Machine Learning Engineer Nanodegree
Capstone Proposal
Lee XA
September 2017

Proposal : Modeling RNA structural interactions with deep learning

Contents

1. Domain Background (Biochemistry)

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.

2. Problem Statement

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 three 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, medium and long which are <8Å , 8-15Å and >15 Å respectively. 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 * 3 categories).

Formula 1, Pairwise distance matrix

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

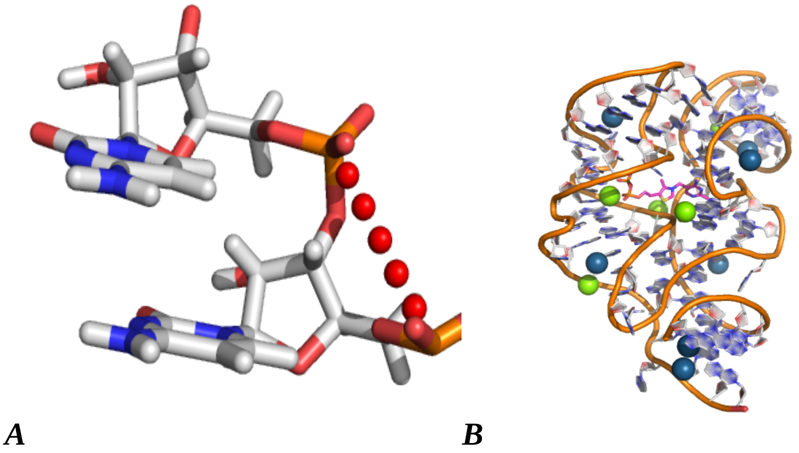


Figure 1a,b : A). Two RNA residues and the distance between phosphate atoms shown in red spheres. B). Example of RNA molecule and its complex structural fold (2cky).

3. Datasets and Inputs

The dataset used here will be RNA structures solved through experimental means from the Protein Data Bank. All of these structures contain the 3D coordinates for the atoms and the residue sequence. **The input variable will be the RNA sequence one-hot encoded and related features (formula 2), while the output variable will be the pairwise distance matrix D.**

RNA sequences consist of only four different building blocks, A, U, G and C, and a sequence of AUGCGG will have an input of the matrix below with dimentions (Number_of_features * length , formula 2). There will also be pair-wise inputs of size (length * length) which will be appropriately merged into the model (refer to section 7). Furthermore, I will include extra features determined by feature engineering using open source bioinformatics software. One tool I will use is a software suite ViennaRNA (Lorenz et al. 2011). These information use a variety of experimental and theoretical methods to further enrich the data, and help the neural network model.

Formula 2 for AUGCGG

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

There are a total of 490 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 unbiased. The data will be augmented by subsampling 35, 50, 75, 100, 150, 200, 400, 600, 800 residues of the entire structure. An example will be a structure of 900 residues long. I will take subsamples of 35 residues, 75 residues,….. and so on. This will increase the size of the train data.

Unfortunately you are not a biochemist and thus you have to assume that the data I propose is most suitable given the problem. The input data is of dimensions **(Num_of_features * Length)** and **(Length * Length * Num_of_features)** and output variable is **(Length * Length * 3)**.

I have included an example how to navigate the database site in the appendix.

4. Solution Statement

It is useful to note that the distance matrix **D** is very similar to an image and I will use CNN to solve the problem with a notable difference that pixel level regression will be done. This might pose some difficulties, but hey if the problem was so simple, it should be trashed.

Why I believe a CNN is suitable is that filters are invariant to input lengths. RNA differ in orders of magnitude in length, and CNN can handle such inputs. Furthermore, certain sequence patterns of RNA are correlated with certain geometries, and this can be picked up through convolution filters and proceed to predict through the distance matrix **D**.

A suitable architecture is below.

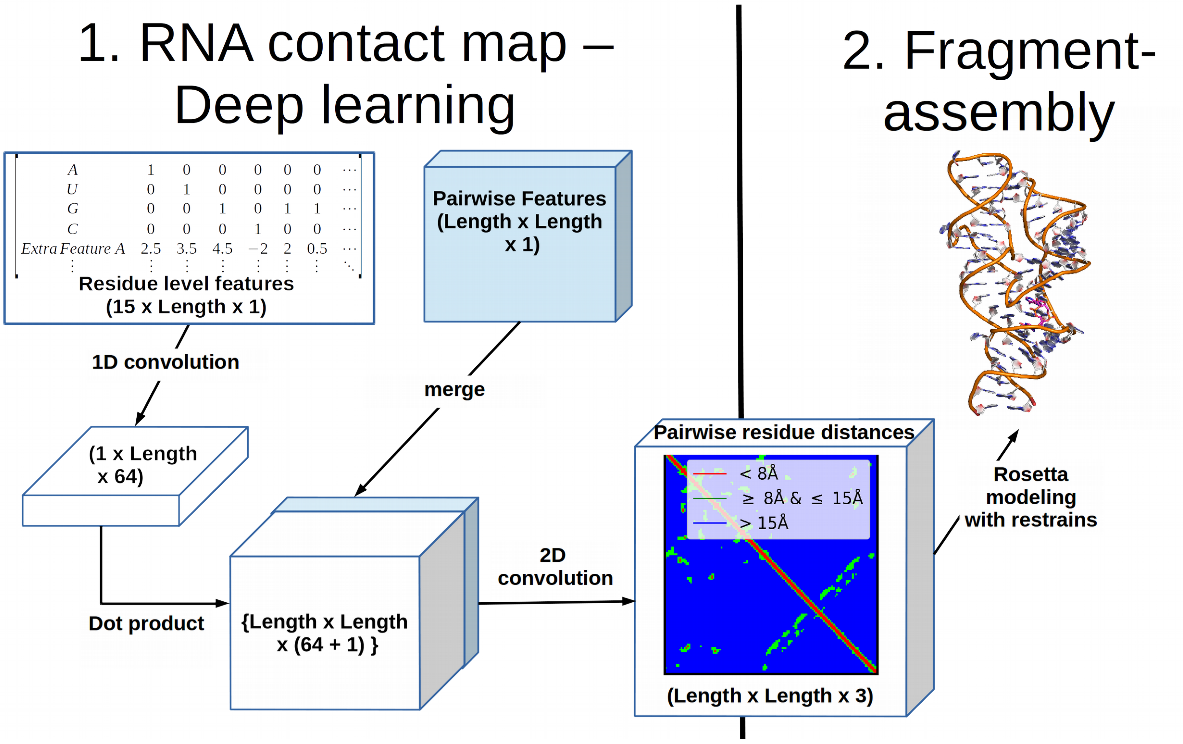

1. RNA contact map – Deep learning
2. Fragment-assembly
A 1 0 0 0 0 0 ⋯
U 0 1 0 0 0 0 ⋯
G 0 0 1 0 1 1 ⋯
C 0 0 0 1 0 0 ⋯
Extra Feature A 2.5 3.5 4.5 −2 2 0.5 ⋯
Residue level features (15 x Length x 1)
Pairwise Features (Length x Length x 1)
1D convolution
merge
(1 x Length x 64)
Dot product
{Length x Length x (64 + 1) }
2D convolution
Pairwise residue distances
< 8Å
≥ 8Å & ≤ 15Å
> 15Å
(Length x Length x 3)
Rosetta modeling with restrains

Figure 2 Architecture of work flow : Machine Learning portion only consist of part 1, part 2 is outside the scope of the project.

5. Benchmark Model

I will use about 15 RNA structures from previous RNA modeling competition also called RNA puzzles (http://ahsoka.u-strasbg.fr/rnapuzzlesv2/). I wish to see how well the method does compared to other state of the art methods used in these competitions. The test set contains relatively few structures and this is a limitation of the field since RNA structure determination is extremely challenging and could take years. I will also explicitly remove these 15 structures from the train set.

There are many different types of RNA modeling tools, however these tools have already been trained on the train dataset, making it not a blind accuracy test. Thus a retrospective analysis RNA competitions are most suitable comparisons for accuracy. Furthermore, these competitions use current state of the art tools and include domain experts and thus are a good gauge of model performance.

It is useful to note that RNA puzzles requires predictions of the entire structure while my model predicts only pair-wise distances. To ensure a fair comparison, I will extract the distance matrix from RNA-puzzles for comparison against my deep learning model.

6. Evaluation metric

I will be evaluating predictions eg. Short, medium or large using a weighted accuracy metric. This is because the ratio of short,medium to large distances are in ratio of 0.04, 0.06 and 0.9 respectively and skewed towards larger distances. The weights for the short,medium and large distances are k/0.04, k/0.06 and k/0.9 where k is a normalizing constant to ensure sum of weights equal to 1. A weighted categorical log-loss using the weighted logits will be used to optimize the model.

7. Project Design

The first thing to be done is to download all non-redundant structures, followed by using a variety of python scripts to extract the necessary information. An example of the data is shown in appendix.

Given that the sequence length varies, CNN networks are most suitable as they can : i) handle inputs of different lengths, ii) are able to capture local information of sequences which is important since local environment plays a role in RNA structure, iii) able to model complex interactions too complex for humans to dissect, iv) shown to work for a related class of biological molecules – proteins.

The raw inputs are sequence, and the rich variety of open source bioinformatics software will be used to give more inputs to the model. Thus the CNN model will receive these three classes of inputs : i) Sequence, ii) single residue features and iii) pairwise residue features. Features i and ii will undergo an outer product operation to become length * length matrix before merging with feature iii. These merge layer would then undergo a deep CNN before outputting the D distance matrix.

For the CNN, I will first try a simple architecture, changing these hyper parameters : I) convolution filter size, ii) number of hidden layers iii) number of filters of each layer. Cross-validation will be used to get the optimal hyper parameters.

One limitation of the study is the small training size and the diversity of RNA structures. This is a limitation of the field, but it is hoped with data augmentation we will be able to generalize RNA structure.

8. References

Das, Rhiju, and David Baker. 2007. "Automated de Novo Prediction of Native-like RNA Tertiary Structures." *Proceedings of the National Academy of Sciences* 104 (37): 14664–14669.

Dev, Sukhendu B. 2015. "Unsolved Problems in biology—The State of Current Thinking." *Progress in Biophysics and Molecular Biology* 117 (2–3): 232–39. doi:10.1016/j.pbiomolbio.2015.02.001.

Ke, A. 2004. "Crystallization of RNA and RNA?protein Complexes." *Methods* 34 (3): 408–14. doi:10.1016/j.ymeth.2004.03.027.

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

Lounnas, Valère, Tina Ritschel, Jan Kelder, Ross McGuire, Robert P. Bywater, and Nicolas Foloppe. 2013. "CURRENT PROGRESS IN STRUCTURE-BASED RATIONAL DRUG DESIGN MARKS A NEW MINDSET IN DRUG DISCOVERY." *Computational and Structural Biotechnology Journal* 5 (6): 1–14. doi:10.5936/csbj.201302011.

Popenda, M., M. Szachniuk, M. Antczak, K. J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, and R. W. Adamiak. 2012. "Automated 3D Structure Composition for Large RNAs." *Nucleic Acids Research* 40 (14): e112–e112. doi:10.1093/nar/gks339.

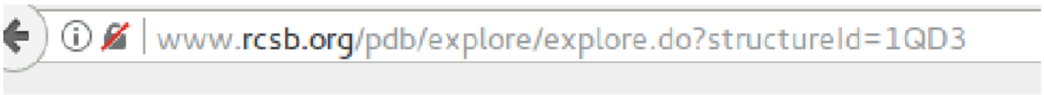
Wan, Yue, Michael Kertesz, Robert C. Spitale, Eran Segal, and Howard Y. Chang. 2011. "Understanding the Transcriptome through RNA Structure." *Nature Reviews Genetics* 12 (9): 641–55. doi:10.1038/nrg3049.

Wang, Sheng, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. 2017. "Accurate de Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model." *PLoS Computational Biology* 13 (1): e1005324.

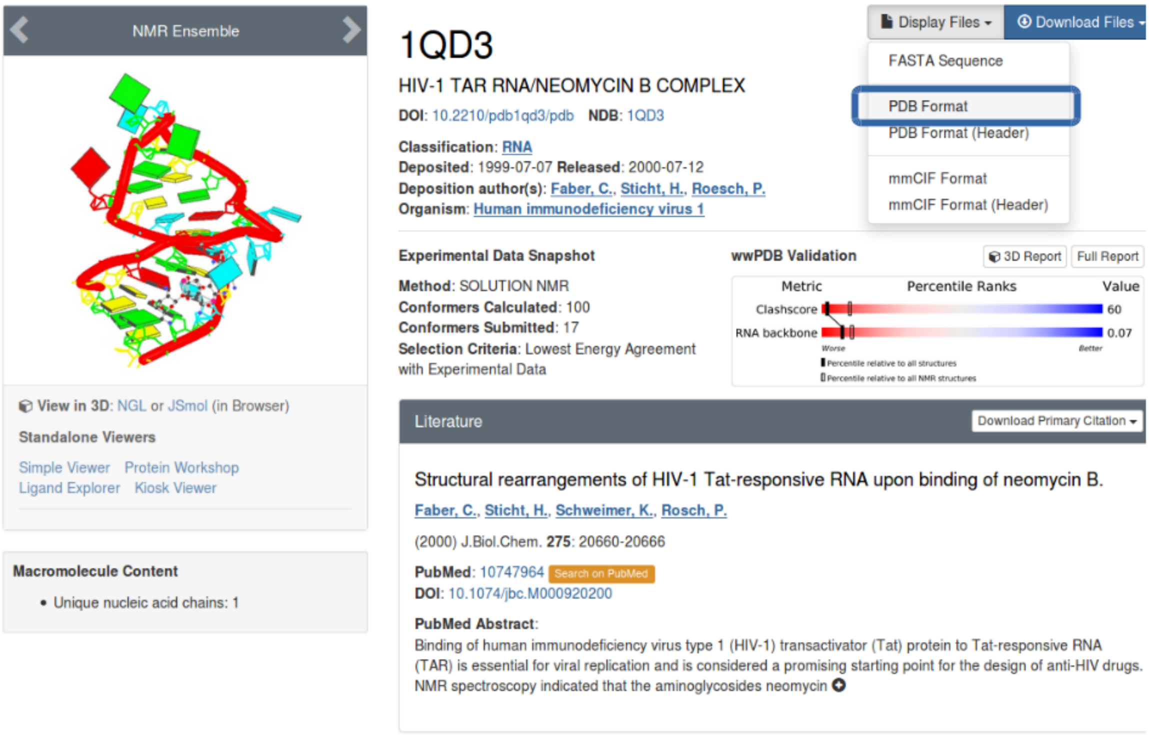
9. Appendix

I have provided a an example of a data source below : HIV-1 TAR RNA molecule in complex with a drug (Neomycin B).

Webpage with RNA : http://www.rcsb.org/pdb/explore/explore.do?structureId=1QD3



www.rcsb.org/pdb/explore/explore.do?structureId=1QD3



NMR Ensemble

View in 3D: NGL or JSmol (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop
Ligand Explorer Kiosk Viewer

Macromolecule Content

- Unique nucleic acid chains: 1

1QD3

HIV-1 TAR RNA/NEOMYCIN B COMPLEX

DOI: 10.2210/pdb1qd3/pdb NDB: 1QD3

Classification: RNA
Deposited: 1999-07-07 Released: 2000-07-12
Deposition author(s): Faber, C., Sticht, H., Roesch, P.
Organism: Human immunodeficiency virus 1

Display Files ▾ | Download Files ▾

- FASTA Sequence
- PDB Format
- PDB Format (Header)
- mmCIF Format
- mmCIF Format (Header)

Experimental Data Snapshot

Method: SOLUTION NMR
Conformers Calculated: 100
Conformers Submitted: 17
Selection Criteria: Lowest Energy Agreement with Experimental Data

wwPDB Validation 3D Report Full Report

Metric	Percentile Ranks	Value
Clashscore		60
RNA backbone		0.07

Worse
Better
▮ Percentile relative to all structures
▯ Percentile relative to all NMR structures

Literature Download Primary Citation ▾

Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B.

Faber, C., Sticht, H., Schweimer, K., Rosch, P.

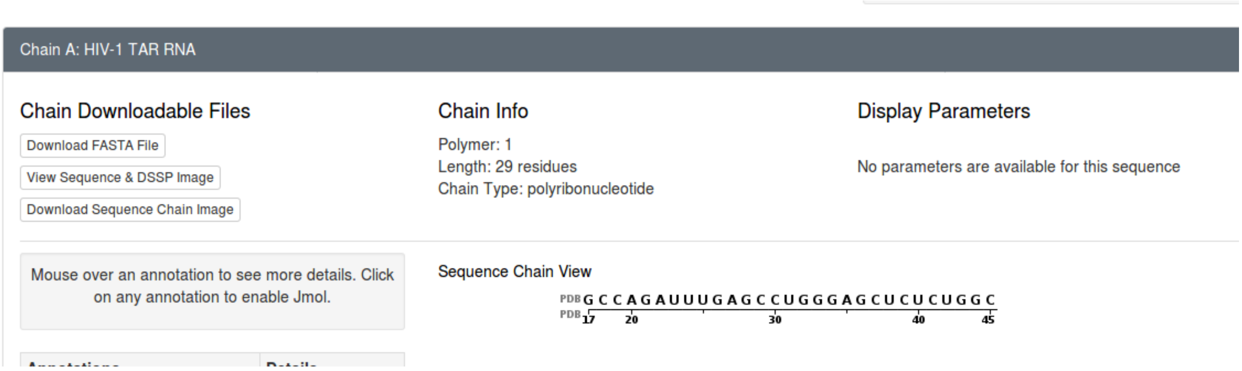
(2000) J.Biol.Chem. **275**: 20660-20666

PubMed: 10747964 Search on PubMed
DOI: 10.1074/jbc.M000920200

PubMed Abstract:
Binding of human immunodeficiency virus type 1 (HIV-1) transactivator (Tat) protein to Tat-responsive RNA (TAR) is essential for viral replication and is considered a promising starting point for the design of anti-HIV drugs. NMR spectroscopy indicated that the aminoglycosides neomycin ⊕

After clicking PDB format, you see a file with gibberish unless you have a PHD in x-ray crystallography. Ignore the gibberish and proceed to lines with the word ATOM. Columns to pay attention are 13-16, 18-20, 21, 23-26, 31-54. These respectively gives atom type, residue type, chain, residue number and xyz coordinates. I will need these information to identify the phosphate atom of each residue and its coordinates.

```
### SAMPLE OF PAGE ###
ATOM     32  P     C A  18      -9.580  -0.134  -9.860  1.00  0.30          P
ATOM     33  OP1   C A  18     -11.026   0.169  -9.719  1.00  0.38           O
ATOM     34  OP2   C A  18      -8.610   0.453  -8.900  1.00  0.35          O
ATOM     35  O5'   C A  18      -9.405  -1.717  -9.844  1.00  0.27          O
##### End of Sample #####
```



Chain A: HIV-1 TAR RNA

Chain Downloadable Files

Chain Info

Polymer: 1
Length: 29 residues
Chain Type: polyribonucleotide

Display Parameters

No parameters are available for this sequence

Mouse over an annotation to see more details. Click on any annotation to enable Jmol.

Sequence Chain View

```
PDB G C C A G A U U U G A G C C U G G G A G C U C U C U G G C
PDB 17  20                  30                  40        45
```

The sequence of the structure is recorded above. Do note that this example was not in the training set because it is shorter than 35 residues long.