Using Deep Learning with Different Architectures to Recognize RNA:DNA Triplex Structures from Histone Modification Features

Joseph L. Tsenum[1], Maria Poptsova[2]

[1]Phystech School of Biological and Medical Physics, Moscow Institute of Physics and Technology (MIPT), Moscow Region, Russia

[2]International Laboratory of Bioinformatics, Higher School of Economics (HSE University), Moscow, Russia

*Corresponding Author: Joseph L. Tsenum; Phystech School of Biological and Medical Physics, Moscow Institute of Physics and Technology (MIPT); tsenumjosephluper@phystech.edu

Abstract

Long non-coding RNAs (lncRNAs) can perform their regulatory roles by forming triple helices through RNA-DNA interaction. Although this has been verified by few in vivo and in vitro methods, in silico approaches that seek to predict the potentials of lncRNAs and DNA sites becoming a triplex forming structure is required. Triplexator have also predicted vast amounts of lncRNAs and DNA sites that has the potentials of becoming a triplex structure. There is also an emerging experimental-evidence that the presence of epigenetic marks at DNA sites and lncRNAs can facilitate the formation of RNA:DNA triplex structures. There is therefore, a huge demand for computational approaches such as deep learning that can make novel predictions about RNA:DNA triplex structure formation. In this study, we developed four (4) deep neural network models that can predict the potentials of lncRNAs and DNA sites to form triple helices genome-wide, by taking histone modification marks as features. Our data was first passed through the Triplexator to screen out lncRNAs and DNA sites with low potentials of forming triple helices. We used different deep learning architectures to build our models, including two-layer convolutional neural networks (CNN) and multilayer perceptron (MLP). Our DNA2_CNN model performed best at a mean AUC of 0.78 at 32 Kernel size and learning rate of 1e-3. Our deep neural network models revealed several novel lncRNAs and DNA sites, including HOTAIR, MEG3, PARTICLE, DACOR1, MIR100HG, FENDRR, ANRIL, TUG1, MALAT1, LINC00599, TINCR, NEAT1, roX2, DHFR, OTX2-AS1, Xist, SNHG16, ATXN8OS, BCYRN1, TERC, Khps1, that have the potential of forming triplex structures, thereby confirming previous experimental results and that of the Triplexator. The performance of our models also supports previous findings that histone modification marks can help in identifying lncRNAs and DNA regions that have the potentials of forming RNA:DNA triplex structures. In conclusion, we showed that different deep learning architectures can recognize lncRNAs and DNA that have the potentials of forming RNA:DNA triplex structures.

Keywords: *Long* non-coding RNAs (lncRNAs), DNA sites, triplex structures, histone modification marks, deep learning architectures

Introduction

Advances in genome sequencing technologies in the past decade have shown the complexity of genome organization in both eukaryotic and prokaryotic organisms (Kapranov *et al.*, 2007; ENCODE Project Consortium, 2004, 2012). These genome sequencing technologies has revealed large amounts of non- coding RNAs (ncRNAs) that were previously coined as 'junk DNA' and a small fraction of protein-coding mRNA in the mammalian DNA.

Efforts have been made in the past years in sequencing ncRNAs and quantifying their various regulatory roles (Quinn and Chang, 2016; Bonasio and Shiekhat-tar, 2014; Batista and Chang, 2013; Lee, 2012). Non coding RNAs (ncRNAs) consists of a broad class of regulatory RNAs; these include small ncRNAs such as small interferingRNAs (siRNAs), PIWI-interacting RNAs, microRNAs (miRNAs), and long noncoding RNAs (lncRNAs), which are made up of more than 200 nucleotide in length (Cech and Steitz, 2014; Kowalczyk *et al.*, 2012). Majority of the mammalian non-coding RNAs are highly conserved, suggesting that they perform significant functions in biological processes (Struhl, 2007).

Long non- coding RNAs (lncRNAs) interacts with epigenetic factors to structure the results of important biological roles such as gene transcription regulation, modulation of nuclear organization, pre-mRNA splicing and X inactivation in females. This has improved our knowledge of the interaction between lncRNAs and epigenetic factors of human diseases as biomarkers of genetic diseases and their therapeutic targets (David J. H *et al.*, 2018).

Long non-coding RNAs (lncRNAs) are capable of forming RNA-DNA hybrid duplexes and RNA-DNA triplexes structures at specific DNA sequences (Wang and Chang, 2011; Li *et al.*, 2016). These triplex-forming lncRNAs are important in biological processes. Both *in-vivo* and *in-vitro* methods have verified the potential of some lncRNAs in forming triplex structures with DNA sequences. Various mathematical, statistical and computational methods have also been applied in predicting lncRNAs that have the potential of forming triplex structures with specific DNA regions. Computationally, classical machine learning and deep neural networks have shown promise in predicting these triplex-forming lncRNAs and DNA sites.

Aims of the Research

Research on the functions of long non-coding RNA as decoys, scaffolds, signals, and guides, to function in gene expression regulation and chromatin states modulation is emerging rapidly (Mercer and Mattick, 2013). Recent sequencing technologies such as CHART-Seq, CHOP-Seq, ChIRP-Seq and RAP-Seq has led to the discovery of large amounts of long non-coding RNAs (lncRNAs). Recent research has shown that long non-coding RNAs (lncRNAs) can form triplex structures with the DNA to perform their regulatory roles (Zhang, *et al*., 2020). They form these structures by interacting with the DNA, other RNAs and proteins in order to guide regulatory proteins, localizing to specific loci thereby shaping the three-dimensional (3D) nuclear organization (Mercer and Mattick, 2013). Previous research that described the canonical B-form DNA double helix also proposed that other possible DNA, DNA-RNA, and RNA structures that plays vital roles as functional genomic elements exists (Albino *et. al*., 2015). Mathematical and statistical base pairing rules such as Triplexator, Triplex Domain Finder (TDF), LongTarget and Triplex-Inspector have been applied over the years in predicting DNA:RNA triplex structure formation (Zhang *et al*., 2020).These methods have been known to recognize vast amounts of potential triplex forming lncRNAs and DNA sites based on canonical rules (Zhang *et al*., 2020). Little amounts of experimentally verified triplex- forming lncRNAs suggests that some of the predicted triplex-forming lncRNAs by canonical rules may not form triplex structures in practice (Zhang *et al*., 2020), however more research is needed to predict the exact triplex-forming lncRNAs and DNA sites. Theoretical triplex potential can be done by computational methods without taking into account *in vitro* and *in vitro* assays verified data.

There is therefore an increasing demand for computational methods such as deep neural networks that can predict triplex-forming lncRNAs and DNA sites. Deep learning has been successfully applied on omics studies in recent years. In recent years, it has been established that convolutional neural networks achieve higher performance in predicting triplex-forming lncRNAsand DNA sites using DNA sequences than classical machine learning methods. However, low performance predictions may suggest that the information stored in the primary DNA sequence is not sufficient for building models effectively. In the present study therefore, we aim at recognizing these triplex-forming lncRNAs and DNA sites by deep learning architectures by taking histone modification marks as features. We would add histone modification marks to our DNA sequences to ascertain if there will be improvement from low quality predictions as reported in other previous works. We intend to use different deep learning architectures such as convolutional neural networks (CNN), and multilayer perceptron (MLP) and compare the performance of our models to know the best performing deep learning architecture in predicting triplex-forming lncRNAs and DNA sites.

Background

Non-Coding RNA and its Types

Sequencing of the human genome became a breakthrough in our knowledge of the major role of RNAs in gene expression regulation and chromatin modulation. Out of the sequenced human genome, 2% accounted for protein-coding genes while 98% accounted for non-coding genes, in spite of more than 90% of the genome being transcribed (Wilusz *et al*., 2009; Birney *et al*., 2007; Katayama *et al*., 2005). This was contrary to earlier believes on how cells of the mammalian origin work. Further research has revealed the roles of non- coding RNAs like lncRNAs in genome regulation and the etiology of human diseases (Maass *et al*., 2014; Rinn *et al*., 2007; Costa 2005). Non coding RNAs such as small ncRNAs (small interfering RNAs (siRNAs), PIWI-interacting RNAs, microRNAs (miRNAs), small nucleolar RNAs (snoRNAs) (Taft *et al*., 2010)); and long noncoding RNAs (lncRNAs), which are made up of more than 200 nucleotide in length were recently profiled as key regulators of gene regulation and chromatin modulation. The recent finding of an interesting group of regulators known as long non-coding RNAs (lncRNAs) has given us novel understanding in biology to explore for the etiology of different diseases ranging from cancer sub-typing, sepsis, aging, cardiovascular diseases to other human diseases. This knowledge has helped in recent times for therapeutic interventions of different ailments. Today, lncRNAs are known to be key classes of the varying RNA family in the living cells as shown in Figure 2.1A. Most definitions describing lncRNAs and their analogue transcripts can be found in Figure 2.1B.

They primarily regulate gene expression and epigenetic control of chromatin in many different cellular systems and pathways, X-chromosome inactivation and imprinting, promoter-specific gene regulation and mRNA stability (Hombach and Kretz, 2016). Mammals use Xist long non-coding RNA to silence (inactivate) the extra X chromosome in females to balance expression level in the two sexes. Xist induces a repressive epigenetic landscape thereby creating the inactiveX chromosome known as Xi (Charles and Ogawa, 2015). These long non-coding RNAs plays important roles in developmental stages of living things.

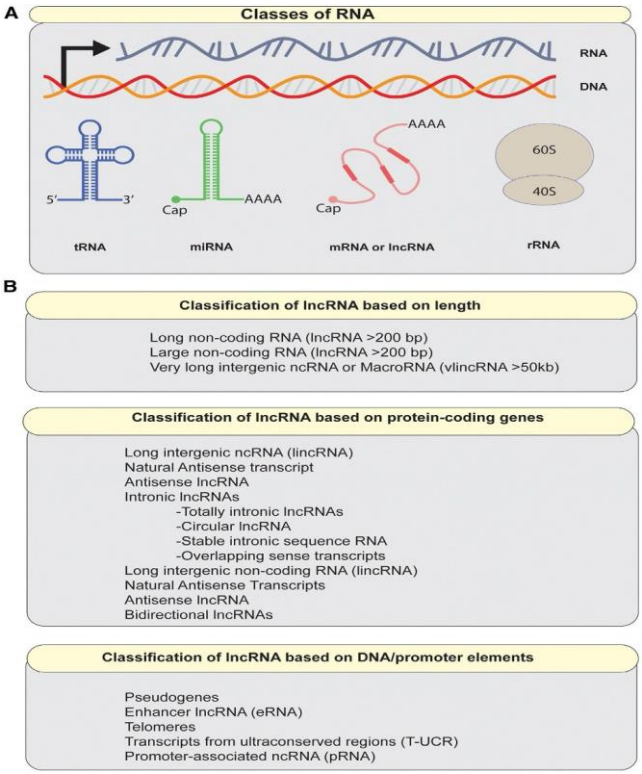

Fig 1.1: Types of RNA in mammalian cells. A; A main classes of RNA. Non coding RNAs may resemble mRNA structurally, but somecan have distinctive features, for example lack of polyA tail. B; Classes of noncoding RNAs >200 bp. qpcr stands for quantitative polymerase chain reaction (Sallam et al., 2018).

Long Non-Coding RNA and its Functions

Recently, the discovery of important biological molecules called long noncoding RNAs (lncRNAs) has brought a new dimension of diagnostic and therapeutic opportunities for the fight against human diseases. NONCODE database is a ncRNAs database (excluding tRNAs and rRNAs), are databases that contain over 100, 000 lncRNAs, out of which majority of the lncRNAs are yet to be annotated (Fang *et al*., 2018). It has long been established that tRNAs and rRNAs have structural role in biological processes without taking part in translation (Rinn and Chang, 2012; Mattick,2011). Long lncRNAs have more than 200 nt in length (Amaral *et al*., 2011). Long non-coding RNAs can be found in animals (Clemson *et al*., 1996; Brown *et al*., 1992), plants (Swiezewski *et al*., 2009), viruses (Reeves *et al*., 2007), yeast (Houseley *et al*., 2008), and prokaryotes (Bernstein *et al*.,1993). In comparison with other well-studied RNAs such as mRNAs, miRNAs, and snoRNAs, (Derrien *et al*., 2012; Pang *et al*., 2006; Wang *et al*., 2004), lncRNA's inter-specie conservation is poor which explains why its functionally is questioned. Research has shown that lncRNAs usually have low expression levels (Derrien *et al*., 2012) thereby being viewed as noise during transcription. This notwithstanding, the last two decades has provided us with several researches signifying that lncRNAs acts in different biological processes (Clark *et al*., 2011, Ponting *et al*., 2009) such as transcription (Martianov *et al*., 2007; Rinn *et al*., 2007), splicing (Tripathi *et al*., 2010; Rintala-Maki and Sutherland, 2009), translation (Beltran *et al*., 2008; Muddashetty *et al*., 2002), protein localization (Campalans *et al*., 2004; Watanabe and Yamamoto, 1994), cellular structure integrity (Sunwoo *et al*., 2009; Kloc *et al*., 2005), imprinting (Bartolomei *et al*., 1991; Brockdorff *et al*., 1991; Brown *et al*., 1991), cell cycle (Mourtada-Maarabouni *et al*., 2008; Ginger *et al*., 2006), and apoptosis (Huarte *et al*., 2010; Reeves *et al*., 2007), stem cell pluripotency (Sheik *et al*., 2010),reprogramming (Loewer *et al*., 2010), and heat shock response (Shamovsky *et al*., 2006; Espinoza *et al*., 2004). Scientists have demonstrated that lncRNAs may play a vital role in regulating cancer progression (Tano and Akimitsu, 2012). The roles of lncRNAs in the development of many other human diseases was also demonstrated by Wapinski and Chang, 2011.

Large amounts of lncRNAs have been profiled by different sequencing techniques. This has given us understanding on how these lncRNAs functions as as signals, decoys, scaffolds, guides, enhancer RNAs, and short peptides (Li *et al*., 2014; Moran *et al*., 2012). As signals, lncRNAs functions as molecular signals in response to stimuli during transcriptional regulation (Pandey *et al*., 2008). Decoy lncRNAs functions to reduce the availability of regulatory factors such as transcription factors, subunits of larger chromatin modifying complexes, miRNAs, and catalytic proteins by introducing what is known as "decoy" binding sites (Kallen *et al*., 2013). As scaffolds, multiple-component complexes, such as ribonucleoprotein (RNP) complexes are assembled on structures provided by lncRNA scaffold transcripts (Yang *et al*., 2014). For ribonucleoprotein (RNP) complexes, transcriptional activation or repression can occur depending on the nature of proteins and RNAs present as soon as the complexes have been fully assembled on the "decoy" binding sites (Aguilo *et al*., 2011; Kotake *et al*., 2011). In this particular case, guide lncRNAs usually directs ribonucleoprotein (RNP) complexes to specific target genes by interacting with them (Grote *et al*., 2013). Another example where these guide lncRNAs are important for the proper localization of the chromosome-modification complexes is in PRC2 and its subsequent repression of gene expression by the lncRNA fetal-lethal non-coding developmental regulatory RNA (*FENDRR*), which serves to bring PRC2 in close association to the promoters of genes linked with the build-up and differentiation of the lateral mesoderm lineage, such as forkhead box F1 (*FOXF1*) and paired-like homeodomain2 (*PITX2*) genes (Grote e*t al*., 2013). Enhancer regions usually produce enhancer RNAs (eRNAs) and these eRNAs may work by modulating chromatin interactions also called 3-dimensional (3D) organization of DNA. Long non-coding RNAs may also work as "tethers" if they are not released from the enhancer regions when performing their role, thereby tethering the interacting proteins to enhancer regions (Li *et al*., 2014). Long non-coding RNAs can also encode short peptides, which may also play vital role in biological processes (Ingolia *et al*., 2011). As more research are being carried out to probe and unveil other functions of

lncRNAs, there is no doubt, that in the near future, more functions of lncRNA will be reported in the fight against human diseases. Figure 2.2 below shows the mechanisms of lncRNAs.

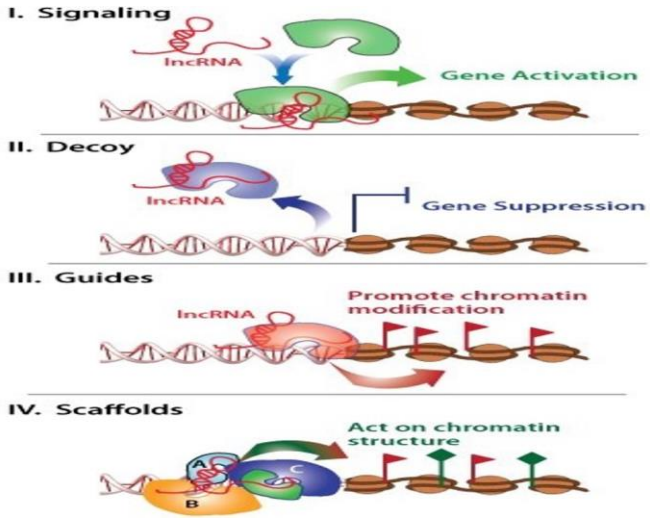

Fig 1.2: The molecular mechanisms of long non-coding RNAs. (Wang and Chang, 2011).

Cis and Trans-Acting lncRNA

Emerging evidence have shown that long non-coding RNAs can regulate gene expression by forming triplex structures either in cis-acting manner (on neighboring genes) or in trans-acting manner (on distant genes) (Guttman and Rinn, 2012; Guil and Esteller, 2012). Cis and trans-encoded RNAs are the most largely studied non-coding RNAs. While the first transcripts are in cis natural antisense relative to the targeted mRNA, the second transcripts in genomic regions are far from the targeted mRNA, therefore only cis-encoded RNAs have a perfect base pairing with the targeted mRNA (Azhikina *et al*., 2015). In the cis-acting manner, lncRNAs are usually transcribed from the promoter regions of target genes, form a triplex with promoter DNA sequences, and regulate gene expression while lncRNAs specifically target binding sites through triplex formation with distant DNA sequences in the trans-acting manner. While PARTICLE, Khps1, FENDRR, pRNA of rRNA gene, and pRNA of *DHFR* forms cis, MEG3 and HOTAIR forms trans conformations (Yue *et al*., 2016). Figure 1.3 indicates gene regulatory functions in cis or trans lncRNA.

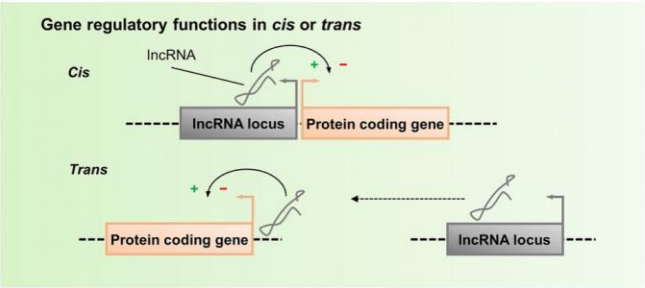

Fig 1.3: Gene regulatory functions in cis or trans lncRNA (Florian Kopp, 2019)

RNA:DNA Triplex Forming Long Non-Coding RNA and their Structure

Several long non-coding RNAs (lncRNAs) have been verified experimentally by sequencing and other computational methods to form RNA:DNA triplex structures. These lncRNAs that have the potential of forming triplex structures include, HOTAIR, MEG3, PARTICLE, DACOR1, MIR100HG, FENDRR, ANRIL, TUG1, MALAT1, LINC00599, TINCR, NEAT1, roX2, DHFR, OTX2-AS1, Xist, SNHG16,
ATXN8OS, BCYRN1, TERC, Khps1, among many others (Kazimierczyk *et al*., 2020). Recently, studies have shown that MEG3 lncRNA exerts their function by interacting with enhancers close to SMAD2, TGFBR1 and TGFB2 by forming triplex structures (Mondal *et al*., 2015). Mondal *et al* reported that by binding to distal regulatory elements, MEG3 is able to regulate the activity ofTGF-β genes. The GA-rich sequences of MEG3 binding sites can form RNA–DNA triplex structures with MEG3 which makes it easier to modulate chromatin activity.
MEG3 binding sites have GA-rich sequences (Mondal *et al*., 2015). HOTAIR lncRNA has been reported to interact with more than 900 genomic regions close to the binding sites of EZH2 and SUZ12 in human cancer cells (Chu *et al*., 2011). It has been reported HOTAIR lncRNA binds with DNA regions in a trans manner thereby recruiting PRC2 and LSD1-CoREST chromatin modifying complexes, resulting to transcription repression (Angrand *et al*., 2015). By forming RNA–DNA triplex structures, HOTAIR lncRNA is able to exert their function on the differentiation of mesenchymal stem cells (MSCs) and its association with senescence-associated DNA-methylation (DNAm) (Kalwa *et al*., 2016). O'Leary *et al* reported that the antisense lncRNA PARTICLE takes part in RNA-DNA triplex structure formation and is expressed and transcribed from the promoter of MAT2A when cells are exposed to low-dose of irradiation (O'Leary *et al*., 2017, O'Leary *et al*., 2015). By forming RNA:DNA triple helices, FENDRR lncRNA recruits PRC2 thereby interacting with Trithorax group and Mixed lineage leukemia (TrxG/Mll) complex (Grote and Herrmann, 2013). The lncRNA ANRIL forms RNA:DNA triple helices by binding with the DNA in a cis manner, thereby recruiting PRC1 and PRC2, leading to transcription repression (Angrand *et al*., 2015). For dosage compensation in mammalian females, the Xist lncRNA (X-inactivation specific transcripts) forms RNA:DNA triple helices with DNA regions and later recruits PRC2 to drop H3K27me3 repressive histone modification marks along the X-chromosome, thereby inactivating the marked copy (Zhao *et al*., 2008; Jeon and Lee, 2011). The formation of RNA:DNA triple helices between MIR100HG lncRNA and DNA regions plays an important role in triple-negative breast cancer (TNBC) progression by regulating p27 gene (Wang *et al*., 2018). The identification of Tug1-binding element (TBE) upstream of the *Ppargc1a* gene and its subsequent binding with Tug1 lncRNA have been linked to increase in *Ppargc1a* promoter activity (Jianyin *et al*., 2016). Research has shown that a direct interaction between Tug1

and PGC-1α regulates mitochondrial bioenergetics in podocytes in the diabetic milieu (Jianyin *et al.*, 2016). MALAT1 lncRNA carrying ENE and A-rich tract have also been reported to increase the levels of intronless β-globin reporter RNA (Jessica *et al.*, 2012). An antisense RNA known as khps1 has been confirmed to form RNA:DNA–DNA triplex structures at the promoter region of Sphingosine kinase 1 (SPHK1), thereby regulating SPHK1 gene through histone acetylation that subsequently result in increase in cell proliferation activity (Postepska-Igielska *et al.*, 2015). The role of NEAT1 lncRNA in leukemia and ovarian cancer through the formation of RNA:DNA triple helices have been revealed. It does this by regulating ADARB2 expression through protein sequestration into paraspeckles and also plays an important role in the knockdown results in inhibition of cell growth (Ke *et al.*, 2016; Hirose *et al.*, 2014). For a given RNA sequence, $R = r1r2\cdots rn$, and DNA sequence $D = d1d2\cdots d0$, the triplex detection problem is the search for (maximum length) substrings *ri…rj and du..dv* from *r* and *d* with minimum length *l* (). Figure 1.4 below shows long non-coding RNAs in prostate cancer.

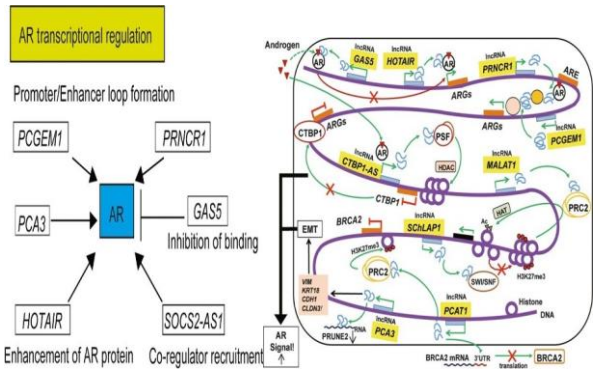

Fig 1.4: Long non-coding RNAs in prostate cancer (Misawa *et al.*, 2017).

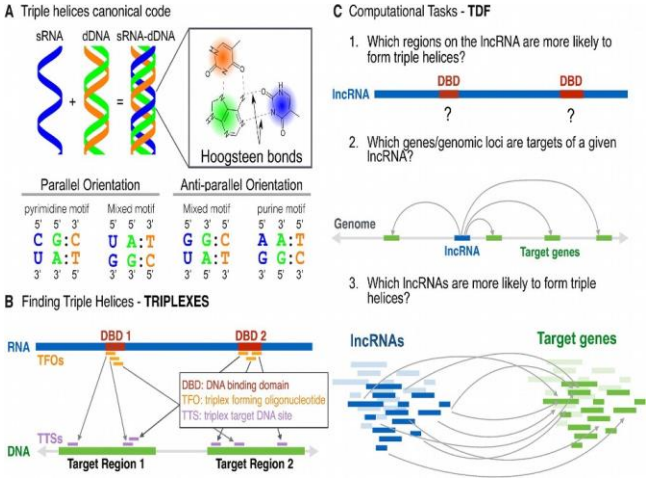

Fig 1.5: The computational framework of *triplexes* and TDF. (A) Triplexes are formed by binding of single-strandedRNA (blue) with a purine-rich strand (green) of a double-stranded DNA through Hoogsteen base pairing rule. For the formation of a triplex in the parallel orientation, a pyrimidine or mixed motifs are needed, but the anti-parallel orientation needs a purine or mixed motifs. (B) For a given RNA and DNA sequence, T *triplexes* identifies candidate triple helices with a minimum size and maximum number of mismatches following one of the canonical codes. Eachtriplex is formed by one RNA sequence (triplex forming oligo – TFO) and a DNA region (triple target sites – TTS). Contiguous regions with overlapping TFOs (marked in red) define a DNA-binding domain. (C) TDF performs statistical tests by combing predictions from *triplexes* to answer the following questions: (1) which regions of an RNA (DBD) are more likely to form triple helices with particular DNA target regions? (2) Which DNA regions (target genes) are more likely to be targeted by the RNA? and (3) which lncRNAs are more likely to form triple helices in a set of target DNA regions? (Kuo *et al.*, 2019).

Long Non-Coding RNA (lncRNA) Interactome

Long non-coding RNAs are capable of interacting with DNA, messsenger RNAs (mRNA), proteins, peptides and micro RNAs (mRNA) – the so called competitive, endogenous RNA (ceRNA) hypothesis.

Interactions of lncRNA with miRNAs - the competitive, endogenous RNA (ceRNA) hypothesis

Vast novel evidence indicates that miRNAs interacts with lncRNAs by binding to specific mRNA 3′ UTR regions thereby modulating the expression of these genes (Krol et al., 2010). Large amounts of experimental data suggests that ncRNAs interacts with miRNAs and this evidence was named as competitive, endogenous RNA (ceRNA) hypothesis (Smillie *et al.*, 2018; Yay *et al.*, 2014; Salmena *et al.*, 2011).

Pairing LncRNAs with Messenger RNAs (mRNA)

Bioinformatics tools such as LncTar and RNAplexn have given us knowledge on the kind of interaction between lncRNA and mRNA (Li *et al.*, 2015; Tafer and Hofacker, 2008). LncRNA interacts with mRNA and the pre-mRNA by direct base pairing, which then plays a vital role in modulating mRNA splicing and editing may participate in regulating translation by affecting mRNA splicing and editing together with its stability (Romero-Barrios *et al.*, 2018; Karapetyan *et al.*, 2013). Long non-coding RNA can regulate pre-mRNA splicing by either blocking spliceosome assembly involving intron-exon junction or by becoming the target for splicing factors (Romero-Barrios *et al.*, 2018). More experimental evidence is needed to confirm lncRNA–mRNA interaction and its role in alternative splicing as research proposed the involvement of MALAT1 lncRNA in modulating alternative splicing of pre-mRNA (Tripathi *et al.*, 2010).

LncRNA–DNA Interactions

There are a lot of approaches depicting the interaction of lncRNAs and DNA regions (Quinodoz andGuttman, 2014). It has been demonstrated that LncRNAs can form RNA-DNA-DNA triple helix (Li *et al.*, 2016; Shaw and Arya, 2008). For example, research has demonstrated the regulatory role of DHFR lncRNA in suppressing the transcription of Dfhr mRNA through the formation of triple helices with the DHFR promoter, thus resulting in the binding of DHFR lncRNA to the TFIIB and blocking the formation of transcription initiation complex (Deng *et al.*, 2009; Blume *et al.*, 2003).

LncRNA–Protein Interactome

Novel evidence has emerged that long non-coding RNAs (lncRNAs) weight ranging from hundreds to tens of thousands of nucleotides plays vital roles as either signals, decoys, guides, and scaffolds for majority of proteins (Sun *et al*., 2017; Wang and Chang 2011).
Today, there are three approaches to study the interaction between lncRNAs and proteins. These include the immunoprecipitation method, crosslinking and immunoprecipitation (CLIP) method and Next Generation Sequencing (NGS) approach. The immunoprecipitation approach has been applied in the preparation of RNA-protein complexes while on the other hand, crosslinking and immunoprecipitation method (for example HITS-CLIP, PAR-CLIP, and iCLIP) works to map out proteins and protein binding sites interacting with a sequence of lncRNA (Jalali *et al*., 2018; Garzia et al., 2017; Yoon and Gorospe, 2016). To accelerate the the analysis of RNA-protein complexes, Next Generation Sequencing (NGS) have been rapidly applied in generating data and studying the interaction between lncRNAs and protein. Computational approaches that seek to predict the structure and function of lncRNA-protein by screening thousands of publicly available data such as annotated sequences and structural motifs (for example, RNA-binding sites) are gaining grounds in recent years.

LncRNA-Peptides and small weight compounds Interactome

Research has shown that small portions of lncRNAs in the cytosol can code for small peptides whose amino acid composition is less than 100 (Chugunova *et al*., 2018; Matsumoto and Nakayama, 2018; Zhu *et al*., 2018). For example, a 16.5 kb long skeletal and muscle-specific lncRNA known as LINC00948 has been confirmed to encode for myoregulin (a 46 amino acid peptide), thus modulating the biosynthesis of skeletal muscles by interacting with sarcoplasmic reticulum Ca 2+ -ATPase (SERCA) (Anderson *et al*., 2015), lysosomal v-ATPase (Matsumoto *et al*., 2017) or removing SERCA inhibitors (Nelson *et al*., 2016). Thetask of profiling lncRNA-origin peptides is a difficult one since *in silico* methods predicts large numbers of Open Reading Frames (ORF), meanwhile very little of them are encoded to yieldproteins. Ruiz-Orera *et al* estimated that only about out of the transcribed lncRNAs, only about 23% of them translated to proteins (Ruiz-Orera *et al*., 2016). The use of bioinformatic algorithms and experimental confirmation has helped in profiling peptide expression. Large amounts of RNAs such as rRNA, mRNA, ribozymes, riboswitches, and aptamers forms small structures containing "pockets" that are capable of binding with compounds with small molecular weights (Warner *et al*., 2018; Gestelanf *et al*., 2005). Figure 1.6 shows the interaction of human lncRNA with cellular biomolecules.

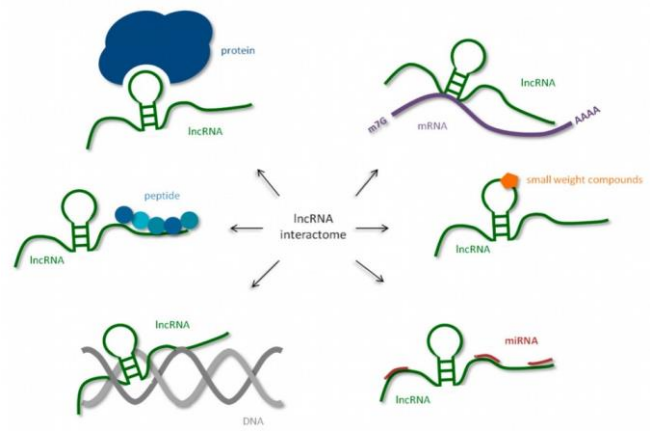

Fig 1.6: Human long non-coding RNAs (lncRNA) interactome, interaction of human lncRNA with cellular biomolecules (Kazimierczyk *et al*., 2020).

Sequencing Technologies for Identifying Long Non-Coding RNA

Either examining mammalian cells, tissues, or serum samples, genome-wide transcriptomic methods is the basis of profiling long non-coding RNAs. Different computational techniques such as classical machine learning and deep learning methods can also identify long non-coding RNAs from RNA-seq data. Generally, RNA-seq data can be analyzed through two different ways: de novo assembly methods by using software packages such as Bridger (Chang *et al*., 2015), SOAP2 (Li *et al*., 2009), Oases (Schulz *et al*., 2012), or Trinity (Grabherr *et al*, 2011), or by alignment against a known standard of defined lncRNAs. The alignment against a known standard of defined lncRNAs method is rapidly used today due to increasing lncRNA annotations.
Even though the understanding about histone modification marks that defines active transcription can be helpful, it is not needed in screening lncRNAs by sequencing approaches. At its promoter, H3K4me3 marks a subunit of lncRNAs showing an element of mammalian conservation while H3K36me3 does so along the region where it is transcribed (Kaikkonen *et al*., 2011; Guttman *et al*., 2009). This chromatin signature is sometimes known as K4k36. Even though histone modification marks can help in identifying lncRNAs, more research is being carried out to know things that differentiate long non-coding RNAs from other non-coding RNAs that regulates transcription for example, enhancer RNAs. The main work of chromatin-binding patterns to define transcription regulators has limitations even though it is one among the core factors that differentiate lncRNAs from other transcription regulatory elements (Li *et al*., 2016; Li *et al*., 2013).
Despite RNA-seq being the most widely used sequencing technique for RNA screening, there are other methods that can be used in detecting RNAs and these include;

(i) serial analysis of gene expression (SAGE) and paired-end tag–expression approaches thattargets polyA tail (Philippe *et al*., 2014; Fullwood *et al*., 2009) ;
(ii) GrRO-seq: is a nuclear run-on assay coupled to deep sequencing to estimate nascent transcription (Gardini,

2017; St Laurent *et al*., 2015).

(iii) direct RNA sequencing, that requires sequencing of undegraded RNA bypassing library preparation (Ozsolak *et al*., 2009); and

(iv) cap analysis of gene expression (CAGE), which profiles RNAs with 5′ cap and can accurately map the 5′ transcript end which makes it more advantageous (Takahashi *et al*., 2012).

Today, several screening microarrays are available in commercial quantities that can help in profiling protein-coding and non-coding transcripts from human or mouse samples (Zhu *et al*.,2017; Shi and Shang, 2016; Chen *et al*., 2015).

Microarrays have proven to be fast and efficient which may be advantageous over RNA-seq.

Nuclear lncRNAs may determine transcriptional products in different ways, which include gene expression regulation by interacting with transcription factors, histone modulation, and its regulatory role in the splicing process and subsequent transport of mRNA out of the nucleus (Wang KC and Chang HY, 2011).

Recently, different sequencing approaches have been developed that can assist in profiling lncRNA interactions (Figure 2.7; 2.7.1 & 2.7.2). The first chromatin affinity approach (chromatin isolation by RNA purification sequencing - ChIRP) to characterize lncRNA interactions was developed by Chu *et al* (Chu *et al*., 2012; Chu *et al*., 2011). Several other research later validated this technique (Sallam *et al*., 2016). Chromatin isolation by RNA purification sequencing (ChIRP) method normally captures lncRNA of interest and probes its interaction by tiling complimentary antisense oligos. By so doing, ChIRP-Seq technique is able to define specific DNA binding sites of a particular lncRNA and by coupling this with mass spectroscopy, it helps in characterizing interacting protein partners. By this approach, we have been able to understand the complex interactions of Xist lncRNA and its relationship with hnRNP K and Spen (Spen family transcriptional repressor), which are needed for mRNA processing (Chu *et al*., 2015). ChIRP-Seq has been applied in studying the interactions between RNA–RNA (Chu *et al*., 2012). RNA antisense purification - RAP (Engreitz *et al*., 2015) and capture hybridization analysis of RNA targets - CHART (Simon *et al*., 2011) are also used in studying the binding motifs based on biotin-tagged oligonucleotide retrieval. While RAP uses longer probes (≈90 base pair designed with 5′ biotin tags), ChIRP uses 20 nucleotide probes with 3′ tags. CHART-Seq also uses 20 mer olignonucleotides, although what it does is to first test individual olignonucleotides that targets ncRNA regions that are accessible by using RNase sensitivity. ChOP-Seq is another sequencing technique that has been applied in studying the kind of interaction lncRNA has with other biological molecules. ChOP-Seq has been applied in studying the complex interactions of kcnq1ot1 lncRNA that take part in the regulation of genes within the *Kcnq1* imprinting domain (Zhang *et al*., 2014).

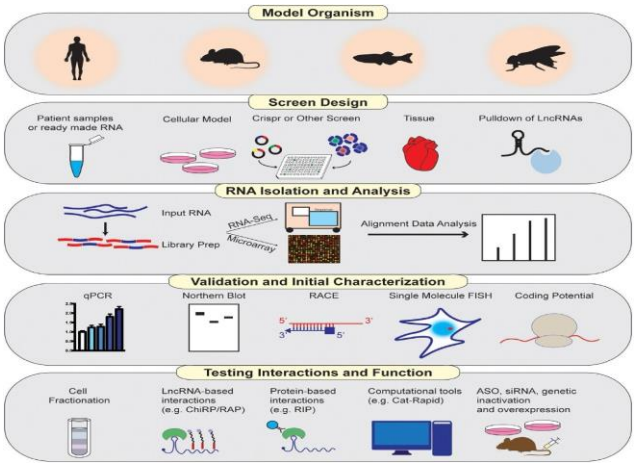

Fig 1.7: Pipeline for long non-coding RNA (lncRNA) discovery and characterization (Sallam *et al*., 2018).

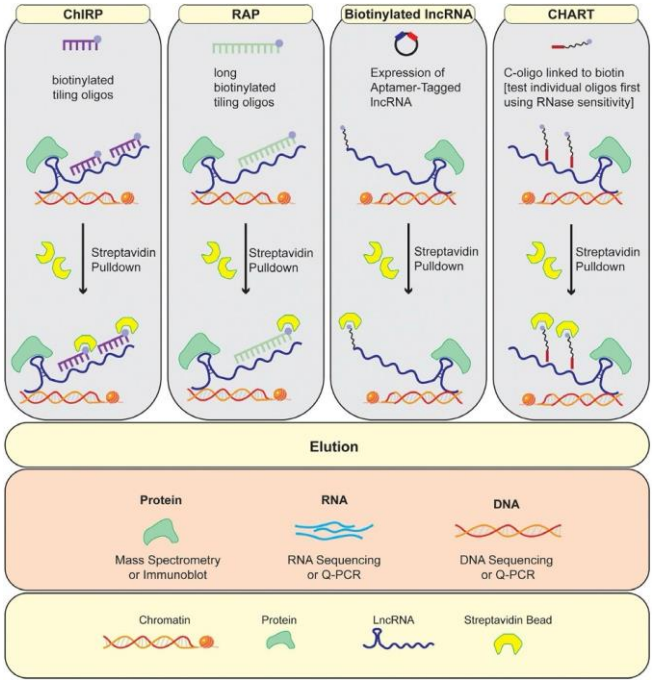

Fig 1.7.1.: Summary of techniques used to probe a long non-coding RNAs (lncRNA) interactome (Sallam*et al*., 2018).

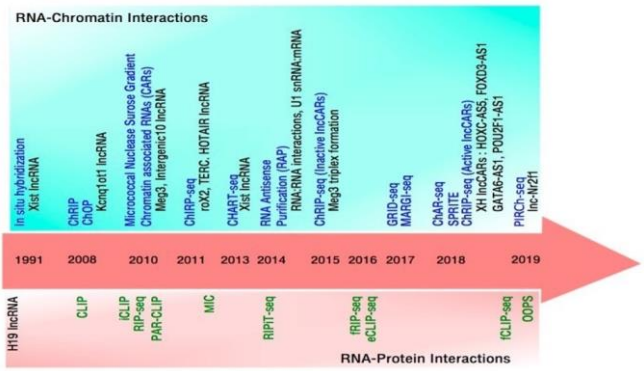

Fig 1.7.2: Timeline of technological advances to study RNA-Protein and RNA-Chromatin interactions. Upper panel (light sea-green box) shows in

chronological order the prominent methods (in blue) to detect RNA-chromatin interactions. Examples of some of the functionally verified lncRNAs from each of these studies are shown (in black) below the corresponding method. Middle panel shows the year in which these methodologies were reported Lower panel (light pink box) likewise shows in chronological order methods (in green) to recognize RNA-proteins interactions (Mishra and Kanduri, 2019).

Base Pairing Rules-Related Mathematics Statistics for Predicting DNA:RNA Triplex Structures

Computational approaches play an important role in recognizing lncRNA triplex structures. Earlier computational approaches did not mark out triplex forming RNA regions, because they were based on searching for purine rich DNA regions (Gaddis *et al*., 2006; Goñi *et al*., 2004). Subsequently, another algorithm known as Triplexator was developed which efficiently recognized lncRNA triplex structures in large DNA sequences (Buske *et al*., 2012). Triplexator was able to categorize all regions of RNA and DNA sequences that would take part in the formation of triplex structures with size larger than l basepair and with k maximum mismatches. Today, Triplexator is a widely used computational tool that recognizes tens of thousands of triplex structures for a single lncRNA, although it provides little statistics to identify significant triple helices. Triplexator is able to recognize vast amounts of lncRNA triplex structures including; HOTAIR, MEG3, PARTICLE, DACOR1, MIR100HG, FENDRR, ANRIL, TUG1, MALAT1, LINC00599, TINCR, NEAT1, roX2, DHFR, OTX2-AS1, Xist, SNHG16, ATXN8OS, BCYRN1, TERC, Khps1, among others (Kazimierczyk *et al*., 2020). Other computational approaches for recognizing triplex structures include, LongTarget, Triplex-Inspector and Triplex Domain Finder (He *et al*., 2014 Buske *et al*., 2013). LongTarget is a web-based technique which estimates only a single DNA region at a time and has been shown recently to be significantly slower than Triplexator and as a result it hinders the analysis of large amounts of RNA or DNA regions (Antonov *et al*., 2018). Triplex-Inspector is another computational method for evaluating triple helices. It is for sequence- specific ligand design and suitable target selectionwhich increases the accuracy of triplex-based genomic manipulations (Buske *et al*., 2013). The algorithm takes a genomic locus of interest and searches for all putative triplex target sites, leveraging our exhaustive search algorithm Triplexator (Buske *et al*., 2012). The last computational approach we mentioned above is the Triplex Domain Finder (TDF) which recognizes the triplex-forming potential between lncRNA and DNA regions such as promoters of differentially regulated genes after knockdown of the lncRNA (Kuo *et al*., 2019).

Classical Machine Learning Methods in Predicting DNA:RNA Triplex Structures

Recently, vast amounts of biological data have posed a huge challenge in understanding and interpreting them correctly. Having knowledge of these big data will not only answer questions about the etiology of animal and plant diseases; it provides an opportunity to appreciate technological advancements in computational sciences. Data generating technologies such as next generation sequencing (NGS) have enabled us to screen large amounts of omics data such as genomic data, gene expression and small RNA abundance data, epigenetic modification data, protein binding motifs, and chromosome conformation in a as shown in (Figure 2.). These omics data come with its own challenges such as heterogeneity, noisy and highly dimensional (Zitnik *etal*., 2019; Altman and Krzywinski, 2018). Today, machine learning/deep learning has become a powerful computational tool in the field of biology. Machine learning-deep learning learns new features from big omics data, thus answering penitent research questions. Classical machine learning tools such as support vector machines, random forests, decision trees, etc have been applied variously on omics data in the past three decades. While most classical machine learning methods are supervised learning, deep learning on the other hand is unsupervised learning. These methods have proven over the years to be fast, efficient and easy to understand and interpret. Figure 1.8 shows machine learning workflow for analyzing complex biological data.

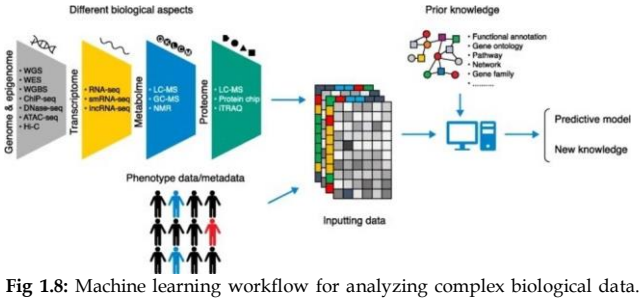

Fig 1.8: Machine learning workflow for analyzing complex biological data. High-throughput data generation techniques for different biological aspects are shown (*left*). *ATAC seq* assay for transposase-accessible chromatin using sequencing, *ChIP-seq* chromatin immunoprecipitation sequencing, *Dnase-seq* DNase I hypersensitive sites sequencing, *GC-MS* gas chromatography- mass spectrometry, *LC-MS* liquid chromatography–mass spectrometry, *lncRNA-seq* long non- coding RNA sequencing, *NMR* nuclear magnetic resonance, *RNA-seq* RNA sequencing, *smRNA- seq* small RNA sequencing, *WES* whole exome sequencing, *WGBS* whole-genome bisulfite sequencing, *WGS* whole genome sequencing, *Hi-C* chromatin conformation capture combined with deep sequencing, *iTRAQ* isobaric tags for relative and absolute quantification (Xu and Jackson, 2019).

Support Vector Machines (SVM)

The support vector machine (SVM) is a computational algorithm that seeks to learn from the data via pattern recognition. This computational tool has been applied on several big omics data including DNA sequences, protein sequences, microarray expression profiles, protein-protein interaction networks, and tandem mass spectra among others (Zitnik *et al*., 2019).

Binary classification is the easiest way of making predictions, thus discriminating between objects that belong to one of two categories, either positive (+1) or

negative (-1). SVMs has two main components of solving this task, either by large margin separation or by kernel functions (Ben-Hur *et al*., 2008). Classification of points in two dimensions forms the basis of large margin separation and is done by drawing a straight line and naming points on one side as positive (+) and on the other side as negative (-) (Ben-Hur *et al*., 2008). This can be expressed mathematically as:

$$y \in \{-1, 1\}$$

where y is the output while 1 is the positive class and -1 is the negative class. Kernel functions usually take two vectors of any dimension as its input and gives an output score that shows similarities between input vectors (Ben-Hur *et al*., 2008). This means that if the dot product is large, then it means that the vectors are similar, but if the dot product is small, it therefore means that the vectors are different. Mathematically, a Kernel function can be written as:

$$K(x, y) = \quad < \phi(x), \phi(y) >$$

where **x** and **y** are input vectors, $\boldsymbol{\phi}$ is a transformation function and <, > denotes dot product operation. $\boldsymbol{\phi}$ only maps the input vector to itself in dot product function.

Random Forests

Random forests (RF) is a powerful algorithm that can be applied on both small and big data. Since it can study interactions between features and interpret its correlation, it is applied on high-dimensional omics data (Wouter *et al*., 2013). Random forest models are fast to train, non-parametric, difficult to over-train, very robust to outliers and noise and can be trained without hyperparameter tuning, even though tuning hyperparameters is good for training machine learning-deep learning models (Wouter *et al*.,2013; Breiman, 2001). Random forests can train an ensemble of decision trees singly according to their samples, variables and class designation (Wouter *et al*., 2013). The name random forest is derived from the idea of using random subset of variables and samples to build every tree. Mathematically, for each classifier; $hk(\boldsymbol{x}) \equiv h(\boldsymbol{x} \mid \Theta\Theta k)$ is a predictor of n

$$y = \pm 1 = \text{outcome with input } \mathbf{x}$$

Decision trees (DTs)

Decision trees (**DTs**) are supervised learning and they are fast to train, non- parametric, simple to understand and interpret, can work on small amount of data, and have been applied on regression and classification tasks in biology (Carl and Steven, 2009). The aim is to generate a model that can predict the value of a target variable. It does this by learning simple decision rules gathered from the data features (Carl and Steven, 2009). The application of decision trees in biology spansfrom predicting splice sites to assigning functions to different protein (Carl and Steven, 2009).

Deep Learning Methods in Predicting DNA:RNA Triplex Structures

In recent years, deep learning (DL) architectures have recorded successes in speech recognition, image recognition and language processing (LeCun Y *et al*., 2015). Deep learning is a subset of machine learning algorithm. Today, different deep neural networks like convolutional neural networks and recurrent neural networks have been applied on different biological tasks including splice site detection, protein-ligand scoring, protein-protein interaction prediction, protein secondary structure prediction, protein contact map prediction, compound toxicity, liver injury prediction, RNA-DNA triplex helix structure recognition, protein design drug repurposing among others (Yongqing *et al*., 2020). Since they can learn new important features from the ever increasing heterogeneous and complex omicsdata via interactions within the layers, it has helped in the diagnosis, treatment, and classification of different complex diseases. Deep neural network's goal is to calculate a mathematical function say f that takes a number of inputs called, x, to their respective outputs known as, y, such that y

$= f(x)$. For a simple feed-forward network, $y = f(x;w) = LN(LN\text{-}1(\ldots L1(x))$ can be defined as a composition of N nonlinear transformations $Li(1<=i<=N)$ where each function Li corresponds to a hidden layer activation, while w is the learnable weight contained in all filter bank layers that are updated in the course of the training process (Martorell-Marugán *et al*., 2019).

Convolutional Neural Network (CNN)

Convolutional neural networks (CNNs) have become a conventional tool for the analysis of omics data as feedforward networks (Waseem *et al*., 2017). The performance of CNNs largely depends on the technical know-how on tuning both internal and external parameters. CNNs have worked well on DNA sequences over the years, hence providing answers to numerous biological questions and paving way for disease sub-typing and treatment. Convolutional layer, pooling layer (subsampling), and fully connected layer are the building block layers of a convolutional neural network architecture despite coming in different forms (Waseem *et al*., 2017). The convolutional layers usually extract features and learns how the neurons in this layer are represented as feature maps (Waseem *et al*., 2017).

Recurrent Neural Network (RNN)

Recurrent neural networks (RNNs) basically make use of sequential data. The network works such that the decision of the first output say time t1 affects the decision of the second output say time t2. RNN have two input sources, the present source and the preceding source, and the

coming together of these two input sources determines the kind of respond to a new data (Martorell- Marugán *et al*., 2019). Evidence from research have shown that recurrent neural networks (RNN) performs better on time-series data than other machine learning and deep learning architectures (Shen *et al*., 2018).

Residual Neural Network (ResNN)

Residual neural networks (ResNN), was first reported by He *et al* in 2016 (He *et al*., 2016) and since then, it has outperformed other machine learning and deep learning architectures. The performance of ResNN in practice can be credited to the skip connections between layers that helps in the propagation of the gradient round the network. Aside from solving the gradient problem in deep networks, the skip connections establishes a dependency between variables in different layers that can be seen as a system state (Ciccone *et al*., 2018; Chang *et al*., 2018; Haber and Ruthotto, 2017; Lu *et al*., 2017; Liao and Poggio, 2016; Ruthotto and Haber, 2018; Chaudhari *et al*., 2017).

Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) is also a class of feedforward artificial neural network and it comprise of at least three layers of nodes namely; an input layer, a hidden layer and an output layer. Apart from the input nodes, each node is a neuron that makes use of a nonlinear activation function. For model training, multilayer perceptron (MLP) uses a supervised learning method known as backpropagation. The distinguishing factor between MLP and a linear perceptron is its multiple layers and non-linear activation. This means that it can distinguish a data that is not linearly separable. Multilayer perceptron (MLP) has been widely applied on omics data. For example, multilayer perceptron (MLP) has been utilized for the analysis of immunohistochemical scores of tissue microarrays for breast cancer biomarker discovery (McKenna *et al*., 2013).

Long-/Short-Term Memory (LSTM)

The main problem of RNNs is that of vanishing gradients, thereby significantly limiting its ability to work with long sequences (Martorell-Marugán *et al*., 2019). To solve this problem, Long-/short-term memory **(**LSTM) and gated recurrent units partially address this issue (Martorell-Marugán *et al*., 2019). Long-/short- term memory (LSTM) seeks to maintain the error that can be back-propagated via time and layers after learning over several steps.

Methods of Hyperparameter Optimization and Others

Hyperparameters are basic building blocks that occurs in all machine learning models to check their structure and performance. Some of the external hyperparameters that are optimized for better accuracy in model training include; learning rate, grid search, training epouchs, Adams optimization algorithm, etc. The learning rate, also known as *step size* defines how much we need to configure the model based on the estimated error when weights are calculated and this is usually done in the range of 0.0 to 1.0 (Park *et al*., 2020). Tensorflow and Pytorch works best at a learning rate of 0.001 while in Natural Language Processing, 0.002 and 0.003 has been reported to perform best. Grid search is an automatic hyperparameter optimization technique that is user-friendly, but may not work efficiently when the number of parameters is large. Epochs specifies the number of rounds of the entire training dataset the machine learning algorithm and deep neural network has completed. Using many epochs to train the model may lead to overfitting, thus making the model to learn even the noise from the parameters. Adam is an optimization algorithm used in place of stochastic gradient descent to update model weights in the training data (Diederik and Jimmy, 2015).

Activation Function

Activation functions are vital in constructing neural networks. Today, ReLU activation function is the default activation function for hidden layers while the output activation function depends on the task you are predicting. Rectified Linear Activation (**ReLU**), Logistic (**Sigmoid**), and Hyperbolic Tangent (**Tanh**) are widely used as hidden layer activation function for deep learning models. Recurrent neural networks uses Sigmoid and Tanh activation functions. Most times, LSTM uses Sigmoid activation function for recurrent connections and Tanh activation as an output activation function. Below are the hidden layer activation functions widely used in MLP, CNN and RNN;

Multilayer Perceptron (MLP): ReLU activation function.

Convolutional Neural Network (CNN): ReLU activation function.

Recurrent Neural Network (RNN): Tanh and/or Sigmoid activation function.

For the output layer, there are three commonly used activation functions, namely; Linear activation function, Logistic (Sigmoid) activation function, and Softmax activation function. Below is how output activation functions can be chosen in different tasks.

Binary Classification: One node, Sigmoid activation.

Multiclass Classification: One node per class, Softmax activation.

Multilabel Classification: One node per class, Sigmoid activation.

Regression: Linear activation function

Performance Metrices

An important step for machine learning methods is the

choice of metrics that can measure the real performance of models. Every performance metric has a particular feature and calculates properties that may be different from the results that has been predicted (Orozco-Arias *et al*., 2020). The performance of regression models can be determined by the following performance metrices; Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 (R-Squared) while the performance of classification models can be calculated by the following metrices; Accuracy, Confusion Matrix (not a metric but fundamental to others), Precision and Recall, F1-score, AU- ROC, and MCC (Orozco-Arias *et al*., 2020).

Loss Functions

The loss function to use on any machine learning-deep learning tasks depends on the kind of task you want to solve. Different loss functions exists for different tasks as follow;
Regression Task: Here a real-value quantity is predicted where the output layer configuration is a one node with a linear activation unit and the loss function is mean squared error (MSE) (Wang *et al*., 2020).
Binary Classification Task: Here you usually classify an example as belonging to one of two the classes where the output layer configuration is a one node with a sigmoid activation unit and the loss function is cross-entropy, also known as Logarithmic loss. The task is designed as predicting the likelihood of an example belonging to class one (Wang *et al*., 2020).

Multi-Class Classification Task:

Here you classify an example as belonging to one of more than two classes where the output layer configuration is a one node for each class using the softmax activation function and the loss function is cross-entropy, also known as Logarithmic loss. The task is designed as predicting the likelihood of an example belonging to each class (Wang *et al*., 2020).

Materials and methods

Data Collection and Preprocessing

We installed biopython after which we cloned the GitHub repository of Zhang *et al.,* 2020 to access the data to a hosted runtime provided by Google (Google Colab).

Data Collection & Preprocessing for our model: Triplex Forming Prediction Model (TriplexFPM-HMM) for Histone Modification Marks-Based Feature

We took *2547* triplex positive DNA regions from Sentürk's paper (GEO Accession No: GSM3417036), *3014* triplex negative DNA regions from Sentürk's paper (GEO Accession No: GSM3417036) as our first negative data and *36597* promoter regions of DNA from Ensembl Release 104 of May 2021 as our second negative data.

Triplex DNA Sites Prediction Positive Dataset for TriplexFPM-HMM

We took *2547* triplex positive DNA regions as a bed file from Sentürk's paper (GEO Accession No: GSM3417036), we then removed all other columns apart from the chromosome number, start and end regions to enable us add histone modification marks.

Triplex DNA Sites Prediction First Negative Dataset for TriplexFPM-HMM

In the same way, we also took *3014* triplex negative DNA regions as a bed file from Sentürk's paper (GEO Accession No: GSM3417036) as our first negative data. Using pandas, we deleted all other columns and the header apart from the chromosome number, start and end regions to enable us add histone modification marks.

We took our 12 histone modification read counts from ENCODE database No. E123 cell line out of k562 cell lines (Inoue F. *et al*., 2017; Kazachenka *et al*., 2018), we converted the 'tagalign.gz'format to 'bam' by following the instruction given on DeepChrome (Ritambhara *et al*., 2016). Wethen added our 12 histone modification marks to both our positive and negative data using bedtools multicov command.

Overall, we had *5561 DNA regions* and 12 histone modification features (H3K36me3, H3K27ac, H3K9me3, H3K4me3, H3K4me2, H3K9ac, H3K9me1, H3K4me1, H2AFZ, H3K27me3, H3K79me2, H4K20me1). *2547* DNA regions were Positive while *3014* were Negative DNA regions. After filtering out sequences greater than 2000 Bps, we decreased the number of regions from *5561 to 5473.*

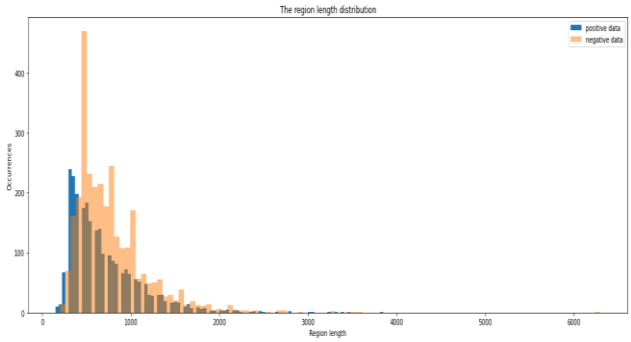

Fig 2.1: Triplex DNA site length distributions for positive & first negative dataset for TriplexFPM-HMM

Model Construction for First Negative Dataset TriplexFPM-HMM
We splitted our filtered data *(5473)* into three sets; train *(0.8), validation (0.1) and test (0.1) sets* and trained MLP and CNN.

Hyperparameter Tuning for First Negative Dataset TriplexFPM-HMM

We set other internal and external hyperparameters as follow: epochs (50), learning rate at (1e-2), input shape (12, 1) Kernel size (32), verbose (1) and used Keras-Tensorflow to train our models. We used six different metrices in assessing the performance of our models, these include: binary accuracy, precision, recall and AUC, F1-score, MCC. We also used ReLU activation as our hidden layer activation function and Sigmoid activation function as our output layer activation function for both MLP and CNN. We used Adam optimization algorithm as our optimizer and since it is a binary classification task, we also used binary cross entropy. We employed early stopping criteria on the optimization procedure when the values of the loss function on the validation set stopped decreasing.

Triplex DNA Sites Prediction Second Negative (Promoters) Dataset for TriplexFPM-HMM

The NN training does not go well - the loss does not decrease through the training epochs. One of the reasons to this problem may be the faulty negative data that we used from the original paper. To handle this, we decided to adapt the Zhang *et al*., 2020 approach, where they used random promoter regions as their negative data.

We took *36 597* promoter regions of DNA from *Ensembl Release 104 of May 2021* as our second negative data (Kevin, *et al*., 2021). Our promoter regions are of Ensembl Regulation 104 of the human regulatory features (GRCh38.p13). We dowloaded only three columns; chromosome number, start and end regions from Ensembl after which we used pandas to remove the header and toadd *chr* to every chromosome number and saved it as a bed file.

We took our *12* histone modification read counts from ENCODE database No. E123 cell line out of k562 cell lines (Inoue *et al*., 2017; Kazachenka *et al*., 2018), we converted the 'tagalign.gz'format to 'bam' by following the instruction given on DeepChrome (Ritambhara *et al*., 2016. We then added our 12 histone modification marks to our second negative data using bedtools multicov command.

Overall, we had *36 597 DNA promoter regions as our negative data* and *12 histone modification features* (H3K36me3, H3K27ac, H3K9me3, H3K4me3, H3K4me2, H3K9ac, H3K9me1, H3K4me1, H2AFZ, H3K27me3, H3K79me2, H4K20me1).

The length distribution between the negative and positive data are different. To match their lengths, we filtered out too long regions from negative data (> 2000 bp), we also dropped all regions from chromosome X (as it was done in Zhang *et al*. 2020). Additionally, we filtered out regions that overlapped with positive data regions. After filtering out sequences greater than 2000 Bps and regions with chromosome X from our negative data, we decreased the number of regions from *36 597 to 20 275.*

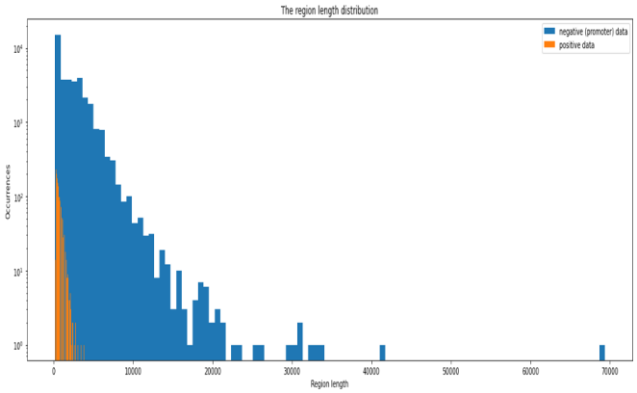

Fig 2.2: Triplex DNA site length distributions for positive & second negative (promoters) dataset for TriplexFPM-HMM

Model Construction for Second Negative (Promoters) Dataset TriplexFPM-HMM

The amount of our negative data is almost 8 times bigger than the amount of positive data. Again, to deal with this unbalanced data, we oversampled our positive data and undersampled our negative data. We fixed the final data size to be 8000. We splitted our fixed data into three sets; *train (0.8), validation (0.1) and test (0.1) sets.*

Hyperparameter Tuning for Second Negative (Promoters) Dataset TriplexFPM-HMM

We set other internal and external hyperparameters as follow: epochs (100), learning rate at (1e-3), input size (12, 1) Kernel size (32), verbose (1) and used Keras-Tensorflow to train our models. We used six different metrices in assessing the performance of our models, these include: binary accuracy, precision, recall and AUC, F1-score, MCC. We also used ReLU activation as our hidden layer activation function and Sigmoid activation function as our output layer activation function for both MLP and CNN. We used Adam optimization algorithm as our optimizer and since it is a binary classification task, we also used binary cross entropy. We used different kernel sizes with different learning rates since it increases the performance of the models. We employed early stopping criteria on the optimization procedure when the values of the loss function on the validation set stopped decreasing.

Results and discussion

In addition to sequence-based feature models published as "Predicting RNA:DNA Triplex Structures from Sequence Features Using Deep Learning Architectures", we developed four models incorporating histone modification marks as input features (TriplexFPM-HMM) to evaluate their predictive potential for RNA:DNA triplex formation. These models were trained using two sets of negative DNA site data: (i) generic negative data (DNA1) and (ii)

promoter regions (DNA2). The results (Table 1.1) indicate that while the initial performance of these models was relatively modest compared to the lncRNA/DNA sequence-based TriplexFPM, grid search optimization substantially improved their predictive accuracy.

Model	Negative Data Source	Feature Type	Mean AUC (Before Grid Search)	Mean AUC (After Grid Search)	Kernel Size	Learning Rate
DNA1_MLP	Negative Data 1	Histone Marks	0.58	0.59	32	1e-2
DNA1_CNN	Negative Data 1	Histone Marks	0.51	0.56	32	1e-2
DNA2_MLP	Negative Data 2 (Prom.)	Histone Marks	0.67	0.76	32	1e-3
DNA2_CNN	Negative Data 2 (Prom.)	Histone Marks	0.57	0.78	32	1e-3

Table 1.1: TriplexFPM-HMM (Histone Modification Mark–based models)

For the DNA1 set, **DNA1_MLP** initially achieved the highest mean AUC of **0.58** at a kernel size of 32 and learning rate of 1e-2, while **DNA1_CNN** performed lowest at **0.51**. Following hyperparameter tuning with grid search, performance improved slightly, with **DNA1_MLP** reaching **0.59** and **DNA1_CNN** improving to **0.56**. Although these gains were modest, they demonstrate the sensitivity of histone-based models to hyperparameter optimization.

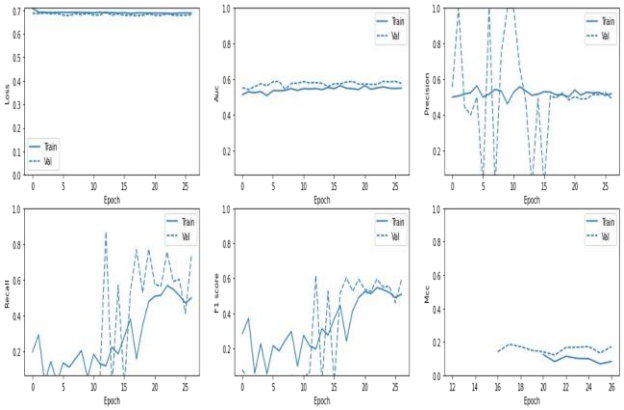

Fig 3.1: MLP model performance plot for first negative DNA site data for TriplexFPM-HMM

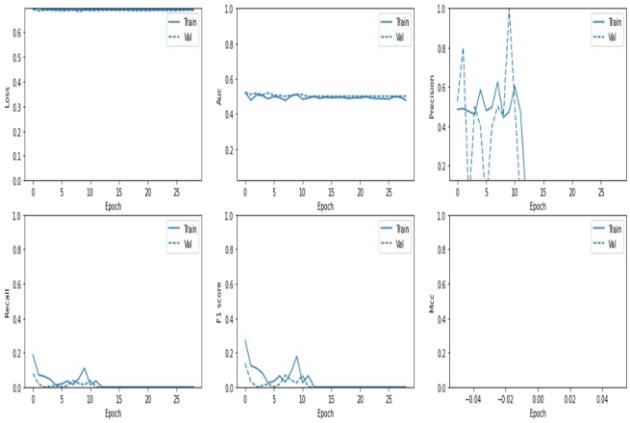

Fig 3.2: CNN model performance plot for first negative DNA site data for TriplexFPM-HMM

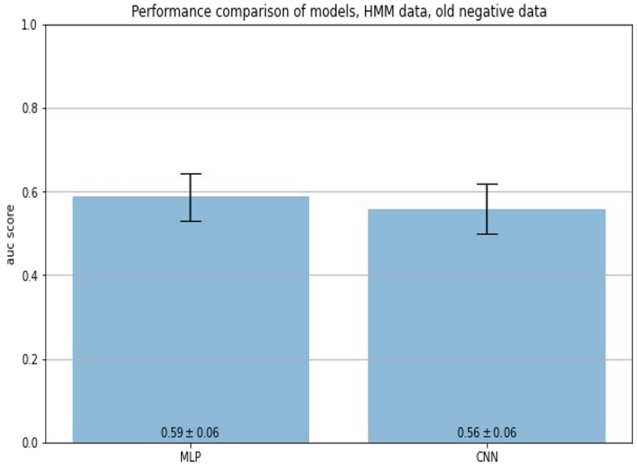

Fig 3.3: Performance comparison of MLP & CNN TriplexFPM-HMM on first negative DNA site data *after* grid search on mean AUC

For the DNA2 promoter set, performance was stronger overall. Before optimization, **DNA2_MLP** achieved a mean AUC of **0.67**, outperforming **DNA2_CNN** at **0.57**. After grid search optimization, both models showed substantial improvement: **DNA2_CNN** outperformed all histone-based models with a mean AUC of **0.78**, while **DNA2_MLP** increased to **0.76**. These findings indicate that promoter regions with histone modification marks carry stronger signals for predicting triplex formation compared to more generic DNA negative sets.

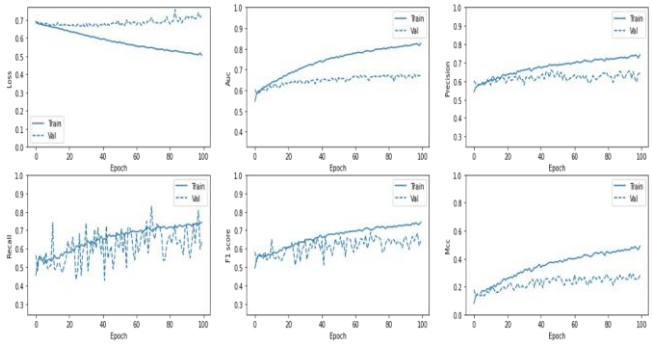

Fig 3.4: MLP model performance plot for second negative DNA site (promoters) data for TriplexFPM-HMM

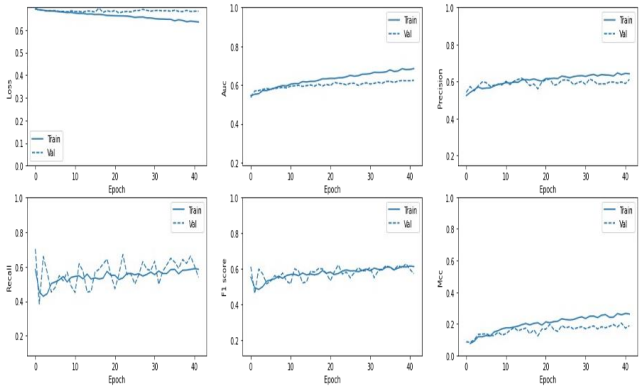

Fig 3.5: CNN model performance plot for second negative DNA site (promoters) data for TriplexFPM-HMM

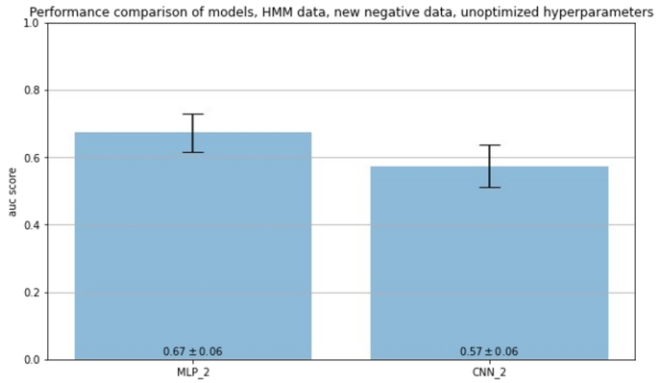

Fig 3.6: Performance comparison of MLP & CNN TriplexFPM-HMM on second negative DNA site (promoters) data *before* grid search on mean AUC

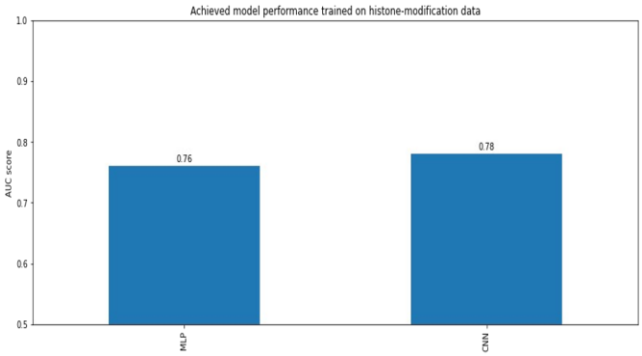

Fig 3.7: Performance comparison of MLP & CNN TriplexFPM-HMM on second negative DNA site (promoters) data *after* grid search on AUC

Collectively, these results suggest that while histone modification features alone may not achieve the same predictive power as lncRNA/DNA sequence-based models, they nonetheless provide complementary information that can enhance triplex prediction. In particular, the superior performance of promoter-based models aligns with previous findings that chromatin states, such as H3K4me3 and H3K36me3 marks, are associated with lncRNA transcriptional regulation and may predispose certain DNA regions to triplex formation.

Limitations. Despite the improvements obtained through optimization, the overall performance of TriplexFPM-HMM remained lower than TriplexFPM sequence-based models in our previous paper. This may be attributed to the complexity of chromatin signatures, the limited size of our negative datasets, and the fact that histone modification marks were used as the sole input features without integration of additional sequence-level information.

Future Directions. Future work should aim to integrate histone modification marks with lncRNA and DNA sequence features in multimodal architectures, potentially leveraging attention-based mechanisms to capture interactions across feature types. Moreover, larger and more diverse histone datasets are needed to increase model robustness. Comparing the performance of deep learning models with classical machine learning approaches such as support vector machines and random forests may also clarify whether simpler algorithms can capture the predictive signal in histone features. Finally, experimental validation remains essential to verify whether the histone modification marks identified by these models indeed correspond to biologically relevant triplex-forming sites.

Declarations

Ethics Approval and Consent to Participate

Not applicable

Consent for Publication

Not applicable

Availability of Data and Materials

The datasets and deep learning models used in this study are openly available on GitHub at https://github.com/Joseph-Luper-Tsenum/Using-Deep-Learning-with-Different-Architectures-to-Recognize-RNA-DNA-Triplex-Structures-from-HMMs/tree/main

Our previously published TriplexFPM sequence-based models can be found on GitHub: https://github.com/Joseph-Luper-Tsenum/Predicting-RNA-DNA-Triplex-Structures-from-Sequence-Features-Using-Deep-Learning-Architectures/tree/main

Conflicts of interest

The authors declare no conflict of interests for this article.

Funding

This research was funded by the federal budget within the quota of the Russian Government, 2019–2021.

Authors' Contribution

JLT designed the study, performed the analyses, and wrote the manuscript. MP supervised the study and contributed valuable feedback.

Acknowledgments

The authors thank the Open Doors Olympiad for providing the scholarship, the Russian Government for funding, Phystech School of Biological and Medical Physics (FBMP) at the Moscow Institute of Physics and Technology (MIPT) for their support, and the members of the International Laboratory of Bioinformatics at HSE University for their valuable assistance and guidance.

Abbreviations

- CNN = Convolutional neural network
- RNN = Recurrent neural network
- ResNN = Residual neural network
- LSTM = Long short-term memory
- MLP = Multilayer perceptron
- RNA = Ribonucleic acid
- DNA = Deoxyribonucleic acid
- lncRNA = long non-coding RNA

• TriplexFPM = Triplex Forming Prediction Model

• AUC = Area under the Receiver operating characteristic (ROC) Curve

• ReLU = Rectified Linear Unit

• lr = Learning rate

• MCC = Matthews correlation coefficient

• k-mer = A k-mer is a subsequence of length k derived from a longer nucleotide or amino acid sequence.

References

Ageeli AA, McGovern-Gooch KR, Kaminska MM, Baird NJ. Finely tuned conformationaldynamics regulate the protective function of the lncRNA MALAT1 triple helix. Nucl Acids Res. 2019;47(3):1468–81.

Aguilo F., Zhou M.M., Walsh M.J. Long noncoding RNA, polycomb, and the ghosts hauntingINK4b-ARF-INK4a expression. *Cancer Res.* 2011;71:5365–5369.

Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods. 2018;15:399–400.

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: A reference database for long noncoding RNAs. Nucleic Acids Res. 2011 Jan; 39(Database issue):D146-51.

Anderson, D.M.; Anderson, K.M.; Chang, C.L.; Makarewich, C.A.; Nelson, B.R.; McAnally, J.R.; Kasaragod, P.; Shelton, J.M.; Liou, J.; Bassel-Duby, R.; et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell 2015, 160, 595–606. [CrossRef]

Angrand, P.-O.; Vennin, C.; Le Bourhis, X.; Adriaenssens, E. The role of long non-coding RNAsin genome formatting and expression. Front. Genet. 2015, 6, 165. [CrossRef] [PubMed] 2

Antonov,I.V., Mazurov,E., Borodovsky,M. and Medvedeva,Y.A. (2018) Prediction of lncRNAs and their interactions with nucleic acids: benchmarking bioinformatics tools. Brief. Bioinformatics, 2018, 1–14.

Azhikina TL, Ignatov DV, Salina EG, Fursov MV, Kaprelyants AS (2015) Role of Small NoncodingRNAs in Bacterial Metabolism. Biochemistry 80: 1633-1646.

Bacolla A, Wang G, Vasquez KM (2015) New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity. PLoS Genet 11(12): e1005696.

Batista, P., and Chang, H. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell 152, 1298–1307.

Beltran M, Puig I, Peña C, García JM, Alvarez AB, Peña R, et al. A natural antisensetranscript regulates Zeb2/Sip1 gene expression during Snail1- induced epithelial-mesenchymal transition. *Genes Dev.* 2008;22:756–69. doi: 10.1101/gad.455708. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support Vector Machinesand Kernels for Computational Biology. PLoS Comput Biol 4(10): e1000173. https://doi.org/10.1371/journal.pcbi.1000173

Bernstein HD, Zopf D, Freymann DM, Walter P. Functional substitution of the signal recognition particle 54-kDa subunit by its Escherichia coli homolog. Proc Natl Acad Sci U S A. 1993 Jun 1; 90(11):5229-33.

Bartolomei MS, Zemel S, Tilghman SM. Parental imprinting of the mouse H19 gene. *Nature.* 1991;351:153–5. doi: 10.1038/351153a0. [PubMed] [CrossRef] [Google Scholar]

Birney E, Stamatoyannopoulos JA, Dutta A, *et al*; ENCODE Project Consortium; NISCComparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute; Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. Nature. 2007;447:799–816.

Blume, S.W.; Meng, Z.; Shrestha, K.; Snyder, R.C.; Emanuel, P.D. The 5 0 - untranslated RNA of the human dhfr minor transcript alters transcription pre- initiation complex assembly at themajor (core) promoter. J. Cell. Biochem. 2003, 88, 165–180. [CrossRef] [PubMed]

Bo Chang, Lili Meng, Eldad Haber, Frederick Tung, and David Begert. Multi- level Residual Networks From Dynamical Systems View. ICLR, pp. 1–14, 2018.

Bonasio, R., and Shiekhattar, R. (2014). Regulation of transcription by long noncoding RNAs. Annu. Rev. Genet. 48, 1–23.

Breiman L. Random Forests, *Mach Learn*, 2001, vol. 45 (pg.5-32) Google Scholar Crossref

Brockdorff N, Ashworth A, Kay GF, Cooper P, Smith S, McCabe VM, et al. Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature.*1991;351:329–31. doi: 10.1038/351329a0. [PubMed] [CrossRef] [Google Scholar]

Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J, Willard HF. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and ishighly localized within the nucleus. Cell. 1992 Oct 30; 71(3):527-42.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature.* 1991;349:38–44. doi:

10.1038/349038a0. [PubMed] [CrossRef] [Google Scholar]

Buske FA, Bauer DC, Mattick JS, Bailey TL. Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci. *Bioinformatics*. 2013;29(15):1895-1897.
Buske,F.A., Bauer,D.C., Mattick,J.S. and Bailey,T.L. (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. Genome Res., 22, 1372–1381.

Campalans A, Kondorosi A, Crespi M. Enod40, a short open reading frame- containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. Plant Cell. 2004 Apr;16(4):1047-59. doi: 10.1105/tpc.019406. Epub 2004 Mar 22. Erratum in: Plant Cell. 2017 Apr;29(4):912. PMID: 15037734; PMCID: PMC412876.

Carl Kingsford, Steven L Salzberg. Nat Biotechnol. Author manuscript; available in PMC 2009 Jun 24. Published in final edited form as: Nat Biotechnol. 2008 Sep; 26(9): 1011–1013. doi:10.1038/nbt0908-1011. PMCID: PMC2701298

Cech, T., and Steitz, J. (2014). The noncoding RNA revolution—trashing old rules to forgenew ones. Cell 157, 77–94.
Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 2015;16:30. doi: 10.1186/s13059-015-0596-2.

Charles Richard J.L., Ogawa Y. (2015) Understanding the Complex Circuitry of lncRNAs atthe X-inactivation Center and Its Implications in Disease

Conditions. In: Morris K. (eds) Long Non-coding RNAs in Human Disease. Current Topics in Microbiology and Immunology, vol 394. Springer, Cham. https://doi.org/10.1007/82_2015_443.

Chu C, Quinn J, Chang HY. Chromatin isolation by RNA purification (ChIRP). J Vis Exp.2012;61. pii: 3912. doi: 10.3791/3912.

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol Cell. 2011;44:667–678. doi: 10.1016/j.molcel.2011.08.027.

Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY. Systematic discovery of Xist RNA binding proteins. Cell. 2015;161:404–416. doi: 10.1016/j.cell.2015.03.025.10. Costa FF. Non-coding RNAs: new players in eukaryotic biology.Gene. 2005;357:83–94. doi: 10.1016/j.gene.2005.06.019.

Chugunova, A.; Navalayeu, T.; Dontsova, O.; Sergiev, P. Mining for small translated ORFs. J. Proteome Res. 2018, 17, 1–11. [CrossRef]
Clark MB, Mattick JS. Long noncoding RNAs in cell biology. *Semin Cell Dev Biol*. 2011;22:366–76. doi: 10.1016/j.semcdb.2011.01.001. [PubMed] [CrossRef] [Google Scholar]

Clemson CM, McNeil JA, Willard HF, Lawrence JB. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. J Cell Biol. 1996 Feb; 132(3):259-75.

Deng, Z.; Norseen, J.; Wiedmer, A.; Riethman, H.; Lieberman, P.M. TERRA RNA binding toTRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. Mol. Cell. 2009, 35, 403–413. [CrossRef] [PubMed]

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, MerkelA, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB,Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ,Notredame C, Harrow J, Guigó R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012 Sep; 22(9):1775- 89.

Diederik P. Kingma and Jimmy Ba Adam: A Method for Stochastic Optimization. 3rd International Conference for Learning Representations, San Diego. (2015). https://arxiv.org/abs/1412.6980

Eldad Haber and Lars Ruthotto. Stable Architectures for Deep Neural Networks. ArXiv, 2017.
ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia of DNA elements) project. Science 306, 636–640.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, Plath K. The Xist lncRNA exploits three- dimensional genome architecture to spread across the X chromosome. Science. 2013;341(6147):1237973.

Espinoza CA, Allen TA, Hieb AR, Kugel JF, Goodrich JA. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat Struct Mol Biol*. 2004;11:822–9. doi:10.1038/nsmb812. [PubMed] [CrossRef] [Google Scholar]

Fang S., Zhang L., Guo J., Niu Y., Wu Y., Li H., Zhao L., Li X., Teng X., Sun X., et al.NONCODEV5: A comprehensive annotation database for long non-coding RNAs. *Nucleic AcidsRes*. 2018;46:D308–D314. doi: 10.1093/nar/gkx1107.

Fan Wang, Pranik Chainani, Tommy White, Jin Yang, Yu Liu & Benjamin Soibam (2018)Deep learning identifies genome-wide DNA binding sites of long noncoding RNAs, RNA Biology, 15:12, 1468-1476, DOI: 10.1080/15476286.2018.1551704.

Kopp F (2019) Molecular functions and biological roles of long non-coding RNAs in humanphysiology and disease. J Gene Med 21(8):e3104

Fullwood MJ, Wei CL, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res. 2009;19:521–532. doi: 10.1101/gr.074906.107.

Gaddis,S.S., Wu,Q., Thames,H.D., DiGiovanni,J., Walborg,E.F., MacLeod,M.C. and Vasquez,K.M. (2006) A web-based search engine for triplex-forming oligonucleotide target sequences. Oligonucleotides, 16, 196–201

Gardini A. Global run-on sequencing (GRO-Seq). Methods Mol Biol. 2017;1468:111–120. doi: 10.1007/978-1-4939-4035-6_9.

Garzia, A.; Meyer, C.; Morozov, P.; Sajek, M.; Tuschl, T. Optimization of PAR- CLIP for transcriptome-wide identification of binding sites of RNA-binding proteins. Methods 2017, 118–119, 24–40. [CrossRef]

Ginger MR, Shore AN, Contreras A, Rijnkels M, Miller J, Gonzalez-Rimbau MF, et al. A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc Natl Acad Sci U S A*. 2006;103:5781–6. doi: 10.1073/pnas.0600745103. [PMC free article] [PubMed][CrossRef] [Google Scholar]

Goñi,J.R., de la Cruz,X. and Orozco,M. (2004) Triplex-forming oligonucleotide target sequences in the human genome. Nucleic Acids Res., 32, 354–360.

Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644– 652. doi: 10.1038/nbt.1883.

Grote, P.; Herrmann, B.G. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. RNA Boil. 2013, 10, 1579–1585. [CrossRef] [PubMed]

Grote P., Wittler L., Hendrix D., Koch F., Wahrisch S., Beisaw A. The tissue- specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell*.2013;24:206–214.

Guil, S., and Esteller, M. (2012). Cis-acting noncoding RNAs: friends and foes. Nat. Struct. Mol. Biol. 19, 1068–1075.

Guttman, M., and Rinn, J. (2012). Modular regulatory principles of large non- coding RNAs. Nature 482, 339–346.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O.,Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227.

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*. 2010;142:409–19. doi: 10.1016/j.cell.2010.06.040. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

He,S., Zhang,H., Liu,H. and Zhu,H. (2014) LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. Bioinformatics, 31, 178–186.

Hirose T, Virnicchi G, Tanigawa A, Naganuma T, Li R, Kimura H, et al. NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. Mol Biol Cell 2014;25:169–83.
Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, Cunningham F. Ensembl variation resources. Database. 2018;1:2018.

Hombach S., Kretz M. (2016) Non-coding RNAs: Classification, Biology and Functioning. In: Slaby O., Calin G. (eds) Non-coding RNAs in Colorectal Cancer. Advances in Experimental Medicine and Biology, vol 937. Springer, Cham. *https://doi.org/10.1007/978-3-319-42059-2_1*.

Ingolia N.T., Lareau L.F., Weissman J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*. 2011;147:789–802.
Inoue F, Kircher M, Martin B, et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res*. 2017;27(1):38-52. PMID: 27831498; PMCID: PMC5204343

Jalali, S.; Gandhi, S.; Scaria, V.S. Distinct and modular organization of protein interactingsites in long non-coding RNAs. Front. Mol. Biosci. 2018, 5,
27. [CrossRef]

Jeon Y, Lee JT. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*. 2011;**146**(1):119–133. doi: 10.1016/j.cell.2011.06.026. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
Jessica A. Brown, Max L. Valenstein, Therese A. Yario, Kazimierz T. Tycowski, Joan A. Steitz. MALAT1 and MENβ RNA 3′ ends form triple helices Proceedings of the National Academy of Sciences. (2012)109 (47) 19202-19207; DOI: 10.1073/ pnas.1217338109

Jianyin Long, … , Paul A. Overbeek, Farhad R. Danesh. The long noncoding RNA *Tug1* connects metabolic changes with kidney disease in podocytes. *J Clin Invest.* 2016;126(11):4205–4218.doi:10.1172/JCI87927.

Kaikkonen MU, Lam MT, Glass CK. Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Res. 2011;90:430–440. doi: 10.1093/cvr/cvr097.

Kallen A.N., Zhou X.B., Xu J., Qiao C., Ma J., Yan L. The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell.* 2013;52:101–112.

Kalwa M, Hänzelmann S, Otto S, Kuo CC, Franzen J, Joussen S, Fernandez- Rebollo E, Rath B, Koch C, Hofmann A, Lee SH. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. Nucl Acids Res. 2016;44(22):10631–43.

Kapranov, P., Willingham, A.T., and Gingeras, T.R. (2007). Genome-wide transcription and the implications for genomic organization. Nat. Rev. Genet. 8, 413–423.

Karapetyan, A.E.; Coen Buiting, C.; Kuiper, R.A.; Marcel, W.; Coolen, M.W. Regulatory rolesfor long ncRNA and mRNA. Cancers 2013, 5, 462–490. [CrossRef] [PubMed]

Katayama S, Tomaru Y, Kasukawa T, *et al*; RIKEN Genome Exploration Research Group;Genome Science Group (Genome Network Project Core Group); FANTOM Consortium. Antisense transcription in the mammalian transcriptome.Science. 2005;309:1564–1566. doi: 10.1126/science.1112009.

Kalwa M, Hänzelmann S, Otto S, et al. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res*. 2016;44(22):10631-10643. doi:10.1093/nar/gkw802

Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, *et al*. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*.2018;175(5):1259-1271.e13. PMID: 30454646; PMCID: PMC6242299

Kazimierczyk M, Kasprowicz MK, Kasprzyk ME, Wrzesinski J. Human Long Noncoding RNA Interactome: Detection, Characterization and Function. Int J Mol Sci. 2020 Feb 4;21(3):1027. doi:10.3390/ijms21031027. PMID: 32033158; PMCID: PMC7037361.

Ke H, Zhao L, Feng X, Xu H, Zou L, Yang Q, et al. NEAT1 is required for survival of breast cancer cells through FUS and miR-548. Gene Regul Syst Bio 2016;10:11–7.

Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson,Kamalkumar Dodiya, Bilal El Houdaigui, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Tiago Grego, Cristina Guijarro-Clarke, Leanne Haggerty, Anmol Hemrom, Thibaut Hourlier, Osagie G Izuogu, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, José Carlos Marugán, Thomas Maurel, Aoife C McMahon, Shamika Mohanan, Benjamin Moore, Matthieu Muffato, Denye N Oheh, Dimitrios Paraschas, Anne Parker, Andrew Parton, Irina Prosovetskaia, Manoj P Sakthivel, Ahamed I Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Emily Steed, Michal Szpak, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Brandon Walts, Andrea Winterbottom, Marc Chakiachvili, AmeyaChaubal, Nishadi De Silva, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth R IIsley, Nick Langridge, Jane E Loveland, Fergal JMartin, Jonathan M Mudge, Joanella Morales, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Fiona Cunningham, Andrew D Yates, Daniel R Zerbino, Paul Flicek Ensembl 2021. Nucleic Acids Res. 2021, vol. 49(1):884–891. PubMed PMID: 33137190.

Kloc M, Wilk K, Vargas D, Shirato Y, Bilinski S, Etkin LD. Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of Xenopus oocytes. *Development*. 2005;132:3445–57. doi: 10.1242/dev.01919. [PubMed] [CrossRef] [Google Scholar]

Kotake Y., Nakagawa T., Kitagawa K., Suzuki S., Liu N., Kitagawa M. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene.*Oncogene.* 2011;30:1956–1962.

Kowalczyk, M.S., Higgs, D.R., and Gingeras, T.R. (2012). Molecular biology: RNA discrimination. Nature 482, 310–311.

Krol, J.; Loedige, I.; Filipowicz, W. The widespread regulation of microRNA biogenesis,function and decay. Nat. Rev. Genet. 2010, 11, 597–610. [CrossRef].
Kung, J., Colognori, D., and Lee, J. (2013). Long noncoding RNAs: past, present, andfuture. Genetics 193, 651–669.

Kuo CC, Hänzelmann S, Sentürk Cetin N, Frank S, Zajzon B, Derks JP, Akhade VS, Ahuja G, Kanduri C, Grummt I, Kurian L. Detection of RNA–DNA binding sites in long noncoding RNAs. NuclAcids Res. 2019;47(6):e32.

Lars Ruthotto and Eldad Haber. Deep Neural Networks motivated by Partial Differential Equations. arXiv, pp. 1–7, 2018.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44. [PubMed] [CrossRef] [Reference list]

Lee, J. (2012). Epigenetic regulation by long noncoding

RNAs. Science 338, 1435–1439.

Li, J.; Ma, W.; Zeng, P.; Wang, J.; Geng, B.; Yang, J.; Cui, Q. LncTar: A tool for predicting the RNA targets of long noncoding RNAs. Brief. Bioinform. 2015. [CrossRef]

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25:1966–1967. doi: 10.1093/bioinformatics/btp336.

Li, T., Mo, X., Fu, L., Xiao, B., and Guo, J. (2016). Molecular mechanisms of longnoncoding RNAs on gastric cancer. Oncotarget 7, 8601–8612.

Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X, Oh S, Kim HS, Glass CK, Rosenfeld MG. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature. 2013;498:516–520. doi: 10.1038/nature12210

Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recentinsights and future perspectives. Nat Rev Genet. 2016;17:207–
223. doi: 10.1038/nrg.2016.4.

Li X.L., Wu Z.Q., Fu X.B., Han W.D. LncRNAs: insights into their function and mechanics in underlying disorders. *Mutat Res Rev Mut Res.* 2014;762:1–21.

Li, Y.; Syed, J.; Sugiyama, H. RNA-DNA triplex formation by longnoncoding RNAs. Cell Chem. Biol. 2016, 23, 1325–1333. [CrossRef]
Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet.* 2010;42:1113–7. doi: 10.1038/ng.710. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Maass PG, Luft FC, Bähring S. Long non-coding RNA in health and disease. J Mol Med(Berl). 2014;92:337–346. doi: 10.1007/s00109-014-1131-8.

Marco Ciccone, Marco Gallieri, Jonathan Masci, Christian Osendorfer, and Faustino Gomez. NAISNET: Stable Deep Networks from Non-Autonomous Differential Equations. arXiv, 2018.

Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature.* 2007;445:666–70. doi: 10.1038/nature05519. [PubMed] [CrossRef] [Google Scholar]

Martorell-Marugán J, Tabik S, B4. enhammou Y, et al. Deep Learning in Omics Data Analysis and Precision Medicine. In: Husi H, editor. Computational Biology [Internet]. Brisbane (AU): Codon Publications; 2019 Nov 21. Chapter 3. Available from: https://www.ncbi.nlm.nih.gov/books/NBK550335/ doi: 10.15586/ computationalbiology.2019.ch3

Matsumoto, A.; Nakayama, K.I. Hidden peptides encoded by putative noncoding RNAs. Cell Struct. Funct. 2018, 43, 75–83. [CrossRef] [PubMed]

Matsumoto, A.; Pasut, A.; Matsumoto, M.; Yamashita, R.; Fung, J.; Monteleone, E.; Saghatelian, A.; Nakayama, K.I.; Clohessy, J.G.; Pandolfi, P.P. mTORC1 and muscle regenerationare regulated by the LINC00961-encoded SPAR polypeptide. Nature 2017, 541, 228–232. [CrossRef]
Mattick JS. The central role of RNA in human development and cognition. FEBS Lett. 2011;585:1600–1616. doi: 10.1016/j.febslet.2011.05.001.

McKenna SJ, Amaral T, Akbar S, Jordan L, Thompson A. Immunohistochemical analysisof breast tissue microarray images using contextual classifiers. *J Pathol Inform.*2013;4(Suppl):S13. Published 2013 Mar
30. doi:10.4103/2153-3539.109871

Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol. 2013;20(3):300.

Misawa A, Takayama KI, Inoue S. Long non-coding RNAs and prostate cancer. CancerScience. 2017 Nov;108(11):2107-2114. DOI: 10.1111/cas.13352.

Mishra K, Kanduri C. Understanding Long Noncoding RNA and Chromatin Interactions: WhatWe Know So Far. *Noncoding RNA*. 2019;5(4):54. Published 2019 Dec 3. doi:10.3390/ncrna5040054

Mondal T, Subhash S, Vaid R, Enroth S, Uday S, Reinius B, Mitra S, Mohammed A, James AR, Hoberg E, Moustakas A. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. Nat Commun. 2015;24(6):7743.

Moran V.A., Perera R.J., Khalil A.M. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res.* 2012;40:6391–6400.

Morris, K., and Mattick, J. (2014). The rise of regulatory RNA. Nat. Rev. Genet. 15, 423–437.

Mourtada-Maarabouni M, Hedge VL, Kirkham L, Farzaneh F, Williams GT. Growth arrest in human T-cells is controlled by the non-coding RNA growth- arrest-specific transcript 5 (GAS5) *J Cell Sci.* 2008;121:939–46. doi: 10.1242/jcs.024646. [PubMed] [CrossRef] [Google Scholar]

Muddashetty R, Khanam T, Kondrashov A, Bundman M, Iacoangeli A, Kremerskothen J, et al. Poly(A)-binding protein is associated with neuronal BC1 and BC200 ribonucleoprotein particles. *J Mol Biol.* 2002;321:433–45. doi:

10.1016/S0022-2836(02)00655-1. [PubMed] [CrossRef] [GoogleScholar]

Navarro C, Cano C, Cuadros M, Herrera-Merchan A, Molina M, Blanco A. A mechanistic study of lncRNA Fendrr regulation of FoxF1 lung cancer tumor supressor. In: International conference on bioinformatics and biomedical engineering 2016 Apr 20 (pp. 781–789). Springer,Cham.

Nelson, B.R.; Makarewich, C.A.; Anderson, D.M.; Winders, B.R.; Troupes, C.D.; Wu, F.;Reese, A.L.; McAnally, J.R.; Chen, X.; Kavalali, E.T.; et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science 2016, 351, 271–275. [CrossRef].

O'Leary VB, Smida J, Buske FA, Carrascosa LG, Azimzadeh O, Maugg D, Hain S, Tapio S, Heidenreich W, Kerr J, Trau M, Ovsepian SV, Atkinson MJ. PARTICLE triplexes cluster in the tumor suppressor WWOX and may extend throughout the human genome. Sci Rep. 2017 Aug 2;7(1):7163. doi: 10.1038/s41598-017-07295-5. PMID: 28769061; PMCID: PMC5541130.

O'Leary VB, Ovsepian SV, Carrascosa LG, Buske FA, Radulovic V, Niyazi M, Moertl S, Trau M, Atkinson MJ, Anastasov N. PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation. Cell Rep. 2015;11(3):474–85.

Orozco-Arias S, Piña JS, Tabares-Soto R, Castillo-Ossa LF, Guyot R, Isaza G. Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements. *Processes*. 2020; 8(6):638. https://doi.org/10.3390/pr8060638

Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. Direct RNA sequencing. Nature. 2009;461:814–818. doi: 10.1038/nature08390.

Pandey R.R., Mondal T., Mohammad F., Enroth S., Redrup L., Komorowski J. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell*. 2008;32:232–246.

Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 2006 Jan; 22(1):1- 5.

Park J, Yi D, Ji S. A Novel Learning Rate Schedule in Optimization for Neural Networks and It's Convergence. *Symmetry*. 2020; 12(4):660. https://doi.org/10.3390/sym12040660.

Philippe N, Bou Samra E, Boureux A, Mancheron A, Rufflé F, Bai Q, De Vos J, Rivals E,Commes T. Combining DGE and RNA-sequencing data to identify new polyA+ non-coding transcripts in the human genome. Nucleic Acids Res. 2014;42:2820–2832. doi:10.1093/nar/gkt1300.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. Cell 136, 629–641.

Postepska-Igielska A, Giwojna A, Gasri-Plotnitsky L, Schmitt N, Dold A, Ginsberg D, et al. LncRNA Khps1 regulates expression of the proto-oncogene SPHK1 via triplex-mediated changes in chromatin structure. *Mol Cell*. 2015;**60**:626–636. doi: 10.1016/j.molcel.2015.10.001. [PubMed] [CrossRef] [Google Scholar]

Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillaume Carlier. Deep Relaxation: partial differential equations for optimizing deep neural networks. Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR, 2017. URL http://arxiv.org/abs/1704.04932.

Qianli Liao and Tomaso Poggio. Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. arXiv, (047):1–16, 2016.

Quinn, J., and Chang, H. (2016). Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. 17, 47–62.

Quinodoz, S.; Guttman, M. Longnoncoding RNAs: An emerging link between gene regulation and nuclear organization. Cell Chem. Biol. 2014, 24, 652–663.

Reeves MB, Davies AA, McSharry BP, Wilkinson GW, Sinclair JH. Complex I binding by a virally encoded RNA regulates mitochondria-induced cell death. Science. 2007 Jun 1; 316(5829):1345-8.

Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem.2012;81:145–166. doi: 10.1146/annurev-biochem-051410-092902.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007;129:1311–1323. doi: 10.1016/j.cell.2007.05.022. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Rintala-Maki ND, Sutherland LC. Identification and characterisation of a novel antisensenon-coding RNA from the RBM5 gene locus. *Gene*. 2009;445:7–16. doi: 10.1016/j.gene.2009.06.009. [PubMed] [CrossRef] [Google Scholar]

Ritambhara Singh, Jack Lanchantin, Gabriel Robins, Yanjun Qi, DeepChrome: deep-learning for predicting gene expression from histone modifications, *Bioinformatics*, Volume 32, Issue 17, 1 September 2016, Pages i639–i648.

Romero-Barrios, N.; Legascue, M.F.; Benhamed, M.; Federico Ariel, F.; Crespi, M. Splicingregulation by long

non-coding RNAs. Nucleic Acids Res. 2018, 46, 2169–2184. [CrossRef] [PubMed]

Ruiz-Orera, J.; Verdaguer-Grau, P.; Villanueva-Canas, J.L.; Messegue, X.; Alba, M.M.Functional and non-functional classes of peptides produced by long non-coding RNAs. biorXiv2016. [CrossRef]

Sallam T, Jones MC, Gilliland T, Zhang L, Wu X, Eskin A, Sandhu J, Casero D, Vallim TQ, Hong C, Katz M, Lee R, Whitelegge J, Tontonoz P. Feedback modulation of cholesterolmetabolism by the lipid-responsive non-coding RNA LeXis. Nature. 2016;534:124–128. doi:10.1038/nature17674.

Salmena, L.; Poliseno, L.; Tay, Y.; Kats, L.; Pandolfi, P.P. A ceRNA hypothesis: The Rosettastone of a hidden RNA language? Cell 2011, 146, 353–
358. [CrossRef]

Sentürk Cetin N, Kuo CC, Ribarska T, Li R, Costa IG, Grummt I. Isolation and genome-wide characterization of cellular DNA: RNA triplex structures. Nucl Acids Res. 2019;47(5):2306–21.

Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA- seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–1092. doi: 10.1093/bioinformatics/bts094.

Shamovsky I, Ivannikov M, Kandel ES, Gershon D, Nudler E. RNA-mediated response to heat shock in mammalian cells. *Nature.* 2006;440:556–60. doi: 10.1038/nature04518. [PubMed] [CrossRef] [Google Scholar]

Shaw, N.N.; Arya, D.P. Recognition of the unique structure of DNA: RNA hybrids. Biochimie2008, 90, 1026–1039. [CrossRef] [PubMed]
Sheik Mohamed J, Gaughwin PM, Lim B, Robson P, Lipovich L. Conserved longnoncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate

pluripotency in mouse embryonic stem cells. *RNA.* 2010;16:324–37. doi: 10.1261/rna.1441510. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Shen, Z., Bao, W. & Huang, DS. Recurrent Neural Network for Predicting Transcription Factor Binding Sites. *Sci Rep* **8,** 15270 (2018). https://doi.org/10.1038/s41598-018-33321-1
Smillie, C.L.; Sirey, T.; Ponting, C.P. Complexities of post-transcriptional regulation and the modeling of ceRNA crosstalk. Crit. Rev. Biochem. Mol. Biol. 2018, 53, 231–245. [CrossRef].

St Laurent G, Wahlestedt C, Kapranov P. The landscape of long noncoding RNA classification. Trends Genet. 2015;31:239–251. doi: 10.1016/j.tig.2015.03.007.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. 14, 103–105.

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* 2009;19:347–59. doi: 10.1101/gr.087775.108. [PMC free article] [PubMed] [CrossRef] [Google Scholar].

Sun, X.; Haider Ali, M.S.S.; Moran, M. The role of interactions of long non- coding RNAsand heterogeneous nuclear ribonucleoproteins in regulating cellular functions. Biochem. J.2017, 474, 2925–2935. [CrossRef]

Swiezewski S, Liu F, Magusin A, Dean C. Cold-induced silencing by long antisensetranscripts of an Arabidopsis Polycomb target. Nature. 2009 Dec 10; 462(7274):799-802.

Tafer, H.; Hofacker, I.L. RNAplex: A fast tool for RNA-RNA interaction search. Bioinformatics 2008, 24, 2657–2663. [CrossRef]
Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non-coding RNAs: regulators ofdisease. J Pathol. 2010;220:126–39. doi: 10.1002/path.2638.

Takahashi H, Kato S, Murata M, Carninci P. CAGE (cap analysis of gene expression): aprotocol for the detection of promoter and transcriptional networks. Methods Mol Biol. 2012;786:181–200. doi: 10.1007/978-1-61779- 292-2_11.
Tano K, Akimitsu N. Long non-coding RNAs in cancer progression. *Front Genet.* 2012;3:219.doi: 10.3389/fgene.2012.00219. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
Tay, Y.; Rinn, J.; Pandolfi, P.P. The multilayered complexity of ceRNA crosstalk andcompetition. Nature 2014, 505, 344–352. [CrossRef]

Tripathi, V.; Ellis, J.D.; Shen, Z.; Song, D.Y.; Pan, Q.; Watt, A.T.; Freier, S.M.; Bennett, C.F.; Sharma, A.; Bubulya, P.A.; et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell 2010, 39, 925–938. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J, Wong GK. Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature. 2004 Oct 14;431(7010):1 p following 757; discussion following 757.
Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncod- ing RNAs. Mol. Cell 43, 904–914. [CrossRef]

Wang, Q., Ma, Y., Zhao, K. *et al.* A Comprehensive Survey of Loss Functions in Machine Learning. *Ann. Data. Sci.* (2020). https://doi.org/10.1007/s40745- 020-00253-5

Wang S, Ke H, Zhang H, Ma Y, Ao L, Zou L, Yang Q, Zhu H, Nie J, Wu C, Jiao B. LncRNA MIR100HG promotes cell proliferation in triple-negative breast cancer through triplex formation with p27 loci. Cell Death Dis. 2018;9(8):1–1. https://doi.org/10.1038/s41419-018-0869-2

Wapinski O, Chang HY. Long noncoding RNAs and human disease. *Trends Cell Biol.* 2011;21:354–61. doi: 10.1016/j.tcb.2011.04.001. [PubMed] [CrossRef] [Google Scholar]

Waseem Rawat, Zenghui Wang; Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput* 2017; 29 (9): 2352–2449. doi: https://doi.org/10.1162/neco_a_00990

Watanabe Y, Yamamoto MS. S. pombe mei2+ encodes an RNA-binding protein essential for premeiotic DNA synthesis and meiosis I, which cooperates with a novel RNA species meiRNA. *Cell.* 1994;78:487–98. doi: 10.1016/0092- 8674(94)90426-X. [PubMed] [CrossRef] [Google Scholar]

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 23, 1494–1504.

Wouter G. Touw, Jumamurat R. Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, Sacha A. F. T. van Hijum, Data mining in the Life Sciences with Random Forest: awalk in the park or lost in the jungle?, *Briefings in Bioinformatics*, Volume 14, Issue 3, May 2013, Pages 315–326, https://doi.org/10.1093/bib/bbs034

Xu, C., Jackson, S.A. Machine learning and complex biological data.

Genome Biol **20,** 76 (2019). https://doi.org/10.1186/s13059-019-1689-0

Yang L., Froberg J.E., Lee J.T. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem Sci*. 2014;39:35–43.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, Massachusetts General Hospital, and Bin Dong.Beyond Finite Layer Neural Networks: Bridging Deep Architectures and Numerical Differential Equations. arxiv, pp. 1–15, 2017.

Yoon, J.H.; Gorospe, M. Cross-Linking Immunoprecipitation and qPCR (CLIP- qPCR) Analysis to map interactions between longnoncoding RNAs and RNA- binding proteins. Methods Mol. Biol. 2016, 1402, 11–17.

Yongqing Zhang, Jianrong Yan, Siyu Chen, Meiqin Gong, Dongrui Gao, Min Zhu* and Wei Gan, "Review of the Applications of Deep Learning in Bioinformatics", Current Bioinformatics 2020; 15(8) . https://doi.org/10.2174/1574893615999200711165743

Yue Li, Junetha Syed, Hiroshi Sugiyama. RNA-DNA Triplex Formation by Long NoncodingRNAs. *Cell Chemical Biology* (2016), 23 (11): 1325-1333

Zhang He, Michael J. Zeitz, Hong Wang, Beibei Niu, Shengfang Ge, Wei Li, Jiuwei Cui, Guanjun Wang, Guanxiang Qian, Michael J. Higgins, Xianqun Fan, Andrew R. Hoffman, Ji-Fan Hu. Long noncoding RNA-mediated intrachromosomal interactions promote imprinting at the *Kcnq1* locus. *J Cell Biol* (2014)204 (1): 61–75. https://doi.org/10.1083/jcb.201304152

Zhang Y, Jia C, Fullwood MJ, Kwoh CK. DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. Brief Bioinform. 2020.

Zhang Y, Jia C, Kwoh CK. Predicting the interaction biomolecule types for lncRNA: an ensemble deep learning approach. Brief Bioinform. 2020.

Zhang, Y., Long, Y. & Kwoh, C.K. Deep learning based DNA:RNA triplex forming potential prediction. BMC Bioinformatics 21, 522 (2020).

Zhao J, et al. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science.* 2008;**322**(5902):750–756. doi: 10.1126/science.1163045. [PMC freearticle] [PubMed] [CrossRef] [Google Scholar]

Zhu, S.; Wang, J.; He, Y.; Meng, N.; Yan, G.R. Peptides/Proteins encoded by non-coding RNA: A novel resource bank for drug targets and biomarkers. Front. Pharmacol. 2018, 9, 1295.[CrossRef]

Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf Fusion. 2019;50:71–91.