Received: 16 August 2022 | Revised: 28 October 2022 | Accepted: 1 November 2022

DOI: 10.1002/pro.4503

TOOLS FOR PROTEIN SCIENCE

THE PROTEIN SOCIETY WILEY

1D2DSimScore: A novel method for comparing contacts in biomacromolecules and their complexes

S. Naeim Moafinejad | **Iswarya P. N. Pandaranadar Jeyeram** | **Farhang Jaryani** | **Niloofar Shirvanizadeh** | **Eugene F. Baulin** | **Janusz M. Bujnicki**

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

Correspondence
Janusz M. Bujnicki, Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109 Warsaw, Poland.
Email: janusz@iimcb.gov.pl

Funding information
European Molecular Biology Organization, Grant/Award Number: ALTF 525-2022; Narodowe Centrum Nauki, Grant/Award Numbers: 2017/25/B/NZ2/01294, 2017/26/A/NZ1/01083

Review Editor: Nir Ben-Tal

Abstract

The biologically relevant structures of proteins and nucleic acids and their complexes are dynamic. They include a combination of regions ranging from rigid structural segments to structural switches to regions that are almost always disordered, which interact with each other in various ways. Comparing conformational changes and variation in contacts between different conformational states is essential to understand the biological functions of proteins, nucleic acids, and their complexes. Here, we describe a new computational tool, 1D2DSimScore, for comparing contacts and contact interfaces in all kinds of macromolecules and macromolecular complexes, including proteins, nucleic acids, and other molecules. 1D2DSimScore can be used to compare structural features of macromolecular models between alternative structures obtained in a particular experiment or to score various predictions against a defined "ideal" reference structure. Comparisons at the level of contacts are particularly useful for flexible molecules, for which comparisons in 3D that require rigid-body superpositions are difficult, and in biological systems where the formation of specific inter-residue contacts is more relevant for the biological function than the maintenance of a specific global 3D structure. Similarity/dissimilarity scores calculated by 1D2DSimScore can be used to complement scores describing 3D structural similarity measures calculated by the existing tools.

Abbreviations: 1D, one dimensional; 2D, two dimensional; 3D, three dimensional; BA, balanced accuracy; CAD, contact area difference; CASP, critical assessment of structure prediction; CSI, critical success index; FM index, Fowlkes–Mallows index; FN, false negative; FP, false positive; FOR, false omission rate; GDT, global distance test; INF, interaction network fidelity; J index, Jaccard index; LDDT, local distance difference test; MCC, Matthew's correlation coefficient; MK, markedness; PT, prevalence threshold; QCS, quality control score; RMSD, root mean square deviation; RPF, recall, precision, and F-measure; TM-Score, template modeling score; TN, true negative; TP, true positive.

KEYWORDS

macromolecular complexes, molecular contacts, protein structure, RNA structure, structural bioinformatics, structure comparison

1 | INTRODUCTION

With a rapidly growing pool of macromolecular three-dimensional (3D) structures obtained by experimental methods and computational predictions, comparing these 3D structures has become increasingly important. The biologically relevant structures of proteins and nucleic acids and their complexes are often dynamic. They combine stable, rigid structural segments that make transient contacts stabilized by weak interactions, as well as structural switches and largely disordered regions whose conformation heavily relies on non-covalent contacts with the environment. Assessing quantitatively structural changes and the variation of contacts between different conformational states is essential for understanding the biological functions of proteins, nucleic acids, and their complexes.

Measuring similarity between different structural models of the same macromolecule is a common task in computational structural biology, especially in the field of 3D structure prediction, as both the development of structure prediction methods and their benchmarking depend on the comparison of computationally predicted and experimentally determined structures.[1] For example, progress in protein structure prediction has been monitored and furthered by new methods for assessing model accuracy that included comparisons of 3D structures.[2] Similarly, progress in the field of RNA 3D structure prediction has been monitored and stimulated by new methods for RNA structure comparison.[3,4] It is well known that the problem of macromolecular structure comparison is multiparametric[5] and that no single, universally acceptable measure can describe all important aspects of the similarities and differences between the compared macromolecular structures. Therefore, different methods of structure comparison and different measures of structure similarity/dissimilarity are applicable for evaluating different aspects of macromolecular structure similarity. Combining several different approaches focusing on different aspects of the structure is considered the preferred approach.

Most existing tools developed for macromolecular structure comparison directly consider the 3D structures. Root mean square deviation (RMSD) of 3D coordinates superimposed as rigid bodies[6] is one of the oldest and still the most commonly used measures of pairwise structural similarity. Other commonly used methods, developed mainly for the comparison of protein 3D structures, include Template Modeling (TM) score,[7] Global Distance Test (GDT) score,[8] Contact Area Difference (CAD) score,[9] Local Distance Difference Test (LDDT),[10] SphereGrinder,[11] Recall, Precision, and F-measure (RPF) score,[12] and Quality Control Score (QCS).[13] Measures typically used for RNA 3D structure comparison include RMSD and TM-Score developed originally for proteins, as well as RNA-specific ones, including Interaction Network Fidelity (INF) and Deformation Index, which combines RMSD and INF.[4]

Many of the measures of structural similarity listed above have been developed for comparing relatively rigid structures, they require superposition of these structures, and they are not easily applicable for the comparison of molecules that exhibit significant flexibility. Some algorithms were developed to deal with flexibility by breaking the structures into smaller units that are superimposed independently, for example, in FATCAT for proteins[14] and SupeRNAlign for RNA.[15] Another approach involves the shifting of the level of comparisons from the entire structures to the individual structural elements, usually down to the level of the local environment of individual residues, which does not require superposition of structures, for example, in QCS, SphereGrinder, CAD-score, LDDT, RPF and INF.

Structural comparisons are important not only when the 3D structures are available but also if the structural information is limited. Recently developed Deep Learning-based methods for 3D structure prediction, in particular, AlphaFold2,[16] provide high-accuracy models for proteins. However, for nucleic acids (in particular RNA) and protein–nucleic acid complexes, the generation of 3D models is not yet a routine procedure. On the other hand, numerous tools have been developed for predicting contacts between proteins and nucleic acids.[17,18] Besides, the most common way of describing RNA structure is in the form of 2D contacts, which may range from only canonical base pairs as in RNA secondary structure, or may also involve non-canonical pairs and stacking interactions. There exist numerous tools for extracting contact information from proteins and nucleic acid structures, for example, PROTMAP2D,[19] RNAMAP2D,[20] DSSR,[21] and ClaRNA,[22] which facilitate the comparison of structural information for macromolecules with known 3D structures, and the ones for which only the contact data are available.

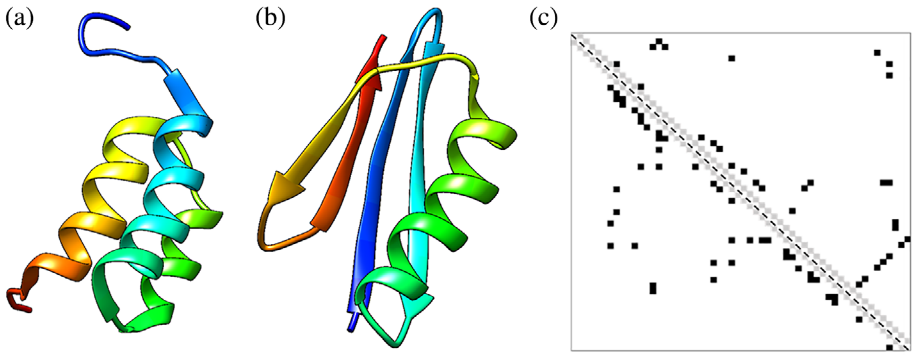
Here, we describe a new tool, 1D2DSimScore, that facilitates the comparison of all kinds of macromolecular complexes, including proteins, nucleic acids, and other molecules, for example, small molecule ligands, with a focus on contacts and contact interfaces. The main area of 1D2DSimScore applications is the comparison of macromolecules that may undergo massive conformational changes and the study of processes where the focus is on local interactions within and between molecules. Structure comparison with 1D2DSimScore may utilize 3D structural information but is not dependent on the availability of 3D structures and has been developed to complement other existing approaches that focus on other aspects of molecular structure.

2 | RESULTS

2.1 | Example application of 1D2DSimScore—Comparison of conformations for very similar molecules

Figure 1 illustrates the structures of two proteins of nearly identical sequence, obtained by computational design to fold into two completely different 3D folds as a result of only a single amino acid difference.[23] Using 1D2DSimScore tools, pairwise non-covalent contacts between amino acid residues were extracted (contacts between consecutive residues were ignored), and residues classified as in contact were indicated. Table 1 shows the results of similarity analysis with 1D2DSimScore for 1D-01 and 2D-01 formats, that is, for contacting residues and for contact pairs, respectively.

The results of comparisons obtained with 1D2DSimScore show that the contacts in the two protein variants exhibit only minimal similarity. While the maps of interacting residues are partially similar between the structures (as evident from moderate values of F1, FM, J, or precision calculated for the 1D-01 format), most of the pairwise interactions have changed, as shown by generally low values of similarity measures for the 2D-01 format. For the 2D-01 format, the high value of specificity as well as a positive value of MCC result from a large number of negative class instances and hence a very large value of TN (as visualized by a large overlap between the white spaces in 2D maps in Figure 1).



(d)

```
TTYKLILNLKQAKEEAIKELVDAGTAEKYIKLIANAKTVEGVWTLKDEIKTFTVTE
X..X.X..XXXXXXXXXXXXX.XXX.X.XX.XX..X.XXXXXXXXXXXXX......

TTYKLILNLKQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE
XXXXXX.XX..XX.XXXXXX.XX.XX.XXX.XX.XXXX..XXX.X.X.X.XXXXX.
```

FIGURE 1 1D2DSimScore input for intra-chain contacts comparison of structures of GA95 and GB95, two designed proteins with only a single amino acid difference but different folds and functions (RCSB PDB codes 2kdl and 2kdm) defined as at least four pairs of atoms found at a distance ≤3 Å between two different residues with the exclusion of consecutive residues. (a, b) Cartoon representations of the GA95 and GB95 structures, respectively; (c) 2D maps of interactions for GA95 (bottom left triangle) and GB95 (upper right triangle), with contacts corresponding to non-covalent interactions shown in black, excluded contacts corresponding to covalent interactions shown in gray and the diagonal indicated by a dashed line; (d) sequence and intra-chain interaction information according to the 1D-01 format

TABLE 1 Contact map similarity between GA95 and GB95 calculated with 1D2DSimScore using 1D-01 (similarity of contacting residues) and 2D-01 (similarity of contact pairs)

	MCC	F_1	FM	J	Specificity	Precision	Recall
1D-01	−0.096	0.667	0.667	0.5	0.222	0.650	0.684
2D-01	0.030	0.053	0.054	0.027	0.980	0.067	0.044

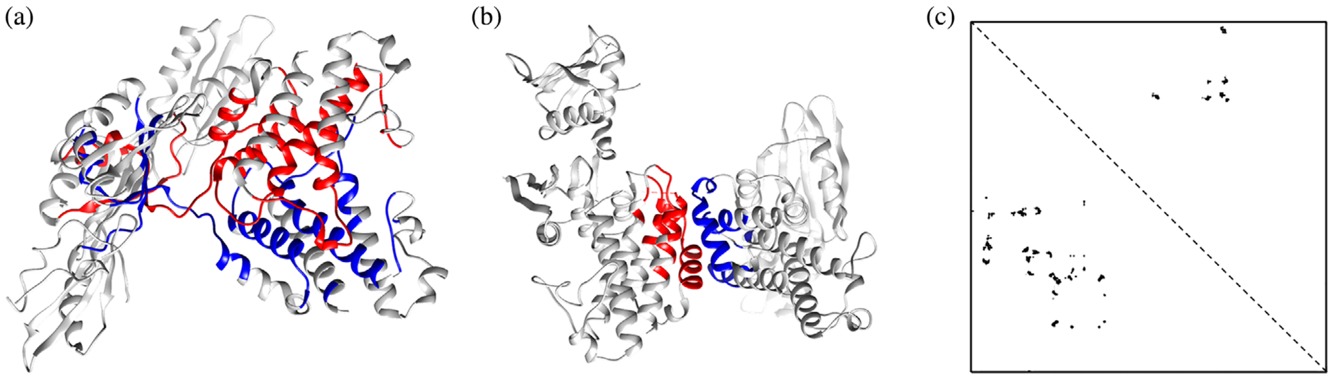


FIGURE 2 1D2DSimScore input for comparison of two alternative dimerization modes of Escherichia coli RnlA endoribonuclease (alternative dimerization is required for activity and inhibition of the HEPN ribonuclease RnlA, RCSB PDB codes 6y2p and 6y2q); (a) cartoon representation of 6y2p (B) cartoon representation of 6y2q; interacting residues are colored in red for chain A and blue for chain B; (c) 2D maps of complex interface for 6y2p (bottom left triangle) and 6y2q (upper right triangle), with contacts corresponding to non-covalent interactions shown in black, and the diagonal indicated by a dashed line

	MCC	F_1	FM	*J*	Specificity	Precision	Recall
1D-01	0.214	0.308	0.368	0.182	0.940	0.679	0.199
2D-01	0.678	0.540	0.069	0.027	0.999	0.144	0.033

TABLE 2 Contact map similarity between two different interfaces in alternative dimers of the same protein (RCSB PDB codes 6y2p and 6y2q) calculated with 1D2DSimScore using 1D-01 (similarity of contacting residues) and 2D-01 (similarity of contact pairs)

2.2 | Example application of 1D2DSimScore—Comparison of two alternative dimerization interfaces in a protein homodimer complex

Figure 2 illustrates the structures of two variants of dimeric structures determined experimentally for *Escherichia coli* RnlA endoribonuclease (RCSB PDB codes 6y2p and 6y2q). While the molecules studied in the two independent experiments are essentially the same, the 3D coordinates are not identical, as in the two structures different sections of the chain are disordered, and in this case, the analysis required the identification of a subset of residues common to both structures. Using 1D2DSimScore tools, pairwise non-covalent contacts between amino acid residues were extracted (contacts between consecutive residues were ignored), and residues classified as in contact were indicated. Table 2 shows the results of similarity analysis with 1D2DSimScore for 1D-01 and 2D-01 formats, that is, for contacting residues and for contact pairs, respectively.

The results of comparisons obtained with 1D2DSimScore show that the dimer interfaces n the two complexes are dramatically different. The interacting residues exhibit very little overlap, and contact pairs are even less similar, as evident from low values of most measures both for the 1D-01 and 2D-01 formats. The high values of specificity as well as positive values of MCC result from a large number of negative class instances and hence a very large value of TN.

2.3 | Example application of 1D2DSimScore—Comparison of an experimentally determined structure with a computational model for a protein–RNA complex

Figure 3 illustrates the experimentally determined structure of a protein–RNA complex phage P22 N peptide complexed with box B RNA (RCSB PDB code, 1a4t), with a computational model of the same molecule. The overall 3D structure of the computational model appears to be qualitatively similar to the reference structure determined experimentally. The quantitative analysis by 1D2DSimScore reveals that indeed the contact interfaces do overlap between the two structures (moderate values of F1, FM, and J, and a relatively high value of precision for the 1D-01 format), but none of the protein–RNA contact pairs were predicted correctly in the computational model (zero values of F1, FM, J, and precision for the 2D-01 format; Table 3).

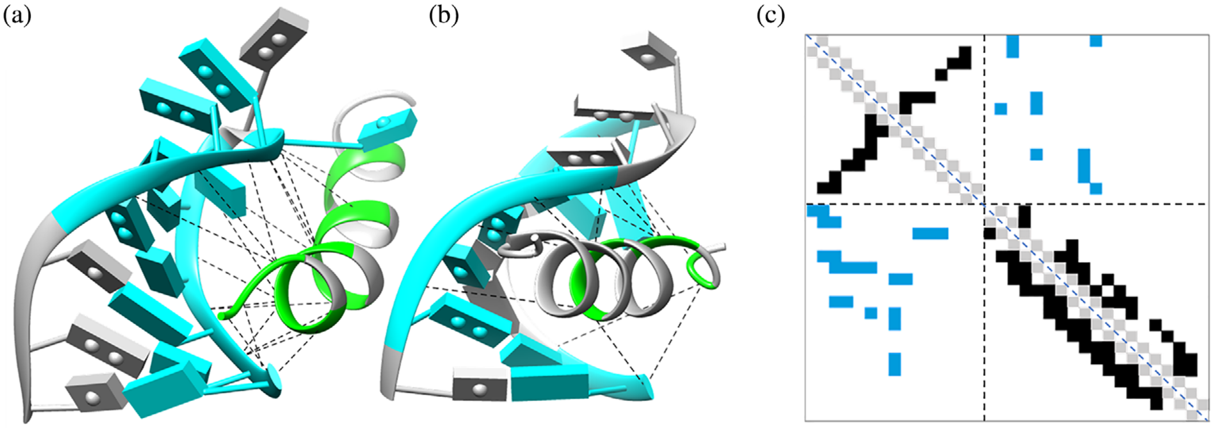


FIGURE 3 Example of the 1D2DSimScore input for a protein-RNA complex (phage P22 N peptide complexed with box B RNA, RCSB PDB code 1a4t). Amino-acid residues and ribonucleotide residues defined as interaction units, a positive class (residue defined as interacting) defined as at least two pairs of atoms found at a distance ≤3.5 Å between two different residues. (a) Native structure of the complex in a cartoon representation; (b) computationally modeled structure of the complex; (c) 2D maps of interactions for the experimentally solved structure (bottom left triangle) and a computational model (upper right triangle); excluded contacts corresponding to covalent interactions are shown in gray, intramolecular non-covalent interactions are shown in black, and protein-RNA interactions (the subject of analysis by 1D2DSimScore) are shown in blue

TABLE 3 Assessment of contact map similarity between the experimentally solved structure and the computational model of phage P22 N peptide complexed with box B RNA

	MCC	F_1	FM	J	Specificity	Precision	Recall
1D-01	0.201	0.571	0.591	0.400	0.75	0.769	0.454
2D-01	−0.030	0	0	0	0.978	0	0

2.4 | Example application of 1D2DSimScore—Comparison of an experimentally determined structure of an RNA molecule and three alternative computational models

Figure 4 illustrates the experimentally determined structure of a 5-hydroxytryptophan aptamer (RCSB PDB code 5kpy) and three computational models selected from a set of models obtained in the course of the blind prediction experiment RNA-Puzzles (Puzzle 9), for the purpose of this work dubbed X, Y, and Z. Table 4 presents the quantitative analysis of contact map similarities between the three models and the reference structure. The visual analysis of three models with respect to the global structure suggests that models X and Y are very similar to the reference structure, while model Z is very different; this is also reflected in the values of RMSD, 6.197, 9.587, and 20.162 Å,[24] respectively. Quantitative comparison of contacts with 1D2DSimScore reveals that all three models recapitulate quite well the canonical base-pairs (which are relatively easiest to predict), with model Y showing the best agreement with the reference structure. However, models Y and Z failed to recapitulate any non-canonical base-pairs, and only model X correctly identified some of them (with respect not only to the identity of interacting residues but also to the exact type of pairing). Still, model Y exhibits the highest similarity to the reference if all base-pairing interactions (canonical and non-canonical are considered). Importantly, model X is better than models Y and Z in predicted stacking interactions, even though all three models are far from ideal in this respect. Considering the similarity measure encompassing all types of interactions, model X is the closest to the reference, in agreement with the 3D similarity assessment. This analysis also reveals that the lack of only a few key contacts between the two hairpin loops in model Z, combined with an imperfect prediction of stacking, made it assume an incorrect global geometry.

2.5 | Example application of 1D2DSimScore—Comparison of macromolecule-small molecule interactions

Figure 5 and Table 5 compare an experimentally determined structure of a protein–ligand complex and a model obtained by docking software.[25] As expected from the visual analysis of the models, the results of the

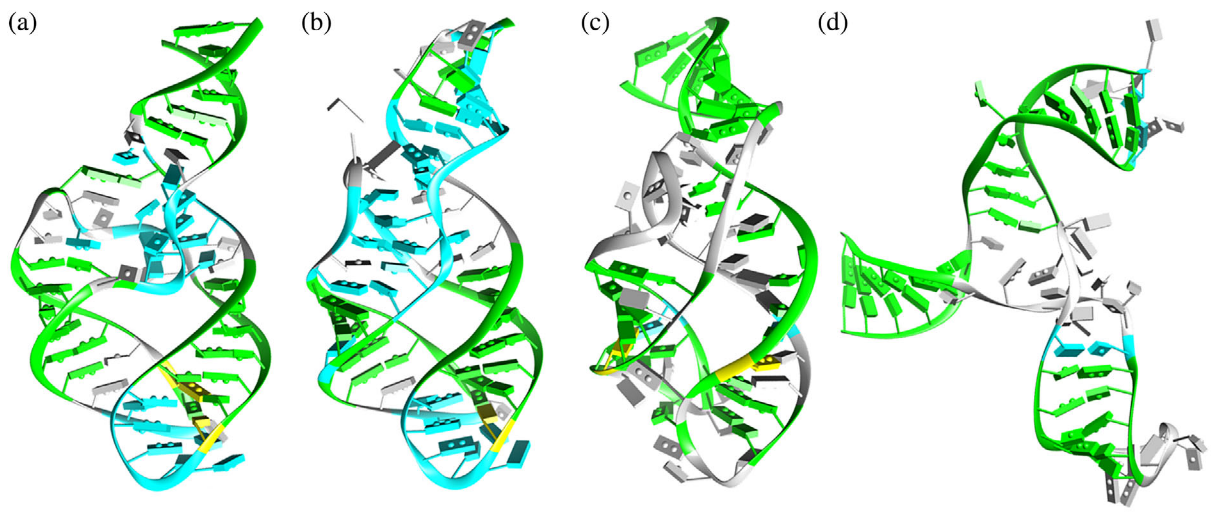


FIGURE 4 Example of the 1D2DSimScore input for the 1D-01 format with multiple classes. (a) Experimentally solved structure (RCSB PDB code, 5kpy); (b-d) three computational models (blind predictions) proposed by different groups in the context of the RNA-Puzzles experiment; canonical base pairs are colored in green, non-canonical base pairs are colored in cyan, residues involved in both canonical and non-canonical base pairs are colored in yellow, and residues that are not involved in any type of base pairs colored in gray

TABLE 4 1D2DSimScore comparison of contact maps between the three models obtained computationally and an experimentally determined crystal structure of an RNA riboswitch (RSCB PDB code 5kpy), considering various classes of contacts using the 2D-N format

	MCC	F_1	FM	J	Specificity	Precision	Recall
Model X versus reference							
Canonical + wobble	0.812	0.823	0.829	0.7	0.993	0.933	0.737
Non-canonical	0.405	0.444	0.449	0.286	0.942	0.387	0.522
Stacking	0.687	0.764	0.773	0.619	0.854	0.666	0.896
All pairs	0.638	0.695	0.696	0.533	0.93	0.655	0.740
All pairs + stacking	0.531	0.72	0.725	0.562	0.73	0.642	0.818
Model Y versus reference							
Canonical + wobble	0.862	0.873	0.878	0.776	0.968	0.791	0.974
Non-canonical	−0.028	0	0	0	0.987	0	0
Stacking	0.545	0.663	0.663	0.495	0.872	0.642	0.685
All pairs	0.664	0.71	0.714	0.55	0.966	0.791	0.644
All pairs+ stacking	0.433	0.649	0.649	0.480	0.788	0.656	0.642
Model Z versus reference							
Canonical + wobble	0.786	0.81	0.81	0.68	0.975	0.8	0.820
Non-canonical	−0.028	0	0	0	0.988	0	0
Stacking	0.541	0.655	0.658	0.487	0.909	0.722	0.6
All pairs	0.581	0.634	0.639	0.464	0.959	0.727	0.561
All pairs + stacking	0.426	0.633	0.637	0.464	0.836	0.711	0.571

comparison of the binding sites at the level of amino acid residues interacting with the ligand reveal high similarity. The difference between some of the scores in the 1D and 2D approaches results mostly from the difference in the number of negatives (TN and FN). The same type of calculations can be done for other complexes involving different types of molecules, for example, RNA–small molecule interactions.[26] The comparison can also be made for different levels of representation, for example, taking individual atoms into account.

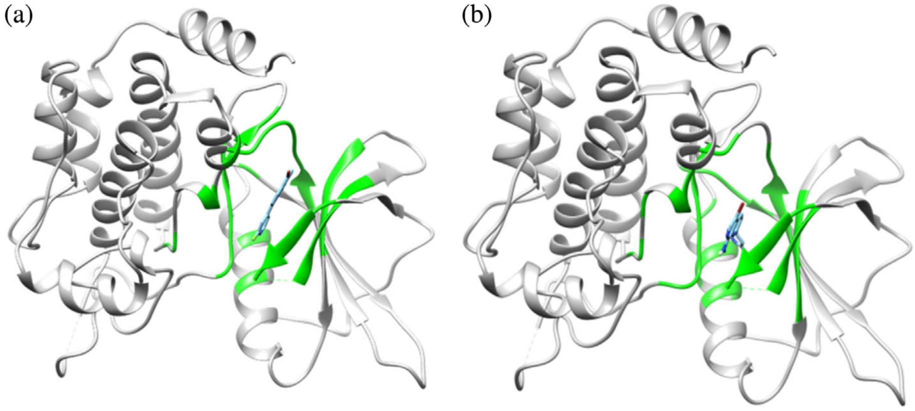


FIGURE 5 Example of interactions between a protein and a small molecule as an input to 1D2DSimScore. (a) Cartoon representation of the experimentally determined structure of the human protein TNIK in complex with a small-molecule ligand 5-bromanyl-2-(2-fluoranylpyridin-4-yl)-1,7-naphthyridin-8-amine (PDB code 6RA5) and (b) a complex of the same ligand docked to the same structure using the Smina docking software. Protein residues interacting with the ligand are colored in green

TABLE 5 Assessment of contact map similarity between the experimentally determined structure of a protein–ligand complex (PDB code 6RA5) and its counterpart where the ligand was redocked computationally

	MCC	F_1	FM	J	Specificity	Precision	Recall
1D-01	0.720	0.731	0.750	0.600	0.97	0.692	0.818
2D-01	0.744	0.744	0.744	0.592	0.999	0.727	0.762

2.6 | Example application of 1D2DSimScore—Comparison of RNA secondary structure patterns for a very long RNA

Table 6 shows the results of the comparison of secondary structure patterns obtained for the SARS-CoV-2 virus genomic RNA (nearly 30,000 residues) based on three different experimental approaches, namely chemical probing with DMS in vitro, and with the SHAPE method in vitro and in vivo.[27]

3 | DISCUSSION

Comparison of macromolecular structures requires consideration of very complex data, and it requires looking at different levels of the molecular organization of the interacting molecules. For the purpose of comparing macromolecular complexes, multiple models have been proposed. 1D2DSimScore is a new tool for the comparison of macromolecular structures at the level of contact information. The software package accepts different types of input, and can compare two different structures of the same macromolecule of any type, structures of two different macromolecules of the same type, or structures of two macromolecular complexes. 1D2DSimScore was able to calculate the similarity measures for two different coronavirus RNA genome secondary structures (containing about 30,000 nucleotides) in dot-bracket notation with each other in less than 1 s. As 1D2DSimScore is written in C++, it is a platform-independent tool that can be easily used on any machine with a C++ compiler.

1D2DSimScore provides a rich set of similarity measures for structure comparisons that are particularly useful for nucleic acids and nucleic acid-protein complexes. It allows separating canonical and non-canonical base pairs, which is especially helpful for de novo RNA 3D structure prediction where most approaches fail to correctly predict non-canonical base pairs, hence being able to compare them among multiple predicted models is of great use. Also, 1D2DSimScore can be used to compare two RNA chains of different lengths, provided their alignment is defined by the user. Furthermore, the user can provide the tool with defined start-end ranges for each RNA sequence to perform comparisons of a specified local alignment region of two RNA structures. This feature may also be available for other types of macromolecules if the user can provide the details of the different types of interactions for them. In the future, we intend to extend 1D2DSimScore to include a classification of contacts within proteins and between proteins and nucleic acids, similar to the classification currently used to describe base-pairing and stacking in nucleic acids used in the 2D-N format.

It is worth noting that some of the metrics should not be used alone, for example, recall/sensitivity without precision and/or sensitivity, as this may lead to misleading conclusions. Recall/sensitivity (also known as true positive rate) refers to the fraction of interactions occurring both in the query and in the reference structures among

TABLE 6 Assessment of similarity between the secondary structures predicted based on three independent experimental probing approaches for the genomic RNA of SARS-CoV2

	MCC	F_1	FM	J	Specificity	Precision	Recall
DMS in vitro versus SHAPE in vivo	0.509	0.770	0.770	0.627	0.722	0.757	0.781
DMS in vitro versus SHAPE in vitro	0.583	0.810	0.810	0.681	0.758	0.798	0.823
SHAPE in vivo versus SHAPE in vitro	0.655	0.835	0.835	0.716	0.772	0.794	0.879

TABLE 7 Runtime of 1D and 2D algorithms for pairwise comparisons of contact vectors or contact maps implemented in 1D2DsimScore, as a function of the sequence length. The result is independent of the type of molecule analyzed

Length [nt]	43	142	650	1,404	2,518	4,307	8,099	15,235	29,903
1D	1 ms	1 ms	1 ms	1 ms	2 ms	5 ms	7 ms	12 ms	26 ms
2D	<1 ms	1 ms	5 ms	10 ms	32 ms	85 ms	300 ms	1 s	3 s

all interactions in the reference. Precision refers to the fraction of interactions occurring both in the query and in the reference among all interactions in the query. Specificity refers to the fraction of non-interactions common to the query and the reference among all non-interactions in the reference. In general, metrics such as MCC and BA that consider all values in the confusion matrix are safer to be used alone. For the 2D algorithm, it is recommended to use the F1-score together with precision and recall (RPF) instead of MCC, since in the 2D comparison, the huge number of negatives, either TN or FN, always dominates the outcome of the MCC calculation.

In Table 7, we examine the running time of the 1D2DSimScore in different situations. Although several features, such as the sequence length and the number of lines describing the contacts, affect the performance of 1D2DSimScore, the runtime is not noticeable when the user chooses the correct package for comparing two structures.

1D2DSimScore allows the handling of multilabel interactions for nucleic acids, that is, a nucleotide residue can present interactions of different types with different partners. Currently, there are several tools that can generate this type of input for nucleic acids, but this has yet to be implemented for other molecules. However, this is an obvious direction for future development.

We use 1D2DSimScore extensively in our in-house applications, and we believe that this software package will be very helpful for other researchers studying the function of molecular interactions within and between macromolecules that involve significant conformational changes, as well as for the development and benchmarking of new methods for predicting macromolecular structures and interactions.

4 | MATERIALS AND METHODS

4.1 | Overview of calculations performed by 1D2DSimScore

1D2DSimScore is a software package that takes as input files with information about interaction maps, that is, intra- and/or intermolecular contacts in two models of a macromolecule or two comparable macromolecules (protein, nucleic acid, and macromolecular complex) or in a set of models, performs a series of pairwise comparisons and generates output files reporting different similarity/dissimilarity measures. Contact information can be defined by the user in several different ways, and the program offers several ways of treating that information (see the sections below for detailed explanations and examples). In the course of each pairwise comparison, 1D2DSimScore considers one structure as a query and another as a reference and analyzes the contact information for all pairs of corresponding interacting units (e.g., residues) between the query and the reference. The program generates a confusion matrix by classifying the pairs of corresponding contact units as true positives (TPs, when the contact status is positive and of the same type for the corresponding contact units in both structures), true negatives (TNs, when the contact status is negative for the corresponding contact units in both structures), false positives (FPs, when the contact status is positive for the contact unit in the query, but the corresponding contact unit in the reference has different contact status, for example, negative or positive of a different type), or false negatives (FNs, when the contact status in the query is negative, but the corresponding contact unit in the reference has a positive contact status).

Based on the confusion matrix resulting from the comparison of corresponding contact units in the query and reference interaction maps, 1D2DSimScore

TABLE 8 Measures of similarity between contact maps calculated by 1D2DSimScore

Similarity measure	Formula
Recall/sensitivity	$TP/(TP+FN)$
Specificity	$TN/(TN+FP)$
Precision	$TP/(TP+FP)$
False omission rate	$FN/(FN+TN)$
Prevalence threshold	$\sqrt{}FPR/(\sqrt{}FPR+\sqrt{}Recall);\ where\, FPR=FP/(FP+TN)$
Balanced accuracy	$(Specificity+Recall)/2$
F1 index	$2\times(Precision\times Recall)/(Precision+Recall)$
Matthew's correlation coefficient	$(TP\times TN-FP\times FN)/\sqrt{}((TP+FP)(TP+FN)(TN+FP)(TN+FN))$
Fowlkes–Mallows index	$\sqrt{}(precision\times Recall)$
Markedness	$Precision-FOR$
Jaccard index	$TP/(TP+FP+FN)$

calculates various measures of pairwise similarity/dissimilarity, including recall, specificity, precision, false omission rate (FOR), prevalence threshold (PT), critical success index (CSI), balanced accuracy (BA), F1-score, Matthew's Correlation Coefficient (MCC), Fowlkes–Mallows index (FM index), markedness (MK), and Jaccard Index (J index) (Table 8). Thereby, 1D2DSimScore allows the user to evaluate various aspects of similarity/dissimilarity between models of any bio-macromolecules at the level of intra- and inter-molecular contacts.

Contact information can be encoded in different ways, including a one-dimensional array (1D vector) or a two-dimensional array (2D matrix). The 1D representation is used to indicate interactions at the level of user-defined interaction units (e.g., domains, amino acid residues, functional groups, and individual atoms), while the 2D representation is used to describe pairs of interaction units. Information about the contact status can be binary (i.e., to indicate whether a given interaction unit is or is not involved in contact), or it can specify different contact types. The 1D representation is suitable for comparing interactions where the user-defined interaction unit can be only involved in one contact at a time. In contrast, the 2D representation is more appropriate for situations where the given interaction unit can be involved in more than one interaction (of the same type or of different types).

4.2 | Type of information used by 1D2DSimScore and format of 1D2DSimScore input files

1D2DSimScore can accept several types of input files that specify different levels of information about contacts. The main categories of inputs include:

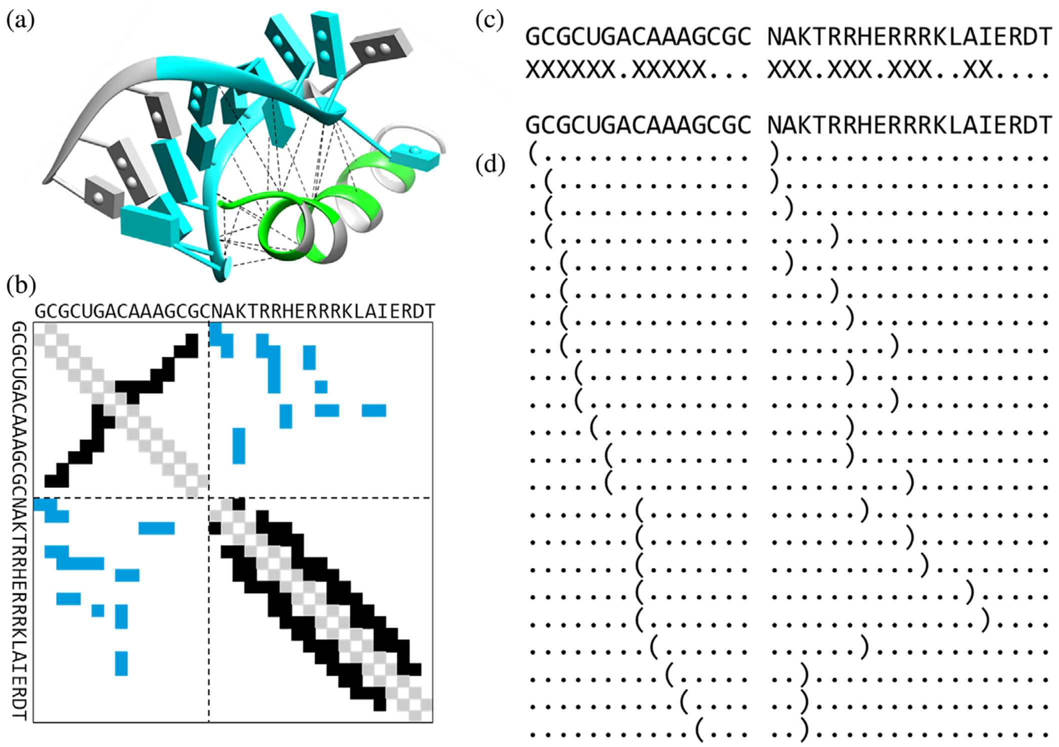
- 1D format with two classes, negative 0 and positive 1 (1D-01)
- 2D format with two classes, negative 0 and positive 1 (2D-01)
- 2D format with multiple different positive classes (2D-N)

4.2.1 | 1D format with two classes, negative 0 and positive 1 (1D-01)

1D format with a binary classification (one negative class, one positive class) describes whether a given interaction unit is or is not in contact with another unit. Here, the interaction unit is typically an amino-acid residue, a nucleotide residue, or another interacting molecule. Still, the user can define any interaction unit and prepare the input file to describe interactions, for example, at the level of individual atoms or functional groups. For the 1D-01 format, 1D2DSimScore accepts a string of 0 and 1 digits, or "." (negative) and "X" (positive) characters (Figure 6).

4.2.2 | 2D format with two classes, negative 0 and positive 1 (2D-01)

2D format with a binary classification (one negative class, one positive class) describes pairs of interacting units and declares whether interaction does or does not occur between two given units. Here, the interaction unit is typically an amino-acid residue, a nucleotide residue, or another interacting molecule, but the user can define any other interaction unit and prepare the input file to describe interactions, for example, at the level of


(a)
(b)
GCGCUGACAAAGCGCNAKTRRHERRRKLAIERDT
GCGCUGACAAAGCGCNAKTRRHERRRKLAIERDT
(c)
(d)

```
GCGCUGACAAAGCGC NAKTRRHERRRKLAIERDT
XXXXXX.XXXXX... XXX.XXX.XXX..XX....

GCGCUGACAAAGCGC NAKTRRHERRRKLAIERDT
(.............. )..................
.(............. )..................
.(............. .).................
.(............. ....)..............
..(............ .).................
..(............ ....)..............
..(............ .....).............
..(............ ........)..........
...(........... .....).............
...(........... ........)..........
....(.......... .....).............
.....(......... .....).............
.....(......... .........).........
.......(....... ......)............
.......(....... .........).........
.......(....... ..........)........
.......(....... .............).....
.......(....... ..............)....
........(...... ......)............
.........(..... ..)................
..........(.... ..)................
...........(... ..)................
```

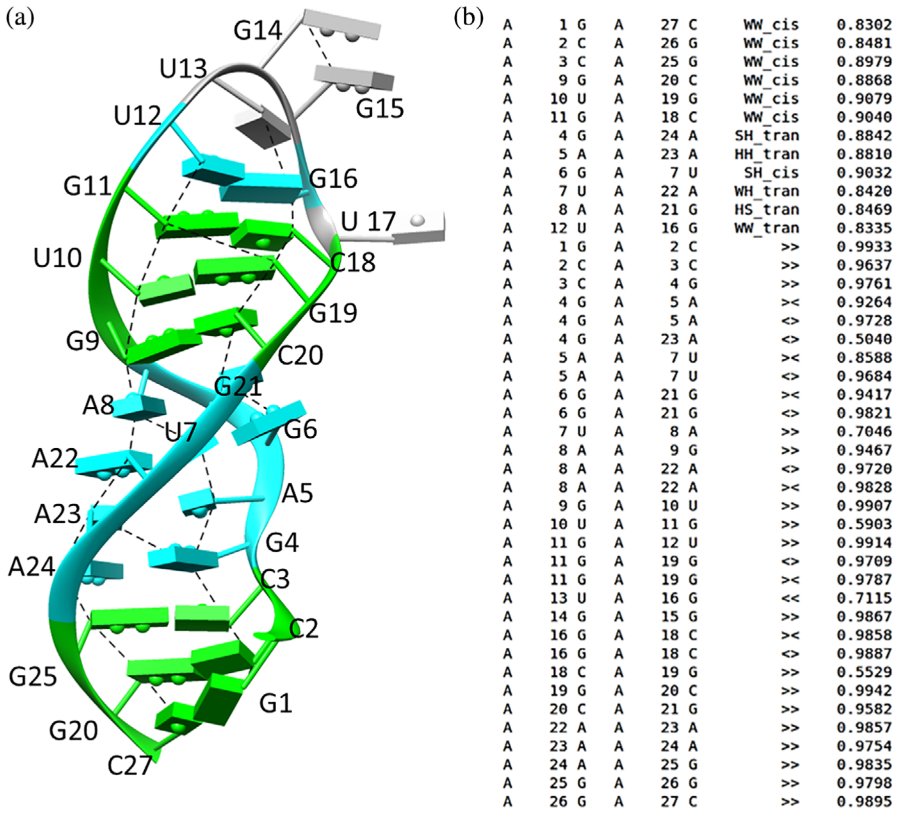
FIGURE 6 Example of 1D2DSimScore input for a protein-RNA complex (phage P22 N peptide complexed with box B RNA, RCSB PDB code 1a4t). Amino-acid residues and ribonucleotide residues are defined as interaction units, a positive class (interacting residues) is defined as at least two pairs of atoms found at a distance ≤3.5 Å between the given amino-acid and nucleotide residues. (a) Structure of the complex in a cartoon representation, interacting residues in RNA and protein colored in cyan and green, respectively, pairwise interactions indicated by dashed lines; (b) 2D map of interactions, with intramolecular interactions shown in gray (covalent) or in black (non-covalent), and RNA -protein interactions shown in turquoise; (c) sequence and interaction information according to the 1D-01 format (interacting residues indicated by "X," non-interacting residues indicated by "."); (d) sequence and interaction information according to the 2D-01 format; interacting residue pairs indicated by pairs of opening and closing brackets (with each contact shown in a separate line for the clarity of presentation), dots used as separators

individual atoms or functional groups. For the 2D-01 format, 1D2DSimScore accepts the dot-bracket (DBN) notation (2D_01 package), commonly used to represent base-pairing interactions in nucleic acids. We have extended this type of notation to all types of interactions, all types of biomolecules and their interfaces (as shown for protein–RNA interactions in Figure 6) or a square matrix of digits, 0–9 (2D_01_CMO package). For comparing contacts representing base pairs in nucleic acids, 1D2DSimScore can accept a nucleic acid sequence as additional information. In such a case, the user may perform calculations separately for different types of interactions (canonical Watson-Crick cis pairs G-C, A-T/A-U, wobble pairs G-U, non-canonical pairs, and all base pairs regardless of the category).

4.2.3 | 2D format with multiple different positive classes (2D-N)

2D format with multiple positive classes of interactions is appropriate for describing complex interactions. It defines not only whether a given interaction unit is or is not in contact with another unit but also enables the classification of contacts depending on the kind of interaction and the interacting partner. Here, the interaction unit is typically a functional group that can make different kinds of interactions. Information stored in this format can be reduced to comparing contacts in 1D if only one of these interactions is considered per interaction unit, which is a common situation in macromolecular structure annotation. In situations where a given interaction unit is allowed to make more than one interaction, the comparison must be carried out in 2D.

The 2D-N format helps analyze contacts in nucleic acid molecules and between nucleic acid molecules and proteins and allows for defining interactions specific to different functional groups. Figure 7 provides an example annotation of an RNA molecule with base-pairing and stacking interactions. In this format, each residue can be annotated with multiple interactions involving the two faces (capable of stacking) and three edges of the base (Watson–Crick, Hoogsteen, and Sugar) capable of forming base pairs. In 2D-N format, only positive



```
A     1 G   A    27 C     WW_cis   0.8302
A     2 C   A    26 G     WW_cis   0.8481
A     3 C   A    25 G     WW_cis   0.8979
A     9 G   A    20 C     WW_cis   0.8868
A    10 U   A    19 G     WW_cis   0.9079
A    11 G   A    18 C     WW_cis   0.9040
A     4 G   A    24 A    SH_tran   0.8842
A     5 A   A    23 A    HH_tran   0.8810
A     6 G   A     7 U     SH_cis   0.9032
A     7 U   A    22 A    WH_tran   0.8420
A     8 A   A    21 G    HS_tran   0.8469
A    12 U   A    16 G    WW_tran   0.8335
A     1 G   A     2 C         >>   0.9933
A     2 C   A     3 C         >>   0.9637
A     3 C   A     4 G         >>   0.9761
A     4 G   A     5 A         ><   0.9264
A     4 G   A     5 A         <>   0.9728
A     4 G   A    23 A         <>   0.5040
A     5 A   A     7 U         ><   0.8588
A     5 A   A     7 U         <>   0.9684
A     6 G   A    21 G         ><   0.9417
A     6 G   A    21 G         <>   0.9821
A     7 U   A     8 A         >>   0.7046
A     8 A   A     9 G         >>   0.9467
A     8 A   A    22 A         <>   0.9720
A     8 A   A    22 A         ><   0.9828
A     9 G   A    10 U         >>   0.9907
A    10 U   A    11 G         >>   0.5903
A    11 G   A    12 U         >>   0.9914
A    11 G   A    19 G         <>   0.9709
A    11 G   A    19 G         ><   0.9787
A    13 U   A    16 G         <<   0.7115
A    14 G   A    15 G         >>   0.9867
A    16 G   A    18 C         ><   0.9858
A    16 G   A    18 C         <>   0.9887
A    18 C   A    19 G         >>   0.5529
A    19 G   A    20 C         >>   0.9942
A    20 C   A    21 G         >>   0.9582
A    22 A   A    23 A         >>   0.9857
A    23 A   A    24 A         >>   0.9754
A    24 A   A    25 G         >>   0.9835
A    25 G   A    26 G         >>   0.9798
A    26 G   A    27 C         >>   0.9895
```

FIGURE 7 Example of the 1D2DSimScore input for an RNA tertiary structure (experimentally determined model of domain IIID of hepatitis C virus internal ribosome entry site, RCSB PDB code 1fqz). (a) RNA structure in the cartoon representation, canonical base pairs colored in green, non-canonical base pairs colored in cyan, nucleotides that are not involved in any type of base pairs colored in gray, stacking interactions indicated by dashed lines. (b) Sequence and interaction information according to the 2D-N format, as generated by the ClaRNA method[22]

information is present since the space of all possible contacts is vast, and negative contact status is assigned automatically in the absence of the positive contact status to each interacting unit. In the case presented in Figure 7, residue U17 is considered non-interacting, according to the dictionary of contacts used to annotate the given structure.

4.3 | Options for defining the space of possible contacts considered by 1D2DSimScore

1D2DSimScore allows the user to select one of two different ways of handling the contact space, either as a one-dimensional array or vector (1D) map of interactions or a two-dimensional array or matrix (2D) map of interactions. One of the main differences between the 1D and 2D algorithms is the calculation of the categories of the confusion matrix.

The main difference between the 1D and 2D approaches is in focus on either the individual interaction units or their pairs, and hence in the number of instances analyzed. In the 1D approach (default for the 1D-01 format), the number of instances equals the number of all interaction units (n), while in the 2D approach (default for the 2D-01 format), the number of instances equals the number of theoretically possible pairs, that is, $n*(n-1)/2$. Hence, for a particular molecular system described by specific interaction units, the fraction of instances classified as negative is much larger in the 2D space than in the 1D space, that is, the fraction of pairs of interaction units that do not interact with each other is much larger than the fraction of interaction units that are non-interacting. The difference in the fraction of negative classes has a particularly large impact on counting true negatives in pairwise comparisons.

Data provided in the 2D-N format can be analyzed in either 1D or 2D mode. Figure 7 presents an example of an RNA molecule, in which each residue is annotated with multiple interactions involving the two faces (capable of stacking) and three edges of the base (Watson–Crick, Hoogsteen, and Sugar) capable of forming base pairs. For comparing an interaction map of this molecule with another interaction map (e.g., for a different model of the same molecule or another molecule with the same number of structural building blocks capable of interacting), the user must choose the interaction unit, which then implies the dimensionality of the contact space. If the user chooses each ribonucleotide to be considered as an interaction unit, there are multiple interactions per interaction unit (i.e., most residues in this structure are involved in two stacking interactions and one base pairing interaction at the same time). Since only one classification (TP, FP, TN, or FN) is allowed per comparison, in case of multiple classes of contacts are assigned to one interaction unit (multiple contacts with other interaction units), the comparison must be carried out at the level of pairs rather than individual units, and the calculation of interaction similarities must be carried out in 2D. However, if each face and each edge are considered as separate interaction units (five interaction units per residue), and if each such unit is allowed to make only one contact, then the contact information from the 2D-N format can be reduced to a 1D vector of interaction units, and the similarity of two interaction maps can be assessed in 1D, at the level of individual interaction units rather than their pairs.

It is important to emphasize that the complex 2D-N data format contains all the information included in the 1D-01 and 2D-01 formats. It can also be used for comparisons that ignore the type of contacts, and in such a case, all types of positive contact status would be reduced to one positive class. In other words, with the 2D-N approach and multiple positive classes, a pair of corresponding contact units (or contact unit pairs) exhibiting different contacts would be classified as false positives, while after the reduction of multiple positive classes to one, the two different contacts would be classified as true positives (as in the 2D-01 approach). Further, extraction of specific subtypes of contacts that can occur only once per interaction unit (e.g., canonical base pairs in nucleic acids, which can occur only once per residue) allows the user to carry out calculations with the same algorithm as used in the case of 1D-01 format.

Using the 2D-N data format for analyzing nucleic acid structures while considering base edges and faces, users can define different numbers of interaction units for determining the space of possible contacts (which will mainly affect the number of non-contacts considered in calculation of true negatives). 1D2DSimScore can use one, three, or five interaction units if similarity measures are derived for canonical and wobble interactions; three or five for non-canonical interactions and all types of base pairs; two interaction units for stacking; and for comparing all types of base-pairing and stacking interactions in nucleic acid structures, the only option is all five edges and faces.

4.4 | Comparisons of data sets with multiple contact sets

In the case of data sets comprising multiple contact maps, the user can calculate the pairwise similarity values for all structures with each other for all types of inputs. For this purpose, the 1D2DSimScore package has different modules for each type of input. The user can specify a directory with all information. Files with the string “01” or “.X” (dot and X) for 1D-01, DBN input files for 2D-01 (in the case of RNA secondary structures of nucleic acids, the sequence of the nucleic acids can also be specified), and .pdb files for 2D-N.

4.5 | Tools for the preparation of 1D2DSimScore input files

The 1D2DSimScore package includes programs for extracting contacting residues and contact pairs from 3D coordinates for all types of macromolecules (including proteins, nucleic acids, and their complexes) and for extracting different classes of contact pairs from nucleic acid structures and nucleic acid-protein complexes. Besides, the relevant information can be obtained with numerous third-party tools for analyzing known macromolecular structures and for predicting contacts from sequences. While it would be impossible to cover all the existing formats, 1D2DSimScore provides a collection of tools for reformatting outputs of several of the existing tools, including predictors of contacts from alignments such as direct coupling analysis methods Gremlin,[28] RNAcmap,[29] and ccmpred.[30]

4.6 | C++ code

The program is C++20 code that can be compiled under the g++ - 11 compiler. 1D2DSimScore is an object-oriented software developed with Standard Template Library (STL). The source code can be found at https://github.com/Naeim-Moafi/1D2DSimScore. The program is available according to the Apache license.

4.7 | Examples of workflows

1D2DSimScore can be used in several scenarios, including comparisons of different conformations of the same molecule, two very similar molecules, or common sections of different molecules. The simplest type of analysis involves two structures (reference and query) either in the PDB format or as contact maps (1D_01, 2D_01, or 2D_01_CMO) and may be applicable to comparing interactions in different conformations obtained under different conditions, different theoretical models of 3D structure, or comparison of a predicted structure with an experimentally determined reference (2D_N). Another application involves a series of pairwise comparisons (many vs. one), for example, in testing computationally predicted structures or contact maps considered as multiple queries with one reference contact map obtained from the experimentally determined structure. Last but not least, for a data set of alternative structures (e.g., a series of conformers obtained from a molecular simulation), a series of all versus all pairwise comparisons can be carried out to generate a matrix of similarities/dissimilarities (1D_01_Dataset, 2D_01_Dataset, and 2D_N_Dataset), which can be further analyzed, for example, to define trajectories or perform clustering analyses.

AUTHOR CONTRIBUTIONS

S. Naeim Moafinejad: Investigation (lead); methodology (equal); software (lead); validation (equal); visualization (equal); writing – original draft (lead); writing – review and editing (equal). **Iswarya P. N. Pandaranadar Jeyeram:** Conceptualization (supporting); methodology (equal); software (supporting). **Farhang Jaryani:** Methodology (supporting); software (equal); writing – original draft (supporting). **Niloofar Shirvanizadeh:** Software (supporting). **Eugene F. Baulin:** Conceptualization (supporting); investigation (equal); methodology (equal); validation (lead); writing – review and editing (supporting). **Janusz M. Bujnicki:** Conceptualization (lead); formal analysis (supporting); funding acquisition (lead); investigation (equal); methodology (equal); project administration (equal); resources (lead); supervision (lead); validation (supporting); writing – original draft (supporting); writing – review and editing (equal).

ACKNOWLEDGMENTS

The authors would like to thank Dr. Filip Stefaniak and Dr. Sunandan Mukherjee (IIMCB) for their help. This study was supported by the Polish National Science Center (NCN, 2017/25/B/NZ2/01294 to J.M.B. 2017/26/A/NZ1/01083 to J.M.B), EMBO fellowship ALTF 525–2022 to E.F.B.; IIMCB statutory funds to J.M.B.; Funding for open access charge: Polish National Science Center, grant 2017/26/A/NZ1/01083 to J.M.B.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study

ORCID

S. Naeim Moafinejad https://orcid.org/0000-0003-0397-2596
Eugene F. Baulin https://orcid.org/0000-0003-4694-9783
Janusz M. Bujnicki https://orcid.org/0000-0002-6633-165X

REFERENCES

1. Olechnovič K, Monastyrskyy B, Kryshtafovych A, Venclovas Č. Comparative analysis of methods for evaluation of protein models against native structures. Bioinformatics. 2019;35(6):937–944.
2. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-round XIV. Proteins. 2021;89(12):1607–1617.
3. Cruz JA, Blanchet M-F, Boniecki M, et al. RNA-puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. RNA. 2012;18(4):610–625.
4. Parisien M, Cruz JA, Westhof E, Major F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. RNA. 2009;15(10):1875–1885.
5. Kufareva I, Ilatovskiy AV, Abagyan R. Pocketome: An encyclopedia of small-molecule binding sites in 4D. Nucleic Acids Res. 2012;40 (Database issue):D535–D540.
6. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Cryst A. 1976;32(5):922–923.
7. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004; 57(4):702–710.
8. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins Suppl. 1999;3:22–29.
9. Olechnovič K, Kulberkytė E, Venclovas C. CAD-score: A new contact area difference-based function for evaluation of protein structural models. Proteins. 2013;81(1):149–162.
10. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013; 29(21):2722–2728.
11. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins. 2014;82(Suppl 2):S7–S13.
12. Huang YJ, Rosato A, Singh G, Montelione GT. RPF: A quality assessment tool for protein NMR structures. Nucleic Acids Res. 2012;40 (Web Server issue):W542–W546.
13. Cong Q, Kinch LN, Pei J, et al. An automatic method for CASP9 free modeling structure prediction assessment. Bioinformatics. 2011;27(24):3371–3378.
14. Ye Y, Godzik A. FATCAT: A web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res. 2004;32 (Web Server issue):W582–W585.
15. Piątkowski P, Jabłońska J, Żyła A, et al. SupeRNAlign: A new tool for flexible superposition of homologous RNA structures and inference of accurate structure-based sequence alignments. Nucleic Acids Res. 2017;45(16):e150.
16. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873): 583–589.
17. Katuwawala A, Kurgan L. Comparative assessment of intrinsic disorder predictions with a focus on protein and nucleic acid-binding proteins. Biomolecules. 2020;10(12). https://doi.org/10.3390/biom10121636.
18. Wang H, Zhao Y. Methods and applications of RNA contact prediction. Chin Phys B. 2020;29(10):108708.
19. Pietal MJ, Tuszynska I, Bujnicki JM. PROTMAP2D: Visualization, comparison and analysis of 2D maps of protein structure. Bioinformatics. 2007;23(11):1429–1430.

20. Pietal MJ, Szostak N, Rother KM, Bujnicki JM. RNAmap2D - calculation, visualization and analysis of contact and distance maps for RNA and protein-RNA complex structures. BMC Bioinform. 2012;13(1):333.
21. Lu X-J, Bussemaker HJ, Olson WK. DSSR: An integrated software tool for dissecting the spatial structure of RNA. Nucleic Acids Res. 2015;43(21):e142.
22. Waleń T, Chojnowski G, Gierski P, Bujnicki JM. ClaRNA: A classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. Nucleic Acids Res. 2014;42(19):e151.
23. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci U S A. 2009;106(50):21149–21154.
24. Miao Z, Adamiak RW, Antczak M, et al. RNA-puzzles round IV: 3D structure predictions of four ribozymes and two aptamers. RNA. 2020;26(8):982–995.
25. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with Smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model. 2013;53(8):1893–1904.
26. Szulc NA, Mackiewicz Z, Bujnicki JM, Stefaniak F. fingeRNAt—A novel tool for high-throughput analysis of nucleic acid-ligand interactions. PLoS Comput Biol. 2022;18(6):e1009783.
27. Manfredonia I, Nithin C, Ponce-Salvatierra A, et al. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. Nucleic Acids Res. 2020;48(22):12436–12452.
28. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013;110(39):15674–15679.
29. Zhang T, Singh J, Litfin T, Zhan J, Paliwal K, Zhou Y. RNAcmap: A fully automatic pipeline for predicting contact maps of RNAs by evolutionary coupling analysis. Bioinformatics. 2021;37(20):3494–3500.
30. Seemayer S, Gruber M, Söding J. CCMpred—Fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics. 2014;30(21):3128–3130.

How to cite this article: Moafinejad SN, Pandaranadar Jeyeram IPN, Jaryani F, Shirvanizadeh N, Baulin EF, Bujnicki JM. 1D2DSimScore: A novel method for comparing contacts in biomacromolecules and their complexes. Protein Science. 2023;32(1):e4503. https://doi.org/10.1002/pro.4503