Machine Learning Engineer Nanodegree
Capstone Proposal
Lee XA
September 2017

Proposal : Modeling RNA structural interactions with deep learning

Domain Background (Biochemistry)
All biology depends on genome sequences which spontaneously fold to form 3D atomic biochemical structures (Wan et al. 2011). Two examples of these highly structured molecules include RNA polymerase, riboswitches whose function is to make a RNA copy of DNA and sense metabolic molecules respectively (Wan et al. 2011). Notwithstanding that, RNA structure has increasingly been found to play a significant role in regulating cellular activities and people are increasingly targeting RNA molecules for therapeutic purposes. As drug 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 with the primary depository of structural data 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. We know much more about proteins yet the folding problem remains unsolved, what more RNA folding (Dev 2015).

One area at least to my knowledge not attempted is deep learning, despite having successfully applied to a similar class of biological molecules – protein 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

I will be predicting pair-wise atom euclidean distances between basic building blocks of RNA, also called *residues* shown in D. The matrix is zero-diagonal symmetric. Figure 1a shows two RNA residues, while figure 1b shows an entire RNA structure with 70 residues. The distances between phosphate atoms as seen by the red spheres in figure 1. I will predict with raw inputs being RNA sequence, based on tje commonly accepted understanding that structure is from sequence alone. (cite)

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

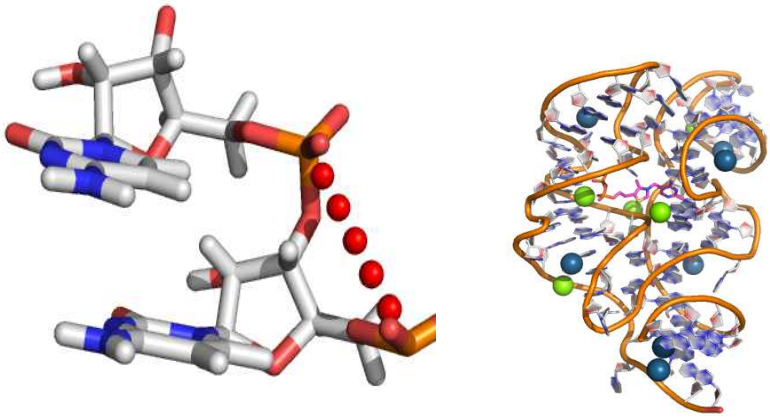


Figure 1a,b in left → right. A). Two RNA residues and the distance between phosphate atoms shown in red spheres. B). Example of RNA molecule and its complex structural fold (2cky).

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 ordinates for the atoms and the residue sequence. **The input variable will be the RNA sequence one-hot encoded and related features (explained in next paragraph), while the output variable will be the pairwise distance matrix D.**

RNA sequences contain of only four difference building blocks, A, U, G and C, and a sequence of AUGCGG will have an input of the matrix below sized n * length. 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).

There are a total of 492 structures in the train dataset, with RNAs from 35-500 residues long.
This number is much reduced 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.

Unfortunately you are not a biochemist and thus you have to assume that the data I propose is most suitable given the problem.

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

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

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.

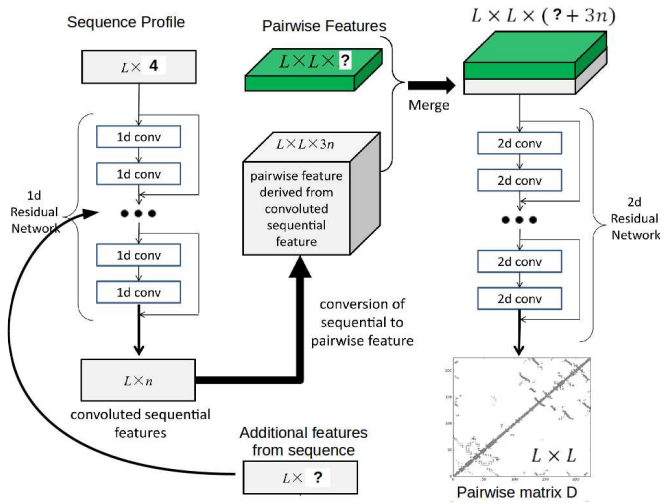


Figure 2a. Architecture adapted from (Wang et al. 2017)

Benchmark Model

I will use 19 RNA structures from previous RNA modelling competition also called RNA puzzles (http://ahsoka.u-strasbg.fr/rnapuzzlesv2/) . I wish to see how well the method does compared to other 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 not include these 19 structures in the train set.

It is useful to note that RNA puzzles model the entire structure while I only predict distance for a single type of atom. To ensure a fair comparison, I will extract the distance matrix from the RNA-puzzles models built for comparison to my prediction

Evaluation metric would be sum of root mean squared error across each unique entry of the pairwise distance matrix. In the event that the root mean squared error is too hard to predict, I will be predicting distance categories eg. Short, medium or large using a categorical entropy loss function.

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. After obtaining the data, I will run some input comparison software since to further reduce redundancies since the non-redundant list still contains redundancies. I will remove sequences which are 90 percent similar as previously done (cite).

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.

Benchmarking will be done using structures from RNA puzzles. There are many different types of RNA modeling tools, however these tools have likely already been trained on the entire train dataset, making it not a blind accuracy test. Thus a retrospective analysis of these RNA puzzles are best suitable comparisons for accuracy.

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.

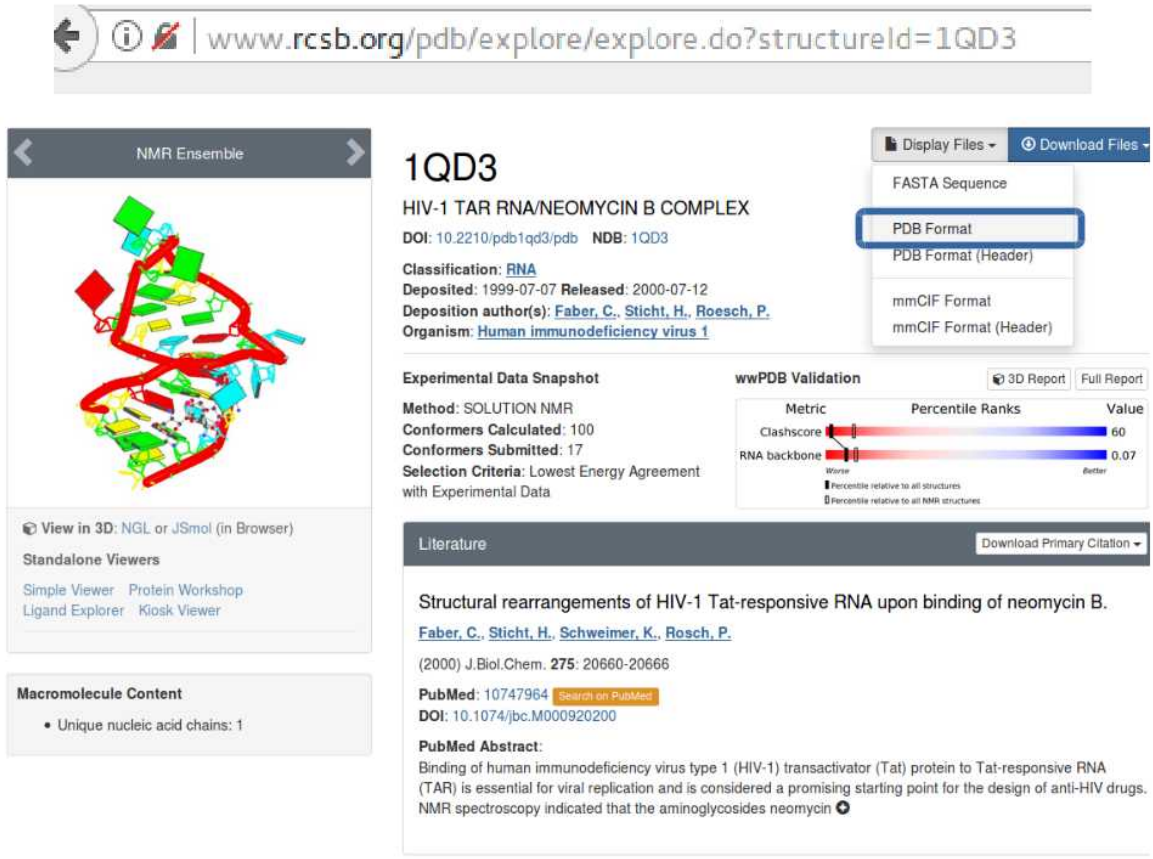
Wan, Yue, Michael Kertesz, Robert C. Spitale, Eran Segal, and Howard Y. Chang. 2011. “Understanding the Transcriptome through RNA Structure.” *Nature Reviews Genetics* 12 (9): 641–55. doi:10.1038/nrg3049.

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.

Appendix

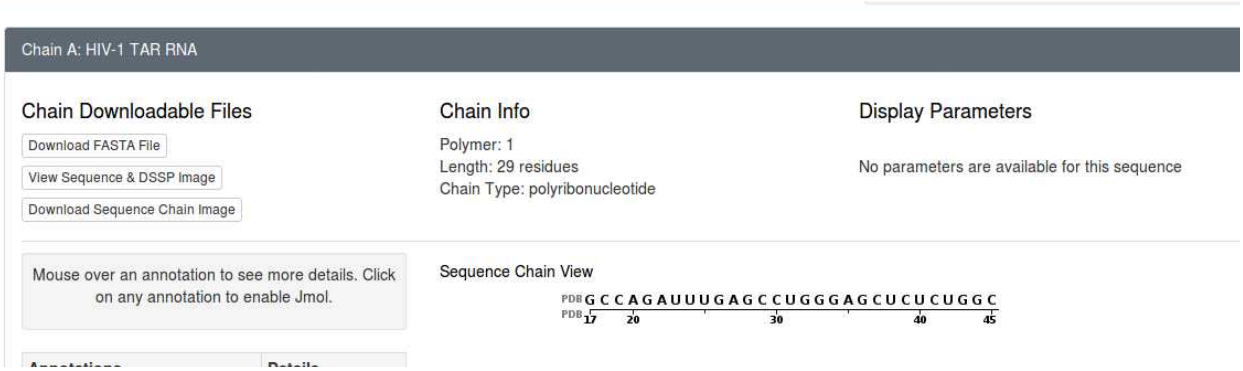
I have provided a an example of a data source below : HIV-1 TAR RNA molecule in complex with a drug (Neomycin B).

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



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



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.