On the prediction of mRNA subcellular localization with machine learning

Zichao Yan

Master of Science

School of Computer Science

McGill University

Montreal,Quebec

2019-06-18

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters of Science, Computer Science

ACKNOWLEDGEMENTS

I would like to give my special thanks to Prof. Mathieu Blanchette, for under the supervision of whom I was firstly introduced to Bioinformatics, a field that I had utterly no experience of initially. His everlasting encouragement and trust on my capability have laid the foundation of my confidence of conducting research in this new direction.

I would also like to thank Faizy Ashan, Christopher JF Cameron, Julie Prost and many other lab mates in the Center of Bioinformatics at McGill University for their genuine friendship.

I also cannot thank enough to Prof. Eric Lécuyer and his team of brilliant biologists from the Montreal Clinical Research Institute for consistently sharing with us their quality experimental data, as well as their knowledge on biological concepts.

In the end, I would like to thank my parents and my friends for their consistent care and support. I would also like to thank Violet Wei, for making me feel special, and for without the transient companionship of whom I might have trodden a very different path.

ABSTRACT

Cells are the basic units of life, and yet they are regulated by many delicate and to some extent, fragile, subcellular processes that are crucial to their survival. A simple genetic mutation could possibly clog up some important regulatory processes, or perturb the function of the product it encodes, which might ultimately bring the demise of the entire system. Therefore, it is important to gain more insights into the many control processes of cell and the regulatory factors associated with them, one prominent example of which would be the mechanism related to the RNA subcellular localization that we would focus on almost exclusively in this study from a computational perspective.

RNA subcellular localization mechanism is one of the most important, yet underappreciated, facets of the broader gene regulatory process, which helps with the cellular organization and regulation on gene expression, via transporting the RNA transcripts to their designated locations where their function, structure or translated proteins are needed. It is generally accepted as a fact that RNA trafficking mechanism is mediated between the trans-regulatory factors such as the RNA binding proteins, and the cis-acting elements — short snippets of the transcript that contain the RBP binding sites — which we call zipcode as they are considered to contain information on its address of delivery.

The release of new RNA subcellular localization dataset has enabled us to build the first computational tool using state-of-the-art deep learning techniques, to predict the localization outcome for the protein-coding RNA from mere transcript sequence, and subsequently to identify the zipcode elements thereof. Our proposed method has achieved good accuracy compared to the baseline methods based on the k-mers features, despite the intrinsic difficulty that arise from the complex and stochastic interactions during trafficking events, as well as the limitations imposed by the available dataset.

ABRÉGÉ

Les cellules sont les unités de base de la vie, et pourtant, elles sont régies par de nombreux processus subcellulaires délicats et, dans une certaine mesure, fragiles, qui sont essentiels à leur survie. Une simple mutation génétique pourrait éventuellement obstruer des processus de régulation importants ou perturber le fonctionnement du produit qu'elle encode, ce qui pourrait éventuellement entraîner la disparition de l'ensemble du système. Par conséquent, il est important de mieux comprendre les nombreux processus de contrôle de la cellule et les facteurs de régulation qui leur sont associés, un exemple frappant étant le mécanisme lié à la localisation sous-cellulaire de l'ARN sur lequel nous nous concentrerons presque exclusivement dans cette étude à partir de une perspective de calcul.

Le mécanisme de localisation sous-cellulaire de l'ARN est l'une des facettes les plus importantes, mais sous-estimée, du processus plus général de régulation des gènes, qui facilite l'organisation cellulaire et la régulation de l'expression des gènes, via le transport des transcrits d'ARN vers leurs emplacements désignés, où leur fonction, leur structure ou des protéines traduites sont nécessaires. Il est généralement admis que le mécanisme du trafic d'ARN est régulé par des facteurs de régulation trans, tels que les protéines de liaison à l'ARN, et les éléments agissant en cis, des extraits courts du transcrit qui contiennent les sites de liaison à la RBP, qui nous appelons zipcode car ils sont censés contenir des informations sur son adresse de livraison.

La publication d'un nouvel ensemble de données de localisation subcellulaire des ARN nous a permis de créer le premier outil de calcul utilisant des techniques d'apprentissage approfondi de pointe, afin de prédire le résultat de la localisation de l'ARN codant pour une protéine à partir d'une simple séquence de transcription, puis d'identifier le éléments de zipcode de ceux-ci. Notre méthode proposée a obtenu une bonne précision par rapport aux méthodes de base basées sur les caractéristiques de k-mers, malgré la difficulté intrinsèque résultant des interactions complexes et stochastiques lors d'événements de gestion de trafic, ainsi que des limitations imposées par l'ensemble de données disponibles.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Overview

This study is focused on the utilization and evaluation of machine learning methods for the inference and prediction of RNA subcellular localization. Relevant biological concepts on the gene regulatory landscape and the roles assumed by RNA subcellular localization are given in Section 1.2, along with a succinct description in Section 1.3 on the various mechanisms behind the multitude of subcellular trafficking events. The importance of RNA intracellular transportation is particularly emphasized in Section 1.4. Our computational models are evaluated on a number of RNA subcellular localization dataset, and we will introduce the relevant technologies that are used to generate these data in Section 1.5. We also provide a sufficient background on the deep learning tools that make up our computational models in Chapter 2, as well as a literature review on the machine learning approaches used to predict RNA-protein interaction and RNA subcellular localization in Chapter 3.

1.2 RNA subcellular localization in the gene regulatory landscape

As framed in the hypothesis of the Central Dogma firstly established by Francis Crick [33], once the information has reached proteins, it cannot be transferred to other proteins or reversed back to nucleic acids. The information encoded in the DNA, on the other hand, may be passed down to the RNA via transcription, and henceforth to protein from messenger RNA (mRNA) via translation [1] . The dogma itself, however, has eluded the discussion on the relevant control mechanism, which is about the rate at which these transfers occur.

The product of a gene, in most cases a kind of protein, undergoes many steps in the process of gene expression in eukaryotes, that are transcription, capping at the 5' end, polyadenylation at the 3' end, RNA splicing, RNA transportation and finally, the translation into a protein, with or without post-translational modification. These procedures, beyond doubt, are highly controlled so that the gene product quantities are maintained at a proper level spatially and temporally, to serve a lot of purposes, one of which is to modulate the complex and intertwined events that often require an appropriate amount and distance of cellular substances and molecules, for example in the stage of embryonic development.

The various control mechanism constitutes the greater gene regulatory landscape, all serving the same purpose — to regulate the gene expression level — but from different angles and at different stages. RNA, and more specifically, the mRNA transportation plays an imperative part in this broad drama. The trafficking of mRNA to where the proteins are needed and then performing rapid on-site translation, indirectly lays its impact on the translation phase that entails a spatial control over the protein distribution in the subcellular territories [22, 14]. It also represents

[1] Although there are other types of information flow, such as the reverse transcription, in this study we would only look at the one that is most well-known , i.e. DNA→RNA→protein, for simplicity.

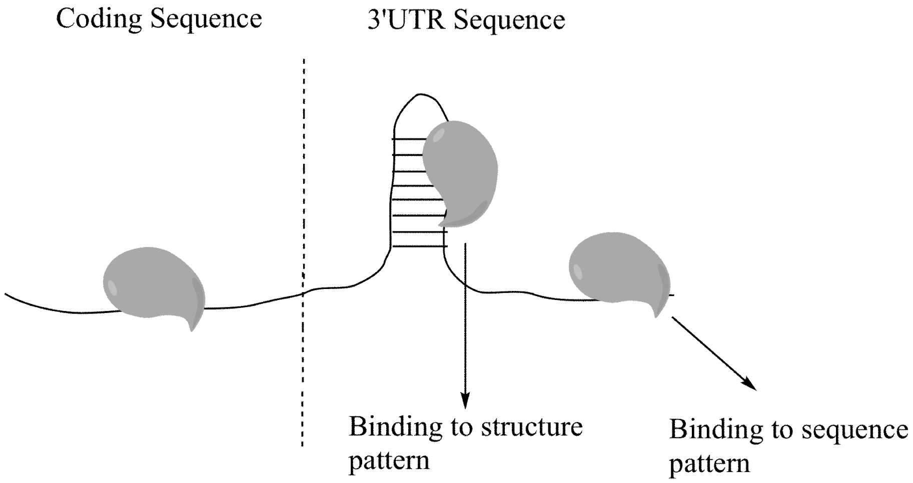

Figure 1–1: RBPs can bind to mRNA possessing specific sequence/structure patterns, at the 3'UTR or the protein-coding regions of the transcript.

an economical version of protein localization, since transporting the RNA templates preserve more energy than its countless protein products.

The RNA localization mechanism is a stochastic process that involves many uncertainty during the trafficking events. A pool of RNA transcripts of a gene can be transported to many different subcellular fractions, although a consistent and asymmetric distribution can often be observed for the majority of the transcriptomes in the previous experiments [12, 60]. Once a gene is transcribed in the nucleus producing a nascent RNA transcript, it will be prepared by different RNA binding proteins (RBP) for the maturation event, forming ribonucleoprotein (RNP) complexes, and then exported out of the nucleus if not retained, by the same regulatory elements but probably of different types, into the cytoplasm and possibly to the membrane, or even excreted beyond the boundary of the cell [22, 14].

It is believed that the RNA localization mechanism involves a diverse population of trans-regulatory factors, the RBPs, which stochastically bind to specific RNA sequence/structure patterns and cooperatively carry the transcript to its intended destination. The RBP binding also happens dynamically during the trafficking event, i.e. an already bound RBP may fall off at some point and another RBP may bind to the transcript and steer it to a new direction. A short and consecutive region containing one or more RBP binding sites is usually identified as the cis-acting element of localization, which has been named in the previous literature as *zipcode* that usually resides in the 3'UTR as well as the protein-coding regions as shown in Figure 1–1.

A set of known sequence-specific RBP binding motifs have been mapped in several previous studies [27, 80, 96, 97], as well as a recent one that augmented sequence motifs with annotated RNA secondary structures [29]. They can be used to align to the computationally imputed RBP binding motifs as an important mean of model verification.

1.3 Cellular mechanisms driving mRNA localization

The mRNA subcellular mechanism has been adequately characterized in various studies of embryonic development. In general, it is mainly driven by three cellular mechanism, that are transporting through the cytoskeleton, diffusion followed by local entrapment and localized from degradation [22, 14], as illustrated in Figure 1–2.

The transportation through the cytoskeleton is via the microfilament network of cells for short range delivery, whereas long range transportation is through the microtubules. Once in the cytoplasm, messenger RNP complexes can be bound by various motor proteins that favours directional movement through the above two pathways, enabling subcellular trafficking. A few examples are given in Figure 1–2: (a) localization of $ASH1$ mRNA through the microfilament network to the bud tip of yeast cell, a highly modulated process involving various RBPs and translational repressors [59]; (b) localization of *β-actin* mRNA from the nucleus to the leading edge in some migratory cell types that enables directed cell migration, accompanied by the zipcode binding proteins (ZBP1 and ZBP2) during the nuclear and cytoplasmic transport [91]; (c) $CAMKII\alpha$ mRNA to the dendrites synapses of neuron cells [39]; (d) localization of *gurken* and *oskar* mRNA to the anterior and posterior poles in Drosophila oocytes [69]. A more detailed and dedicated explanation on these above mRNA trafficking examples are given in [14].

The diffusion and entrapment scheme is simpler to the extent that in general fewer transportation agents are involved. The RNA transcripts simply diffused to the lower concentration area of the cell and subsequently become anchored to that location, albeit still could be assisted by a few RBPs. An example is provided in Figure 1–2 (e), demonstrating the METRO pathway for Xcat mRNA diffusion in Xenopus oocytes [67].

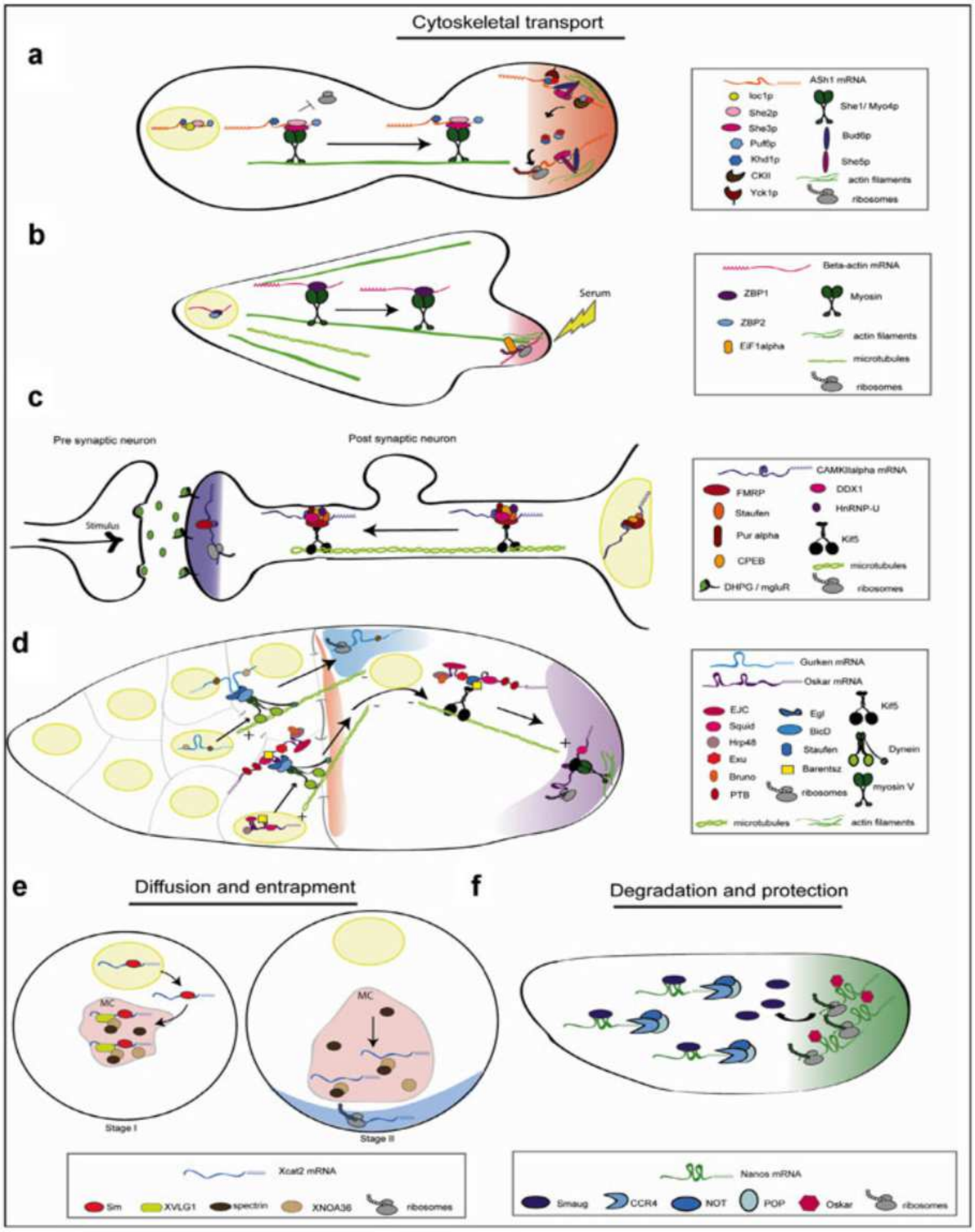

Figure 1–2: The three mechanism that drives mRNA intracellular transport, that are transporting via cytoskeletal pathway (a)-(d), diffusion followed by entrapment (e) and protection from degradation (f). Reproduced from [14].

The third mechanism deals with protection from degradation, in which RNA transcripts are degraded globally except for some specific subcellular sites where the degradation is disabled, therefore allowing mRNA transported there to be translated into proteins. An example is given in Figure 1–2 (f), where the mRNA are degraded in the cytoplasm and are only protected in the posterior pole of the Drosophila embryo during its development [11].

1.4 Importance of the RNA localization

RNA subcellular localization takes on significant roles in the gene regulatory landscape, especially in its ability to coordinate gene regulatory events in the post-transcriptional phase, which necessitates the transportation of certain regulatory RNAs to the periphery of or partitioned from their targets, such as the group of microRNA (miRNA), base-pairing with the mRNA to obstruct translation [3, 10], and the group of small interference RNA (siRNA) to cleave and degrade the mRNA before translation [17]. Other examples from the short non-coding RNA (ncRNA) include small nuclear RNA (scRNA), small nucleolar RNA (snoRNA), as well as many long non-coding RNA (lncRNA) [22, 14], which all exhibit unique localization pattern in the subcellular territories.

Recent studies have also suggested the existence of a certain RNA species called enhancer RNA (eRNA), another kind of short ncRNA, that promotes gene expression, although the exact mechanism of their functionality has not been known [65, 100].

As for the family of mRNA, its localization accounts in part for the major asymmetry observed in the protein distribution [110, 22, 14], since localizing the mRNA templates is a more efficient and energy-preserving option than transporting the countless copies of translated proteins, a feature especially relied on by the neuron cells for dendritic plasticity [22, 14]. It can also prevent protein from being translated at positions that can have a wasteful or deleterious effect [22, 14]. mRNA may also take on non-coding roles similar to those of ncRNA, such as acting as a scaffold to RNP complexes [22, 14].

The significance of RNA subcellular localization can also be appreciated from the related disease pathogenesis, should the normal RNA trafficking pathway be affected either on the transporting agents leading to critical RBP malfunctions, or on the cis-acting zipcode regions of mRNA. The related disease includes cancer, neurological and muscular dysfunction [30, 22, 110]. One notable example would be the Alzheimer's Disease (AD), a dementia with symptoms of cognitive decline and memory loss. According to the prevailing explanation, AD is triggered by the abnormal accumulation of amyloid β-peptide (Aβ) plaques [51], the exposure of axons to which further leads to increased localization and translation of ATF4 encoding mRNA in the nucleus, which ultimately turns on a neurodegenerative program [9]. Additionally, the aberrant mis-localization of the lncRNA BC200 is also found to be highly correlated with AD, whose implicated regulatory function leads to the synaptodendritic deterioration [88].

Thus, understanding further into the RNA subcellular localization mechanism would equip us with the necessary knowledge to design smart medicines that could possibly have a remedial effect on the malfunctioned transportation agents.

1.5 Technologies to map RNA localization

To the extent of our knowledge, the study into transcriptome subcellular localization can either proceed with possibly time-lapsed microscopic imaging with the help of fluorescent tagging directly or indirectly to the mRNA molecules to visualize its movement, or through biochemical manipulation followed by high-throughput microarray or RNA-seq sequencing to measure the enrichment of RNA transcripts in some specific subcellular compartments [14]. We will later discuss more on the latter approach, where two major technologies have emerged that enabled the evaluation of our computational models, in Section 1.5.2 and Section 1.5.3. A complementary overview on the microscopy based methods will be also given, although it is less relevant to this study.

1.5.1 Microscopy imaging

This family of approaches has evolved chronologically from directly using exogenous fluorescently labeled mRNA, to using fluorescent probes, then to indirectly labelling mRNA via fluorescent RBPs.

The earliest approach via injecting exogenous fluorescent-labelled mRNA can be traced back to the experiment on maternal mRNA Vg1 in Xenopus oocytes [119]. Despite its welcoming simplicity, this approach has some major disadvantages, in that the introduction of external mRNA molecules may have a deleterious effect, or possibly alter the actual intracellular trafficking dynamics due to mRNA saturation.

Improvements have soon been made, introducing the fluorescent probes, which only labels an mRNA when the transfected fluorescent oligos have successfully attached to the mRNA molecules. No longer interfering with the endogenous dynamics of the cell, however, it still brings up a new obstacle, which is the high background noise due to the unattached fluorescent oligos. A broad line of works have thus been published to circumvent or suppress the background noise, such as the fluorescence only co-occurs with the oligo binding events [68, 70]. We would not discuss its many variants in detail, and we would refer the interested readers to a review given in [14] Section 5.2.

Finally, a newer and more appreciated technique has emerged that relies on the interplay between two components: the Green Fluorescent Proteins (GFP) fusioned to MS2 bacteriophage coat protein forming the MS2-GFP fusion protein, and RNAs that contain the multimerized MS2-hairpin elements. The MS2-GFP fusion protein can bind to the MS2-elements of RNA, when both are co-expressed, therefore identifying the RNA trafficking pathways along its movement inside the cell. Key disadvantages of this method can be attributed to the concatenation of multimerized MS2-hairpin to the target RNAs may interfere with their regulatory role inside the cell, as well a high background noise due to unbound GFP-MS2.

The general disadvantages related to the microscopy based methods lie in their relatively low throughput, and the requirement of more complex and unaffordable facilities. Therefore, our work only uses data generated via biochemical manipulations coupled with deep sequencing.

1.5.2 Fractionation based approaches

Compartment specific RNA profiling is generally achieved through subcellular fractionation followed by RNA deep sequencing. A series of work based on the biochemical fractionation approach have successfully identified RNA associated to a number of subcellular fractions, such as nucleus and cytoplasm [74, 18, 113], to name a few. Frac-Seq [107, 84] and CeFra-Seq [75] are both specific examples belonging to this category.

Fractionation based methods have also been used to study protein subcellular localization, for example in a recent paper the authors proposed a method called SubCellBarCode [90] that separates a subcellular-wide proteome into five subcellular fractions in duplicates for each of the five different human cancer cell lines, resulting in 50 samples per protein. The fractionation step is followed by a mass spectrometry to identify the proteins localized to each of the isolated compartments.

According to a clustering analysis based on t-SNE dimensionality reduction, 15 clusters of the proteins localization profiles are identified, which are further grouped into 4 subcellular compartments, that are secretory, nucleus, cytosol and mitochondria, via aligning to the compartment enrichment annotations given by Gene Ontology (GO) and Uniprot. A machine learning method is then developed to predict a single subcellular localization site for each protein, that appears to possess notable predictive capacity.

In the RNA domain, on the other hand, CeFra-Seq is another fractionation based method that combines biochemical fractionation and high-throughput RNA sequencing to map RNA in specific subcellular compartments [75]. Biochemical fractionation part first breaks the cell membrane using a mild hypotonic lysis and later undergoes the Douce homogenization. Matters from a number of subcellular compartments are subsequently separated with centrifugation, recovering the nucleus from the purification of sucrose cushioned ultracentrifugation, as well as the soluble cytosolic fraction from the supernatant. The re-solublilized portion of the pellet subject to the Triton X-100 (1%) and ultracentrifugation is assigned to the endomembrane, whereas the residuals goes to the insoluble fraction.

Then followed by high-throughput RNA sequencing, the overall protocol is able to measure the transcriptome enrichment in four intracellular compartments, that are nucleus, cytosol, endomembrane and cytoplasmic insolubles. A later study based on the CeFra-Seq protocol has further revealed that around 80% of the RNA transcripts have exhibited asymmetric subcellular localization pattern [12].

The limitation of fractionation based approaches mainly lies in the difficulty of mapping RNA in some organelles or compartments that are ineffective under the classical biochemical purification scheme. Also, the fractionation step is susceptible to contamination and loss of materials, from or to other subcellular components.

1.5.3 Protein-RNA cross-linking based approaches

Another category of biochemical manipulation for RNA profiling follows from a general scheme of cross-linking proteins and RNAs to form RNP, which can be later precipitated with the help of immunoprecipitation (Clip or RIP) and purified to enable RNA deep sequencing.

The first experiment of this kind was done on the mouse brain [115], which firstly undergoes an ultraviolet in-situ cross-linking step, forming covalent bonds between the protein and RNA that are in direct contact. Then following the cell lysis, protein immunoprecipition and proteinase K pull-down in the purification step, the RNA can be subsequently isolated and sequenced.

The main disadvantage of this scheme is that it only enables cross-linking between protein and RNA that are in direct contact, which is too restrictive to reveal a broader picture of RNA subcellular distribution.

An newer protocol called APEX-RIP has been recently developed to map transcriptome enrichment in membrane-enclosed cellular organelles, e.g. the nucleus and mitochondrial, as well spaces in between the membranes such as the cytosolic and endoplasmic reticulum but only up to a limited efficacy [60].

This technology combines APEX and RIP, that are peroxidase catalyzed proximal endogenous protein biotinylation, and RNA Immunoprecipitation to target on RNP for the subsequent protein pull-down and sequencing. APEX is an artificial peroxidase that can be genetically integrated to a number of subcellular compartments, via implanting APEX-encoding DNA to the cell which can then be translated into many APEX protein copies that exhibit deterministic localization traits. Upon the attachment of its substrates, APEX catalyzes the formation of biotin-phenoxyl radicals which later triggers the biotinylation of endogenous proteins within close proximity (a few nanometers). The APEX-tagged proteins can be used to bind the co-localized RNA transcripts in their vicinity, via cross-linking with formaldehyde. The bound RNA transcript can be later enriched and identified with streptavidin pull-down and deep sequencing.

APEX-RIP was also used to map RNA enrichment in four subcellular compartments, which are nucleus, cytoplasm, endoplasmic reticulum and mitochondrial.

While it is not straightforward to compare in terms of the accuracy between CeFra-Seq or APEX-RIP, it is generally believe that APEX-RIP represents a more promising direction of RNA intracellular trafficking mapping protocol.

A more precise approach to directly make use of APEX enabled biotin-tagging on RNA, instead of cross-linking RNA to the biotinylated protein, has been developed and named as APEX-Seq [42], which addresses the problem of poor cross-linking resolution in non-membrane enclosed cellular space.

1.6 Learning RNA subcellular localization and identifying the zipcodes

In this study we propose the first computational method to predict mRNA localization from RNA sequence (with or without secondary structural annotation) inputs, using advanced deep learning techniques such as convolutional neural nets (CNN) [71] and long short term memory (LSTM) [56]. A review on these deep learning components will be given in Chapter 2.

We name our method RNATracker, which will be presented in Chapter 4, along with the complementary analysis of some learned model components including a visualization of sequence motifs learned by its first convolutional layer, attention weights to highlight that the 3'UTR regions of RNA transcripts are more informative to infer localization, as well as a mask test to identify zipcodes regions from the RNA sequences in the dataset.

1.7 Motivation

In light of the recent advances of high-throughput sequencing technologies and improved biochemical manipulation methods to identify RNA localized at some specific organelles or compartments, an unprecedented opportunity has emerged, where a systematic and large-scale training and evaluation of bioinformatics/machine learning methods on RNA subcellular localization have become possible. These algorithmic progression on the available RNA localization dataset may provide the system biology community with a clearer understanding into the underlying RNA trafficking dynamics, and a more comprehensive identification of the associated regulatory elements.

CHAPTER 2
Background on deep learning

Deep learning is often praised for its powerful end-to-end and automatic feature-extracting capability, provided that the data structure of the inputs has been adequately exploited, which usually requires carefully designed neural architecture and reasonably selected hyper-parameters, as well as a sufficient amount of training data.

We will first provide the background on two of the most widely used building blocks in deep learning — Convolutional Neural Net (CNN) and Long Short-Term Memory (LSTM). Then we will talk about the residual neural net, which is a more advanced organization of CNN. Finally, we will touch briefly on the deep generative adversarial neural net and activation maximization, the combination of which would enable us to design RNA zipcodes.

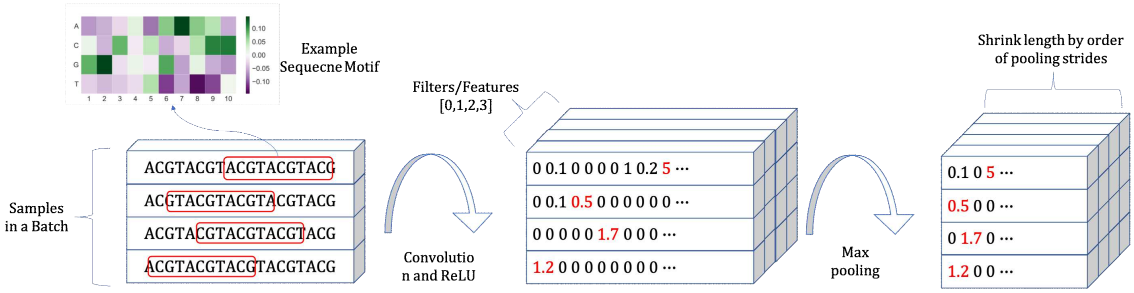

Figure 2–1: A nutshell of convolution and pooling operations in sequence scanning.

2.1 Convolutional Neural Net

Since their first invention, CNNs [71] have gained great popularity inside the machine learning community. The original CNN is designed to classify digit images, which surpassed the previous methods at the time. A number of its distinguished advantages compared to a simple feed-forward neural net are listed as follows:

- CNN exploits the topological information of the inputs, such as the order and structure in a patch of pixels which would identify an object, and the order and structure of a snippet in a sequence of nucleotides which deliver an RBP binding motif.
- Each neuron is computed from a small patch of neurons in the previous layer (called a local receptive field). This sparsity in connectivity enables CNN to compute relatively faster than a fully-connected neural net, and forces the CNN to learn a sparse solution which acts as a strong regularization.
- The same set of kernel is used to parse every patch of the input, thus facilitating weight sharing, which in turn reduces overfitting.
- CNN is also able to parse patches of inputs in parallel, allowing for efficient training and evaluation of high dimensional data.

CNN usually consists of a convolution operation, coupled with a non-linear activation and a pooling stage. The size of the output is usually smaller, unless specifically chosen to retain the same size as the input. Down-sampling the input is, in general, preferable since a smaller output would improve the computational efficiency in the subsequent layers, albeit suffering from a minor information loss, unless the down-sampling ratio is grossly out of proportion.

However, CNNs can also be used in an up-sampling fashion, namely to increase the size of output, resulting in, for example, larger images or longer genomic sequences. This genre of convolution is called transposed-convolution.

An example of 1-dimensional convolution on a batch of genomic sequences is given in Figure 2–1, complemented with Rectified Linear Units (ReLU) as the non-linear activation function, and a max-pooling stage. We will talk about each component in more details in the rest of this section.

2.1.1 Convolution

Denoting the 1-dimensional convolution kernel as K of shape (L, D_{in}, D_{out}), where L is the length of the kernel, D_{in} is the number of the channels in the input and D_{out} is the number of filters/channels in the output, the convolution operation can be expressed as follows,

$$O_j = \sum_{i=1}^{D_{in}} I_i \bigoplus K_{i,j}$$

where O_j is the j_{th} channel of the output for $j = 1...D_{out}$, and I_i is the i_{th} channel of the input. Note that for simplicity, the final addition of a bias vector has been omitted.

$\bigoplus$ is the convolution operation, which moves $K_{i,j}$, a vector of length L to be thought as a scanning window, along I_i. At each scanning window position, $K_{i,j}$ is element-wise multiplied with a local patch of data in I_i, whose sum is filled in for that position as the convolution outcome.

Additional parameters of the convolution operation include the length of paddings to either size of the sequence denoted as p, and the length of strides when moving the scanning window denoted as s. The length of the output o can be then given as follows, denoting the input length as i and the length of kernel as k,

$$o = \lfloor \frac{i + p - k}{s} \rfloor + 1$$

Therefore, to obtain the output with the same length as the input, the least amount of paddings is determined to be

$$p = (i - 1)s + k - i$$

Note that in this case, $i + p - k$ is a multiple of s, and for all $p' = p + j$ where $j = 1...s - 1$ would result in the same output length. The down-sampling of an input feature map can be achieved with $s \geq 2$.

2.1.2 Pooling

Pooling provides small invariance to the translation of inputs. There are a few types of pooling, such as max pooling and mean pooling, just to name a few. The pooling operation splits the input into patches, and each patch is summarized with the max operator or the mean operator accordingly. The parameters associated with pooling are the length of the patches k and the size of the stride s, without trainable weights. The length of the output o can thus be determined as

$$o = \lfloor \frac{i - k}{s} \rfloor + 1$$

The role taken by pooling layers is mainly for dimensionality reduction on the feature map, although this function can still be fulfilled with a down-sampling convolution layer with stride greater than one. Down-sampling with convolution layer is usually more computationally efficient than pooling. For example, the same downsampling effect of a uni-stride convolution layer followed by a stride two pooling layer, can be equivalently replaced with a single convolution with stride two that consumes at least four times less computation.

2.1.3 Transposed-convolution

Considering the usual convolution as a mapping from the input to the feature map, fulfilling its role as a encoding operation, then its inverse operation would be to recover the input from the feature map as a decoder while maintaining the same convolutional connectivity patterns between the input and the feature map. The inverse operation is usually termed as the transposed-convolution, which has been proven useful in the design of convolutional decoders and generators, where an up-sampling operation is usually performed to map a latent-encoding to an input data point.

Since convolution is in essence an affine mapping, and to preserve the same convolutional connectivity, one only needs to transpose the weight matrix implied by the convolution kernel, which takes place during the backward pass. Therefore, the transposed convolution can be then thought as the gradients of some convolution with respect to its input [41].

2.2 Long short-term memory

Despite their great applicability and pattern extraction capacity, CNNs are not yet powerful enough to infer from time series data when used by themselves. Whereas CNN looks for similar patterns recurring at different patches of the input or feature map, a Recurrent Neural Net (RNN) investigates the correlation between different locations (or time step) of the input with a built-in order of information that have been processed sequentially.

A Long short-term memory (LSTM) network is a specific type of RNN that is used more frequently in practice, along with other types of RNN such as the Gated Recurrent Unit (GRU) [24].

A diagram of a standard LSTM cell, a time step from an unrolled LSTM, is shown in Figure 4–2 (B), accompanied with the formulas that define its operations from Eq. 4.1 to Eq. 4.3. The LSTM maintains a so-called cell-state, which memorizes information accumulated from the previous time steps. The LSTM is also autoregressive, in that the output from the last time step is also an input to the current LSTM cell/time step.

The LSTM prevails in discovering long range dependencies and correlations in the input, implemented by its use of the forget gate which eliminates information from the cell state that is deemed no longer relevant, the input gate to update the cell state with new information from the current input, as well as the output gate which decides what portion of the current cell state should be present in the output.

A more detailed LSTM explanation can be found in Section 4.4.

2.3 Residual neural net

Due to the sparse-connectivity of the convolution operation in CNN, each neuron in the output is only connected to a (small) patch of elements in the input, which is called the local receptive field of that neuron. The size of the local receptive field is fixed for all neurons in one convolution layer, and by stacking multiple convolution layers, each having a small local receptive field to capture the fine-grained features, a broader receptive field can be obtained step-by-step, allowing the network to compose these fine-grained local features into a more meaningful global representation of the inputs. Therefore, the depth of the convolutional network is crucial to achieving a better accuracy, as has been discussed in many previous papers on image classification tasks [105, 112].

However, naively stacking many convolution layers can also impede the training as a side effect, due to the vanishing gradient or the exploding gradient problems. Although these problems can be to some extent alleviated with various normalization techniques such as the batch normalization [58] or layer normalization [77], the general effect of adding more convolution layers has still been observed to be deleterious over the shallower networks [52], even accompanied with the normalization layers.

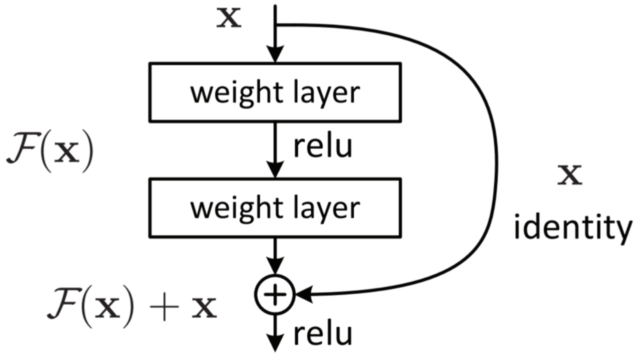

Figure 2–2: A basic residual neural net building block — adding skip connections at the output of the convolution layer. Reproduced from [52].

In order to bring out the full anticipated capacity of deeper convolutional neural architecture and to ease the difficulty that arises during the training, a residual neural net (resnet) architecture [52] has been proposed that adds skip connections (or equivalently, residual connections) interleaving the convolution layers, feeding the input together with the output to the next convolution layer, as shown in Figure 2–2. To intuitively explain the benefits of adding skip connections, one could imagine that in the beginning of training, the residual net can learn to skip some weighted convolution layers, thanks to the skip connection, to learn some simpler features and use them to make a prediction, which helps speed the convergence. Once better features can be learned or manipulated at the skipped layers, the network will shift its attention back to the training of those layers.

We note to the readers that an actual residual block can be more complex than the one shown in Figure 2–2. In particular, if $\mathcal{F}$ performs down-sampling/up-sampling, or alters the number of channels, the shape of $\mathcal{F}(x)$ will not equal to that of x. In this case, an additional 1×1 convolution layer is usually placed at the shortcut to match the shape and dimension. Also, a batch normalization layer is often placed after each weight layer in the residual block.

The above paradigm of organizing the convolution layers to build deeper networks enables researchers to stack tens or even hundreds of convolution layers that expedites the development of many areas of deep learning application, such as objective detection and text mining. Its applicability, beyond doubt, also extends to genomic sequence analysis tasks.

2.4 Activation maximization

Once having trained and fixed a predictive model that maps an input (image/genomic sequence) to a prediction, say a vector of categorical probabilities in a multiclass classification task, a question that one could ask in the context of activation maximization, is to devise an example (or a population of examples) that maximally activates a neuron at some intermediate layer or a logit at the output layer.

We are interested in maximizing the output of a specific logit, since it gives rise to designing an input that maximizes its probability of belonging to the corresponding category.

Expressing the predictive model as $o = f(\theta^*, x)$, where o is the vector of logits at the output layer given input x and fixed weights θ^*, $\frac{\partial o_i}{\partial x} = \frac{\partial f(\theta^*, x)_i}{\partial x}$ provides the gradients to maximize the objective function $f(\theta^*, x)_i$. Therefore,

$$x = x + \lambda \cdot \frac{\partial f(\theta^*, x)_i}{\partial x}$$

gives the general formula for activation maximization, where x is a randomly initialized input, or chosen randomly from the dataset.

This naive approach has a significant drawback, in that the optimized input x^* may not resemble any real data, or in other words, has very low fidelity. To capture the prior on real data distribution, [89, 64] have proposed to stack the predictive model on top of a generator function capable of mapping an input code to a realistic output, and the activation maximization is thus done solely on the input code, fixing the generator function as well as the predictive model.

In addition to the striking simplicity of the method, pretraining the generator as well as the predictor separately before the actual activation maximization begins also gives room for a lot of flexibility. However, considering that the optimization on the input code is a long iterative process, and that a whole new optimization process is needed each time to obtain a new batch of input code, this method is unsurprisingly inefficient, hence unfit for the purpose of performing large-scale sampling.

2.5 Deep generative models

The club of deep generative models has been expanding and accepting new members ever since the boost of deep learning. A number of its notable members include the autoregressive models (e.g. LSTM), variational autoencoders (VAE) [66], generative adversarial nets (GAN) [47], as well as their predecessors the deep belief nets (DBN).

The VAE operates under a general probabilistic setting, assuming that each data point is associated with a latent variable, whose prior and likelihood function are from a parametric family of distributions, which are continuous and differentiable almost everywhere. However, the true posterior can be intractable, which necessitate its construction of a probabilistic encoder that models the distribution of the latent variable given an input to approximate its true posterior. A probabilistic decoder is also employed to learn the likelihood function. A variational lower bound can be then established on the marginal log data likelihood, together with the reparameterization trick, one could easily adopt a standard off-the-shelf gradient descent technique to jointly optimize the decoder and encoder, which learns better approximate distribution to the true posterior as well as improves the marginal data likelihood.

A GAN, on the other hand, does not have an encoder component that explicitly models a distribution on the latent variable, nor does it optimizes directly or indirectly on the marginal data likelihood. It has a generator/decoder function that maps a latent encoding to a data structure that mimics the real data, and a discriminator function to differentiate the real and fake data. The interplay between the generator and discriminator, to put it simply, is for the discriminator to assign higher scores to the real data and to penalize the generated data, whereas the goal of the generator is to fool the discriminator into assigning higher scores to its solutions. Usually, the equilibrium is reached when the distribution P_θ implied by the generator function equals to the real data distribution P_r.

For brevity concern, in this section we would only mention two types of GAN — the original GAN whose objective can be framed as minimizing the Jensen-Shannon Divergence (JSD) between P_r and P_θ, as well as the Wasserstein GAN (WGAN) which minimizes the Wasserstein distance between the above two distributions. There are other good formulations of GAN, and various conditional GANs using style transfer, as well as add-on tricks to stabilize training, such as spectral normalization [85].

2.5.1 Original GAN

The training objective for the original GAN for both two components is

$$\min_G \max_D \mathbb{E}_{x \sim p_r}[log(D(x))] + \mathbb{E}_{x \sim p_\theta}[log(1 - D(x))]$$

where G stands for the generator and D stands for the discriminator. However, this formula suffers from the vanishing gradients on the generation function, when the discriminator is too close to optimal, as pointed out in [47, 111], which impedes the convergence.

One way to salvage this is to carefully mitigate the discriminator and generator updates such that the discriminator won't overwhelm the generator, albeit how to manipulate and schedule the updates are more of an engineering issue. Another alternative is to replace the generator objective $\min_G \mathbb{E}_{x\sim p_\theta}[log(1-D(x))]$ with $\min_G \mathbb{E}_{x\sim p_\theta}[-log(D(x)]$, which unfortunately, makes the update unstable when the discriminator is imperfect [111].

2.5.2 Wasserstein GAN

Wasserstein GAN (WGAN) is a new formulation of GAN that equivalently minimizes the Wasserstein distance between the real data distribution p_r and p_θ implied by the generator, that is,

$$W(p_r, p_\theta) = \inf_{\gamma\sim\Pi(p_r,p_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$$

. Much of its theoretical background is borrowed from optimal transport [117], whereby the original intractable Wasserstein distance is allowed to be transformed into an optimizable objective function with some additional constraints, via the Kantorovich duality theorem,

$$W(p_r, p_\theta) = \sup_{\|d_\phi\|_L \leq K} \frac{1}{K}\mathbb{E}_{x\sim p_r}[d_\phi(x)] - \mathbb{E}_{x\sim p_\theta}[d_\phi(x)]$$

where $\|d_\phi\|_L \leq K$ means the discriminator function is K-Lipschitz continuous. The constant K can be subsumed into the learning rate.

It is shown in [4] that the Wasserstein distance assumes a weaker topology than JSD or KL divergence, which will either be discontinuous or not well-defined, when the data generating distribution p_θ is supported by a low dimensional manifold in a way either it does not exist or it has disjoint support to p_r. The Wasserstein distance, on the other hand, is still continuous and differentiable almost everywhere.

The Lipschitz continuity constraints can be met with weight-clipping up to some constant K, by practically clamping all weight parameters into a compact set e.g. $[-0.01, 0.01]$. An alternative and even better technique is to use gradient penalty [49] which adds an additional term to the loss function,

$$\mathbb{E}_{x \sim p_{int}}[(\|\nabla_x d_\phi(x)\|_2 - 1)^2]$$

where p_{int} is a uniform distribution defined on the line segment between two points, each sampled from p_r and p_θ, thus a sample from p_{int} is a linear interpolation between a real data point and a fake fake data point.

Gradient penalty allows for a greater expressivity in the discriminator function, so that the generator can also learn to generate data with richer features. In practice, people would always use gradient penalty instead of weight-clipping when training a WGAN.

CHAPTER 3
Machine learning to study RNA-protein interaction and RNA localization

3.1 Deep learning for RNA-protein interaction

There have been a grand series of works that leveraged the advances of deep learning in the context of predicting DNA/RNA-protein interaction, ever since its first successful application in the DeepBind paper [2] in 2015, which predicts binding specificity on short snippets of DNA or RNA sequences to certain DNA or RNA binding proteins. DeepBind is a deep neural net based on the convolutional architecture, and can be fitted with dataset from a multitude of sources, such as the Protein binding microarrays (PBM) which reveals the transcription factor binding specificities [86], and the RNAcompete assays [96] generated in vitro that informs on RBP binding sites. The output from DeepBind is a binding intensity given by a sigmoid function that can be used to perform classification.

DeepBind has been demonstrated to be superior over the traditional methods based on position weight matrices (PWM) of binding proteins, in terms of the efficiency on large-scale training and inference, as well as the prediction accuracy. The authors also revealed that the DeepBind model, having its parameters fitted with in-vitro dataset such as the PBM or RNAcompete, was still able to make good predictions on in-vivo data collected from ChIP-Seq or CLIP-Seq. This acts as an extra validation that given sufficiently large training set, end-to-end deep learning methods are able to learn meaningful biological assumptions from plain supervised genomics sequence data, without any built-in prior knowledge on the underlying regulatory process, or resorting to manually-curated feature sets.

Despite the great success of DeepBind, it invites to improvements from various aspects. One of the most obvious would be to build a more optimal neural architecture to better explore the inherent structure of the sequence data, with the addition of complementary features such as the RNA secondary structures in the RBP binding specificity prediction tasks.

One another possible improvement, though not having been addressed enough in the line of works despite its significant relevance to the understanding of the functional roles of these genomics, is to improve the model interpretability. A more robust and meaningful inference is often lacking in the deep learning application literature, compared to the more traditional machine learning methods.

Afterwards, a broad family of works have emerged on making use of deep learning to predict DNA/RNA sequence functions or properties, such as the DanQ model [95] on DNA functions that stacks a bidirectional LSTM networks on top of convolution networks, and the iDeep [93] that integrates another sets of manually imputed features such as the basepairing probability of each nucleotide into the prediction pipeline.

It has been long known that RNA secondary structure has a significant influence on the RNA-protein interaction, and many algorithms for RBP binding predictions and motif identifications have attempted to integrate the structural information into the model pipeline.

3.2 RNA secondary structure enhanced RBP motifs discovery and binding prediction

MEMERIS [55] is the first algorithm of its kind that searched for sequence motifs with general structural properties, rather than sequence-structure pairs with specific composition or constrains. This method favors searching sequence motifs on single-stranded RNA, such as the hairpin or the bulge between two stems, over the ones located in double-stranded regions. MEMERIS is viewed as an extension over the MEME motif searching program [7] that modifies the Expectation Maximization (EM) algorithm with priors on the likeliness of a substring being located on single-stranded RNA, in a way that the EM algorithm will prefer to identify sequence motifs on single-stranded RNA areas over double-stranded areas.

The limitation of MEMERIS, such as the restricted scope of structural preference, as well as its reliance on unrealistic minimum free energy, has been addressed in RNAContext [62], which notably uses an ensemble of all possible secondary structures estimated by SFOLD [37] and annotates structural context for each nucleotide with a fixed vocabulary set. The probability of a motif binding to a substring is jointly estimated with the sequence and structural contexts, according to the standard biophysical model [102]. The predicted binding probability is then linearly mapped to a binding affinity, and the model parameters are estimated using Limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [20] on the least square cost function between the estimated binding affinity and its supervised target.

GraphProt [83] informatively represents an abstracted RNA secondary structure obtained from RNAshape [106] as a hypergraph that encodes relationship among groups of stable subgraphs such as hairpin, stem, bulge, and the basic nucleotides.

It is derived from a predecessor model called GraphClust [54], making a few improvements and adaptations in the RBP binding scenarios in addition to the hypergraph representation mentioned earlier, such as enforcing the order of RNA folding from 5'UTR to 3'UTR on a directed graph, and focusing on local RNA folding around the RBP binding sites.

The same graph kernel, Neighborhood Subgraph Pairwise Distance kernel (NSPD) [32], is used as in GraphClust, where two features are computed for a pair of subgraphs in relatively close proximity, that are the maximum size of the subgraphs (radius) and the maximum distance between the two subgraphs. A standard support vector machine is then used to classify RBP binding with a set of very high dimensional features (up to millions) extracted by this kernel.

The integration of secondary structural information of RNA into deep learning pipelines is usually achieved by some sequence based structural annotation instead of exploiting its full graphical capacities, such as in [92]. One of the more promising directions of leveraging RNA secondary structural information in deep learning would be to adopt the graph neural net [103] approaches that implicitly learn graph kernels on RNA graphs.

3.3 Predictive models for RNA localization

There have been a few works that attempt to classify lncRNA localization. LncRNA is a type of non-coding RNA whose length exceeds 200 nucleotides, with regulatory roles well-studied and characterized at the gene transcription level, as well as various post-transcriptional steps including RNA splicing [114], mRNA translation [120] and degradation [46]. LncRNA may also take on certain trans-regulatory roles in the context of localizing mRNA, via base-pairing with the mRNA and indirectly allowing the RBP attached to itself to maneuver the mRNA [22]. Mutation in the lncRNA function may also hold implication for a variety of disease [5]. Therefore, it is of equal importance to understand the underlying mechanism of localizing lncRNA.

These previous studies have mainly relied on an online database called RNALocate [121], which contains manually recorded RNA localization entries for 42 subcellular locations across 65 species. RNALocate only associates a discrete subcellular localization label per RNA for a certain cell type in a certain species, without providing any further intensities or enrichment values. However, RNALocate is indeed a very convenient database for developing and evaluating machine learning methods.

LncLocator [21] and iLoc-lncRNA [109] are two machine learning methods that are developed under the RNALocate database, trained and evaluated on less than 1000 lncRNA data. LncLocator firstly learns a hidden representation of the lncRNA sequences with an autoencoder, then makes prediction with an ensemble of machine learning algorithms. iLoc-lncRNA improves fairly drastically over LncLocator with a simpler multi-class support vector machine, but with pseudo 8-tuple nucleotide compositions (PseKNC) as features.

DeepLncRNA [48] is a deep neural network built for large-scale lncRNA subcellular localization classification, with 8678 lncRNA samples collected from the ENCODE project [26]. Each lncRNA is associated with a binary label of either nuclear retention or cytosolic export, after applying a certain threshold on the weighted and averaged log2 fold-change values across all cell types.

The authors in [48] have followed a feature-engineering approach. The inputs to DeepLncRNA are manually selected features including k-mers (k=2..5), lncRNA subtypes annotated by the Ensembl database [1], chromosome location and RBP binding annotations computed from known RBP binding motifs given in [97]. DeepLncRNA is composed of three fully connected layers, with ReLU activation function, dropout and L1, L2 regularization. It is shown that DeepLncRNA outperforms LncLocator and iLoc-lncRNA, in terms of accuracy and sensitivity — the ability to correctly classify lncRNA retained at the nucleus, when evaluated on a very small test set under a train-test split.

In a recent paper [124], B. Zuckerman et al. identify splicing efficiency as the predominant factor influencing the nuclear retention and export mechanism of cells, along with the other factors including the transcript length, sequence content and etc.

It has been empirically observed that lncRNAs tend to be retained in the nucleus since their regulatory roles are best characterized in the nuclear and chromatin areas [116, 87], whereas mRNAs are usually exported to the cytoplasm where their products can serve certain structural or functional roles after translation. A recent studied also revealed that lncRNAs are more subject to alternative splicing than mRNAs, leading to more possible isoforms that are inefficiently spliced [36].

Motivated from the above observations, the authors gathered the RNA-seq data in the cytoplasmic and nuclear compartments of nine human cell lines covered by the ENCODE project, and obtained a number of features for the downstream predictions. Gene level splicing efficiency were estimated as the ratio between exon-exon reads and the sum of exon-exon and exon-intron junction reads as defined in [87], selecting only the non-overlapping introns with the most confident support in the RNA-seq data. Gene level splicing specificity was also estimated which reflects the frequency of the predominant splicing pattern.

A preliminary study on the (Spearman's) correlation of the individual feature to the cytoplasmic/nuclear target ratio has given a lot of insights. Splicing efficiency is observed to be negatively correlated with the nuclear enrichment for both lncRNA and mRNA, and is strongly correlated to the target ratio for the mRNA faction. Other factors also contribute non-trivially, for example cytoplasmic enrichment is correlated with gene expression level, splicing specificity and etc.

The authors also tried a linear regression model to learn a mapping of the features to the target ratios, to see if a combination of features would lead to more accurate prediction, but with limited success. A random forest classifier is later used to classify the samples into three binned groups (cytoplasmic, intermediate and nuclear), after applying appropriate thresholds on the ratios. The random forest classifiers have attained satisfying accuracy.

3.4 Conclusion

With the emergence of larger scale and higher resolution RNA subcellular localization dataset obtained via biochemical manipulation and RNA deep sequencing, the timing is right to devise more catered end-to-end machine learning methods to gain deeper insights into the underlying mechanism driving RNA subcellular localization. One major advantage of adopting a sequence based predictive pipeline that is possibly enhanced with RNA structural information, would be the identification of fine-grained sequence determinants considered as zipcodes.

CHAPTER 4
Prediction of mRNA subcellular localization using deep recurrent neural networks

4.1 Preface

The following Chapter is taken from a recent publication: Zichao Yan, Eric Lécuyer and Mathieu Blanchette. Prediction of mRNA subcellular localization using deep recurrent neural networks. Intelligent System on Molecular Biology, 2019.

The contributions came from Zichao Yan who implemented and evaluated the computational models and wrote parts of the paper, as well as Eric Lécuyer and Mathieu Blanchette who validated the biological concepts, analyzed the biological interpretation and also wrote parts of the paper.

4.2 Abstract

Motivation: Messenger RNA subcellular localization mechanisms play a crucial role in post-transcriptional gene regulation. This trafficking is mediated by trans-acting RNA-binding proteins interacting with cis-regulatory elements called zipcodes. While new sequencing-based technologies allow the high-throughput identification of RNAs localized to specific subcellular compartments, the precise mechanisms at play, and their dependency on specific sequence elements, remain poorly understood.

Results: We introduce RNATracker, a novel deep neural network built to predict, from their sequence alone, the distributions of mRNA transcripts over a predefined set of subcellular compartments. RNATracker integrates several state-of-the-art deep learning techniques (e.g. CNN, LSTM and attention layers) and can make use of both sequence and secondary structure information. We report on a variety of evaluations showing RNATracker's strong predictive power, which is significantly superior to a variety of baseline predictors. Despite its complexity, several aspects of the model can be isolated to yield valuable, testable mechanistic hypotheses, and to locate candidate zipcode sequences within transcripts.

4.3 Introduction

RNA subcellular localization constitutes a key but underappreciated aspect of gene regulation [23]. Once transcribed, capped, spliced, polyadenylated, mRNA can be shuttled to different parts of the nucleus, or exported to the cytoplasm, where it can further be transported to specific sites, or even excreted in extracellular vesicles (Figure 4–1). In the case of messenger RNA (mRNA), subcellular localization can control how much will be available for translation by ribosomes and where translation will occur, thereby allowing both a quantitative and spatial control over protein production. In particular, this mechanism represents an economical mean of protein localization, by transporting the messenger to the site where the protein is needed and performing on-site translation. While the importance of RNA subcellular localization is best characterized in embryonic development [73] and neuronal dendrites [19], it is also highly prevalent in other cell types, with more than 80 % of human transcripts showing asymmetrical localization in human and insect cultured cells

[13]. Defective RNA trafficking, due to mutations either in the cis- or trans-acting molecules, are linked to a number of muscular and neurodegenerative diseases, as well as cancer [31]. Improving our understanding of the mechanisms of mRNA localization, and its dependency on transcript sequence or structure, is thus important for the fundamental understanding of molecular biology and has profound biomedical implications.

The RNA trafficking process is mainly driven by a diverse population of trans-regulatory factors called RNA binding proteins (RBPs) [43, 44, 98, 38], which stochastically, cooperatively, and dynamically bind to specific RNA sequence/structure patterns. While non-specific protein-RNA interactions are common and help stabilize mRNAs, sequence-specific binding to short sequence/structure patterns allows transcript-specific regulation [15]. Indeed, sequence motifs have been mapped for a large set of RBPs [28, 81].

mRNA localization cis-regulatory elements (also known as zipcodes) are short (20-200 nt) RNA regions that harbor binding sites for one or more RBP that help mediate the transport mRNAs to their intended destination, either actively along the cytoskeleton, diffusion, or compartment-specific degradation. Although the number of well-characterized zipcodes remains very limited (only about a dozen in human), most are observed to be located in the 3' UTR (but many exceptions exist) [15].

While the importance and prevalence of mRNA subcellular localization has been known for a long time based on experiments such as fluorescent in-situ hybridization (FISH) (Lecuyer et al. 2007), it is only more recently that high-throughput sequencing-based assays emerged. APEX-RIP is a technique that takes advantage of protein proximity based biotinylation, mediated by a compartment-specific APEX2 fusion protein, to identify localized transcriptomes [61]. The organelle-localized APEX2 fusion protein will biotinylate proximal interacting proteins and, following cross-linking and streptavidin pull-down, co-localizing mRNAs can be identified by deep sequencing. This technology was recently used to map the transcriptome of the nucleus, cytoplasm, endoplasmic reticulum (ER), and mitochondria. CeFra-seq is an alternate technology relying on biochemical separation of subcellular components, followed by RNA-seq [76, 13]. It was used to map transcript abundance in the nucleus and cytosol, as well as those associated to endomembranes (ER, golgi, etc.) and those left in the insoluble fraction, consisting of mRNAs associated to cytoskeletal and mitotic apparatus-associated proteins. Both technologies yield reproducible assessments of relative mRNA abundance in the subcellular component they probe and demonstrate the breadth of localization patterns observed in a variety of human cell types.

In this paper, we aim to build a predictive model of mRNA localization that will quantitatively determine the relative expression of a given transcript among a predetermined set of cellular compartments, based only on sequence information. Such a model is essential to generate testable mechanistic hypotheses about the cis- and trans-regulatory molecules at hand and predict the impact of mutations on this key step of gene regulation.

The computational identification of functional regulatory elements within biological sequences is one of the key problems addressed by bioinformatics approaches. Recently, new types of machine learning approaches emerged for sequence function prediction. Those are based on deep neural networks, and often combined convolutional [72] and recurrent neural networks (e.g. Long-Short Term Memory (LSTM) [57]). These approaches were shown to be highly effective at deciphering complex regulatory mechanisms such as alternative splicing [78], transcriptional regulation [2, 95, 122], RBP binding [93, 79], and RNA polyadenylation [35]. In those approaches, feature extraction and learning are combined in an end-to-end fashion that often yields better performance compared to conventional feature engineering approaches. The advantage of CNNs lies in their capability of performing automatic and parallel feature extraction by learning parameterized sequence motifs analogous to the position weight matrices (PWM) commonly used in classical sequence analysis algorithms. LSTMs, on the other hand, are more suitable for analyzing sequential data to discover correlations between different positions, allowing to capture sequence context and cooperative binding.

To our knowledge, no computational predictor of mRNA subcellular localization exists to date. This is the challenge we tackle in this paper. We introduce, evaluate, and interpret RNATracker, a deep neural network predictor of subcellular localization combining two convolutional layers, a bidirectional LSTM layer, and an attention module. Although the architecture of our model has some similarities with previously proposed approaches [95, 93, 79], mRNA subcellular localization differs from most previous applications of deep learning to biological sequence function prediction in several aspects that make it particularly challenging. First, the process of subcellular localization is a long chain of complex events mediated by a large number of protein-RNA and RNA-RNA interactions, and may depend on both primary sequence and secondary structure. Second, our goal is to learn a multi-output function that predicts the expression distribution of a given transcript across several cellular fractions, instead of a single positive/negative label. Third, most mRNAs only exhibit a moderate degree of subcellular asymmetry, and experimental measurements are somewhat noisy and potentially biased. Finally, transcripts have greatly variable lengths, an issue generally not encountered in previous applications.

In this paper, we introduce the RNATracker model and demonstrate its superior ability to predict subcellular localization on two recently published data sets obtained by CeFra-seq [13] and APEX-RIP [61]. We then dissect the trained models to learn new biology about the mechanisms involved. Finally, we use a sliding window masking strategy to identify the regions most likely to be conferring the observed localization pattern, and present evidence in support of the regulatory function of those regions.

4.4 Methods

The goal of RNATracker is to predict an mRNA's subcellular localization profile from its sequence alone (including possibly its secondary structure inferred from the sequence). To this end, we designed a convolutional bidirectional Long Short-Term Memory (LSTM) neural network with attention mechanism, inspired from previous work on the prediction of protein-mRNA interactions [2, 93, 79] and DNA function [95]. Here, we introduce the methodological aspects of training data, feature encoding, model architecture, training, and evaluation.

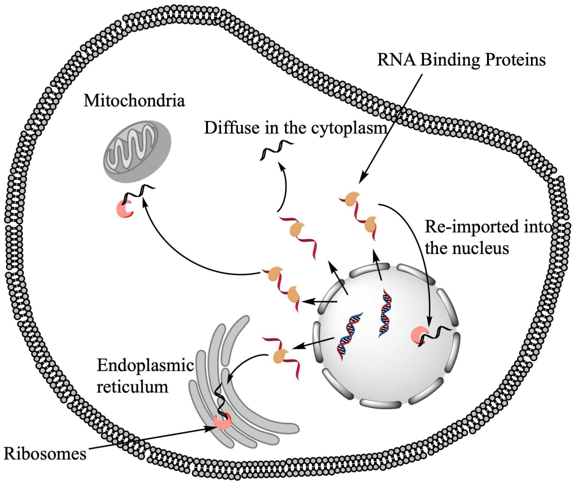

Figure 4–1: Schematic representation of RNA trafficking mechanisms and outcomes in eukaryotes.

Subcellular localization data

Messenger RNA subcellular localization data was obtained from CeFra-Seq [13] and APEX-RIP [61] experimental data, in the form of normalized expression values (FPKM) for each annotated human protein-coding gene. The first data set covers four subcellular fractions ($\mathcal{F}$ = {cytosol, nuclear, membranes, insoluble}), whereas the second one identified transcripts enriched in a different set of compartments ($\mathcal{F}$ = {endoplasmic reticulum, mitochondrial, cytosol, nuclear}). Although FPKM normalisation can sometimes distort relative expression values across samples, this was not a major concerned here because most genes had similar expression across fractions.

We averaged replicates and excluded genes with low total expression, keeping only those whose total FPKM expression across all fractions exceeds 1. This resulted in a set of 11,373 localization-annotated transcripts in the CeFra-Seq dataset and

13,860 in the APEX-RIP dataset. Let $e(g, f)$ denote the expression level of gene g in fraction $f \in \mathcal{F}$, expressed in FPKM. The normalized localization value for gene g in fraction $f \in \mathcal{F}$ was defined as $loc(g, f) = \frac{e(g,f)}{\sum_{f' \in \mathcal{F}} e(g,f')}$, which measures the relative abundance of g in each fraction.

Sequences and RNA secondary structure

Messenger RNA sequences were downloaded from the Ensembl database [1], keeping only the longest protein-coding isoform. We inferred RNA secondary structure information for each transcript using RNAplfold [16] (window size=150, span=100). The output of RNAplfold, which is a list of base pairing probabilities, are converted to an intermediate dot-bracket annotation by greedily creating as many nested basepairs as possible. The resulting predicted structure was parsed using the forgi library [63], part of the Vienna RNA package [82], to annotate each position as belonging to an internal loop (I), hairpin loop (H), multi-loop (M), dangling start (F), dangling end (T) or stem (S).

Feature encoding

RNA nucleotides are represented using 1-hot encoding over 4 bits. When RNA secondary structure is considered, a 6-bit encoding of the structural state is used, or a 24-bit encoding of the joint representation of sequence and structural states.

Input sequence length varies from ∼200 nt to more than 30,000 nt. RNATracker can either operate on individual input sequence of arbitrary lengths, or on fixed length inputs, the latter allowing a variety of mini-batch optimizations and normalizations. In the fixed-length mode, sequences longer than 4000 nt are truncated at the 5' end (working under the assumption that localization signals are more often found in a transcript's 3' end [15]). Sequences shorter than 4000 nt are left-padded with empty nucleotides encoded as 0000. We also investigated fixing the length at 1000, 2000, and 8000 nt, but obtained reduced prediction accuracy at 1000 and 2000 nt, and little accuracy benefits at 8000 nt.

Model Architecture

RNATracker is a convolutional neural network (CNN) coupled with a Long Short-Term Memory (LSTM) recurrent neural network with attention mechanism. The overall structure of our model structure is shown in Figure 4–2. Each component is described in details below.

Our network includes two sets of CNN+pooling layers (Figure 4–2 A). Each CNN layer consists of 32 convolutional filters of length 10 with ReLU activation, initialized with Xavier uniform. Each pooling layer takes a window of size 3 and a stride of 3, to aggregate local information along the sequence as well as to effectively downsample the sequence by a factor of roughly 9 before passing it on to the subsequent LSTM layers. A network with a single convolutional layer was also evaluated but proved less accurate.

The output of CNN+pooling layers is fed into the subsequent LSTM layer (Figure 4–2 B), which is a recurrent neural network that allows information to flow from positions to position, while being updated based on the data at the current position,

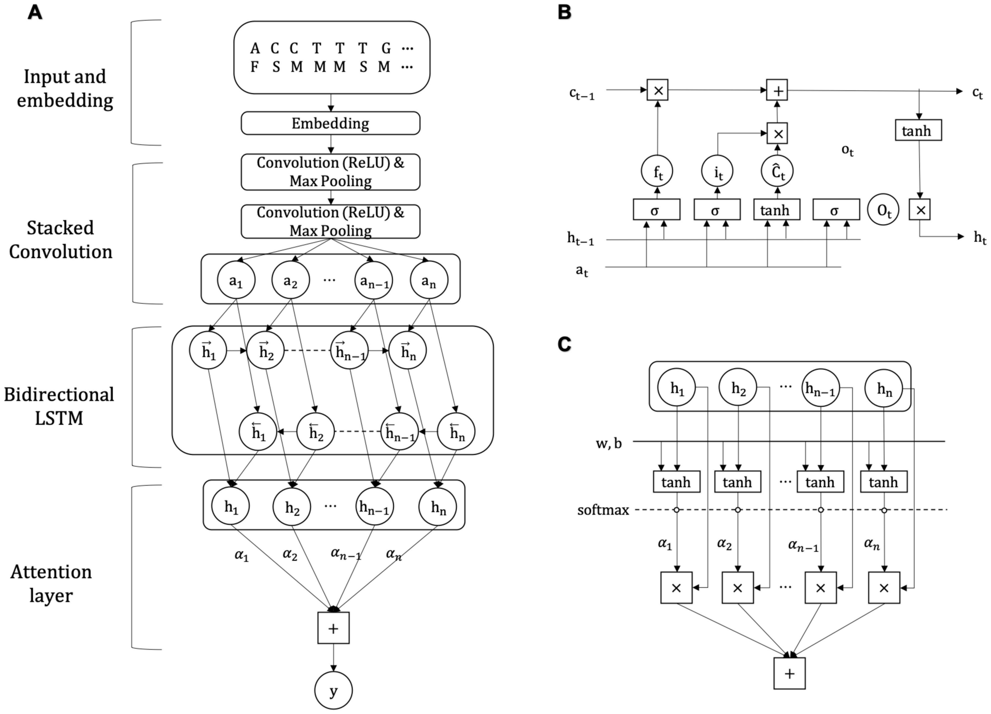
A
Input and embedding
A C C T T T G ···
F S M M M S M ···
Embedding
Stacked Convolution
Convolution (ReLU) & Max Pooling
Convolution (ReLU) & Max Pooling
a_1 a_2 ··· a_{n-1} a_n
Bidirectional LSTM
$\vec{h}_1$ $\vec{h}_2$ $\vec{h}_{n-1}$ $\vec{h}_n$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_{n-1}$ $\overleftarrow{h}_n$
Attention layer
h_1 h_2 ··· h_{n-1} h_n
α_1 α_2 α_{n-1} α_n
+
y
B
c_{t-1}
c_t
×
+
tanh
o_t
f_t i_t $\hat{C}_t$
σ σ tanh σ O_t
h_{t-1}
h_t
a_t
C
h_1 h_2 ··· h_{n-1} h_n
w, b
tanh tanh ··· tanh tanh
softmax
α_1 α_2 α_{n-1} α_n
× × ··· × ×
+

Figure 4–2: Structure of the RNATracker deep neural network. (A) Top-down model architecture, from the feature encoding, convolution and LSTM layers to the attention module. (B) Details of a LSTM cell. (C) Details of the attention module employed in this study.

according to the following equations:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{C}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \circledcirc \begin{pmatrix} W_i \\ W_f \\ W_o \\ W_c \end{pmatrix} [h_{t-1}, x_t] + \begin{pmatrix} b_i \\ b_f \\ b_o \\ b_c \end{pmatrix} \tag{4.1}$$

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \tag{4.2}$$

$$h_t = o_t * \tanh(C_t) \tag{4.3}$$

where i_t, f_t and o_t denote the input, forget and output gate respectively, each as an independent function of previous cell output h_{t-1} and input to the current cell x_t. C_t is the cell memory, composed in part of $\hat{C}_t$ which is the candidate cell memory for time step t, whose element-wise multiplication with the input gate i_t determines how much information to update into the current cell memory C_t. Similarly f_t controls how much information to forget from previous cell memory C_{t-1}, therefore $f_t * C_{t-1}$ makes up the other part of C_t. Finally o_t controls the information of the current cell output h_t. $\circledcirc$ stands for component-wise function composition.

The use of bidirectional LSTM has previously been shown to be advantageous compared to ordinary unidirectional LSTM, since they are able to aggregate information from both directions [104]. Our network includes both a forward (5' to 3') and a reverse (3' to 5') direction LSTM. For each time step, the output of the bidirectional LSTM is the concatenation of the outputs of the two directional LSTMs.

Attention Mechanism

Based on previous studies [23], we expect the localization signals contained within most mRNAs to be confined to a relatively short contiguous portion of the sequence, often (but not always) located in the 3' UTR. To take advantage of this, RNATracker integrates the notion of attention mechanism [6], which is a popular add-on technique for multiple tasks in fields such as document classification [118] and relation classification [123]. This allows RNATracker to learn to pay more attention to regions of the sequence that convey more relevant information about localization. The details of the attention module are shown in Figure 4–2 C.

Let us denote output of the bidirectional LSTM layer at time step t as $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The attention layer performs the following computation:

$$s_t = \tanh(w \cdot h_t + b) \tag{4.4}$$

$$\alpha_t = \frac{exp(s_t)}{\sum_{i=1}^{l} exp(s_i)} \tag{4.5}$$

$$c = \sum_{i=1}^{l} \alpha_i h_i \tag{4.6}$$

where w is a trainable weight vector in lieu of a context vector, l denotes the length of the output from the biLSTM layer, and c is the vector that summarizes the output at different time steps in h weighted by α_t.

Finally, we attach a fully connected layer with softmax activation after the attention module, to form a 4-categorical output.

Loss function and regularization

The entire network is trained to minimize the Kullback-Leibler divergence between the predicted and true subcellular distributions p and q:

$$KL(p, q) = \sum_{i}^{N} \sum_{j \in \mathcal{F}} p_{ij} \ \log \frac{p_{ij}}{q_{ij}}$$

where N is the size of batch, and p is the observed distribution of normalized localization values across the subcellular fractions. Regularization is achieved using dropout units after convolutional layers, with a ratio empirically determined at 0.2.

When using fixed-length input sequences, we use a mini-batch of size 256, which significantly speeds up training. We have investigated the use of batch normalization [58], which in other contexts has been shown to speed up convergence. However, we observe that with our 5' zero-padding of short sequences, this leads to extra input variability being introduced at the 5' end when the sequences in the batch have unequal lengths, resulting in slightly decreased prediction accuracy. Therefore in practice we choose not to use batch normalization, which however would be worth considering if training efficiency is more of a concern, or in situations where input sequences are of equal lengths.

The set of hyper-parameters reported in this study are selected based on the previous literature [93, 79] and subject to a small amount of manual tuning. Overall, we found our model robust to the choice of reasonable hyper-parameters.

Use of RNA secondary structure

To assess the extent to which RNA secondary structure can be used to inform subcellular localization prediction, we trained three variants of RNATracker: (i) RNATracker$_{\text{seq}}$ uses only primary sequence information; (ii) RNATracker$_{\text{seq}\times\text{struct}}$ represents sequence and structure information jointly using 1-hot encoding over $4 \times 6 = 24$ bits/nt; and (iii) RNATracker$_{\text{seq+struct}}$, which uses different encodings for the sequence and secondary structure, and processes them via different convolutional layers, whose outputs are concatenated before going through the LSTMs.

Training and evaluation

Our model is implemented using Keras [25]. Training uses the Adam optimizer with Nesterov momentum [40]. For all experiments we used 10-fold cross-validation to evaluate our models. A maximum of 100 epochs is used for training each fold, and a validation set consisting of 10% of the training data is used to monitor the loss in the training process to detect overfitting.

The variable length of mRNA transcripts poses a unique challenge to this study in terms of training time, as this prevents the use of mini-batches. Training examples thus need to be presented one at a time, which results in slow training (7 days for 10-fold cross-validation on a single GTX1080Ti graphic card, using a learning rate of 10^{-4}). Skipping the LSTM layers allows somewhat faster training (2 days), but at a small cost in terms of accuracy (see Results). Sequence truncation/padding to 4 kb allows batch training, which yields significant gains in training time (8 hours for 10-fold cross-validation, with a learning rate of 10^{-3}).

Baseline predictors

Since we are not aware of any previous work on the prediction of mRNA subcellular localization, we chose to compare the different versions of RNATracker to two baseline predictors based on the popular k-mer representation. The simplicity of k-mer based approach stems from the fact that the ordering information is lost in this representation. However, it has proved effective for related types of sequence function prediction, such as transcription factor binding [45]. Here, we use a feature vector of k-mer counts that combines features from 1-mer to 5-mer extracted from the full RNA sequence, resulting in a 1367-dimensional input vector. We actually investigated going up to 7-mers, but obtained no benefit in terms of accuracy.

Two types of predictors were trained: a fully connected neural network (DNN-5Mer) with two hidden layers of size equal to the input dimension, each followed by ReLU activation and dropout, and a smaller neural network (NN-5Mer) with no hidden layer.

Locating zipcodes within individual transcripts

RNATracker can be used to quantify the extent to which specific subsequences of a given transcript contribute to the localization prediction, thereby identifying candidate zipcode elements. This is achieved by temporarily masking (zeroing-out) the sequence of a given portion of the transcript, and computing the Kullback-Leibler distance between RNATracker's localization predictions on the original and masked sequences. We use a mask of 100 nt and slide it (with 1 nt stride) along the transcript's sequence to obtain a relative importance vector. Because all the masked sequences have the same length, they can be evaluated in batch, which considerably speeds up the execution. We also experimented with another masking scheme where the masked portion is randomized rather than zeroed-out (100 repetitions), but this did not significantly change the results, while taking significantly longer. Therefore, the results presented here are for the zero-masking approach.

4.5 Results

The different versions of RNATracker were evaluated on two mRNA subcellular localization data sets. The first was obtained by CeFra-seq in HepG2 cells, and contains 11,373 transcripts analyzed in the nuclear, cytosolic, membranes, and insoluble fractions [13]. The second was produced using APEX-RIP on HEK 293T cells, and contains 13,860 analyzed in the endoplasmic reticulum, mitochondrial, cytosolic, and nuclear fractions [61]. Figure 4–3 shows the distribution of normalized localization values for each of the four CeFra-seq subcellular fractions, confirming the previously made observation that the cytoplasmic, nuclear, and insoluble fractions contain a larger number of strongly localized transcripts, compared to the membrane fraction.

Normalized localization values of different fractions are generally negatively correlated, except for the cytosolic and membrane fractions, which are unsurprisingly positively correlated due to physical colocation (Suppl. Fig. S2). This will have important consequences on the results presented later. Furthermore, transcripts localized to the cytosol tend to be shorter. See also Suppl. Fig. S4 and S3 for analogous analyses of APEX-RIP data.

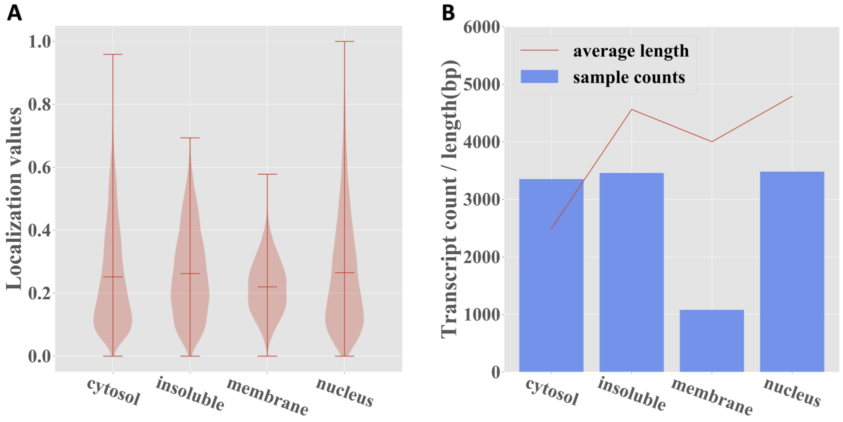

Figure 4–3: Summary statistics for the CeFra-Seq dataset. (A) Distribution of the normalized localization values for each subcellular fraction. (B) Number and average length of transcripts whose predominant localization is in each of the four fractions.

Performance of RNATracker

We used 10-fold cross-validation to evaluate the performance of the different versions of RNATracker and the two baseline k-mer profile predictors, on both the CeFra-seq and APEX-RIP data sets. To limit computational burden, more detailed analyses of some key model components such as the attention weights and the learned sequence motifs were performed exclusively on the CeFra-Seq dataset.

Figure 4–4 compares the true localization values to those predicted by RNATracker on the ceFra-seq dataset (see Suppl. Fig. S5 for analysis of the APEX-RIP dataset). Correlation coefficients obtained vary from 0.54 for the nuclear and membrane fractions to 0.705 for the cytoplasm faction, and all are significantly different from zero (p-value≈0). In APEX-RIP data, the accuracy is slightly lower, ranging from 0.456 (nuclear fraction) to 0.626 (endoplasmic reticulum), but again all are highly significant (p-value≈0).

Table 4–1 compares the Pearson correlation coefficients between the experimental and predicted localization values of the combined folds, obtained by different predictors. This reveals several observations. First, for both data sets and across all fractions, the best results are obtained using RNATracker applied to full-length sequences (i.e. no trimming/padding) and without RNA secondary structure information. These correlation coefficients are consistently 10 to 25% higher than those obtained by the k-mer based neural network, and 2-14% higher than those obtained by RNATracker operating on fixed-length sequences. Gains compared to fixed-length sequences are particularly significant for the membrane fraction (CeFra-seq) and endoplasmic reticulum (APEX-RIP), suggesting that localization to those fractions may often be mediated by sequences located in the 5' end of the transcript. This makes sense since targeting to the ER membrane is known to be mediated by the signal sequence that can be found in mRNAs encoding secreted proteins [53]. We also observe that the two variants using RNA secondary structure information consistently perform 1-3% worse than the version using sequence information alone (analysis only performed in the fixed-length setting, for running time reasons).

Our LSTM-based RNATracker was also compared to a pure CNN model (NoLSTM), revealing a consistent 3-7% increase in correlation coefficients due to the LSTM component. Similarly, a version of RNATracker without the attention module was evaluated but performed significantly worse than its attention-based counterpart (esp. on APEX-RIP data, where the difference ranges from 25 to 30%). These results show that both the LSTM and attention layers are essential for good prediction accuracy.

However, the significantly shorter training time makes the fixed-length training a viable alternative when resources are limited.

Table 4–1: Pearson correlation coefficients by subcellular fraction of various model and input settings. NoLSTM and NoAttention are the two ablation tests without the bidirectional LSTM or the attention module.

		Full-length RNA Inputs		Fixed-length Inputs (4 kb)				5Mer Inputs	
Dataset	Compartment	RNATracker$_{\text{seq}}$	NoLSTM	RNATracker$_{\text{seq}}$	NoAttention	Seq+Struct	Seq×Struct	DNN-5Mer	NN-5Mer
CeFra-Seq	Cytosol	**0.705**	0.676	0.685	0.625	0.666	0.652	0.637	0.558
	Insoluble	**0.641**	0.626	0.619	0.557	0.604	0.591	0.552	0.478
	Membrane	**0.540**	0.509	0.469	0.306	0.451	0.409	0.421	0.384
	Nuclear	**0.542**	0.515	0.502	0.379	0.475	0.449	0.485	0.432
APEX-RIP	ER	**0.626**	0.554	0.485	0.150	0.469	0.394	0.407	0.368
	Mitocondria	**0.482**	0.449	0.423	0.139	0.376	0.320	0.292	0.224
	Cytosol	**0.561**	0.522	0.501	0.259	0.493	0.423	0.446	0.363
	Nuclear	**0.456**	0.402	0.397	0.235	0.384	0.338	0.332	0.238

We next assessed the ability of RNATracker to identify the predominant localization of a given transcript, defined as the fraction where the transcript's expression is the highest. Instead of retraining RNATracker for this new classification task, we simply turned this regressor into a classifier by making it output the fraction with the highest predicted localization value. Suppl. Fig. S6 reports the receiver operating characteristic (ROC) and precision-recall (PR) curves for each predictor, micro-averaged across the four fractions. Consistent with the results on the regression task, RNATracker trained with full-length sequences slightly outperforms all

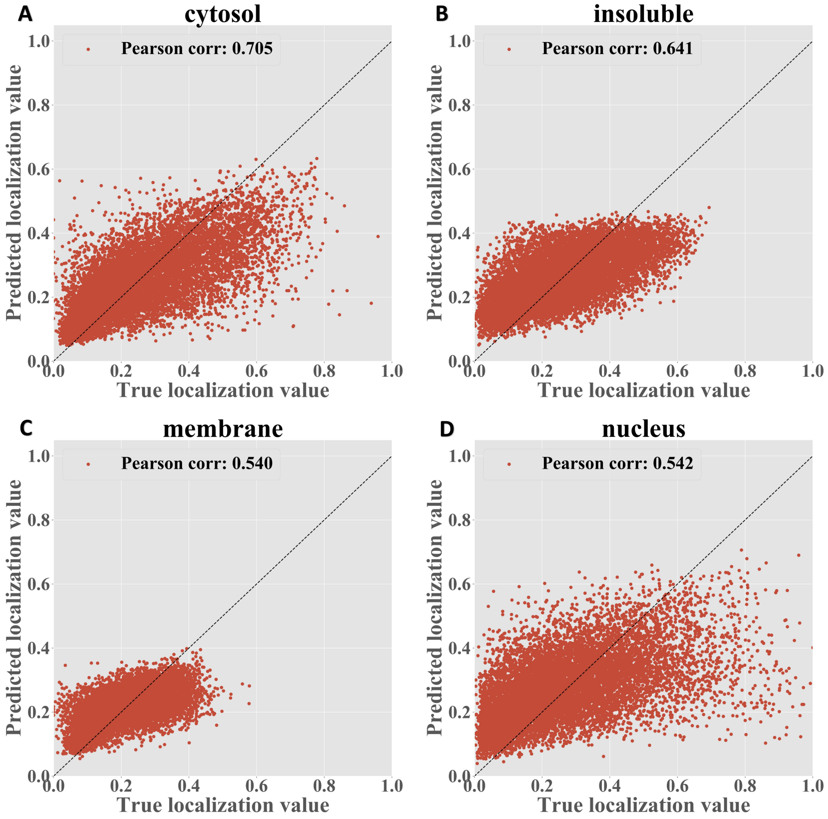

Figure 4–4: RNATracker$_{\text{seq}}$ predictions for the CeFra-Seq dataset by fractions, trained with full-length transcripts. Each point is a transcript with its true localization value shown on the x-axis and the predicted value shown on the y-axis.

other models, although by a narrow margin compared to the fixed-length version. These results also confirm the strong benefit of the attention module, and the slightly deleterious impact of including RNA secondary structure information. Similar observations can made for the APEX-RIP dataset (Suppl. Fig. S7).

To better illustrate the difference between various models, we used Delong's test from the R package pROC [101] to compare the ROC curves, confirming that the performance gain from fixed-length to full-length version is statistically significant (p-value $= 6.1\times10^{-9}$), and so are the benefits of the LSTM and the attention module (both p-values $< 2.2 \times 10^{-16}$).

Given its slightly superior performance, for the rest of this section, we focus analyzing RNATracker with full-length input sequences but no RNA secondary structure, and with LSTM and attention layers. Figure S6 (C) and (D) dissects the prediction performance per subcellular fraction. Consistent with correlation results previously shown in Figure 4–4, RNATracker has the best performance for the cytosolic fraction (ROC AUC = 0.851, PR AUC = 0.716), slightly better than results on the insoluble and nuclear factions, and much better than those on the membrane fraction. Several factors may explain these differences. First, very few transcripts ($\sim$1000) are predominantly found in the membrane fraction, and almost none have membrane localization value greater than 0.5 (see Figure 4–3 (A)). Second, transcripts predominantly localized to the cytoplasmic fraction tend to be significantly shorter than others (see Figure 4–3 (B)), which is a clue our predictor takes advantage of.

Dissecting the attention module

As demonstrated earlier, the attention mechanism is beneficial to predicting localization profiles. To better understand its role, we studied how the attention weights α_i vary along the sequence, under the fixed-length setting. Figure 4–5 shows that most of the attention weight concentrates at the $\sim$400 nt at the 3' end of the transcript. This is likely caused by two factors. First, the few well characterized cis-acting localization regulatory elements tend to be located in the 3' UTR [23], so it is likely that this is where the most meaningful signal is located. Second, the

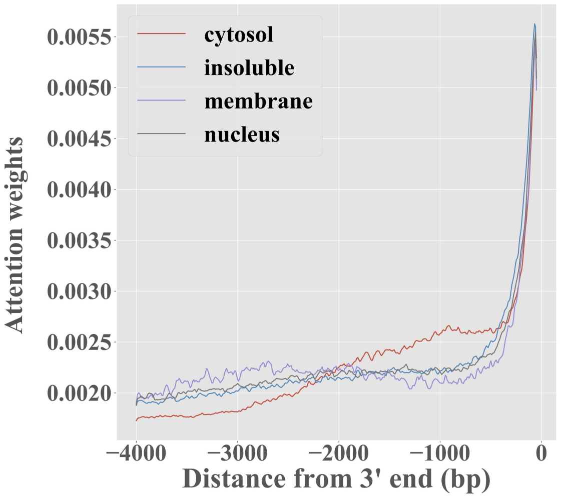

Figure 4–5: Attention weights α_i, for RNATracker with fixed-length inputs, averaged over the transcripts predominantly localized to each of the four fractions, as a function of position in transcript.

zero-padding introduced in transcripts shorter than 4 kb is always introduced at the 5' end, making this region generally less informative. It is worth noting, however, that RNATracker is fully able to identify zipcodes located outside that region (see Suppl. Fig. S1).

Analysis of sequence motifs

The weights learned by the 32 filters from the first CNN layer are akin to position-weight matrices used in classical sequence analysis. We used weblogo [34] to visualized the learned motifs, and Tomtom [8] to map learned motifs to binding preferences of known RBPs [98] (keeping in mind the caveat that this is an incomplete catalog and that matching motifs to RBPs is error-prone). 9 of the 30 convolutional filters were found to match a the binding profile of a known RBP (Tomtom pvalue

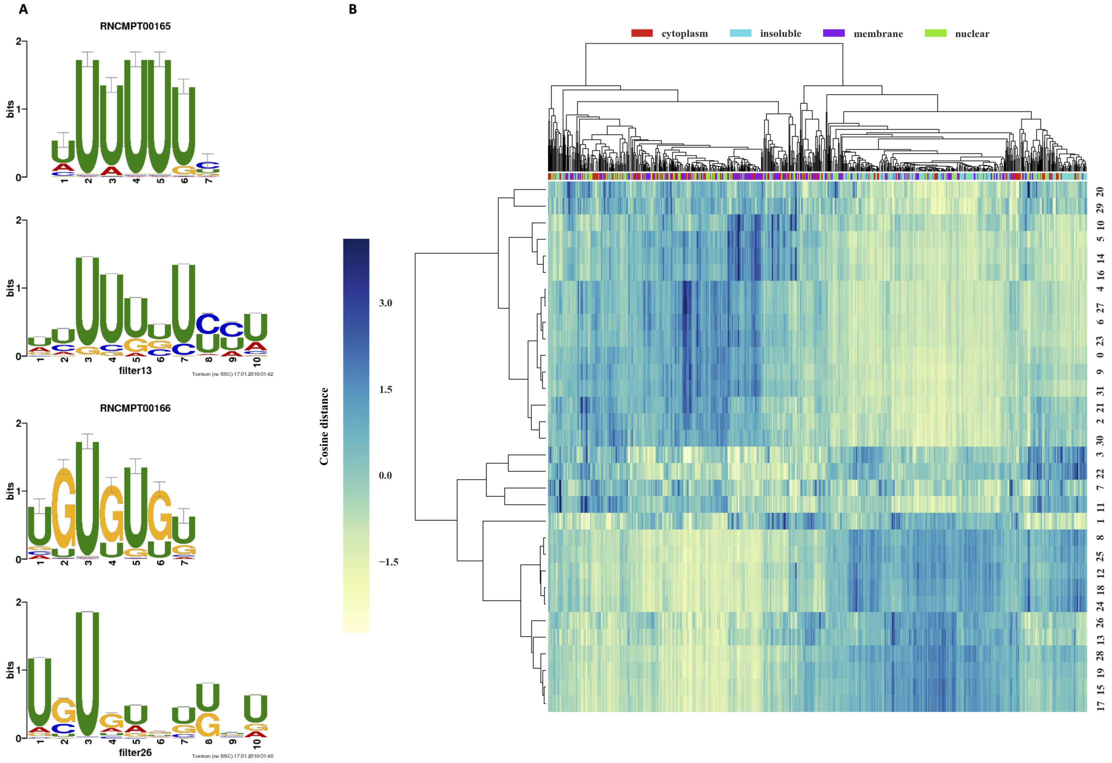

Figure 4–6: (A) Visualization of selected learned sequence motifs (above) mapped to those of known RBPs (below) from [98] that are TIA1 (up) and BRUNOL5 (down) . (B) Hierarchical clustering of 32 filters with 1024 strongly localized transcripts (256 transcripts per fraction), using the cosine distance between the 1024-dimensional vectors of average activation values, averaged across the transcript length.

< 0.05). Representative examples are shown in Figure 4–6 (A), with strong matches to RBPs TIA1 (p-value=7.63×10^{-4}) and BRUNOL5 (p-value=1.64×10^{-6}).

To better understand the role of the 32 motifs learned by RNATracker, and the way in which it combines them to obtain predictions, we clustered them based on their co-occurrences across a subset of 1024 transcripts consisting of the 256 transcripts most strongly localized to each of the four fractions. Two broad sets of motifs emerge. The first (top half of heatmap), contains several C/G-rich motifs as well as more complex motifs, which are strongly associated to cytoplasmic transcripts. The second (bottom half of heatmap), is characterized by A/U-rich motifs, as well as A-G or U-G dinucleotide repeats, which are mostly found in transcripts from the nuclear and insoluble fractions.

To study how RNATracker uses individual sequence motifs to obtain its localization predictions, we iteratively zeroed-out the output of all but one of the filters, and computed the Pearson correlation coefficient between the predicted localization values in the full and zeroed-out model, separately for each fraction. In this way, we are able to crudely isolate the contribution of each single convolution filter to the final prediction.

Locating zipcodes within transcripts

RNA subcellular localization is generally believed to be linked to the presence of discrete contiguous regulatory elements called localization zipcodes. By iteratively masking small portions of a transcript and studying how the predicted localization changes, one can identify candidate zipcodes, defined as regions whose masking significantly alters the localization prediction (see Methods and Figure S1 for examples on specific transcripts). A candidate zipcode can further be assigned an enhancing or repressive label for a given fraction, depending on whether its masking results in a reduction or increase in the predicted localization score for that fraction. Figure 4–7 shows the number of positive and negative zipcode regions identified at different stringency levels (KL cutoff). At the KL cutoff of 0.0075, we identify 374 unique positive zipcodes, but only 167 unique negative zipcodes.

Because the number of experimentally characterized zipcodes is very small (less than a dozen in human), we had to rely on indirect measures to assess the validity of the predicted zipcode elements. Due to their important role in regulating proper gene expression, we would expect most zipcodes to be under negative selection, and thus to be more highly conserved across species than their neighboring regions.

We thus used PhyloP conservation score [94], calculated from the multiple genome alignments of 100 vertebrates and available from the UCSC Genome Browser [50]. Focusing on the 2392 transcripts exhibiting strong subcellular localization (maximum localization value > 0.5), we compared the distribution of average PhyloP scores within the top 541 predicted zipcodes to the PhyloP score distribution of regions of 3' UTRs not predicted to be zipcodes (Figure 4–8). While the two distributions largely overlap, large conservation scores (>1) are roughly two times more frequent in candidate zipcodes than elsewhere, and the two distributions have means that are significantly different (p-value close to 0 using a Kolmogorov-Smirnov (KS) test). This shows that predicted zipcodes are under stronger negative selection than the rest of the 3' UTRs, although this may be caused by functions other than localization. Varying the KL threshold used to identify zipcodes, we observe that higher KS statistics (i.e. higher interspecies conservation values) are obtained for our most confidence predictions (Figure 4–7). With the caveat mentioned above, this suggests that RNATracker's KL score can be used as indicators of zipcode prediction reliability.

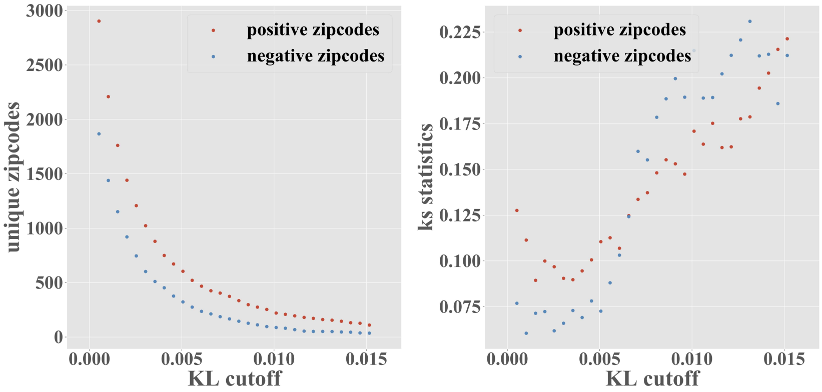

Figure 4–7: Number (left) and inter-species conservation (measured using the KS statistics (right) of enhancing and repressive candidate zipcode regions identified at increasingly strict KL cutoffs.

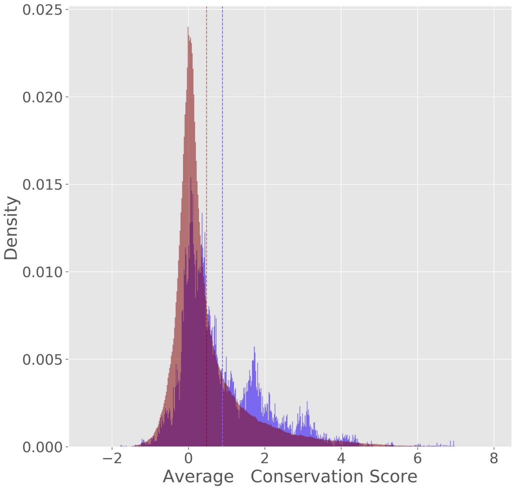

Figure 4–8: Distribution of average PhyloP scores for 541 regions predicted to be zipcode elements (KL score > 0.0076, in blue) and 3688436 regions predicted not to be (KL score ≤ 0.0076, in red). Dotted vertical lines indicate the means of the two distributions.

4.6 Discussion and Conclusion

Along with two recently published approaches ([124] and [48]), RNATracker is among the first computational predictors of mRNA subcellular localization. It achieves satisfactory (but certainly perfectible) performance on two of the largest subcellular localization data sets currently available, thanks to its use and adaptation of cutting-edge machine learning approaches such as LSTM and attention modules, without which prediction accuracy is generally inferior. Although the problem of predicting localization from sequence has some similarity to other sequence-based function prediction, its difficulty stands out because of the complexity of the mechanisms at play and the relative weakness and noisiness of the localization signal of most transcripts, among other reasons. The variable length of transcripts also leads to new challenges, both in terms of generalization and computational efficiency. Beyond being able to predict subcellular localization of full-length transcripts, RNATracker is able to locate candidate cis-regulatory regulatory regions (zipcodes) in strongly localized transcripts. In the absence of a large set of experimentally identified zipcodes, validating these predictions is challenging, but an analysis of inter-species sequence conservation, used a proxy for negative selection and thus function, indicates that many of our predicted zipcode are under stronger selection than surrounding 3' UTR regions.

Somewhat surprisingly, and despite our best attempts, we were unable to demonstrate significant benefits from the consideration of RNA secondary structure. This may be explained by a number of factors, and certainly does not suggest that structure plays no role in localization. First, our ability to accurately characterize secondary structure is imperfect, and our use of RNAplfold, which only considers relatively short-range interactions, may be limiting; the probabilistic structure profile proposed by [29] may be good alternative. Second, incorporating RNA structure information increases the size of the input feature space, from 4 bit per position for pure sequence, to 10 or 24 depending on whether the seq+struct or seq×struct encoding is used. This may more easily lead to overfitting, thereby negating the benefits of this potentially valuable information. More condensed encodings (e.g. paired/unpaired) may prove beneficial. Finally, rather than feeding as input precomputed structural information, one may consider letting the model learn to reconstruct them from some lower-level sequence/structural features.

Several factors may be limiting the accuracy of RNATracker. First and foremost, the quantity and specificity of RNA localization data remains relatively low, which limits the sophistication of the models learned from it and forces the use of strict regularization (limitation in model complexity, early stopping, dropout) to avoid too severe overfitting, which in turn limits the space of reasonable hyper-parameters. This is in part due to the fact that isoforms are currently not distinguished (all expression data is mapped to the longest annotated isoform), although this could be addressed by more advanced processing of future ceFra-seq/APEX-like data, provided higher sequencing depth is obtained. Second, localization data produced by ceFra-seq/APEX is inherently noisy and may sometimes inaccurately reflect a transcripts true localization. Combined with the fact that many transcripts exhibit only slightly asymmetrical localization or strong localization to more than one subcellular fraction, this makes for hard data to train from.

Improvements to our current approach could be considered in several directions, most of which are currently being explored. First, we may be able to take advantage of transfer learning to exploit models trained for other types of prediction tasks relevant to mRNA localization, such as the easier prediction of RBP binding [2, 93, 79] or possibly alternative splicing [78]. This would involve building a predictive model initialized from a model previously trained for one of these tasks, or re-using certain components of it, such as its convolution filters. Our initial attempts in that direction, based on re-using the convolutional filters trained to predict RBP binding events from Clip-Seq data [108], did not provide improved accuracy. Indeed, the convolution filters only take up a small proportion of all trainable weights. Alternatively, we could directly use prior knowledge about RBP binding affinities, e.g. from [98, 38], to initialize convolutional filters.

Second, in this study, we used inter-species conservation as an indirect valuation of our zipcode predictions. One could instead make direct use of this information as an input to the predictor or to its attention module.

Finally, bootstrapping techniques, e.g. reconstruction loss [99], can be integrated into the training to account for the noise of the targets, together with unlabeled RNA sequences.

With mRNA subcellular localization increasingly recognized as a key player in regulating gene expression, new and improved data sets will rapidly become available, and the power of approaches such as RNATracker will increase. At the same time, the predictions made by RNATracker, both in terms of location of zipcode elements and the way in which individual motifs combine to results in its localization predictions, constitute testable hypotheses that will fuel discovery in the field. All in all, this represents a rich, promising, and challenging area for future research in bioinformatics and machine learning.

Acknowledgements

We thank Faizy Ashan for insightful discussions, Louis Philip Benoit Bouvrette for providing assistance to the mapping of transcript-genome coordinates, and Jérôme Waldispühl for advice on RNA secondary structure prediction.

Funding

This work has been supported by a team grant from the Fond de Rechreche Québécois sur la Nature et les Technologies (FRQNT) to MB and EL, and a technology development grant from the Institut de Valorisation des Données (IVADO) to MB and EL.

CHAPTER 5
Conclusion

The main focus of this study revolves around the prediction of RNA subcellular localization and the identification of sequence determinants from the input RNA transcripts from a machine learning standpoint. It is important to gain a broader and clearer understanding into the underlying mechanisms, which has a strong implication over the regulation of cellular activities and subcellular protein distribution. Another important aspect is to fully understand the pathogenesis related to RNA localization, and to develop treatments that can for example take advantage of the existing RNA trafficking pathways.

Our proposed model, RNATracker, is able to achieve a good accuracy evaluated under two datasets collected from previous experiments. RNATracker is based on deep learning, performing end-to-end learning on RNA transcript sequences with or without annotated secondary structure. Due to the fact that RNATracker is heavily parameterized, in contrast to the limited quantity and quality of data, the model is susceptible to overfitting and only demonstrated limited prediction efficacy, despite its reliance on powerful deep learning tools such as CNN and LSTM.

A number of future directions are worth exploring, such as to more effectively incorporating the RNA secondary structural information into the prediction pipeline. It may be better to sample from an ensemble of all possible secondary structures instead of keeping only one that has the minimum free energy, and abstraction of the subgraph is also necessary to cope with excessively long RNA transcripts. A graph neural net representation of the RNA secondary structures would also enable a richer exploration of the structural motifs associated to RBP binding and RNA trafficking, possibly leading to more accurate predictions. On the other hand, to address the noise levels intrinsically involved in RNA localization and experimental protocols used to obtain the data, it is necessary to evaluate the model with larger and better dataset. A more carefully chosen neural architecture would also hypothetically suppress overfitting along with more and better training data.

It is also important to integrate the biological factors related to RNA subcellular localization into the design of the predictive model pipeline. For example alternative splicing should be taken into account to make a more biologically coherent gene-level localization prediction, instead of merely representing each gene with only its longest protein-coding RNA transcript.

Finally, having obtained a predictive model which associates a localization profile to each RNA transcript, the design of RNA zipcodes can be achieved from the opposite direction, which is to synthesize RNA oligos that possess some preferred localization target according to the predictive model. This inverse design operation can be aided by the deep generative framework described in the background section, and represents a whole new direction of RNA design.

Appendix

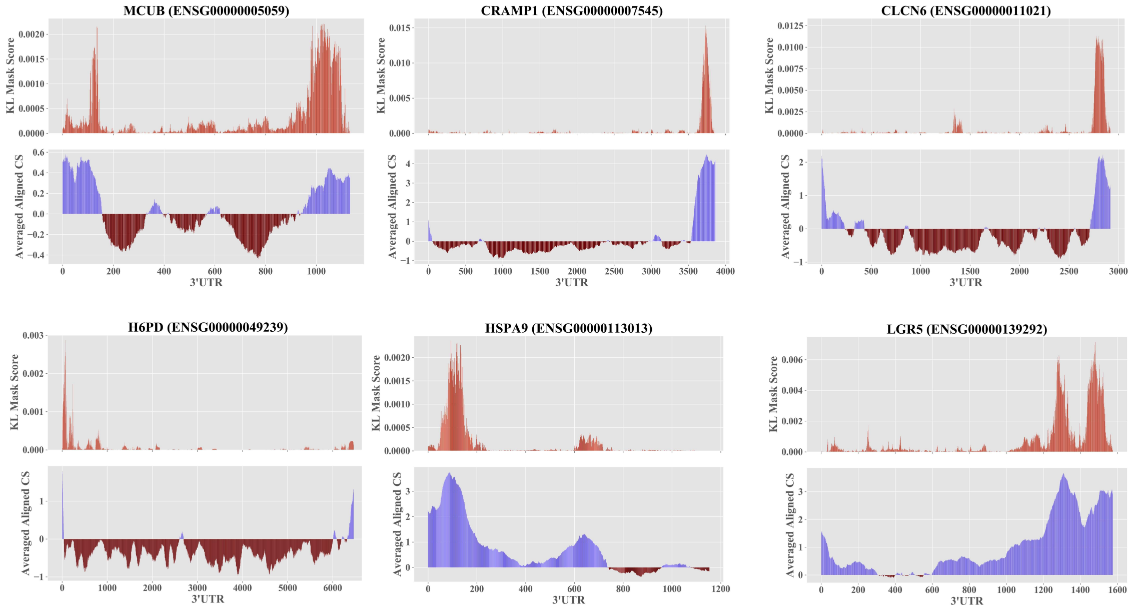

Figure S1: Alignments of KL scores and average conservation scores for 6 selected genes.

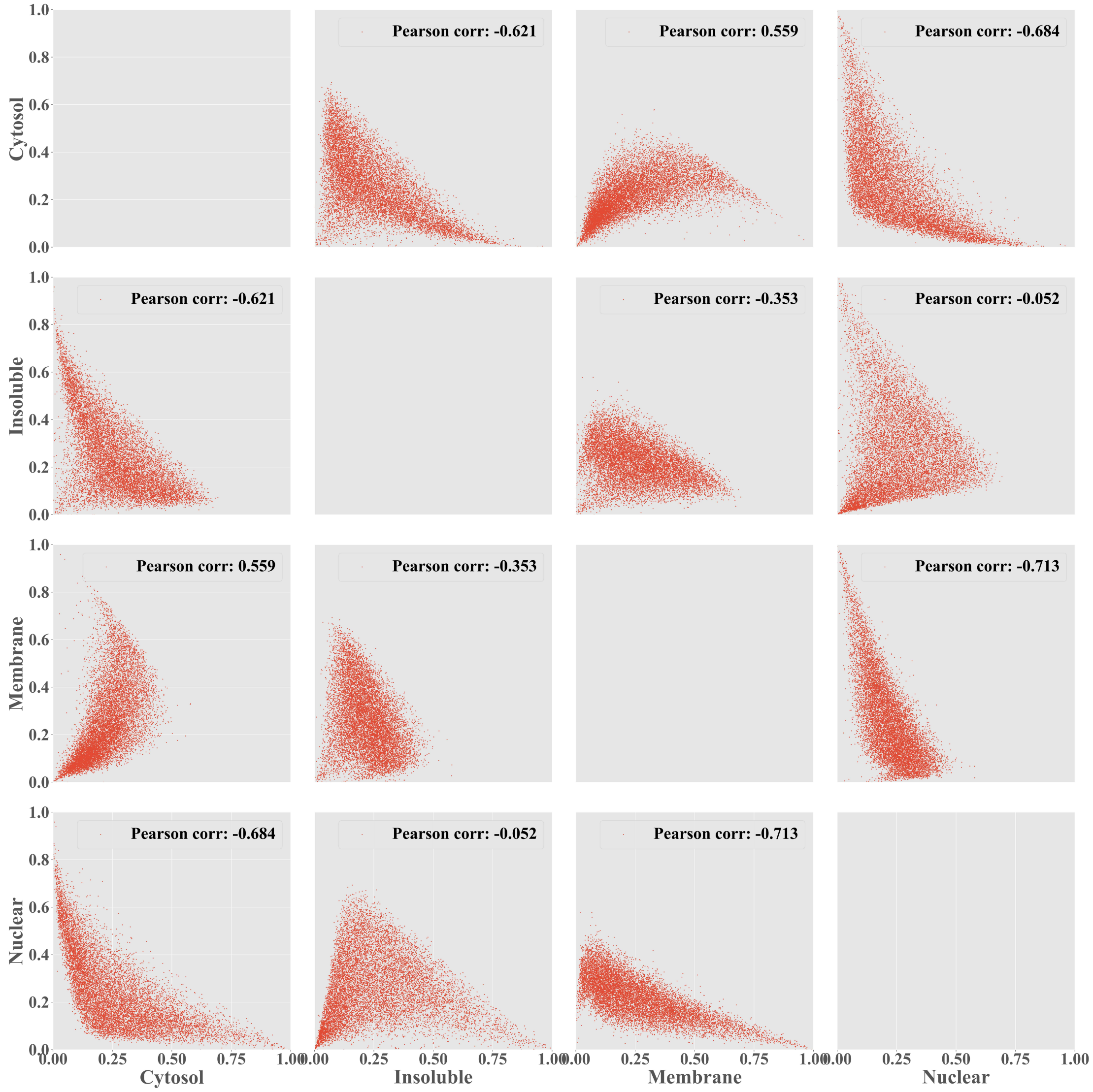

Figure S2: Pairwise joint distribution of the normalized expression values between different subcellular fractions over all genes in the CeFra-Seq dataset.

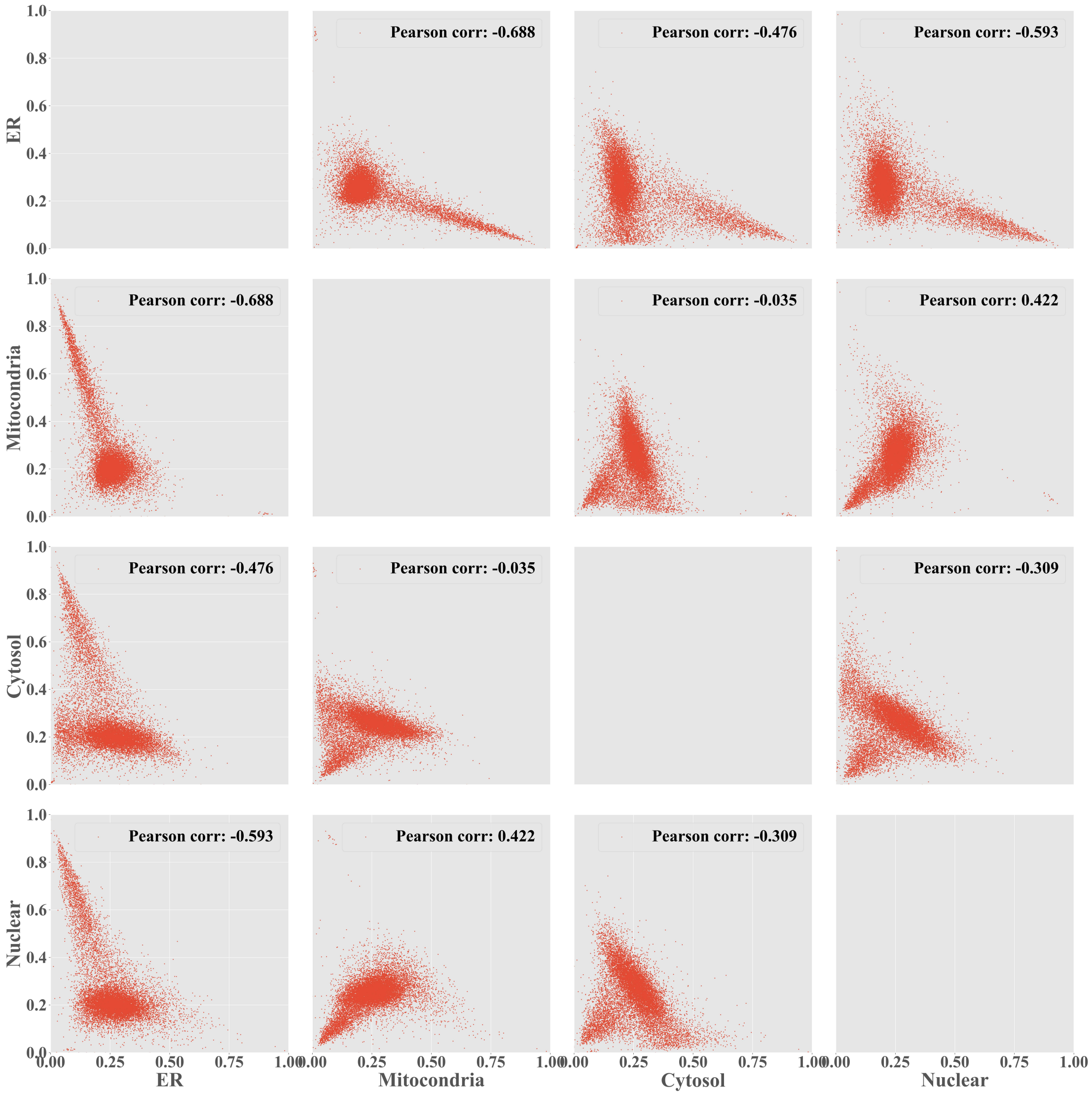

Figure S3: Pairwise joint distribution of the normalized expression values between different subcellular fractions over all genes in the APEX-RIP dataset.

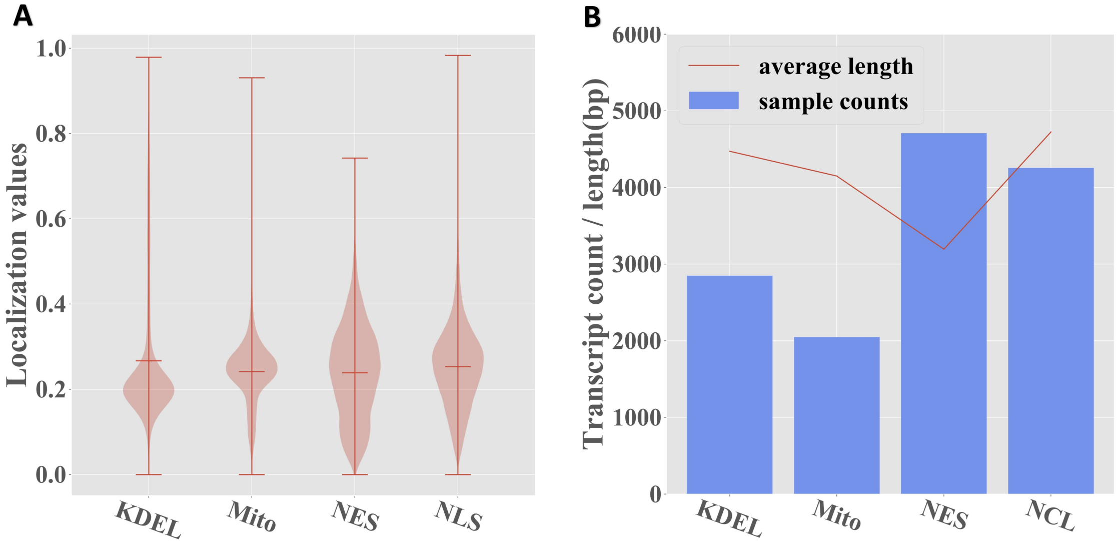

Figure S4: Distribution of the localization values for the APEX-RIP dataset. KDEL: endoplasmic reticulum, Mito: mitochondria, NES: cytoplasm, NLS: nucleus.

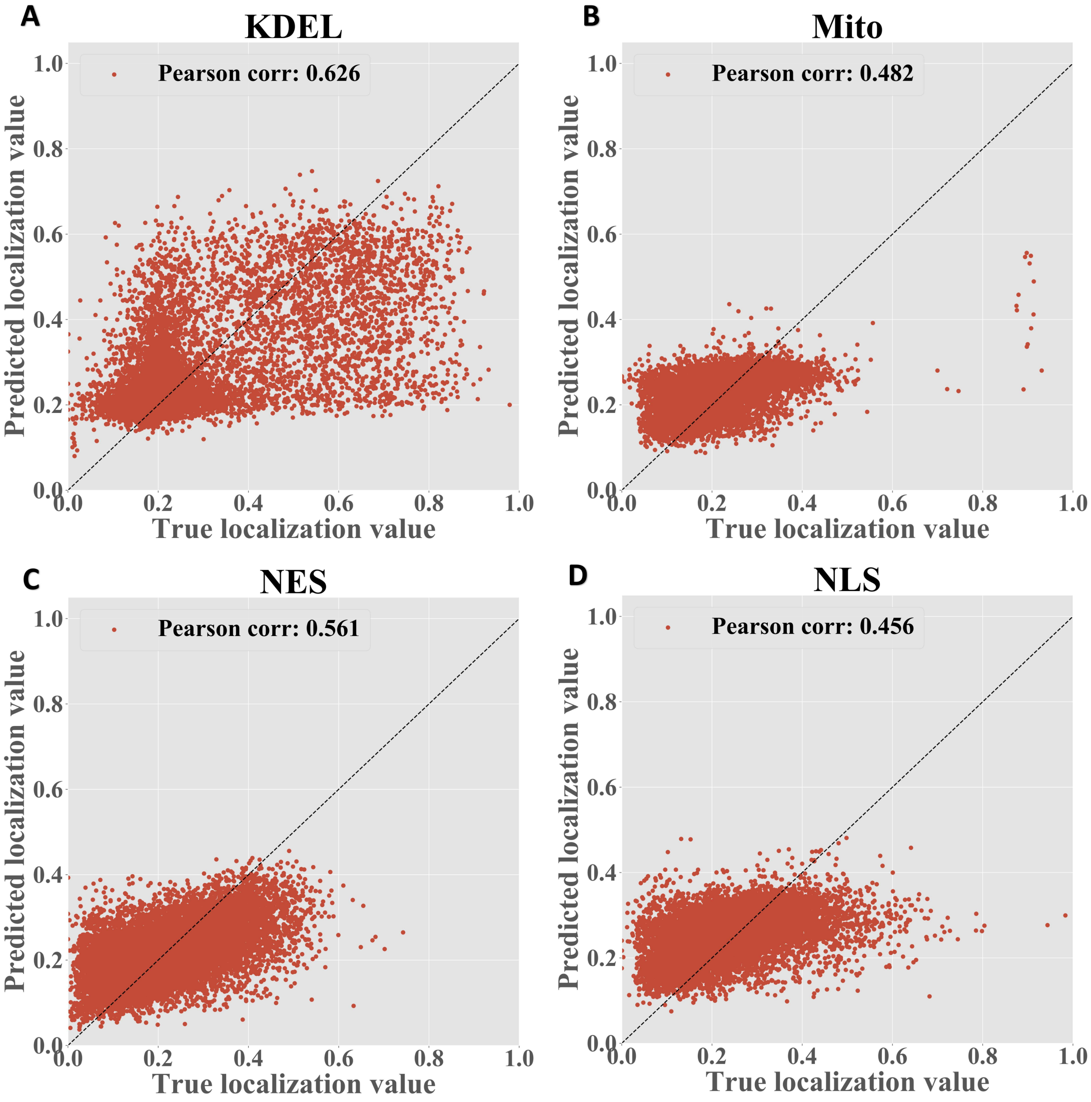

Figure S5: Scatter plots for RNATracker applied to full-length sequences in the APEX-RIP dataset.

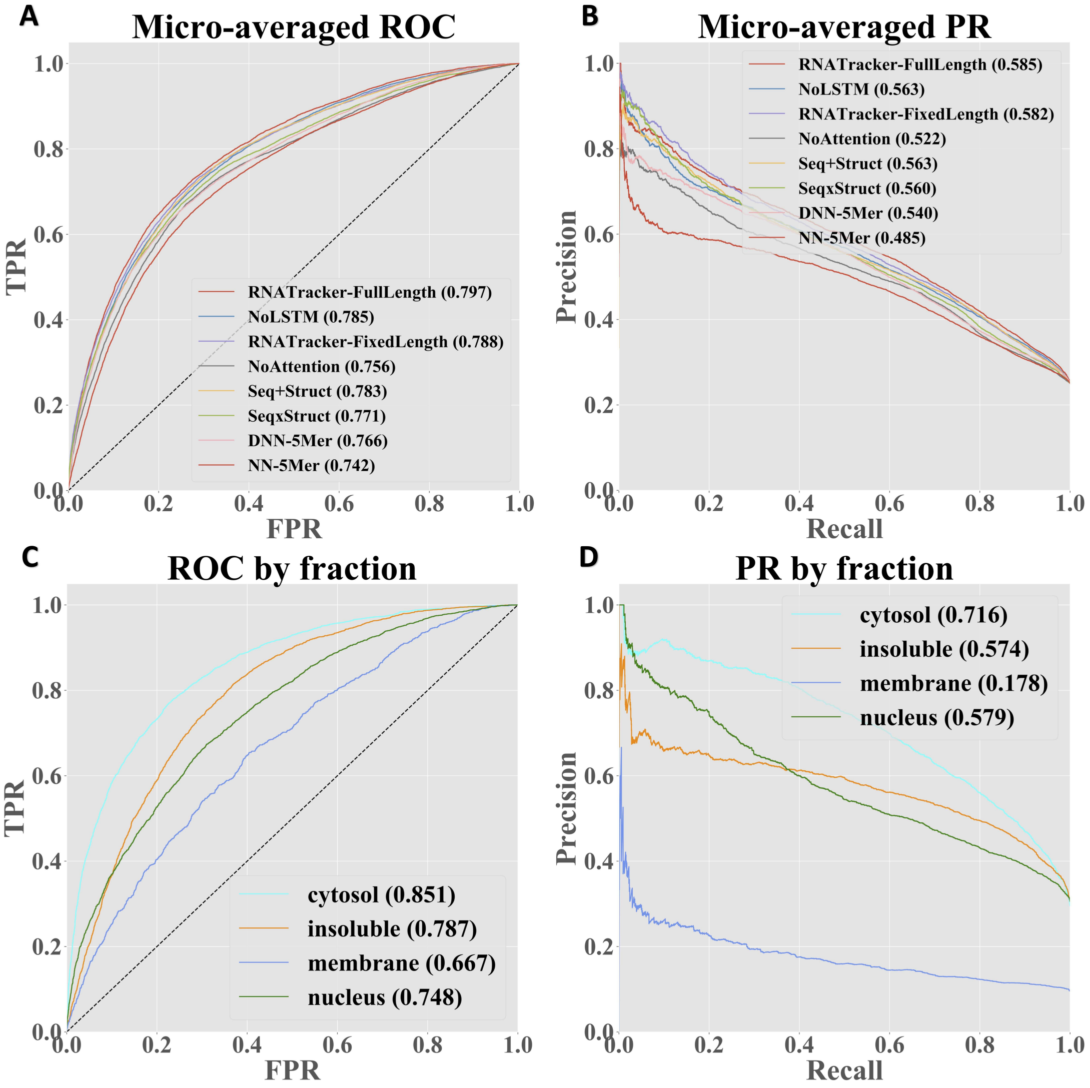

Figure S6: Analysis of RNATracker variants evaluated on their ability to predict the predominant localization of a transcript. (A) and (B) present the micro-averaged ROC and PR curves for the four fractions of the different modalities of RNATracker and baselines. (C) and (D) ROC curves and PR curves for RNATracker$_{\mathrm{seq}}$, for each fraction.

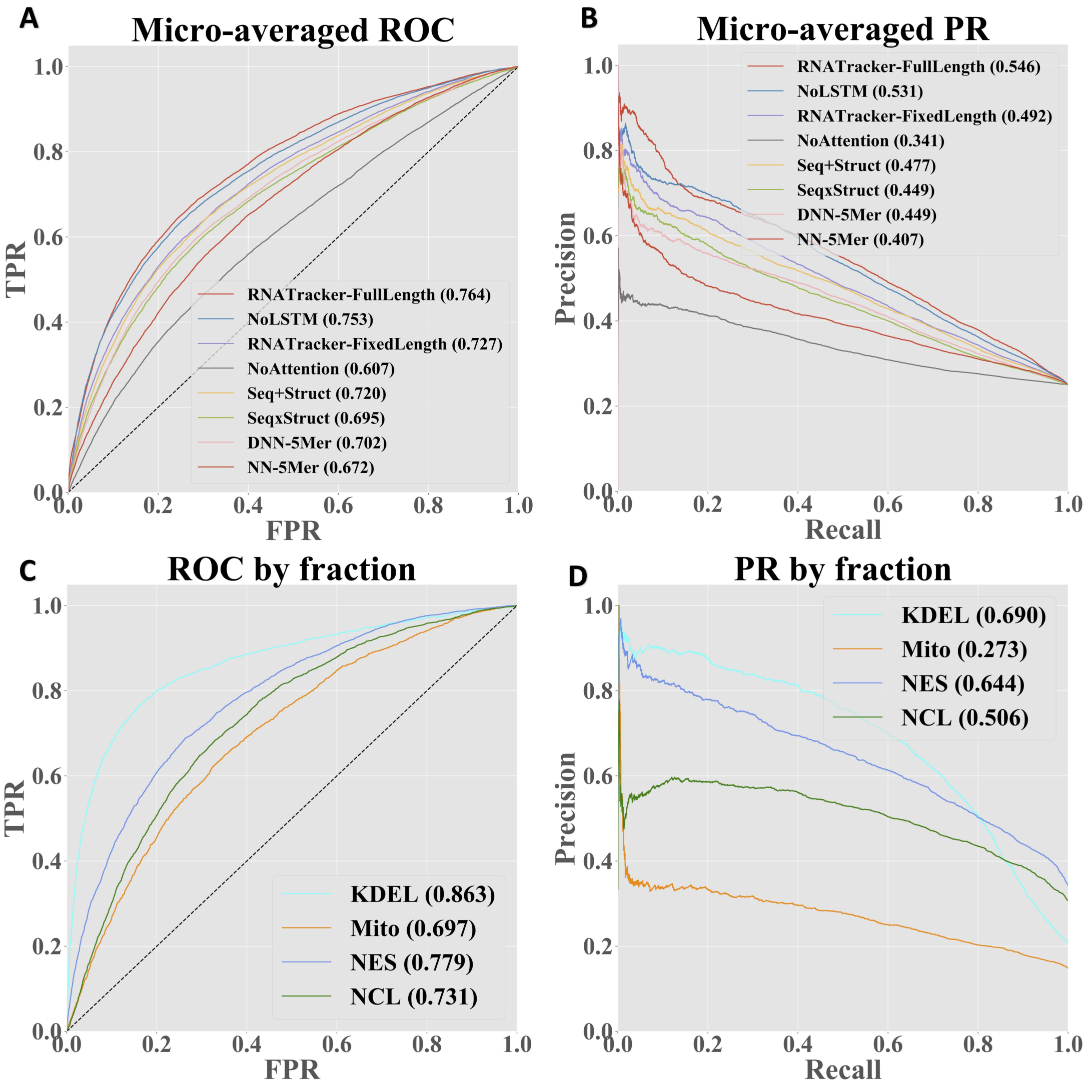

Figure S7: (A) and (B) compared the micro-averaged ROC and PR curves of different RNATracker modalities and baselines. (C) and (D) presents ROC curve and PR curve on a fraction basis, for the best performing RNATracker model trained with full-length sequences.

References

[1] Bronwen L Aken, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, et al. Ensembl 2017. *Nucleic Acids Research*, 45(D1):D635–D642, 2016.

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

[3] V. Ambros. The functions of animal microRNAs. *Nature*, 431(7006):350–5, 2004.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[5] D Ayers. Long non-coding RNAs: novel emergent biomarkers for cancer diagnostics. *J Cancer Res Treat*, 1(2):31–35, 2013.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[7] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–73, 2006.

[8] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl_2):W202–W208, 2009.

[9] J. Baleriola, C. A. Walker, Y. Y. Jean, J. F. Crary, C. M. Troy, P. L. Nagy, and U. Hengst. Axonally Synthesized ATF4 Transmits a Neurodegenerative Signal across Brain Regions. *Cell*, 158(5):1159–1172, 2014.

[10] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, 2004.

[11] Arash Bashirullah, Susan R Halsell, Ramona L Cooperstock, Malgorzata Kloc, Angelo Karaiskakis, William W Fisher, Weili Fu, Jill K Hamilton, Laurence D Etkin, and Howard D Lipshitz. Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster. *The EMBO journal*, 18(9):2610–2620, 1999.

[12] L. P. Benoit Bouvrette, N. A. L. Cody, J. Bergalet, F. A. Lefebvre, C. Diot, X. Wang, M. Blanchette, and E. Lecuyer. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells. *RNA*, 24(1):98–113, 2018.

[13] Louis Philip Benoit Bouvrette, N. A. L. Cody, Julie Bergalet, Fabio Alexis Lefebvre, Cédric Diot, Xiaofeng Wang, Mathieu Blanchette, and Eric Lécuyer. Cefra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in drosophila and human cells. *RNA*, 24(1):98–113, 2018.

[14] J. Bergalet and E. Lecuyer. The functions and regulatory principles of mRNA intracellular trafficking. *Adv Exp Med Biol*, 825:57–96, 2014.

[15] J. Bergalet and E. Lécuyer. The functions and regulatory principles of mRNA intracellular trafficking. *Adv Exp Med Biol*, 825:57–96, 2014.

[16] Stephan H Bernhart, Ivo L Hofacker, and Peter F Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, 2005.

[17] E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–6, 2001.

[18] D. M. Bhatt, A. Pandya-Jones, A. J. Tong, I. Barozzi, M. M. Lissner, G. Natoli, D. L. Black, and S. T. Smale. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, 150(2):279–90, 2012.

[19] Clive R Bramham and David G Wells. Dendritic mRNA: transport, translation and function. *Nature Reviews Neuroscience*, 8(10):776, 2007.

[20] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[21] Z. Cao, X. Pan, Y. Yang, Y. Huang, and H. B. Shen. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics*, 34(13):2185–2194, 2018.

[22] A. Chin and E. Lecuyer. RNA localization: Making its way to the center stage. *Biochim Biophys Acta Gen Subj*, 1861(11 Pt B):2956–2970, 2017.

[23] Ashley Chin and Eric Lecuyer. RNA localization: Making its way to the center stage. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 2017.

[24] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[25] François Chollet et al. Keras, 2015.

[26] Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[27] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Research*, 39(Database issue):D301–8, 2011.

[28] Kate B Cook, Hilal Kazan, Khalid Zuberi, Quaid Morris, and Timothy R Hughes. Rbpdb: a database of RNA-binding specificities. *Nucleic Acids Research*, 39(suppl_1):D301–D308, 2010.

[29] Kate B Cook, Shankar Vembu, Kevin CH Ha, Hong Zheng, Kaitlin U Laverty, Timothy R Hughes, Debashish Ray, and Quaid D Morris. RNAcompete-S: Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection. *Methods*, 126:18–28, 2017.

[30] T. A. Cooper, L. Wan, and G. Dreyfuss. RNA and disease. *Cell*, 136(4):777–93, 2009.

[31] Thomas A Cooper, Lili Wan, and Gideon Dreyfuss. RNA and disease. *Cell*, 136(4):777–793, 2009.

[32] Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 255–262. Omnipress, 2010.

[33] F. Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561, 1970.

[34] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004.

[35] Andrew Delong, Michael K K Leung, and Brendan J Frey. Inference of the human polyadenylation code. *Bioinformatics*, 34(17):2889–2898, 04 2018.

[36] I. W. Deveson, M. E. Brunck, J. Blackburn, E. Tseng, T. Hon, T. A. Clark, M. B. Clark, J. Crawford, M. E. Dinger, L. K. Nielsen, J. S. Mattick, and T. R. Mercer. Universal Alternative Splicing of Noncoding Exons. *Cell Syst*, 6(2):245–255 e5, 2018.

[37] Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, 31(24):7280–301, 2003.

[38] Daniel Dominguez, Peter Freese, Maria S Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J Lambert, Eric L Van Nostrand, Gabriel A Pratt, et al. Sequence, structure, and context preferences of human RNA binding proteins. *Molecular Cell*, 70(5):854–867, 2018.

[39] M. Doyle and M. A. Kiebler. Mechanisms of dendritic mRNA transport and its role in synaptic tagging. *EMBO J*, 30(17):3540–52, 2011.

[40] Timothy Dozat. Incorporating Nesterov momentum into Adam. In *Proceedings of 4th International Conference on Learning Representations, Workshop Track*, 2016.

[41] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[42] Furqan M Fazal, Shuo Han, Pornchai Kaewsapsak, Kevin R Parker, Jin Xu, Alistair N Boettiger, Howard Y Chang, and Alice Y Ting. Atlas of subcellular RNA localization revealed by APEX-seq. *BioRxiv*, page 454470, 2018.

[43] Fabrizio Ferrè, Alessio Colantoni, and Manuela Helmer-Citterich. Revealing protein-lncRNA interaction. *Briefings in Bioinformatics*, 17(1):106–116, 2016.

[44] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829, 2014.

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

[46] C. Gong and L. E. Maquat. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 470(7333):284–8, 2011.

[47] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 27, 2014.

[48] Brian L Gudenas and Liangjiang Wang. Prediction of LncRNA subcellular localization with deep learning from sequence features. *Scientific Reports*, 8(1):16385, 2018.

[49] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[50] Maximilian Haeussler, Ann S Zweig, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Angie S Hinrichs, Jairo Navarro Gonzalez, et al. The ucsc genome browser database: 2019 update. *Nucleic Acids Research*, 47(D1):D853–D858, 2018.

[51] J. Hardy and D. J. Selkoe. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science*, 297(5580):353–6, 2002.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[53] Orith Hermesh and Ralph-Peter Jansen. Take the (RN)A-train: localization of mRNA to the endoplasmic reticulum. *Biochim Biophys Acta.*, 1833(11):2519–25, 2013.

[54] S. Heyne, F. Costa, D. Rose, and R. Backofen. GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*, 28(12):I224–I232, 2012.

[55] M. Hiller, R. Pudimat, A. Busch, and R. Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, 34(17), 2006.

[56] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

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

[58] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[59] R. P. Jansen and D. Niessing. Assembly of mRNA-protein complexes for directional mRNA transport in eukaryotes–an overview. *Curr Protein Pept Sci*, 13(4):284–93, 2012.

[60] P. Kaewsapsak, D. M. Shechner, W. Mallard, J. L. Rinn, and A. Y. Ting. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *Elife*, 6, 2017.

[61] Pornchai Kaewsapsak, David Michael Shechner, William Mallard, John L Rinn, and Alice Y Ting. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *eLife*, 6:e29224, 2017.

[62] H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, and Q. Morris. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*, 6:e1000832, 2010.

[63] Peter Kerpedjiev, Christian Höner Zu Siederdissen, and Ivo L Hofacker. Predicting RNA 3d structure using a coarse-grain helix-centered model. *RNA*, 21(6):1110–1121, 2015.

[64] Nathan Killoran, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan J Frey. Generating and designing dna with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.

[65] T. K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–U65, 2010.

[66] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[67] M. Kloc and L.D. Etkin. Two distinct pathways for the localization of RNAs at the vegetal cortex in Xenopus oocytes. *Development*, 121(2):287–297, 1995.

[68] Olaf Köhler, Dilip Venkatrao Jarikote, and Oliver Seitz. Forced intercalation probes (FIT Probes): thiazole orange as a fluorescent base in peptide nucleic acids for homogeneous single-nucleotide-polymorphism detection. *ChemBioChem*, 6(1):69–77, 2005.

[69] J. Krauss, S. Lopez de Quinto, C. Nusslein-Volhard, and A. Ephrussi. Myosin-V regulates oskar mRNA localization in the Drosophila oocyte. *Curr Biol*, 19(12):1058–63, 2009.

[70] S. Kummer, A. Knoll, E. Socher, L. Bethge, A. Herrmann, and O. Seitz. Fluorescence Imaging of Influenza H1N1 mRNA in Living Infected Cells Using Single-Chromophore FIT-PNA. *Angewandte Chemie-International Edition*, 50(8):1931–1934, 2011.

[71] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 1989.

[72] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[73] Eric Lécuyer, Hideki Yoshida, Neela Parthasarathy, Christina Alm, Tomas Babak, Tanja Cerovina, Timothy R Hughes, Pavel Tomancak, and Henry M

Krause. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131(1):174–187, 2007.

[74] F. A. Lefebvre, L. P. Benoit Bouvrette, J. Bergalet, and E. Lecuyer. Biochemical Fractionation of Time-Resolved Drosophila Embryos Reveals Similar Transcriptomic Alterations in Replication Checkpoint and Histone mRNA Processing Mutants. *J Mol Biol*, 429(21):3264–3279, 2017.

[75] F. A. Lefebvre, N. A. L. Cody, L. P. B. Bouvrette, J. Bergalet, X. Wang, and E. Lecuyer. CeFra-seq: Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing. *Methods*, 126:138–148, 2017.

[76] Fabio Alexis Lefebvre, N. A. L. Cody, Louis Philip Benoit Bouvrette, Julie Bergalet, Xiaofeng Wang, and Eric Lécuyer. Cefra-seq: Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing. *Methods*, 126:138–148, 2017.

[77] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[78] Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

[79] Shuya Li, Fanghong Dong, Yuexin Wu, Sai Zhang, Chen Zhang, Xiao Liu, Tao Jiang, and Jianyang Zeng. A deep boosting based approach for capturing the sequence binding preferences of RNA-binding proteins from high-throughput clip-seq data. *Nucleic Acids Research*, 45(14):e129–e129, 2017.

[80] Y. Liu, S. Sun, T. Bredy, M. Wood, R. C. Spitale, and P. Baldi. MotifMap-RNA: a genome-wide map of RBP binding sites. *Bioinformatics*, 33(13):2029–2031, 2017.

[81] Yu Liu, Sha Sun, Timothy Bredy, Marcelo Wood, Robert C Spitale, and Pierre Baldi. Motifmap-RNA: a genome-wide map of rbp binding sites. *Bioinformatics*, 33(13):2029–2031, 2017.

[82] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[83] D. Maticzka, S. J. Lange, F. Costa, and R. Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*, 15(1):R17, 2014.

[84] T. R. Mercer, S. Neph, M. E. Dinger, J. Crawford, M. A. Smith, A. M. Shearwood, E. Haugen, C. P. Bracken, O. Rackham, J. A. Stamatoyannopoulos, A. Filipovska, and J. S. Mattick. The human mitochondrial transcriptome. *Cell*, 146(4):645–58, 2011.

[85] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[86] S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36(12):1331–9, 2004.

[87] N. Multherjee, L. Calviello, A. Hirsekorn, S. de Pretis, M. Pelizzola, and U. Ohler. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nature Structural and Molecular Biology*, 24(1):86–96, 2017.

[88] E. Mus, P. R. Hof, and H. Tiedge. Dendritic BC200 RNA in aging and in Alzheimer's disease. *Proc Natl Acad Sci U S A*, 104(25):10679–84, 2007.

[89] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 29, 2016.

[90] L. M. Orre, M. Vesterlund, Y. Pan, T. Arslan, Y. Zhu, A. Fernandez Woodbridge, O. Frings, E. Fredlund, and J. Lehtio. SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol Cell*, 73(1):166–182 e7, 2019.

[91] F. Pan, S. Huttelmaier, R. H. Singer, and W. Gu. ZBP2 facilitates binding of ZBP1 to beta-actin mRNA during transcription. *Mol Cell Biol*, 27(23):8340–51, 2007.

[92] X. Y. Pan, P. Rijnbeek, J. C. Yan, and H. B. Shen. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *Bmc Genomics*, 19, 2018.

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

[94] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[95] Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107–e107, 2016.

[96] D. Ray, H. Kazan, E. T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B. J. Blencowe, Q. Morris, and T. R. Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7):667–70, 2009.

[97] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–7, 2013.

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

[99] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[100] B. Ren. TRANSCRIPTION Enhancers make non-coding RNA. *Nature*, 465(7295):173–174, 2010.

[101] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source

package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(1):77, 2011.

[102] H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–41, 2007.

[103] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Trans Neural Netw*, 20(1):61–80, 2009.

[104] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[105] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[106] P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

[107] T. Sterne-Weiler, R. T. Martinez-Nunez, J. M. Howard, I. Cvitovik, S. Katzman, M. A. Tariq, N. Pourmand, and J. R. Sanford. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res*, 23(10):1615–23, 2013.

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

[109] Z. D. Su, Y. Huang, Z. Y. Zhang, Y. W. Zhao, D. Wang, W. Chen, K. C. Chou, and H. Lin. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics*, 34(24):4196–4204, 2018.

[110] B. Suter. RNA localization and transport. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms*, 1861(10):938–951, 2018.

[111] Ilya Sutskever, Rafal Jozefowicz, Karol Gregor, Danilo Rezende, Tim Lillicrap, and Oriol Vinyals. Towards principled unsupervised learning. *arXiv preprint arXiv:1511.06440*, 2015.

[112] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[113] H. Tilgner, D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigo. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for IncRNAs. *Genome Research*, 22(9):1616–1625, 2012.

[114] V. Tripathi, J. D. Ellis, Z. Shen, D. Y. Song, Q. Pan, A. T. Watt, S. M. Freier, C. F. Bennett, A. Sharma, P. A. Bubulya, B. J. Blencowe, S. G. Prasanth, and K. V. Prasanth. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*, 39(6):925–38, 2010.

[115] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaž Ule, and Robert B Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215, 2003.

[116] I. Ulitsky and D. P. Bartel. lincRNAs: genomics, evolution, and mechanisms. *Cell*, 154(1):26–46, 2013.

[117] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[118] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[119] Joel K Yisraeli and DA Melton. The maternal mRNA Vg1 is correctly localized following injection into Xenopus oocytes. *Nature*, 336(6199):592, 1988.

[120] J. H. Yoon, K. Abdelmohsen, S. Srikantan, X. Yang, J. L. Martindale, S. De, M. Huarte, M. Zhan, K. G. Becker, and M. Gorospe. LincRNA-p21 suppresses target mRNA translation. *Mol Cell*, 47(4):648–55, 2012.

[121] T. Zhang, P. Tan, L. Wang, N. Jin, Y. Li, L. Zhang, H. Yang, Z. Hu, L. Zhang, C. Hu, C. Li, K. Qian, C. Zhang, Y. Huang, K. Li, H. Lin, and D. Wang. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res*, 45(D1):D135–D138, 2017.

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

[123] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212, 2016.

[124] B. Zuckerman and I. Ulitsky. Predictive models of subcellular localization of long RNAs. *RNA*, 25(5):557–572, 2019.