Lee Xiong An
Project 2

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

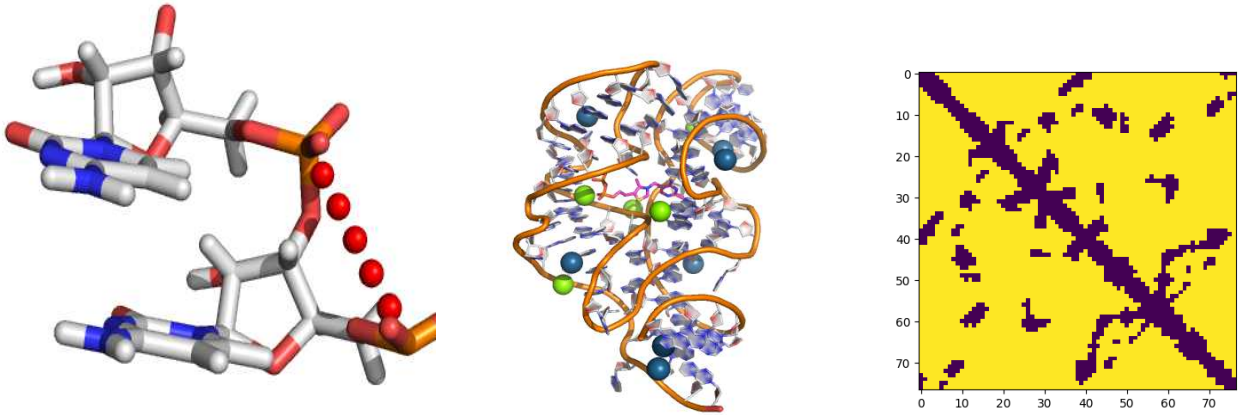
I will be predicting pair-wise atom euclidean distances between basic building blocks of RNA, also called *residues* shown in matrix **D** (refer to formula 1). The matrix is zero-diagonal symmetric. The matrix has dimensions length * length , with length being number of residues in the RNA sequence.

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 at <=16Å and >16Å respectively. 1Å is an atomic unit of measurement, which equals 10^{-10}m. The distance matrix hence is a tensor of size (Length * Length * 2 categories). The categories are split instead of using a binary category to facilitate adding of more categories in the future.

Formula 1, Pairwise distance matrix

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



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. 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, where the sequence will be one-hot encoded. Extra features will be added are using open source bioinformatics software. These extra features uses a variety of experimental and theoretical methods to further enrich the data, and help the neural network model. The final input is a matrix with dimensions (Number_of_features * length , formula 2). **I have also added another feature containing pair-wise inputs of size (length * length * 1)** which will be appropriately merged into the model (refer to section 7)

Formula 2 for AUGCGG

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

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 not biased by duplicates. The data will also be augmented by subsampling 35, 50, 75, 100, 150, 200, 500 residues of the entire structure and increases the size of the training set by 30 fold. The test set was not augmented.

4. Model Architecture

Matrix **D** is very similar to an image and I will use CNN to solve the problem.

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 the distance matrix **D**.

The convolutions layer follows inception 3 architecture since it saves parameters yet allows large convolution filter size. The model was made larger than necessary to prevent and can be optimized smaller in the future.

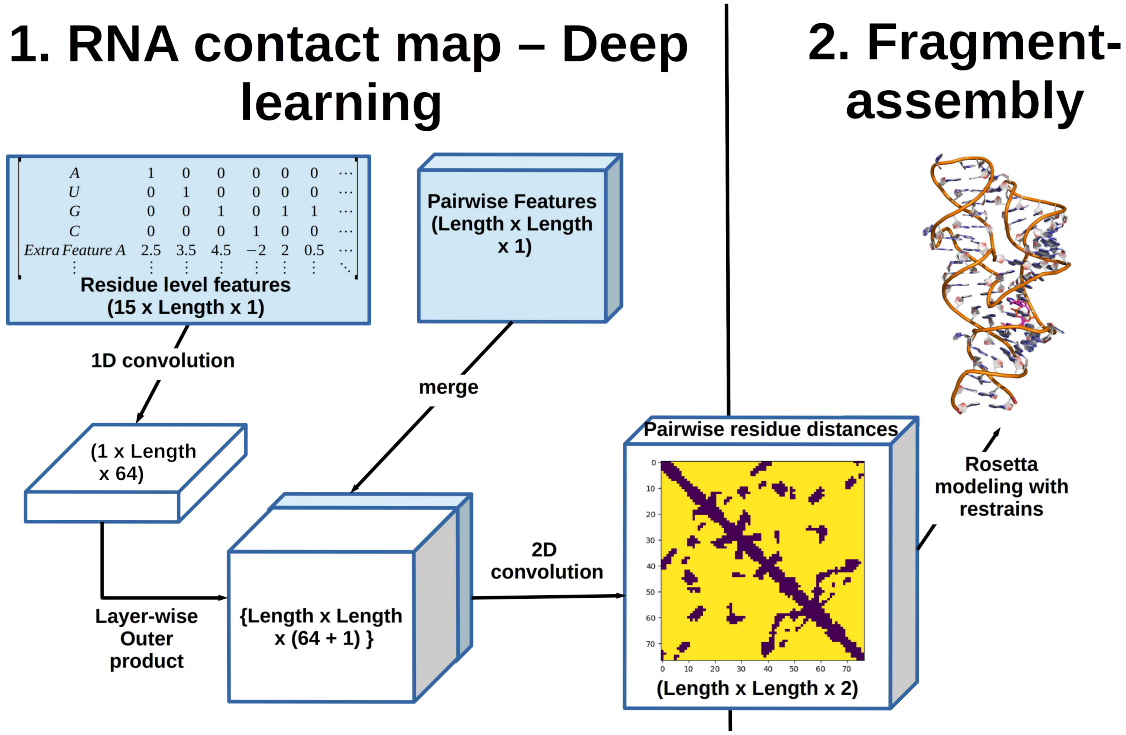


Figure 2 Architecture of model : Inputs are shown in blue, and the output is a pair-wsie distance matrix. Machine Learning portion only consist of part 1, part 2 is outside the scope of the project but completes modeling of the RNA structure.

5. Model training and evaluation metric

Since the number of training examples is relatively small and vary in length, I used stochastic gradient descent with a cosine decay on the learning rate. The model was split into 2 : 1 for train and validation, with the validation set used to stop training. The evaluation metric was categorical cross entropy.

6. Results

Train balanced accuracy : 80%
Test balanced accuracy : 74%

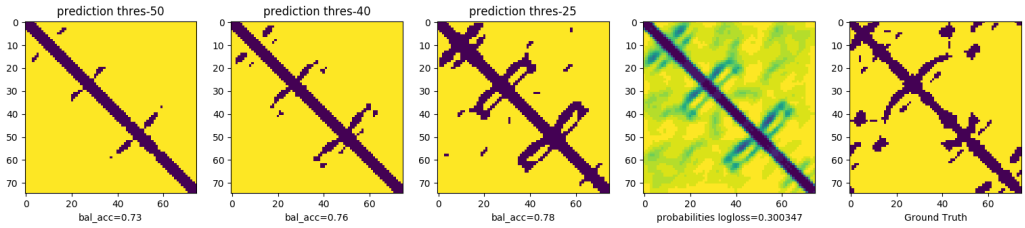
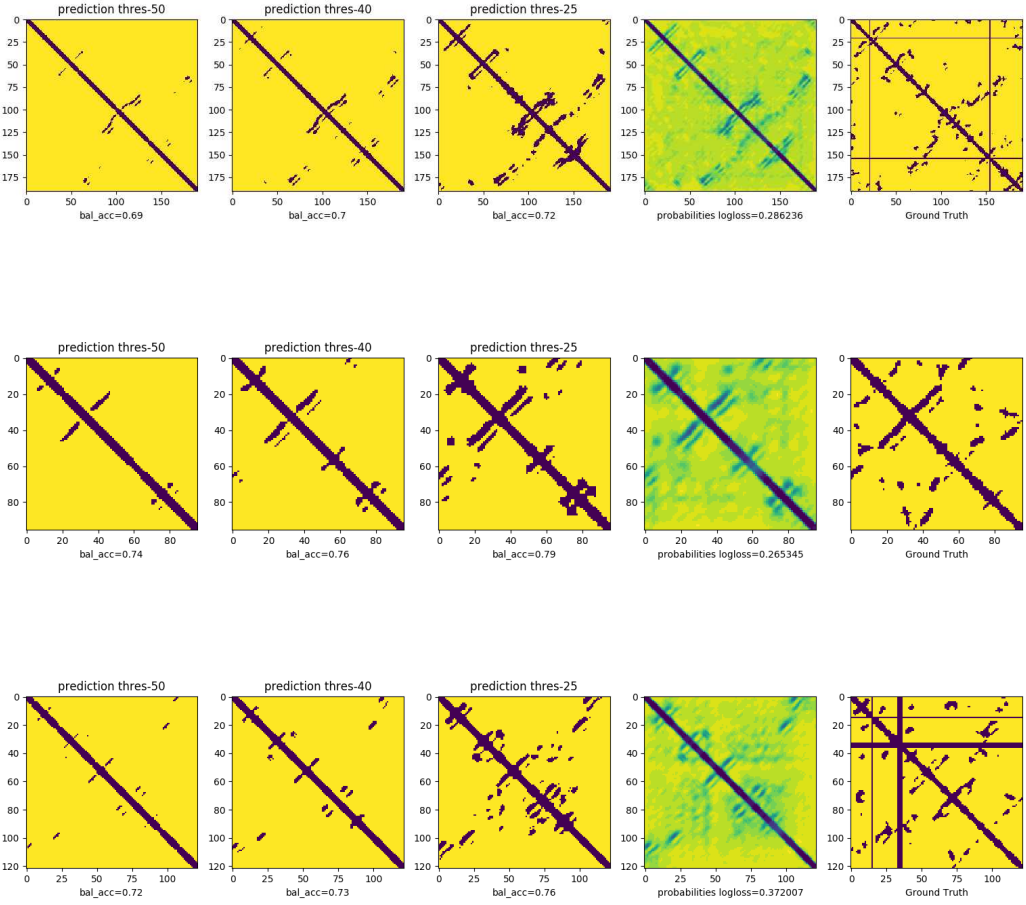
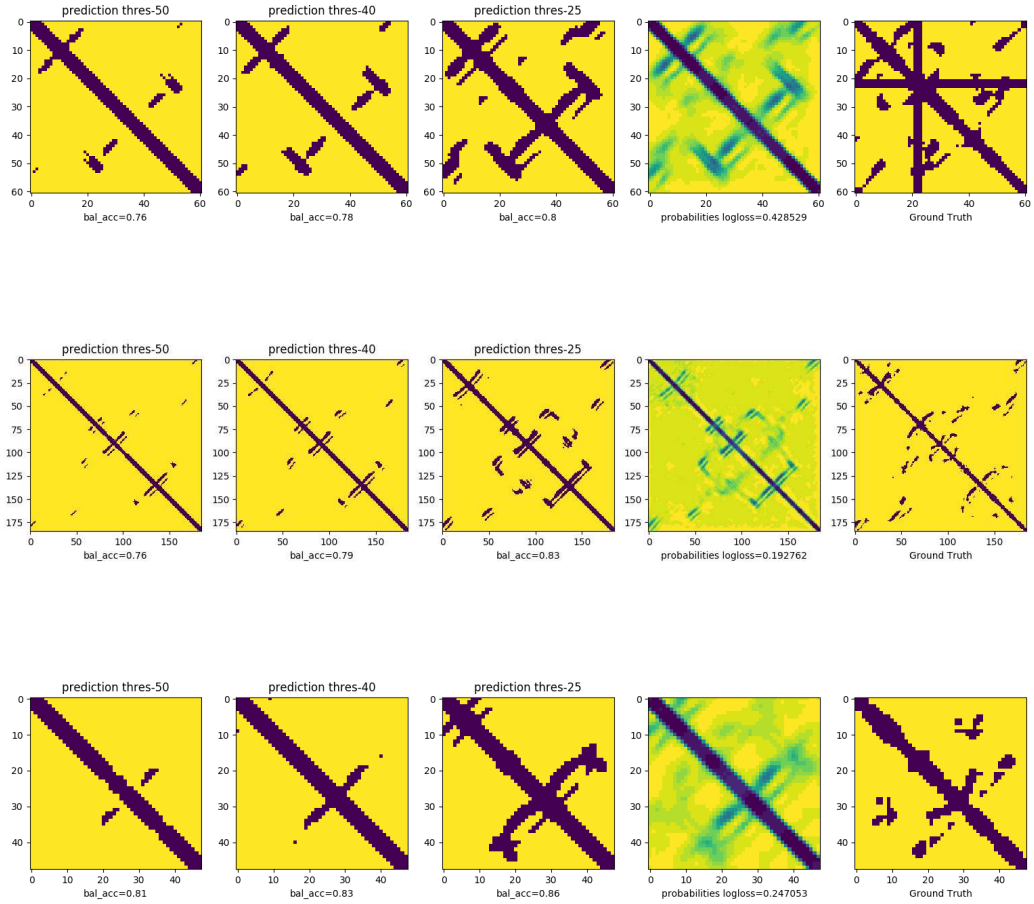


Figure 3 Here shows the predictions with various thresholds, log probabilities and also the ground truth. Generally, we get the diagonals predicted with high confidence since these are easy. The regions far from the diagonals correspond to areas far apart in sequence usually far apart in 3D. However, these regions occasionally interact in 3D space and are difficult to predict accurately due to flexible nature of the RNA molecule though the model seems to predict some correct. The predictions are symmetric after adding it's transpose.

More examples of results :





7. 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.

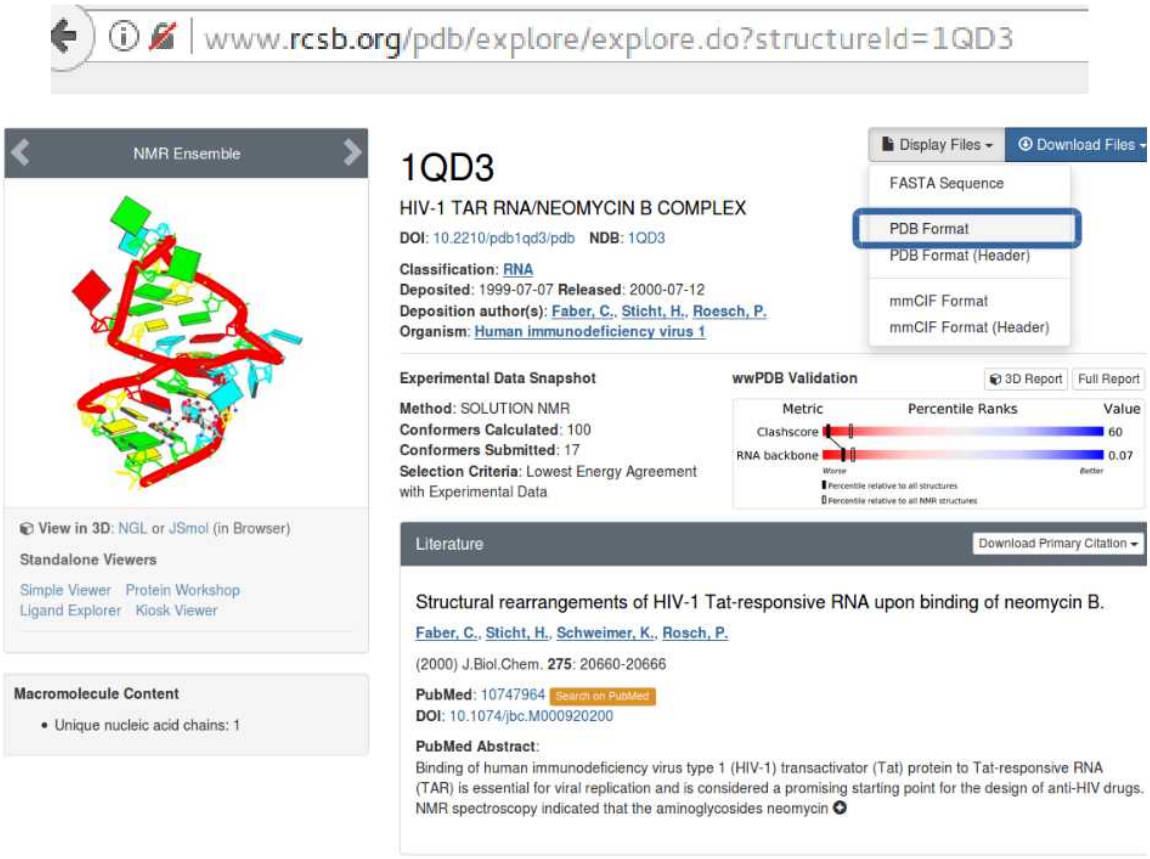
Wan, Yue, Michael Kertesz, Robert C. Spitale, Eran Segal, and Howard Y. Chang. 2011. “Understanding the Transcriptome through RNA Structure.” *Nature Reviews Genetics* 12 (9): 641–55. doi:10.1038/nrg3049.

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.

8. Appendix

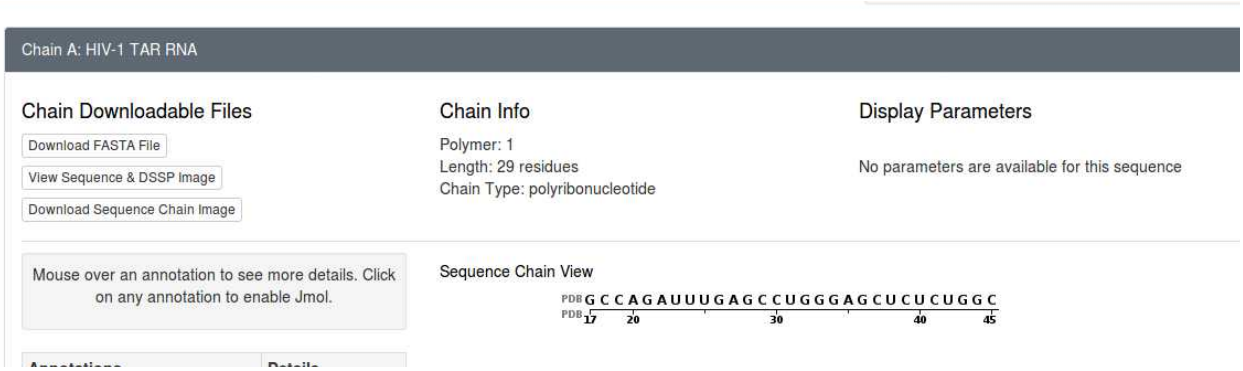
I have provided a an example of a data source below : HIV-1 TAR RNA molecule in complex with a drug (Neomycin B).

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.