FACTA UNIVERSITATIS
Series: **Physics, Chemistry and Technology** Vol. 18, N° 1, 2020, pp. 39 - 45
https://doi.org/10.2298/FUPCT2001039A

EVALUATION OF THE CONSTRUCTED 3D MODELS OF RNAS: A REVIEW

UDC 577.217.337.2

Biljana Arsić[1], Boris Furtula[2], Marjan Ranđelović[1]

[1]Faculty of Sciences and Mathematics, University of Niš, Niš, Serbia
[2]Faculty of Sciences and Mathematics, University of Kragujevac, Kragujevac, Serbia

Abstract. *The development of new experimental techniques, such as cryo-electron spectroscopy, enables insight into the structural features inside cells. However, in specific cases, it is still not possible to get the cryo images. Therefore, the development of the scores for the evaluation of the quality of the constructed RNAs, similarly to the proteins, is a prerequisite for the investigation of the diseases caused by the organisms not well investigated. Here, we are providing a summary of the evaluation scores in use for the prediction of the quality of the constructed 3D models of the RNAs.*

Key words: *scores, RNAs, evaluation*

1. INTRODUCTION

The RNAs are attracting attention since the 1990s due to numerous roles inside living organisms, but also because of good catalytic activities. The secondary RNA structures are available for the majority of known living organisms. Also, 3D structures of RNAs from various organisms in the form of cryo-electron microscopy images and X-ray images are available alone or in the complexes. However, in certain cases, it is almost impossible to isolate particular organelles containing RNAs. Then, it is necessary to construct the 3D structures of RNAs based on the secondary structures. The construction of the interesting segments or the whole RNAs is not so complicated nowadays and computationally expensive. Various programming packages are in use, such as ModeRNA (contain commands that enable changes on the different entities: the entire molecule, a particular region, and a single residue (Rother et al., 2011)), and SimRNA web (a web server for RNA 3D structure modeling with optional restraints (Magnus et al., 2016)). Besides these programs, other software packages were created for RNA 3D structure prediction, such as YAMPP (Malhotra et al., 1994), NAB (Macke and Case, 1998),

Received February 16th, 2020; accepted December 27th, 2020

Corresponding author: Biljana Arsić
Faculty of Sciences and Mathematics, University of Niš, Niš, Serbia
E-mail: biljana.arsic@pmf.edu.rs

ERNA-3D (Zwieb et al., 1998), MANIP (Massire and Westhof, 1998), S2S (Jossinet and Westhof, 2005), FARNA (Das and Baker, 2007), MC-Fold/MC-Sym (Parisien and Major, 2008), RNA2D3D (Martinez et al., 2008), iFOLDRNA (Sharma et al., 2008; Krokhotin et al., 2015), NAST (Jonikas et al., 2009), Assemble (Jossinet et al., 2010), HiRE-RNA (Pasquali and Derreumaux, 2010), FARFAR (Das et al., 2010), RNABuilder (Flores and Altman, 2010), OxRNA (Sulc et al., 2014), 3dRNA (Zhao et al., 2012), *etc.* A different approach from previously developed software programs is used in EvoClustRNA (guided *in silico* modeling by seeking a global helical arrangement for the target sequence that is shared across *de novo* models of numerous sequence homologs (Magnus et al., 2019)).

The structural organization of RNAs is going in discrete states or transitions that involve the organization of secondary (2D) structure or base-pairing and a cooperative transition to the 3D structure. This fact is important in computational approaches leading to the *in silico* construction of RNAs. Various program packages give various answers even for a short RNA sequence (Schlick and Pyle, 2017). Some programs worth mentioning serve for searching the specific motifs inside RNAs, like JAR3D. Its aim consists of finding possible 3D geometries for the hairpin and internal loops by matching with RNA 3D Motif Atlas when it is possible. Probabilistic scoring and other distance criteria are used for novel sequences. The score shows the ability to form the same pattern of interactions.

Some of the knowledge-based potentials used for proteins can be used for the evaluation of the models of RNA tertiary structures (Yang et al., 2016), but some new-ones were developed for RNAs. It is worth to mention the initiatives, such as RNA-Puzzles which aim is assessing the cutting edge of RNA structure prediction techniques, comparing the different methods and tools (with the elucidation of their relative strengths and weaknesses), and clarification of their limits in terms of sequence length and the complexity of the structures, the determination to get the ultimate solution to the structure prediction problem, and the promotion of the available methods which guide users for the suitable tools for different problems, and encouraging the RNA structure prediction community in their efforts to improve the current tools (Cruz et al., 2012).

2. GRAPH MODELS OF RNAS

There are several graph models developed from the 1970s. Some of the developers are Tinoco, Nussinov, and Waterman (Schlick and Pyle, 2017). Several web servers emerged recently. RAG-3D extends the RNA-As-Graphs (RAG) catalog to 3D graphs and provides the link of the solved PDB structures and 3D graphs (Zahran et al., 2015). RAGTOP program (RNA-As-Graphs-TopologyPrediction) uses the coarse-grained representation of graphs for efficient sampling of the associated conformational space (Laing et al., 2013; Kim et al., 2014; Kim et al., 2015).

3. EVALUATION

The success of the proposed 3D RNA models must follow two general criteria. The predicted model must be geometrically and topologically very close to the experimentally determined structure (used as a reference). The correctness and of the crystal structure or

NMR structure is assumed. Secondly, the stereochemistry of the predicted model must be correct. Stereochemical correctness can be checked using MolProbity (Davis et al., 2007). For the evaluation of RNA models (not applicable to proteins), firstly Capriotti et al. developed Ribonucleic Acids Statistical Potential (RASP) in 2011 (Capriotti et al., 2011; Norambuena et al., 2013). Afterward, the coarse-grained and all-atom KB potentials were proposed by Bernauer et al. in 2011 (Bernauer et al., 2011). The third in the row who developed the statistical potential (3dRNA score) was Yi Xiao's group in 2015 (Wang et al., 2015). For the evaluation of the predicted models, the root means square deviation (*RMSD*) is used together with the Interaction Network Fidelity (*INF*).

KB potential is a distance-dependent statistical potential that uses a Dirichlet process mixture model to obtain distance distributions (Bernauer et al., 2011; Neal, 2000). RASP is also a distance-dependent statistical potential which is a detailed full-atom potential including the representation of local and non-local interactions in RNA structures (Capriotti et al., 2011; Melo et al., 2002). 3dRNA score beside distance-dependent potential also uses a dihedral-dependent potential involving seven RNA dihedral angles (Wang et al., 2015). Different from all previously mentioned scores, Precision Training RNA Mark (PTRNAmark) not only considers non-bonded interactions (Yang et al., 2016). *RMSD* is defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{1}$$

where δ is the Euclidean distance between a given pair of corresponding atoms. *RMSD* is calculated for all heavy atoms, and *N* is the number of all pairs of heavy atoms.

The interaction network fidelity (*INF*) is calculated using the formula:

$$INF = \sqrt{\left(\frac{TP}{TP+FP}\right)\cdot\left(\frac{TP}{TP+FN}\right)} \tag{2}$$

where *TP* is the number of correctly predicted base-base interactions, *FP* is the number of predicted base-base interactions with no correspondence in the solution model, and *FN* is the number of base-base interactions in the solution model not present in the predicted model (Magnus et al., 2019). Then, the deformation index (*DI*) is calculated as:

$$DI = \frac{RMSD}{INF} \tag{3}$$

Deformation profile is a distance matrix computed as the average *RMSD* between the individual bases of the predicted and the reference models superposing each nucleotide of the predicted RNA model over the corresponding nucleotide of the reference model (Cruz et al., 2012). The *p*-value is calculated according to the following formula (Cruz et al., 2012):

$$p-value = \frac{1+erf\left(\frac{(RMSD-<RMSD>)/1.8}{\sqrt{2}}\right)}{2} \tag{4}$$

where $< RMSD > \; = a \cdot N^{0.41} - b$; N is the number of nucleotides; a and b are fitting parameters that depend on whether the secondary structure information is provided as an input to molecular dynamic simulation.

3.1. The K-category correlation coefficient

3.1.1. Correlation coefficient, R_K

R_K correlation coefficient is defined as:

$$R_K = \frac{cov(X,Y)}{\sqrt{cov(X,X) \cdot cov(Y,Y)}} \qquad (5)$$

In this case, Pearson's correlation coefficient is naturally extended by the application of the dot (inner) product between corresponding columns for each position in the rows. The range of values goes from [-1,1].

3.2. Accuracy measures

Assessing the prediction accuracy achieved by a given RNA secondary structure prediction procedure is measured using sensitivity, positive predictive value (*PPV*), and the *F-measure* (Aghaeepour and Hoos, 2013). *Sensitivity* can be defined as a ratio of several correctly predicted base-pairs to the number of base-pairs in the reference structure:

$$Sensitivity = \frac{\#Correctly\ Predicted\ Base - Pairs}{\#Base - Pairs\ in\ the\ Reference\ Structure} \qquad (6)$$

PPV is the ratio of the number of correctly predicted base-pairs to the number of base-pairs in the predicted structure:

$$PPV = \frac{\#Correctly\ Predicted\ Base - Pairs}{\#Base - Pairs\ in\ the\ Predicted\ Structure} \qquad (7)$$

F is defined as a harmonic mean of sensitivity and *PPV*:

$$F - measure = \frac{2 \cdot sensitivity\ \cdot PPV}{sensitivity + PPV} \qquad (8)$$

The *F-measure*, sensitivity, and *PPV* for the prediction of any structure are in the interval [0,1], where 1 characterizes a perfect prediction.

The Matthews correlation coefficient (*MCC*) is used as a single score summarizing both sensitivity and *PPV* (Matthews, 1975). It is defined as:

$$MCC = \frac{(TP\ \cdot TN - FP\ \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (9)$$

where *TP*, *TN*, *FP*, and *FN* represent the number of true positive, true negative, false positive, and false negative base pairs.

3.3. Metrics of the scoring function

To rank the near-native RNA structures, the *ES* (enrichment score) can be used which is based on the top 10% scoring ($E_{top10\%}$) and the best 10% *RMSD* values ($R_{top10\%}$), then the evaluation of their degree of overlap (this choice is arbitrary). It is defined as:

$$ES = \frac{|E_{top10\%} \cap R_{top10\%}|}{0.1 \cdot 0.1 \cdot N(decoys)} \tag{10}$$

where $E_{top10\%}$ is the number of structures with energies (scores given by scoring function) in the lowest 10% of the energy range. For *RMSD*-based *ES*, $R_{top10\%}$ is the number of structures with *RMSD* in the lowest 10%. $|E_{top10\%} \cap R_{top10\%}|$ is the intersection of $E_{top10\%}$ and $R_{top10\%}$. If the relationship is random, *ES* is equal to 1, so:

$$\text{ES}=\begin{cases} 10 & \text{perfect scoring} \\ 1 & \text{perfectly random} \\ < 1 & \text{bad scoring} \end{cases} \tag{11}$$

3.4. Evaluation of the secondary structure prediction (base-pairing and topology)

As the most accurate secondary structure prediction method it was proven the multiple sequence analysis or shape-directed to find the conserved motifs (Tan et al., 2017). The physics-based free energy minimization secondary structure predictions are still in use among the biologists' community. Current algorithms for the prediction of secondary structures prefer canonical base pairs (A-U, C-G, and G-U base pairs). Unfortunately, there are 10% of the 2028 native base pairs are non-canonical base pairs (A-G, A-C, A-A, U-U, C-U, C-C, and G-G base pairs). According to this, the accuracy of base-pair predictions cannot be more than 90% (Zhao et al., 2018). On average, only about 38% of the predicted secondary structures have identical topologies with native (Zhao et al., 2018).

4. Conclusion

In silico experiments, and subsequent evaluations show that the exact prediction of the secondary structure of RNAs is very hard, and not accurate. Therefore, further investigations must be performed on the prediction methods from a different point of view because the most successful approaches for predicting RNA tertiary structures are based on secondary structures.

References

Aghaeepour, N., Hoos, H.H., 2013. BMC Bioinform., 14, 139. doi:10.1186/1471-2105-14-139

Bernauer, J., Huang, X., Sim, A.Y.L., Levitt, M., 2011. RNA, 17, 1066-1075. doi: 10.1261/rna.2543711

Capriotti, E., Norambuena, T., Marti-Renom, M.A., Melo, F., 2011. Bioinformatics, 27, 1083-1093. doi: 10.1093/bioinformatics/btr093

Cruz, J.A., Blanchet, M.-F., Boniecki, M., Bujnicki, J.M., Chen, S.-J., Cao, S., Das, R., Ding, F., Dokholyan, N.V., Flores, S.C., Huang, L., Lavender, C.A., Lisi, V., Major, F., Mikolajczak, K., Patel, D.J., Philips, A., Puton, T., Santalucia, J., Sijenyi, F., Hermann, T., Rother, K., Rother, M., Serganov, A., Skorupski, M., Soltysinski, T., Sripakdeevong, P., Tuszynska, I., Weeks, K.M., Waldsich, C., Wildauer, M., Leontis, N.B., Westhof, E., 2012. RNA, 18, 610-625. doi: 10.1261/rna.031054.111

Das, R., Baker, D., 2007. Proc. Natl. Acad. Sci. USA, 104, 14664-14669. doi: 10.1073/pnas.0703836104

Das, R., Karanicolas, J., Baker, D., 2010. Nat. Methods, 7, 291-294. doi: 10.1038/nmeth.1433

Davis, I.W., Murray, L. W., Richardson, J. S., Richardson, D. C., 2004. Nucleic Acids Res., 32, W615-W619. doi: 10.1093/nar/gkh398

Flores, S.C., Altman, R.B., 2010. RNA, 16, 1769-1778. doi: 10.1261/rna.2112110

Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D., Altman, R.B., 2009. RNA, 15, 189-199. doi: 10.1261/rna.1270809

Jossinet, F., Ludwig, T.E., Westhof, E., 2010. Bioinformatics, 26, 2057-2059. doi: 10.1093/bioinformatics/btq321

Jossinet, F., Westhof, E., 2005. Bioinformatics, 21, 3320-3321. doi: 10.1093/bioinformatics/bti504

Kim, N., Laing, C., Elmetwaly, S., Jung, S., Curuksu, J., Schlick, T., 2014. Proc. Natl. Acad. Sci. USA, 111, 4079-4084. doi: 10.1073/pnas.1318893111

Kim, N., Zahran, M., Schlick, T., 2015. Methods Enzymol., 553, 115-135. doi: 10.1016/bs.mie.2014.10.054

Krokhotin, A., Houlihan, K., Dokholyan, N.V., 2015. Bioinformatics, 31, 2891-2893. doi: 10.1093/bioinformatics/btv221

Laing, C., Jung, S., Kim, N., Elmetwaly, S., Zahran, M., Schlick, T., 2013. PLoS One, 8, e71947. doi: 10.1371/journal.pone.0071947

Macke, T.J., Case, D.A., 1998, in: Modeling unusual nucleic acid structures. Vol. 682 of ACS Symposium Series, pp. 379-393. doi: 10.1021/bk-1998-0682.ch024

Magnus, M., Boniecki, M.J., Dawson, W.K., Bujnicki, J.M., 2016. Nucleic Acids Res., 44, W315-W319. doi: 10.1093/nar/gkw279

Magnus, M., Kappel, K., Das, R., Bujnicki, J.M., 2019. BMC Bioinform., 20, 512. doi: 10.1186/s12859-019-3120-y

Malhotra, A., Tan, R.K., Harvey, S.C., 1994. Biophys. J., 66, 1777-1795. doi: 10.1016/S0006-3495(94)80972-5

Martinez, H.M., Maizel, J.V., Shapiro, B.A., 2008. J. Biomol. Struct. Dyn., 25, 669-683. doi: 10.1080/07391102.2008.10531240

Massire, C., Westhof, E., 1998. J. Mol. Graphics Modell., 16, 197-205. doi: 10.1016/s1093-3263(98)80004-1

Matthews, B.W., 1975. Biochem. Biophys. Acta, 405, 442-451. doi: 10.1016/0005-2795(75)90109-9

Melo, F., Nchez, R., Sali, A., 2002. Protein Sci., 11, 430-448. doi: 10.1002/pro.110430

Neal, R.M., 2000. J. Comp. Graph. Stat., 9, 249-265. doi: 10.2307/1390653

Norambuena, T., Cares, J.F., Capriotti, E., Melo, F., 2013. Bioinformatics, 29, 2649-2650. doi: 10.1093/bioinformatics/btt441

Parisien, M., Major, F., 2008. Nature, 452, 51-55. doi: 10.1038/nature06684

Pasquali, S., Derreumax, P., 2010. J. Phys. Chem. B, 114, 11957-11966. doi: 10.1021/jp102497y

Rother, M., Rother, K., Puton, T., Bujnicki, J.M., 2011. Nucleic Acids Res., 39, 4007-4022. doi: 10.1093/bioinformatics/btr400

Schlick, T., Pyle, A.M., 2017. Biophys. J., 113, 225-234. doi: 10.1016/j.bpj.2016.12.037

Sharma, S., Ding, F., Dokholyan, N.V., 2008. Bioinformatics, 24(17), 1951-1952. doi: 10.1093/bioinformatics/btn328

Šulc, P., Romano, F., Ouldridge, T.E., Doye, J.P., Louis, A.A., 2014. J. Chem. Phys., 140, 235102. doi: 10.1063/1.4881424

Tan, Z., Fu, Y., Sharma, G., Mathews, D.H., 2017. Nucleic Acids Res., 45, 11570-11581. doi: 10.1093/nar/gkx815

Wang, J., Zhao, Y., Zhu, C., Xiao, Y., 2015. Nucleic Acids Res., 43, e63. doi: 10.1093/nar/gkv141

Yang, Y., Gu, Q., Shi, Y.-Z., 2016. bioRxiv. doi: 10.1101/076000

Zahran, M., Sevim Bayrak, C., Elmetwaly, S., Schlick, T., 2015. Nucleic Acids Res., 43, 9474-9488. doi: 10.1093/nar/gkv823

Zhao, Y., Wang, J., Zeng, C., Xiao, Y., 2018. Biophys. Rep., 4, 123-132. doi: 10.1007/s41048-018-0058-y

Zhao, Y.J., Huang, Y.Y., Gong, Z., Wang, Y.J., Man, J.F., Xiao, Y., 2012. Sci. Rep., 2, 734. doi: 10.1038/srep00734

Zwieb, C., Gowda, K., Larsen, N., Muller, F., 1998. In: Comparative modeling of the three-dimensional structure of signal recognition particle RNA. Vol. 682 of ACS Symposium Series. Amer. Chemical Soc., 405-413. doi: 10.1021/bk-1998-0682.ch026

EVALUACIJA KONSTRUISANIH 3D MODELA RNK: PRIKAZ

Razvoj novih eksperimentalnih tehnika, kao što je krio-elektronska mikroskopija, omogućava uvid u strukturne karakteristike unutar ćelije. Međutim, u specifičnim slučajevima, i dalje nije moguće dobiti krio slike. Zato, razviće skorova za evaluaciju kvaliteta konstruisanih modela RNK, slično proteinima, preduslov je za istraživanje bolesti koje su uzrokovane organizmima koje nisu dovoljno ispitani. Ovde su izloženi skorovi za evaluaciju koji se koriste za predviđanje kvaliteta konstruisanih 3D modela RNK.

Ključne reči: *skorovi, RNK, evaluacija*