Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

APPLICATIONS NOTE

13Check_RNA: A tool to evaluate ^{13}C chemical shifts assignments of RNA

A. A. Icazatti [1,*], **O. A. Martin** [1], **M. Villegas** [1], **I. Szleifer** [2,3,4] **and J. A. Vila** [1,*]

[1] Instituto de Matemática Aplicada San Luis, Universidad Nacional de San Luis, CONICET, Avenida Italia 1556, 5700, San Luis–Argentina,
[2] Department of Biomedical Engineering,
[3] Chemistry of Life Processes Institute, and
[4] Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Chemical shifts (CS) are an important source of structural information of macromolecules such as RNA. In addition to the scarce availability of CS for RNA, the observed values are prone to errors due to a wrong re-calibration or miss assignments. Different groups have dedicated their efforts to correct CS systematic errors on RNA. Despite this, there are not automated and freely available algorithms for correct assignments of RNA ^{13}C CS before their deposition to the BMRB or re-reference already deposited CS with systematic errors.
Results: Based on an existent method we have implemented an open source python module to correct ^{13}C CS (from here on $^{13}\mathrm{C}_{exp}$) systematic errors of RNAs and then return the results in 3 formats including the nmrstar one.
Availability: This software is available on GitHub at https://github.com/BIOS-IMASL/13Check_RNA under a MIT license.
Contact: ale.icazatti@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

CS provide very valuable information about the chemical and structural environment of magnetically active nuclei such as ^{13}C. Nevertheless, as every experimental measurement, CS usually have errors. These errors could arise from signal miss assignments or an incorrect $^{13}\mathrm{C}_{exp}$ indirect re-calibration (Wishart, 1995). Such errors interfere with establishing a reliable relationship between the experimental data and the macromolecular structure. Different groups have dedicated great efforts to detect and correct these errors (Aeschbacher *et al.*, 2012; Brown *et al.*, 2015; Frank *et al.*, 2013). There are methods that allow detection and correction of CS errors in proteins (Wang and Wishart, 2005; Zhang,H. *et al.*, 2003), but not in RNA. In this scenario, Aeschbacher et al 2012 proposed a method to validate experimental $^{13}\mathrm{C}_{exp}$ databases of RNA. For this purpose, Aeschbacher *et al.*, 2012 selected a total of 5 carbon nuclei that are present in most in-vitro synthesized RNAs determined by NMR. The experimental CS values of the selected nuclei have a remarkable stability to temperature, pH and buffer variations making them very robust internal references to assess the existence of systematic errors in RNA $^{13}\mathrm{C}_{exp}$.

Based on the above mentioned method we have implemented in Python an algorithm for correct assignments, or automatic detection and correction of systematic errors in RNA $^{13}\mathrm{C}_{exp}$. Following the distinction between *structure-based* and *chemical shift-based* approaches made by (Frank *et al.*, 2013, 2014), the current algorithm must be considered as a *chemical shift-based* approach. This represent an enormous advantage of the method because no additional structural information is required. In other words, it only makes use, as input, of the chemical shifts listed in a nmrstar file.

2 Methods

2.1 Code implementation

The code was written in Python and was tested in versions 2.7, 3.5 and 3.6. Has as dependencies Pandas (McKinney *et al.*, 2012), NumPy (van der Walt *et al.*, 2011) and PyNMRSTAR (Wedell, 2017). We used Jupyter Notebooks during developing and testing (Kluyver *et al.*, 2016).

2.2 Code tests

From the BMRB server, we downloaded all nmrstar files with RNA $^{13}\mathrm{C}_{exp}$. A set of 174 nmrstar files was obtained and analyzed to search for systematic errors in $^{13}\mathrm{C}_{exp}$ using the algorithm shown in Figure 1 (Supplementary Information).

2.3 Code features

The user provides a path to a file with new $^{13}\mathrm{C}_{exp}$ data in nmrstar format. For data already deposited in the BMRB, an entry number can be given as input. If the file is not present in the working directory it will be automatically previously downloaded from the BMRB database. Alternatively, a path to a file can be provided for data downloaded from the BMRB. If a systematic error is found, the algorithm will return the corrected $^{13}\mathrm{C}_{exp}$ in one of 3 formats, two of them are files: nmrstar (default) and csv. The third option returns a Pandas DataFrame for further processing.

2.4 Workflow description

The $^{13}\mathrm{C}_{exp}$ are re-calibrated following the flow chart shown in (Figure 1 in Supplementary Information). Firstly, all $^{13}\mathrm{C}_{exp}$ are obtained from the nmrstar file. The sequence of 5' and 3' terminal nucleotides is extracted, and only if 5'–GG and a 3'–C are present in the RNA sequence, the file is analyzed. Otherwise, a warning message alerts the user. Next, the algorithm counts the number of available reference $^{13}\mathrm{C}_{exp}$. The program continues only if there are 2 or more reference values. If all the reference values fall inside the expected ranges, the dataset is reported as correct. If all values fall outside the expected ranges, the program searches for a systematic error. If a portion of the reference values is correct and another portion falls outside the expected ranges, the algorithm evaluates the existence of a systematic error in nitrogenous bases or in ribose $^{13}\mathrm{C}_{exp}$ separately, i.e. if part of the data is correct and another part has an offset. If all the previous listed conditions are fulfilled and a systematic error is likely to exist, the difference Δ, between the experimental and the average <x> of the expected chemical shift values, respectively, is computed for the corresponding reference nuclei (see Aeschbacher *et al.*, 2012 for the expected reference values). From Δ the program computes its standard deviation σ_Δ. If σ_Δ is ≤ 0.5 ppm, the $^{13}\mathrm{C}_{exp}$ are re-referenced as follows:

$$^{13}C_{corrected} = ^{13}C_{exp} - \mu_\Delta \quad (1)$$

where μ_Δ is the mean of the previously calculated Δ's. A cutoff value of σ_Δ = 0.5 ppm was optimized to increase the sensitivity to discriminate between systematic and non-systematic errors . Alternatively, the user can provide the value of this cutoff as an argument. A series of messages informs the user about problems during the execution of the algorithm, namely, if: (i) a nmrstar file cannot be read, (ii) the reference nuclei are not enough, (iii) $^{13}\mathrm{C}_{exp}$ have a non-systematic error (iv) the RNA molecule lacks the terminal sequence with the internal references necessary to apply the method. In any of these four cases the algorithm will stop.

3 Results

From the 174 nmrstar files a total of 187 $^{13}\mathrm{C}_{exp}$ datasets where analyzed. The difference is a consequence of some nmrstar files having two or more chemical shift datasets. After running the algorithm, 113 datasets where found to have the terminal 5'–GG/3'–C sequences necessary to apply the method. The remaining 74 have not the reference sequence and hence the method doesn't apply for them. Within the group of 113 datasets with the reference sequence, 107 datasets have $^{13}\mathrm{C}_{exp}$ information for at least two reference nuclei i.e. they can be analyzed. From this subset, 28 datasets were reported as correct because their reference 13C CS fall inside the expected ranges. From the remaining datasets, for which their reference $^{13}\mathrm{C}_{exp}$ fall outside the expected ranges, 19 datasets gave a $\sigma_\Delta \leq 0.5$ ppm indicating the existence of systematic errors. The remaining 60 structures gave a $\sigma_\Delta > 0.5$ ppm indicating that they have non-systematic errors i.e. a systematic correction cannot be computed. Table 1 in Supplementary Information lists the BMRB accession numbers of the 19 structures with systematic errors. BMRB entries which coincides with those reported in Aeschbacher et al 2012 are highlighted in boldface. From these 19 datasets, 2 have a systematic error but only for part of the data, i.e. the ribose sugar $^{13}\mathrm{C}_{exp}$: BMRB entry 5932, which also is reported with approximately the same systematic error (∼2.6), for part of the data in Brown *et al.*, 2015 and as 'part of the data usable' in Aeschbacher *et al.*, 2012, and BMRB entry 5919 which here we report as having a small systematic error of -0.55. As highlighted by the mentioned authors, BMRB entries with an offset near -2.7 ppm correspond to structures where indirect re-calibration to DSS of $^{13}\mathrm{C}_{exp}$ (Wishart *et al.*, 1995) was incorrectly or not applied. The source of other systematic errors remains unknown to us, but a further analysis of the experimental conditions may reveal the origin of the observed offsets.

4 Conclusion

To carry out a proper structural analysis of RNA molecules it is crucial to have access to reliable data. For this reason we automated the correction of systematic errors in $^{13}\mathrm{C}_{exp}$ of RNA, based on a method proposed by Aeschbacher et al. (2012). The user will be able to provide as input (i) new data in nmrstar format for correct assignments, (ii) a BMRB entry or (iii) a BMRB file, for re-referencing $^{13}\mathrm{C}_{exp}$ systematic errors. Additionally, the user can choose among three options for the re-referenced $^{13}\mathrm{C}_{exp}$ values output, namely as a Pandas DataFrame, a nmrstar file-type (set by default) or a csv file-type. The code and the corrected nmrstar files are available on GitHub at https://github.com/BIOS-IMASL/13Check_RNA. We encourage users to use this code in their own applications and submit issues.

Acknowledgements

We thank Jon Wedell, for helping us with the PyNMRSTAR library

Funding

This work was supported by Consejo Nacional de Investigaciones Científicas y Técnicas (PIP-0087) and Agencia Nacional de Promoción Científica y Tecnológica (PICT-0767, PICT-0218), from Argentina.

Conflict of Interest: none declared.

References

Aeschbacher,T., Schubert, M. and Allain, F.H. (2012) A procedure to validate and correct the ^{13}C chemical shift calibration of RNA datasets. *J. Biomol. NMR*, **52**, 179–190.

Breaker,R.R., (2012) Riboswitches and the RNA World. *Cold Spring Harb Perspect Biol*, **4**, 1–16.

Brown,J.D., Summers,M.F. and Johnson,B.A. (2015) Prediction of hydrogen and carbon CS from RNA using database mining and support vector regression. *J. Biomol. NMR*, **63**, 39–52.

Eddy,S.R. (2001) Non-coding rna genes and the modern rna world *Nat Rev Genet*, **2**, 919–929.

Frank,A.T., Stelzer,A.C. and Bae,S. (2013) Prediction of RNA 1H and 13C Chemical Shifts: A Structure Based Approach.*J Phys Chem B*, **117**, 13497–13506.

Frank,A.T., Law,S.M. and Brooks,C.L. (2014) A Simple and Fast Approach for Predicting H-1 and C-13 Chemical Shifts: Toward Chemical Shift-Guided Simulations of RNA. *J. Phys. Chem. B*, **118**, 12168–12175.

Geisler,S. and Coller,J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts *Nat Rev Mol Cell Biol*, **14**, 699–712.

Jore,M.M, Brouns,S.J.J. and van der Oost,J. (2011) RNA in Defense: CRISPRs Protect Prokaryotes against Mobile Genetic Elements. *Cold Spring Harb Perspect Biol*, doi: 10.1101/cshperspect.a003657

Kluyver,T. *et al.* (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. *Position. Power Acad. Publl. Play. Agents Agendas*, 87–90.

McKinney,W. (2012) pandas: A foundational Python library for data analysis and statistics. *O′Reilly Media, Inc.*

Mortimer,S.A., *et al.* (2014) Insights into RNA structure and function from genome–wide studies. *Nat Rev Genet*, **15**, 469–479.

Sabin,L.R., Delás,M.J. and Hannon,G.J. (2013) Dogma Derailed: The Many Influences of RNA on the Genome *Mol Cell.*, **43**(Issue 5), 783–794.

van Der Walt,S., Colbert,C. and Varoquaux,G.(2011) The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, **13**, (2) 22–30.

Wang,Y. and Wishart,D.S. (2005) A simple method to adjust inconsistently referenced ^{13}C and ^{15}N chemical shift assignments of proteins. *J. Biomol. NMR*, **31**, 143–148.

Wedell,J. (2017) PyNMRSTAR. [Online].

Wishart,D.S. *et al.* (1995) ^{1}H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR*, **6**, 135–140.

Zhang,H. *et al.* (2003) RefDB: A database of uniformly referenced protein CS. *J. Biomol. NMR*, **25**, 173–195.