ACTA ABP BIOCHIMICA POLONICA

Vol. 63, No 4/2016
753–757
https://doi.org/10.18388/abp.2016_1333

Regular paper

StructAnalyzer – a tool for sequence *versus* structure similarity analysis

Jakub Wiedemann[1*] and Maciej Miłostan[1,2*]✉

[1]Institute of Computing Science, Poznan University of Technology, Poznań, Poland; [2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

In the world of RNAs and proteins, similarities at the level of primary structures of two comparable molecules usually correspond to structural similarities at the tertiary level. In other words, measures of sequence and structure similarities are in general correlated – a high value of sequence similarity imposes a high value of structural similarity. However, important exceptions that stay in contrast to this general rule can be identified. It is possible to find similar structures with very different sequences, as well as similar sequences with very different structures. In this paper, we focus our attention on the latter case and propose a tool, called StructAnalyzer, supporting analysis of relations between the sequence and structure similarities. Recognition of tertiary structure diversity of molecules with very similar primary structures may be the key for better understanding of mechanisms influencing folding of RNAs or proteins, and as a result for better understanding of their function. StructAnalyzer allows exploration and visualization of structural diversity in relation to sequence similarity. We show how this tool can be used to screen RNA structures in Protein Data Bank (PDB) for sequences with structural variants.

Key words: sequence similarity, structural similarity, RNA

Received: 30 May, 2016; revised: 28 June, 2016; accepted: 19 July, 2016; available on-line: 02 November, 2016

INTRODUCTION

Despite technological progress in laboratory pipelines, computing methods and computational facilities, determination of three dimensional structures of RNAs and proteins in-situ or in-silico is not a trivial task (cf. Lukasiak *et al*., 2010). Comparison of deposition statistics between Protein Data Bank (PDB) (Berman *et al*., 2000) and NCBI's RefSeq (Pruitt *et al*., 2012), shows how large is the gap between the known sequences and structures. In-silico methods attempt to reduce this gap but as the RNA-Puzzles (Miao *et al*., 2015) competition has shown, they are still far from being perfect.

Nowadays, the most successful structure prediction methods are often somehow based on correlations between the sequence and structure similarities, for example they are transformed in the form of libraries of fragments like in the RNA Composer (Popenda *et al*., 2012) and FARNA (Cheng *et al*., 2015; Das & Baker, 2007). It is a known fact that the similarity in structure (cf. (Zok *et al*., 2014) for information about structural similarities) of molecules, like proteins or RNAs, highly correlates with sequence similarities, under assumption that all of the structures compared where obtained under similar conditions. Similar conditions are important from the perspective of thermodynamics – changes in conditions are the driving force of folding and unfolding. However, in practice it is not feasible to impose the same conditions for all molecules in the process of structure determination because of various factors, e.g. physiological conditions of molecular activity and stability. Let us stress that the RNA structures, in comparison to proteins, are more flexible and less thermodynamically stable due to a larger number of degrees of freedom (Rother *et al*., 2011) (e.g. torsional angles in the backbone). Thus, we can assume that even small changes in the environment may cause a substantial change in the RNA conformation. The intriguing question is: *how structurally diverse are similar RNA sequences whose structures are deposited in PDB*?

Thus, the primary aim of our work is to provide a tool, called StructAnalyzer, that allows us to explore and visualize structural diversity in relation to sequence similarity for RNAs and proteins. In contrast to other similar tools, like RNAlyzer (Lukasiak *et al*., 2013) or RNAssess (Lukasiak *et al*., 2015), our aim is not to assess quality of the model versus the reference structure, but rather the analysis of structural diversity of the real structures determined by biochemical experiments (e.g. crystallography or NMR). This exploration should allow to identify twilight zones where the high sequence similarity does not impose structural identity. It is worth to note that, purposely, we would like to analyze only sets of highly similar sequences (90–100% of pairwise similarity). We do not want to construct a minimal library of structural fragments that covers as large area of the sequence space as possible (in such a case, it is common to keep the sequence similarity to below some level). We would like to support identification and visualization of structural variants of almost identical sequences. We assume that within clusters obtained by grouping molecules by sequence similarity, it should be possible to find diverse structures. Moreover, within these structures it should be possible to identify fragments with a relatively high and low stability. It is worth to note that some of the structures stored in PDB were obtained as complexes or in the presence of metal ions or with ligands and immersed in different chemical solutions. Interactions between proteins, RNAs, ions and ligands may lead to substantial structural changes. Experiments with proteins (Alexander *et al*., 2009) showed that sometimes even a point mutation, or a small set of point mutations, in the sequence

✉e-mail: Maciej.Milostan@cs.put.poznan.pl
*Contributed equally to this work
Abbreviations: PDB, Protein Data Bank

Table 1. Molecules' PDB IDs and general description

PDB ID	Description from PDB database
2O3X	Crystal structure of the prokaryotic ribosomal decoding site complexed with paromamine derivative NB30
3BNT	Crystal structure of the homo sapiens mitochondrial ribosomal decoding site in the presence of [CO(NH3)6]CL3 (A1555G mutant, BR-derivative)
1FYO	Eukaryotic decoding region A-site RNA

can switch the structure into a totally different structural fold. We believe that such cases also exist in RNAs and our tool may help to identify them.

MATERIALS AND METHODS

Data sources. We show features and test performance of StructAnalyzer on 3 datasets. First set consists of two proteins differing in single amino acid and originating from the paper by Alexander and coworkers (2009). The last two sets contain only RNAs and were generated by the following approach.

In the first step, we generated pairwise sequence alignments for all possible pairs of RNA structures deposited in PDB and computed a matrix of relevant similarity scores. For that purpose, we used the MUSCLE software (Edgar, 2004; http://www.drive5.com/muscle/) which accepts FASTA files as input. Be aware of the fact that the FASTA sequences stored in the PDB database sometimes differ from the sequences contained in the structure files (in particular if we consider a specific chain). Thus, we extracted sequences of the RNA molecules directly from the files containing structures (*.pdb) by means of a self-written Python script. This script generates one FASTA sequence file for each chain of molecules stored in a particular pdb file.

In the second step, based on the above mentioned matrix of sequence similarities, we constructed two sets of molecules. The first one contains all the pairs of structures having 100% sequence similarity and the second one consists of all the pairs of structures with sequence similarity over or equal to 90%, but less than 100%. Relations between pairs of molecules from each set had been depicted in the form of a graph (see Supplementary Data for details). Molecules are denoted by vertices which are labelled using relevant PDB IDs. Edges connect the molecules (denoted by vertices) with a similarity score over the defined cut off. It is worth noting that in case of both datasets we obtained graphs containing disjoint subgraphs.

The results of procedures described above for both sets are presented in the Supplementary Data (at www.actabp.pl). From the first set, containing pairs of sequences with 100% sequence similarity, the algorithm created 383 subgraphs. For the second set (sequences with similarity above 90%, but less than 100%) we obtained 93 subgraphs. From both sets we chose one subgraph to show features of the presented tool.

Algorithm description. The tool presented here allows to perform both, one-to-many and many-to-many sequence and structure comparisons. Our program uses PDB and Multi-FASTA files as input. On the basis of the data obtained and computational analysis, StructAnalyzer generates graphical interpretation of the results. The general workflow of StructAnalyzer is shown in Fig. 1.

In the first stage, our algorithm generates sequence alignment using the MUSCLE software. This alignment is the basis for further analysis. We can distinguish two general modes of comparisons: many-to-one (one sequence is treated as the reference one) and many-to-many. Results of each mode are visualized in a different manner.

In both cases (many-to-one and many-to-many) the algorithm selects corresponding fragments of sequences based on the sequence alignment. Selected fragments are aligned with corresponding fragments of the reference structure and the algorithm calculates their structural similarity. RMSD is used as a measure of structural similarity. The program also allows merging of spatially neighbouring fragments into larger entities to increase the number of atoms used to perform the structural comparisons. To do this, the algorithm searches the spatial neighbourhood of each of the atoms of the previously obtained fragments. The scope of the spatial neighbourhood is restricted by the user defined radius (in Angstroms). The identified neighbours are added to the base fragment. Fragments extended by the added atoms are aligned and similarity of their structures is calculated.

In case of pairwise comparison, besides the previously described function, StructAnalyzer allows to perform a comparison of all fragments with a predetermined length of one molecule, to fragments (with the same length) of another molecule. The predetermined length is further referenced as the frame.

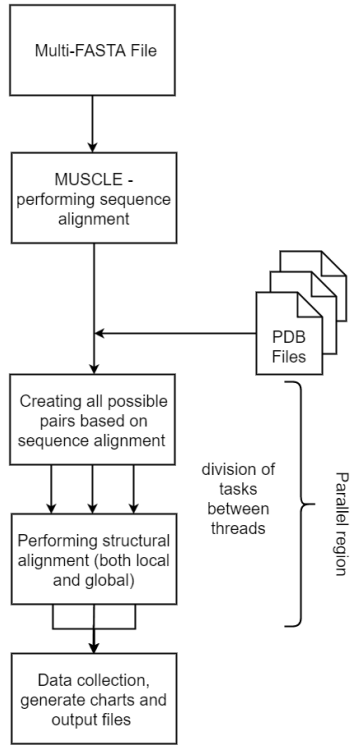

Figure 1. StructAnalyzer workflow.
The scheme presents the most important steps of the analysis performed by StructAnalyzer.

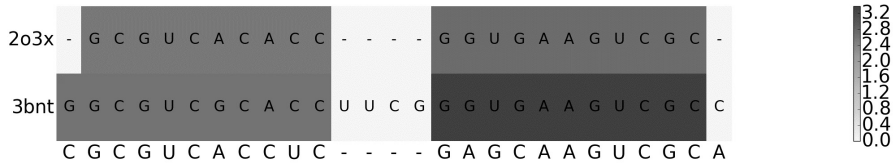

Figure 2. Example of a linear alignment for 1fyo molecule against other structures (Table 1) in the set of molecules with sequence similarity 70%.

RESULTS AND THEIR REPRESENTATION

StructAnalyzer allows the user to save their results to a .csv file but its undeniable advantage is an ability to visualize them. For global alignment of structures, our tool presents the results using heat maps and a linear diagram (see Fig. 2). The linear diagram is particularly useful for comparing a sequence with low similarity or discontinuous fragments. In local alignment, we need to consider two cases. The first one is when dealing with large gaps in the alignment is necessary. The results can be visualized as a linear alignment with scores for each fragment determined individually. The second considered case is a situation when the local alignment is determined using a previously described frame. In this case, the results are shown on a heat map.

DISCUSSION

In order to show the capability of our tool, we conducted analysis for the three previously described sets. For each set, the StructAnalyzer determined the RMSD value for all molecules by performing both, a local and global alignment.

The first set consists of two protein structures differing in a single amino acid: 2KDM and 2KDL (Fig. 3). For molecules with such high sequence similarity, the results are surprising. The heat map (Fig. 4) for the global comparison shows the RMSD value is above 12 Angstroms. If we consider local comparison of this structures (Fig. 5), we can see some resemblance at the diagonal (or regions close to the diagonal) of the heat map. As we can easily deduce, high similarity scores at the diagonal indicate the identity of local structures for alignment under consideration, while similarities at the regions surrounding the diagonal can signal potential mismatches in the proposed alignment. The case under consideration shows that even a point mutation can influence the structure and function to a large extent. From the perspective of function, it is worth to stress that both proteins have affinity to bind different molecules (see Table 2).

The second set consists of six RNA structures. As shown in the heat map (Fig. 6), the RMSD values within the set range from 0.5 to about 4 Angstroms. The results are quite unexpected considering the sequence similarity in the presented collection (equal to 100%). It is easy to spot, in that case, how large influence on the structure of the RNA the environmental conditions have. In order to demonstrate factors affecting the development of the analysed molecules, a brief description (extracted from PDB) of the structures has been gathered in Table 3. Despite large differences at the level of global alignment (see Fig. 6), it is worth to take a look at the differences at the local level. The results of the local alignment are presented in heat maps (see Figs. 7 and 8), generated for the structure pairs 2CD3 and 2CD6 (with the frame sizes equal to 5; Fig. 7; and 7; Fig. 8). As we can see, at the local alignment level there are many fragments with either good or very bad RMSD values. Based on these results we can deduce that despite big differences between the molecules observed from a global perspective, when we consider the local perspective, e.g. smaller fragments of structures, we can find many similarities. These similar fragments in globally different structures can stand for conservative regions which are characterized by low volatility and may determine similar functions of the considered molecules. On the other hand, sequence fragments which in many structures are characterized by a significant diversity, may designate potentially disordered regions. Another example of local comparison is presented in a heat map (Fig. 9) for the 1F7G and 1F7I structures (Fig. 10). In

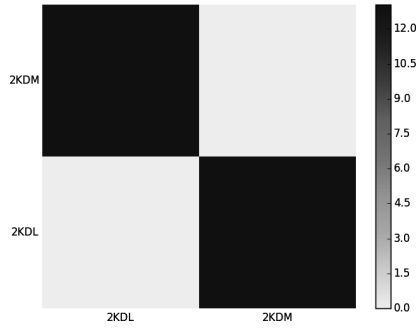

Figure 4. Heat map for global comparison of 2KDL and 2KDM structures.
The value of RMSD determines the colour.

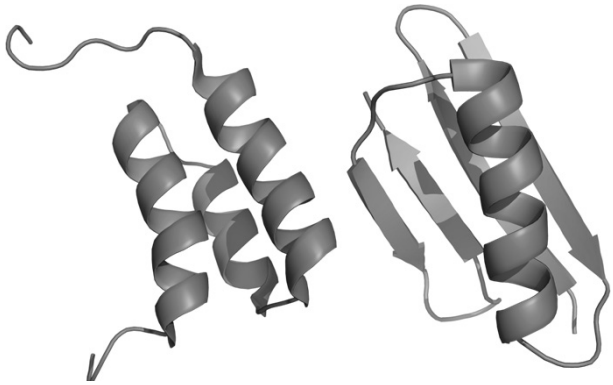

Figure 3. 2KDL (left) and 2KDM (right) spatial structures.

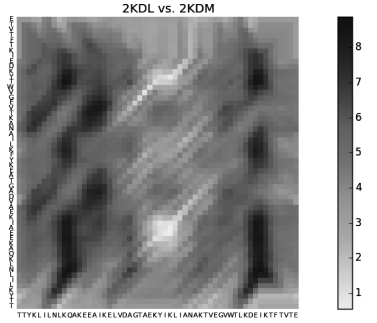

Figure 5. Heat map for structures (PDB IDs = 2KDL, 2KDM) differing by only one amino acid.
The heat map was generated by using a frame size equal to 15.

Table 2. Molecules' PDB Ids, general description and classification

PDB ID	Description from PDB database	Classification
2KDM	NMR structures of GA95 AND GB95, two designed proteins with 95% sequence identity but different folds and functions	IGG binding protein
2KDL	NMR structures of GA95 AND GB95, two designed proteins with 95% sequence identity but different folds and functions	Human serum albumin binding protein

Table 3. Molecules' PDB IDs and general description

PDB ID	Description from PDB database
1F6Z	Solution structure of the RNAse P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide
1F7I	Solution structure of the RNAse P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide complexed with cobalt (III) hexamine, NMR, ensemble of 12 structures
1F7G	Solution structure of the RNAse P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide, ensemble of 17 structures
1F79	Solution structure of RNAse P RNA (M1 RNA) P4 stem C70U mutant oligoribonucleotide complexed with cobalt (III) hexammine, NMR, minimized average structure
2CD3	Refinement of RNAse P P4 stemloop structure using residual dipolar coupling data – C70U mutant
2CD6	Refinement of RNAse P P4 stemloop structure using residual dipolar coupling data, C70U mutant cobalt (III) hexammine complex

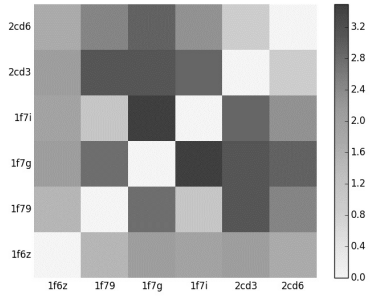

Figure 6. Heat map for all molecules against each other.
The value of RMSD determines the colour.

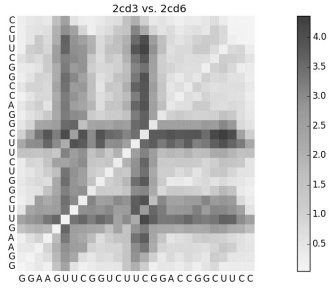

Figure 7. Heat map for structures (PDB IDs = 2CD3, 2CD6) with the same sequence.
The heat map was generated by using a frame size equal to 5.

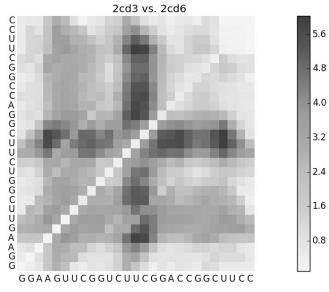

Figure 8. Heat map for structures (PDB IDs = 2CD3, 2CD6) with the same sequence.
The heat map was generated by using a frame size equal to 7.

this case, at the diagonal (and near the diagonal) of the heat map we can see similar and dissimilar fragments. Dissimilarities can be the result of a metal (cobalt) presence during the structure determination process of the 1F7G molecule. This case shows how the environment can influence folding of the structure and also how even small structural differences at the local level can change overall fold of the analysed molecule.

The third set contains 3 RNA structures. Sequence similarity between molecules that were at the edges of the subgraph containing the analysed structures are shown in Table 4. As in the previous example, despite high sequence similarity we can observe significant structural differences between all molecules (see Fig. 11). From the analysis of local comparison we can see a huge range of RMSD, from 0 to almost 6 Angstroms. It is worth noting the obvious fact that in the case of the analysed molecules, the best RMSD values for the fragments compared are most frequently located at the diagonal of the presented heat map (see Fig. 12). Consideration of values outside of the diagonal may be useful in detection of misalignments and when we look for reoccurring local spatial motifs between the analysed molecules – e.g. larger or smaller affinity for the sequence to adopt some spatial structure.

CONCLUSIONS

StructAnalyzer is a new, promising tool for structural analysis of RNA and proteins. This tool is still under active development, and thus new features will be incorporated shortly; this tool will be available for the general

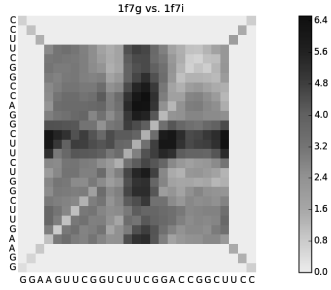

Figure 9. Heat map for structures (PDB IDs = 1F7G, 1F7I) with the same sequence.
The heat map was generated by using a frame size equal to 7.

Table 4. Sequence similarity between molecules in the analyzed set

PDB ID:CHAIN ID	PDB ID 2:CHAIN ID	Sequence similarity
1N66:A	2GRW:A	0.95
1N66:A	2GV4:A	0.95
2GRW:A	2GV4:A	0.95
PDB ID:CHAIN ID	Description from PDB database	
1N66:A	Structure of the pyrimidine-rich internal loop in the Y-domain of poliovirus 3′-UTR	
2GRW:A	Solution structure of the poliovirus 3′-UTR Y-stem	
2GV4:A	Solution structure of the poliovirus 3′-UTR Y-stem	

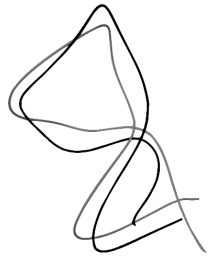

Figure 10. Superposition of spatial structures of 1F7G (black) and 1F7I (grey).

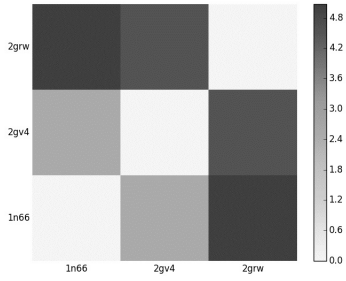

Figure 11. Heat map for all molecules against each other.
The value of RMSD determines the colour.

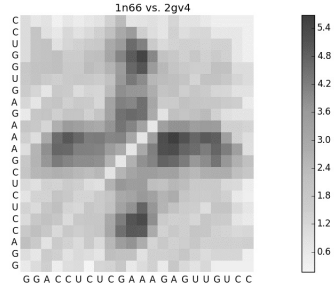

Figure 12. Heat map for structures (PDB IDs = 2F88, 2LPT) with sequence similarity equal to 95%.
The heat map was generated by using a frame size equal to 5.

public (structanalyzer.cs.put.poznan.pl). In the current release it can perform both, global and local structure comparisons on the basis of sequence alignment and visualize the obtained results in an attractive manner. The presented approach enables to examine, for example, how different conditions or sequence differences affect development of the structures. Moreover, a visual representation of the results makes them much easier to interpretate. Global comparison of structures shows us, in general, if there are any differences. In large sets of structures it allows us to screen through the whole set, so there is no need to examine the structures one by one, and it immediately indicates where and how big these structural differences are. After global comparison, we can decide for which structures we want to run the comparison locally or we can terminate the job. Results of local comparison provide us information about influence of the local differences, like point mutations or deletions, on the global shape of the molecule. Those results also allow identification of potential conservative or disordered regions. Another important feature of the tool presented here, is support for parallel processing which significantly reduces the duration of analysis.

Acknowledgements

Supported by the Polish Ministry of Science and Higher Education, under the KNOW program.

REFERENCES

Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA* **106**: 21149–21154. doi: 10.1073/pnas.0906408106

Berman HM, Westbrook J, Feng Z., Gilliland G., Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242. doi: 10.1093/nar/28.1.235

Cheng CY, Chou F-C, Das R, (2015) Chapter Two – modeling complex RNA tertiary folds with Rosetta. In *Methods in Enzymology*, Chen S-J, Burke-Aguero DH, eds, **55**: 35–64. Academic Press. doi:10.1016/bs.mie.2014.10.051

Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci USA* **104**: 14664–14669. doi:10.1073/pnas.0703836104

Edgar RC (2004a) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797. Print 2004. PubMed PMID: 15034147. doi: 10.1093/nar/gkh340

Edgar RC (2004b) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113. PubMed PMID: 15318951. doi: 10.1186/1471-2105-5-113

Lukasiak P, Blazewicz J, Milostan M (2010) Some operations research methods for analyzing protein sequences and structures. *Annals Operations Res* **175**: 9–35

Lukasiak P, Antczak M, RatajczakT,Szachniuk M, Popenda M, Adamiak RW, Blazewicz J (2015) RNAssess – a webserver for quality assessment of RNA 3D structures. *Nucleic Acids Res* **43**: W502–W506. doi:10.1093/nar/gkv557

Lukasiak P, Antczak M, Ratajczak T, Bujnicki JM, Szachniuk M, Popenda M, AdamiakRW,Blazewicz J (2013) RNAlyzer – novel approach for quality analysis of RNA structural models. *Nucleic Acids Res* **41**: 5978–5990. doi:10.1093/nar/gkt318

Miao Z, Adamiak RW, Blanchet M-F, Boniecki M, Bujnicki JM, Chen S-J, Cheng C, Chojnowski G, Chou F-C, Cordero P, Cruz JA, Ferre-D'Amare A, Das R, Ding F, Dokholyan NV, Dunin-Horkawicz S, Kladwang W, Krokhotin A, Lach G, Magnus M, Major F, Mann TH, Masquida B, Matelska D, Meyer M, Peselis A, Popenda M, Purzycka KJ, Serganov A., Stasiewicz J, Szachniuk M, Tandon A, Tian S, Wang J, Xiao Y, Xu X, Zhang J, Zhao P, Zok T, Westhof E (2015) RNA-puzzles round II: Assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **21**: 1–19. doi:10.1261/rna.049502.114

Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW, (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res* **40**: e112. doi:10.1093/nar/gks339

Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130–D5. PMID: 22121212. PMCID: PMC3245008. doi:10.1093/nar/gkr1079

Rother K, Rother M, Boniecki M, Puton T, Bujnicki JM (2011) RNA and protein 3D structure modeling: similarities and differences. *J Mol Model* **17**: 2325–2336. doi:10.1007/s00894-010-0951-x

Zok T, Popenda M, Szachniuk M (2014)MCQ4Structures to compute similarity of molecule structures.*Central Eur J Operations Res* **22**: 457–474. doi:10.1007/s10100-013-0296-5