.abstract

Asad, S., Fang, Y., Wycoff, K. L., & Hirsch, A. M. (1994). Isolation and characterization of cDNA and genomic clones of MsENOD40; transcripts are detected in meristematic cells of alfalfa. *Protoplasma*, *183*(1), 10–23. https://doi.org/10.1007/BF01276808

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., & Couso, J.-P. (2014). Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *ELife*, *3*, e03528–e03528. https://doi.org/10.7554/eLife.03528

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(suppl_2), W202–W208. http://dx.doi.org/10.1093/nar/gkp335

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, *98*(18), 10037 LP – 10041. https://doi.org/10.1073/pnas.181342398

Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W. E. J., Kundaje, A., Gunawardena, H. P., Yu, Y., Xie, L., Krajewski, K., Strahl, B. D., Chen, X., Bickel, P., Giddings, M. C., Brown, J. B., & Lipovich, L. (2012). Long noncoding RNAs are rarely translated in two human cell lines. *Genome Research*, *22*(9), 1646–1657. https://doi.org/10.1101/gr.134767.111

Bazin, J., Baerenfaller, K., Gosai, S. J., Gregory, B. D., Crespi, M., & Bailey-Serres, J. (2017). Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proceedings of the National Academy of Sciences*, *114*(46), E10018 LP-E10027. https://doi.org/10.1073/pnas.1708433114

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., & Giraldez, A. J. (2014a). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal*, *33*(9), 981 LP – 993. http://emboj.embopress.org/content/33/9/981.abstract

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., & Giraldez, A. J. (2014b). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal*, *33*(9), 981 LP – 993. http://emboj.embopress.org/content/33/9/981.abstract

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bhartiya, D., Jalali, S., Ghosh, S., & Scaria, V. (2014). Distinct patterns of genetic variations in potential functional elements in long noncoding RNAs. *Human Mutation*, *35*(2), 192–201. https://doi.org/10.1002/humu.22472

Bi, P., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J. R., Shelton, J. M., Sánchez-Ortiz, E., Bassel-Duby, R., & Olson, E. N. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. *Science*, *356*(6335), 323 LP – 327. http://science.sciencemag.org/content/356/6335/323.abstract

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. https://books.google.com/books?id=qWPwnQEACAAJ

Blanvillain, R., Young, B., Cai, Y., Hecht, V., Varoquaux, F., Delorme, V., Lancelin, J.-M., Delseny, M., & Gallois, P. (2011). The Arabidopsis peptide kiss of death is an inducer of programmed cell death. *The EMBO Journal*, *30*(6), 1173–1183. https://doi.org/10.1038/emboj.2011.14

Brylinski, M., & Skolnick, J. (2009). FINDSITELHM: a threading-based approach to ligand homology modeling. *PLoS Computational Biology*, *5*(6).

Burkholder, W. F., Kurtser, I., & Grossman, A. D. (2001). Replication Initiation Proteins Regulate a Developmental Checkpoint in Bacillus subtilis. *Cell*, *104*(2), 269–279. https://doi.org/https://doi.org/10.1016/S0092-8674(01)00211-2

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., … Zardecki, C. (2018). RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, *47*(D1), D464–D474. https://doi.org/10.1093/nar/gky1004

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* , *25*(18), 1915–1927. https://doi.org/10.1101/gad.17446611

Cai, B., Li, Z., Ma, M., Wang, Z., Han, P., Abdalla, B. A., Nie, Q., & Zhang, X. (2017). LncRNA-Six1 Encodes a Micropeptide to Activate Six1 in Cis and Is Involved in Cell Proliferation and Muscle Growth. *Frontiers in Physiology*, *8*, 230. https://doi.org/10.3389/fphys.2017.00230

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., & Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, *13*(2), 165.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., … Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, *309*(5740), 1559–1563. https://doi.org/10.1126/science.1112014

Casson, S. A., Chilley, P. M., Topping, J. F., Evans, I. M., Souter, M. A., & Lindsey, K. (2002). The POLARIS Gene of Arabidopsis Encodes a Predicted Peptide Required for Correct Root Growth and Leaf Vascular Patterning. *The Plant Cell*, *14*(8), 1705 LP – 1721. https://doi.org/10.1105/tpc.002618

Chan, B. Y., & Kibler, D. (2005). Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC Bioinformatics*, *6*(1), 262. https://doi.org/10.1186/1471-2105-6-262

Chen, L.-L. (2016). Linking Long Noncoding RNA Localization and Function. *Trends in Biochemical Sciences*, *41*(9), 761–772. https://doi.org/https://doi.org/10.1016/j.tibs.2016.07.003

Chew, G.-L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F., & Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development*, *140*(13), 2828–2834. https://doi.org/10.1242/dev.098343

Chikhi, R., Sael, L., & Kihara, D. (2010). Real-time ligand binding pocket database search using local surface descriptors. *Proteins: Structure, Function, and Bioinformatics*, *78*(9), 2007–2028.

Chilley, P. M., Casson, S. A., Tarkowski, P., Hawkins, N., Wang, K. L.-C., Hussey, P. J., Beale, M., Ecker, J. R., Sandberg, G. K., & Lindsey, K. (2006). The POLARIS peptide of Arabidopsis regulates auxin transport and root growth via effects on ethylene signaling. *The Plant Cell*, *18*(11), 3058–3072. https://doi.org/10.1105/tpc.106.040790

Chu, D., & Wei, L. (2019). Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC Cancer*, *19*(1), 359. https://doi.org/10.1186/s12885-019-5572-x

Chugunova, A., Navalayeu, T., Dontsova, O., & Sergiev, P. (2018). Mining for Small Translated ORFs. *Journal of Proteome Research*, *17*(1), 1–11. https://doi.org/10.1021/acs.jproteome.7b00707

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3642–3649.

Cohen, S. M. (2014). Everything old is new again: (linc)RNAs make proteins! *The EMBO Journal*, *33*(9), 937 LP – 938. http://emboj.embopress.org/content/33/9/937.abstract

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, *8*(12), 1229–1231. https://doi.org/10.1101/gr.8.12.1229

Couso, J.-P., & Patraquim, P. (2017). Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology*, *18*, 575. http://dx.doi.org/10.1038/nrm.2017.58

Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., & Menschaert, G. (2013). Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics*, *14*(1), 648. https://doi.org/10.1186/1471-2164-14-648

Crespi, M. D., Jurkevitch, E., Poiret, M., d'Aubenton-Carafa, Y., Petrovics, G., Kondorosi, E., & Kondorosi, A. (1994). enod40, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth. *The EMBO Journal*, *13*(21), 5099–5112.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., … Flicek, P. (2018). Ensembl 2019. *Nucleic Acids Research*, *47*(D1), D745–D751. https://doi.org/10.1093/nar/gky1113

D'Lima, N. G., Ma, J., Winkler, L., Chu, Q., Loh, K. H., Corpuz, E. O., Budnik, B. A., Lykke-Andersen, J., Saghatelian, A., & Slavoff, S. A. (2017). A human microprotein that interacts with the mRNA decapping complex. *Nature Chemical Biology*, *13*(2), 174–180. https://doi.org/10.1038/nchembio.2249

De Silva, D. R., Nichols, R., & Elgar, G. (2014). Purifying Selection in Deeply Conserved Human Enhancers Is More Consistent than in Coding Sequences. *PLOS ONE*, *9*(7), e103357. https://doi.org/10.1371/journal.pone.0103357

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* , *22*(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

Dhar, V. (2013). Data Science and Prediction. *Commun. ACM*, *56*(12), 64–73. https://doi.org/10.1145/2500499

Dong, X., Wang, D., Liu, P., Li, C., Zhao, Q., Zhu, D., & Yu, J. (2013). Zm908p11, encoded by a short open reading frame (sORF) gene, functions in pollen tube growth as a profilin ligand in maize. *Journal of Experimental Botany*, *64*(8), 2359–2372. https://doi.org/10.1093/jxb/ert093

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L., & Liu, S.-Q. (2016). Insights into protein–ligand interactions: mechanisms, models, and methods. *International Journal of Molecular Sciences*, *17*(2), 144.

Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., … Institute, T. T. G. R. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Eddy, S. R. (2005). A Model of the Statistical Power of Comparative Genome Sequence Analysis. *PLOS Biology*, *3*(1), e10. https://doi.org/10.1371/journal.pbio.0030010

Farrell, C. M., O'Leary, N. A., Harte, R. A., Loveland, J. E., Wilming, L. G., Wallin, C., Diekhans, M., Barrell, D., Searle, S. M. J., Aken, B., Hiatt, S. M., Frankish, A., Suner, M.-M., Rajput, B., Steward, C. A., Brown, G. R., Bennett, R., Murphy, M., Wu, W., … Pruitt, K. D. (2014). Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Research*, *42*(Database issue), D865-72. https://doi.org/10.1093/nar/gkt1059

Fields, A. P., Rodriguez, E. H., Jovanovic, M., Stern-Ginossar, N., Haas, B. J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S. A., Ingolia, N. T., Regev, A., & Weissman, J. S. (2015). A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Molecular Cell*, *60*(5), 816–827. https://doi.org/https://doi.org/10.1016/j.molcel.2015.11.013

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., … Searle, S. M. J. (2013). Ensembl 2014. *Nucleic Acids Research*, *42*(D1), D749–D755. https://doi.org/10.1093/nar/gkt1196

Frank, M. J., & Smith, L. G. (2002). A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells. *Current Biology : CB*, *12*(10), 849–853. https://doi.org/10.1016/s0960-9822(02)00819-9

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., & Couso, J. P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biology*, *5*(5), e106. https://doi.org/10.1371/journal.pbio.0050106

Gao, M., & Skolnick, J. (2013). APoc: large-scale identification of similar protein pockets. *Bioinformatics*, *29*(5), 597–604.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P. K.-H., Tsui, L.-C., Waye, M. M. Y., Wong, J. T.-F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., … Group, M. (2003). The International HapMap Project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323.

Guo, B., Zhai, D., Cabezas, E., Welsh, K., Nouraini, S., Satterthwait, A. C., & Reed, J. C. (2003). Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature*, *423*(6938), 456–461. https://doi.org/10.1038/nature01627

Guo, P., Yoshimura, A., Ishikawa, N., Yamaguchi, T., Guo, Y., & Tsukaya, H. (2015). Comparative analysis of the RTFL peptide family on the control of plant organogenesis. *Journal of Plant Research*, *128*(3), 497–510. https://doi.org/10.1007/s10265-015-0703-1

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., & Lander, E. S. (2013). Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell*, *154*(1), 240–251. https://doi.org/https://doi.org/10.1016/j.cell.2013.06.009

Hamilton, M. (2011). *Population genetics*. John Wiley & Sons.

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., & Shiu, S.-H. (2010). sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics (Oxford, England)*, *26*(3), 399–400. https://doi.org/10.1093/bioinformatics/btp688

Handler, A. A., Lim, J. E., & Losick, R. (2008). Peptide inhibitor of cytokinesis during sporulation in Bacillus subtilis. *Molecular Microbiology*, *68*(3), 588–599. https://doi.org/doi:10.1111/j.1365-2958.2008.06173.x

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., Zhang, B., Zhang, D., Qin, Y., Yang, F., & Chen, R. (2018). SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Briefings in Bioinformatics*, *19*(4), 636–643. http://dx.doi.org/10.1093/bib/bbx005

Harte, R. A., Farrell, C. M., Loveland, J. E., Suner, M.-M., Wilming, L., Aken, B., Barrell, D., Frankish, A., Wallin, C., Searle, S., Diekhans, M., Harrow, J., & Pruitt, K. D. (2012). Tracking and coordinating an international curation effort for the CCDS Project. *Database : The Journal of Biological Databases and Curation*, *2012*, bas008. https://doi.org/10.1093/database/bas008

Hartford, C. C. R., & Lal, A. (2020). When Long Noncoding Becomes Protein Coding. *Molecular and Cellular Biology*, *40*(6). https://doi.org/10.1128/MCB.00528-19

Hashimoto, Y., Niikura, T., Tajima, H., Yasukawa, T., Sudo, H., Ito, Y., Kita, Y., Kawasumi, M., Kouyama, K., Doyu, M., Sobue, G., Koide, T., Tsuji, S., Lang, J., Kurokawa, K., & Nishimoto, I. (2001). A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer&#039;s disease genes and Aβ. *Proceedings of the National Academy of Sciences*, *98*(11), 6336 LP – 6341. https://doi.org/10.1073/pnas.101133498

Heo, L., Shin, W.-H., Lee, M. S., & Seok, C. (2014). GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Research*, *42*(W1), W210–W214.

Hoffmann, B., Zaslavskiy, M., Vert, J.-P., & Stoven, V. (2010). A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics*, *11*(1), 99.

Huang, J.-Z., Chen, M., Chen, D., Gao, X.-C., Zhu, S., Huang, H., Hu, M., Zhu, H., & Yan, G.-R. (2017). A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Molecular Cell*, *68*(1), 171-184.e6. https://doi.org/https://doi.org/10.1016/j.molcel.2017.09.015

Ingolia, N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics*, *15*, 205. https://doi.org/10.1038/nrg3645

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. *Cell*, *165*(1), 22–33. https://doi.org/https://doi.org/10.1016/j.cell.2016.02.066

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., Wills, M. R., & Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports*, *8*(5), 1365–1379. https://doi.org/10.1016/j.celrep.2014.07.045

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, *324*(5924), 218 LP – 223. http://science.sciencemag.org/content/324/5924/218.abstract

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011a). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802. https://doi.org/10.1016/j.cell.2011.10.002

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011b). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802. https://doi.org/10.1016/j.cell.2011.10.002

Jha, P., Lu, D., & Xu, S. (2015). Natural Selection and Functional Potentials of Human Noncoding Elements Revealed by Analysis of Next Generation Sequencing Data. *PLOS ONE*, *10*(6), e0129023. https://doi.org/10.1371/journal.pone.0129023

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5’UTRs, and pseudogenes are translated and some are likely to express functional proteins. *ELife*, *4*, e08890. https://doi.org/10.7554/eLife.08890

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*(1), 192. https://doi.org/10.1186/1471-2105-9-192

Johnson, J. M., Edwards, S., Shoemaker, D., & Schadt, E. E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics*, *21*(2), 93–102. https://doi.org/https://doi.org/10.1016/j.tig.2004.12.009

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., & Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, *45*(W1), W12–W16. http://dx.doi.org/10.1093/nar/gkx428

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., … Gingeras, T. R. (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, *316*(5830), 1484 LP – 1488. https://doi.org/10.1126/science.1138341

Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., … Gerstein, M. (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science*, *342*(6154), 1235587. https://doi.org/10.1126/science.1235587

Kikuchi, K., Fukuda, M., Ito, T., Inoue, M., Yokoi, T., Chiku, S., Mitsuyama, T., Asai, K., Hirose, T., & Aizawa, Y. (2009). Transcripts of unknown function in multiple-signaling pathways involved in human stem cell differentiation. *Nucleic Acids Research*, *37*(15), 4987–5000. https://doi.org/10.1093/nar/gkp426

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Koch, M., Chitayat, S., Dattilo, B. M., Schiefner, A., Diez, J., Chazin, W. J., & Fritz, G. (2010). Structural basis for ligand recognition and activation of RAGE. *Structure*, *18*(10), 1342–1352.

Kondo, T, Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., & Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science (New York, N.Y.)*, *329*(5989), 336–339. https://doi.org/10.1126/science.1188158

Kondo, Takefumi, Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., & Kageyama, Y. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology*, *9*(6), 660–665. https://doi.org/10.1038/ncb1595

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, *35*(Web Server issue), W345-9. https://doi.org/10.1093/nar/gkm391

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, C., Zeng, J., Drew, B. G., Sallam, T., Martin-Montalvo, A., Wan, J., Kim, S.-J., Mehta, H., Hevener, A. L., de Cabo, R., & Cohen, P. (2015). The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance. *Cell Metabolism*, *21*(3), 443–454. https://doi.org/https://doi.org/10.1016/j.cmet.2015.02.009

Li, B., Turuvekere, S., Agrawal, M., La, D., Ramani, K., & Kihara, D. (2008). Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins: Structure, Function, and Bioinformatics*, *71*(2), 670–683.

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, *27*(13), i275–i282. http://dx.doi.org/10.1093/bioinformatics/btr209

Louie, E., Ott, J., & Majewski, J. (2003). Nucleotide Frequency Variation Across Human Genes. *Genome Research* , *13*(12), 2594–2601. https://doi.org/10.1101/gr.1317703

Ma, Jiao, Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., Budnik, B. A., Kellis, M., & Saghatelian, A. (2014). Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *Journal of Proteome Research*, *13*(3), 1757–1765. https://doi.org/10.1021/pr401280w

Ma, Jinxia, Yan, B., Qu, Y., Qin, F., Yang, Y., Hao, X., Yu, J., Zhao, Q., Zhu, D., & Ao, G. (2008). Zm401, a short-open reading-frame mRNA or noncoding RNA, is essential for tapetum and microspore development and can regulate the floret formation in maize. *Journal of Cellular Biochemistry*, *105*(1), 136–146. https://doi.org/doi:10.1002/jcb.21807

Ma, L., Bajic, V. B., & Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biology*, *10*(6), 924–933. https://doi.org/10.4161/rna.24604

Ma, L., Cao, J., Liu, L., Du, Q., Li, Z., Zou, D., Bajic, V. B., & Zhang, Z. (2019). LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Research*, *47*(D1), D128–D134. https://doi.org/10.1093/nar/gky960

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., & Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, *16*(1), 179. https://doi.org/10.1186/s13059-015-0742-x

Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., & Couso, J. P. (2013a). Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science*, *341*(6150), 1116 LP – 1120. http://science.sciencemag.org/content/341/6150/1116.abstract

Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., & Couso, J. P. (2013b). Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science*, *341*(6150), 1116 LP – 1120. http://science.sciencemag.org/content/341/6150/1116.abstract

Makarewich, C. A., Baskin, K. K., Munir, A. Z., Bezprozvannaya, S., Sharma, G., Khemtong, C., Shah, A. M., McAnally, J. R., Malloy, C. R., Szweda, L. I., Bassel-Duby, R., & Olson, E. N. (2018). MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. *Cell Reports*, *23*(13), 3701–3709. https://doi.org/https://doi.org/10.1016/j.celrep.2018.05.058

Makarewich, C. A., & Olson, E. N. (2017). Mining for Micropeptides. *Trends in Cell Biology*, *27*(9), 685–696. https://doi.org/https://doi.org/10.1016/j.tcb.2017.04.006

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K. I., Clohessy, J. G., & Pandolfi, P. P. (2016). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, *541*, 228. http://dx.doi.org/10.1038/nature21034

McVean, G. A., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Geneva, U. of. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Mercer, T. R., Dinger, M. E., & Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, *10*(3), 155–159. https://doi.org/10.1038/nrg2521

Morris, R. J., Najmanovich, R. J., Kahraman, A., & Thornton, J. M. (2005). Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, *21*(10), 2347–2355.

Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K., & Gerstein, M. B. (2011). Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Research*, *39*(16), 7058–7076. https://doi.org/10.1093/nar/gkr342

Mudge, J. M., & Harrow, J. (2016). The state of play in higher eukaryote gene annotation. *Nature Reviews Genetics*, *17*, 758. https://doi.org/10.1038/nrg.2016.119

Najmanovich, R., Kurbatova, N., & Thornton, J. (2008). Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, *24*(16), i105–i111.

Narita, N. N., Moore, S., Horiguchi, G., Kubo, M., Demura, T., Fukuda, H., Goodrich, J., & Tsukaya, H. (2004). Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in Arabidopsis thaliana. *The Plant Journal : For Cell and Molecular Biology*, *38*(4), 699–713. https://doi.org/10.1111/j.1365-313X.2004.02078.x

Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., Reese, A. L., McAnally, J. R., Chen, X., Kavalali, E. T., Cannon, S. C., Houser, S. R., Bassel-Duby, R., & Olson, E. N. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, *351*(6270), 271 LP – 275. http://science.sciencemag.org/content/351/6270/271.abstract

Nicolau Jr, D. V, Paszek, E., Fulga, F., & Nicolau, D. V. (2014). Mapping hydrophobicity on the protein molecular surface at atom-level resolution. *PloS One*, *9*(12).

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schönbach, C., Gojobori, T., Baldarelli, R., … management:, S. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, *420*(6915), 563–573. https://doi.org/10.1038/nature01266

Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., & Menschaert, G. (2016). sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research*, *44*(D1), D324–D329. http://dx.doi.org/10.1093/nar/gkv1175

Olexiouk, V., Van Criekinge, W., & Menschaert, G. (2018). An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research*, *46*(D1), D497–D502. https://doi.org/10.1093/nar/gkx1130

Pauli, A., Norris, M. L., Valen, E., Chew, G.-L., Gagnon, J. A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., Tsai, S. Q., Joung, J. K., Saghatelian, A., & Schier, A. F. (2014). Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science*, *343*(6172). https://doi.org/10.1126/science.1248636

Plaza, S., Menschaert, G., & Payre, F. (2017). In Search of Lost Small Peptides. *Annual Review of Cell and Developmental Biology*, *33*(1), 391–416. https://doi.org/10.1146/annurev-cellbio-100616-060516

Polycarpou-Schwarz, M., Groß, M., Mestdagh, P., Schott, J., Grund, S. E., Hildenbrand, C., Rom, J., Aulmann, S., Sinn, H.-P., Vandesompele, J., & Diederichs, S. (2018). The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene*, *37*(34), 4750–4768. https://doi.org/10.1038/s41388-018-0281-5

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., Hart, E., Suner, M.-M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., … Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, *19*(7), 1316–1323. https://doi.org/10.1101/gr.080531.108

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *ELife*, *5*. https://doi.org/10.7554/eLife.13328

Rees, D. C., Williams, T. N., & Gladwin, M. T. (2010). Sickle-cell disease. *The Lancet*, *376*(9757), 2018–2031. https://doi.org/https://doi.org/10.1016/S0140-6736(10)61029-X

Roche, D. B., Buenavista, M. T., & McGuffin, L. J. (2013). The FunFOLD2 server for the prediction of protein–ligand interactions. *Nucleic Acids Research*, *41*(W1), W303–W307.

Roche, D. B., Tetchner, S. J., & McGuffin, L. J. (2011). FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*, *12*(1), 160.

Rogozin, I. B., Gertz, E. M., Baranov, P. V, Poliakov, E., & Schaffer, A. A. (2018). Genome-Wide Changes in Protein Translation Efficiency Are Associated with Autism. *Genome Biology and Evolution*, *10*(8), 1902–1919. https://doi.org/10.1093/gbe/evy146

Rohrig, H., Schmidt, J., Miklashevichs, E., Schell, J., & John, M. (2002). Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(4), 1915–1920. https://doi.org/10.1073/pnas.022664799

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., & Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *ELife*, *3*, e03523. https://doi.org/10.7554/eLife.03523

Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X., & Albà, M. M. (2018). Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nature Ecology & Evolution*, *2*(5), 890–896. https://doi.org/10.1038/s41559-018-0506-6

Sael, L., & Kihara, D. (2012). Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins: Structure, Function, and Bioinformatics*, *80*(4), 1177–1195.

Scarsi, M., Majeux, N., & Caflisch, A. (1999). Hydrophobicity at the surface of proteins. *Proteins: Structure, Function, and Bioinformatics*, *37*(4), 565–575.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003

Schwab, S. R., Li, K. C., Kang, C., & Shastri, N. (2003). Constitutive Display of Cryptic Translation Products by MHC Class I Molecules. *Science*, *301*(5638), 1367 LP – 1371. http://science.sciencemag.org/content/301/5638/1367.abstract

Simhadri, V. L., Hamasaki-Katagiri, N., Lin, B. C., Hunt, R., Jha, S., Tseng, S. C., Wu, A., Bentley, A. A., Zichel, R., Lu, Q., Zhu, L., Freedberg, D. I., Monroe, D. M., Sauna, Z. E., Peters, R., Komar, A. A., & Kimchi-Sarfaty, C. (2017). Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. *Journal of Medical Genetics*, *54*(5), 338 LP – 345. https://doi.org/10.1136/jmedgenet-2016-104072

Skarshewski, A., Stanton-Cook, M., Huber, T., Al Mansoori, S., Smith, R., Beatson, S. A., & Rothnagel, J. A. (2014). uPEPperoni: an online tool for upstream open reading frame location and analysis of transcript conservation. *BMC Bioinformatics*, *15*, 36. https://doi.org/10.1186/1471-2105-15-36

Slavoff, S A, Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L., & Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology*, *9*(1), 59–64. https://doi.org/10.1038/nchembio.1120

Slavoff, Sarah A, Heo, J., Budnik, B. A., Hanakahi, L. A., & Saghatelian, A. (2014). A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *The Journal of Biological Chemistry*, *289*(16), 10950–10957. https://doi.org/10.1074/jbc.C113.533968

Smith, J. E., Alvarez-Dominguez, J. R., Kline, N., Huynh, N. J., Geisler, S., Hu, W., Coller, J., & Baker, K. E. (2014a). Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Reports*, *7*(6), 1858–1866. https://doi.org/10.1016/j.celrep.2014.05.023

Smith, J. E., Alvarez-Dominguez, J. R., Kline, N., Huynh, N. J., Geisler, S., Hu, W., Coller, J., & Baker, K. E. (2014b). Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Reports*, *7*(6), 1858–1866. https://doi.org/10.1016/j.celrep.2014.05.023

Snyder, P. W., Mecinović, J., Moustakas, D. T., Thomas, S. W., Harder, M., Mack, E. T., Lockett, M. R., Héroux, A., Sherman, W., & Whitesides, G. M. (2011). Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proceedings of the National Academy of Sciences*, *108*(44), 17889–17894.

Sousa, M. E., & Farkas, M. H. (2018). Micropeptide. *PLOS Genetics*, *14*(12), e1007764. https://doi.org/10.1371/journal.pgen.1007764

Stein, C. S., Jadiya, P., Zhang, X., McLendon, J. M., Abouassaly, G. M., Witmer, N. H., Anderson, E. J., Elrod, J. W., & Boudreau, R. L. (2018). Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Reports*, *23*(13), 3710-3720.e8. https://doi.org/https://doi.org/10.1016/j.celrep.2018.06.002

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., & Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*, *41*(17), e166–e166. http://dx.doi.org/10.1093/nar/gkt646

UniProt: a hub for protein information. (2015). *Nucleic Acids Research*, *43*(Database issue), D204-12. https://doi.org/10.1093/nar/gku989

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Wang, D., Li, C., Zhao, Q., Zhao, L., Wang, M., Zhu, D., Ao, G., & Yu, J. (2009). Zm401p10, encoded by an anther-specific gene with short open reading frames, is essential for tapetum degeneration and anther development in maize. *Functional Plant Biology*, *36*(1), 73–85. https://doi.org/10.1071/FP08154

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, *41*(6), e74–e74. http://dx.doi.org/10.1093/nar/gkt006

Wang, R. F., Parkhurst, M. R., Kawakami, Y., Robbins, P. F., & Rosenberg, S. A. (1996). Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *The Journal of Experimental Medicine*, *183*(3), 1131 LP – 1140. http://jem.rupress.org/content/183/3/1131.abstract

Ward, L. D., & Kellis, M. (2012). Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science, 337*(6102), 1675 LP – 1678. https://doi.org/10.1126/science.1225057

Wen, J., Lease, K. A., & Walker, J. C. (2004). DVL, a novel class of small polypeptides: overexpression alters Arabidopsis development. *The Plant Journal : For Cell and Molecular Biology*, *37*(5), 668–677. https://doi.org/10.1111/j.1365-313x.2003.01994.x

Wikipedia contributors. (2020). *Genetic variation --- Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Genetic_variation&oldid=937860463

Wilusz, J. E., Sunwoo, H., & Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes & Development* , *23*(13), 1494–1504. https://doi.org/10.1101/gad.1800909

Xie, Z.-R., & Hwang, M.-J. (2015). *Methods for Predicting Protein–Ligand Binding Sites BT - Molecular Modeling of Proteins* (A. Kukol (Ed.); pp. 383–398). Springer New York. https://doi.org/10.1007/978-1-4939-1465-4_17

Yeasmin, F., Yada, T., & Akimitsu, N. (2018). Micropeptides Encoded in Transcripts Previously Identified as Long Noncoding RNAs: A New Chapter in Transcriptomics and Proteomics. *Frontiers in Genetics*, *9*, 144. https://doi.org/10.3389/fgene.2018.00144

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. http://dx.doi.org/10.1093/nar/gkx1098

Zhang, H., Li, P., Zhong, H.-S., & Zhang, S.-H. (2013). Conservation vs. variation of dinucleotide frequencies across bacterial and archaeal genomes: evolutionary implications. *Frontiers in Microbiology*, *4*, 269. https://doi.org/10.3389/fmicb.2013.00269

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309.

Zhu, M., & Gribskov, M. (2019). MiPepid: MicroPeptide identification tool using machine learning. *BMC Bioinformatics*, *20*(1), 559. https://doi.org/10.1186/s12859-019-3033-9

Zhu, X., Xiong, Y., & Kihara, D. (2015). Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2. 0. *Bioinformatics*, *31*(5), 707–713.

PUBLICATIONS

Zhu, M., & Gribskov, M. (2019). MiPepid: MicroPeptide identification tool using machine learning. BMC Bioinformatics, 20(1), 559. https://doi.org/10.1186/s12859-019-3033-9

Zhu, M., & Gribskov, M. Analysis of genetic variation in micropeptides and long noncoding RNAs. In submission to Bioinformatics.

Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., & Kihara, D. (2019). Protein docking model evaluation by 3D deep convolutional neural networks. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz870