* Final product of some genes is RNA itself.
* However, most genes in a cell produce mRNA molecules that serve as intermediaries on the pathway to proteins.
* Cell converts info from mRNA to protein, this is translation
* Cells read this code using the ribosome
* DNA to RNA is like written text to typed text. RNA to protein is like translating to a different language.
* mRNA to protein is converted through something called the genetic code.
* Each group of 3 nucleotides is called a codon.
* Genetic code is uniform in most organisms except for a few differences in the DNA of mitochondria.
* Sequence of 3 is called codon, but it isn’t directly translated into an amino acid – tRNA (around 80 nucleotides) helps in converting.
* tRNA causes wobble base pairing between codons and anticodons (complementary codon), allowing codons to be translated into amino acids with upto one mismatch.
* tRNAs are covalently modified before they exit from nucleus.
* Some tRNA contains introns.
* Enzyme synthetase and and tRNAs are both important in converting codon to amino acid.
* Editing by RNA synthetase ensures accuracy in amino acid.
* Amino Acids Are Added to the C-terminal End of a Growing Polypeptide Chain.
* To maintain the correct reading frame and to ensure accuracy (about 1 mistake every 10,000 amino acids), protein synthesis is performed in the ribosome
* The mRNA nucleotide sequence is translated into an amino acid sequence using the tRNAs as adaptors to add each amino acid in the correct sequence to the end of the growing polypeptide chain.
* A ribosome contains four binding sites for [RNA](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5756/) molecules: one is for the [mRNA](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5496/) and three (called the A-site, the P-site, and the E-site) are for tRNAs
* A [tRNA](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5902/) [molecule](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5486/) is held tightly at the A- and P-sites only if its [anticodon](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A4829/) forms [base](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A4875/) pairs with a [complementary](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5012/) [codon](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A4997/) (allowing for wobble) on the mRNA molecule that is bound to the ribosome. The A- and P-sites are close enough together for their two tRNA molecules to be forced to form base pairs with adjacent codons on the mRNA molecule. This feature of the ribosome maintains the correct [reading frame](https://www.ncbi.nlm.nih.gov/books/n/mboc4/A4754/def-item/A5718/) on the mRNA.
* When a stop codon is encountered, the ribosome releases the finished protein, its two subunits separate again. These subunits can then be used to start the synthesis of another protein on another mRNA molecule.
* A different type of variation, sometimes called translational recoding, occurs in many cells.
* Another form of recoding is translational frameshifting. This type of recoding is commonly used by retroviruses, a large group of eucaryotic viruses, in which it allows more than one protein to be synthesized from a single mRNA.
* A stop codon at the end of the gag coding sequence can be bypassed on occasion by an intentional translational frameshift that occurs upstream of it.
* This frameshift occurs at a particular codon in the mRNA and requires a specific recoding signal.
* Chromosomes made of long strands of DNA
* On that strand, we have sequences called genes
* Each gene codes for proteins (sequences of DNA)
* Gene=DNA, protein=amino acids
* How do you go from genes(encoded in DNA) to protein (made up of polypetides, made of amino acids)
* Transcription- Gene to mRNA transcripts to RNA
* then mRNA to protein, this is translation, into a polypeptide sequence
* New amino acid attaches to polypeptide chain
* mRNA contains all the nucleotides- CAUCAUCAU etc.
* There's a ribosome sitting on this chain (say sitting on CAUCAU). This is RPF
* tRNA (anticodons) matches codon - and then translates to amino acid
* tRNA starts binding to A site- aminoacyl, then shifts the RPF sequence to the right while amino acid attaches to polypeptide
* One mutation in DNA leads to point mutation in RNA which changes the protein
* Each codon sequence - makes an ORF- open reading frame
* In eukaryotes, introns are useless, exons are useful, introns aren't read into ORF
* 5' capping 3' poly-acetylation- splicing- mature mRNA separates all the introns and makes a proper nucleotide sequence from which ORF can be derived
* ORF is for ribosomes to read the stretch of DNA nucleotides, which has a protein synthesis initiation sites and a termination site
* Each open reading frame adds one amino acid
* IF in CATCATCAT - one nucleotide changes/gets deleted/gets added - shit happens!!
* So frameshift happens if a nucleotide is added/deleted/read at a wrong frame

**How to know correct ORF?**

1. **Presuming that the different ORFs are splice variants, it is usual to take the longest ORF.**
2. **The ORF finder is a program available at NCBI website. It identifies all ORF or possible protein coding region from six different reading frame. It shows 6 horizontal bars corresponding to one of the possible reading frame. In each direction of the DNA there would be 3 possible reading frames. So total 6 possible reading frame (6 horizontal bars) would be there for every DNA sequence. The 6 possible reading frames are +1, +2, +3 and -1, -2 and -3 in the reverse strand. The resultant amino acids can be saved and search against various protein databases using BLAST for finding similar sequences or amino acids. The result displays the possible protein sequence and the length of the open reading frame etc.**
3. **BLAST- raw**

**Notes on paper**

Ribosome profiling - Uses mRNA sequencing to determine which mRNAs are being actively translated. It produces a “global snapshot” of all the ribosomes active in a cell at a particular moment, known as a translatome. Consequently, this enables researchers to identify the location of translation start sites, the complement of translated ORFs in a cell or tissue, the distribution of ribosomes on an mRNA, and the speed of translating ribosomes.

RNA-Seq (RNA sequencing), also called whole transcriptome shotgun sequencing[2] (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment.[3][4]

RNA-Seq - Facilitates the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusion, mutations/SNPs and changes in gene expression over time, or differences in gene expression in different groups or treatments. In addition to mRNA transcripts, RNA-Seq can look at different populations of RNA to include total RNA, small RNA, such as miRNA, tRNA, and ribosomal profiling. RNA-Seq can also be used to determine exon/intron boundaries and verify or amend previously annotated 5' and 3' gene boundaries.

Ribosome stalling – ribosomes accumulate at specific codon positions of mRNAs, aka ribosomes piling up at specific positions on mRNAs

ROSE– RibosOme Stalling Estimator - estimates probability of a ribosome stalling at a specific genomic (codon) location

Translational elongation – ribosome scans the A site and attaches amino acids to P site, elongating the polypeptide chain.

Assumption: A ribosome stalling event is primarily deter- mined by its surrounding sequence. The codon position of interest, i.e., the ribosome A site, was first extended both upward and downward by 30 codons, which yielded the codon sequence profile of the expected stalling event.

ROSE relied on a number of motif detectors (i.e., convolution operators) to scan the input sequence and integrated those stalling-relevant motifs to capture the intrinsic contextual features of ribosome stalling

RSS – ribosome stalling score - the RSS can be considered as an estimate of the likelihood of ribosome stalling. A higher RSS generally indicates a higher predicted probability of ribosome stalling at the corresponding codon position.

Check STAR Method- the codon sites with normalized footprint densities beyond the threshold were labeled as positive samples, while an equal num- ber of codon sites randomly chosen from the remaining were labeled as negative samples

Questions:

1. What is normalization?
2. In the dataset – LCL – is it timepoint data (from what I understand, yes)
3. From what it seems, they have some data which they know frameshift has happened and they’ve labelled the positions. Do we have such data? Can we use their data? (HeLa and LCL)
4. Ramp regions – is it important? Note that here we excluded all the reads of the ramp regions (i.e., the first 50 codons at the 50 ends of coding sequences)

Preprocessing:

In our framework, we first encode the input codon sequence using the one-hot encoding technique (Pedregosa et al., 2011), that is, the mth codon type is encoded as a binary vector of length 64, in which the mth position is one while the others are zeros, after index- ing all 64 codon types. Then the encoded information is fed into one convolutional layer and one pooling layer to learn the hidden features.

CNN:

Then the encoded information is fed into one convolutional layer and one pooling layer to learn the hidden features. In the convolutional layer, several one-dimensional convolution operations are performed over the 64-channel input data, in which each channel corresponds to one dimension of the input vector, and the weight matrix (i.e., kernel) can be regarded as the position weight matrix (PWM).

Next, the rectified linear activation function (ReLU) is used to imitate the neuron activation, that is, the output of the convolutional layer is further processed by the activation function Y=ReLU(X).

After convolution and rectification, we reduce the dimension of matrix Y using the max pooling operation, which computes the maximum value within a scanning window of size three and step size two.

To calculate the final probability of a ribosome stalling event, the unified representation is directly fed to a sigmoid layer, which computes

../../../Screenshot%202018-12-26%20at%201.07.10%20AM.png

The above calculated probability p(s) is defined as the intergenic ribosome stalling score (also termed interRSS), which measures the likelihood of ribosome stalling at a codon position. To eliminate the interRSS bias among different genes, we further define the intragenic ribosome stalling score (also termed intraRSS) as follows:

../../../Screenshot%202018-12-26%20at%201.15.29%20AM.png

where interRSS(position) represents the interRSS of the codon position of interest and mean(gene) stands for the mean interRSS of the corresponding gene. When computing mean(gene), we exclude those codon positions in the ramp regions (i.e., the first 50 co- dons at the 50 ends of coding sequences).

Given the training samples {(si,yi)}i, the loss function of our model is defined as the sum of the negative log likelihoods (NLLs)

../../../Screenshot%202018-12-26%20at%201.16.38%20AM.png

where si is the input codon sequence and yi is the true label.

**Hyperparameter optimization:**

Although we can achieve this goal using the conventional cross-validation strategies, it is generally time-consuming to test all possible combinations of these hyperparameters. To conquer this difficulty, here we propose a one-way model selection strategy for automatic and efficient hyperparameter calibration. In this strategy, we first arbitrarily choose the initial values of the hyperparameters from a candidate set. Then, we separate the hyperparameters into two groups, including those describing the network structure (denoted by H1), such as the kernel size and the kernel number, and those describing the optimization procedure (denoted by H2), such as the base learning rate and the weight decay coefficient. Next, by fixing the values of the hyperparameters in H2, we calibrate those hyperparameters in H1 using a three-fold cross-validation (CV) procedure, and determine their optimal values that achieve the best CV performance. Similarly, the hyperparameters in H2 are also calibrated via the three-fold CV procedure after fixing the previously determined values of the hyper- parameters in H1.