McGill University

Doctoral Thesis

Learning From Watching Evolution

Author: Faizy Ahsan

Supervisor: Dr. Mathieu Blanchette

co-Supervisor: Dr. Doina Precup

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computer Science

School of Computer Science

McGill University, Montreal

October 25, 2021

This document is dedicated to my parents Rukhsana and Abdul Qaiyum

Abstract

The computational prediction of functions associated with biological sequences is of high importance in bioinformatics research. It is crucial in understanding complex cellular mechanisms by complementing wet-lab experiments and to study diseases related to mutations in such sequences. However, the existing sequence-based predictors are quite inferior compared to the analogous wet-lab experiments. In this thesis, we present two approaches that make use of evolutionary information in order to boost the performance on biological sequence function prediction tasks. Our first approach, PhyloReg, is a semi-supervised approach that regularizes a given supervised model, e.g. logistic regression or convolutional neural network, such that the resulting model predicts similar scores over the neighbouring orthologous sequences in the phylogenetic tree. The second approach, PhyloPGM, is applicable to the inference stage. It combines prediction scores of a previously trained classifier on orthologous sequences to boost the prediction accuracy. Lastly, we provide a web-interface to compute PhyloPGM scores on given genomic location(s) that will allow researchers to focus on the PhyloPGM outcomes.

The results with 422 ChIP-seq datasets show that PhyloReg improves the transcription factor binding sites (TFBSs) prediction accuracy significantly. Similarly, PhyloPGM is shown to effectively boost the RNA binding prediction accuracy in 31 CLIP-seq datasets. The web interface provides a practical application of PhyloPGM where user can analyze whether a genomic location binds to a RBP. We showcase our methodologies w.r.t. the TFBSs and RNA binding prediction problems, however, both PhyloReg and PhyloPGM are, in principle, applicable to any supervised learning algorithms and other sequence

function prediction tasks such as miRNA target gene and mRNA subcellular localization predictions.

Abrégé

La prédiction informatique des fonctions associées aux séquences biologiques est d'une grande importance dans la recherche en bioinformatique. Elle est cruciale pour comprendre les mécanismes cellulaires complexes en complétant les expériences en laboratoire humide et pour étudier les maladies liées aux mutations dans de telles séquences. Cependant, les prédicteurs basés sur les séquences existants sont assez inférieurs aux expériences analogues en laboratoire humide. Dans cette proposition, nous présentons deux approches qui utilisent des informations évolutives afin d'améliorer les performances des tâches de prédiction de fonction de séquence biologique. Notre première approche, PhyloReg, est une approche semi-supervisée qui régularise un modèle supervisé donné, par ex. régression logistique ou réseau neuronal convolutif de telle sorte que le modèle résultant prédit des scores similaires sur les séquences orthologues voisines dans l'arbre phylogénétique. La deuxième approche, PhyloPGM, est applicable à l'étape d'inférence qui combine les scores de prédiction du classificateur préalablement formé sur des séquences orthologues pour augmenter la précision de la prédiction. Enfin, nous fournissons une interface Web pour calculer les scores PhyloPGM sur un ou plusieurs emplacements génomiques donnés, ce qui permettra aux chercheurs de se concentrer sur les résultats de PhyloPGM.

Les résultats avec 422 ensembles de données ChIP-seq montrent que PhyloReg améliore considérablement la précision de prédiction des sites de liaison aux facteurs de transcription. De même, il est démontré que PhyloPGM augmente efficacement la précision de la prédiction de la liaison à l'ARN dans 31 ensembles de données CLIP-seq. L'interface

Web fournit une application pratique de PhyloPGM où l'utilisateur peut analyser si un emplacement génomique se lie à un RBP. Nous présentons nos méthodologies w.r.t. les problèmes de prédiction de liaison de TFBS et d'ARN, cependant, PhyloReg et PhyloPGM sont, en principe, applicables à tous les algorithmes d'apprentissage supervisé et à d'autres tâches de prédiction de fonction de séquence par exemple les prédictions de localisation subcellulaire du mARN et les prédictions du gène cible du mi-ARN.

Acknowledgements

I would not have been able to follow the journey of a doctoral student without the support of many individuals. I would like to express my sincere gratitude to all who helped shape my thoughts and reasoning over the years, including James Wagner, David Becerra, Philip Bouvrette, Alex Butyaev, Chris Drogaris, Haji Mohammad Saleem, Rola Dali, Vladimir Reinharz, Nikou Sefat, Ayrin Tabibi, Elliot Layne, Dongjoon Kim, Zichao Yan, Yanlin Zhang, Akash Singh, Lucas hamilton-Bourezg, Johnny Liu, Elizabeth O'Meara, Alexandre Drouin and many others whom I may have forgotten, but are appreciated nonetheless.

In addition, to the members of my Ph.D. committee, Yue Li and Hamed Najafabadi and the members of my Ph.D. defence examiners, Reihaneh Rabbany and Abdoulaye Diallo. Furthermore, to my mother Rukhsana, father Abdul Qaiyum and brothers Kashif Rizvee and Simaab Amir for their moral and emotional support.

Finally, I am extremely grateful to my advisors Mathieu Blanchette and Doina Precup for the guidance and mentorship throughout the program and Jerome Waldispuhl for the elaborative discussions and shared lab space.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The binding of proteins to DNA or RNA sequence is an integral part of biological functions within a cell of a living organism. The computational prediction of protein bindings to DNA or RNA sequences will allow to comprehend associated functions and to study related diseases that may occur due to erroneous binding. For example, the problem of whether a given DNA sequence will bind to a protein called transcription factor (TF) is relevant for the study of gene regulatory networks and the diseases associated with the mutations in the DNA regulatory regions [Slattery et al., 2014, Spielmann and Mundlos, 2013]. Similarly, the binding prediction of a protein called RNA binding protein (RBP) to an RNA sequence is significant in comprehending post-transcriptional regulation and the associated diseases [Stefl et al., 2005, Lukong et al., 2008a].

Recent years have seen an explosion of machine learning (ML) tools, especially deep learning algorithms, for various sequence-function prediction tasks and have superseded the results of classical computational approaches [Alipanahi et al., 2015, Quang and Xie, 2016, Pan and Shen, 2018]. Although machine learning based methods have outperformed classical computational methods, existing computational methods are yet to replace their experimental counterparts due to high false positive prediction rates.

One broad area that is less explored for the biological sequence-function prediction tasks is the use of evolutionary information, which not only will augment the training data, but will allow identifying TF or RBP binding sites in the light of evolution. Furthermore, the orthologous regions in different organisms i.e. the genomic regions derived from the same common ancestor, are indeed observed to be under selection as per the orthologous conjecture [Shiraishi et al., 2001, Shabalina et al., 2004, Papatsenko et al., 2006, Cooper and Brown, 2008, Chen and Zhang, 2012, Stamboulian et al., 2020]. However, the raw integration of sequence conservation information may not improve sequence-function prediction models due to a phenomenon called binding sites turnover [Sinha and Siggia, 2005, Moses et al., 2006]. Under this phenomenon, the binding property of a given region such as the number of binding sites, is maintained, but the sequence itself is not conserved. Moreover, a study by Kheradpour et al. [2013] suggests that the conservation of transcription factor binding sites (TFBSs) is more crucial for enhancer activity than the overall sequence conservation. Therefore, a more sophisticated integration of sequence conservation information is required for a robust computational model to predict TF and RBP binding sites.

In this chapter, we first describe biology of transcriptional regulation and wet-lab experimental identification of TFBSs in § 1.2 and § 1.3. Then, we describe biology of RBP and wet-lab experimental identification of RBP binding sites in § 1.4 and § 1.5. We provide a brief survey on existing computational models to predict TF or RBP binding sites § 1.6. In § 1.7, we discuss some of the major challenges associated with the problem of predicting TF or RBP binding sites. We describe comparative genomics approaches to predict TF or RBP binding sites in § 1.8. Finally, we present the thesis outline and mention the author contributions in § 1.9.

1.2 Biology of Transcriptional Regulation

A particular DNA segment is transcribed into a RNA molecule through a process called transcription, which has a significant part in the synthesis, regulation and processing of proteins. A transcription factor (TF) is a protein that binds to regulatory regions in DNA and regulates the transcription process of adjacent gene. Several regulatory regions in the DNA e.g. promoter, enhancer and silencer, help to direct and regulate transcription of the DNA segment.

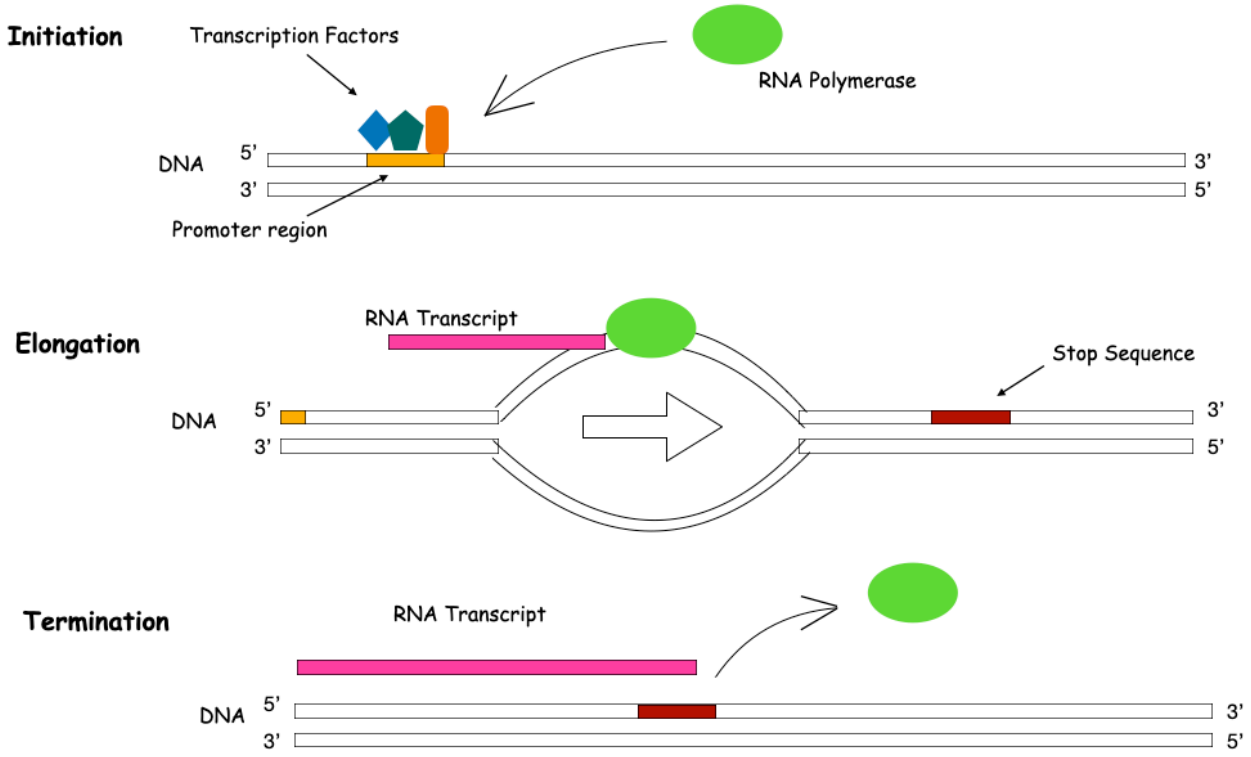


Figure 1.1: Transcription process. Multiple TFs bind to promoter region in DNA and recruit RNA polymerase to initiate the transcription process. During the elongation step, the RNA polymerase move in 5′ to 3′ direction and a RNA transcript is formed. The transcription process terminates after a stop sequence is reached by the RNA polymerase.

The DNA segment where transcription process starts is called transcription start site (TSS). The promoter region of a gene is usually 100-1000 base pairs long and is located near the TSS [Pacheco, 2013]. A transcription process, in general, is consist of three steps:

initiation, elongation and termination. In the initiation step, TFs bind to promoter region and initiate the transcription process. The bound TFs recruit RNA polymerase enzyme in the promoter region and a transcription bubble is created. In the elongation step, the transcription bubble divides the DNA strand while moving in 5′ to 3′ direction and simultaneously a RNA transcript is formed. The transcription process terminates when a stop sequence is detected by the RNA polymerase. Figure 1.1 describes a transcription process.

A gene may have enhancers and silencers as regulatory regions in DNA. The enhancer regions are usually 50-1500 base pairs long and can be located in either direction from TSS as far as 1,000,000 base pairs away [Pennacchio et al., 2013]. Some TFs can bind to enhancer regions and increase the transcription rate [Griffiths et al., 2005]. The silencer regions have similar features as enhancers in terms of length and location [Riethoven, 2010]. The transcription rate can decrease if some TFs bind to silencer region [Lieberman and Marks, 2009].

In human genome, there are around 1600 TFs as suggested by Lambert et al. [2018] and the number of observed genes are about one order higher in magnitude ($\sim$20,000 [Pennisi, 2012]. Several TFs, instead of a single TF, work together for the production of a gene. Similar to other proteins, TFs are transcribed from a DNA segment into an RNA molecule and then translated to a protein. The TFs are translated in cell's cytoplasm in eukaryotes and need to be relocated to nucleus, where DNA is located, in order to regulate transcription process. Many other proteins direct TFs to the nucleus through nuclear localization signals [Whiteside and Goodbourn, 1993]. A TF can even regulate itself by regulating the gene that produces it.

A TF has a signal sensing domain (SSD) that can make it active or inactive and only an active TF can take part in gene regulation. There are many ways that a TF can be activated, e.g. ligand binding, phosphorylation [Bohmann, 1990], interaction with other TFs [Massagué et al., 2005, Glass and Rosenfeld, 2000]. Apart from SSD, a TF has two main binding domains: DNA-binding domain (DBD) and trans-activating domain (TAD). The

DBD allows a TF to bind with regulatory regions and the TAD helps to bind with other proteins such as coactivators or corepressors.

The specific DNA sequences to which a TF binds are known as transcription factor binding sites (TFBSs) and are usually 6 to 20 base pairs long Zambelli et al. [2012]. In general, a TF binds to a TFBS in a sequence specific way and the specific pattern to which a TF binds is known as its motif. A TF may not bind to all the bases in TFBS and binding strength with different bases may differ. Moreover, a motif can be degenerate i.e. a nucleotide occurring at a paricular position in TFBS is not fixed. Thus, a TF can bind to a subset of closely related sequences. It should be noted that TFs binding is observed to be highly clustered i.e. TFBSs of many TFs are usually present in relatively same genomic regions [Yan et al., 2013].

A TF may bind in one cell type and not the other for a given genomic location even if its motif is present. The cell-type specificity of a TF binding is determined by multitudes of factors. A TF may require cooperative binding with other TFs and the required TFs may be absent in some cell-types [Panne, 2008, Wasson and Hartemink, 2009, Meijsing et al., 2009, Kitayner et al., 2010, Siggers et al., 2011, Slattery et al., 2011]. A TF may need to compete with other TFs and proteins in order to bind with a genomic location [Miller and Widom, 2003, Mirny, 2010, Teif and Rippe, 2010]. Moreover, some regulatory regions may not be accessible to TFs [Bai and Morozov, 2010, Pique-Regi et al., 2011]. Therefore, chromatin context is an important factor for cell-type specificity of TF binding. For example, addition of methyl groups to histones, known as histone methylation, can either increase or decrease the rate of TF binding [Gupta et al., 2010]. Additionally, DNA can temporarily become accessible to TFs due to thermal fluctuations that can partially unwrap the nucleosome [Cuesta-López et al., 2011]. One other reason of cell-type specificity of TF binding is the presence of pioneer TFs that are capable of binding with inaccessible nucleosomal DNA and making it accessible to other TFs [Glatt et al., 2011, Barozzi et al., 2014].

Mutations in TFBSs may disrupt the gene regulatory network and may lead to a disease [Spielmann and Mundlos, 2013]. For example deletions in the regulatory regions of FOXL2 gene is linked to blepharophimosis syndrome that affects eyelids development [Beysen et al., 2005]. Benko et al. [2009] report that deletions in the regulatory regions of SOX9 gene are associated with Pierre Robin syndrome that causes small jaw and cleft palate. A study by Lettice et al. [2011] suggests that changes in enhancer activities of SHH gene may result into holoprosencephaly spectrum (HPES) disorder that causes limb malformation. Similarly, a study by Dathe et al. [2009] suggests that duplication of a regulatory region of BMP2 gene may lead to autosomal-dominant brachydactyly type A2 (BDA2) and result into limb malformation. Alterations in regulatory regions may cause transcriptional dysregulation and, possibly, lead to cancer causing genetic alterations [Bradner et al., 2017].

1.3 Experimental Identification of Transcription Factor Binding Sites

A ChIP-seq experiment is an *in vivo* wet lab experiment to identify the TFBSs for a TF in the entire genome of a cell [Solomon et al., 1988, Johnson et al., 2007, Robertson et al., 2007]. The TF of choice is cross-linked with DNA, usually, by treating cell with formaldehyde. The DNA is then fragmented and the TF-bound DNA fragments are immunoprecipitated using a specific antibody. The precipitated TF-DNA complexes are recovered, sequenced and aligned to a reference genome. The regions with enriched alignments indicate TFBSs and are identified through peak-calling programs [Robertson et al., 2007, Fejes et al., 2008, Zhang et al., 2008]. A typical ChIP-seq experiment results into 2-20 million genomic regions of 200-300 bps in length where a TF can bind in the given type of cell [Pepke et al., 2009]. One ChIP-seq experiment costs around 500-100$ in general. The ENCODE consortium [Consortium et al., 2012] has provided the ChIP-seq based TFBSs for hundreds of transcription factors in more than eighty cell lines. It should be noted

that ChIP-seq experiments are unable to identify the exact TFBSs due to experimental limitations during the sonication stage. A ChIP-exo experiment is a refined version of ChIP-Seq that uses enzymes called exonuclease to trim immunoprecipitated DNA precisely from the cross-linked site and, in theory, can identify TFBSs in single nucleotide resolution [Rhee and Pugh, 2011].

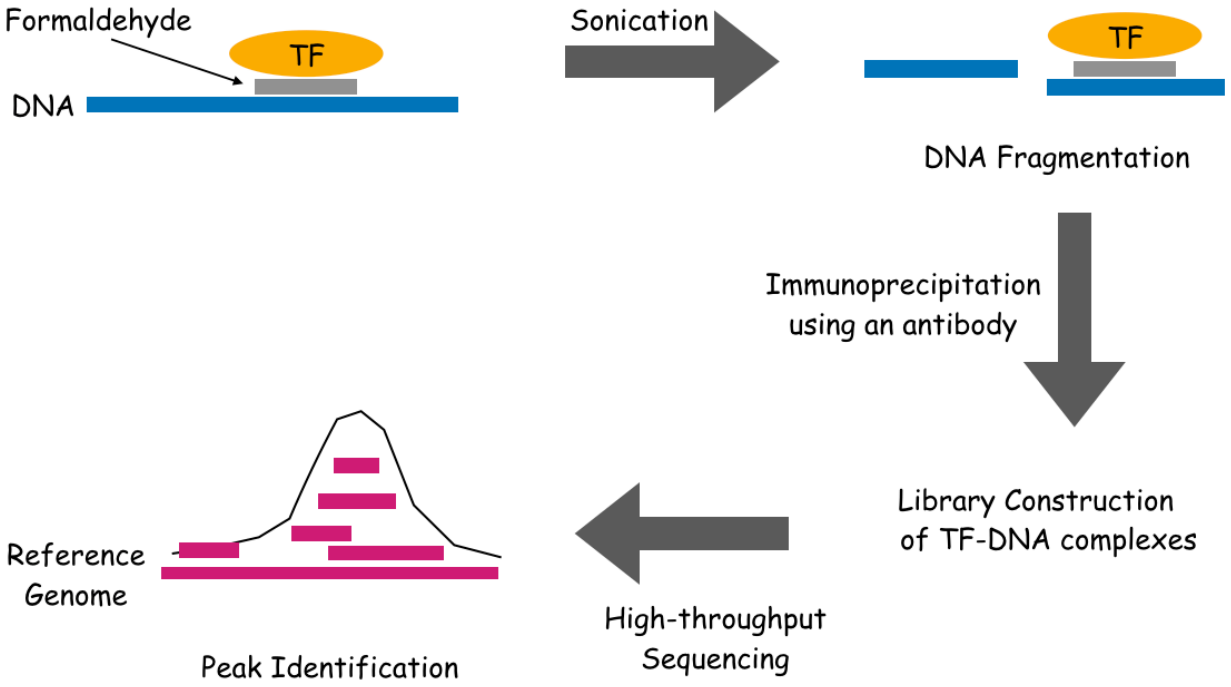


Figure 1.2: ChIP-seq workflow. In the ChIP step, a TF is cross-linked with DNA using formaldehyde. The DNA is then fragmented by sonication. The cross-linked DNA segments are then immunoprecipitated using an antibody. A library is constructed with the precipitated TF-DNA complexes. The recovered DNA fragments are sequenced and aligned to get the genomic positions. The genomic regions with enriched alignments are identified through peak calling programs and are considered as TFBSs. The figure is from Liu et al. [2010]

The discernment of gene regulatory network will require ChIP-seq experiments to be repeated for all pairs of TF and cell type and will be immensely expensive and time consuming. Moreover, ChIP-seq identified regions may be false positive TFBSs or inconsequential binding [Vanhille et al., 2015, Barakat et al., 2018]. Despite of the shortcomings,

the ChIP-seq experiments are preferred wet-lab experiments for genome-wide studies of identifying *in vivo* TFBSs at the time of this thesis also discussed in a recent study [Ferraz et al., 2021].

1.4 Biology of RNA Binding Proteins

An RNA binding protein (RBP) is a protein that binds to a specific sequence motif in a RNA sequence and plays a significant regulatory process in the post-transcriptional phase of gene expression such as RNA stability, splicing, subcellular localization and translation [Keene, 2001, Stefl et al., 2005] (see figure 1.3). The length of a human RBP motif roughly varies from 4-10 bps [Gabut et al., 2008, Cook et al., 2010]. The RBPs mostly bind in the 3' untranslated region (UTR) of RNA whose average length in human is about 800 bps [Mignone and Pesole, 2018]. In fact, multiple RBPs bind to the 3' UTR of RNA and their unison regulates the gene expression [Quattrone and Dassi, 2019]. There are around 2000 RBPs reported in human [Castello et al., 2012, Quattrone and Dassi, 2019, Benoit Bouvrette et al., 2020]. An RBP has a RNA binding domain that recognizes a RNA region to bind to [Stefl et al., 2005]. The most abundant RNA binding domains are RNA-recognition motif (RRM), double stranded RNA-binding motif (dsRBM), and the nucleic-acid-binding domain, the CCHH-type zinc-finger domain [Stefl et al., 2005].

A RRM domain interacts with RNA, mostly, in a sequence specific way [Handa et al., 1999]. RBPs with more than one RRMs require cooperative binding of at least two RRMs with the RNA for the high-affinity binding [Stefl et al., 2005]. RBPs with RRM domain play an essential role in many cellular functions, such as RNA processing, splicing, export and stability [Dreyfuss et al., 2002]. In contrast to RRM domain, a dsRBM binding is determined by the specific shape of the target RNA [Stefl et al., 2005]. RBPs with dsRBM domain play an essential role in RNA processing, editing, interference, localization and repression [Doyle and Jantsch, 2002, Saunders and Barber, 2003]. A study by Lu et al. [2003] found that the zinc fingers first interact with the backbone double helix structure

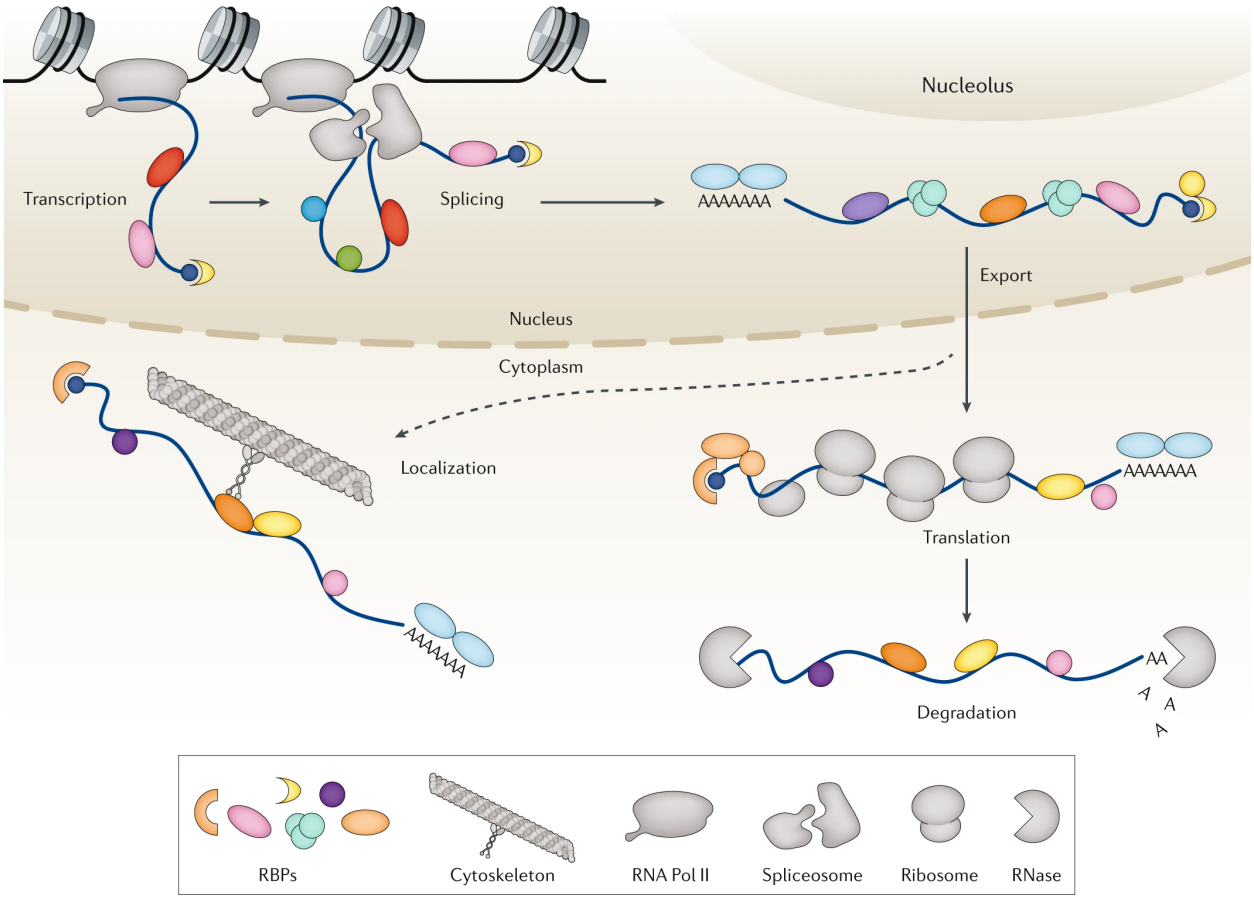


Figure 1.3: The binding of RBPs to RNAs has important biological functions e.g. splicing, export, localization, translation etc. Figure is from Gebauer et al. [2020]

and then specifically recognize the individual bases in the RNA loop regions. RBPs with the CCHH-type zinc finger domain play an essential role in transcription regulation, RNA processing and degradation [Lu et al., 2003, Hudson et al., 2004].

The structure of RNA sequence plays a crucial role in the RNA binding to RBPs [Lee et al., 1997, Montange and Batey, 2008, Mortimer et al., 2014, De Groot et al., 2019]. RNA structure can be predicted from its sequence and can be represented into two levels: secondary structure that involves canonical base-pairing and tertiary structure, the three dimensional shape of RNA molecule [Mathews et al., 1999]. The RNA secondary structure are usually determined by free energy minimization [Lück et al., 1996, Mathews et al., 1997]. RNAfold computes the minimum free energy and uses dynamic programming to backtrace the optimal secondary structure [Lorenz et al., 2011]. RNAfold and many

other tools to compute optimal, sub-optimal, locally stable RNA secondary structures are available in the viennaRNA package [Lorenz et al., 2011]. The secondary structure with minimum free energy may not represent the complete possible structures [Ding and Lawrence, 2003]. SFOLD provides ensemble of secondary structure based on statistical sampling [Ding and Lawrence, 2003]. The RNAShape tool can categorize various secondary structures of a RNA sequence into a simplified form [Steffen et al., 2006].

Computing tertiary structure of RNA is challenging due to the effect of environmental conditions (e.g. ion concentrations and temperature) [Pucci and Schug, 2019]. Some approaches to predict RNA tertiary structure use secondary structure as input and match parts of the secondary structure with templates in a database of known RNA tertiary structures e.g. RNAComposer [Biesiada et al., 2016], Vfold3D [Zhao et al., 2017], 3dRNA [Zhao et al., 2012]. Other methods make use of specific conformations (iFoldRNA [Ding et al., 2008], NAST [Flores and Altman, 2010], SimRNA [Boniecki et al., 2016]) or matching templates based on structure and geometry (RNABuilder [Flores et al., 2010], ModeRNA [Rother et al., 2011]).

Aberrations in RBP bindings to RNA may result in neurological disorders and cancer [Lukong et al., 2008b]. A fragile X syndrome (FXS) may result from undesired changes in the 5′ UTR of the FMR1 gene [Chelly and Mandel, 2001]. Similarly, myotonic dystrophy type 1 (DM1) may result from erratic functions of RBPs MBNL1 and CUGBP1 due to changes in the 3′ UTR of myotonic dystrophy protein kinase (DMPK) messenger RNA [Wang and Cooper, 2007]. Studies have suggested that erroneous regulation of RBPs are linked to cancer e.g. over expressed RBP Sam68 is linked to breast and prostate cancers [Lukong et al., 2005, Busa et al., 2007, Paronetto et al., 2010], under expressed RBP QKI is linked to gliomas [Chénard and Richard, 2008]. Hong [2017] reviewed probable roles of RBPs in cancer development.

1.5 Experimental Identification of RBP Binding Sites

Much similar to ChIP-seq, a CLIP-seq (Cross-linking immunoprecipitation Sequencing) experiment is an *in vivo* wet-lab experiment that can locate the RBP binding sites for a given RBP [Ule et al., 2003, Licatalosi et al., 2008b, Chi et al., 2009, Darnell, 2010]. In a CLIP process, the RNA and RBP are *in vivo* cross-linked using ultraviolet (UV) light that forms covalent bonds between RBP and RNA nucleotides. Then, the cells are lysed and RNAs are fragmented. The RNA-RBP complex are separated through immunoprecipitation. The nucleotides that were cross-linked are then converted to complementary DNAs (cDNAs), which in turn are sequenced and mapped to the genome to obtain the genomic locations of RBP binding sites. Optinally, peak calling programs such as Piranha [Uren et al., 2012], CLIPper [Lovci et al., 2013] and PureCLIP [Krakau et al., 2017] may be used to identify the locations of RBP binding sites. The identified RBP binding sites are roughly 100 nucleotides long. A typical CLIP-seq method is described in the figure 1.4.

The CLIP-seq experiments vary based on the CLIP protocols applied for the cross-linking and immunoprecipitation. The photoactivable ribonucleoside-enhnced cross-linking and immunoprecipitation (PAR-CLIP) method incorporates photoreactive ribonucleosides analogs like 4-thiouridine and 6-thioguanosine into RNA for enhanced recovery of RBP binding sites [Hafner et al., 2010b]. Although PAR-CLIP can locate RBP binding sites with high accuracy, it is limited to cultured cells and may cause nuclear stress response [Sheng and Cai, 2012, Burger et al., 2013]. The individual nucleotide-resolution cross-linking and immunoprecipitation (iCLIP) method reverse transcribes RNAs to identify the RBP binding sites at high resolution [König et al., 2010, Wang et al., 2010, Tollervey et al., 2011]. However, the effect of reverse transcription on CLIP method is not completely known. Additionally, the identification of cross-linked induced mutation sites (CIMS) along with the CLIP method is shown to identify the RBP binding sites [Zhang and Darnell, 2011].

Stražar *et al.* [Stražar et al., 2016] compiled a RBP binding sites dataset for 31 RBPs obtained from various CLIP-seq experiments. Each dataset consists of around 3,283-6,000

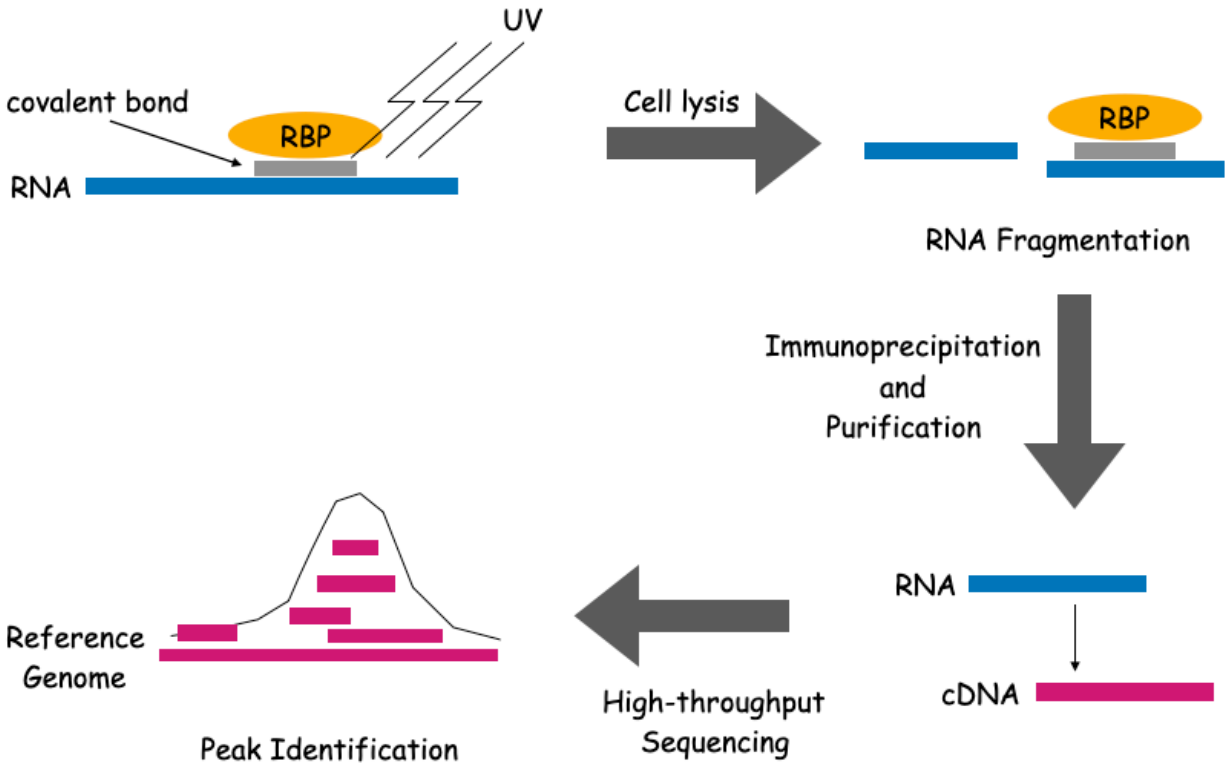


Figure 1.4: CLIP-Seq experiment. First, RNAs and RBPs are cross-linked to form covalent bonds in between them in the presence of ultraviolet(UV) radiation. After cell lysis, RNAs are fragmented. RNA-RBP complexes are separated through immunoprecipitation. Obtained RNAs are purified and converted to complementary DNAs (cDNAs). The cDNAs are sequenced and mapped to a reference genome. The peaks in the mapped regions, which are identified through peak-calling programs, are termed as RBP binding sites.

positive examples and 23,672-26,214 negative examples in train set and 1,892-2,000 positive examples and 7,725-7,991 negative examples in test set. The positive examples are the regions of the transcriptome identified as cross-linking sites in the CLIP-seq experiments. The negative examples are sampled regions from genes that were not identified as cross-linking sites in any of the CLIP-seq experiments. The length of each example is 101 nucleotides.

In general, a CLIP-seq experiment costs around 500-1000. Similar to ChIP-seq experiments, multiple CLIP-seq experiments for each RBP across each cell line is required to comprehend the post-transcriptional regulatory network, which will be expensive and

time consuming. Moreover, CLIP-seq experiment involves multiple steps with some steps posing optimization difficulties and low efficiencies [König et al., 2012, Moore et al., 2014, Ule et al., 2005]. Furthermore, some sequence bias in cross-linking step of CLIP-seq due to UV radiation is reported , but, the effect of bias is not clear [Sugimoto et al., 2012]. Nonetheless, CLIP-seq approaches are considered state-of-the-art wet-lab experiment to identify genome-wide RBP binding sites [Uhl et al., 2020]. The ENCORE project [Van Nostrand et al., 2020] used CLIP-seq methods as one of the tools to compile 1,223 data sets of 356 RBPs for characterizing and mapping RBPs in terms of their binding preferences, sub-cellular localization and the function associated with RBP binding sites [Deakyne and Mazin, 2011].

1.6 Computational Approaches to Predict TF and RBP binding sites

Computational approaches to predict TF and RBP binding sites offer several benefits over the associated wet-lab experiments. The computational approaches will allow to reduce the time and cost related to wet-lab experiments. The TF and RBP binding site locations obtained from wet-lab experiments are with respect to a reference genome. An individual genome may differ from the reference genome and using a computational model to know the binding sites will be more affordable than the wet-lab experiments in terms of time and cost. It will be easier to learn the impact of mutations with a computational model. Moreover, computational models will allow to comprehend the mechanisms determining the binding. It is possible with computational approaches to rapidly evaluate candidate sequences to assist therapeutic approaches. Finally, computational models will be quite useful in the scenarios where wet-lab experiments are virtually impossible e.g. to determine binding sites in the brain cell of a live individual or to determine binding sites in species whose genomes are difficult to obtain.

1.6.1 Shallow Learning Approaches

The classical computational approaches to predict TFBS are based on motif models that use consensus sequence or position weight matrices to represent the motifs [Stormo et al., 1982]. A consensus sequence is the sequence of predominant bases in the alignment of experimentally determined binding sites. Similarly, the position weight matrix denotes the relative frequency of bases (A,C,G,T) at each location of the aligned binding sites. Apart from aligning sequences of known binding sites, there are computational approaches that aim to discover new motifs in terms of consensus sequence or PWM from a given set of sequences and are known as motif-discovery approaches e.g. MEME [Bailey et al., 2006], REFINE [Bailey et al., 1994] and HOMER [Heinz et al., 2010]. Motif based models are applied in a sliding window fashion across the entire genome to obtain a score at each location and the locations with scores higher than a selected threshold are deemed as binding sites. One major drawback of motif based models is that the nucleotides at different positions in a consensus sequence or a PWM are assumed to be independent of each other, which is not always true [Man and Stormo, 2001, Bulyk et al., 2002, Maerkl and Quake, 2007]. The other major drawback is that they do not consider the sequence context and will indicate all genomic locations as binding sites that match with a chosen consensus sequence or PWM resulting into a high false positive rate. Moreover, the motif-based models are unable to distinguish between cell-type specific binding sites i.e. the locations where a certain TF binds in one cell type and not in the other, since the nucleotides at each position are same. Day and McMorris [1992], Stormo [2000] provide a detailed overview of motif-based models along with their strength and weakness. It should be noted that profile-HMMs, which can convert a multiple sequence alignment into a position-specific scoring system, can be used to identify motifs and locate transcription factor binding sites [Eddy, 1998]. For example, Mapper [Marinescu et al., 2005] used HMMER [Eddy, 1998, Bateman et al., 2002] on the alignment of known TFBSs to search for putative TFBSs. Mathelier and Wasserman [2013] used HMMs with ChIP-seq

data to graphically represent motifs in order to identify TFBSs. HMM-based models to predict TFBSs are reviewed in Slattery et al. [2014].

The motif-based models are now significantly outperformed by machine learning based approaches that allow to learn variable sized motifs and inter-dependent relations among the nucleotides [Li et al., 2015]. In general, DNA sequences of fixed lengths are labelled as bound or non-bound sites through wet-lab experiments, e.g. ChIP-seq. The sequence features are, often, represented as k-mer counts, which are used to train a supervised model as a binary classifier to predict TFBSs. For example Arvey et al. [2012] used support vector machines (SVM) [Cortes and Vapnik, 1995] and k-mer features with certain mis-matches allowed, Ghandi et al. [2014] allowed gaps within k-mers with SVM, and Sharmin et al. [2016] used k-mer features with ensemble models to predict TFBSs in a given cell type.

The sequence-based computational models to predict TFBSs are applicable to predict RBP binding sites [Tacke and Manley, 1995, Pérez et al., 1997, Liu et al., 1998, Sanford et al., 2008, 2009, Agostini et al., 2014]. For example Hogan et al. [2008] used MEME [Bailey et al., 2006] and REFINE [Bailey et al., 1994] to identify RBP binding motifs in yeast. RNACompete [Ray et al., 2009] used AlignACE [Roth et al., 1998, Hughes et al., 2000], MEME [Bailey et al., 2006] and MEMERIS [Hiller et al., 2006] to identify *in vitro* and *in vivo* RNA binding motifs. RBPmap is based on PWMs to predict RBP binding sites in human, mouse and Drosophila [Paz et al., 2014]. RPISeq applies SVM and Random Forest models to k-mer features of both RNA and RBP sequences to predict RBP binding sites [Muppirala et al., 2011].

There are many computational tools that use both sequence and structure information to predict RBP binding sites (catRAPID [Agostini et al., 2013], Livi and Blanzieri [2014], RCK [Orenstein et al., 2016]). For example, RNAContext [Kazan et al., 2010] annotated sequence nucleotides into paired, hairpin loop, unstructed and miscellaneous using SFOLD [Ding and Lawrence, 2003]. GraphProt [Maticzka et al., 2014] used RNAShape with graph kernels [Costa and De Grave, 2010] and SVM. CapR [Fukunaga et al., 2014] computed the binding preference probabilities for each bases within the secondary struc-

ture contexts and RPIBind [Luo et al., 2017] used local conformation structures of RNA to predict the RBP binding sites.

1.6.2 Deep Learning Approaches

The deep learning based approaches further improved the TFBS prediction accuracy over the shallow models [Koo and Ploenzke, 2020]. The deep learning based approaches differ in terms of model architectures. Convolutional neural networks (CNNs) are used to predict TFBSs by representing genomic sequences as image-like data [Alipanahi et al., 2015, Zhou and Troyanskaya, 2015, Zeng et al., 2016]. Typically, an l-sized sequence is transformed into a $l \times 4$ matrix by one-hot encoding of each nucleotide in the sequence. A CNN architecture is consist of a set of filters, known as convolutional filters or convolutional kernels that act as motif detectors, like PWMs. Each filter is represented as a $m \times 4$ matrix whose values are learned using training data. First, the convolutional filters are applied to the one-hot encoded input. Then, a pooling operation is applied by taking the maximum of the outputs of each convolutional filters, which is known as global-max pooling. It should be noted that pooling stage may involve other computations such as taking average or both maximum and average etc. The goal of the pooling operation is to aggregate the motif signals detected from convolutional filters. Afterwards, the output from pooling stage is passed through one or multiple fully connected neural networks to obtain the final prediction score. Alipanahi et al. [2015] used a CNN architecture to predict both TF and RBP binding sites. Zeng et al. [2016] explored several CNN architectures for TFBSs prediction problem.

One major drawback with CNN architectures is its inability to capture the positional dynamics of motifs in a sequence due to the indirect assumption of at most one motif existing in a sequence [Hassanzadeh and Wang, 2016]. A recurrent neural network (RNN) is able to capture temporal/spatial dynamics in sequence-like data and is shown to predict TFBSs [Shen et al., 2018]. An RNN architecture has connections between its units that forms a directed cycle, which facilitates to have an internal state and to learn dy-

namic temporal and spatial patterns in the data. However, traditional RNN suffers from a problem called vanishing gradient effect that inhibits to back-propagate the errors from a distant point of time [Hochreiter, 1998]. A long short term memory network (LSTM) is a type of RNNs that aims to solve the vanishing gradient problem with the help of a gating mechanism that allows it to retain or discard information [Hochreiter and Schmidhuber, 1997]. A gated recurrent unit (GRU) is another RNN architecture that uses gating mechanism to solve the vanishing gradient problem [Chung et al., 2014]. A bi-directional LSTM (BLSTM) or a bi-directional GRU applies two LSTMs or GRUs from both sides of a sequence and their outputs are combined to obtain the final output. Shen et al. [2018] trained bi-directional GRUs with k-mer embeddings on ChIP-seq datasets to predict TFBSs. Koo and Ploenzke [2020] provide summary of several deep learning architectures that are use in genomics with a focus on TFBS prediction.

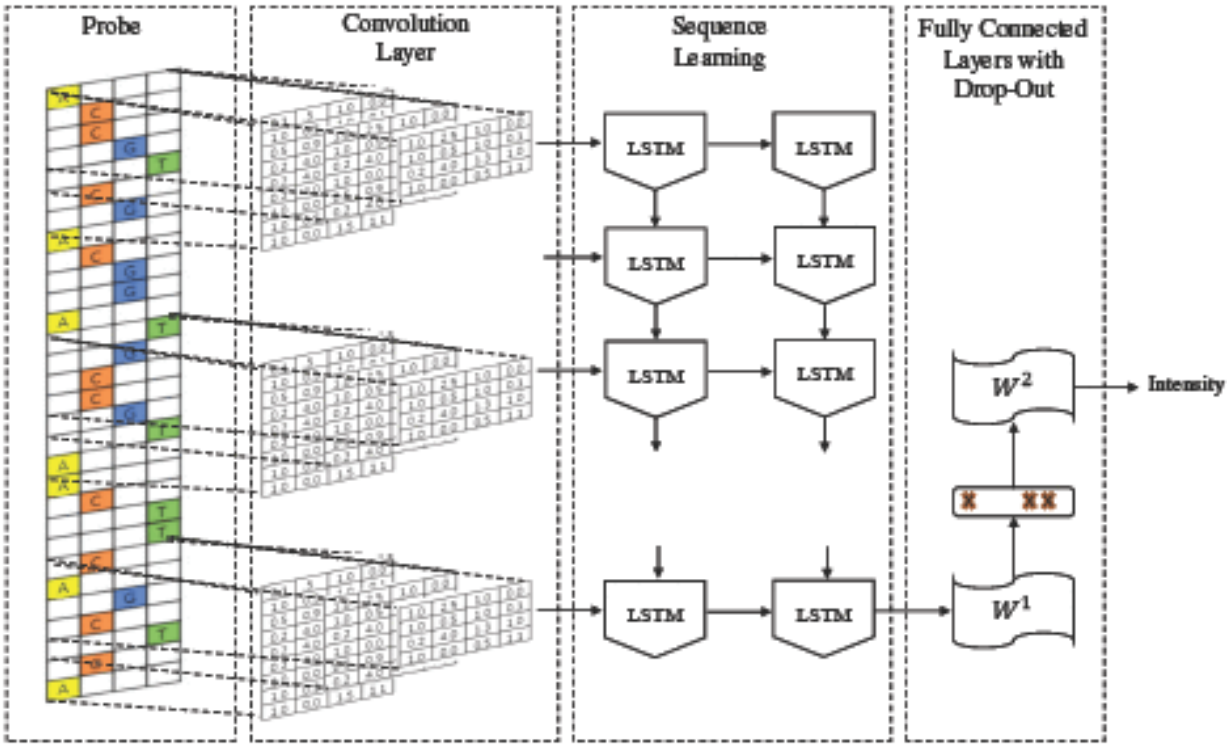


Figure 1.5: Example of a hybrid of CNN and RNN architecture, Deeperbind, which was designed to predict TFBS. Figure is from the figure 2 of Hassanzadeh and Wang [2016]

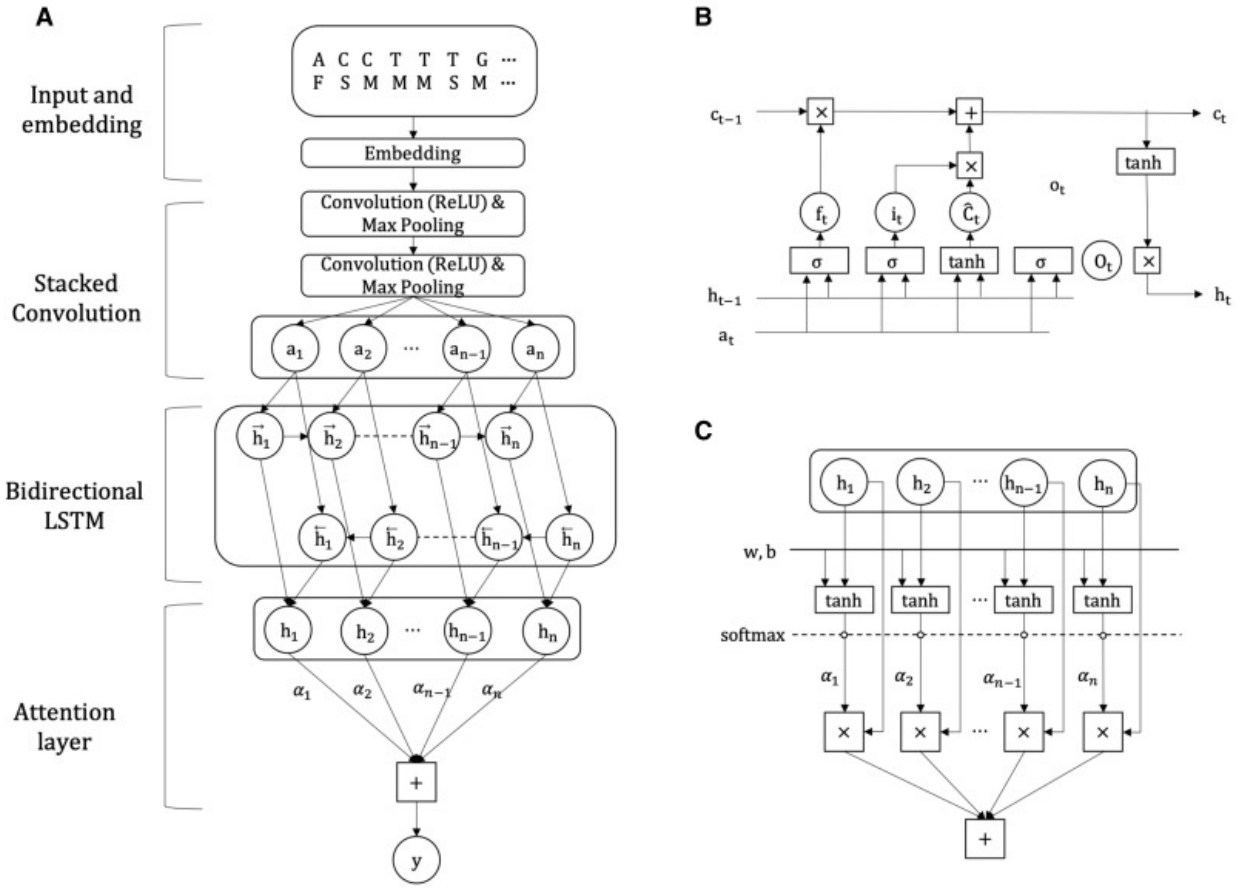


Figure 1.6: Example of a hybrid of CNN and RNN, RNATracker, which was designed to predict mRNA sub-cellular localization. Figure is from the figure 2 of Yan et al. [2019]

A hybrid of CNN and RNN architectures benefit from both type of networks and is shown to outperform a CNN or a RNN on TF and RBP binding site prediction tasks [Hassanzadeh and Wang, 2016, Quang and Xie, 2016, Pan et al., 2018, Yan et al., 2019, Quang and Xie, 2019, Park et al., 2020]. For example Deeperbind [Hassanzadeh and Wang, 2016] applies a convolutional layer to one-hot enocoded input sequence, followed by a bi-directional LSTM layer and a fully-connected neural network to predict TFBSs. Figure 1.5 shows a typical Deeperbind architecture. FactorNet [Quang and Xie, 2019] is another CNN-RNN hybrid architecture that can use DNA sequence information to predict TFBSs. An input to FactorNet is a sequence and its reverse complement. FactorNet applies a convolutional layer of 32 filters to both inputs. Then a dropout layer is applied

where a certain number of randomly selected neurons are set to zero in order to avoid overfitting. Afterwards, a max pool layer and a bi-directional LSTM layer are applied. A dropout layer is applied for one more time. Finally, both outputs from the previous layers are passed through a fully-connected neural network and averaged to obtain the final prediction score. In terms of predicting RBP binding sites from sequence input, RNATracker [Yan et al., 2019] is one such example of a CNN-RNN hybrid architecture. Although RNATracker is designed to predict sub-cullular localization of messenger RNA (mRNA), RNATracker should be suitable for other prediction tasks related to RNA such as RBP binding site prediction. RNATracker applies two conolutional layers to an one-hot encoded sequence input. Then a pooling layer is applied, which is followed by a bi-directional LSTM layer with attention. Afterwards, a fully-connected neural network is used to obtain the final prediction score. RNATracker applies a dropout layer after each convolutional and bi-directional LSTM layer to avoid overfitting. Figure 1.6 shows the RNATracker architecture.

1.6.3 Model Evaluation

Most of the computational TF and RBP binding site predictors are build as a supervised binary classifiers with a goal to predict whether a binding site is present or not in a given input sequence. There are some deep-learning based approaches, such as Avsec et al. [2021] that predicts the number of ChIP-seq reads at each position of the input sequence to identify TF motifs. The commonly used metrics to evaluate all these models are area under curve of receiver operating characteristic (AUROC) and area under curve of precision-recall (AUPR). A specificity of a model is defined as the correctly identified number of non-bound sequences and a recall score measures the correctly identified bound sequences. A precision score measures the correct predictions made on the total number of identified bound sequences. The specificity, recall and precision scores of a model can vary based on the threshold used over the model prediction scores. An AUROC score measures the are under the curve of specificity and recall scores, while an AUPR score

measures the area under the curve of precision and recall scores. An AUPR score is a better evaluation metrics than an AUROC score on the imbalanced datasets where number of non-bound sequences are much larger than the bound sequences. A false discovery rate (FDR) measures the proportion of incorrect binding sites identified by a model. The genome-wide number of binding sites of a TF or an RBP is much smaller ($<<$0.1%) than the number not-bound sites. Thus, a recall score at a lower FDR score (e.g. 0.1% or 0.05%) should provide an insight on genome-wide application of a binding site predictor for a TF or an RBP.

1.6.4 ENCODE-DREAM Competition

A recent DREAM competition [Kundaje et al., 2021] was organised to predict *in vivo* binding sites of 31 TFs in 13 cell lines with the use of DNA sequences, *in vitro* DNA shape parameters, *in vivo* chromatin accessibility profiles and gene expression data. The participants were provided with ChIP-seq datasets from ENCODE and were evaluated on held-out cell lines and held-out chromosomes. Apart from the main task of predicting TFBSs, participants were evaluated on several other parameters, such as influence of training and testing contexts on predictor performance, the extent to which guaranteed performance can be obtained, effect of cellular contexts on TFs and families of TFs, and relative contribution of sequence and chromatin features for binding prediction. The models were evaluated on AUPR scores and recall scores at various FDR thresholds.

Out of the top three winning submissions [Li et al., 2019, Keilwagen et al., 2019, Quang and Xie, 2019], only FactorNet was capable of taking only sequence data as input if required. Despite using sequence and non-sequence information related to TFBSs, none of the approaches could replicate the wet-lab experiments results. Moreover, the evolutionary data related to TF or RBP binding sites were neglected. A sequence-based TF or RBP binding sites predictor is preferred due to easier access to sequence data compared to non-sequence information. Moreover, sequence-based predictors could be integrated with other computational approaches that use sequence and other non-sequence data

e.g. chromatin accessibility, DNA shape etc. Our main focus in this thesis is to develop sequence-based approaches to predict TF and RBP binding sites that integrate evolutionary information with deep learning techniques.

1.7 Challenges in predicting TFBS and RNA-RBP binding

The computational prediction of TF or RBP binding sites is extremely challenging due to numerous reasons. The TF binding to DNA is determined by multitude of factors e.g. cooperative binding with other TFs [Kitayner et al., 2010, Meijsing et al., 2009, Panne, 2008, Wasson and Hartemink, 2009, Siggers et al., 2011, Slattery et al., 2011], chromatin accessibility [Bai and Morozov, 2010, Pique-Regi et al., 2011], competition with other TFs and proteins [Miller and Widom, 2003, Mirny, 2010, Teif and Rippe, 2010], and presence of pioneer TFs [Glatt et al., 2011, Barozzi et al., 2014]. The RBP binding to RNA also depends on similar factors e.g. cooperative binding of RRMs Stefl et al. [2005], presence of inter-domain linker [Stefl et al., 2005].

The epigenetics factors play a major role in TFs binding to DNA, e.g. DNA methylation (addition of methyl group to DNA strand) can recruit certain TFs in the promoter region and silence a gene expression [Lazarovici et al., 2013, Bird, 2002], histone modification can increase or decrease the transcription rate [Gupta et al., 2010]. Epigenetics can also affect RBP binding to RNA e.g. Dor and Cedar [2018] review the post-transcriptional regulation due to RNA methylation.

The RNA structure is quite flexible compared to DNA and are observed to play major role in RNA binding to RBPs [Ray et al., 2013, Daubner et al., 2013, Gupta and Gribskov, 2011]. Although a DNA sequence structure is more constrained than an RNA sequence, studies have suggested DNA structure as one of the determining factor of TF binding to DNA [Joshi et al., 2007, Rohs et al., 2009, White et al., 2013].

Finally, the present state-of-the-art computational methods for the TF or RBP binding site prediction problem are, mostly, based on binary classification problem i.e. the goal is

to predict whether a given DNA or RNA sequence will bind to a TF or RBP. Such methods, often, require training data sets to be roughly balanced i.e. equal number of binding and non-binding sites. Similarly, the trained models are, in general, evaluated on roughly balanced testing sets. However, the actual number of TF or RBP binding sites are much less ($<<$0.1%) than the non-bound sites. Therefore, a model that predicts even 1% of data as false positive will result into a large number of false positive ($\sim$99%) genome-wide.

1.8 Comparative Genomics Data and Approaches

One of the goal of comparative genomics is to analyze genomes from different species and identify the regions under selection. Such regions are supposed to be associated with important biological functions in species. The sequencing of genomes of multiple species has allowed the study of many biological functional activities in the light of evolution; for example to identify the conserved regions associated with certain biological functions in enteric bacteria [McClelland et al., 2000], yeast [Cliften et al., 2001, 2003, Kellis et al., 2003], mouse and humans [Consortium et al., 2002], caenorhabditis brigase and caenorhabditis elegans genomes [Stein et al., 2003], Arabidobsis and rice [Movahedi et al., 2011], fungi [Li and Breaker, 2017], drosophila [Berman et al., 2004], and to identify conserved non-coding segments in a wide range of vertebrates [Thomas et al., 2003], and within closely related species such as primates [Gumucio et al., 1992, Boffelli et al., 2003, Lawrie and Petrov, 2014]. The application of comparing genomes of different species are extensively reviewed [Collins et al., 2003, Hardison, 2003, Alföldi and Lindblad-Toh, 2013, Xiao et al., 2014, Yu et al., 2017, Eichler, 2019].

The comparative genomics approaches to predict binding sites are mostly based on the phylogenetic footprinting method [Tagle et al., 1988]. The main principle is that the function of binding sites should be conserved across species and the corresponding region should show differential selection pressure. The workflow consists of aligning the region of interest with its orthologs and analyzing the conserved portions. Pairwise alignment

tools such as BLASTZ [Schwartz et al., 2003b], LASTZ [Harris, 2007] can be used to align sequences of two species. To align sequences of more than two species, usually, a reference sequence is selected and all other sequences are compared and aligned accordingly in order to align the multiple sequences from different species [Bray and Pachter, 2003, Brudno et al., 2003, Schwartz et al., 2003a]. MultiZ allows to have a reference sequence in local regions and sub-groups of species [Blanchette et al., 2004b]. AVID [Bray et al., 2003] and LAGAN [Brudno et al., 2003] recursively align the given sequences and look for anchors, ie non-crossing and non-overlapping matches. The short-comings and advantages of the sequence alignment algorithms are reviewed in multiple surveys (e.g. Kumar and Filipski [2007], Li and Homer [2010], Wang et al. [2015]).

The conservation score at different positions are commonly measured through tree-based markov models e.g. PhastCons, PhyloP [Felsenstein and Churchill, 1996, Siepel et al., 2005, Siepel and Haussler, 2005, Yang, 1995]. PhastCons [Siepel et al., 2005] uses a hidden Markov model on multiple sequences aligned to compute the probability of each nucleotide in the alignment column of being conserved. PhastCons takes flanking regions into account for computing the conservation score at a given position. PhyloP [Siepel et al., 2006] applies similar techniques, but, focus on individual columns and ignore the effects of flanking columns. PhyloP can distinguish between fast-evolving and slow-evolving regions. PhastCons is more suited for detecting long continuous conserved elements and PhyloP is more effective in identifying evolutionary selection at specified class of nucleotides such as third codon positions.

There are many techniques that applied phylogenetic footprinting to find TFBSs or RBP binding sites. Kellis et al. [2003] computed motif conservation score based on the count of conserved regions to identify short conserved motifs and the process is repeated with neighboring regions to find a large set of motifs in yeast. The rVISTA tool [Loots and Ovcharenko, 2004] identified binding sites of multiple TFs in a given region based on motif-matching and filters the final TFBSs candidates based on sequence conservation of orthologous regions. MONKEY [Moses et al., 2004] used evolutionary distances in

probabilistic framework to identify conserved TFBSs. Li et al. [2008b] found that the probable functional TFBSs identified through PWM matching had higher PhastCons scores in Drosophila. Nettling et al. [2017] combined phylogenetic footprinting with intra-motif dependencies that is measured through computing mutual information of neighbouring regions in order to classify TFBSs. Ray et al. [2013] used RNACompete analysis and PhyloP scores to identify the functional roles of RBP binding sites.

The aligned sequences from closely related species may pose difficulties in identifying the binding sites, which are relatively short sequences as the non-functional regions may be equally conserved [Cliften et al., 2001, Blanchette and Tompa, 2002]. On the other hand, the sequences from distant related species may become too diverged to be accurately aligned and the short conserved regions may become difficult to be indentified [Cliften et al., 2001, Blanchette and Tompa, 2002]. Blanchette and Tompa [2002] make use of phylogenetic tree with phylogenetic footprinting and parsimony-based approach to identify the most conserved k-mers to circumvent the problem caused by sequence alignment. CONREAL [Berezikov et al., 2004] used PWMs to identify TFBSs candidates in different species, which are then aligned using anchoring technique similar to AVID and LAGAN to produce ordered set of most conserved TFBSs.

Ancestral sequence reconstruction (ASR) is another step of comparative genomics that could be used to identify the regions in genome that are associated with important biological functions [Sadri et al., 2011, Gumulya and Gillam, 2017]. The ASR requires multiple sequence alignment of orthologous regions from related extant species and a phylogenetic tree that relates those extant species. Then, the ancestral bases may be reconstructed based on greedy algorithm [Blanchette et al., 2008], maximum likelihood (e.g. PAML [Yang, 2007], Ancestors 1.0 [Diallo et al., 2010], FastML [Ashkenazy et al., 2012]), or bayesian approaches (e.g. LAZARUS [Hanson-Smith et al., 2010], PhyloBot [Hanson-Smith and Johnson, 2016]) that can explain the insertions, deletions and substitutions in the descendant species. In this thesis, we use ancestral sequences generated by Ancestors 1.0 [Diallo et al., 2010] to predict TF or RBP binding sites. Constructing an ances-

tor sequence that minimizes the number of insertions or deletions of nucleotides in the descendants is an NP-hard problem [Chindelevitch et al., 2006]. Ancestors 1.0 [Diallo et al., 2010] uses a heuristics approach based on tree-HMM [Diallo et al., 2007] to infer the insertions and deletions of nucleotides during the construction of ancestral sequences. Gumulya and Gillam [2017] provided a survey on the ASR approaches and their applications. Although the intermediate steps involved in ASR are error-prone (e.g. in sequence alignment [Anisimova et al., 2010], in phylogenetic tree [DeGiorgio and Rosenberg, 2016, Stadler et al., 2016]), the ancestral sequences are shown to have provided useful insights to functional sites in genome [Zhang and Rosenberg, 2002, Ugalde et al., 2004, Voordeckers et al., 2012, Bar-Rogovsky et al., 2013, Sadri et al., 2011]. Sadri et al. [2011] used MULTIZ and ancestor 1.0 programs to find orthologous regions to human sequences and composed features of count of conserved and substituted sites, which are then used with SVM to predict whether a position in human genome will mutate or not. Blanchette [2012] used inferred ancestral mammalian genomes with PWMs to predict TFBSs. MirAncesTar [Leclercq et al., 2017] used ancestral genomes and mammalian orthologs to boost the accuracy of existing human micro RNA target sites predictors by taking account of binding sites turnover. The ancestral sequences offer a large amount of data and evolutionary history that should be useful with sophisticated approaches like deep learning to identify TFBSs and RNA binding sites.

1.9 Thesis Roadmap and Author Contributions

In Chapter 2, we propose several ML methods that use evolutionary information to boost the performance of computational models for the biological sequence function prediction tasks. First, we develop a semi-supervised regularization approach called PhyloReg for the problem of TFBSs prediction. The PhyloReg approach assumes that the orthologous regions have same label i.e. whether the regions bind to a given TF and requires a base TFBS predictor that minimizes a loss function e.g. cross-entropy loss function. A

phylogenetic loss is added such that the base model is penalized for mis-predicting the orthologous regions. The resulting trained model is shown to improve the accuracy of base model in simulated and real data. The PhyloReg principle, in theory, is applicable to any sequence-based model for any sequence-function prediction task that minimizes some loss function.

In Chapter 3, we propose a probabilistic aggregation approch called PhyloPGM that combines the predictions of an existing TFBS predictor or RNA binding predictor on orthologs of a given human genomic sequence. The goal of PhyloPGM is to boost the base model prediction accuracy on the given human sequence. Unlike PhyloReg, model training is not required in the PhyloPGM approach. We show that the PhyloPGM significantly improved the accuracy of the base models for the TFBS and RNA binding prediction problem. Interestingly, the PhyloPGM model efficiently predicted more functional regions in human i.e. the TFBSs or RBP binding sites that are related to the fitness of human species than the base models.

In Chapter 4, we provide a web service called PhyloPGM-Web to analyze and assess a submitted human genomic sequence using PhyloPGM. The web service is developed using Cloudgene `http://www.cloudgene.io/`, nginx and python on a linux platform. A public url is provided where the user can submit the desired genomic location. In the background, the web service executes PhyloPGM with a number of pre-trained TF and RBP binding site predictors. Once the execution is complete, the user is presented with the predicted outcomes (a TF binding probability or a RBP binding probability).

In Chapter 5, we summarize our results and analyses with PhyloReg and phyloPGM and discuss the future aspects.

1.9.1 Author's contribution

This thesis includes the text and figures from three scientific articles and each article has been published, submitted, or is in preparation for submission to a journal. Faizy Ahsan

is the first author of each article found in Chapter 2-4. Listed below are the articles in their order of appearance throughout the thesis.

Chapter 2 is the extended version of

Ahsan, Faizy, Alexandre Drouin, François Laviolette, Doina Precup, and Mathieu Blanchette. "Phylogenetic Manifold Regularization: A semi-supervised approach to predict transcription factor binding sites." In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 62-66. IEEE, 2020.

FA performed the computational analysis and prepared the manuscript draft. FA and MB contributed equally to the final manuscript draft. All authors were involved in the discussion and development of the project.

Chapter 3

Faizy Ahsan, Zichao Yan, Doina Precup, and Mathieu Blanchette (2021) PhyloPGM: Boosting Regulatory Function Prediction Accuracy Using Evolutionary Information. *In preparation for submission to Nature Methods.*

FA performed the computational analysis and prepared the manuscript draft. ZY helped in the implementation of RNATracker model. FA, DP and MB were involved in the discussion leading to the development of PhyloPGM approach. FA and MB contributed equally to the final manuscript draft.

Chapter 4

Faizy Ahsan, Zichao Yan, and Mathieu Blanchette (2021) PhyloPGM As A Web Service. *In preparation for submission to Bioinformatics Application Note.*

FA implemented the web service and prepared the manuscript under the guidance and supervision of MB. ZY helped in the implementation of RNATracker model.

Chapter 2

Learning Transcription Factor Binding Site Predictor Using Evolutionary Data

In the following manuscript, I show that evolutionary data in terms of orthologous regions are helpful in improving the prediction accuracy of a sequence-based transcription factor binding site (TFBS) predictor. I propose a regularization technique, PhyloReg, that can be applied in a semi-supervised fashion for the TFBS prediction task. First, I design and implement an algorithm to create artificial datasets of a set of orthologs. The algorithm mimics the evolution of a biological function (activity) associated with a sequence. Then, I show that PhyloReg technique is helpful in predicting the activity if sufficient amount of selection pressure and training examples are present. Afterwards, I process and compile a previously published TFBS prediction datasets and extract orthologs/ancestors. I apply PhyloReg to a previously published deep neural network on the compiled dataset. I show that PhyloReg is more accurate in predicting TFBSs than the base model and sequence conservation techniques. I find that PhyloReg is more useful in the cases where base model has relatively lower accuracy or the number of training examples is relatively smaller. I observe that for certain transcription factors (TFs), the species that were closer to human in terms of evolutionary distance were more relevant in PhyloReg results, while for other TFs even distant species were relevant.

Citation: Ahsan, Faizy, Alexandre Drouin, François Laviolette, Doina Precup, and Mathieu Blanchette. "Phylogenetic Manifold Regularization: A semi-supervised approach to predict transcription factor binding sites." In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 62-66. IEEE, 2020.

2.1 Manuscript 1: Phylogenetic Manifold Regularization: A semi-supervised approach to predict transcription factor binding sites

Authors: Faizy Ahsan[1], Alexandre Drouin[2], François Laviolette[2], Doina Precup[1], and Mathieu Blanchette[1]

[1]School of Computer Science, McGill University, Montreal, Quebec, Canada

[2]Université Laval, Quebec, Canada

2.2 Abstract

The computational prediction of transcription factor binding sites remains a challenging problems in bioinformatics, despite significant methodological developments from the field of machine learning. Such computational models are essential to help interpret the non-coding portion of human genomes, and to learn more about the regulatory mechanisms controlling gene expression. In parallel, massive genome sequencing efforts have produced assembled genomes for hundred of vertebrate species, but this data is underused. We present PhyloReg, a new semi-supervised learning approach that can be used for a wide variety of sequence-to-function prediction problems, and that takes advantage of hundreds of millions of years of evolution to regularize predictors and improve accuracy. We demonstrate that PhyloReg can be used to better train a previously proposed deep learning model of transcription factor binding. Simulation studies further help de-

lineate the benefits of the approach. Gains in prediction accuracy are obtained over a broad set of transcription factors and cell types.

2.3 Introduction

The specific binding of transcription factors (TFs) to genomic regions called transcription factor binding sites (TFBSs) are key events in the regulation of gene expression. These protein-DNA interactions define gene regulatory networks and control biological functions. Mutations in TFBSs are known to cause numerous diseases [Slattery et al., 2014, Spielmann and Mundlos, 2013]. The information defining TF binding sites is present both in the DNA sequence itself [Barolo, 2016] and its cell-type specific context (e.g. DNA accessibility and methylation, and presence of other bound TFs) [Arvey et al., 2012]. Computational models to predict TFBSs and their cell-type specificity from sequence alone are of high importance in genomics, computational and molecular biology as they would enable fast screening of personal genomes for non-coding regulatory mutations, among other benefits. In some cases, these models may also suggest biological factors and mechanisms that may impact binding, which could then be tested experimentally.

Humans have $\sim$2000 TFs, each with its own complex affinity to DNA [Brivanlou and Darnell, 2002]. TFs bind to specific 6-20np DNA motifs that may be located either proximally (promoters) or distally (enhancers) from the gene they regulate [Zambelli et al., 2012]. Classically, these motifs are represented as consensus sequences or position weight matrices (PWMs), which can be used to weakly discriminate between bound and unbound sites w.r.t. a given TF [Stormo, 2000, Stormo et al., 1982]. However, these simple motif-based models ignore sequence context dependencies associated with co-operative binding of TFs or nucleosome positioning. Machine-learning-based models attempt to predict a TF's binding sites by considering a broader region (e.g. 100-200 bp regions obtained by ChIP-Seq experiments) surrounding a candidate site. Shallow ML-based models (e.g. string-kernel SVM [Arvey et al., 2012], gapped kernel SVM [Ghandi et al., 2014],

and ensemble approaches [Sharmin et al., 2016]) represent candidate DNA sequences as sets of k-mers, which allows to capture some interactions among motifs. More recently, deep learning approaches based on convolutional and recurrent neural network (CNN, RNN) models improved on shallow models by capturing wide range of motif features from the sequences [Alipanahi et al., 2015, Zhou and Troyanskaya, 2015, Zeng et al., 2016]. However, current approaches are far from having sufficient accuracy to replace laboratory experiments, as shown by a recent DREAM competition [Kundaje et al., 2021].

One approach to improving existing approaches for sequence function prediction tasks is to exploit evolutionary information. Because of selective pressure, functional regions tend to be more conserved across species than surrounding non-functional regions [Hardison, 2003, Chen et al., 2007]. Phylogenetic footprinting approaches have successfully taken advantage of this phenomenon to identify non-coding functional elements [Tagle et al., 1988, Blanchette and Tompa, 2002, Moses et al., 2004, Loots and Ovcharenko, 2004, Boffelli et al., 2003]. However, the assumption that TFBSs are conserved across species is not completely accurate, as shown in multiple studies [Moses et al., 2006, Lawrie et al., 2011, Li et al., 2008a, Halfon et al., 2011, Wang et al., 2016]. This is particularly due to a phenomenon called binding site turnover, where a binding site for a given TF may become replaced by another functionally analogous binding site (either for the same TF or for a different one) located nearby [Sinha and Siggia, 2005, Moses et al., 2006]. More sophisticated phylogenetic footprinting models are more robust to this type of event and report accuracy gains in both yeast [Hawkins et al., 2009] and human [Blanchette, 2012]. Although individual binding sites are often lost due to mutations, selective pressure often results in the overall function of a promoter or enhancer to remain conserved over long evolutionary periods [Shelest and Wingender, 2005]. Based on this criterion, methods have been developed that use the high-level feature information across different species rather than simply the low-level sequence conservation. This includes the idea that TFBSs often occur as dense clusters in the regulatory regions and various studies make use of presence of densely clustered binding sites in conserved regions to predict TFBSs in

Drosophila [Berman et al., 2004, Ross et al., 2018, Crocker et al., 2015], mouse [Boulling et al., 2013] and human [Blanchette et al., 2006].

Deep neural network models require a sufficiently large number of training examples to perform well. However, the number of positive examples (binding sites) for a given TF in a given human cell type typically ranges from 10,000 to 100,000, which is insufficient to take full advantage of the power of these sophisticated models. On the other hand, binding sites from the genome of interest (e.g. human) have orthologs in many other mammals. Those can be identified using whole-genome alignments [Kent et al., 2002] and ancestral genomes can be reconstructed with high accuracy [Blanchette et al., 2004a, Paten et al., 2008, Diallo et al., 2009, 2007, Westesson et al., 2012]. Thus, the use of evolutionary information such as orthologous regions in other species offers the possibility to augment the training data. However, this data is mostly unlabelled due to a lack of wet-lab experimental results in most species. In addition, orthologous and ancestral sequences are related through a phylogenetic tree, thus, violating the i.i.d. assumptions of many supervised ML models.

In this paper, we introduce a phylogenetic regularization approach called PhyloReg, to improve the training of deep learning TFBS predictors by making use of evolutionary data to reduce overfitting. PhyloReg is a semi-supervised ML approach based on manifold regularization [Belkin et al., 2006]. It is well suited for scenarios where relatively few labelled examples but abundant unlabelled examples (orthologous regions in extant and ancestral species) are available. We demonstrate with simulated and real data that PhyloReg consistently produces models that are more accurate than those based on single-species data. In addition, PhyloReg is applicable to any sequence-based function predictor that optimizes a continuous loss function (e.g. all CNN and recurrent neural networks, and many k-mer based models).

2.4 Methods

2.4.1 Phylogenetic Regularization

In general, machine learning models perform better when trained with large amounts of data. Our phylogenetic regularization approach aims to augment the data used to train TFBS prediction models in humans with orthologous sequences from other modern mammals and ancestors. Although, the labels of such orthologs are generally unknown, PhyloReg relies on the assumption that orthologs and ancestors from closely related species (according to a given phylogenetic tree) tend to share similar functional properties. Exploiting this concept, and the fact that for mammals orthologous and ancestral sequences can accurately be identified/inferred (see below), PhyloReg is a regularization approach that encourages a model to make predictions that are phylogenetically consistent, i.e. that do not (or rarely) change too drastically between neighboring species.

Concretely, a regularization term is added to the training objective of the supervised learning algorithm of our choice. Let $L(\hat{y}, y)$ be a function that measures prediction error (e.g., cross-entropy or mean squared error) between the predicted and observed target values. PhyloReg augments this loss function as follows:

$$\frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y_i) \;\; + \beta \cdot \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{|E^i|}\sum_{e\in E^i}(f(e^p) - f(e^c))^2\right) \tag{2.1}$$

where we have a set of N labeled examples $\{(x_i, y_i)\}_1^N$, with sequence x_i and label y_i. For each example x_i, we have a phylogenetic tree containing x_i as a leaf, made of a set of branches E^i connecting x_i to its unlabelled orthologous and ancestral sequences. e^p and e^c are the sequences associated to the parent and child nodes of edge e. Finally, f is a learned model and β controls the importance of the phylogenetic regularization term.

The key notion of equation (2.1) is to shape the model's predictions based on the phylogenetic manifold. The first term uses the labelled examples to train the model to get a good fit, while the second term encourages that similar predictions be made on orthologs.

A key benefit of PhyloReg is that the number of unlabelled ortologous examples can be much larger than the number of labelled examples, leading to a biologically-principled data augmentation strategy.

2.4.2 PhyloReg implementation

In our study, we implemented the PhyloReg loss term for a regression task on a simulated dataset and a classification task on TFBSs dataset. We observed that the optimizing the loss terms in equation 2.1 became easier if the phylogenetic loss term is considered in the loss function only at a certain epoch c rather than at every epoch. This helps with both running time and convergence. In our experiments, we tried c from $[1, 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50]$ and selected $c = 25$ based on accuracy and training time.

For the prediction of TFBS, we use a CNN-RNN hybrid architecture recently proposed by Quang and Xie [2019] called FactorNet that came as one of the best performing sequence-based model in a recent DREAM competition [Kundaje et al., 2021]. Although, FactorNet can incorporate non-sequence information as well, in this study we used the FactorNet architecture that makes use of solely sequence information. FactorNet takes a sequence and its reverse complement as input. The first layer is a convolution layer with 32 filters of size 26 and ReLU activation layer. Then a dropout layer of $p = 0.1$ is used. The convolution layer is followed by a max pool layer of filter size 13. The output from these layers is passed through a single bidirectional LSTM layer of hidden size 32. Again, a dropout layer of $p = 0.5$ is used. Then, the output is passed through a fully connected layer of size 128 with ReLu activation function, which is followed by a last dropout layer of $p = 0.5$. Finally, a fully connected layer of size 1 with sigmoid activation function yields the output. The final output of FactorNet is the mean of the outputs of the sequence and its reverse complement. The FactorNet models are trained batch-wise with batch size of 128 labelled examples along with their orthologs (in case of PhyloReg version). A validation set with early stopping is used to

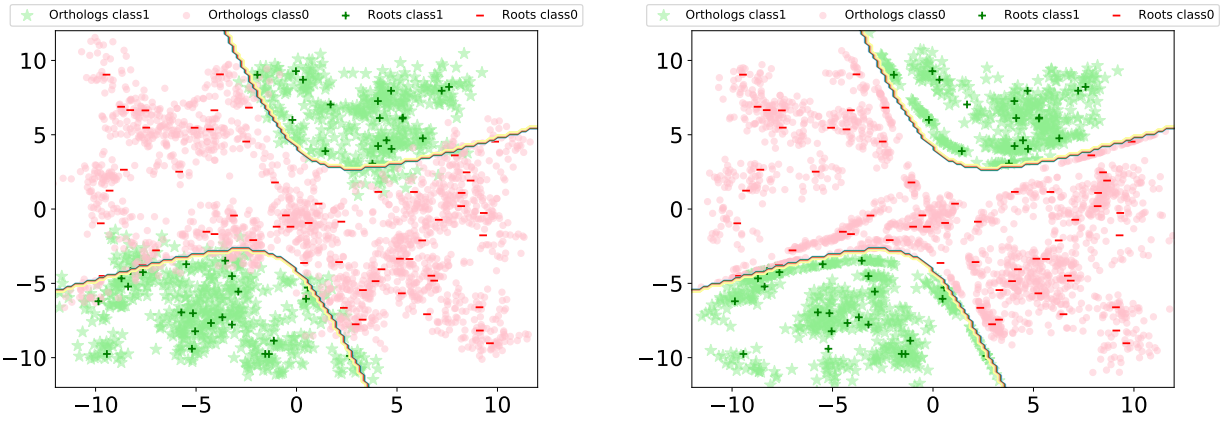
avoid overfitting. Weights are learned by minimizing the equation (2.1) using Adam optimizer [Kingma and Ba, 2014] with its default parameters. The models used in our study are implemented using pytorch package [Paszke et al., 2017]. The code is available at `https://github.com/BlanchetteLab/PhyloReg` and the supplementary materials are available at `https://github.com/BlanchetteLab/PhyloReg/tree/master/Supplementary`.

2.4.3 Datasets

We use a dataset based on 422 ChIP-Seq experiments, originally assembled by Zeng *et al.* Zeng et al. [2016], who used it to train and assess the CNN model they proposed, available at `http://cnn.csail.mit.edu/motif\_occupancy/`, originally produced by the Encode project Consortium et al. [2012], to evaluate the performance of the PhyloReg approach. Each positive example is a 101 bp region centered on a ChIP-seq peak for a given transcription factor in a give human cell type. For each TF, Zeng *et al.* randomly selected as negative examples an equal number of genomic regions, ensuring to match the GC-content and motif-binding affinity of the positive examples. In this study, we extended each sequence to 1000 bp by adding the neighboring 450 bp on both sides of a given sequence. We elongated the sequences because the original FactorNet is shown to perform well on 1000 bp input and the binding sites in the orthologs may have altered its position from the ChIP-Seq peak position in human.

We use blat [Kent, 2002] and liftover to map regions from the hg19 assembly to the hg38 assembly. The examples with overlapping regions (even by 1 bp) within each experiment are excluded using BEDOPS [Neph et al., 2012]. We randomly divided each data set in 4:1 ratio to create training and test sets. A portion of the training set of size min(500, $0.1 \times$ train_size) was further set aside as validation set, to detect and avoid overfitting.

We extracted mammalian orthologous regions for each training and test examples using a 100-way vertebrate whole-genome alignment available from the UCSC genome browser [Kent et al., 2002, Blanchette et al., 2004b]. We complemented those sequences



(a) Selection Pressure = 0, Mutation Rate = 1.

(b) Selection Pressure = 100, Mutation Rate = 1.

Figure 2.1: Example of a two-dimensional simulated data set. 100 points (marked with plus and minus signs) are chosen to initialize the evolutionary process. The orthologs evolve randomly along the branches of a complete binary tree of depth 5, producing a set of 31 points. The activity function is $A(x_1, x_2) = \sigma(x_1^2 \cdot w_1 + x_1 \cdot x_2 \cdot w_2 + x_2^2 \cdot w_3 + w_4)$. (a) Data set generated using a selection coefficient of zero. Orthologs cluster around their ancestor. (b) At higher selecion coefficient (100), orthologs tend to stay on the same contour line as their ancestor. This effect is most notable in portions of the space with a strong gradient, (e.g. near the black curves).

with computationally predicted ancestral sequences produced by Ancestor1.0 ([Diallo et al., 2009]). The resulting orthologous and ancestral sequences may not all be of exactly 1000 bp. When needed sequences were truncated or zero-padded to yield sequences of fixed size. We ignore the orthologs sequences of size < 500 bp. The average number of examples in ChIP-seq experiments used in this study is 65,000 with a minimum of 600 and a maximum of 246,266. Each human sequence has on average 80 orthologous/ancestral examples.

2.4.4 Phylogenetic Simulation

To evaluate the effectiveness of PhyloReg, we use a simple simulation process that generates a set of orthologous feature vectors, evolved along the branches of a given phy-

logenetic tree (in our case, a complete binary tree with equal branch lengths). Evolution of feature vectors is constrained by selective pressure (with selection coefficient S) on the value of a user-defined function we call the *activity* function, which tends to be preserved during evolution. The activity of a feature vector x aims to mimic the level of biological activity (e.g. TF binding affinity) of x. Briefly, the simulation procedure starts by selecting a random feature vector x_r at the root of the tree. It then evolves this vector along the branches of T by perturbing each feature value with independent Gaussian noise with standard deviation m (herein called the mutation rate), and accepting or rejecting the change based on the difference in the value of the activity function of the proposed descendant x_c compared to that of the parent x_p:

$$\Pr_{accept}(x_c|x_p) \quad = \quad \frac{\text{Norm}(S \cdot (\text{act}(\text{x}_p) - \text{act}(\text{x}_c)); \mu = 0, \sigma^2 = 1)}{\text{Norm}(0; \mu = 0, \sigma^2 = 1)}.$$

Hence if the product of the change in activity and the selection coefficient is large, it is unlikely to be accepted. If the proposed change is rejected, another perturbation is attempted on x_p, until the change is accepted. The resulting set of orthologous feature vectors and their activity values can then be used to assess the impact of phylogenetic regularization on prediction accuracy, depending on various simulation parameters. An example of two simulated data sets is shown in Figure 2.1.

2.5 Results

In this section, we first present the results obtained with PhyloReg on both artificial and real data sets, and compare them to other approaches.

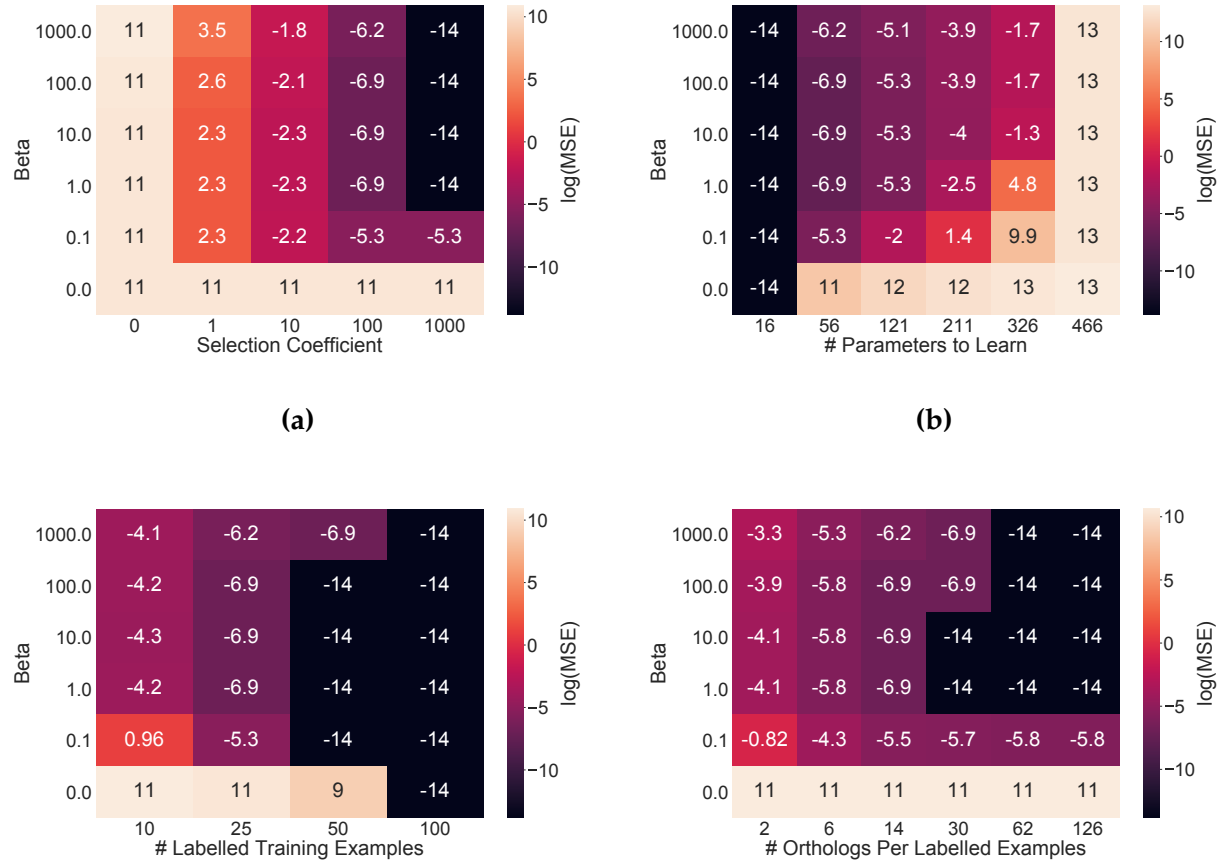


Figure 2.2: Mean-squared error (MSE) on test data, for simulated data sets with default values as follows: Number of labeled examples $n = 25$; Number of unlabeled orthologs per training example $m = 14$; Selection coefficient $S = 100$; dimensionality (d=10), which yields $\binom{d}{2}+1$ weight parameters to be estimated. The results reported are the median over 100 repetitions. For each panel, different values of regularization weight β are considered, and one of the experiment's parameters is varied, leaving the others to their default value.

2.5.1 Results on simulated data

To assess the ability of phylogenetic regularization to help with generalization and reduce prediction error, we designed a simulation study where an activity function A is a quadratic function of a set of d variables $X = x_1, ..., x_d : A(X) = sigmoid(\sum_i \sum_j w_{i,j} x_i x_j + w_0)$. In the simulation, vector X is initialized randomly at the root of a complete binary phylogenetic tree, and then evolves randomly along the branches of the tree according

to a Brownian motion, but subject to selective pressure to maintain an activation value similar to that of the parent node (see Methods). The value of $w_{i,j}$ are chosen randomly ahead of the simulation and remain fixed. See Methods for details. Hence, at high values of the selection coefficient, examples tend to evolve along the contour lines of the activity function (see Fig 2.1). This process mimics natural selection on a biological process encoded by a set of genetically defined variable X, where there is selective pressure on each example to have an activity that is similar to its parent. For example, a given regulatory region may be under selective pressure to maintain a given affinity for a certain transcription factor, or a gene's promoter might be under selection to yield expression near a given optimal value (neither too high nor too low).

This evolutionary process is repeated n times, every time resulting in a group $O^i = X_1^i, ..., X_m^i$ of m orthologous vectors (one for each node of the tree), and a single real-valued activity value y_i associated with a pre-selected leaf of the tree (mimicking the reference species from which experimental data is available). We then train a logistic regression model to estimate the $\binom{d}{2} + 1$ weight parameters, using a loss function that includes the phylogenetic regularization term, weighted by regularization weight β. Finally, the model's performance is evaluated using the mean-squared error (MSE) on left-out data.

We used our simulation data to assess the benefits of phylogenetic regularization in the context of different parameters, including the selection coefficient (S), the dimensionality of the feature space ($\binom{d}{2} + 1$), the number of training examples (n), and the number of orthologs (m) per labeled example. Note that setting $\beta = 0$ yields a standard regression model trained on the n labeled examples, which is our baseline. Additionally, we added a ℓ_2 penalty term to the loss term and observed a decrease in model performance. Therefore, we excluded the ℓ_2 penalty term in the experiments with the simulated data. Fig. 2.2 (a) shows that phylogenetic regularization is able to take advantage of orthologous data, compared to a standard regression model ($\beta = 0$), which, based on the small number of training examples available, is unable to generalize. Furthermore, the impact of phylogenetic regularization increases with the selection coefficient, and the ideal

choice of phylogenetic regularization weight β depends on that coefficient. Fig. 2.2 (b) illustrates that when dealing with a limited number of labeled examples, phylogenetic regularization enables learning in higher dimensional spaces. For example, with phylogenetic regularization and under strong selection coefficients, 25 labeled examples suffice to accurately learn models with up to $\binom{25}{2} + 1 = 326$ parameters. Similarly (Fig. 2.2 (c)), at a fixed dimensionality, the number of labeled examples needed to learn an accurate model is much lower with phylogenetic regularization than without. Finally, the benefits of phylogenetic regularization increase with the size of the tree (Fig. 2.2 (d)).

2.5.2 Experiments with real datasets

Phylogenetic regularization is applicable to any supervised ML algorithm that explicitly minimizes some loss function. The experiments with simulated data validated phylogenetic regularization assumptions in a regression problem setting. Here, we elected to test our approach on the well studied binary classification problem of transcription factor binding site prediction. We used 422 ChIP-seq data sets (106 TFs across 88 cell types) from the Encode project [Consortium et al., 2012] that were used in previous publications and available at `http://cnn.csail.mit.edu/motif\_occupancy/` [Zeng et al., 2016]. Positive examples are 1000 bp regions centered around ChIP-seq peaks, and negative examples are the regions not bound by the TF in the same cell type with similar GC content and motif strength as that of the positive examples.

We trained several FactorNet models with the architecture described in Section 2.4, taken from Quang and Xie [2019]. The first is the standard FactorNet model without phylogenetic regularization, trained on the human data alone. The second, called PhyloFactorNet, uses phylogenetic regularization with $\beta = 1000$. The choice of $\beta = 1000$ was made empirically selected from $[0.1, 1, 10, 100, 1000, 10000]$ based on a simple grid search of a separate validation subsets. Phylogenetic regularization was based on either all mammals (PhyloFactorNet (mammals)) or only primates (PhyloFactorNet (primates)).

Since inter-species sequence conservation has long been used as a guide toward functional regions, we trained a model called FactorNet+PhastCons, which uses the aforementioned FactorNet model trained on human data only, and combined its output with PhastCons conservation score computed across vertebrates [Siepel et al., 2005, Siepel and Haussler, 2005], using a simple logistic regression trained on each data set. We also considered the PhastCons score alone as a predictor.

For each of the 422 ChIP-seq data sets, all models were trained on the same train/test splits and were evaluated using the Area Under the receiving-operating curve (AUC) scores on the test data. Furthermore, we computed Recall scores at different False Discovery Rates for the improvement of PhyloFactorNet (mammals) over PhyloFactorNet (human). The evaluation metrics are computed using scikit-learn [Pedregosa et al., 2011a].

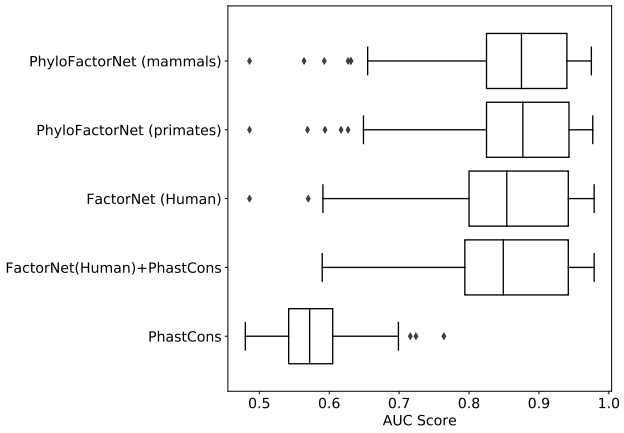


Figure 2.3: Test AUC scores of 8 different types of predictors, across the 422 ChIP-Seq experiments used in this study.

Figure 2.3 show the distribution of AUC values obtained for the different models. PhyloFactorNet(mammals) outperfoms the FactorNet(human) by 2% (median AUCs)(p-value: 2.08×10^{-27}; Wilcox signed-rank test); a very significant margin considering that recent papers in the field rarely exhibit gains of more than a couple percent over predecessors. An improvement in prediction accuracy is observed across 289 (PhyloFactorNet(primates)) and 278 (PhyloFactorNet(mammals)) of the 422 data sets. It should be noted that the improvement from PhyloReg is observed with large magnitudes in smaller

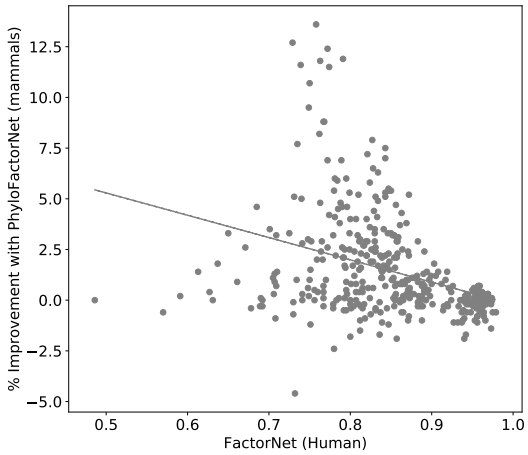


Figure 2.4: AUC score improvement in percentages over FactorNet (human) using PhyloFactorNet (mammals) on 422 ChIP-Seq data sets. X-axis: AUC score of FactorNet trained on human data alone; Y-axis: FactorNet predictor trained using phylogenetic regularization.

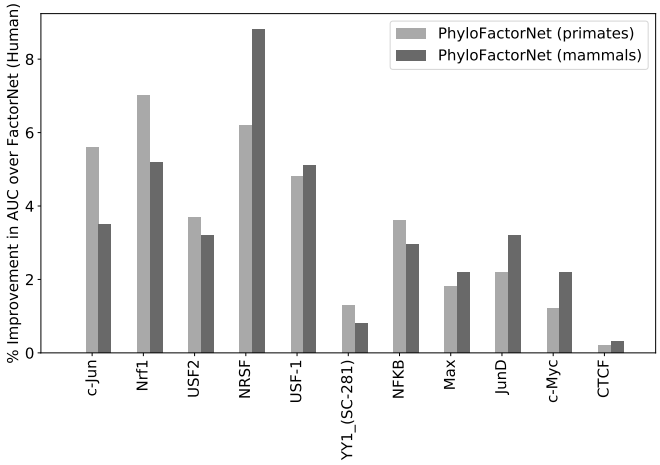


Figure 2.5: Median percentage improvement in AUC score over FactorNet (Human) using PhyloFactorNet (primates) and PhyloFactorNet (mammals) for the TFs present in at least 5 of the 422 ChIP-Seq experiments and train size $>= 10,000$.

datasets as shown in the Supplementary Figure 2.7. Adding the PhastCons score to the FactorNet(human) predictor does not significantly improve the predictions.

As visible in Figure 2.4, the gains provided by the phylogenetic regularization is most notable for TFs for which the FactorNet(human) accuracy is moderate, suggesting that our approach is a good way to improve prediction accuracy for TFs for which the de-

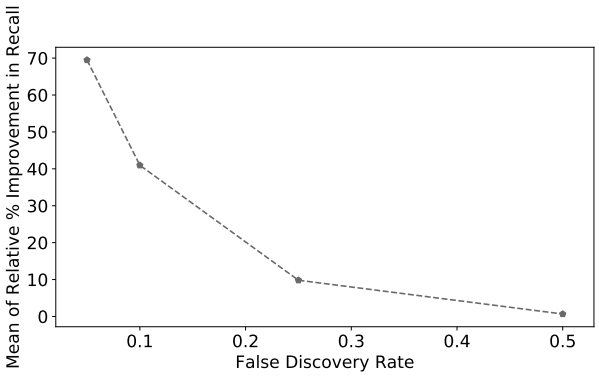


Figure 2.6: Mean of relative percentage improvement in Recall score with PhyloFactorNet (mammals) over FactorNet (human).

terminants of binding are complex (e.g. those dependent on co-factors or nucleosome positioning). Figure 2.5 shows the AUC gain for the TFs. For several TFs (e.g. Nuclear Respiratory Factor1 (Nrf1) and Neuron-Restrictive Silencer Factor (NRSF)) the gains provided by PhyloFactorNet exceed 5%. Notably, on this subset of datasets, the number of examples in the test set is sufficiently large, which provides a low-variance estimate of the true AUC.

One should keep in mind that in a classification problem where negative examples actually outnumber positive examples by approximately 1000-to-1 (as they do in the human genome, for a typical TF), even setting a relatively low false positive rate threshold (e.g. 10^{-2}) would yield a high false discovery rate (FDR) (approximately 90%). Thus, to estimate the performance of a predictor, one should pay particular attention to the Recall score in very low FDR regimes. Figure 2.6 shows that the Recall gains obtained by PhyloFactorNet approach is particularly strong in those range, with relative gains of more than 70% at FDR=0.05.

2.5.3 Species informativity

To further understand when and how phylogenetic regularization is particularly powerful, we applied the PhyloFactorNet(mammals) trained models to obtain prediction scores

for each ortholog and ancestral sequence of each data set. We then measured the extent to which the prediction score obtained on sequences from a given species/ancestor are informative for the prediction of the label of the corresponding human sequence, effectively asking the question: how well can the human label be predicted by looking only at the ortholog's sequence in species X? Focusing on the K562, GM12878, cell types, Supplementary Figures 2.8, 2.9, show that, unsurprisingly, species closely related to human (particularly old-world monkeys) generally yield most accurate predictions of human labels. However, for many transcription factors/cell types, similar performances are obtained using much more distantly related mammals. For example, the binding of transcription factors such as NF-YA, Elk1, Nrf1, and ATF3 in human is predicted with nearly as good accuracy using orthologous data from any mammalian species. We note that these TFs tend to exhibit relatively ubiquitous expression across cell types. In contrast, for transcription factors such as CEBPB, NF-E2, PU.1, MafK, and MafF, non-primate species yield predictions that are much less accurate than those obtained on human. Remarkably, most of these transcription factors are bZip transcription factors active in the erythroid lineage (which K562 cells belong to). Similar observations are made in GM12878, H1-hESC, and HepG2 (see Supplementary material), with cell-type specific TFs (e.g. NFKB in GM12878, derived from B-lymphocytes) exhibiting reduced prediction power in non-primate species. We speculate that the binding sites of some of these TFs may be under weaker selection, which would reduce the benefits offered by PhyloReg.

2.6 Discussion and Conclusion

We present PhyloReg, a semi-supervised approach for TFBSs predictions that complements labeled sequence data in a given species of interest (e.g. human) with a potentially very large number of readily available orthologous and ancestral sequences. PhyloReg takes advantage of the selective pressure on regulatory regions to enable models that generalize better than those trained on human data alone. We report substantial improve-

ments obtained by using phylogenetic regularization to improve the training of a recently published machine learning approach based on convolutional neural networks for transcription factor binding prediction. PhyloReg enable one to increase by more than 2% the performance (AUC) of a state-of-the-art FactorNet neural network.

Although the main application featured here is a classification task, we demonstrate using simulated data that the approach is also applicable to regression tasks, in a context where selective pressure results in a conservation of the activity level of a given sequence. Remarkably, in this setting, PhyloReg enables learning complex predictive models from high-dimensional data from significantly fewer training examples than would be needed if only labeled examples were used.

Although this work did not explore questions related to model interpretability, this is an important area of research, especially for bioinformatics applications of ML. We speculate that phylogenetic regularization will help in this direction, by providing predictors that are not only more accurate, but also more evolutionarily robust, and hence probably better disentangled models.

Another exciting direction is to design automated approaches to learn to weigh the changes in values of f along the different branches of our phylogenetic tree. For example, it may be that certain species may be much less relevant than others in studying certain types of functions (binding of specific transcription factors, or other types of functions), e.g. because of different conditions that species is exposed to, or because of genetic differences (e.g. the transcription factor may not even exist in that genome, or may not be expressed in that cell type). The approach presented in Supplementary Figures 2.8, 2.9, 2.10 and 2.11 are a step in that direction, but more sophisticated approaches may be beneficial and informative from an evolution standpoint.

A major benefit of phylogenetic regularization is that, in theory, it is easily applicable to any sequence-to-function prediction task where an ML model explicitly aims to minimizes some loss function (e.g. most neural network based approaches). This family of problems includes many other types of interaction prediction tasks (microRNA binding

target sites, binding sites of RNA-binding proteins to RNA, protein-protein interactions, nucleosome localization, and protein post-translational modifications, etc.), as well as higher-level functions such as control of gene expression levels by regulatory regions, mRNA splicing and stability. Notably, machine learning models have already been proposed for each of these tasks, and more. PhyloReg offers the potential benefit of boosting the accuracy of these approaches at very little extra cost.

Acknowledgement

Computational resources for the project were made available to our group thanks to a Compute Canada Resource Allocation to MB. The author will like to thank Blanchette Lab members Zichao Yan, Mohammad Nikou Sefat, Amaury Leroy, Elliot Layne, Samy Coulombe, Dongjoon Lim, Christopher JF Cameron, Ayrin Ahia-Tabibi and Rola Dali for the support and constructive discussions.

Supplementary Materials

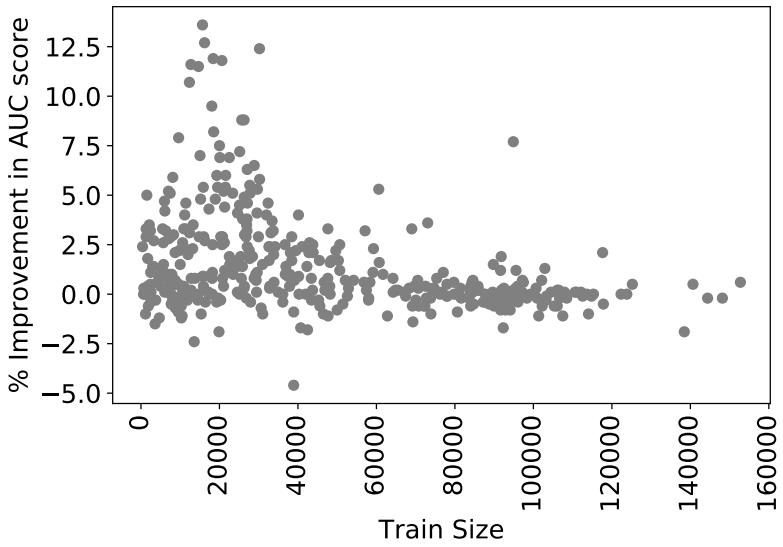


Figure 2.7: Percentage improvement in AUC score over FactorNet (human) using Phylo-FactorNet (mammals) in the 422 ChIP-Seq experiments.

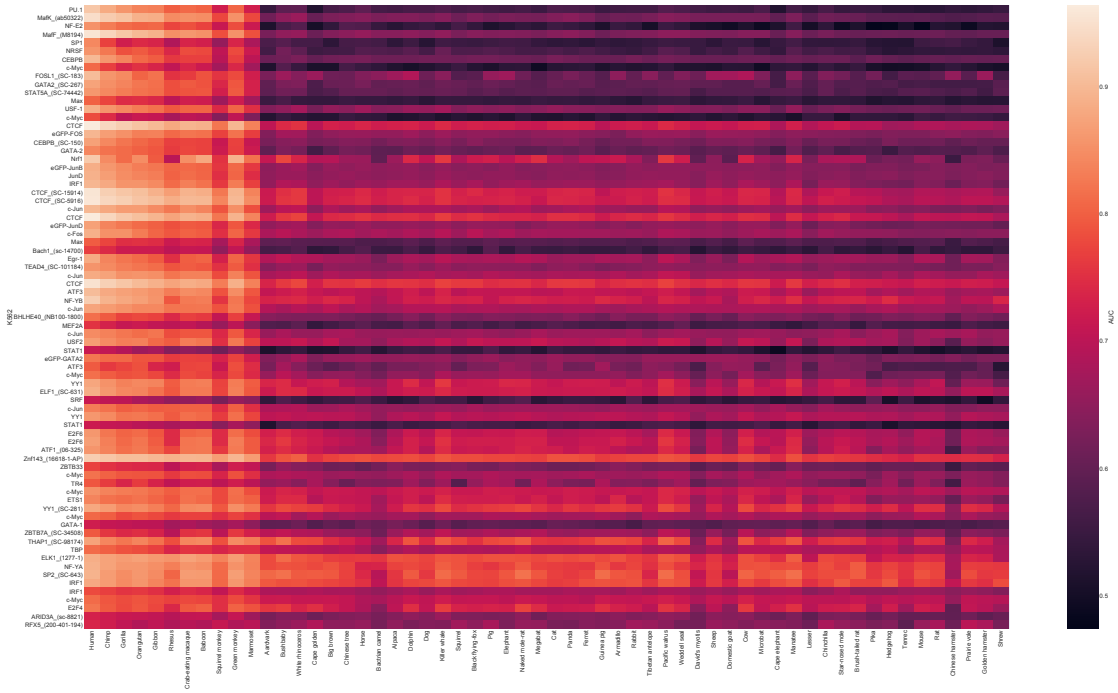


Figure 2.8: The test AUC scores obtained by using the score of the trained predictor applied to individual orthologous sequence, against the label of the human ortholog in K562 cells.

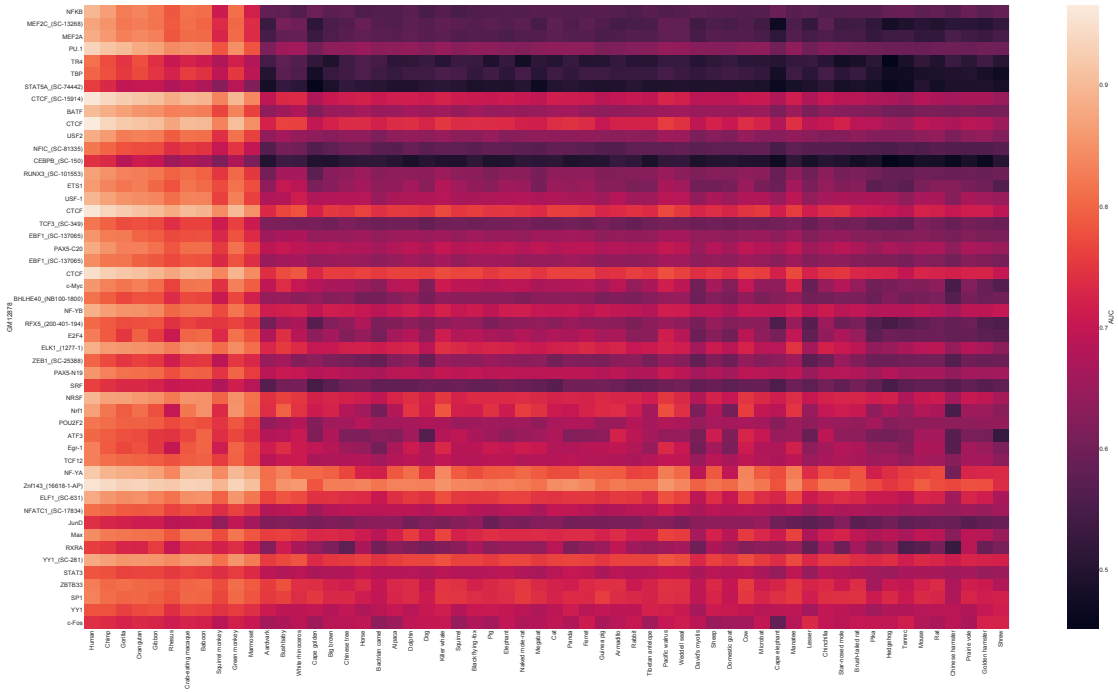


Figure 2.9: The test AUC scores obtained by using the score of the trained predictor applied to individual orthologous sequence, against the label of the human ortholog in GM12878 cells.

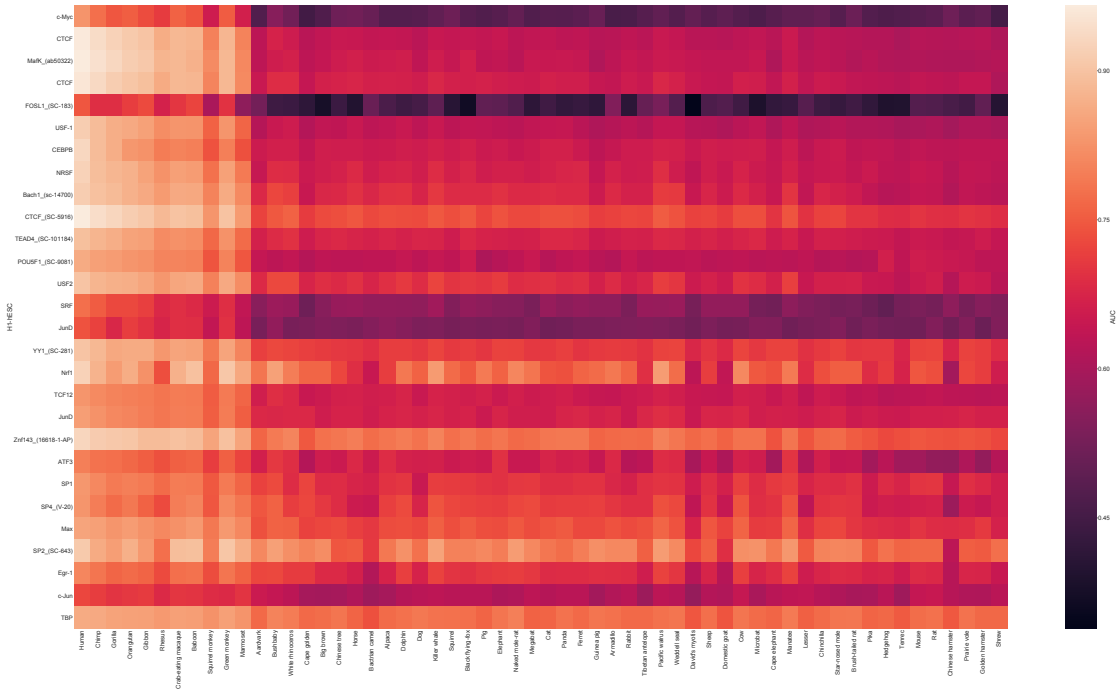


Figure 2.10: The test AUC scores obtained by using the score of the trained predictor applied to individual orthologous sequence, against the label of the human ortholog in H1-hESC cells.

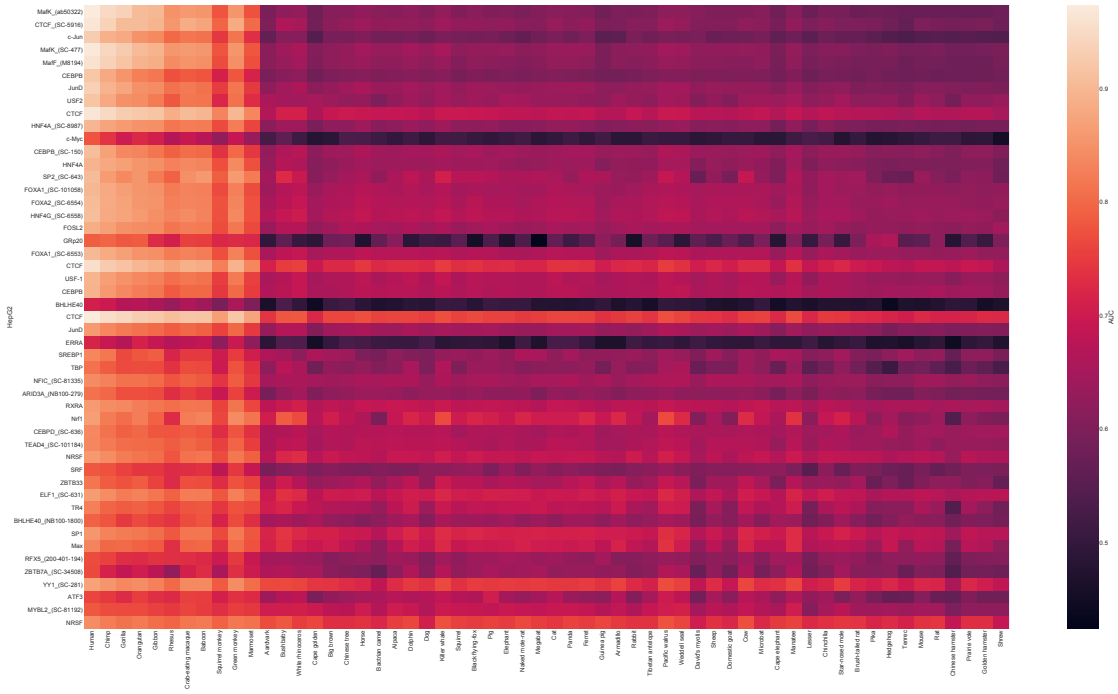


Figure 2.11: The test AUC scores obtained by using the score of the trained predictor applied to individual orthologous sequence, against the label of the human ortholog in HepG2 cells.

Chapter 3

Inferring Regulatory Function Using Evolutionary Data

In the following manuscript, I propose a learning algorithm, PhyloPGM, to boost the accuracy of previously trained transcription factor (TF) and RNA binding protein (RBP) binding site predictors. PhyloPGM assumes that the regulatory function is maintained in the orthologous regions and is applicable at the inference stage of a machine learning pipeline. First, I process and compile previously published transcription factor binding site (TFBS) and RBP binding site datasets. I extract the corresponding orthologs/ancestors sequences. I apply previously trained TF and RBP binding site predictors on the datasets to obtain orthologous prediction scores. These scores are used in the PhyloPGM pipeline to obtain the PhyloPGM score. I show that PhyloPGM improves the accuracy of base models and the amount of improvement is larger for the cases where base models have relatively lower accuracy. Then, I process and compile datasets of human genomic regions where mutations are linked to adverse effects. I find that PhyloPGM is more helpful than base model or sequence conservation approach to identify the disease-causing human non-coding variants. I observe that different branches in the phylogenetic tree have different impact on PhyloPGM results.

Citation: Faizy Ahsan, Zichao Yan, Doina Precup, and Mathieu Blanchette (2021) PhyloPGM: Boosting Regulatory Function Prediction Accuracy Using Evolutionary Information. *In preparation for submission to Nature Methods.*

3.1 Manuscript 2: PhyloPGM: Boosting Regulatory Function Prediction Accuracy Using Evolutionary Information

Authors: Faizy Ahsan[1], Zichao Yan[1], Doina Precup[1], and Mathieu Blanchette[1]

[1]School of Computer Science, McGill University, Montreal, Quebec, Canada

3.2 Abstract

The computational prediction of a function associated with a genomic sequence is of utter importance in -omics study. The binding prediction of transcription factor (TF) to a regulatory region will allow to understand the gene regulatory mechanism. The binding prediction of RNA and RNA binding protein will allow to comprehend the post-transcriptional gene expression. However, the existing computational methods for transcription factor binding sites (TFBSs) and RNA-RBP binding proteins suffer from high false positive rates and seldom use the evolutionary information. The vast amount of available orthologous data across multitudes of extant and ancestral genomes present an opportunity to improve the accuracy of existing computational methods.

We present a novel probabilistic approach called PhyloPGM that aggregates the prediction scores of a previously trained TFBS predictor or RNA-RBP binding predictors on orthologous regions to boost their respective prediction accuracy. PhyloPGM significantly improves the prediction accuracy of RNATracker, a sequence-based RNA-RBP binding

predictor and of Factornet, a sequence-based TFBS predictor. PhyloPGM is simple to implement, yet, yields remarkable results.

3.3 Introduction

The binding of transcription factors (TFs) to specific DNA regions, determines the gene regulatory network. The TF binding is determined by both the presence of motifs, specific $\sim$6-20 bps patterns in the DNA sequence and the cell-type specific context such as DNA accessibility and methylation, and presence of other bound TFs) [Slattery et al., 2014]. The sequence and cell-type specific interaction between proteins and either DNA or RNA drives both transcriptional and post-transcriptional regulation [Stefl et al., 2005]. Some representative examples include RNA splicing that prepares the nascent RNA transcript for maturation, and the subsequent localization which transports the messenger RNA (mRNA) to certain subcellular compartments where their products are needed. These regulatory processes are mediated by a diverse population of RNA binding proteins (RBPs), each having an affinity for a specific RNA motif, and aberrations from their usual interaction scheme with the RNAs are known to implicate a series of neurological disorders and possibly cancer [Lukong et al., 2008a]. Therefore, it is crucial to characterize the TF and the RBP binding specificity in order to comprehend the gene regulatory network, to scrutinize the associated disease pathways and possibly, to develop related therapeutic approaches.

The ChIP-Seq experiment is an *in-vivo* experiment that can identify the binding sites of one transcription factor in one cell type within a resolution of $\sim$200 bps [Johnson et al., 2007]. The ENCODE consortium has produced ChIP-Seq experiments data for hundreds of transcription factors in dozens of cell types [Consortium et al., 2012]. Similarly, the wet-lab experiments called CLIP-Seq (abbreviation of cross-linking immunoprecipitation and high throughput RNA sequencing;PARCLIP [Hafner et al., 2010a], HITSCLIP [Licatalosi et al., 2008a], ICLIP [Konig et al., 2010]) can identify the *in vivo* RNA binding to a

given RBP. In the CLIP-Seq experiment, a RBP and RNA are cross-linked with UV light, which is followed by lysing, immunoprecipation and sequencing. Although, the CLIP-Seq experiments yield a resolution of $\sim$100 bps in RNA that bound with the RBP, the exact location is not known, which is also true for the ChIP-Seq experiments. Moreover, it is impractical to conduct ChIP-Seq and CLIP-Seq experiments for each combination of TFs and cell-types and RNA and RBPs in order to characterize the binding specificity of the TFs and RBPs. Therefore, a computational method is required to predict the TFBSs and RNA-RBP binding to profile the sequence specificity of the TFs and RBPs.

The recent computational methods to predict TFBSs and RNA-RBP binding are heavily dominated with deep learning based approaches in terms of prediction accuracy e.g. convolutional neural networks [Alipanahi et al., 2015] or a hybrid of computational neural network and recurrent neural network [Pan et al., 2018]. In general, a DNA sequence of roughly 1000 bps or a RNA sequence of $\sim$100 bps is represented as a one-hot encoded tensor, which is then passed through the deep neural network of choice to predict whether the DNA sequence will bind to the TF or the DNA sequence will bind to the RBP of interest or not. Although, the deep learning approaches [Zhang et al., 2016, Quang and Xie, 2016, Pan and Shen, 2017, 2018] have outperformed the classical computational methods and shallow machine learning approaches [Hiller et al., 2006, Kazan et al., 2010, Li et al., 2010, Maticzka et al., 2014, Fukunaga et al., 2014, Pietrosanto et al., 2016] , they are often prone to high false positive rate and are yet to be established as wet-lab alternatives.

Due to the biological importance of TFBSs and RNA-RBP binding, the features present in the RNA sequence that allow to bind with the RBP or the regulatory regions to the TFs should be conserved during the evolution. Indeed, the sequence function across the orthologous regions in different organisms are observed to be conserved according to the orthologs conjecture [Shiraishi et al., 2001, Shabalina et al., 2004, Papatsenko et al., 2006, Cooper and Brown, 2008, Chen and Zhang, 2012, Stamboulian et al., 2020]. Intuitively, the sequence function conservation based approaches should yield a better model for RNA-RBP binding prediction. However, the conservation based approaches may suffer from

binding sites turnover phenomenon, where the number of binding sites in a given region is maintained, but the sequence itself is not conserved in the ortholgous regions [Sinha and Siggia, 2005, Moses et al., 2006]. Therefore, a more sophisticated operation is required to use the sequence function conservation property rather than the crude combination of sequence conservation score with the deep learning methods [Ahsan et al., 2020].

In this study, we present an aggregation approach called PhyloPGM which aims to boost the accuracy of a pre-trained base predictor for a specific type of function (see Figure 3.1). The base predictor is a machine learning predictor that assigns a prediction score (real number) to a given input sequence. In this paper, we use PhyloPGM for two types of functional prediction tasks: transcription factor (using FactorNet [Quang and Xie, 2019] as base predictor) and RNA-binding protein (using RNATracker [Yan et al., 2019]) occupancy prediction. To obtain a prediction on a given human sequence, the base predictor is first applied to that sequence and its orthologous regions from up to 58 other mammalian species as well as up to 57 computationally reconstructed ancestral sequences. PhyloPGM then aggregates the prediction scores using a phylogenetically-informed, probabilistic graphical model, essentially computing a likelihood ratio test contrasting the hypotheses that the human sequence is a positive (Y=1) or negative (Y=0) example. PhyloPGM takes advantage of the fact that selective pressure makes changes in sequence function rare. Hence, predictions made on orthologous and ancestral sequences are informative about the function of the given human sequence. In cases where the base predictor is relatively inaccurate, and where function changes are relatively rare, PhyloPGM could in principle use predictions made on orthologous/ancestral sequences to "correct" the prediction made on the human sequence, especially when the latter is borderline. Importantly, because PhyloPGM treats the base predictor as a black box (i.e. it does not need any information about the base predictor's inner workings), it is highly flexible and applicable to a wide variety of sequence function prediction problems for which the community has developed base predictors.

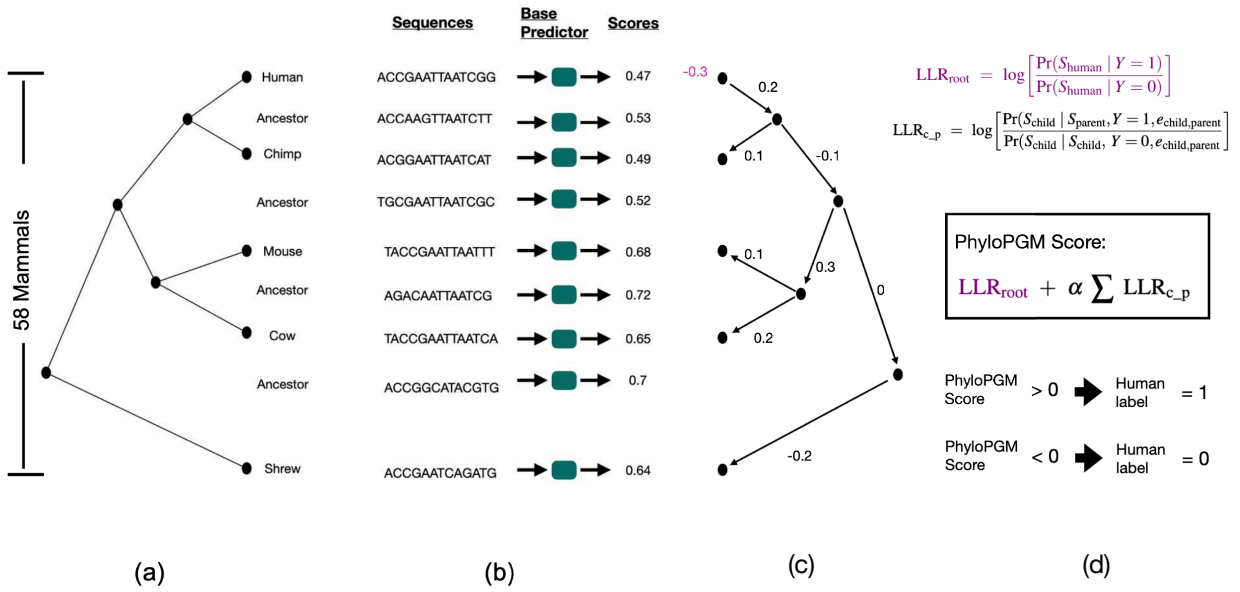


Figure 3.1: PhyloPGM workflow. (a) the phylogenetic tree, (b) the input human sequence and its orthologs are fed to trained base predictor in order to obtain the orthologous prediction scores, (c) Each branch weight denotes the log likelihood ratio of child score given parent score (LLR_{c_p} obtained by treating human as the root species and the human weight denotes the log likelihood ratio of root (LLR_{root}, (d) the equations used to compute the the log likelihood ratio and the PhyloPGM score. $e_{child,parent}$ is the evolutionary distance between child and parent species. The human sequence label is assigned 1 (binding site) if the PhyloPGM score ¿ 0, otherwise 0 (non-binding site).

The goal of PhyloPGM is to combine the prediction scores on a set of related orthologous and inferred ancestral sequences. Multi-Instance Learning (MIL) is a class of ML approaches that work under a similar setting i.e. the task is to classify a group of instances, termed as bag [Dietterich et al., 1997]. The MIL classifier labels the bag as positive if at least one of the instance is positive, otherwise negative. The classical MIL algorithms assume instances to be i.i.d., though there are MIL algorithms that handle non i.i.d. cases as well [Ping et al., 2010, Zhou et al., 2009]. Gao and Ruan [2015] used MIL with DNA structure data for *in vitro* TFBSs predictions in mouse without phylogenetic context. The MIL algorithms are extensively reviewed in Amores [2013], Foulds and Frank [2010]. Our requirement for the aggregating approach differs from the classical MIL in two principal

ways: (1) the instances are prediction scores obtained from a base predictor (rather than raw sequences or feature vectors); (2) the goal is to predict the label of a specific example from each bag (corresponding to the human sequence); (3) instances in the bag are not i.i.d. but are phylogenetically related through a known and fixed tree.

3.4 Results

In this section, we present and evaluate the results obtained using PhyloPGM for the tasks of binary TF and RBP occupancy prediction. In both cases, given a short (101-nt) DNA or RNA sequences), the goal is to predict whether a given TF or RBP would bind this sequence in a given cell type. The input sequence is much longer than the putative binding site itself, which provides important sequence context (e.g. for the presence of binding sites for co-factors, or structural RNA elements) to the base predictor.

3.4.1 PhyloPGM improves predictors' performance

We first applied PhyoPGM to the task of TF occupancy prediction, using FactorNet [Quang and Xie, 2019] as base predictor. FactorNet is a recently developed hybrid of convolutional and recurrent neural network architectures, which performed particularly well on a recent ENCODE-DREAM challenge [Kundaje et al., 2021]. We used a set of 13 ChIP-Seq datasets, obtained from the ENCODE-DREAM website `https://www.synapse.org/#!Synapse:syn6131484/wiki/402026`. The data sets originate from four different cell types and contain 56,700 to 423,218 positive examples and 50,356,411 to 51,164,150 negative examples (see Methods).

The performance of the predictors is evaluated on the test set provided by ENCODE-DREAM [Kundaje et al., 2021], using the Area Under the Precision-Recall curve (AUPR), which better reflects the true predictor's performance with imbalanced data sets, compared to the more traditional AUC score. Overall, PhyloPGM improves the AUPR scores of the FactorNet models by approximately 30% (FactorNet median test AUPR: 0.13, Phy-

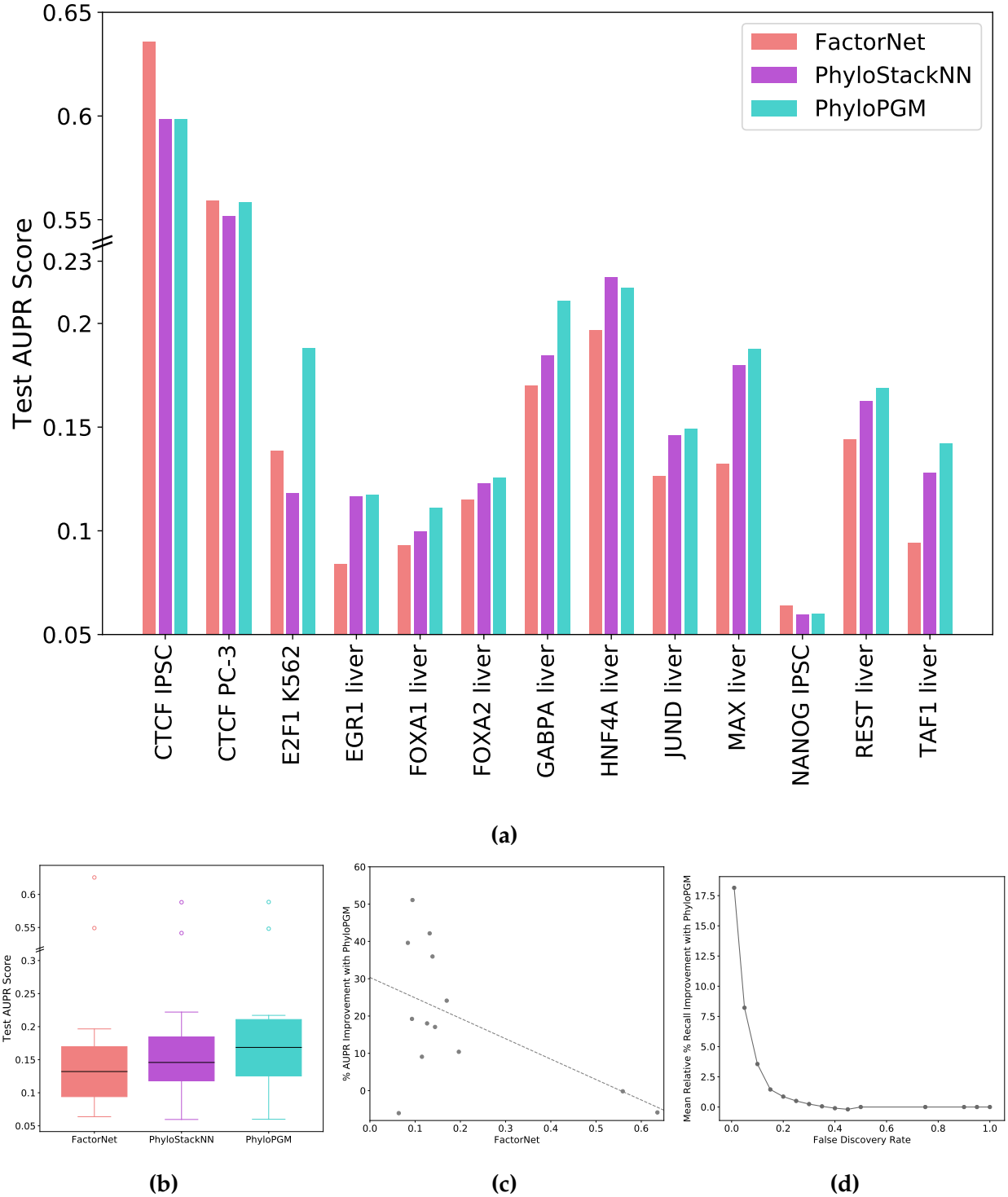


Figure 3.2: Model comparison for TFBS prediction problem. (a) Test AUPR scores of FactorNet, PhyloStackNN and PhyloPGM over 13 ChIP-Seq datasets. (b) Distribution of test AUPR. (c) Test AUPR improvement percentage of PhyloPGM over FactorNet. (d) Mean relative percentage improvement of PhyloPGM test recall score over FactorNet for different false discovery rate thresholds.

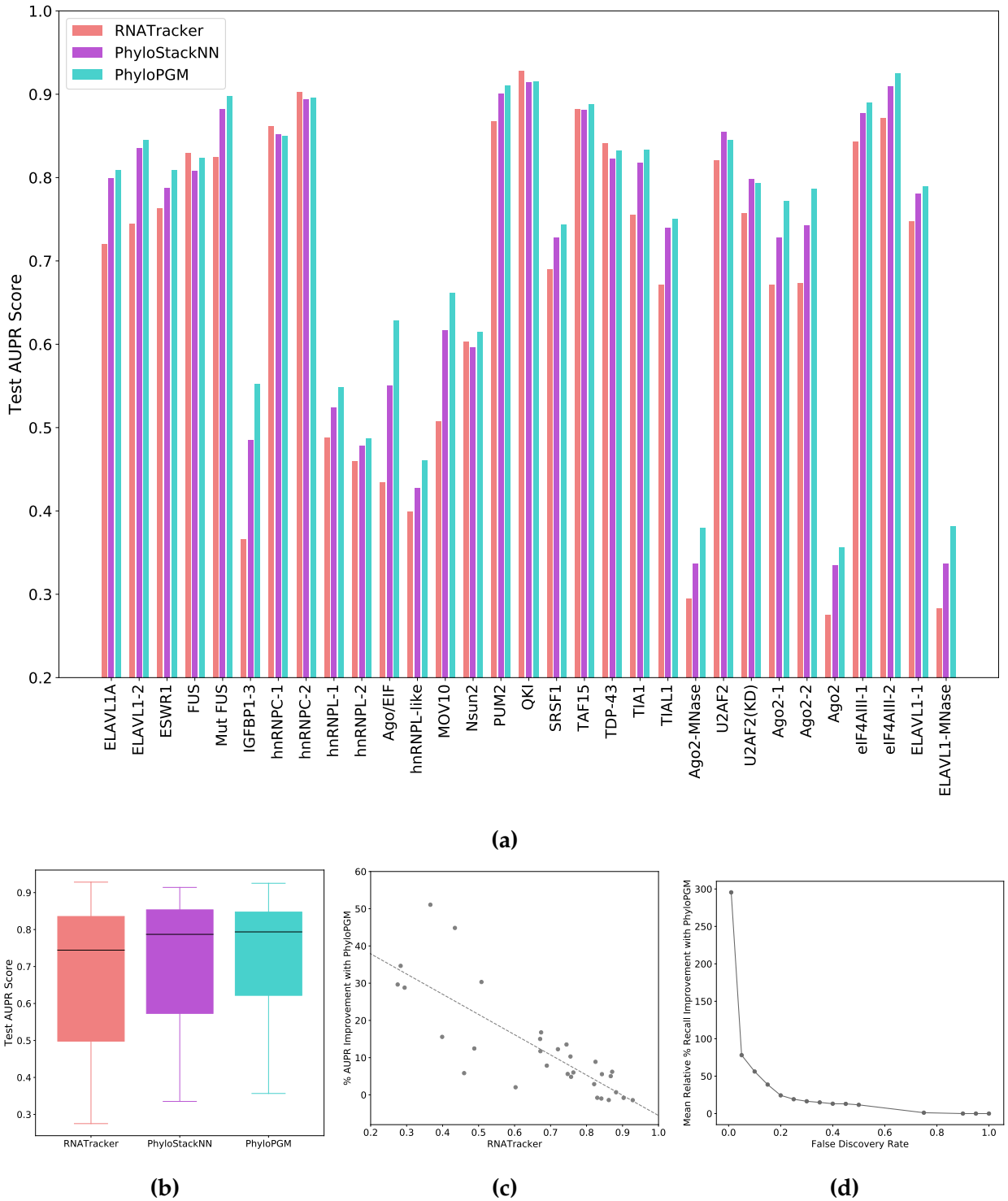


Figure 3.3: Model comparison for RNA binding prediction problem. (a) Test AUPR scores of RNATracker, PhyloStackNN and PhyloPGM over 31 CLIP-Seq datasets. (b) Distribution of test AUPR. (c) Test AUPR improvement percentage of PhyloPGM over RNATracker. (d) Mean relative percentage improvement of PhyloPGM test recall score over RNATracker for different false discovery rate thresholds.

loPGM median test AUPR: 0.17; wilcoxon signed rank test p-value: 0.019) (see Figure 3.2a and 3.2b). We also evaluated another approach, called PhyloStackNN, which uses a neural network to learn to optimally combine the base prediction scores, but without prior knowledge of the phylogenetic tree (see Methods). PhyloPGM outperforms PhyloStackNN by a smaller margin (PhyloStackNN median test AUPR: 0.15; Wilcoxon signed rank test p-value: 0.0058), which shows that utilizing the phylogenetic tree to combine the orthologous scores is indeed helpful. Notably, PhyloPGM seems particularly effective at improving prediction accuracy in liver, and less so in induced pluripotent stem cells (IPSC).

We repeated a similar evaluation for the RBP occupancy prediction task based on 31 CLIP-Seq datasets from Stražar et al. [2016] . These data were collected in HEK293, HeLa, and U266 cell types and contain 3283 to 6000 positive examples and 23672 to 26214 negative examples (see Methods). Here, we used RNATracker [Yan et al., 2019], a hybrid of convolutional and recurrent neural networks, as such architectures have shown remarkable prediction accuracy with sequence function prediction tasks [Pan et al., 2018, Quang and Xie, 2019]. Again, we find that PhyloPGM outperforms the base predictor (RNATracker median test AUPR: 0.74, PhyloPGM median test AUPR: 0.793; Wilcoxon signed rank test p-value: 8.65×10^{-6}) (see Figures 3.3a and 3.3b). PhyloPGM improves upon the base model in 26 of the 31 data sets; for the remaining 5 data sets (FUS, hnRNPC-1/2, QKI, TDP-43), the AUPR scores differ by less than 1%. Similarly to the TFBS prediction problem, we find that the PhyloPGM appraoch performs better than the PhyloStackNN approach where the phylogenetic relationship is not used (PhyloStackNN median test AUPR: 0.787; Wilcoxon signed rank test p-value: 6.57×10^{-6}).

3.4.2 Improvement to the Recall score

In general, TF and RBP binding predictors suffer from high false discovery rates, due to the fact that the ratio of negative to positive examples. Thus, apart from AUPR scores, such models should also be evaluated on the recall score at different false discovery rates

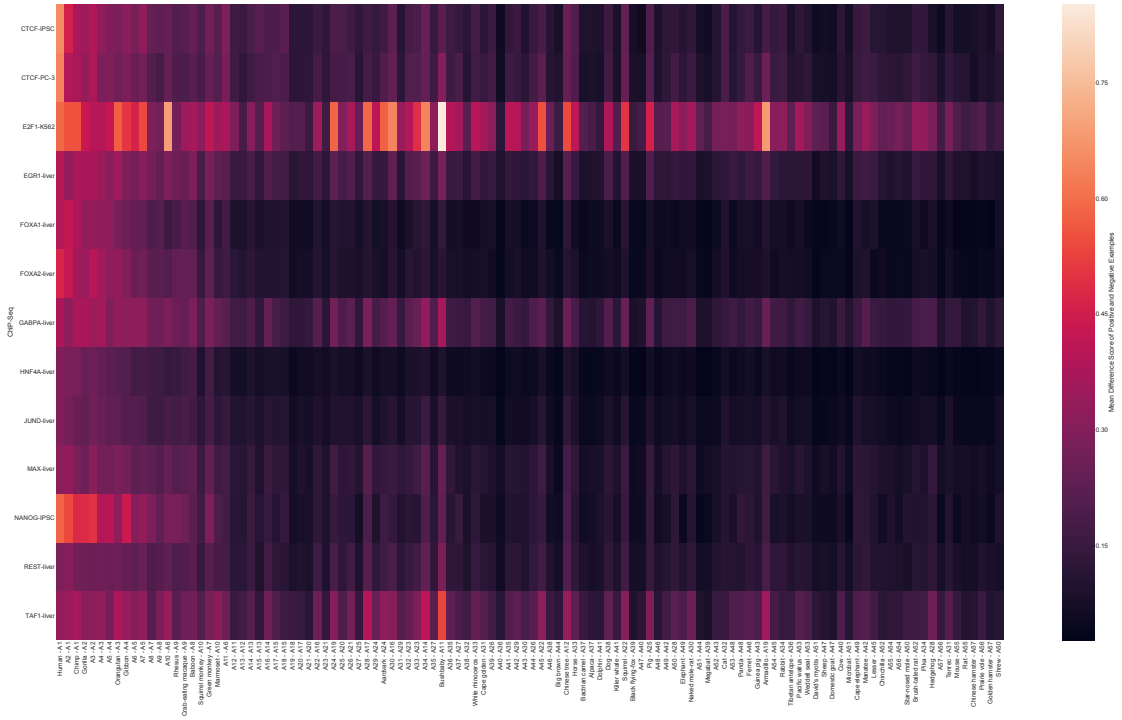
(FDRs). Figures 3.2d and 3.3d report the relative percentage improvement in the mean recall scores of PhyloPGM over FactorNet and RNATracker at different FDR thresholds. We find that PhyloPGM yields particularly large gains in recall at low FDR ranges (FDR¡0.1), which is there range of particular interest for genome-wide applications. The relative improvement in the recall score at 1% FDR is ∼18% with PhyloPGM over FactorNet in the ChIP-Seq data sets and ∼300% over RNATracker in the CLIP-Seq data sets.

3.4.3 PhyloPGM most significantly improves weaker models

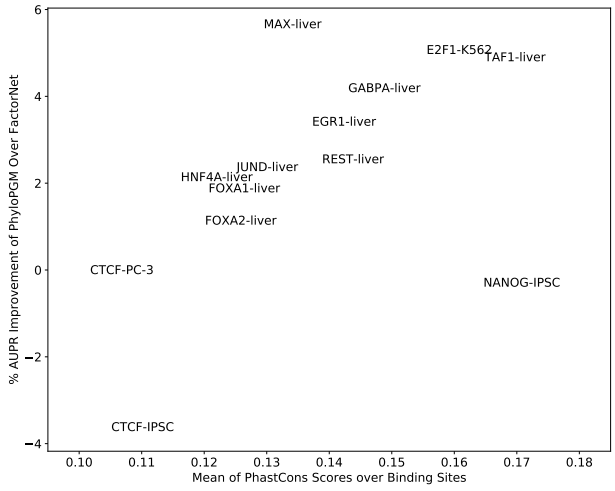
One of the main PhyloPGM design principle is to make use of orthologous examples in correctly predicting the labels that are difficult to classify with the base predictor. We observe that the amounts of improvement with PhyloPGM over FactorNet and RNATracker are larger for weaker base models, i.e. for datasets where the base models obtain a low AUPR scores (see Figures 3.2c and 3.3c). This confirms that the information from orthologous/ancestral sequences is particularly beneficial for hard-to-predict TFs and RBPs.

3.4.4 Contribution of each phylogenetic tree branches

The PhyloPGM score is essentially a sum of log-likelihood ratios over the branches of the tree, with the change in prediction score observed along each branch contributing to nudging the final prediction toward the positive or negative class. Hence it is meaningful to investigate which branch of the tree contributes most to the boost of prediction accuracy obtained by PhyloPGM. To this end, we computed, for each data sets and each branch in the tree, the mean difference of the branch log likelihood ratio of positive and negative examples (see Figures 3.4a and 3.5a). The branches most beneficial to the PhyloPGM predictions are those where this difference is largest. Notably, nearly all branches are at least minimally useful for all data sets, justifying the use of the full phylogenetic tree. However, the extent of branch-specific signals are beneficial varies significantly. For TF occupancy prediction tasks (Figures 3.4a), branches closest to human generally the

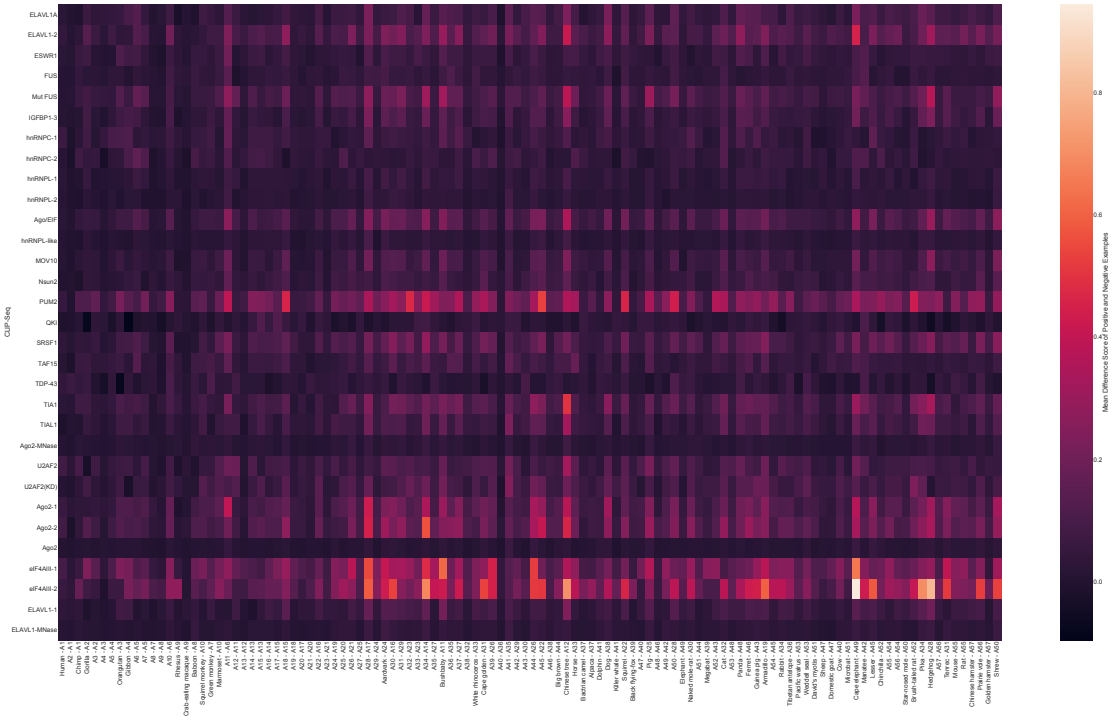


(a)

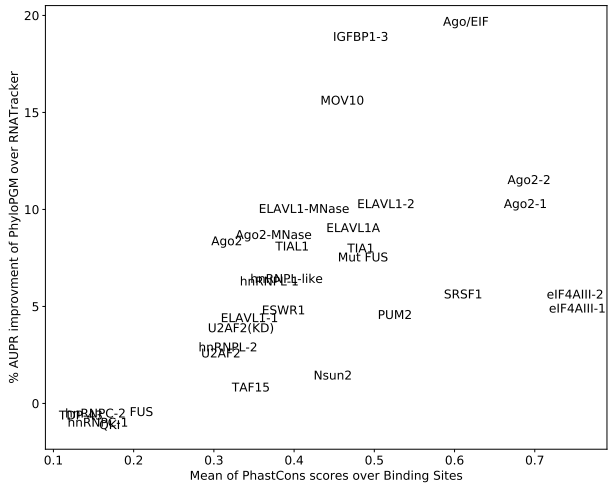


(b)

Figure 3.4: Each cell represents difference of mean of branch likelihood ratio of positive and negative examples for a branch of the phylogenetic tree in a ChIP-Seq experiment. The examples are represented as positive or negative based on the human orthologue. The branch likelihood ratio is computed from the FactorNet scores on the orthologous examples. The columns are sorted w.r.t evolutionary distance of the branch from human. Each column is named as species followed by its direct parent. The species A_i denotes ancestral species, where i indicates the evolutionary distance from human (i=1 is closest to human).



(a)



(b)

Figure 3.5: Each cell represents difference of mean of branch likelihood ratio of positive and negative examples for a branch of the phylogenetic tree in a CLIP-Seq experiment. The examples are represented as positive or negative based on the human ortholog. The branch likelihood ratio is computed from the RNATracker scores on the orthologous examples. The columns are sorted w.r.t evolutionary distance of the branch from human. Each column is named as species followed by its direct parent. The species A_i denotes ancestral species, where i indicates the evolutionary distance from human (i=1 is closest to human).

most predictive value. This is particularly true for CTCF and Nanog, which also happen to be those data sets obtained from IPSCs, and for which PhyloPGM underperforms. We hypothesize that many of the human binding sites for these proteins may have arisen recently during primate evolution, as suggested by Ni et al. [2012] and Scerbo et al. [2014]. On the contrary, transcription factors such as E2F1, GAPBA, and TAF1 (all assayed in liver) display a high level of branch informativeness across much of the mammalian tree. This suggests a lower turnover of regulatory regions for those TFs.

To investigate the role of conservation in more details, we compared the percentage improvement from PhyloPGM in AUPR with the mean of PhastCons scores in the bound examples for each dataset (see Figures 3.4b and 3.5b). We observe that the amount of improvement in AUPR is highly correlated with the PhastCons scores. Moreover, the majority of CLIP-seq data have higher PhastCons scores than the ChIP-Seq data, which is expected due to RNA binding sites being generally more conserved than the TFBSs [Payne et al., 2018]. Therefore, PhyloPGM seems to be more effective in boosting the binding prediction accuracy of TFs and RBPs whose binding sites are more conserved. The major exception to this trend is NANOG and IPSC data, may be due to the absence of binding sites in the orthologous regions [Scerbo et al., 2014].

3.4.5 PhyloPGM helps identifying disease-causing human non-coding variants

ChIP-Seq and CLIP-Seq experiments are limited to the question of whether a given protein binds a certain genomic region or not, but does not reveal information on the functional consequences of this interaction. Indeed many binding events appear to have no or only limited consequences on gene expression [Vanhille et al., 2015, Barakat et al., 2018], and hence be evolutionarily neutral. Because PhyloPGM indirectly measures the level of selective pressure to maintain the binding potential of a region for a given TF/RBP, it stands to reason that regions with high PhyloPGM scores not only have a higher chance

of being bound, but also that this binding event is more likely to be of functional consequences.

To test this hypothesis, we used a variety of external data sources to identify binding events that are more likely to be of functional consequences, including: (i) the non-coding portion of the ClinVar database [Landrum et al., 2016], which human mutations associated to diseases; (ii) the non-coding human variants linked to phenotypic consequences through several publications [Biggs et al., 2020]; (iii) the list of deleterious non-coding variants identified through machine learning and other computational techniques [Wells et al., 2019]. Regions of the human genome bound by a TF/RBP and overlapping at least of those data sets are deemed more likely to harbor functional binding events and are called *putatively functional*.

We then measured, for each TF/RBP, the extent to which the bound regions that are assigned the highest PhyloPGM scores (top 30%) overlap the set of putatively functional sites. The same procedure was applied to the top regions ranked based on the base predictor (FactorNet or RNATracker) or a simpler measure of sequence conservation (PhastCons).

Figure 3.6 shows that for 11 of the 12 TF data sets, high-scoring putatively functional TF binding siteds are more commonly found within high-scoring PhyloPGM sites regions assigned a high score by PhyloPGM overlap a larger number of putatively functional sites

Figure 3.7 show that PhyloPGM is more effective in predicting the functional regions e.g. in (E2F1, K562), (EGR1,liver), (GABPA, liver) of the ChIP-Seq datasets and in Mut FUS, SRSF1, Ago2-1 of the CLIP-Seq datasets. This is an interesting benefit of PhyloPGM because PhyloPGM not only boost the base model performance, but, is also more predictive of the functional aspects.

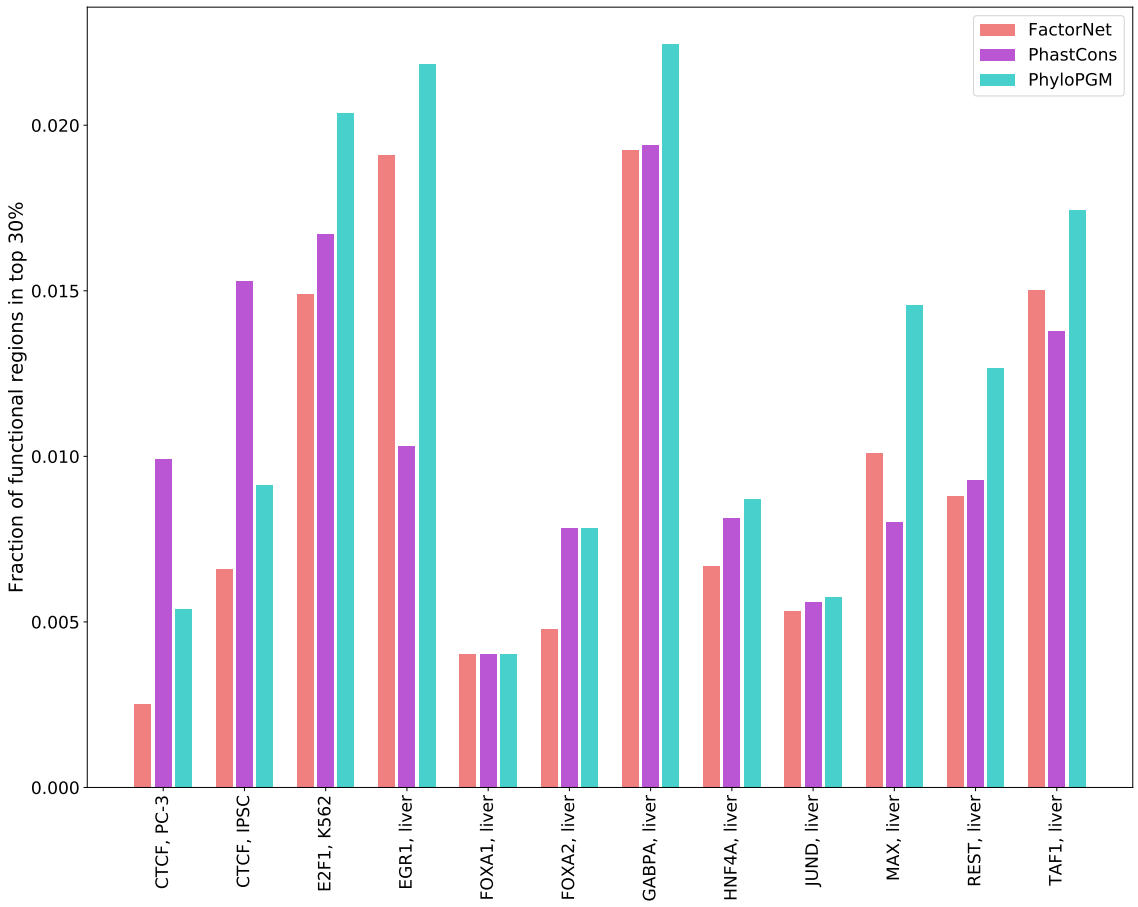


Figure 3.6: Mean of relative percentage improvement in recall score with PhyloPGM over FactorNet (human).

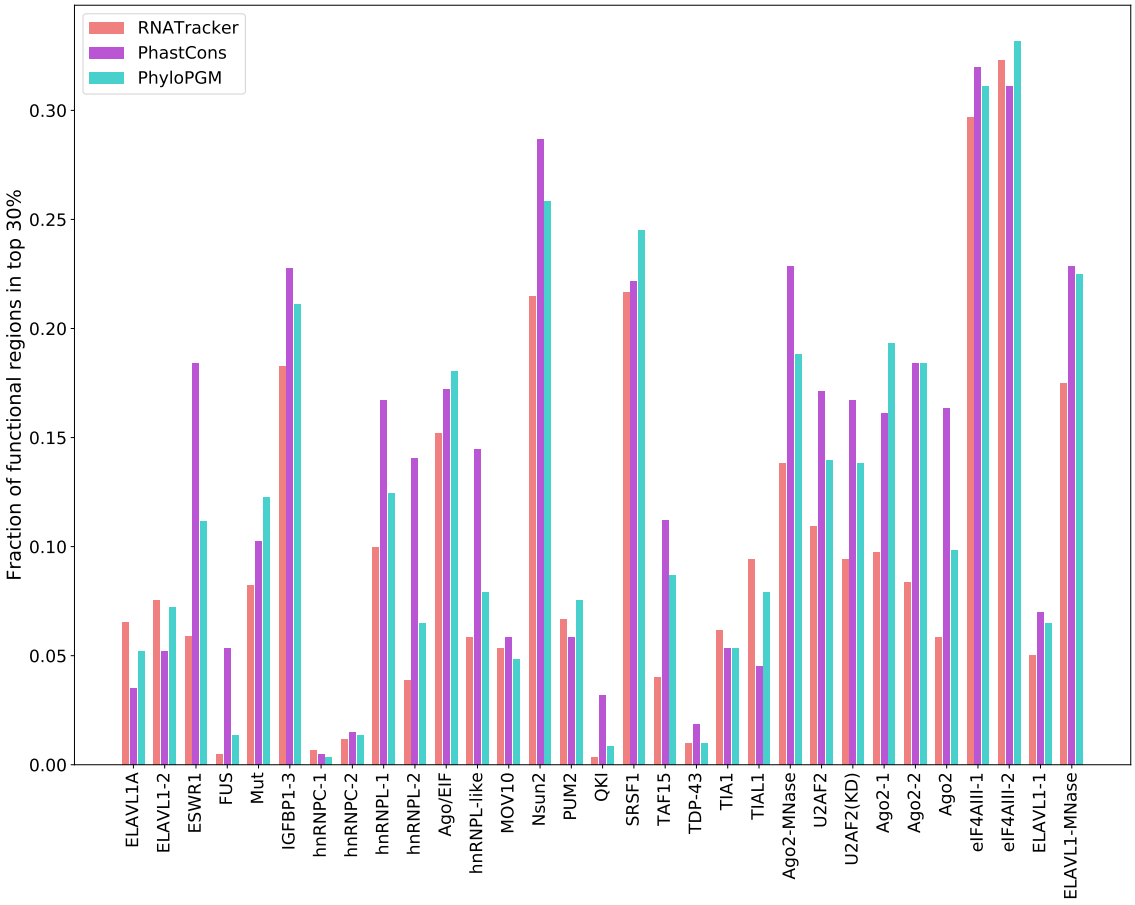


Figure 3.7: Mean of relative percentage improvement in recall score with PhyloPGM over RNATracker (human).

3.5 Discussion

We present PhyloPGM, an aggregation approach to boost the prediction accuracy of a previously trained TF or RBP binding predictor. We show that PhyloPGM significantly improves the median AUPR scores of FactorNet and RNATracker models trained on human sequences by more than 4% in 13 ChIP-Seq datasets and 5% in 31 CLIP-Seq datasets. PhyloPGM, in principle, is designed to improve the prediction accuracy of the labelled examples that are difficult to classify i.e. the examples that lie closer to the decision boundary. Indeed, our analysis show that the log-likelihood ratio of the parents and descendants in the orthologous set improve the prediction quality of such examples. The most improvements in the AUPR score with PhyloPGM are observed on datasets where FactorNet or RNATracker performed relatively poorly. Moreover, we show that the explicit use of the phylogenetic tree provides significant gain for PhyloPGM over PhyloStackNN, which combines the orthologous predcition scores with a neural network without taking into account of the phylogenetic relationship. Additionally, PhyloPGM is shown to have better recall scores at lower false discovery rates than the base models in both ChIP-Seq and CLIP-Seq datasets.

We find that the datasets showing more improvement with PhyloPGM over base models have relatively higher PhastCons scores i.e. the sequences are more conserved. We observe that PhyloPGM improves the base model relatively more in CLIP-Seq data compared to ChIP-Seq data. The RNA binding sites are mostly observed in the 3' UTR region, which are generally more conserved than transcriptional regulatory regions. This may explain the comparatively better performance of PhyloPGM in CLIP-Seq data. Furthermore, binding site turnover may affect transcriptional regulatory regions more than 3' UTRs, which may cause loss of or larger shifting of binding sites in the orthologs.

The comparison of branches in the phylogenetic tree in terms of the impact of the likelihood ratio on PhyloPGM shows that the branches that are farther from human are relatively more useful. However, the branch likelehood ratio seems to be less/not useful after

a certain distance from the human, which may indicate loss of binding sites in such orthologs. In the similar direction, we should explore other phylogenetic relationships such as the effect of subset of species on the PhyloPGM accuracy, relationship between the regulatory function associated with a sequence and its conservation across different species. More of such investigations should allow to identify important evolutionary changes that had impact on regulatory regions. Additionally, this should open up the possibility of using PhyloPGM as a potential comparative genomics tool that can be applied in many other related areas e.g. therapeutic approaches, studying evolution of regulatory activities and other functions related to biological sequences. Furthermore, the application of PhyloPGM on a subset of useful species rather than the entire orthologs shall reduce the computational run time of PhyloPGM with some loss or gain in the accuracy.

An important observation from the analysis with the ClinVar datasets is that PhyloPGM is more predictive of the human genomic regions where mutations are linked to diseases. One aspect of results with ClinVar datasets is that PhyloPGM is capable of identifying deleterious regions. Moreover, one can compare base model predictions on reference genome and an individual genome to filter the regions with significant prediction differences. Then, PhyloPGM can be applied on these selected regions of an individual genome to detect regions with any concerned mutations. The other aspect of results with ClinVar datasets is that the regions where mutations are linked to diseases could be considered as functional, in the sense that mutations in such regions could affect the fitness of species. Such regions should be associated with some regulatory activities. Now, the ChIP-seq and CLIP-seq experiments are not free from noise (e.g. false TF or RBP binding sites, inconsequential binding etc) [Vanhille et al., 2015, Barakat et al., 2018, König et al., 2012, Moore et al., 2014, Ule et al., 2005]. It can be a case that such wet-lab experiments identify a genomic location as a potential binding site for a TF or RBP, however, a TF or RBP binding to such location has no impact on any regulatory activity. Improving the wet-lab experiment data with more functional regions (i.e. identified binding sites has some role in a regulatory activity) may result into further improvement in the accuracy

with PhyloPGM. The improved PhyloPGM scores can further be used to identify regions associated with regulatory activities.

At present, PhyloPGM is presented for a binary classification task, but, is potentially extendable to multi classification tasks (e.g. with one-vs-all setting). This should allow PhyloPGM to be applicable to other sequence function prediction tasks that involves more than one labels, for example protein function prediction [Kulmanov and Hoehndorf, 2020], mRNA subcellular localization [Yan et al., 2019]. PhyloPGM is inherently designed for classification tasks and will require modifications in order to be applicable to regression-based sequence function prediction tasks e.g. predicting gene expression value from a sequence. The discretization of regression values may allow the application of PhyloPGM in the regression tasks. Furthermore, the use of beta distribution and conditional multivariate distribution in place of multinomial distribution may allow to better fit the log-likelihood ratio of the branches in PhyloPGM pipeline. Although many sequence function prediction tasks have computational models and datasets (e.g. [Kulmanov and Hoehndorf, 2020, Yan et al., 2019, Leclercq et al., 2017]), applying PhyloPGM to them will require necessary adjustments w.r.t. the base predictors and datasets. The datasets size and base predictor forms vary from one sequence function prediction tasks to another. Furthermore, the improvement in accuracy and evolutionary insights from PhyloPGM for a given sequence function prediction task depends on the base predictor and the datasets (s.t. sequence function is maintained during evolution).

3.6 Methods

We define the problem of aggregating prediction scores in order to improve prediction accuracy as,

Given: a set of prediction scores on the orthologous and ancestral genomic sequences obtained from a base model, B, which is previously trained using human genomic sequences only, and a phylogenetic tree that relates the involved species.

Goal: to predict the label of a human genomic sequence such that the resulting prediction improves the accuracy of B.

We first describe the ChIP-Seq data, CLIP-Seq data and orthologous data that are used to demonstrate the efficiency of PhyloPGM. Then, we detail the Factornet and RNA-Tracker models, which are used as base predictors. We conclude by describing the PhyloPGM and PhyloStackNN algorithms.

3.6.1 ChIP-Seq data

A recent DREAM challenge "ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge" provided ChIP-Seq data from ENCODE for various problems related to TFBSs prediction [Kundaje et al., 2021]. One of the labelling problems is to build a TFBS prediction model for a given cell-type. The data consist of 13 TF/cell-type pairs from 12 TFs and three cell-types (liver, PC-3, induced pluripotent stem cell (IPSC)). The train and test sets both belong to the same cell-type. The test examples are from chromosomes 1, 8 and 21 for a given cell-type and the other chromosomes form train examples. The test set size is 8 million examples. We sub-sampled negative examples to the same number of positive examples in the train set during the training phase. We use 20% of the train set as validation set.

3.6.2 CLIP-Seq data

The protein binding data is originally proposed in Strazar et al. [2016] and includes results of 31 RBP binding experiments conducted under the CLIP-Seq protocol. Each of the experiment provides 8000 positive examples that contain a binding site for a specific RBP, and 32000 negative (unbound) examples, where an example is an RNA sequence of 101 nt. A partition into a fixed train-test split is then used in the original paper, each containing 20% positive examples. The positive binding sites are identified through several

variants of the CLIP-Seq protocol such as PAR-CLIP [Hafner et al., 2010a], iCLIP [Konig et al., 2010] and HITS-CLIP [Licatalosi et al., 2008a].

3.6.3 Orthologous data

The orthologous regions of each human genomic region in other mammals are extracted using mafsInRegion program (`https://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/mafsInRegion`) from a 100-way vertebrate whole-genome alignment available from the UCSC genome browser [Kent et al., 2002]. Only the 58 mammalian sequences were used in this study. The orthologous regions are complemented with computationally predicted ancestral sequences produced by Ancestor1.0 [Diallo et al., 2009]. The collected orthologous regions are symmetrically trimmed or joined with surrounding regions to yield sequences sized 1000 bps (TFBS prediction problem) or 101 bps (RBP binding site prediction problem). Each example has on average 80 orthologous and ancestral sequences. We ignore orthologous regions that are smaller than 70% of the corresponding human sequence.

3.6.4 FactorNet as base predictor

FactorNet [Quang and Xie, 2019] is one of the best performing sequence-based model in a recent DREAM competition[Kundaje et al., 2021]. In this study, we used the FactorNet architecture that takes sequence information as sole input. The input is a genomic sequence and its reverse complement, which are passed to a convolution layer of 32 filters. The size of each filter is 26 and the resulting output is passed through a ReLU activation layer. A dropout layer of $p = 0.1$ is applied, which is followed by a max pooling layer with a filter size of 13. Then, a single birectional LSTM layer of hidden size 32 is used with a dropout layer of $p = 0.5$. Afterwards, a fully connected layer of size 128 with ReLU activation function is used. The output of fully connected layer is then passed through a dropout layer of $p = 0.5$. The final output layer is a fully connected layer of size 1 with a sigmoid

activation function. The mean of the FactorNet outputs from the given genomic sequence and its reverse complement is the final output of the FactorNet. In this study, we trained FactorNet batch-wise (batch size = 128) with early stopping using a validation set.

3.6.5 RNATracker as base predictor

Yan et al. [2019] proposed a hybrid of convolutional and recurrent neural network architecture, called RNATracker, to predict the mRNA localization. A mRNA sequence is represented as a one-hot encoded vector, which is then passed through two convolutional layers. Then, a pooling layer is used to aggregate the motif scores. Finally, a bi-directional LSTM with attention is used to aggregate the motif features. The resulting output is passed through a fully connected layer followed by a linear layer to predict the RNA-RBP binding. Although, RNATracker was developed to predict mRNA localization, the architecture is equally capable of predicting RNA-protein binding. The RNATracker architecture used in this study has two convolutional layers, where each convolution layer has 32 filters of length 10 followed by a max pooling layer of window size 3 and stride 3. The subsequent bidirectional LSTM layer has 100 hidden units and the following fully connected layer has 128 units. The output layer has one unit with sigmoid activation function that gives the final prediction score. A dropout layer ($p = 0.1$) is used after each convolutional and bidirectional LSTM layer.

3.6.6 PhyloPGM: Probabilistic Aggregation Approach

PhyloPGM is a prediction score aggregation approach, which is inspired from the probabilistic graphical models [Koller et al., 2009]. Consider a model trained for the TF or RBP binding prediction problem and a phylogenetic tree, ψ, where each node represents the real-valued score assigned by a base predictor to the corresponding orthologous sequence. A simple way to combine the predictions would be to take a weighted-average. However, this ignores the dependencies modeled by the tree structure and a smaller num-

ber of strong predictions can get undermined by a majority of weak predictions. A better approach could be to utilize a probabilistic graphical model view of combining the scores.

Consider a phylogenetic tree, ψ, with n nodes, where index 1 is the root , s_i denotes the base model score assigned to node i, and e_{ij} is the evolutionary distance between parent i and descendant j. Let the label of the root species be Y. Then the probability of $Y = y$ given the set of prediction scores is:

$$P[Y = y \mid s_1, s_2, ..., s_n] \propto \; P[s_1, s_2, ..., s_n| \; Y = y] \cdot P[Y = y]$$

$$= P[s_1|Y = y] \cdot \prod_{(p,c) \in edges(\psi)} P[s_c|s_p, Y = y, e_{p,c}] \tag{3.1}$$

where p, c are parent-descendant pairs and $e_{p,c}$ is the evolutionary distance between them in ψ.

The final combined score to predict Y is the log likelihood ratio of eq. 3.1 with $Y = 1$ and $Y = 0$, where 1 and 0 denotes positive and negative labels respectively:

$$\text{PhyloPGM\_Score} = \log \left(\frac{P[Y = 1 \mid s_1, s_2, ..., s_n]}{P[Y = 0 \mid s_1, s_2, ..., s_n]} \right) \tag{3.2}$$

$$= \log \left(\frac{P[s_1|Y = 1]}{P[s_1|Y = 0]} \right) + \sum_{(p,c) \in edges(\psi)} \log \left(\frac{P[s_c|s_p, Y = 1, e_{p,c}]}{P[s_c|s_p, Y = 0, e_{p,c}]} \right) \tag{3.3}$$

$$\propto \; \log \frac{P[s_1|Y = 1]}{P[s_1|Y = 0]} + \; \alpha \; \cdot \sum_{(p,c) \in edges(\psi)} \log \frac{P[s_c|s_p, Y = 1, e_{p,c}]}{P[s_c|s_p, Y = 0, e_{p,c}]} \tag{3.4}$$

where α is a model hyper parameter to balance the effect of likelihood ratio of non-root species.

The conditional probabilities ($P[s_c|s_p, Y = y, e_{p,c}]$) of the base model score on a descendant species given the parent score, label and the evolutionary distance is difficult to compute. We estimate the conditional probabilities of scores on root node and over each edge empirically from the scores in the training dataset, T. In order to empirically

estimate the conditional probabilities, the base prediction scores, which are assumed to be between 0 and 1, are rounded to first decimal place and binned in 12 bins (0-1, and one extra bin for the missing values). The required probabilities in the eq. 3.4 are estimated as,

$$P[s_1|Y=y] = \frac{\sum_{i\in T} 1_{s_1^i=s_1 \wedge l(i)=y} + \epsilon}{\sum_{i\in T} 1_{l(i)=y} + 12\epsilon}$$

$$P[s_c|s_p, Y=y] = \frac{\sum_{i\in T} 1_{s_c^i=s_c \wedge s_p^i=s_p \wedge l(i)=y} + \epsilon}{\sum_{i\in T} 1_{s_p^i=s_p \wedge l(i)=y} + 12\epsilon}$$

where ϵ is a pseudo count set to $\epsilon = 1$

It should be noted that for a given example, s_c or s_p may be missing due to the absence of orthologous regions in the corresponding species. The missing values are ignored in such cases. In this study, we use the phylogenetic tree available from `https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/` and rerooted the tree so that human becomes the root node. We empirically selected $\alpha = 0.1$ from $[0.001, 0.01, 0.1, 1, 10, 100, 1000]$.

3.6.7 PhyloStackNN Approach

The goal of the stacking approach, PhyloStackNN, is to test the importance of the explicit use of the phylogenetic tree in the PhyloPGM approach. PhyloStackNN is a simple multi-layer perceptron that takes base predictor scores $s_1, s_2, ..., s_n$ as input and is trained to predict the label. It is trained on the same train/test split as PhyloPGM. The MLP architecture is chosen from the hyper-parameter search over {'hidden_layer_sizes': [(32,), (100,), (64, 32)], 'ℓ_2 penalty': [0.1, 1, 10]} using 10-fold cross validation.

3.6.8 Implementation Details and Availability

We used PyTorch [Paszke et al., 2019] v1.4.0 to train RNATracker and FactorNet models. The scikit-learn [Pedregosa et al., 2011b] is used to implement PhyloStackNN and

to compute AUPR, precision and recall scores. The PhyloPGM package is available at `https://github.com/BlanchetteLab/PhyloPGM`.

Chapter 4

A Web Interface to PhyloPGM

In the following manuscript, I develop and implement PhyloPGM-Web, a web interface for predicting transcription factor (TF) and RNA binding protein (RBP) binding sites. First, I configure and manage a web server. Then, I integrated software framework to the web server that allows user to register, submitting and monitoring a computational job. I process and compiled previously published TF and RBP binding sites data, which are used to train a previously published deep learning model. I extracted the orthologs/ancestors sequences that are used in PhyloPGM pipeline. I find that PhyloPGM improved the accuracy of base predictor in both TF and RBP binding site datasets. Afterwards, I implemented PhyloPGM pipeline within the software framework. PhyloPGM-Web allows users to submit human genomic locations, executes PhyloPGM pipeline and outputs predictions from 31 RBP and 435 TF binding predictors. The results are shown with interactive plots and an e-mail notification is sent to the user once the results are complete. I showcase PhyloPGM-Web with an analysis of Polypyrimidine Tract Binding Protein 3 (PTBP3) gene. I observe that branches with longer evolutionary distance can have more impact on PhyloPGM result even if the associated species are farther from human in terms of evolutionary distance.

Citation: Faizy Ahsan, Zichao Yan, and Mathieu Blanchette (2021) PhyloPGM As A Web Service. *In preparation for submission to Bioinformatics Application Note.*

4.1 Manuscript 3: PhyloPGM-Web: An online platform for evolutionarily-boosted prediction of protein-DNA/RNA interactions

Authors: Faizy Ahsan[1], Zichao Yan[1], Mathieu Blanchette[1]

[1]School of Computer Science, McGill University, Montreal, Quebec, Canada

4.2 Abstract

We present PhyloPGM-Web, a user-friendly web service to predict human transcription factor (TF) and RNA binding protein (RBP) binding sites by taking advantage of deep multiple genome alignments to boost prediction accuracy. PhyloPGM-Web uses deep learning approach called RNATracker as base model, applies it to the human sequence of interest as well as to its orthologs and ancestors, and combines the prediction scores obtained using PhyloPGM. The web service allows a user to submit a set of genomic locations of interest and predicts the RBPs or TFs binding sites of 31 RBPs and 435 TFs in multiple cell lines. Along with the RNATracker and PhyloPGM prediction scores, a user may investigate the contribution of individual species and branches of the phylogenetic tree towards the PhyloPGM score.

4.3 Introduction

The computational prediction of transcription factor (TF) and RNA binding protein (RBP) binding sites will allow to comprehend transcriptional and post-transcriptional gene regulatory networks. It has several advantages over wet-lab experiments, such as time and cost reductions, rapid evaluation of candidate sequences, and, possibly, assisting therapeutic approaches. Although computational models based on deep learning approaches have been shown to outperform classical computational models to predict TF and RBP

binding sites, they are far from replacing wet-lab experiments in terms of prediction accuracy [Alipanahi et al., 2015, Yan et al., 2019, Quang and Xie, 2019, Ahsan et al., 2020]. Even if computational models are imperfect, they have many advantages over wet-lab experiments, for example motif analysis, candidate selection, studying impact of mutation, and analyzing genomes that are difficult to obtain.

Using computational models often requires a high level computational expertise, which represents a major obstacle to biologists benefiting from these tools. A web-interface to a computation model that predicts TF or RBP binding sites in a given input genomic location will allow a user to focus on the model outcomes without dealing with the complex computational pipeline. For example LASAGNA-Search [Lee and Huang, 2013] is a web tool that uses position weight matrix (PWM) models to predict TFBSs in a given input sequence and RNAsite [Su et al., 2021] is a web-interface that uses sequence and structure information to predict RBP binding sites.

We recently introduced PhyloPGM, a tool that improves the prediction accuracy of a previously trained predictor for both TF and RBP binding prediction problem [Ahsan et al., 2021]. PhyloPGM works by aggregating the prediction scores on the target sequence and its orthologous regions from a pretrained predictor to predict TF or RBP binding sites in human. PhyloPGM implementation is a complex pipeline, relies on a set of pretrained predictors and requires very large alignment files that are not easy to distribute. In this chapter, we present a web-interface for PhyloPGM where a user may submit a set of genomic locations to find binding sites for a set of TF or RBP. In addition to providing accurate prediction of TF/RBP binding sites, PhyloPGM scores capture the level of selection pressure on those sites, and hence provides a way to assess which binding site may be more critical to fitness.

4.4 Results

4.4.1 Overview of PhyloPGM-Web pipeline

PhyloPGM-Web provides easy access to the PhyloPGM algorithm to predict TF binding sites in a given set of human DNA sequences, and RBP binding sites in the RNA sequences that would be transcribed from them. The PhyloPGM-Web pipeline is described in Figure 4.1. A user is first required to register on the web-interface. Registration enables the system to keep track of user's ongoing jobs and past results. The user starts by providing as input a bed file of human genomic regions (hg38 assembly), optionally providing a transcriptional strand for each region (relevant only for RBP binding prediction). They then select one or more TF or RBPs for which they want predictions to be made. Currently, PhyloPGM-Web allows selecting models for 435 TFs across multiple cell types (from the DREAM datasets [Kundaje et al., 2021] and the MIT-CSAIL datasets [Zeng et al., 2016]) and 31 RBPs (from the CLIP-Seq datasets). To maximize prediction accuracy, PhyloPGM-Web applies RNATracker to regions of 1000 bp (for TFBS prediction) or 101 nt (for RBP binding prediction). Once a user submits the bed file, the web-interface extends the input location(s) on both sides by 500 bps or 50 bps. Then the corresponding orthologs/ancestors sequences from 58 mammals and 57 computationally reconstructed ancestors are fetched. Afterwards pretrained RNATracker models are applied to each orthologs/ancestor sequence and the RNATracker scores are combined using PhyloPGM. The RNATracker/PhyloPGM score assigned to a given position of an input sequence is the score of the window centered at that position. The user is sent an e-mail notification when the PhyloPGM scores are computed. The web interface presents the results as interactive plots of PhyloPGM and RNATracker prediction scores for the selected TFs/RBPs at each position along the sequence. The user may select which TFs or RBPs should be shown. A slider bar with false discovery rate (FDR) threshold enables the user to control the quantity of results that are reported. Furthermore, the user is provided with the interactive heatmaps of RNATracker prediction scores for each species and log-likelihood ratio

of each branch of the phylogenetic tree. The log-likelihood ratio provides phylogenetic information that contrasts the hypothesis that the given human sequence is a TF/RBP binding site or not.

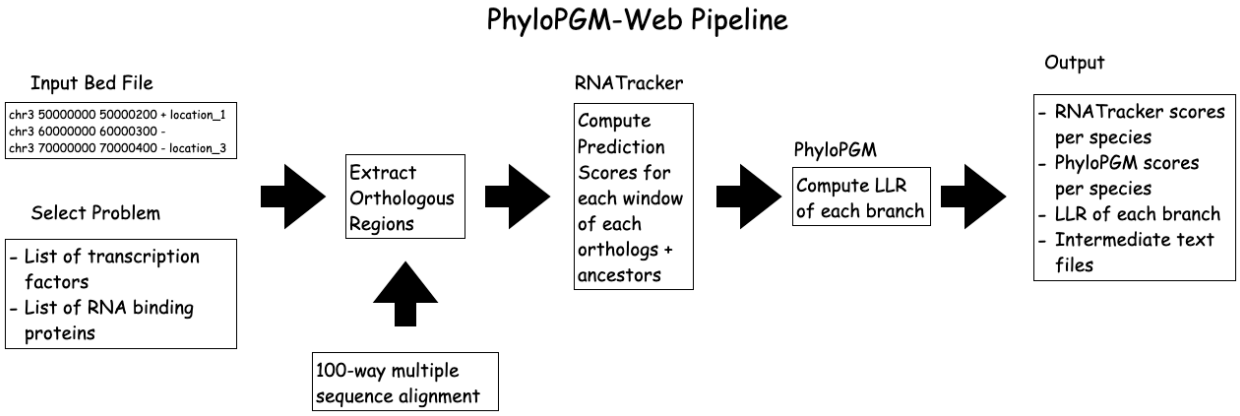


Figure 4.1: PhyloPGM-Web pipeline. The user is required to submit a bed file with input location(s) and to choose whether to predict binding sites for transcription factors, RNA binding proteins, or both. Then, orthologous regions in extant and ancestral species are extracted. RNATracker scores and PhyloPGM scores are computed for each window in the input location(s). PhyloPGM-Web provides RNATracker and PhyloPGM scores, log-likelihood ratio (LLR) of each branch of the phylogenetic tree and the intermediate text files.

4.4.2 Evaluation of Trained Models on PhyloPGM-Web

We use RNATracker to train TFBS predictors on ChIP-seq data for 13 TF/cell-type pairs from the ENCODE-DREAM challenge data [Kundaje et al., 2021] and 422 TF/cell-type pairs from the MIT-CSAIL data [Zeng et al., 2016]. The trained TFBS RNATracker predictors are used to build PhyloPGM models. We find that PhyloPGM improves the median test AUPR score of RNATracker on both the ENCODE-DREAM challenge data [Kundaje et al., 2021] (RNATracker median test AUPR: 0.13, PhyloPGM median test AUPR: 0.14; Wilcoxon signed rank test p-value: 0.46) and MIT-CSAIL data [Zeng et al., 2016] (RNATracker median test AUPR: 0.85 , PhyloPGM median test AUPR: 0.86; Wilcoxon signed

rank test p-value: 1.98×10^{-59}). It should be noted that the test set size and the ratio of binding and non-binding sites in test sets are different for both TFBS data. The test sets in MIT-CSAIL data [Zeng et al., 2016] are approximately balanced with an average number of 68,044 examples, while test sets in ENCODE-DREAM challenge data [Kundaje et al., 2021] are highly imbalanced ($<< 1\%$ binding sites) with an average number of $\sim$8 million examples.

We observe that the amount of improvement in test AUPR score is larger for the cases where RNATracker has relatively lower test AUPR score (see Figures 4.5 and 4.6). The test AUPR scores of RNATracker and PhyloPGM for both TFBS data are shown in Table 4.1 and Table 4.2. Similarly, we previously trained RNATracker models on 31 CLIP-seq data, which are then used as base models to build PhyloPGM models for predicting RBP binding sites. Ahsan et al. [2021] found that PhyloPGM improves the test AUPR score of RNATracker on 31 CLIP-seq data (RNATracker median test AUPR: 0.74, PhyloPGM median test AUPR: 0.793; Wilcoxon signed rank test p-value: 8.65×10^{-6}.

4.4.3 Analysis of PTBP3 3'UTR with PhyloPGM-Web

We illustrate the use of PhyloPGM-Web to analyse the 3' UTR of the human Polypyrimidine Tract Binding Protein 3 (PTBP3) gene (chr9:112217716-112223851, reverse strand). The regulation of PTBP3 was used to showcase the integrated data analysis with ENCORE pipeline [Van Nostrand et al., 2020]. The protein encoded by PTBP3 plays a role in the regulation of cell differentiation [Yamamoto et al., 1999, Sadvakassova et al., 2009]. Mutations in PTBP3 are associated with fanconi anemia, which can cause malformations in major organ systems, affect bone marrow and pose a high risk to cancer [Deakyne and Mazin, 2011].

Figure 4.2 illustrates the main output of PhyloPGM-Web, with RNATracker and PhyloPGM scores computed for every window of 101 nt, with 50 nt offsets. For comparison, we included in the figure a snapshot of the UCSC genome browser showing the binding sites identified by eCLIP for the same RBPs, from the ENCORE project [Van Nostrand

et al., 2020]. At a 5% FDR threshold, we find that RNATracker is able to identify 5 out of the 10 RBPs that have eCLIP peaks in this sequence, while PhyloPGM-Web find 9. Note the PhyloPGM-Web does not attempt to pinpoint the precise location of binding sites, but instead assesses the binding potential of 101-nt windows. Hence, it is expected that the position of the eCLIP peaks only approximately match the predictions. The comparison shows that the PhyloPGM approach is able to boost the RNATracker scores to identify more potential RBPs that can bind to the input location.

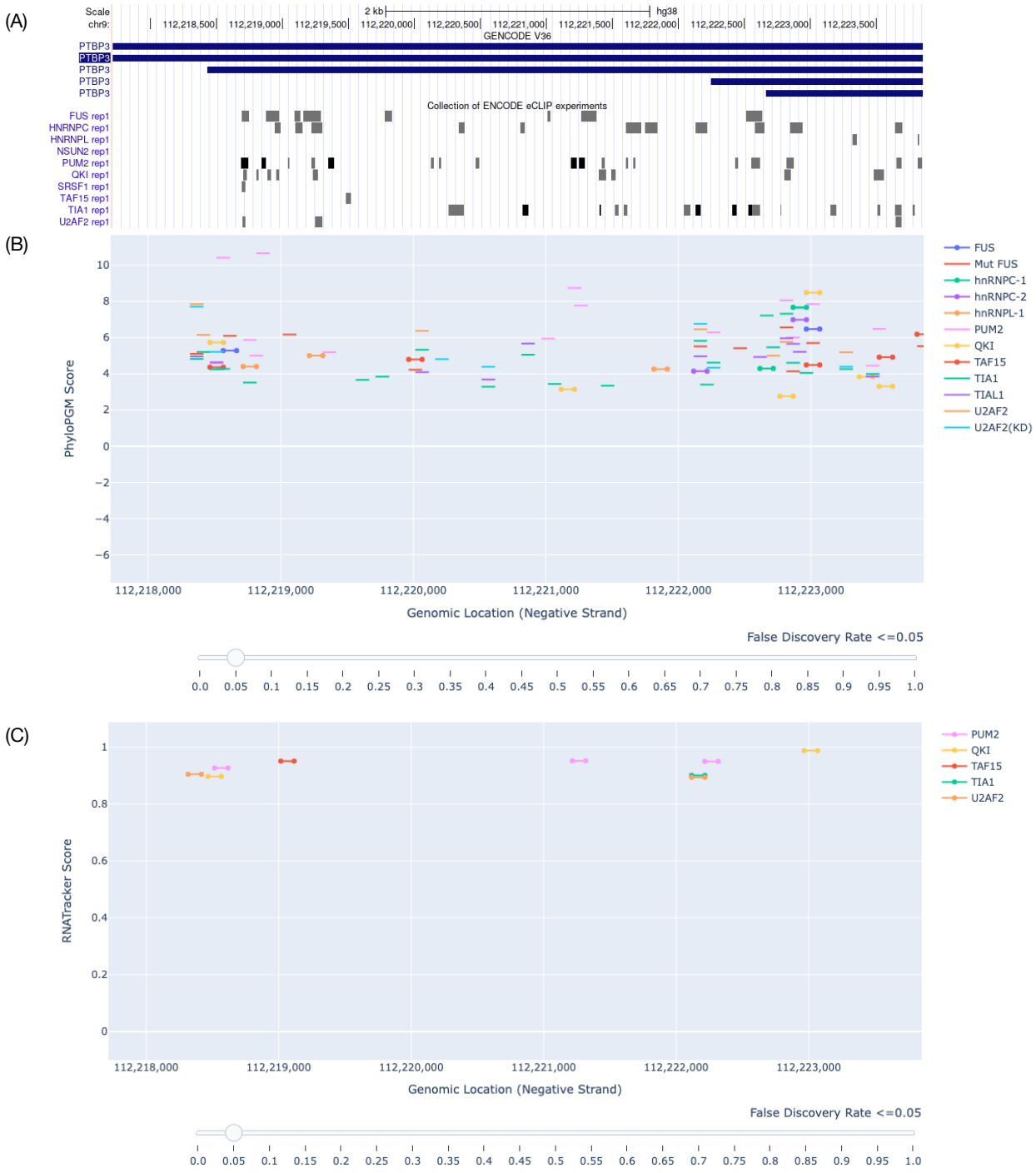


Figure 4.2: RBP binding prediction results for the PTBP3 3'UTR. (A) UCSC genome browser eCLIP results (K562 cell-type, replicate 1 only) for the 31 RBPs used in our study. (B) PhyloPGM RBP binding prediction, and (C) RNATracker RBP binding prediction from PhyloPGM-Web at every 50 nt over the window of 101 nt at 5% FDR.

In Figure 4.3, we present how PhyloPGM-Web allows a detailed analysis of the putative binding sites of PUM2, and their evolution across mammals. Figure 4.3(B) shows the prediction scores from the RNATracker model trained on PUM2 CLIP-seq data for the 58 mammals and their ancestors. Figure 4.3(C) shows the likelihood ratio of each branch in the phylogenetic tree, as computed by PhyloPGM. We observe that the log-likelihood ratio scores of longer branches (e.g. rodents) are higher than those obtained for the relatively shorter branches (e.g. primates) in the highly conserved regions. This result could imply that regulatory functions are maintained at different levels in a species and its parent across the phylogenetic tree. More investigation into species analysis with PhyloPGM should reveal relevant information regarding the evolution of regulatory regions and its impact on human.

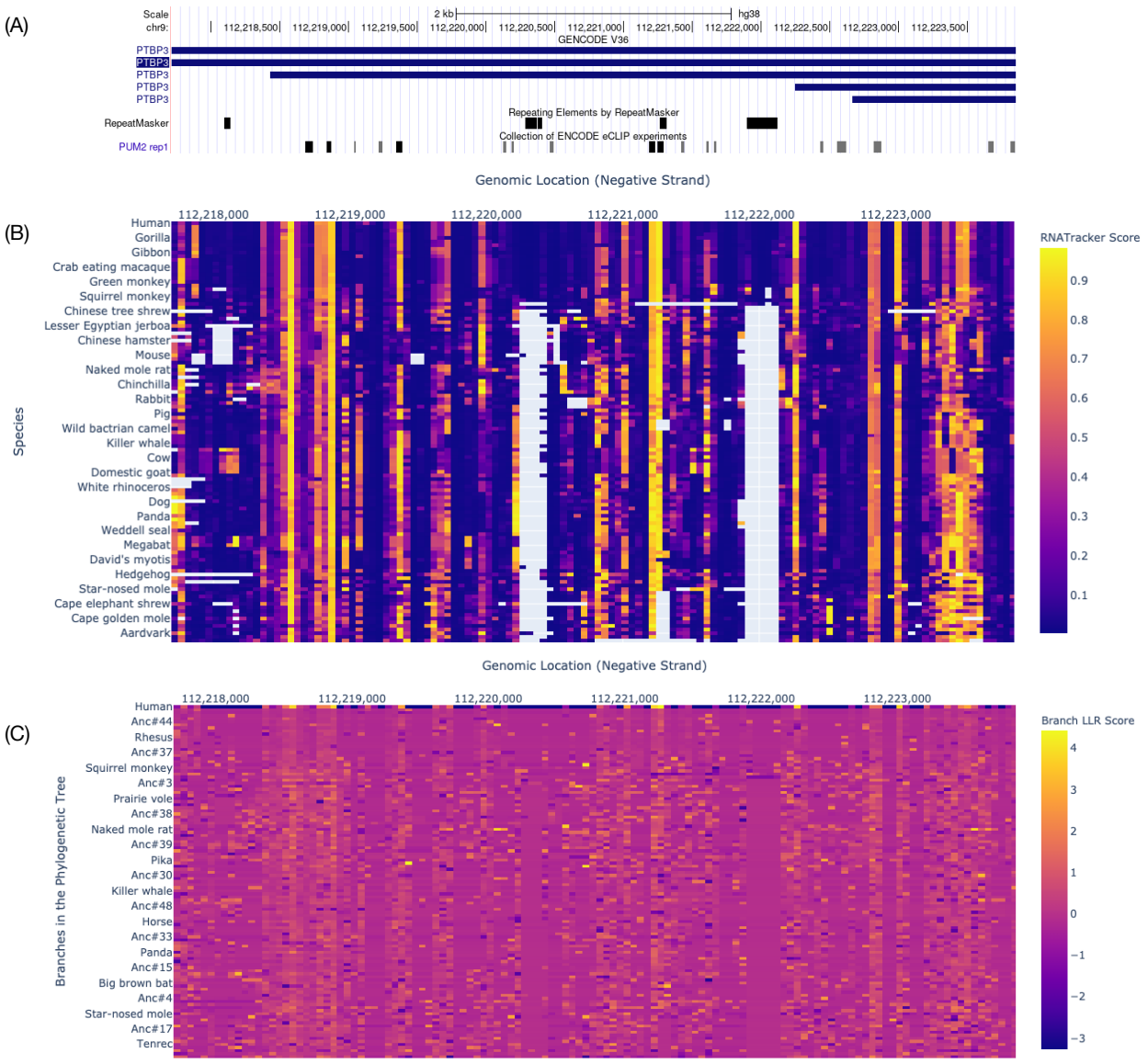


Figure 4.3: Comparison of eCLIP peaks of PUM2 (K562 cells, replicate 1) from UCSC genome browser (A) and heat maps obtained from PhyloPGM-Web for the 3'UTR of PTBP3. (B) The heat map shows the RNATracker prediction scores on the 58 mammals and their ancestors obtained from a RNATracker model trained on PUM2 CLIP-seq data. The gray regions are the missing ortholgs. The two genomic regions with missing orthologs outside primates correspond to two Alu transposable elements. (C) Heat map showing the log-likelihood ratio of each branch in the phylogenetic tree. The row "Human" is the likelihood ratio of human species as per the PhyloPGM approach.

4.5 Methods

4.5.1 Datasets

For the TF binding prediction task, we use one of the ENCODE-DREAM challenge data [Kundaje et al., 2021] (available at `https://www.synapse.org/#!Synapse:syn6131484/wiki/402026`) that consist of 13 TF/cell-type pairs (12 TFs and three cell types:liver, PC-3 and induced pluripotent stem cell) to train RNATracker [Yan et al., 2019] models for TFBS prediction problem. The training sets consist of human genomic sequences of 200 bp from each chromosome except for chromosomes 1, 8 and 21 and test sets are from chromosomes 1, 8 and 21. The labels of both training and test sets are from same cell types. The average number of test examples per TF is $\sim$8 million. In our study, we sub-sample negative examples (sequences with no binding sites for a given TF) in training sets to match the number of positive examples. The average number of training examples after sub-sampling is 400,482 of which 20% is set aside as a validation set. We elongate each sequence from both sides to get a sequence of 1000 bp because binding sites of a given TF may shift in orthologous regions.

We further expand the list of TF binding predictors on PhyloPGM-Web by including the 422 ChIP-seq datasets available at `http://cnn.csail.mit.edu/motif\_occupancy/`. The dataset was originally produced by ENCODE [Consortium et al., 2012] and assembled by Zeng et al. [2016]. Each example is a 101 bp human genomic region, where positive examples are centered on a ChIP-Seq peak and negative examples are randomly selected genomic region with matching GC-content and motif-binding affinity as the positive examples. The average number of examples in the training set is 68,044 with a minimum of 600 and a maximum of 692,340. The average number of examples in the test set is 17,012 with a minimum of 150 and a maximum of 173,086. Again, each example is extended to 1000 bp. We exclude training examples that overlap by even 1 bp with examples from the test set using BEDOPS tool [Neph et al., 2012]. The combination of both TFBS dataset provides 435 TF predictors to the PhyloPGM-Web.

For the RBP binding prediction task, we use 31 CLIP-Seq datasets from Strazar et al. [2016] to train RNATracker. The data is from cell types HEK293, HeLa and U266. The train and test sets consist of approximately 30,000 and 10,000 human genomic sequences, where each sequence is 101 bps long. Similar to TFBS problem, we sub-sample negative examples to obtain a balanced dataset.

Each RNATracker model is trained on human training examples with early stopping using a validation set. In all the three datasets mentioned above, the genomic regions are mapped from the hg19 assembly to the hg38 assembly using liftover. We extract orthologous regions of human sequences in 58 mammals using a 100-way vertebrate whole-genome alignment from the UCSC genome browser [Kent et al., 2002]. The extant and ancestral orthologous regions are extracted using mafsInRegion program (`https://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/mafsInRegion`). Additionally, we complement the orthologous regions with computationally predicted ancestral sequences from Ancestor1.0 [Diallo et al., 2007]. We ignore an orthologous sequence whose length is less than 70% of their human counterpart. The trained models compute orthologous prediction scores that are aggregated using PhyloPGM approach (see Chapter 3 Methods). A PhyloPGM score above zero indicates that the TF or RBP binds in the given location and below zero indicates otherwise. Figure 4.4 shows the tree used in our study. We compute FDR as the ratio of false positive to the sum of true positive and false positives. The true positives are the number of correctly predicted binding sites and false positives are the number of examples incorrectly predicted as binding sites at a given threshold.

Figure 4.4: The phylogenetic tree of 58 mammals and their ancestors that we use in our study.

4.5.2 Implementation

PhyloPGM-Web is implemented with cloudgene [Schönherr et al., 2012], nginx, BASH and python 3.8.3 on a linux platform and is available at `http://webtext.cs.mcgill.ca`. We used plotly [Inc., 2015] to create interactive plots. All accuracy measures are computed using scikit-learn [Pedregosa et al., 2011a].

4.6 Discussion and Conclusion

We present a user-friendly web-interface for PhyloPGM approach to predict the TFs or RBPs that bind to a given genomic location. Although we use RNATracker as base model in the PhyloPGM approach, other predictors could easily be used. Using PhyloPGM-Web, a user can compare the binding preferences of 115 TFs and 31 RBPs and perform phylogenetic analysis for 58 mammals and their ancestors with interactive plots. In the future, more models for more TFs and RBPs will be added, using ChIP-seq and eCLIP data. Furthermore, we will explore motif analysis and model interpretation. Other areas to explore with the web-interface is to facilitate other sequence function prediction tasks such as protein-function prediction, mRNA sub-cellular localization, and micro-RNA target binding sites prediction. We will also extend PhyloPGM-Web to use a recently published 200-mammal alignment [Armstrong et al., 2020]. Finally, one more useful addition will be to allow users to train their own base models or to submit their trained models that can be used with the PhyloPGM approach.

Supplementary Materials

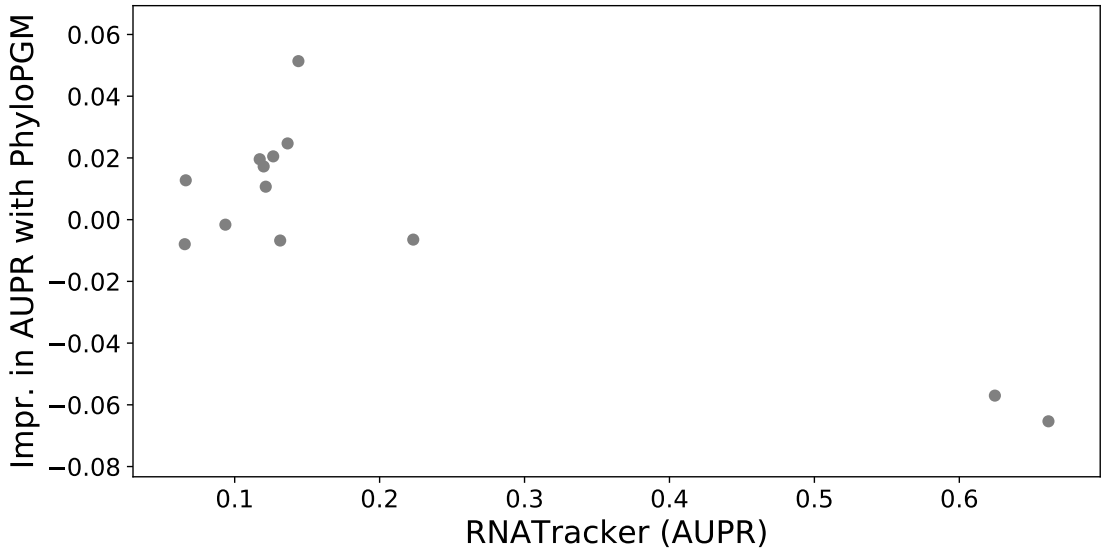


Figure 4.5: Scatter plot of AUPR scores of RNATracker and improvement in AUPR score with PhyloPGM for the 13 ChIP-Seq data from the ENCODE-DREAM challenge [Kundaje et al., 2021].

Table 4.1 PhyloPGM Results with ENCODE-DREAM Challenge

Transcription Factor	Cell Type	RNATracker(AUPR)	PhyloPGM(AUPR)
CTCF	PC-3	0.62	0.57
CTCF	IPSC	0.66	0.60
E2F1	K562	0.14	0.20
EGR1	Liver	0.07	0.08
FOXA1	Liver	0.09	0.09
FOXA2	Liver	0.13	0.12
GABPA	Liver	0.14	0.16
HNF4A	Liver	0.22	0.22
JUND	Liver	0.12	0.14
MAX	Liver	0.13	0.15
NANOG	IPSC	0.07	0.06
REST	Liver	0.12	0.14
TAF1	Liver	0.12	0.13

Table 4.1: Test AUPR scores of RNATracker and PhyloPGM on the 13 ChIP-seq data of ENCODE-DREAM challenge [Kundaje et al., 2021].

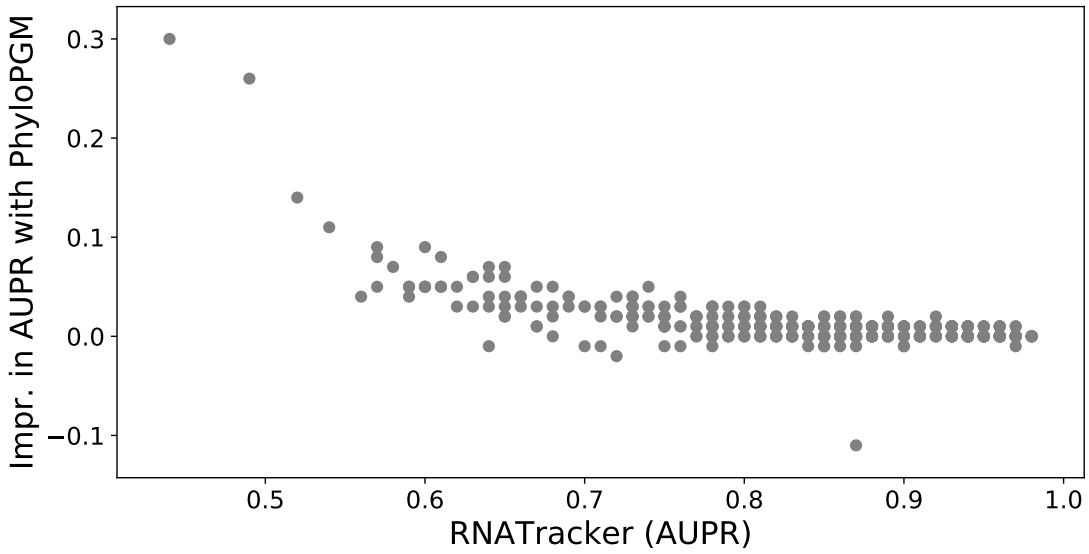


Figure 4.6: Scatter plot of AUPR scores of RNATracker and improvement in AUPR score with PhyloPGM for the MIT-CSAIL datasets [Zeng et al., 2016].

Table 4.2 PhyloPGM Results with MIT-CSAIL datasets

Transcription Factor	Cell Type	RNATracker(AUPR)	PhyloPGM(AUPR)
CTCF	Dnd41	0.96	0.96
CTCF	GM12878	0.93	0.94
CTCF	H1-hESC	0.93	0.93
CTCF	HeLa-S3	0.96	0.96
CTCF	HepG2	0.96	0.97
CTCF	HMEC	0.96	0.96
CTCF	HSMM	0.93	0.94
CTCF	HSMMtube	0.92	0.92
CTCF	HUVEC	0.94	0.95
CTCF	K562	0.95	0.95
CTCF	NH-A	0.93	0.94
CTCF	NHDF-Ad	0.96	0.97
CTCF	NHEK	0.93	0.94
CTCF	NHLF	0.97	0.97
CTCF	Osteobl	0.96	0.96
ATF3	A549	0.65	0.71
CREB1_(SC-240)	A549	0.88	0.88
CTCF_(SC-5916)	A549	0.97	0.98
CTCF_(SC-5916)	A549	0.94	0.95
ELF1_(SC-631)	A549	0.84	0.85
ETS1	A549	0.91	0.91
FOSL2	A549	0.9	0.91
FOXA1_(SC-101058)	A549	0.83	0.83
GR	A549	0.54	0.65
GR	A549	0.74	0.76

GR	A549	0.72	0.76
GR	A549	0.8	0.81
NRSF	A549	0.77	0.79
TCF12	A549	0.85	0.85
USF-1	A549	0.74	0.77
USF-1	A549	0.73	0.75
USF-1	A549	0.8	0.81
YY1_(SC-281)	A549	0.87	0.87
ZBTB33	A549	0.85	0.86
CTCF_(SC-5916)	ECC-1	0.97	0.97
ERalpha_a	ECC-1	0.64	0.71
ERalpha_a	ECC-1	0.66	0.7
ERalpha_a	ECC-1	0.62	0.67
FOXA1_(SC-6553)	ECC-1	0.79	0.82
GR	ECC-1	0.56	0.6
ATF3	GM12878	0.44	0.74
BATF	GM12878	0.87	0.87
CEBPB_(SC-150)	GM12878	0.65	0.72
EBF1_(SC-137065)	GM12878	0.75	0.76
Egr-1	GM12878	0.83	0.83
ELF1_(SC-631)	GM12878	0.86	0.86
ETS1	GM12878	0.87	0.87
MEF2A	GM12878	0.84	0.85
MEF2C_(SC-13268)	GM12878	0.86	0.86
NFATC1_(SC-17834)	GM12878	0.59	0.64
NFIC_(SC-81335)	GM12878	0.86	0.87
NRSF	GM12878	0.73	0.75

PAX5-C20	GM12878	0.82	0.83
PAX5-N19	GM12878	0.82	0.83
POU2F2	GM12878	0.73	0.75
PU.1	GM12878	0.94	0.94
RUNX3_(SC-101553)	GM12878	0.85	0.86
RXRA	GM12878	0.64	0.67
SP1	GM12878	0.88	0.88
SRF	GM12878	0.78	0.79
STAT5A_(SC-74442)	GM12878	0.83	0.84
TCF12	GM12878	0.79	0.79
TCF3_(SC-349)	GM12878	0.8	0.81
USF-1	GM12878	0.78	0.79
YY1_(SC-281)	GM12878	0.84	0.85
ZBTB33	GM12878	0.85	0.87
ZEB1_(SC-25388)	GM12878	0.78	0.79
PAX5-C20	GM12891	0.83	0.83
POU2F2	GM12891	0.77	0.78
PU.1	GM12891	0.93	0.93
YY1_(SC-281)	GM12891	0.88	0.88
PAX5-C20	GM12892	0.84	0.85
YY1	GM12892	0.87	0.88
ATF3	H1-hESC	0.75	0.77
CTCF_(SC-5916)	H1-hESC	0.94	0.94
Egr-1	H1-hESC	0.83	0.83
FOSL1_(SC-183)	H1-hESC	0.76	0.75
JunD	H1-hESC	0.7	0.73
NRSF	H1-hESC	0.73	0.76

POU5F1_(SC-9081)	H1-hESC	0.6	0.65
RXRA	H1-hESC	0.64	0.68
SP1	H1-hESC	0.86	0.86
SP2_(SC-643)	H1-hESC	0.92	0.93
SP4_(V-20)	H1-hESC	0.88	0.88
SRF	H1-hESC	0.79	0.79
TCF12	H1-hESC	0.81	0.82
TEAD4_(SC-101184)	H1-hESC	0.85	0.85
USF-1	H1-hESC	0.86	0.85
YY1_(SC-281)	H1-hESC	0.84	0.85
YY1_(SC-281)	HCT-116	0.86	0.87
ZBTB33	HCT-116	0.88	0.89
NRSF	HeLa-S3	0.7	0.73
ATF3	HepG2	0.85	0.85
BHLHE40	HepG2	0.66	0.7
CEBPB_(SC-150)	HepG2	0.75	0.76
CEBPD_(SC-636)	HepG2	0.66	0.7
CTCF_(SC-5916)	HepG2	0.95	0.95
ELF1_(SC-631)	HepG2	0.85	0.84
FOSL2	HepG2	0.81	0.81
FOXA1_(SC-101058)	HepG2	0.86	0.86
FOXA1_(SC-6553)	HepG2	0.88	0.88
FOXA2_(SC-6554)	HepG2	0.87	0.87
HNF4A_(SC-8987)	HepG2	0.89	0.9
HNF4G_(SC-6558)	HepG2	0.89	0.89
JunD	HepG2	0.63	0.66
MYBL2_(SC-81192)	HepG2	0.8	0.81

NFIC_(SC-81335)	HepG2	0.8	0.82
NRSF	HepG2	0.67	0.7
NRSF	HepG2	0.85	0.85
RXRA	HepG2	0.8	0.82
SP1	HepG2	0.83	0.84
SP2_(SC-643)	HepG2	0.8	0.83
SRF	HepG2	0.82	0.82
TCF12	HepG2	0.52	0.66
TEAD4_(SC-101184)	HepG2	0.74	0.76
USF-1	HepG2	0.82	0.83
YY1_(SC-281)	HepG2	0.88	0.88
ZBTB33	HepG2	0.88	0.89
ZBTB7A_(SC-34508)	HepG2	0.77	0.79
ATF3	K562	0.89	0.91
CEBPB_(SC-150)	K562	0.77	0.77
CTCF_(SC-5916)	K562	0.97	0.96
E2F6	K562	0.85	0.85
Egr-1	K562	0.9	0.9
ELF1_(SC-631)	K562	0.84	0.85
ETS1	K562	0.86	0.86
FOSL1_(SC-183)	K562	0.72	0.74
GATA2_(SC-267)	K562	0.79	0.79
Max	K562	0.78	0.78
MEF2A	K562	0.67	0.72
NRSF	K562	0.78	0.79
PU.1	K562	0.81	0.82
SP1	K562	0.87	0.88

SP2_(SC-643)	K562	0.92	0.92
SRF	K562	0.72	0.7
STAT5A_(SC-74442)	K562	0.65	0.67
TEAD4_(SC-101184)	K562	0.72	0.74
THAP1_(SC-98174)	K562	0.9	0.91
USF-1	K562	0.85	0.85
YY1_(SC-281)	K562	0.89	0.9
YY1	K562	0.84	0.84
ZBTB33	K562	0.84	0.85
ZBTB7A_(SC-34508)	K562	0.8	0.8
NRSF	PANC-1	0.65	0.69
FOXP2	PFSK-1	0.86	0.87
NRSF	PFSK-1	0.82	0.82
FOXP2	SK-N-MC	0.87	0.88
NRSF	SK-N-SH	0.82	0.83
NRSF	SK-N-SH	0.78	0.81
CTCF	SK-N-SH_RA	0.98	0.98
USF1_(SC-8983)	SK-N-SH_RA	0.92	0.92
YY1_(SC-281)	SK-N-SH$_RA$	0.87	0.89
CTCF_(SC-5916)	T-47D	0.94	0.94
ERalpha_a	T-47D	0.57	0.66
ERalpha_a	T-47D	0.6	0.65
ERalpha_a	T-47D	0.6	0.65
FOXA1_(SC-6553)	T-47D	0.87	0.87
GATA3_(SC-268)	T-47D	0.78	0.78
NRSF	U87	0.83	0.84
BHLHE40	A549	0.68	0.73

CEBPB	A549	0.94	0.94
Max	A549	0.82	0.82
NFKB	GM10847	0.87	0.87
BHLHE40_(NB100-1800)	GM12878	0.81	0.82
c-Fos	GM12878	0.78	0.81
CTCF_(SC-15914)	GM12878	0.96	0.96
E2F4	GM12878	0.89	0.9
EBF1_(SC-137065)	GM12878	0.81	0.81
ELK1_(1277-1)	GM12878	0.89	0.9
JunD	GM12878	0.58	0.65
Max	GM12878	0.85	0.84
NF-E2_(SC-22827)	GM12878	0.49	0.75
NFKB	GM12878	0.89	0.89
NF-YA	GM12878	0.86	0.88
NF-YB	GM12878	0.86	0.86
Nrf1	GM12878	0.88	0.89
RFX5_(200-401-194)	GM12878	0.73	0.77
STAT1	GM12878	0.76	0.79
STAT3	GM12878	0.8	0.83
TBP	GM12878	0.73	0.76
TR4	GM12878	0.75	0.78
USF2	GM12878	0.88	0.89
YY1	GM12878	0.82	0.84
Znf143_(16618-1-AP)	GM12878	0.92	0.94
NFKB	GM12891	0.89	0.89
NFKB	GM12892	0.83	0.84
NFKB	GM15510	0.85	0.86

NFKB	GM18505	0.85	0.86
NFKB	GM18526	0.8	0.82
NFKB	GM18951	0.85	0.86
NFKB	GM19099	0.84	0.85
NFKB	GM19193	0.87	0.88
Bach1_(sc-14700)	H1-hESC	0.81	0.81
CEBPB	H1-hESC	0.87	0.87
c-Jun	H1-hESC	0.82	0.84
c-Myc	H1-hESC	0.87	0.88
JunD	H1-hESC	0.77	0.79
MafK_(ab50322)	H1-hESC	0.97	0.97
Max	H1-hESC	0.81	0.82
Nrf1	H1-hESC	0.89	0.9
RFX5_(200-401-194)	H1-hESC	0.57	0.65
TBP	H1-hESC	0.83	0.85
USF2	H1-hESC	0.91	0.91
Znf143_(16618-1-AP)	H1-hESC	0.85	0.86
TCF7L2	HCT-116	0.81	0.82
ELK4	HEK293	0.91	0.91
TCF7L2	HEK293	0.84	0.85
ZNF263	HEK293-T-REx	0.84	0.83
AP-2alpha	HeLa-S3	0.77	0.77
AP-2gamma	HeLa-S3	0.81	0.81
CEBPB	HeLa-S3	0.82	0.83
c-Fos	HeLa-S3	0.86	0.88
c-Jun	HeLa-S3	0.92	0.92
c-Myc	HeLa-S3	0.81	0.83

E2F1	HeLa-S3	0.85	0.87
E2F4	HeLa-S3	0.88	0.89
E2F6	HeLa-S3	0.89	0.9
ELK1_(1277-1)	HeLa-S3	0.9	0.91
ELK4	HeLa-S3	0.91	0.91
HA-E2F1	HeLa-S3	0.84	0.85
JunD	HeLa-S3	0.89	0.9
MafK_(ab50322)	HeLa-S3	0.96	0.96
Max	HeLa-S3	0.82	0.82
NF-YA	HeLa-S3	0.86	0.86
NF-YB	HeLa-S3	0.83	0.84
Nrf1	HeLa-S3	0.88	0.89
PRDM1_(9115)	HeLa-S3	0.73	0.76
RFX5_(200-401-194)	HeLa-S3	0.64	0.7
STAT1	HeLa-S3	0.78	0.79
STAT3	HeLa-S3	0.81	0.83
TBP	HeLa-S3	0.78	0.81
TCF7L2_C9B9_(2565)	HeLa-S3	0.81	0.82
TCF7L2	HeLa-S3	0.73	0.77
TR4	HeLa-S3	0.74	0.79
USF2	HeLa-S3	0.87	0.87
Znf143_(16618-1-AP)	HeLa-S3	0.82	0.84
ARID3A_(NB100-279)	HepG2	0.73	0.75
BHLHE40_(NB100-1800)	HepG2	0.78	0.79
CEBPB	HepG2	0.78	0.79
CEBPB	HepG2	0.93	0.93
c-Jun	HepG2	0.95	0.95

ERRA	HepG2	0.68	0.71
GRp20	HepG2	0.87	0.76
HNF4A	HepG2	0.81	0.82
HSF1	HepG2	0.7	0.69
JunD	HepG2	0.92	0.93
MafF_(M8194)	HepG2	0.97	0.97
MafK_(ab50322)	HepG2	0.98	0.98
MafK_(SC-477)	HepG2	0.98	0.98
Max	HepG2	0.78	0.79
Nrf1	HepG2	0.87	0.89
RFX5_(200-401-194)	HepG2	0.69	0.72
SREBP1	HepG2	0.81	0.84
TBP	HepG2	0.76	0.79
TCF7L2	HepG2	0.6	0.69
TR4	HepG2	0.76	0.8
USF2	HepG2	0.91	0.92
c-Fos	HUVEC	0.9	0.91
c-Jun	HUVEC	0.86	0.87
GATA-2	HUVEC	0.87	0.88
Max	HUVEC	0.85	0.85
CEBPB	IMR90	0.9	0.91
CTCF_(SC-15914)	IMR90	0.94	0.95
MafK_(ab50322)	IMR90	0.96	0.96
ARID3A_(sc-8821)	K562	0.61	0.69
ATF1_(06-325)	K562	0.88	0.89
ATF3	K562	0.75	0.74
Bach1_(sc-14700)	K562	0.73	0.74

BHLHE40_(NB100-1800)	K562	0.8	0.8
CEBPB	K562	0.89	0.89
c-Fos	K562	0.61	0.66
c-Jun	K562	0.63	0.69
c-Jun	K562	0.65	0.67
c-Jun	K562	0.71	0.74
c-Jun	K562	0.63	0.69
c-Jun	K562	0.68	0.7
c-Myc	K562	0.81	0.83
c-Myc	K562	0.83	0.84
c-Myc	K562	0.8	0.81
c-Myc	K562	0.81	0.82
c-Myc	K562	0.83	0.83
c-Myc	K562	0.84	0.85
CTCF_(SC-15914)	K562	0.93	0.94
E2F4	K562	0.86	0.87
E2F6	K562	0.87	0.87
ELK1_(1277-1)	K562	0.91	0.92
GATA-1	K562	0.59	0.64
GATA-2	K562	0.62	0.65
IRF1	K562	0.72	0.74
IRF1	K562	0.76	0.77
IRF1	K562	0.78	0.79
IRF1	K562	0.88	0.89
JunD	K562	0.75	0.76
MafF_(M8194)	K562	0.97	0.97
MafK_(ab50322)	K562	0.98	0.98

Max	K562	0.78	0.79
NF-E2	K562	0.9	0.89
NF-YA	K562	0.87	0.87
NF-YB	K562	0.87	0.86
Nrf1	K562	0.91	0.92
RFX5_(200-401-194)	K562	0.69	0.73
STAT1	K562	0.64	0.63
STAT1	K562	0.68	0.68
STAT1	K562	0.66	0.69
STAT1	K562	0.67	0.68
TBP	K562	0.79	0.81
TR4	K562	0.79	0.8
USF2	K562	0.79	0.8
YY1	K562	0.81	0.83
Znf143_(16618-1-AP)	K562	0.71	0.74
ZNF263	K562	0.71	0.7
c-Fos	MCF10A-Er-Src	0.9	0.9
c-Fos	MCF10A-Er-Src	0.91	0.91
c-Fos	MCF10A-Er-Src	0.89	0.89
c-Fos	MCF10A-Er-Src	0.92	0.93
c-Myc	MCF10A-Er-Src	0.84	0.85
c-Myc	MCF10A-Er-Src	0.87	0.88
E2F4	MCF10A-Er-Src	0.88	0.88
STAT3	MCF10A-Er-Src	0.93	0.93
STAT3	MCF10A-Er-Src	0.94	0.94
STAT3	MCF10A-Er-Src	0.88	0.89
STAT3	MCF10A-Er-Src	0.93	0.94

STAT3	MCF10A-Er-Src	0.94	0.94
GATA3_(SC-269)	MCF-7	0.76	0.77
GATA3_(SC-268)	MCF-7	0.57	0.62
HA-E2F1	MCF-7	0.87	0.87
TCF7L2	MCF-7	0.74	0.77
c-Myc	NB4	0.84	0.84
Max	NB4	0.79	0.79
YY1	NT2-D1	0.91	0.91
TCF7L2	PANC-1	0.82	0.83
GATA-1	PBDEFetal	0.59	0.63
GATA-1	PBDE	0.75	0.76
GATA-2	SH-SY5Y	0.86	0.86
GATA3_(SC-269)	SH-SY5Y	0.87	0.87
eGFP-FOS	K562	0.75	0.77
eGFP-GATA2	K562	0.78	0.77
eGFP-JunB	K562	0.61	0.66
eGFP-JunD	K562	0.65	0.68
CTCF	A549	0.95	0.96
CTCF	Fibrobl	0.95	0.95
CTCF	Gliobla	0.96	0.96
c-Myc	GM12878	0.9	0.89
CTCF	GM12878	0.97	0.97
CTCF	GM12891	0.94	0.94
CTCF	GM12892	0.94	0.94
CTCF	GM19238	0.94	0.94
CTCF	GM19239	0.95	0.95
CTCF	GM19240	0.95	0.95

c-Myc	H1-hESC	0.67	0.68
CTCF	H1-hESC	0.96	0.96
c-Myc	HeLa-S3	0.69	0.73
CTCF	HeLa-S3	0.97	0.97
c-Myc	HepG2	0.75	0.77
CTCF	HepG2	0.98	0.98
c-Myc	HUVEC	0.82	0.82
CTCF	HUVEC	0.98	0.98
c-Myc	K562	0.86	0.86
CTCF	K562	0.94	0.94
c-Myc	MCF-7	0.71	0.73
c-Myc	MCF-7	0.8	0.8
c-Myc	MCF-7	0.77	0.79
c-Myc	MCF-7	0.81	0.82
CTCF	MCF-7	0.94	0.94
CTCF	MCF-7	0.92	0.93
CTCF	MCF-7	0.94	0.94
CTCF	MCF-7	0.94	0.94
CTCF	MCF-7	0.95	0.96
CTCF	NHEK	0.96	0.96
CTCF	ProgFib	0.93	0.94
CTCF	A549	0.93	0.94
CTCF	AG04449	0.96	0.96
CTCF	AG04450	0.93	0.94
CTCF	AG09309	0.94	0.94
CTCF	AG09319	0.96	0.96
CTCF	AG10803	0.96	0.96

CTCF	AoAF	0.93	0.94
CTCF	BE2_C	0.93	0.94
CTCF	BJ	0.92	0.93
CTCF	Caco-2	0.93	0.94
CTCF	GM06990	0.94	0.94
CTCF	GM12801	0.78	0.8
CTCF	GM12864	0.97	0.97
CTCF	GM12865	0.97	0.97
CTCF	GM12872	0.95	0.95
CTCF	GM12873	0.94	0.95
CTCF	GM12874	0.94	0.95
CTCF	GM12875	0.93	0.94
CTCF	GM12878	0.93	0.94
CTCF	HAc	0.96	0.96
CTCF	HA-sp	0.95	0.95
CTCF	HBMEC	0.93	0.93
CTCF	HCFaa	0.93	0.94
CTCF	HCM	0.96	0.96
CTCF	HCPEpiC	0.96	0.96
CTCF	HCT-116	0.94	0.94
CTCF	HEEpiC	0.96	0.97
CTCF	HEK293	0.95	0.95
CTCF	HeLa-S3	0.95	0.95
CTCF	HepG2	0.94	0.94
CTCF	HFF	0.97	0.98
CTCF	HFF-Myc	0.94	0.95
CTCF	HL-60	0.9	0.91

CTCF	HMEC	0.93	0.94
CTCF	HMF	0.94	0.95
CTCF	HPAF	0.96	0.97
CTCF	HPF	0.94	0.95
CTCF	HRE	0.93	0.94
CTCF	HRPEpiC	0.97	0.97
CTCF	HUVEC	0.94	0.94
CTCF	HVMF	0.94	0.95
CTCF	K562	0.97	0.97
CTCF	MCF-7	0.97	0.97
CTCF	NB4	0.95	0.96
CTCF	NHDF-neo	0.93	0.93
CTCF	NHEK	0.93	0.93
CTCF	NHLF	0.96	0.96
CTCF	RPTEC	0.96	0.96
CTCF	SAEC	0.95	0.95
CTCF	SK-N-SH_RA	0.95	0.95
CTCF	WERI-Rb-1	0.96	0.97
CTCF	WI-38	0.98	0.98

Table 4.2: Test AUPR scores of RNATracker and PhyloPGM on the MIT-CSAIL datasets [Zeng et al., 2016]

Chapter 5

Conclusion

Deep learning based approaches have improved the prediction accuracy of TF and RBP binding site predictors, but remain far from replacing the wet-lab experiments [Kundaje et al., 2021, Pan et al., 2018]. Many studies have shown that biological functions associated with a DNA or RNA sequences are mostly conserved. However, computational models to predict a function associated with a sequence e.g. TF or RBP binding site prediction, seldomly utilize sequence conservation information.

Comparative genomics techniques offer a wide range of data and tools to explore functions associated with genomic sequences. The multiple sequence alignment algorithms allow to identify similar regions within a group of species. The ancestral reconstructions algorithms can provide orthologous regions from extinct species and reveal valuable information about sequence function. One of the major challenges to the integration of such evolutionary information with sequence binding prediction models is binding sites turnover phenomenon, where a genomic region may maintain its binding property e.g. number of binding sites, while the region itself may not be conserved during the evolution. More sophisticated approaches that utilize phylogenetic relationship and ancestral sequences are shown to counter such effects [Sadri et al., 2011, Blanchette, 2012, Leclercq et al., 2017].

In this thesis, we develop a semi-supervised learning approach, PhyloReg, that uses labeled human examples and unlabeled orthologs to produce a robust TFBS predictor. Then, we present a probabilistic aggregation approach, PhyloPGM, that can boost the prediction accuracy of previously trained RNA-RBP binding predictor. Finally, we provide a web interface for PhyloPGM that can predict the binding of TFs and RBPs to given genomic locations.

5.1 Summary of Contributions

To our knowledge, the methodologies developed in our study are first of a kind that bridge machine learning and evolution for studying regulation and they will, hopefully, serve as basis for valuable future developments.

In chapter 2, we address the problem of predicting whether a given genomic location binds to a TF in a cell-type by using evolutionary information. We develop a semi-supervised regularization approach called PhyloReg approach. The major advantage of PhyloReg is that it allows to use the vast amount of unlabelled examples (orthologous regions in extant and ancestral species) with machine learning techniques to build a robust model. The orthologous regions of human genomic regions are relatively cheaper to obtain compared to the labelled examples that require expensive wet-lab experiments. Although PhyloReg loss is simple to integrate with a loss function of a supervised learning model, the learning of combined loss can become computationally intensive with a large amount of unlabelled data. We use heuristics like updating the PhyloReg loss at certain intervals to reduce the computation time. One other key observation is that in the cases where the accuracy of a model trained on human data is lower, the amount of improvement with PhyloReg is larger. Finally, PhyloReg assumes that a sequence label (e.g. whether it binds to a TF) is maintained in its orthologs. Therefore, PhyloReg should perform better than the models trained on human data alone when selective pressure is present in the regulatory regions.

In chapter 3, we develop a probabilistic aggregation approach called PhyloPGM for the TF and RBP binding prediction tasks. We show that PhyloPGM boosts the prediction accuracy of previously trained models on human data using evolutionary information. Similar to PhyloReg, the amount of improvement in accuracy with PhyloPGM is more for base models with relatively lower accuracy. Unlike PhyloReg that requires training, PhyloPGM is applicable at inference stage. PhyloPGM integrates the log-likelihood ratio (LLR) of base model prediction scores on the branches of the phylogenetic tree and accuracy improvement with PhyloPGM shall signify the contribution of LLR from each branch. Interestingly, we find that the branches that are relatively evolutionary farther from human have more impact on the PhyloPGM score for the cases where PhyloPGM provides relatively large amount of improvement. Thus, PhyloPGM is able to capture important long-range evolutionary changes. Furthermore, we find that the PhyloPGM is more capable than the base model or a conservation-based approach of predicting genomic regions where alterations can cause a disease. Therefore, PhyloPGM can be used to analyze the impact of mutation in a given genomic region. For example, the predictions of a base model on a reference genome and a genome from an individual can be compared. The PhyloPGM score on the regions with prediction discrepancies can be used to identify deleterious mutations.

The major challenges with PhyloPGM is its dependence on a pretrained predictor and computation involving large alignment files. In chapter 4, we present PhyloPGM-Web, a user-friendly web interface to PhyloPGM for predicting TF and RBP binding sites. A user may submit genomic regions of interest to find possible TFs or RBPs that can bind from a list of 115 TFs and 31 RBPs. PhyloPGM-Web handles the complex pipeline of PhyloPGM in background so that a may focus on the phylogenetic analysis and binding preferences of the submitted genomic regions.

5.2 Perspectives on Future Work

One major goal in terms of future work with PhyloReg is to explore the aspects related to model interpretability. We suppose that phylogenetic regularization should allow to study the role of evolution on regulatory regions. In one such analysis with PhyloReg, we find that primate species are more relevant for the TF binding prediction in human for certain TFs (e.g. CEBPB, MafK). While distant related species of human are found as relevant as closely related for certain TFs (e.g. Elk1, ATF3). However, more investigation is needed in this regard to have firm conclusions.

Another feature to add in PhyloReg is to allow adaptive weights to each branch in the phylogenetic tree during the computation of PhyloReg loss. It is possible that certain species may be less relevant than others for a function linked with a human genomic location (e.g. TF or RBP binding sites). Different weights with different species should allow PhyloReg to learn more robust models. In the similar direction, we should consider evolutionary distances between species while calculating PhyloReg loss function.

PhyloReg should be explored with other sequence function prediction tasks, such as protein function prediction, RNA localisation, microRNA target sites prediction and with other supervised learning algorithms. Adapting PhyloReg to other sequence function prediction tasks poses a challenge of suitable integration of PhyloReg loss with the loss function of existing models for such problems that may involve some heuristics. For example predicting protein functions often requires building a multi-class predictor [Kulmanov and Hoehndorf, 2020] with a different set of metrics to evaluate model performance e.g. CAFA3 evaluation [Zhou et al., 2019]. Other sequence function prediction tasks such as mRNA subcellular localization requires to predict multiple regression values that represent expression distribution across several cellular fraction for a given input [Yan et al., 2019]. Additionally, the training data of other sequence function prediction tasks may be highly imbalanced, especially in multi-class setting, where some classes

may have most examples and others relatively few. Learning PhyloReg with such data will require necessary adjustments and caliberations.

One of the core components of PhyloPGM is to estimate conditional probability of a descendant species prediction score given its parent species' prediction score and human label using multinomial distribution. We should explore other distributions e.g. univariate normal distribution or beta distribution to estimate such conditional probabilities to better fit the underlying conditional distribution. Another aspect to explore with PhyloPGM approach is to use a subset of species rather than the entire mammalian orthologs to reduce the computation time. This may increase or decrease the accuracy as we have observed different species have differing impact on PhyloPGM accuracy for certain TFs or RBPs. We implement PhyloPGM as a binary classifier in our study. One possible direction in future work should include extension of PhyloPGM approach to regression and muli-class prediction problems. This will allow PhyloPGM to be applicable to other sequence function prediction tasks that require regression (e.g. mRNA subcellular localization) or multi-class (e.g. protein function prediction) predictors.

We plan to add more TF and RBP binding site predictors to PhyloPGM-Web in future. The present version of web-interface uses RNATracker as base model and an interesting extension will be to allow users to submit their own pretrained base models. In the similar direction, we should enhance the web-interface to handle other sequence function prediction tasks. This will add several benefits to PhyloPGM-Web application such as comparison of newly developed base models, staying up to date with the state-of-the-art and studying evolution of sequence function based on the base model features. However, adding the feature of allowing users' pretrained base models will require to solve scalability and security challenges. Nevertheless, a user submitting a desired base model for a designated sequence function prediction task to get the evolutionary insights and improved prediction scores will indeed be a useful application of PhyloPGM-Web.

With the sequencing of new genomes (eg. [Armstrong et al., 2020]), we can expect PhyloReg and PhyloPGM to be more powerful in producing more robust models for sequence

function prediction tasks as more evolutionary evidences will be available. However, sufficient amount of selection pressure is required to be present in data for the effective working of PhyloReg and PhyloPGM. The availability of more genomic data not only present computational challenge of integrating vast amount of data with the complex processing of PhyloReg or PhyloPGM, but also an intricate challenge of distinguishing between noise and evolutionary signatures related to sequence functions.

Finally, we aim to design model interpretation techniques and more phylogenetic operations in PhyloPGM-Web. Some of the useful additions will be to represent the identified motifs in the input sequence as sequence logo [Schneider and Stephens, 1990] and to show the impact of various subset of species on the PhyloPGM score. Furthermore, allowing user to submit synthetic sequences or a genomic region from an individual's genome as input should help studying the effect of mutation and, hopefully, should be beneficial for therapeutic developments.

Bibliography

Federico Agostini, Andreas Zanzoni, Petr Klus, Domenica Marchese, Davide Cirillo, and Gian Gaetano Tartaglia. cat rapid omics: a web server for large-scale prediction of protein–rna interactions. *Bioinformatics*, 29(22):2928–2930, 2013.

Federico Agostini, Davide Cirillo, Riccardo Delli Ponti, and Gian Gaetano Tartaglia. Seamote: a method for high-throughput motif discovery in nucleic acid sequences. *BMC genomics*, 15(1):1–9, 2014.

Faizy Ahsan, Alexandre Drouin, François Laviolette, Doina Precup, and Mathieu Blanchette. Phylogenetic manifold regularization: A semi-supervised approach to predict transcription factor binding sites. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 62–66. IEEE, 2020.

Faizy Ahsan, Zichao Yan, Doina Precup, and Mathieu Blanchette. Phylopgm: Boosting regulatory function prediction accuracy using evolutionary information. manuscript submitted to nature methods. 2021.

Jessica Alföldi and Kerstin Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome research*, 23(7):1063–1068, 2013.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105, 2013.

Maria Anisimova, Gina Cannarozzi, and David A Liberles. Finding the balance between the mathematical and biological optima in multiple sequence alignment. *Trends in Evolutionary Biology*, 2(1):e7–e7, 2010.

Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T Fiddes, Adam M Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020.

Aaron Arvey, Phaedra Agius, William Stafford Noble, and Christina Leslie. Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome research*, 22(9):1723–1734, 2012.

Haim Ashkenazy, Osnat Penn, Adi Doron-Faigenboim, Ofir Cohen, Gina Cannarozzi, Oren Zomer, and Tal Pupko. Fastml: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic acids research*, 40(W1):W580–W584, 2012.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021.

Lu Bai and Alexandre V Morozov. Gene regulation by nucleosome positioning. *Trends in genetics*, 26(11):476–483, 2010.

Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, 34(suppl_2): W369–W373, 2006.

Hagit Bar-Rogovsky, Adrian Hugenmatter, and Dan S Tawfik. The evolutionary origins of detoxifying enzymes: the mammalian serum paraoxonases (pons) relate to bacterial homoserine lactonases. *Journal of Biological Chemistry*, 288(33):23914–23927, 2013.

Tahsin Stefan Barakat, Florian Halbritter, Man Zhang, André F Rendeiro, Elena Perenthaler, Christoph Bock, and Ian Chambers. Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell stem cell*, 23(2):276–288, 2018.

Scott Barolo. How to tune an enhancer. *Proceedings of the National Academy of Sciences*, 113 (23):6330–6331, 2016.

Iros Barozzi, Marta Simonatto, Silvia Bonifacio, Lin Yang, Remo Rohs, Serena Ghisletti, and Gioacchino Natoli. Coregulation of transcription factor binding and nucleosome occupancy through dna features of mammalian enhancers. *Molecular cell*, 54(5):844–857, 2014.

Alex Bateman, Ewan Birney, Lorenzo Cerruti, Richard Durbin, Laurence Etwiller, Sean R Eddy, Sam Griffiths-Jones, Kevin L Howe, Mhairi Marshall, and Erik LL Sonnhammer. The pfam protein families database. *Nucleic acids research*, 30(1):276–280, 2002.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

Sabina Benko, Judy A Fantes, Jeanne Amiel, Dirk-Jan Kleinjan, Sophie Thomas, Jacqueline Ramsay, Negar Jamshidi, Abdelkader Essafi, Simon Heaney, Christopher T Gordon, et al. Highly conserved non-coding elements on either side of sox9 associated with pierre robin sequence. *Nature genetics*, 41(3):359–364, 2009.

Louis Philip Benoit Bouvrette, Samantha Bovaird, Mathieu Blanchette, and Eric Lécuyer. ornament: a database of putative rna binding protein target sites in the transcriptomes of model species. *Nucleic acids research*, 48(D1):D166–D173, 2020.

Eugene Berezikov, Victor Guryev, Ronald HA Plasterk, and Edwin Cuppen. Conreal: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome research*, 14(1):170–178, 2004.

Benjamin P Berman, Barret D Pfeiffer, Todd R Laverty, Steven L Salzberg, Gerald M Rubin, Michael B Eisen, and Susan E Celniker. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in drosophila melanogaster and drosophila pseudoobscura. *Genome biology*, 5(9):R61, 2004.

DIANE Beysen, Jeroen Raes, BP Leroy, A Lucassen, JRW Yates, J Clayton-Smith, H Ilyina, S Sklower Brooks, S Christin-Maitre, Marc Fellous, et al. Deletions involving long-range conserved nongenic sequences upstream and downstream of foxl2 as a novel disease-causing mechanism in blepharophimosis syndrome. *The American Journal of Human Genetics*, 77(2):205–218, 2005.

Marcin Biesiada, Katarzyna J Purzycka, Marta Szachniuk, Jacek Blazewicz, and Ryszard W Adamiak. Automated rna 3d structure prediction with rnacomposer. In *RNA Structure Determination*, pages 199–215. Springer, 2016.

Harry Biggs, Padmini Parthasarathy, Alexandra Gavryushkina, and Paul P Gardner. ncvardb: a manually curated database for pathogenic non-coding variants and benign controls. *Database*, 2020, 2020.

Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16 (1):6–21, 2002.

Mathauieu Blanchette, Abdoulaye Baniré Diallo, Eric D Green, Webb Miller, and David Haussler. Computational reconstruction of ancestral dna sequences. In *Phylogenomics*, pages 171–184. Springer, 2008.

Mathieu Blanchette. Exploiting ancestral mammalian genomes for the prediction of human transcription factor binding sites. *BMC bioinformatics*, 13(19):S2, 2012.

Mathieu Blanchette and Martin Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome research*, 12(5):739–748, 2002.

Mathieu Blanchette, Eric D Green, Webb Miller, and David Haussler. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome research*, 14(12):2412–2423, 2004a.

Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715, 2004b.

Mathieu Blanchette, Alain R Bataille, Xiaoyu Chen, Christian Poitras, Josée Laganière, Céline Lefèbvre, Geneviève Deblois, Vincent Giguère, Vincent Ferretti, Dominique Bergeron, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome research*, 16(5):656–668, 2006.

Dario Boffelli, Jon McAuliffe, Dmitriy Ovcharenko, Keith D Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, 2003.

D Bohmann. Transcription factor phosphorylation: a link between signal transduction and the regulation of gene expression. *Cancer cells (Cold Spring Harbor, NY: 1989)*, 2(11): 337–344, 1990.

Michal J Boniecki, Grzegorz Lach, Wayne K Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M Rother, and Janusz M Bujnicki. Simrna: a coarse-grained method for rna folding simulations and 3d structure prediction. *Nucleic acids research*, 44(7):e63–e63, 2016.

Arnaud Boulling, Linda Wicht, and Daniel F Schorderet. Identification of hmx1 target genes: a predictive promoter model approach. *Molecular vision*, 19:1779, 2013.

James E Bradner, Denes Hnisz, and Richard A Young. Transcriptional addiction in cancer. *Cell*, 168(4):629–643, 2017.

Nick Bray, Inna Dubchak, and Lior Pachter. Avid: A global alignment program. *Genome research*, 13(1):97–102, 2003.

Nicolas Bray and Lior Pachter. Mavid multiple alignment server. *Nucleic acids research*, 31 (13):3525–3526, 2003.

Ali H Brivanlou and James E Darnell. Signal transduction and the control of gene expression. *Science*, 295(5556):813–818, 2002.

Michael Brudno, Chuong B Do, Gregory M Cooper, Michael F Kim, Eugene Davydov, Eric D Green, Arend Sidow, Serafim Batzoglou, NISC Comparative Sequencing Program, et al. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome research*, 13(4):721–731, 2003.

Martha L Bulyk, Philip LF Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30(5):1255–1261, 2002.

Kaspar Burger, Bastian Mühl, Markus Kellner, Michaela Rohrmoser, Anita Gruber-Eber, Lukas Windhager, Caroline C Friedel, Lars Dölken, and Dirk Eick. 4-thiouridine inhibits rrna synthesis and causes a nucleolar stress response. *RNA biology*, 10(10):1623–1630, 2013.

R Busa, MP Paronetto, D Farini, E Pierantozzi, F Botti, DF Angelini, F Attisani, G Vespasiani, and C Sette. The rna-binding protein sam68 contributes to proliferation and survival of human prostate cancer cells. *Oncogene*, 26(30):4372–4382, 2007.

Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M Beckmann, Claudia Strein, Norman E Davey, David T Humphreys, Thomas Preiss, Lars M Steinmetz, et al. Insights into rna biology from an atlas of mammalian mrna-binding proteins. *Cell*, 149(6):1393–1406, 2012.

Jamel Chelly and Jean-Louis Mandel. Monogenic causes of x-linked mental retardation. *Nature Reviews Genetics*, 2(9):669–680, 2001.

Christina TL Chen, Jen C Wang, and Barak A Cohen. The strength of selection on ultraconserved elements in the human genome. *The American Journal of Human Genetics*, 80 (4):692–704, 2007.

Xiaoshu Chen and Jianzhi Zhang. The ortholog conjecture is untestable by the current gene ontology but is supported by rna sequencing data. *PLoS Comput Biol*, 8(11): e1002784, 2012.

Carol Anne Chénard and Stéphane Richard. New implications for the quaking rna binding protein in human disease. *Journal of neuroscience research*, 86(2):233–242, 2008.

Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute hits-clip decodes microrna–mrna interaction maps. *Nature*, 460(7254):479–486, 2009.

Leonid Chindelevitch, Zhentao Li, Eric Blais, and Mathieu Blanchette. On the inference of parsimonious indel evolutionary scenarios. *Journal of bioinformatics and computational biology*, 4(03):721–744, 2006.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A Cohen, and Mark Johnston. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *science*, 301(5629):71–76, 2003.

Paul F Cliften, LaDeana W Hillier, Lucinda Fulton, Tina Graves, Tracie Miner, Warren R Gish, Robert H Waterston, and Mark Johnston. Surveying saccharomyces genomes to identify functional elements by comparative dna sequence analysis. *Genome research*, 11 (7):1175–1186, 2001.

Francis S Collins, Eric D Green, Alan E Guttmacher, and Mark S Guyer. A vision for the future of genomics research. *Nature*, 422(6934):835, 2003.

ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

Mouse Genome Sequencing Consortium et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520, 2002.

Kate B Cook, Hilal Kazan, Khalid Zuberi, Quaid Morris, and Timothy R Hughes. Rbpdb: a database of rna-binding specificities. *Nucleic acids research*, 39(suppl_1):D301–D308, 2010.

Gregory M Cooper and Christopher D Brown. Qualifying the relationship between sequence conservation and molecular function. *Genome research*, 18(2):201–205, 2008.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *ICML*, 2010.

Justin Crocker, Namiko Abe, Lucrezia Rinaldi, Alistair P McGregor, Nicolás Frankel, Shu Wang, Ahmad Alsawadi, Philippe Valenti, Serge Plaza, François Payre, et al. Low affin-

ity binding site clusters confer hox specificity and regulatory robustness. *Cell*, 160(1-2): 191–203, 2015.

Santiago Cuesta-López, Hervé Menoni, Dimitar Angelov, and Michel Peyrard. Guanine radical chemistry reveals the effect of thermal fluctuations in gene promoter regions. *Nucleic acids research*, 39(12):5276–5283, 2011.

Robert B Darnell. Hits-clip: panoramic views of protein–rna regulation in living cells. *Wiley Interdisciplinary Reviews: RNA*, 1(2):266–286, 2010.

Katarina Dathe, Klaus W Kjaer, Anja Brehm, Peter Meinecke, Peter Nürnberg, Jordao C Neto, Decio Brunoni, Nils Tommerup, Claus E Ott, Eva Klopocki, et al. Duplications involving a conserved regulatory element downstream of bmp2 are associated with brachydactyly type a2. *The American Journal of Human Genetics*, 84(4):483–492, 2009.

Gerrit M Daubner, Antoine Cléry, and Frédéric HT Allain. Rrm–rna recognition: Nmr or crystallography… and new findings. *Current opinion in structural biology*, 23(1):100–108, 2013.

William HE Day and FR McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic acids research*, 20(5):1093–1099, 1992.

Natalia Sánchez De Groot, Alexandros Armaos, Ricardo Graña-Montes, Marion Alriquet, Giulia Calloni, R Martin Vabulas, and Gian Gaetano Tartaglia. Rna structure drives interaction with proteins. *Nature communications*, 10(1):1–13, 2019.

JS Deakyne and AV Mazin. Fanconi anemia: at the crossroads of dna repair. *Biochemistry (Moscow)*, 76(1):36–48, 2011.

Michael DeGiorgio and Noah A Rosenberg. Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure. *Theoretical population biology*, 110:12–24, 2016.

Abdoulaye Banire Diallo, Vladimir Makarenkov, and Mathieu Blanchette. Exact and heuristic algorithms for the indel maximum likelihood problem. *Journal of Computational Biology*, 14(4):446–461, 2007.

Abdoulaye Banire Diallo, Vladimir Makarenkov, and Mathieu Blanchette. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, 26(1):130–131, 2009.

Abdoulaye Banire Diallo, Vladimir Makarenkov, and Mathieu Blanchette. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, 26(1):130–131, 2010.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Feng Ding, Shantanu Sharma, Poornima Chalasani, Vadim V Demidov, Natalia E Broude, and Nikolay V Dokholyan. Ab initio rna folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *Rna*, 14(6):1164–1173, 2008.

Ye Ding and Charles E Lawrence. A statistical sampling algorithm for rna secondary structure prediction. *Nucleic acids research*, 31(24):7280–7301, 2003.

Yuval Dor and Howard Cedar. Principles of dna methylation and their implications for biology and medicine. *The Lancet*, 392(10149):777–786, 2018.

Michael Doyle and Michael F Jantsch. New and old roles of the double-stranded rna-binding domain. *Journal of structural biology*, 140(1-3):147–153, 2002.

Gideon Dreyfuss, V Narry Kim, and Naoyuki Kataoka. Messenger-rna-binding proteins and the messages they carry. *Nature reviews Molecular cell biology*, 3(3):195–205, 2002.

Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

Evan E Eichler. Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine*, 381(1):64–74, 2019.

Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven JM Jones. Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008.

Joseph Felsenstein and Gary A Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, 13(1):93–104, 1996.

Ricardo André Campos Ferraz, Ana Lúcia Gonçalves Lopes, Jessy Ariana Faria da Silva, Diana Filipa Viana Moreira, Maria João Nogueira Ferreira, and Sílvia Vieira de Almeida Coimbra. Dna–protein interaction studies: a historical and comparative analysis. *Plant Methods*, 17(1):1–21, 2021.

C Flores and RB Altman. Coarse-grained modeling of large rna molecules with knowledge-based potentials and structural filters. *RNA*, 15(9):1769–1778, 2010.

Samuel C Flores, Yaqi Wan, Rick Russell, and Russ B Altman. Predicting rna structure by multiple template homology modeling. In *Biocomputing 2010*, pages 216–227. World Scientific, 2010.

James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010.

Tsukasa Fukunaga, Haruka Ozaki, Goro Terai, Kiyoshi Asai, Wataru Iwasaki, and Hisanori Kiryu. Capr: revealing structural specificities of rna-binding protein target recognition using clip-seq data. *Genome biology*, 15(1):R16, 2014.

Mathieu Gabut, Sidharth Chaudhry, and Benjamin J Blencowe. Snapshot: The splicing regulatory machinery. *Cell*, 133(1):192–192, 2008.

Zhen Gao and Jianhua Ruan. A structure-based multiple-instance learning approach to predicting in vitro transcription factor-dna interaction. *BMC genomics*, 16(4):S3, 2015.

Fátima Gebauer, Thomas Schwarzl, Juan Valcárcel, and Matthias W Hentze. Rna-binding proteins in human genetic disease. *Nature Reviews Genetics*, pages 1–14, 2020.

Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, 10(7):e1003711, 2014.

Christopher K Glass and Michael G Rosenfeld. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes & development*, 14(2):121–141, 2000.

Sebastian Glatt, Claudio Alfieri, and Christoph W Müller. Recognizing and remodeling the nucleosome. *Current opinion in structural biology*, 21(3):335–341, 2011.

John F Griffiths, Anthony JF Griffiths, Susan R Wessler, Richard C Lewontin, William M Gelbart, David T Suzuki, Jeffrey H Miller, et al. *An introduction to genetic analysis*. Macmillan, 2005.

DL Gumucio, H Heilstedt-Williamson, TA Gray, SA Tarle, DA Shelton, DA Tagle, JL Slightom, M Goodman, and FS Collins. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Molecular and cellular biology*, 12(11):4919–4929, 1992.

Yosephine Gumulya and Elizabeth MJ Gillam. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro'approach to protein engineering. *Biochemical Journal*, 474(1):1–19, 2017.

Aditi Gupta and Michael Gribskov. The role of rna sequence and structure in rna–protein interactions. *Journal of molecular biology*, 409(4):574–587, 2011.

Swati Gupta, Se Y Kim, Sonja Artis, David L Molfese, Armin Schumacher, J David Sweatt, Richard E Paylor, and Farah D Lubin. Histone methylation regulates memory formation. *Journal of Neuroscience*, 30(10):3589–3599, 2010.

M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, Jr. Ascano, M., A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–41, 2010a. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2010.03.009. URL `https://www.ncbi.nlm.nih.gov/pubmed/20371350`.

Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr, Anna-Carina Jungkamp, Mathias Munschauer, et al. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, 141(1):129–141, 2010b.

Marc S Halfon, Qianqian Zhu, Elizabeth R Brennan, and Yiyun Zhou. Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules. *BMC genomics*, 12(1):578, 2011.

Noriko Handa, Osamu Nureki, Kazuki Kurimoto, Insil Kim, Hiroshi Sakamoto, Yoshiro Shimura, Yutaka Muto, and Shigeyuki Yokoyama. Structural basis for recognition of the tra mrna precursor by the sex-lethal protein. *Nature*, 398(6728):579–585, 1999.

Victor Hanson-Smith and Alexander Johnson. Phylobot: A web portal for automated phylogenetics, ancestral sequence reconstruction, and exploration of mutational trajectories. *PLoS computational biology*, 12(7):e1004976, 2016.

Victor Hanson-Smith, Bryan Kolaczkowski, and Joseph W Thornton. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular biology and evolution*, 27(9):1988–1999, 2010.

Ross C Hardison. Comparative genomics. *PLoS biology*, 1(2):e58, 2003.

Robert S Harris. *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University, 2007.

Hamid Reza Hassanzadeh and May D Wang. Deeperbind: Enhancing prediction of sequence specificities of dna binding proteins. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 178–183. IEEE, 2016.

John Hawkins, Charles Grant, William Stafford Noble, and Timothy L Bailey. Assessing phylogenetic motif models for predicting transcription factor binding sites. *Bioinformatics*, 25(12):i339–i347, 2009.

Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using rna secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research*, 34(17):e117–e117, 2006.

Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

Suntaek Hong. Rna binding protein as an emerging therapeutic target for cancer prevention and treatment. *Journal of cancer prevention*, 22(4):203, 2017.

Brian P Hudson, Maria A Martinez-Yamout, H Jane Dyson, and Peter E Wright. Recognition of the mrna au-rich element by the zinc finger domain of tis11d. *Nature structural & molecular biology*, 11(3):257–264, 2004.

Jason D Hughes, Preston W Estep, Saeed Tavazoie, and George M Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *Journal of molecular biology*, 296(5):1205–1214, 2000.

Plotly Technologies Inc. Collaborative data science, 2015. URL `https://plot.ly`.

David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

Rohit Joshi, Jonathan M Passner, Remo Rohs, Rinku Jain, Alona Sosinsky, Michael A Crickmore, Vinitha Jacob, Aneel K Aggarwal, Barry Honig, and Richard S Mann. Functional specificity of a hox protein mediated by the recognition of minor groove structure. *Cell*, 131(3):530–543, 2007.

Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. Rnacontext: a new method for learning the sequence and structure binding preferences of rna-binding proteins. *PLoS Comput Biol*, 6(7):e1000832, 2010.

Jack D Keene. Ribonucleoprotein infrastructure regulating the flow of genetic information between the genome and the proteome. *Proceedings of the National Academy of Sciences*, 98(13):7018–7024, 2001.

Jens Keilwagen, Stefan Posch, and Jan Grau. Accurate prediction of cell type-specific transcription factor binding. *Genome biology*, 20(1):9, 2019.

Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241, 2003.

W James Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S Mikkelsen, and Manolis Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research*, 23(5):800–811, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Malka Kitayner, Haim Rozenberg, Remo Rohs, Oded Suad, Dov Rabinovich, Barry Honig, and Zippora Shakked. Diversity in dna recognition by p53 revealed by crystal structures with hoogsteen base pairs. *Nature structural & molecular biology*, 17(4):423, 2010.

Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–U166, 2010. ISSN 1545-9993. doi: 10.1038/nsmb.1838. URL <GotoISI>://WOS:000279631500021.

Julian König, Kathi Zarnack, Gregor Rot, Tomaž Curk, Melis Kayikci, Blaž Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iclip reveals the function of hnrnp particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909, 2010.

Julian König, Kathi Zarnack, Nicholas M Luscombe, and Jernej Ule. Protein–rna interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, 13(2):77–83, 2012.

Peter K Koo and Matt Ploenzke. Deep learning for inferring transcription factor binding sites. *Current opinion in systems biology*, 19:16–23, 2020.

Sabrina Krakau, Hugues Richard, and Annalisa Marsico. Pureclip: capturing target-specific protein–rna interaction footprints from single-nucleotide clip-seq data. *Genome biology*, 18(1):1–17, 2017.

Maxat Kulmanov and Robert Hoehndorf. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422–429, 2020.

Sudhir Kumar and Alan Filipski. Multiple sequence alignment: in pursuit of homologous dna positions. *Genome research*, 17(2):127–135, 2007.

A Kundaje, N Boley, R Kuffner, L Heiser, J Costello, G Stolovitzky, T Norman, B Hoff, and S Friend. Encode-dream in vivo transcription factor binding site prediction challenge. synapse. 10.7303/syn6131484. `https://www.synapse.org/#!Synapse:syn6131484/wiki/402026`, 2021. [Online; accessed 16-July-2021].

Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

David S Lawrie and Dmitri A Petrov. Comparative population genomics: power and principles for the inference of functionality. *Trends in Genetics*, 30(4):133–139, 2014.

David S Lawrie, Dmitri A Petrov, and Philipp W Messer. Faster than neutral evolution of constrained sequences: the complex interplay of mutational biases and weak selection. *Genome biology and evolution*, 3:383–395, 2011.

Allan Lazarovici, Tianyin Zhou, Anthony Shafer, Ana Carolina Dantas Machado, Todd R Riley, Richard Sandstrom, Peter J Sabo, Yan Lu, Remo Rohs, John A Stamatoyannopoulos, et al. Probing dna shape and methylation state on a genomic scale with dnase i. *Proceedings of the National Academy of Sciences*, 110(16):6376–6381, 2013.

Mickael Leclercq, Abdoulaye Baniré Diallo, and Mathieu Blanchette. Prediction of human mirna target genes using computationally reconstructed ancestral mammalian sequences. *Nucleic acids research*, 45(2):556–566, 2017.

Chih Lee and Chun-Hsi Huang. Lasagna-search: an integrated web tool for transcription factor binding site search and visualization. *Biotechniques*, 54(3):141–153, 2013.

KangSeok Lee, Shikha Varma, John SantaLucia Jr, and Philip R Cunningham. In vivo determination of rna structure-function relationships: analysis of the 790 loop in ribosomal rna. *Journal of molecular biology*, 269(5):732–743, 1997.

Laura A Lettice, Sarah Daniels, Elizabeth Sweeney, Shanmugasundaram Venkataraman, Paul S Devenney, Philippe Gautier, Harris Morrison, Judy Fantes, Robert E Hill, and David R FitzPatrick. Enhancer-adoption as a mechanism of human developmental disease. *Human mutation*, 32(12):1492–1499, 2011.

Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 2010.

Hongyang Li, Daniel Quang, and Yuanfang Guan. Anchor: trans-cell type prediction of transcription factor binding sites. *Genome research*, 29(2):281–292, 2019.

Qian-Ru Li, Anne-Ruxandra Carvunis, Haiyuan Yu, Jing-Dong J Han, Quan Zhong, Nicolas Simonis, Stanley Tam, Tong Hao, Niels J Klitgord, Denis Dupuy, et al. Revisiting the saccharomyces cerevisiae predicted orfeome. *Genome research*, 2008a.

Sanshu Li and Ronald R Breaker. Identification of 15 candidate structured noncoding rna motifs in fungi by comparative genomics. *BMC genomics*, 18(1):785, 2017.

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

Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, et al. Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS Biol*, 6(2):e27, 2008b.

Yifeng Li, Chih-yu Chen, Alice M Kaye, and Wyeth W Wasserman. The identification of cis-regulatory elements: A review from a machine learning perspective. *Biosystems*, 138:6–17, 2015.

D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. N. Wang, J. C. Darnell, and R. B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–U22, 2008a. ISSN 0028-0836. doi: 10.1038/nature07488. URL `<GotoISI>://WOS:000261170500030`.

Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, et al. Hits-clip yields genome-wide insights into brain alternative rna processing. *Nature*, 456(7221): 464–469, 2008b.

Michael Lieberman and Allan D Marks. *Marks' basic medical biochemistry: a clinical approach*. Lippincott Williams & Wilkins, 2009.

Edison T Liu, Sebastian Pott, and Mikael Huss. Q&a: Chip-seq technologies and the study of gene regulation. *BMC biology*, 8(1):1–6, 2010.

Hong-Xiang Liu, Michael Zhang, and Adrian R Krainer. Identification of functional exonic splicing enhancer motifs recognized by individual sr proteins. *Genes & development*, 12(13), 1998.

Carmen M Livi and Enrico Blanzieri. Protein-specific prediction of mrna binding using rna sequences, binding motifs and predicted secondary structures. *BMC bioinformatics*, 15(1):1–11, 2014.

Gabriela G Loots and Ivan Ovcharenko. rvista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic acids research*, 32(suppl_2):W217–W221, 2004.

Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):1–14, 2011.

Michael T Lovci, Dana Ghanem, Henry Marr, Justin Arnold, Sherry Gee, Marilyn Parra, Tiffany Y Liang, Thomas J Stark, Lauren T Gehman, Shawn Hoon, et al. Rbfox proteins regulate alternative mrna splicing through evolutionarily conserved rna bridges. *Nature structural & molecular biology*, 20(12):1434–1442, 2013.

Duo Lu, M Alexandra Searles, and Aaron Klug. Crystal structure of a zinc-finger–rna complex reveals two modes of molecular recognition. *Nature*, 426(6962):96–100, 2003.

Rupert Lück, Gerhard Steger, and Detlev Riesner. Thermodynamic prediction of conserved secondary structure: application to the rre element of hiv, the trna-like element of cmv and the mrna of prion protein. *Journal of molecular biology*, 258(5):813–826, 1996.

K. E. Lukong, K. W. Chang, E. W. Khandjian, and S. Richard. RNA-binding proteins in human genetic disease. *Trends Genet*, 24(8):416–25, 2008a. ISSN 0168-9525 (Print) 0168-9525 (Linking). doi: 10.1016/j.tig.2008.05.004. URL `https://www.ncbi.nlm.nih.gov/pubmed/18597886`.

Kiven E Lukong, Kai-wei Chang, Edouard W Khandjian, and Stéphane Richard. Rna-binding proteins in human genetic disease. *Trends in Genetics*, 24(8):416–425, 2008b.

Kiven Erique Lukong, Daniel Larocque, Angela L Tyner, and Stéphane Richard. Tyrosine phosphorylation of sam68 by breast tumor kinase regulates intranuclear localization and cell cycle progression. *Journal of Biological Chemistry*, 280(46):38639–38647, 2005.

Jiesi Luo, Liang Liu, Suresh Venkateswaran, Qianqian Song, and Xiaobo Zhou. Rpi-bind: a structure-based method for accurate identification of rna-protein binding sites. *Scientific reports*, 7(1):1–13, 2017.

Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

Tsz-Kwong Man and Gary D Stormo. Non-independence of mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (qumfra) assay. *Nucleic acids research*, 29(12):2471–2478, 2001.

Voichita D Marinescu, Isaac S Kohane, and Alberto Riva. Mapper: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC bioinformatics*, 6(1):1–20, 2005.

Joan Massagué, Joan Seoane, and David Wotton. Smad transcription factors. *Genes & development*, 19(23):2783–2810, 2005.

Anthony Mathelier and Wyeth W Wasserman. The next generation of transcription factor binding site prediction. *PLoS computational biology*, 9(9):e1003214, 2013.

DAVID H Mathews, ALOKE R Banerjee, DONGMEI D Luan, THOMAS H Eickbush, and DOUGLAS H Turner. Secondary structure model of the rna recognized by the reverse transcriptase from the r2 retrotransposable element. *Rna*, 3(1):1–16, 1997.

David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *Journal of molecular biology*, 288(5):911–940, 1999.

Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. Graphprot: modeling binding preferences of rna-binding proteins. *Genome biology*, 15(1):1–18, 2014.

Michael McClelland, Liliana Florea, Ken Sanderson, Sandra W Clifton, Julian Parkhill, Carol Churcher, Gordon Dougan, Richard K Wilson, and Webb Miller. Comparison of the escherichia coli k-12 genome with sampled genomes of a klebsiella pneumoniae and three salmonella enterica serovars, typhimurium, typhi and paratyphi. *Nucleic acids research*, 28(24):4974–4986, 2000.

Sebastiaan H Meijsing, Miles A Pufall, Alex Y So, Darren L Bates, Lin Chen, and Keith R Yamamoto. Dna binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324(5925):407–410, 2009.

Flavio Mignone and Graziano Pesole. (January 2018) mRNA Untranslated Regions (UTRs). *eLS*, 2018.

Joanna A Miller and Jonathan Widom. Collaborative competition mechanism for gene activation in vivo. *Molecular and cellular biology*, 23(5):1623–1632, 2003.

Leonid A Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences*, 107(52):22534–22539, 2010.

Rebecca K Montange and Robert T Batey. Riboswitches: emerging themes in rna structure and function. *Annu. Rev. Biophys.*, 37:117–133, 2008.

Michael J Moore, Chaolin Zhang, Emily Conn Gantman, Aldo Mele, Jennifer C Darnell, and Robert B Darnell. Mapping argonaute and conventional rna-binding protein interactions with rna at single-nucleotide resolution using hits-clip and cims analysis. *Nature protocols*, 9(2):263–293, 2014.

Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. Insights into rna structure and function from genome-wide studies. *Nature Reviews Genetics*, 15(7):469–479, 2014.

Alan M Moses, Derek Y Chiang, Daniel A Pollard, Venky N Iyer, and Michael B Eisen. Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome biology*, 5(12):R98, 2004.

Alan M Moses, Daniel A Pollard, David A Nix, Venky N Iyer, Xiao-Yong Li, Mark D Biggin, and Michael B Eisen. Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS computational biology*, 2(10):e130, 2006.

Sara Movahedi, Yves Van de Peer, and Klaas Vandepoele. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in arabidopsis and rice. *Plant physiology*, pages pp–111, 2011.

Usha K Muppirala, Vasant G Honavar, and Drena Dobbs. Predicting rna-protein interactions using only sequence information. *BMC bioinformatics*, 12(1):1–11, 2011.

Shane Neph, M Scott Kuehn, Alex P Reynolds, Eric Haugen, Robert E Thurman, Audra K Johnson, Eric Rynes, Matthew T Maurano, Jeff Vierstra, Sean Thomas, et al. Bedops: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, 2012.

Martin Nettling, Hendrik Treutler, Jesus Cerquides, and Ivo Grosse. Combining phylogenetic footprinting with motif models incorporating intra-motif dependencies. *BMC bioinformatics*, 18(1):1–10, 2017.

Xiaochun Ni, Yong E Zhang, Nicolas Negre, Sidi Chen, Manyuan Long, and Kevin P White. Adaptive evolution and the birth of ctcf binding sites in the drosophila genome. *PLoS biology*, 10(11):e1001420, 2012.

Yaron Orenstein, Yuhao Wang, and Bonnie Berger. Rck: accurate and efficient inference of sequence-and structure-based protein–rna binding models from rnacompete data. *Bioinformatics*, 32(12):i351–i359, 2016.

José Carlos Ribeiro Pacheco. *PGP: prokaryote gene prediction software*. PhD thesis, 2013.

Xiaoyong Pan and Hong-Bin Shen. Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC bioinformatics*, 18(1):136, 2017.

Xiaoyong Pan and Hong-Bin Shen. Learning distributed representations of rna sequences and its application for predicting rna-protein binding sites with a convolutional neural network. *Neurocomputing*, 305:51–58, 2018.

Xiaoyong Pan, Peter Rijnbeek, Junchi Yan, and Hong-Bin Shen. Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC genomics*, 19(1):511, 2018.

Daniel Panne. The enhanceosome. *Current opinion in structural biology*, 18(2):236–242, 2008.

Dmitri Papatsenko, Andrey Kislyuk, Michael Levine, and Inna Dubchak. Conservation patterns in different functional sequence categories of divergent drosophila species. *Genomics*, 88(4):431–442, 2006.

Sungjoon Park, Yookyung Koh, Hwisang Jeon, Hyunjae Kim, Yoonsun Yeo, and Jaewoo Kang. Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. *Scientific reports*, 10(1):1–10, 2020.

Maria Paola Paronetto, Manuela Cappellari, Roberta Busà, Simona Pedrotti, Roberta Vitali, Clay Comstock, Terry Hyslop, Karen E Knudsen, and Claudio Sette. Alternative splicing of the cyclin d1 proto-oncogene is regulated by the rna-binding protein sam68. *Cancer research*, 70(1):229–239, 2010.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

Benedict Paten, Javier Herrero, Stephen Fitzgerald, Kathryn Beal, Paul Flicek, Ian Holmes, and Ewan Birney. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research*, 2008.

Joshua L Payne, Fahad Khalid, and Andreas Wagner. Rna-mediated gene regulation is less evolvable than transcriptional regulation. *Proceedings of the National Academy of Sciences*, 115(15):E3481–E3490, 2018.

Inbal Paz, Idit Kosti, Manuel Ares Jr, Melissa Cline, and Yael Mandel-Gutfreund. Rbpmap: a web server for mapping binding sites of rna-binding proteins. *Nucleic acids research*, 42(W1):W361–W367, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011a.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011b.

Len A Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A Nobrega, and Gill Bejerano. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4):288–295, 2013.

Elizabeth Pennisi. Encode project writes eulogy for junk dna, 2012.

Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for chip-seq and rna-seq studies. *Nature methods*, 6(11):S22–S32, 2009.

Ismael Pérez, James G McAfee, and James G Patton. Multiple rrms contribute to rna binding specificity and affinity for polypyrimidine tract binding protein. *Biochemistry*, 36(39):11881–11890, 1997.

Marco Pietrosanto, Eugenio Mattei, Manuela Helmer-Citterich, and Fabrizio Ferre. A novel method for the identification of conserved structural patterns in rna: From small scale to high-throughput applications. *Nucleic acids research*, 44(18):8600–8609, 2016.

Wei Ping, Ye Xu, Kexin Ren, Chi-Hung Chi, and Furao Shen. Non-iid multi-instance dimensionality reduction by learning a maximum bag margin subspace. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome research*, 21(3):447–455, 2011.

Fabrizio Pucci and Alexander Schug. Shedding light on the dark matter of the biomolecular structural universe: Progress in rna 3d structure prediction. *Methods*, 162:68–73, 2019.

Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11): e107–e107, 2016.

Daniel Quang and Xiaohui Xie. Factornet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 166:40–47, 2019.

Alessandro Quattrone and Erik Dassi. The architecture of the human rna-binding protein regulatory network. *IScience*, 21:706–719, 2019.

Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Pena Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. Rapid and systematic analysis of the rna recognition specificities of rna-binding proteins. *Nature biotechnology*, 27(7):667–670, 2009.

Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of rna-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.

Jean-Jack M Riethoven. Regulatory regions in dna: promoters, enhancers, silencers, and insulators. *Computational biology of transcription factor binding*, pages 33–42, 2010.

Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, et al. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–657, 2007.

Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

Jermaine Ross, Alexander Kuzin, Thomas Brody, and Ward F Odenwald. Mutational analysis of a drosophila neuroblast enhancer governing nubbin expression during cns development. *genesis*, 56(8):e23237, 2018.

Frederick P Roth, Jason D Hughes, Preston W Estep, and George M Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature biotechnology*, 16(10):939–945, 1998.

Magdalena Rother, Kristian Rother, Tomasz Puton, and Janusz M Bujnicki. Moderna: a tool for comparative modeling of rna 3d structure. *Nucleic acids research*, 39(10):4007–4022, 2011.

Javad Sadri, Abdoulaye Banire Diallo, and Mathieu Blanchette. Predicting site-specific human selective pressure using evolutionary signatures. *Bioinformatics*, 27(13):i266–i274, 2011.

Gulzhakhan Sadvakassova, Monica C Dobocan, Marcos R Difalco, and Luis F Congote. Regulator of differentiation 1 (rod1) binds to the amphipathic c-terminal peptide of thrombospondin-4 and is involved in its mitogenic activity. *Journal of cellular physiology*, 220(3):672–679, 2009.

Jeremy R Sanford, Pedro Coutinho, Jamie A Hackett, Xin Wang, William Ranahan, and Javier F Caceres. Identification of nuclear and cytoplasmic mrna targets for the shuttling protein sf2/asf. *PloS one*, 3(10):e3369, 2008.

Jeremy R Sanford, Xin Wang, Matthew Mort, Natalia VanDuyn, David N Cooper, Sean D Mooney, Howard J Edenberg, and Yunlong Liu. Splicing factor sfrs1 recognizes a functionally diverse landscape of rna transcripts. *Genome research*, 19(3):381–394, 2009.

Laura R Saunders and Glen N Barber. The dsrna binding protein family: critical roles, diverse cellular functions. *The FASEB Journal*, 17(9):961–983, 2003.

Pierluigi Scerbo, Gabriel V Markov, Céline Vivien, Laurent Kodjabachian, Barbara Demeneix, Laurent Coen, and Fabrice Girardot. On the origin and evolutionary history of nanog. *PloS one*, 9(1):e85104, 2014.

Thomas D Schneider and R Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990.

Sebastian Schönherr, Lukas Forer, Hansi Weißensteiner, Florian Kronenberg, Günther Specht, and Anita Kloss-Brandstätter. Cloudgene: a graphical execution platform for mapreduce programs on private and public clouds. *BMC bioinformatics*, 13(1):1–9, 2012.

Scott Schwartz, Laura Elnitski, Mei Li, Matt Weirauch, Cathy Riemer, Arian Smit, NISC Comparative Sequencing Program, Eric D Green, Ross C Hardison, and Webb Miller. Multipipmaker and supporting tools: Alignments and analysis of multiple genomic dna sequences. *Nucleic acids research*, 31(13):3518–3524, 2003a.

Scott Schwartz, W James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C Hardison, David Haussler, and Webb Miller. Human–mouse alignments with blastz. *Genome research*, 13(1):103–107, 2003b.

Svetlana A Shabalina, Aleksey Y Ogurtsov, Igor B Rogozin, Eugene V Koonin, and David J Lipman. Comparative analysis of orthologous eukaryotic mrnas: potential hidden functional signals. *Nucleic Acids Research*, 32(5):1774–1782, 2004.

Mahfuza Sharmin, Héctor Corrada Bravo, and Sridhar Hannenhalli. Heterogeneity of transcription factor binding specificity models within and across cell lines. *Genome research*, 26(8):1110–1123, 2016.

Ekaterina Shelest and Edgar Wingender. Construction of predictive promoter models on the example of antibacterial response of human epithelial cells. *Theoretical biology and medical modelling*, 2(1):2, 2005.

Zhen Shen, Wenzheng Bao, and De-Shuang Huang. Recurrent neural network for predicting transcription factor binding sites. *Scientific reports*, 8(1):1–10, 2018.

Zu-Hang Sheng and Qian Cai. Mitochondrial transport in neurons: impact on synaptic homeostasis and neurodegeneration. *Nature Reviews Neuroscience*, 13(2):77–93, 2012.

Takeshi Shiraishi, Teresa Druck, Koshi Mimori, Jacob Flomenberg, Lori Berk, Hansjuerg Alder, Webb Miller, Kay Huebner, and Carlo M Croce. Sequence conservation at human

and mouse orthologous common fragile regions, fra3b/fhit and fra14a2/fhit. *Proceedings of the National Academy of Sciences*, 98(10):5722–5727, 2001.

Adam Siepel and David Haussler. Phylogenetic hidden markov models. In *Statistical methods in molecular evolution*, pages 325–351. Springer, 2005.

Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

Adam Siepel, Katherine S Pollard, and David Haussler. New methods for detecting lineage-specific selection. In *Annual International Conference on Research in Computational Molecular Biology*, pages 190–205. Springer, 2006.

Trevor Siggers, Michael H Duyzend, Jessica Reddy, Sidra Khan, and Martha L Bulyk. Non-dna-binding cofactors enhance dna-binding specificity of a transcriptional regulatory complex. *Molecular systems biology*, 7(1):555, 2011.

Saurabh Sinha and Eric D Siggia. Sequence turnover and tandem repeats in cis-regulatory modules in drosophila. *Molecular biology and evolution*, 22(4):874–885, 2005.

Matthew Slattery, Todd Riley, Peng Liu, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, Remo Rohs, Barry Honig, Harmen J Bussemaker, et al. Cofactor binding evokes latent differences in dna binding specificity between hox proteins. *Cell*, 147(6): 1270–1282, 2011.

Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, 39(9):381–399, 2014.

Mark J Solomon, Pamela L Larsen, and Alexander Varshavsky. Mapping proteindna interactions in vivo with formaldehyde: Evidence that histone h4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, 1988.

Malte Spielmann and Stefan Mundlos. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays*, 35(6):533–543, 2013.

Tanja Stadler, James H Degnan, and Noah A Rosenberg. Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? *Systematic biology*, 65(4):628–639, 2016.

Moses Stamboulian, Rafael F Guerrero, Matthew W Hahn, and Predrag Radivojac. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics*, 36(Supplement_1):i219–i226, 2020.

Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. Rnashapes: an integrated rna analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

R. Stefl, L. Skrisovska, and F. H. Allain. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep*, 6(1):33–8, 2005. ISSN 1469-221X (Print) 1469-221X (Linking). doi: 10.1038/sj.embor.7400325. URL `https://www.ncbi.nlm.nih.gov/pubmed/15643449`.

Lincoln D Stein, Zhirong Bao, Darin Blasiar, Thomas Blumenthal, Michael R Brent, Nansheng Chen, Asif Chinwalla, Laura Clarke, Chris Clee, Avril Coghlan, et al. The genome sequence of caenorhabditis briggsae: a platform for comparative genomics. *PLoS biology*, 1(2):e45, 2003.

Gary D Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1): 16–23, 2000.

Gary D Stormo, Thomas D Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research*, 10(9):2997–3011, 1982.

M. Strazar, M. Zitnik, B. Zupan, J. Ule, and T. Curk. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, 32(10): 1527–1535, 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw003. URL <GotoISI>://WOS:000376656900012.

Martin Stražar, Marinka Žitnik, Blaž Zupan, Jernej Ule, and Tomaž Curk. Orthogonal matrix factorization enables integrative analysis of multiple rna binding proteins. *Bioinformatics*, 32(10):1527–1535, 2016.

Hong Su, Zhenling Peng, and Jianyi Yang. Recognition of small molecule–rna binding sites using rna sequence and structure. *Bioinformatics*, 37(1):36–42, 2021.

Yoichiro Sugimoto, Julian König, Shobbir Hussain, Blaž Zupan, Tomaž Curk, Michaela Frye, and Jernej Ule. Analysis of clip and iclip methods for nucleotide-resolution studies of protein-rna interactions. *Genome biology*, 13(8):1–13, 2012.

R Tacke and James L Manley. The human splicing factors asf/sf2 and sc35 possess distinct, functionally significant rna binding specificities. *The EMBO journal*, 14(14):3540–3551, 1995.

Danilo A Tagle, Ben F Koop, Morris Goodman, Jerry L Slightom, David L Hess, and Richard T Jones. Embryonic ε and γ globin genes of a prosimian primate (galago crassicaudatus): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of molecular biology*, 203(2):439–455, 1988.

Vladimir B Teif and Karsten Rippe. Statistical–mechanical lattice models for protein–dna binding in chromatin. *Journal of Physics: Condensed Matter*, 22(41):414105, 2010.

JW Thomas, JW Touchman, RW Blakesley, GG Bouffard, SM Beckstrom-Sternberg, EH Margulies, M Blanchette, AC Siepel, PJ Thomas, JC McDowell, et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424 (6950):788, 2003.

James R Tollervey, Tomaž Curk, Boris Rogelj, Michael Briese, Matteo Cereda, Melis Kayikci, Julian König, Tibor Hortobágyi, Agnes L Nishimura, Vera Župunski, et al. Characterizing the rna targets and position-dependent splicing regulation by tdp-43. *Nature neuroscience*, 14(4):452–458, 2011.

Juan A Ugalde, Belinda SW Chang, and Mikhail V Matz. Evolution of coral pigments recreated. *Science*, 305(5689):1433–1433, 2004.

Michael Uhl, Van Dinh Tran, and Rolf Backofen. Improving clip-seq data analysis by incorporating transcript information. *BMC genomics*, 21(1):1–8, 2020.

Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaž Ule, and Robert B Darnell. Clip identifies nova-regulated rna networks in the brain. *Science*, 302(5648):1212–1215, 2003.

Jernej Ule, Kirk Jensen, Aldo Mele, and Robert B Darnell. Clip: a method for identifying protein–rna interaction sites in living cells. *Methods*, 37(4):376–386, 2005.

Philip J Uren, Emad Bahrami-Samani, Suzanne C Burns, Mei Qiao, Fedor V Karginov, Emily Hodges, Gregory J Hannon, Jeremy R Sanford, Luiz OF Penalva, and Andrew D Smith. Site identification in high-throughput rna–protein interaction data. *Bioinformatics*, 28(23):3013–3020, 2012.

Eric L Van Nostrand, Peter Freese, Gabriel A Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M Blue, Jia-Yu Chen, Neal AL Cody, Daniel Dominguez, et al. A large-scale binding and functional map of human rna-binding proteins. *Nature*, 583(7818):711–719, 2020.

Laurent Vanhille, Aurélien Griffon, Muhammad Ahmad Maqbool, Joaquin Zacarias-Cabeza, Lan TM Dao, Nicolas Fernandez, Benoit Ballester, Jean Christophe Andrau, and Salvatore Spicuglia. High-throughput and quantitative assessment of enhancer activity in mammals by capstarr-seq. *Nature communications*, 6(1):1–10, 2015.

Karin Voordeckers, Chris A Brown, Kevin Vanneste, Elisa van der Zande, Arnout Voet, Steven Maere, and Kevin J Verstrepen. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol*, 10(12):e1001446, 2012.

Guey-Shin Wang and Thomas A Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761, 2007.

Ming-Shan Wang, Rong-wei Zhang, Ling-Yan Su, Yan Li, Min-Sheng Peng, He-Qun Liu, Lin Zeng, David M Irwin, Jiu-Lin Du, Yong-Gang Yao, et al. Positive selection rather than relaxation of functional constraint drives the evolution of vision during chicken domestication. *Cell research*, 26(5):556, 2016.

Xiao-Dan Wang, Jin-Xing Liu, Yong Xu, and Jian Zhang. A survey of multiple sequence alignment techniques. In *International Conference on Intelligent Computing*, pages 529–538. Springer, 2015.

Zhen Wang, Melis Kayikci, Michael Briese, Kathi Zarnack, Nicholas M Luscombe, Gregor Rot, Blaž Zupan, Tomaž Curk, and Jernej Ule. iclip predicts the dual splicing effects of tia-rna interactions. *PLoS Biol*, 8(10):e1000530, 2010.

Todd Wasson and Alexander J Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome research*, 19(11):2101–2112, 2009.

Alex Wells, David Heckerman, Ali Torkamani, Li Yin, Jonathan Sebat, Bing Ren, Amalio Telenti, and Julia di Iulio. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature communications*, 10(1):1–9, 2019.

Oscar Westesson, Gerton Lunter, Benedict Paten, and Ian Holmes. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS One*, 7(4):e34572, 2012.

Michael A White, Connie A Myers, Joseph C Corbo, and Barak A Cohen. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of chip-seq peaks. *Proceedings of the National Academy of Sciences*, 110(29):11952–11957, 2013.

Simon T Whiteside and Stephen Goodbourn. Signal transduction and nuclear targeting: regulation of transcription factor activity by subcellular localisation. *Journal of cell science*, 104(4):949–955, 1993.

Shu Xiao, Xiaoyi Cao, and Sheng Zhong. Comparative epigenomics: defining and utilizing epigenomic variations across species, time-course, and individuals. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 6(5):345–352, 2014.

Hanako Yamamoto, Kappei Tsukahara, Yoshihide Kanaoka, Shigeki Jinno, and Hiroto Okayama. Isolation of a mammalian homologue of a fission yeast differentiation regulator. *Molecular and Cellular Biology*, 19(5):3829–3841, 1999.

Jian Yan, Martin Enge, Thomas Whitington, Kashyap Dave, Jianping Liu, Inderpreet Sur, Bernhard Schmierer, Arttu Jolma, Teemu Kivioja, Minna Taipale, et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, 154(4):801–813, 2013.

Z. C. Yan, E. Lecuyer, and M. Blanchette. Prediction of mrna subcellular localization using deep recurrent neural networks. *Bioinformatics*, 35(14):I333–I342, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz337. URL <GotoISI>://WOS:000477703600038.

Ziheng Yang. A space-time process model for the evolution of dna sequences. *Genetics*, 139(2):993–1005, 1995.

Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

Jia Yu, Jochen Blom, SP Glaeser, S Jaenicke, T Juhre, O Rupp, O Schwengers, S Spänig, and Alexander Goesmann. A review of bioinformatics platforms for comparative genomics. recent developments of the edgar 2.0 platform and its utility for taxonomic and phylogenetic studies. *Journal of biotechnology*, 261:2–9, 2017.

Federico Zambelli, Graziano Pesole, and Giulio Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, 14(2):225–237, 2012.

Haoyang Zeng, Matthew D Edwards, Ge Liu, and David K Gifford. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127, 2016.

Chaolin Zhang and Robert B Darnell. Mapping in vivo protein-rna interactions at single-nucleotide resolution from hits-clip data. *Nature biotechnology*, 29(7):607–614, 2011.

Jianzhi Zhang and Helene F Rosenberg. Complementary advantageous substitutions in the evolution of an antiviral rnase of higher primates. *Proceedings of the National Academy of Sciences*, 99(8):5486–5491, 2002.

Sai Zhang, Jingtian Zhou, Hailin Hu, Haipeng Gong, Ligong Chen, Chao Cheng, and Jianyang Zeng. A deep learning framework for modeling structural features of rna-binding protein targets. *Nucleic acids research*, 44(4):e32–e32, 2016.

Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1–9, 2008.

Chenhan Zhao, Xiaojun Xu, and Shi-Jie Chen. Predicting rna structure with vfold. In *Functional Genomics*, pages 3–15. Springer, 2017.

Yunjie Zhao, Yangyu Huang, Zhou Gong, Yanjie Wang, Jianfen Man, and Yi Xiao. Automated and fast building of three-dimensional rna structures. *Scientific reports*, 2(1):1–6, 2012.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931, 2015.

Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23, 2019.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM, 2009.