Machine Learning Engineer Nanodegree

Capstone Project

Joe Udacity
December 31st, 2050

I. Definition

(approx. 1-2 pages)

Project Overview

In this section, look to provide a high-level overview of the project in layman's terms. Questions to ask yourself when writing this section:

- *Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?*
- *Has enough background information been given so that an uninformed reader would understand the problem domain and following problem statement?*

All biology depends on nucleotide and amino acid sequences which spontaneously fold to form 3D atomic biochemical structures, with the former relatively less studied. (Wan et al. 2011). However, RNA structure has increasingly been found to play a significant role in regulating cellular activities where two examples include RNA polymerase, riboswitches whose function is to make a RNA copy of DNA and sense metabolic molecules respectively (Wan et al. 2011). As RNA and associated molecules bind specifically through complementary shapes and energetically favorable interactions (Van der Waals interaction, hydrogen bonds, ionic pairing) (Lounnas et al. 2013), there is a need to study RNA structure. The primary depository of structural data is the Protein Data Bank (https://www.rcsb.org/pdb/home/home.do).

Among the 100 thousand structures deposited into the Protein Data Bank, only three thousand contains RNA -- RNA is inherently flexible and difficult to resolve to a good resolution (Ke 2004). There is now a need to develop of various computational tools to predict RNA tertiary structure, where notable developments include statistical mechanics and machine translation methods (Das and Baker 2007; Popenda et al. 2012). It must be noted that RNA contains at least 20 * length atoms and the 3D space to explore when predicting an RNA structure is not a trivial task.

One area at least to my knowledge not attempted is deep learning, despite having successfully applied to a similar class of biological molecules – proteins which are largely analogous to RNA in structure and function (Wang et al. 2017). It is hoped that deep learning can be used to help predict RNA structure or least reduce the complexity of RNA structure problem for other algorithms.

Problem Statement

In this section, you will want to clearly define the problem that you are trying to solve, including the strategy (outline of tasks) you will use to achieve the desired solution. You should also thoroughly

discuss what the intended solution will be for this problem. Questions to ask yourself when writing this section:

- *Is the problem statement clearly defined? Will the reader understand what you are expecting to solve?*
- *Have you thoroughly discussed how you will attempt to solve the problem?*
- *Is an anticipated solution clearly defined? Will the reader understand what results you are looking for?*

Output of model

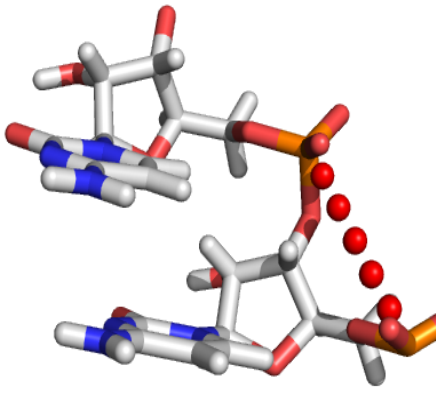
I will be predicting pair-wise atom euclidean distances between basic building blocks of RNA, also called *residues* shown in matrix **D** (refer to formula 1). The matrix is zero-diagonal symmetric. The matrix has dimensions length * length , with length being number of residues in RNA sequence in the pair-wise distance matrix **D**.

Figure 1a shows exactly what distances are measured, which is the distance between two RNA phosphate atoms, while figure 1b shows an entire RNA structure with 70 residues and a complex fold for you to gain an intuition of the complexity of the problem.

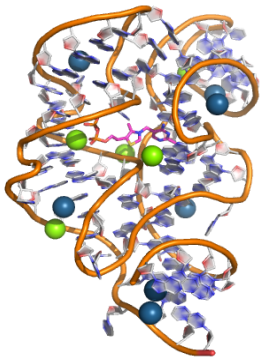
To be specific about matrix **D**, since distance based regression may be difficult, I would predict categorical values instead of actual euclidean distances. The other reason why I used distance categorical values is that I would be able to assign probabilities on the predictions. These distance categories are short and long which are < 16 Å and >16 Å respectively and rationale for the distance is in the analysis. 1Å is a unit of measurement of the field, which equals 10^{-10}m. The distance matrix hence is a tensor of size (Length * Length * 1 for two categories).

Formula 1, Pairwise distance matrix

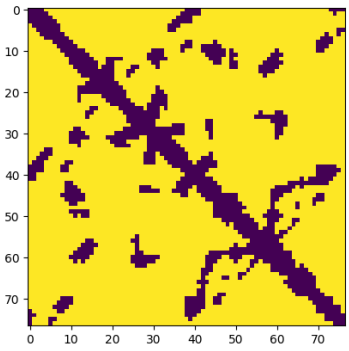
$$\boldsymbol{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & 0 \end{bmatrix}$$



A



B



C

Figure 1a,b : A). Two RNA residues and the distance between phosphate atoms shown in red

spheres. B). Example of an RNA molecule and its complex structural fold (2cky). C). Distance matrix, (dark, light for short and long distances respectively). This RNA is the TPP riboswitch which is a potential antibiotic target.

3. Inputs

ii. Pairwise features

There are certain features that are already (length * length) without the outer product transformation which transforms (Number_of_features * length) to (length * length).

They include secondary structure, pairwise residue from both sequence and consensuses sequence. Figure 3 gives an example of a pairwise residues. They will give 21 (length * length) features.

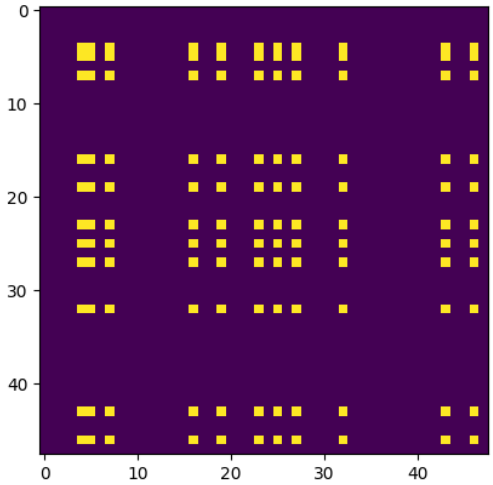


Figure 3 – Here shows presence of both A and G residues among residue pairs depicted by yellow dot.

Dataset.

There are a total of 492 structures in the train dataset, with RNAs from 35 to 500 residues long. This number is much reduced compared a total of 3000 structures due to redundancies in the protein data bank as each structure may occur multiple times. We used a manually curated list of non-redundant structures found here of version **2.141** to ensure our data is not biased by duplicates. The train data will also be augmented by subsampling 35, 50, 75, 100, 150, 200, 500 residues of the entire structure and this increases the size of the training set by 30 fold. We made a train,validation, test split of 370 : 110 : 12 structures. Model is trained only on train set, training stopped when validation performance is maxed and model finally evaluated against the 12 structures. These 12 structures were part of an RNA modeling competition and benchmarks against the best methods.

Metrics

Model predicts categorical distance matrix and the metric used to train the model will be sigmoid log loss averaged over each entry in the distance matrix. The model however will be evaluated using balanced accuracy. The formula is (TP/Positives + TN/negatives)/2. This gives a balanced metric that is unaffected by the size of the RNA. This is because the pair-wise distance categorical matrix contains mostly negatives and larger RNA contains relatively more negatives compared to a smaller RNA. The balanced accuracy thus gives a metric that can be compared across RNAs of different sizes.

The probability threshold will be 0.3, where distances predicted a 30% probability will be deemed sufficient to be called a close distance. This was seen in our cross validations where a lower threshold results in a higher balanced accuracy.

II. Analysis

(approx. 2-4 pages)

Algorithms and Techniques

We will be using deep learning to predict the categorical distance matrix from the sequence and features generated from the sequence. Deep learning automates features and current image processing techniques are now state of the art and exceed human performance (cite). I believe a CNN is the easiest and most direct way to solve the problem because the RNA sequence varies in length in two orders of magnitude and existing machine learning methods cannot easily take in inputs of varying sizes. Furthermore, the problem is extremely complicated and almost impossible to model by ‘intuition’ . This makes is difficult to design features to model using machine learning methods. Given the extreme difficulty for a different machine learning model and the robustness of CNN in image processing, I have decided to use other methods as a benchmark instead of machine learning.

Data Exploration

i. Sequence inputs

RNA sequences consist of only four different building blocks, A, U, G and C, where the sequence will be one-hot encoded into A, U, G and C in formula 2. Extra features added are using a variety of experimental and theoretical methods to further enrich the data, and help the neural network model. The input is a matrix with dimensions (Number_of_features * length , formula 2). This matrix will undergo 1D convolution along the horizontal axis and finally transformed to a (length * length * channels) tensor using inverse singular value decomposition .

The features are the following :
1. Residue A, U, G, C and others (5 features)
2. Predicted secondary structure and its associated base pair probabilities, positional entropy, mean free energy structure and positions of base pairs from ViennaRNA. (5 features)
3. Consensuses sequence[1] and confidence of prediction from infernal. (5 features)

Formula 2 for AUGCGG

$$\boldsymbol{Inputs} = \begin{bmatrix} \textit{Legend} & \textit{Residue}\,1 & \textit{Residue}\,2 & \ldots & \ldots & \ldots & \ldots \\ A & 1 & 0 & 0 & 0 & 0 & 0 \\ U & 0 & 1 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 1 & 0 & 1 & 1 \\ C & 0 & 0 & 0 & 1 & 0 & 0 \\ \textit{Others} & 0 & 0 & 0 & 0 & 0 & 0 \\ \textit{Extra Feature A} & 2.5 & 3.5 & 4.5 & -2 & 2 & 0.5 \\ \textit{Extra Feature B} & 9 & 0.5 & -4.5 & 2 & 9 & -9.5 \end{bmatrix}$$

1 Consensus sequence is the representative sequence of the RNA family

The biological and physical significance of AUGC is that sequence determines structure (quote) and thus it is possible to predict structure (and its distance matrix) from the sequence alone. Secondly, given that secondary structure predictions are rather robust (RNA composer), we use these algorithms to predict our secondary structure. Secondary structure helps to determine and simplify which residues are close and form hydrogen bonds, however this is not exhaustive

and there are other forms of structure that needs to be predicted. The final groups of features are that of the consensus sequences. RNA can be very different in sequence, yet adopt similar folds if they are from the same RNA family. This prediction will help predict related RNAs that differ very much in sequence. I have included an environment to explore the features.

ii. Pairwise inputs

There are certain features that are already (length * length) without the matrix transformation which transforms (Number_of_features * length) to (length * length). These features include which pairs of residues form secondary structures (also known as base pairs), which can be predicted reliably by ViennaRNA. These also consist of pairwise residue from both the sequence and consensuses sequence. Figure 3 gives an example of a pairwise residues. They will give 21 (length * length) features. It is hoped this can help prediction of certain combination of residue pairs that interact strongly together.

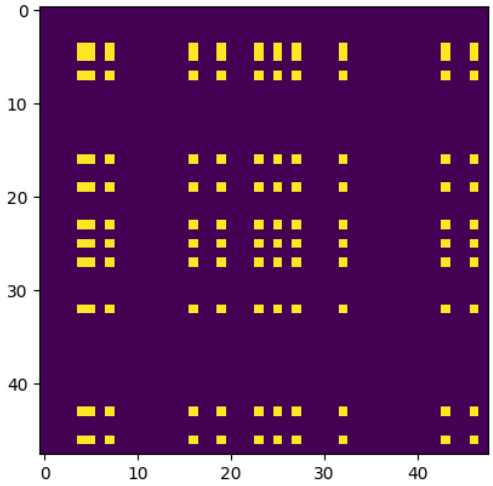


Figure 3 – Here shows presence of both A and G residues among residue pairs depicted by yellow dot.

Since deep learning does automated feature selection, unimportant inputs will have little significance in the model and vice versa for important inputs. The convolution filters should pick up important features for the model which are too difficult for humans to visualize and model.

Chosing of distance 16Å

I plotted all the distances between residues and found that there are three peaks in 8Å, 12 Å and 18Å. I tried different models with different distances. The 8Å model was too few to have any predictive power while and 12Å model did not converge due to the class imbalance. After reweighing the minority class, the model converged, but the categorical distance matrix did not look informative which is why I settled for a 16Å model.

Due to experimental limitations, each RNA had an average of 6% of residues not resolved with a std of 14%. This means there are many structures which contain many missing distances. I will mask these missing distances from the loss by doing an element-wise multiplication of zero for these missing distances. The diagonals and its immediate neighbors will also be masked out of the log-

loss so that the model can concentrate to predict the other distances since distances near the diagonals are always < 16Å .

Exploratory Visualization

Provided in environment

Benchmark

I have used RNA modelling competitions and converted the structures into a distance matrix for equal comparison to my deep learning method. There were 21 RNA competitions, with 16 having released results. Among these 16, one had an erroneous sequence between the predicted and modeled structure (Puzzle 13), one was trivial (puzzle 1), two contained more than one RNA chain (puzzle 2 and 10) and outside the scope of our train set. We were left with 12 structures to benchmark our method against others. Since the other methods were outside machine learning, the details will be left out, but I will highlight each method's average rank across the competitions with a brief description of the method.

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

- *Has some result or value been provided that acts as a benchmark for measuring performance?*
- *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*