# 1 Abstract

RMDetect, written in 2011, has issues with its implementation for RNA Modular Search. Technology has evolved since then. JAR3D and BayesPairing are two recent tools. This paper will take data sets from the ZIP file given as Supplementary Text/Software. The reasons for these said issues will be further analysed. We are interested in finding out the advantages and disadvantages of the listed software. While RMDetect still has issues with its computational power and JAR3D has not been designed to be best used for scoring applications and not searching, it is said that BayesPairing achieves around 70% in identification accuracy.

# 2 Introduction

First question is what are Sequence Motifs. We know that RNA is made up of Watson-Crick Pairs: Guanine–Cytosine and Adenine–Uracil. According to the RMDetect Paper, there are sets or "stacked arrays" of ordered non-Watson-Crick Base Pairs embedded in between these said Watson-Crick Pairs.[1] RNA Motifs, while having a similar name, are sets of Secondary Structure Elements as described in the same RMDetect Paper. The difference is not highlighted in the JAR3D Paper.[2] In fact, there seems to be a whole debate on this definition, and that may be a reason in the discrepancy among the results of the three software. BayesPairing uses a "flexible" definition for RNA 3D modules allowing it to process complex modules and to discover intricacies that were not previously discoverable from the previous software.[3]

Furthermore, the authors of the RMDetect Paper are Biochemists. According to their CVs, the first gentleman is a Portuguese National, and the second gentleman is a French National who was born in the Brussels Capital Area. The fact that these two gentlemen were Biochemists could allude to why RMDetect is suboptimal. Their background is not in Computer Science. Therefore, there would be limitations. The goal of this paper is **to explore on the limitations and advantages of each software**. Each of the papers use the information from the previous papers to build up and to improve on the previous software. Therefore, it will be interesting to see the expected output based on what is presented in the papers. It is imperative that these constraints be analysed to allow for further improvements and modifications. No piece of software is perfect, and therefore

1. José Almeida Cruz and Eric Westhof, "Sequence-based identification of 3D structural modules in RNA with RMDetect", Article, *Nature Methods* 8 (8 May 2011): 513 EP, https://doi.org/10.1038/nmeth.1603.

2. James Roll et al., "Identifying novel sequence variants of RNA 3D motifs", *Nucleic Acids Research* 43, no. 15 (29 June 2015): 7504–7520, ISSN: 0305-1048, doi:10.1093/nar/gkv651, eprint: http://oup.prod.sis.lan/nar/article-pdf/43/15/7504/17434148/gkv651.pdf, https://doi.org/10.1093/nar/gkv651.

3. Roman Sarrazin-Gendron et al., "Automated, customizable and efficient identification of 3D base pair modules with BayesPairing", 4 March 2019, doi:10.1093/nar/gkz102, eprint: http://oup.prod.sis.lan/nar/advance-article-pdf/doi/10.1093/nar/gkz102/28007114/gkz102.pdf, https://doi.org/10.1093/nar/gkz102.

each and every single one of the three described software has its own advantages and limitations.

# 3 Methods

Originally, the plan was to run all the FASTA files presented in the RMDetect software in both BayesParing and JAR3D. Unfortunately to my surprise, there is only one FASTA file presented in the ZIP containing RMDetect. Therefore, I have decided to look online for additional files. On the JAR3D Webserver, there is a series of FASTA files. I have decided to run some of these. My goal is to run a custom dataset in order to test. RMDetect makes this difficult as well because there is no webserver. Each file will have to be analysed in order to figure out the intricacies. I plan on describing the advantages and disadvantages of each software.

To begin, I will try to make my dataset based on the expected modules. I will try to make three of each type. It was suggested that 10 to 20 modules be created for this task. I will then run the dataset on the web server. Unfortunately with RMDetect as stated earlier, there is no web server, and I will likely have to download the full ZIP file and run each file through the script. It is expected that BayesPairing and JAR3D will perform better than RMDetect. This is mainly due to the fact that both were written more recently. Two biochemists wrote RMDetect in 2011. Technology has evolved since then. It is expected that the efficiency and the algorithm have been optimised over the past few years. As well, these two gentleman did not specialise in Computer Science. It is not unexpected why their software would be suboptimal.

After running the dataset through the web servers, I will interpret the results. It is expected that there will be a difference in accuracy due to implementations. Due to the fact that several years have passed since the release of RMDetect, it is expected that BayesPairing and JAR3D have improved on the shortcomings of RMDetect. It is obviously now expected that both BayesPairing and JAR3D will not only run faster, but they will be able to categorise more modules. Because of the definition of 'module', JAR3D only analyses a portion of what BayesPairing analyses, but also continues to remain a superior scoring tool because it was specifically designed for this.[4] The accuracy of the algorithms will likely be analysed through a box plot which would display the medians. The runtimes have been analysed before, so that will not be necessary, but it will be something interesting to look at in the analysis. I will likely have to download the Python Files and Insert a Timer or look at how long it took to load the results on the webserver.

4. Sarrazin-Gendron et al., "Automated, customizable and efficient identification of 3D base pair modules with BayesPairing".

# 4 Results

Trying to find the `FASTA` Files was annoying. I wanted to analyse the Tandem-GA Modules because that was the type of module that `RMDetect` had the most trouble interpreting. Then, I tried googling 'Tandem-GA Module' directly to no avail. I then came across the `RFam` Database, and I found that `RF02540` is a 'family match' for the module. Attached is the sample output from `BayesPairing / JAR3D` and the corresponding input.

# 5 Conclusion

## 5.1 RMDetect Problems

Unfortunately after multiple tries, `RMDetect` could not function correctly. It can be assumed that this could be a result of two biochemists writing a suboptimal piece of software. It is said that a majority of the false positives were a result of the Tandem-GA Module. The false discovery rate in the single sequence search was greater than 0.50; in the multiple sequence search, 0.23 with most of those errors by the Tandem-GA Module.[5] This is why my focus was put on this specific module. I wanted to know what was causing the problems. I had an idea that it had to do with the calculation of the scores.

$$score_{ij} = \log_2 \left( \frac{\Pr(sp_{ij}|M_{BN})}{\Pr(sp_{ij}|M_{GC})} \right) \tag{1}$$

Equation 1, also known as the Score Equation, calculates the score based off the conditional probabilities of the sequence pairs with respect to the Bayesian Network Model and the Null Model. The offset of the score is likely caused by the small size of the Tandem-GA Module. There is a second equation which may play a minor role in this:

$$BPP_{ij} = \frac{\exp\left(-\frac{FE_{ij}}{kT}\right)}{\exp\left(-\frac{FE_{all}}{kT}\right)} \tag{2}$$

$\exp(x)$ just means $e^x$. This notation was used to save space as it would have probably been really confusing to see multiple fractions within fractions. Using exponential rules, Equation 2, also known as the Base-Pair Probability Equation, can be simplified to:

$$BPP_{ij} = e^{\frac{FE_{all} - FE_{ij}}{kT}} \tag{3}$$

This simplification makes sense because one is taking the proportion of the $FE$ in the pair with respect to the total $FE$. Equation 3 can also be skewed because of the proximity of the base pairs since the size of the Tandem-GA Module is small.

5. Cruz and Westhof, "Sequence-based identification of 3D structural modules in RNA with RMDetect".

## 5.2 BayesPairing and JAR3D

After running the Tandem-GA Modules into BayesPairing and JAR3D, it is clear that these two pieces of software perform better at interpreting these said modules. As stated above, this makes sense because both pieces of software were written more recently than RMDetect. It is also clear from the results that JAR3D is a better tool for scoring these modules. This as stated earlier, is due to the fact that these computer programs do not use the same definition for a 'sequence motif'.

JAR3D has an obvious advantage in speed. I personally had to run BayesPairing several times before I could get output. Each run took around ten to fifteen minutes to run. For some reason, there was a clear issue in parsing the giving FASTA File. Attached are two output files as PDFs and the input FASTA File. This is one clear advantage of JAR3D.

It is clear that BayesPairing is slower than JAR3D, but that may be due to design. As stated earlier in the paper, JAR3D was designed with a limited definition for 'motif'. Therefore, it may be the fact that BayesPairing is taking a more through look through the given FASTA File that is causing the slowdown. BayesPairing is probably interpreting the files and considering any additional information provided by the FASTA File.

For Reference: I have personally attached the input FASTA File along with the two outputs as PDFs. To replicate the data, one just has to copy and paste the contents into the JAR3D webserver. For BayesPairing, it is slightly more tricky as the secondary structure cannot be inputted directly. One would have to take the secondary structure out of the top line of the FASTA File and paste it into the space for the secondary structure.

On a unrelated note, this class has been very interesting. I would like to thank both instructors for the opportunity to explore bioinformatics. IT has truly been an honour to be a member of the course.