Published online 7 February 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 10* ***4007–4022***
doi:10.1093/nar/gkq1320

ModeRNA: a tool for comparative modeling of RNA 3D structure

Magdalena Rother[1,2], Kristian Rother[1,2], Tomasz Puton[1,2] and Janusz M. Bujnicki[1,2,*]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw and [2]Laboratory of Structural Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

Received December 11, 2009; Revised November 24, 2010; Accepted December 11, 2010

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

ABSTRACT

RNA is a large group of functionally important biomacromolecules. In striking analogy to proteins, the function of RNA depends on its structure and dynamics, which in turn is encoded in the linear sequence. However, while there are numerous methods for computational prediction of protein three-dimensional (3D) structure from sequence, with comparative modeling being the most reliable approach, there are very few such methods for RNA. Here, we present ModeRNA, a software tool for comparative modeling of RNA 3D structures. As an input, ModeRNA requires a 3D structure of a template RNA molecule, and a sequence alignment between the target to be modeled and the template. It must be emphasized that a good alignment is required for successful modeling, and for large and complex RNA molecules the development of a good alignment usually requires manual adjustments of the input data based on previous expertise of the respective RNA family. ModeRNA can model post-transcriptional modifications, a functionally important feature analogous to post-translational modifications in proteins. ModeRNA can also model DNA structures or use them as templates. It is equipped with many functions for merging fragments of different nucleic acid structures into a single model and analyzing their geometry. Windows and UNIX implementations of ModeRNA with comprehensive documentation and a tutorial are freely available.

INTRODUCTION

Ribonucleic acid (RNA) is a group of macromolecules involved in essential processes such as communication of biological information between DNA and proteins, regulation of cellular processes and catalysis of biochemical reactions (1). The function of RNA depends on its structure and dynamics, which in turn is encoded in the linear sequence. Atomic coordinates of experimentally determined RNA three-dimensional (3D) structures are collected in the Protein Data Bank (PDB) (2) and the Nucleic Acid Data Base (NDB) (3) (with a total number of 1848 RNA structures as of 23 September 2010). Nonetheless, there is a large and growing gap between the number of known 3D structures and known RNA sequences. For instance, for the structurally best characterized family of RNAs, i.e. tRNAs, there are 1 101 833 sequences in the Rfam database (4) and only 170 structures. This situation is analogous to that observed for protein sequences and structures, where the number of available sequences greatly exceeds the number of experimentally determined structures [reviews: (5,6)].

The lack of experimentally determined structures for the majority of biological macromolecules has prompted the development of computational methods for 3D structure prediction. The development of structural bioinformatics tools has been driven mostly by the growth of structural databases, which thus far has been much faster for proteins than for RNAs.

There are two main approaches for 3D structure prediction of biomacromolecules. The first approach is knowledge-based; it relies on the database of experimentally solved structures and empirically observed dependencies between sequence and structural similarities. One of the types of knowledge-based modeling is comparative (or homology-) modeling, based on the empirical observation that evolutionarily related macromolecules usually retain similar 3D structure despite the divergence on the sequence level [review: (7)]. The two main limiting criteria that need to be observed to use this approach are that the modeling of the 'target' structure starts with another known structure of a homologous molecule to be used as a 'template', and that each element of the target sequence is

*To whom correspondence should be addressed. Tel: +48 22 597 0750; Fax: +48 22 597 0715; Email: iamb@genesilico.pl

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

aligned to the homologous element in the template sequence/structure. In particular, homologous residues should be aligned with each other. High sequence similarity is not a prerequisite. In fact, it is possible to create good homology models even if the sequence identity between the target and the template is zero (8). However, obviously the higher the sequence similarity, the easier it is to generate a correct alignment (to find homologous residues between the target and the template). On average, molecules with higher sequence similarity tend to exhibit more similar structures (9). Therefore, using templates with higher sequence similarity is recommended. Apart from sequence divergence, structures may also change because of environmental factors, e.g. the binding of other molecules or the composition of the solution (salt, pH) (10). It is generally the responsibility of the user of the homology modeling software to choose a template, whose biological state corresponds best to the desired biological state of the target to be modeled. With an incorrectly chosen template and/or wrong alignment, the model will be always very far from the native structure. These limitations concern all homology modeling tools, as templates and alignments are always necessary in this approach (11).

Comparative analyses of homologous structures of non-coding RNAs, including tRNAs, rRNAs and riboswitches [e.g. (12)], revealed patterns of conservation that are analogous to those observed in proteins: the secondary and tertiary structure is usually more conserved than sequence, and core regions important for stability and function tend to be more conserved at all levels. In general, it can be stated that in families of homologous RNAs the 3D fold is conserved and alignment of sequences and secondary-structure patterns can be used to recognize such structural conservation, in analogy to 'fold recognition' often used in protein structure prediction. On the other hand, not all homologous RNAs preserve the same structure, e.g. some degree of topological variability outside the conserved core has been identified in the RNA subunits of RNase P from *Escherichia coli* (type A) and *Bacillus subtilis* (type B) (13). Nonetheless, similar variability of folds has been also observed in proteins (10).

It must be emphasized that RNA has some unique structural features that hamper a direct application of protein-like modeling methodology. The secondary structure of RNA is defined by canonical (Watson–Crick) base pairs, while the tertiary (3D) structure is formed mostly by non-canonical base pairs that form 3D motifs (14). Stacking interactions determine helical structures, and are responsible for coaxial stacking of helices not connected linearly. Finally, the folding and dynamics of many structured RNA molecules depends on the presence of ions. Therefore, comparative modeling of RNA 3D structures shares some challenges with protein modeling, but also presents its unique ones.

One variant of knowledge-based modeling that has been particularly useful in RNA structural bioinformatics is manual modeling by human experts. It usually involves interactive (user-guided) manipulation of macromolecular structures based on assembly of fragments derived from various experimentally determined structures that are predicted to be similar to different parts of the target. The computational tools facilitate the choice, the manipulation, and the visualization of fragments, and often provide specialized algorithms for local optimization of geometry to seal breaks in the chain or relieve steric clashes. Graphics-based methods developed for this purpose include S2S/Assemble (15,16), ERNA-3D (17) and RNA2D3D (18). This type of knowledge-based modeling has been used in the past to generate models of a number of RNAs, including 5S rRNA (19), U1 snRNA (20), group I catalytic intron (21), parts of the signal recognition particle (17), *Thermus thermophilus* 23S rRNA (22), the class I ligase ribozyme (23) and tmRNA (24). Actually, the field of computational modeling of RNA structure is heavily dominated by manual modeling by experts, a situation resembling the early days of protein modeling, before the advent of the acclaimed CASP experiment (5). Nonetheless, similar methods have been also applied to model protein structures (25) [review: (26)].

The second approach for structure prediction is *ab initio* modeling based on biophysical rules, which simulates the folding process in search of the conformation with the minimal free energy [review: (27)]. The main problems with this method are the extreme ruggedness of the energy landscape (multiple local minima), the complexity of the function with which to calculate the free energy of the system (high cost of evaluating each conformation), imperfections of the energy function (the calculated energy does not necessarily correspond to the real energy) and the need for huge computational power to sample the conformational space (due to a large number of degrees of freedom it is not feasible to check all possible conformations). For these reasons, the use of *ab initio* methods is rather limited to smaller molecules, and even then the user cannot be sure whether a native-like conformation has been generated during the folding simulation, and whether it was scored better than those less native-like ones. To increase the efficiency of computations, full atom models are sometimes replaced by coarse-grained models, which treat groups of atoms as single interaction centers, so that smaller number of interactions must be evaluated [review: (28)].

Combination of the knowledge- and physics-based approaches resulted in the development of the so-called *de novo* folding methods, which assembly the target structure from small fragments derived from other (not necessarily homologous) known structures [review: (26)]. Here, the assembly relies on searching the conformational space as in the *ab initio* approach, and while the number of the degrees of freedom is limited by the use of a restricted set of conformers, this method shares most problems with the *ab initio* approach, including high computational cost and uncertainty as to identification of the most native-like structure in a large number of alternative models. FARNA was the first program that successfully used this approach to fold RNA molecules of up to 50-nt length (29). MC-Fold/MC-Sym is another method that assembles RNA structures from a library of fragments (30).

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

For proteins, many methods for automated 3D structure prediction have been developed and made freely available to the scientific community. They can produce reasonably accurate and practically useful models, in particular, based on the comparative modeling approach (31). There are numerous freely available tools (standalone programs and internet servers) for modeling of protein structure, e.g. comparative modeling programs Modeller (32) and Swiss-Model (33). In comparison to the field of protein structural bioinformatics, there is a paucity of comparable tools for RNA modeling. This situation prompted us to develop ModeRNA, a scriptable tool for prediction of RNA 3D structures by comparative modeling. It allows both for simplistic 'protein-like' construction of models from a set of templates and alignments, and for user-controlled manipulations of structures including fragment assembly. Special emphasis was placed on features unique to RNA, such as detecting and constructing base pairs, adding RNA-specific secondary-structure elements and modeling post-transcriptional modifications. The concept underlying ModeRNA has been inspired by the protein modeling tool SwissModel (33).

MATERIALS AND METHODS

Overview

As a minimal input ModeRNA requires the 3D coordinates of a template structure and a pairwise sequence alignment between the sequences of the template and the target RNA to be modeled. It must be emphasized that ModeRNA does not infer the alignment by itself; the alignment must be supplied by the user. Needless to say, the accuracy of the alignment will ultimately determine the quality of the resulting model—exactly as is the case with all methods for comparative modeling of protein structure. Casual modelers must be warned that for large RNA molecules with complex structures, the development of a good alignment may require laborious manual preparation of the input data based on previous expertise of the respective RNA family.

For each position in the target–template sequence alignment, ModeRNA infers a set of operations necessary to generate the model of the target from the structure of the template. These include: copying coordinates of residues that are invariant between the target and the template, introducing substitutions for aligned residues that differ, adding or removing post-transcriptional modifications, processing insertions/deletions (indels) and adding structural fragments for short regions without a template. ModeRNA generates coordinates of the modeled target RNA and a report with detailed information about all steps of the modeling process. The whole program is implemented in Python and freely available as open source. Below, we explain in detail all operations performed by ModeRNA.

Base exchange operations

For residues that are identical between the template and the target, the coordinates of all atoms are copied from the template residue to the model, without any changes (at least initially). When a substitution in the alignment occurs, coordinates are also copied for the whole residue, followed by a base exchange. The target base is loaded into the model and superimposed onto three atoms of the template base adjacent to the glycosidic bond (e.g. N9, C8 and C4 for adenine), then the atoms of the template base are removed. Both transversions (replacement of purine by purine and pyrimidine by pyrimidine) and transitions (replacement of purine by pyrimidine and vice versa) are modeled this way. This operation preserves the conformation of the backbone (ribose and phosphate) as well as the torsion angle of the glycosidic bond.

Post-transcriptionally modified nucleosides

One of the features of ModeRNA that distinguishes it from most other modeling programs is that it can recognize modified nucleosides in the template structure and in the sequence alignment and preserve, add, or remove them accordingly in the model building process. Post-transcriptional modifications of nucleosides are crucial for the function of RNAs; they appear to be as important as post-translational modifications are for function of many proteins. In tRNA, by far the most abundantly modified RNA, they aid in folding into a well-defined tertiary structure, in fine-tuning the recognition by aminoacyl-tRNA synthetases, and allow for multi-codon specificity for the anticodon loop (34). In rRNA, modifications also increase the efficiency of translation, and are responsible for bacterial resistance to ribosome-targeting antibiotics (35). Modifications have also been observed in mRNA, and various types of non-coding RNAs, including snRNA, snoRNA and miRNA. To date, 115 different nucleotide modifications have been characterized, and this number is still growing (36). About half of them are methylations, which can occur at almost every atom of standard bases and/or at the ribose 2′OH group. More complex modifications such as aminoacylations, formylations, sulfurylations, isoprenylations and combinations of multiple modifications have been also observed. In most modifications, functional groups are added or substituted (e.g. O to S, NH_2 to O), but e.g. pseudouridine formation requires an isomerization, and queosine formation requires a replacement of the original base by an independently synthesized new base in the course of a transglycosylation reaction.

Several different naming schemes for nucleotide modifications have been used, e.g. for the base 5-methylcytidine the abbreviations 5mC, m5C, m^5C, mC^5 and mC have been used in literature. For representation at the sequence level, one-letter abbreviations for some modified bases have been introduced by Sprinzl and coworkers (37), but the number of currently known modifications exceeds the number of letters in the Latin alphabet. To allow for alignments containing all possible modified nucleotides, ModeRNA can recognize not only the one-character symbols, but also an unambiguous numbering scheme recently introduced in the MODOMICS database (36). The PDB is also inconsistent in naming different modified residues (e.g. in the PDB entry 1F7U

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

1–methyladenosine from chain B, residue 958 is named '1MA', while the chemically identical residue from the PDB entry 1OB2, chain B, residue 58 bears the name 'MAD'). To recognize modified nucleotides in RNA structures, a subgraph matching algorithm that matches the topology of atoms in each residue to patterns representing each of the 115 modifications was implemented. Thus, modified nucleosides with non-standard or even incorrect names can be identified based on their chemical structures. In the output file, ModeRNA by default applies these names of modifications that most frequently occur in the PDB.

For adding modifications to standard bases, a set of 67 small fragments covering all chemical groups in the currently known 115 modifications has been created. Each such fragment contains atoms belonging to a modification, and a triplet of connecting atoms that are used to fit the fragment onto an existing standard base. For removing modifications (or atoms that are replaced in the course of a modification), either the excess atoms are removed (e.g. for a small functional group such as a methyl) or an unmodified base is added by superimposing it onto the original base, followed by removal of the original base.

Modeling of indels

Modeling of insertions and deletions is probably the most challenging and crucial step in comparative modeling. In the case of insertions in the target sequence (gaps in the template), additional nucleotide residues must be introduced to the RNA model being created. Examples of such situations include the introduction of a bulge into a helical stem, loop enlargement, extension of a helix or of a terminal tail, or even introduction of an entire new element of secondary structure. In the case of deletions in the target sequence (gaps in the target), the relevant residues have to be removed from the template, and the resulting ends must be sealed to restore the continuity of the backbone. Such operations may involve the replacement of a longer segment of sequence (e.g. a loop) by a shorter one that has a more extended conformation.

Indel modeling in ModeRNA follows the fragment insertion approach, similar to the one widely used in comparative modeling of proteins (38) and implemented e.g. in SwissModel (33). A fragment includes the residue(s) to be inserted and counterparts of residues that flank the indel in the template. The default distribution of the ModeRNA package allows for inserting fragments up to 17 residues long (not counting the flanks). The choice of the maximum length was conditioned by the size of the library file (20 MB). The fragment library includes 131 316 fragments (n-grams) of RNA structure that are 2–19 residues long and have a continuous backbone. It has been derived from the representative set of 172 RNA tertiary structures in the RNADB2005 set (39), which provides manually curated, non-redundant RNA structures from different families, including large structures e.g. the ribosome, and is expected to cover all known types of local RNA structure. For modeling of longer insertions, a larger library covering fragments up to 100-nt long, derived from the same database, is available for download from the ModeRNA website. While this article was under review, a similar approach has been validated as a reliable method for modeling of 3–32-nt long loops by another group (40).

For each indel, ModeRNA attempts to identify a backbone fragment with an appropriate length and superimposes its flanking residues onto the corresponding anchoring residues in the template structure as to maximize its fit to the anchor and to minimize steric clashes with the rest of the molecule. The fragment search includes a pre-filtering stage, where the geometry of the flanking residues is compared to all fragments of appropriate length from a library, and a fitting stage where the 50 most promising candidates are evaluated by inserting them into the model. If the gap cannot be closed by the above-mentioned procedure, e.g. if an extended fragment of the template is to be deleted and the resulting ends are too far from each other, ModeRNA will generate a model with an unsealed gap, and generate a warning that the model is discontinuous. Such situations often occur when modeling is attempted for a wrong template or in regions where the target–template alignment is erroneous. In the pre-filtering stage of the fragment search, the geometrical fit between the two flanking residues from the 5′-end (r5) and from the 3′-end (r3) is sought with the terminal nucleotides of each fragment of appropriate length in the library. The geometrical fit is evaluated by comparing six atom–atom distances ($O5'_{r5}$–$O5'_{r3}$, $C5'_{r5}$–$C5'_{r3}$, $C4'_{r5}$–$C4'_{r3}$, $C3'_{r5}$–$C3'_{r3}$, $C1'_{r5}$–$C1'_{r3}$, $N1_{r5}/N9_{r5}$–$N1_{r3}/N9_{r3}$)—see Figure 1. The sum of the square deviations of these values between the fragment ends and the flanking sites in the model roughly approximates the RMSD, and can be calculated rapidly for a huge number of candidate fragments.

After the pre-filtering stage, the spatial complementarity of the 50 best-scoring fragment candidates is

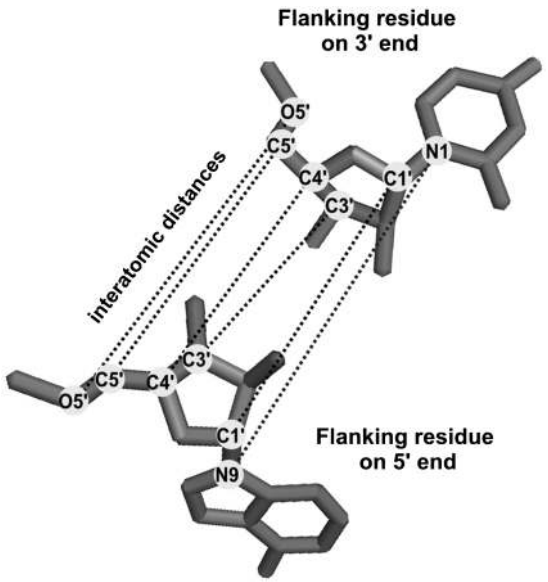

Figure 1. Distances used in the pre-filtering stage of the fragment search.

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

examined: first, each of them is inserted into the model by superimposing the $C1'_{r5}$, $C4'_{r5}$, $C3'_{r5}$, $O3'_{r5}$ and $N1_{r5}/N9_{r5}$ atoms of the terminal residues at the 5′-end, and $C1'_{r3}$, $C4'_{r3}$, $C3'_{r3}$, $O5'_{r3}$, $N1_{r3}/N9_{r3}$ and $C5_{r3}'$ at the 3′ end. Second, the bases in the fragment are exchanged to obtain the correct sequence by adding the coordinates of bases in the same way as described in the base exchange and nucleotide modification sections above. For ranking the 50 candidates, the pre-filtering score is combined with the RMSD value of the superposition of r5 and r3 of the fragment with the model, the sequence dissimilarity (between the sequence to be inserted and the sequence of the database fragment) and the number of interatomic clashes:

$$\text{score} = a \times \text{RMSD} + b \times \text{seq\_dissimilarity} + c \times \#\text{clashes}$$

using the weight parameters $a = 10.0$, $b = 1.0$, $c = 2.0$. The sequence dissimilarity between the target and the fragment is calculated using a matrix containing 0.0 for identical nucleotides, 0.1 for nucleotides versus their modifications, 0.5 for transitions and 1.0 for transversions and nucleotide substitutions involving modifications. Finally, the best-scoring candidate is inserted into the model, retaining two residues r5 and r3 flanking the insertion site from the template.

A major limitation of the above-mentioned fragment insertion procedure is that ModeRNA cannot infer by itself base pairs between the fragment and the rest of the molecule, unless they are explicitly specified by the user (see the section on modeling of secondary structure). ModeRNA also cannot extensively optimize van der Waals interactions. Further optimization of the local conformation and scoring of additional models with alternative loop conformations may be achieved with more specialized software, such as methods for molecular dynamics. The development of methodology for *ab initio* loop modeling and other types of refinement of RNA structural models is however beyond the scope of this manuscript and will be the subject of our future studies.

Backbone remodeling

Possible discontinuities in the backbone (resulting e.g. from an imperfect match of fragment ends to the flanking regions of the template), are repaired using the Full Cyclic Coordinate Descent (FCCD) algorithm that connects two ends with a minimal number of operations (41). ModeRNA rebuilds coordinates of the RNA backbone atoms between two residues, aiming to restore the following native-like features, ordered according to the priority (i) acceptable bond lengths, (ii) absence of interatomic clashes, (iii) acceptable bond angles, (iv) acceptable torsion angles. Acceptable values of bond lengths and angles have been taken from a statistical analysis of structures in our fragment library. Acceptable torsion angles were directly taken from Richardson *et al.* (39). To avoid clashes, 42 RNA suites defined by Richardson *et al.* (39) are tried one after another as starting conformations, until a clash-free loop closure is found. Subsequently, the positions of the most flexible P and O5′ atoms are optimized by a simple stochastic search algorithm trying to satisfy angle and dihedral constraints. For generating coordinates at various stages of the procedure, the NeRF algorithm used in ROSETTA (42) has been implemented. In case the entire procedure fails to close the backbone, details about the kind of distortion for the residues flanking the problematic site are reported.

Modeling secondary structure

To enable modeling of base pairing interactions, a number of commands have been included in ModeRNA. First, the secondary structure of any loaded RNA chain can be detected and saved in the dot-bracket format. Second, helical structures can be inserted as extensions of any canonical Watson–Crick base pair present in the starting structure that can serve as an anchor. Helices can be also extended or shortened by adding or removing internal base pairs. Third, any custom structural element can be introduced between four residues defined as anchors. Fourth, the search for fragments to be inserted from the single chain fragment database can be restricted with a user-defined secondary structure (e.g. enabling insertion of stem–loop elements). Fifth, individual nucleotide residues can be added as base-pairing partners to defined unpaired nucleotides in the starting structure.

Implementation

ModeRNA is available under the GPL Open Source license, with implementations for UNIX and Windows systems. The program requires Python and the BioPython library (43). Several functions for numerical calculations that are part of the PyCogent library (44) have been included in the ModeRNA code. For Windows, an executable version is available that does not require installation of any additional software. It does not require large computational resources, for instance one tRNA model can be built in 2–20 s on a standard PC with one 2.4 GHz processor, with the exact time depending mostly on the number and size of indels that must be modeled by the fragment insertion procedure.

The program is accompanied by an extensive test suite, which consists of test functions that check whether individual parts of the code (functions, classes, modules) produce a predefined output given known input data. Separate sets of tests are available for each of the 37 scripting functions, for the command-line interface (acceptance tests), and for each underlying program component (unit tests). Many of these tests use data from tRNA modeling cases (see 'Results' section), or particularly difficult examples found during ModeRNA development. Others were introduced to ensure the absence of particular program bugs. In total, 548 test functions have been created. The purpose of automatic tests is to guarantee that existing functionalities stay intact while new features are added to the program. A user can also execute the test suite in order to check whether ModeRNA has been installed correctly.

Documentation

The ModeRNA web page (http://iimcb.genesilico.pl/moderna) provides the source code and the binary

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

version of the program, as well as comprehensive documentation. In the tutorial section, two examples of modeling *Escherichia coli* tRNAs (Asp^{QUC} and Tyr^{GUA}) are described. Further, the FAQ section provides answers to many basic question like 'How to write an input script?' or 'Which command-line options are available?', as well as an introduction to more sophisticated features of ModeRNA, such as modeling of DNA structures. The website also includes a complete list of available commands with usage examples, and a basic introduction for Python programmers willing to customize ModeRNA for their own purpose. Detailed instructions about input preparation and ModeRNA usage can be also found in the Supplementary File 1.

RESULTS

Models

In order to test ModeRNA, we created models of 99 tRNAs with known (experimentally solved) structure. We used each of them as a target to be modeled on each of the other 98 structures as templates ($99 \times 98 = 9702$ models total). Additionally, we created a control set consisting of 99 tRNA models, where each sequence was modeled on its 'own' structure as a template. The tRNA family was selected for this analysis, as a large set of experimentally solved structures is available, and despite the high structure conservation a lot of local variation exists. Besides, tRNAs contain many modified nucleotides, which enabled us to test the procedures of modification modeling. Pairwise target–template alignments used for tRNA modeling were extracted from the Rfam database (4). In this test, we used a fragment library, from which tRNA fragments had been excluded. Likewise, in other examples analyzed below, the native structures being modeled have been also removed from the fragment library. In the ModeRNA release available for download, all these fragments has been included to provide the widest possible diversity of local structures.

We managed to build all tRNA models automatically within 48 h on a 2 GHz Intel Xeon processor. For completeness of their sequence 9675 models passed checks and were subject to automated geometry optimization. One hundred and twenty-six models have not been generated, as it turned out all of them involved the structure of tRNA (Tyr^{GUA}) in complex with tyrosyl-tRNA synthetase (1H3E) as a template or its sequence as a target. In this structure, residue 16 labeled as an unmodified cytidine has no base coordinates (i.e. represents an abasic site), hence presents an inconsistency that cannot be resolved automatically by the present version of the method and must be dealt with manually. It is expected that additional intervention of the user either in the modeling process or in verification and curation of all target–template alignments may lead to improvement of many models analyzed in this work. However, the purpose of this benchmark is the assessment of ModeRNAs functionality for automated modeling rather than the assessment of our own qualifications as experts in manual modeling, hence we have not attempted any such intervention.

There are two ways in which the quality of the models can be assessed with ModeRNA. First, a report on the geometry of nucleotides can be generated: ModeRNA calculates bond lengths, bond angles and values of dihedrals, compares them to reference values derived from high-resolution structures, and reports outliers. Thereby, discontinuities of the backbone and unnatural backbone torsion angles can be detected. It has to be pointed out, though, that such outliers in PDB structures are not uncommon (45). Thus, the geometry check of the model makes sense only in the light of an analogous check of the template structure used. Second, ModeRNA can detect clashes between Van der Waals spheres of atoms and thereby highlight cases where modifications, substitutions or indels resulted in base or backbone atoms to collide.

Evaluation of tRNA models

In order to provide a reasonable evaluation of our tRNA models, the Global Distance Test—Total Score (GDT_TS) was calculated, as it is widely used for the evaluation of protein models (46). The GDT_TS score denotes the sum of percent of residues that are within the 1, 2, 4 and 8 Å sphere between a superimposed model and native reference structure, divided by 4. The result can vary from 0 to 1 where a higher value denotes a more accurate model. The distance between two residues was calculated as the average of distances between the corresponding P and C4′ atoms (template versus model). In case of the 9675 tRNA models built by ModeRNA, the average GDT_TS value was equal to 0.5. The detailed results are shown in Figure 2.

On the average, models and native structures exhibit an all-atom RMSD of 5.6 Å (Figure 2). The average RMSD of the P and the C4′ atoms was calculated from all model-reference structure pairs as 5.2 Å. For comparison, the average RMSD for all template–template pairs is 4.9 Å. Thus, both the models and the experimentally solved tRNA structures exhibit similar structural diversity. Figure 2 shows that most of the models are roughly as similar to the native structure as the template structures used (which is a general characteristic of homology modeling). It is worth emphasizing that in 2135 cases the models had the RMSD with respect to the original structure equal to or lower than the template.

The acceptor stem and the anticodon loop of tRNA are known to be highly flexible and to change their conformation upon complex formation with proteins (e.g. enzymes acting on tRNA) or in the ribosome (47). This can be observed in the model of the ribosome-bound *E. coli* $tRNA_{Phe}$ 2J00_W (PDB-ID 2J00, chain W) built on the template 2HGP_D (Figure 3), where the ribosome is in post-initiation conformation. The resulting all-atom RMSD is as high as 3.6 Å despite the two molecules have 100% identical sequences. This high RMSD value is mostly caused by residues 17 (RMSD = 8.5) and 47 (RMSD = 6.7), where the bases have different orientations in the model and in the native structure, and the two residues preceding the CCA 3′-end. A similar observation can be made for the model of *E. coli* $tRNA_{Thr}$ built on *E. coli* $tRNA_{Phe}$ (1B23_R) as the template (Figure 3).

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

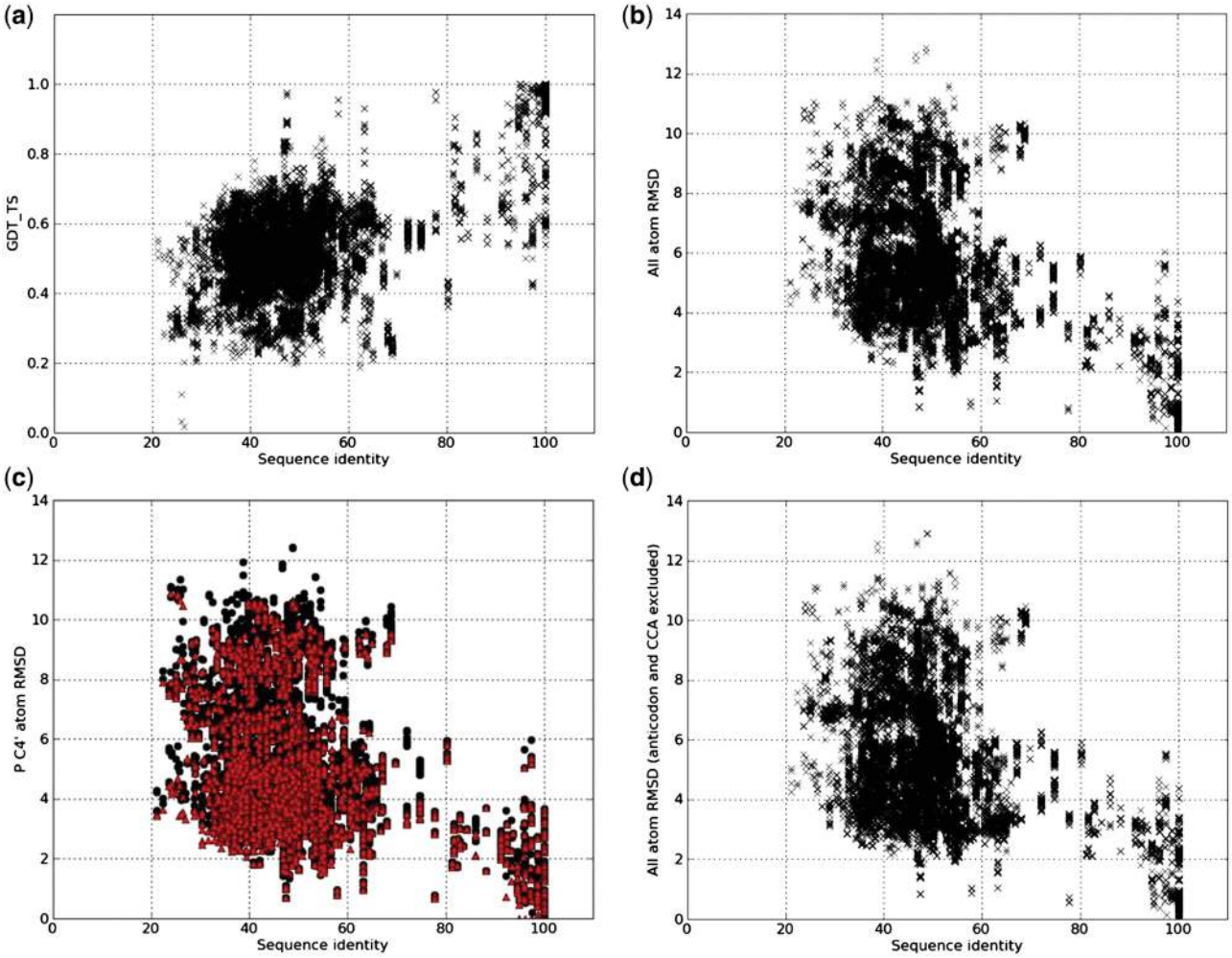

Figure 2. Evaluation of tRNA models generated from templates and alignments—relation between the sequence identity and (**a**) the GDT_TS score, (**b**) all atom RMSD of models against experimentally solved structures, (**c**) the P atom and the C4′ atom RMSD of models against experimentally solved structures (in black) and experimentally solved structures (templates) against experimentally solved structures (targets) (in red) and (**d**) all atom RMSD of models against experimentally solved structures where the anticodon and the CCA regions are excluded.

The native structure of $tRNA_{Thr}$ (1QF6_B) is interacting with threonyl-tRNA synthetase, while the template structure is in contact with the translation elongation factor EF-Tu, resulting in a mutual shift of the acceptor stem and the anticodon loop by several Å. This illustrates that the accuracy of comparative models of tRNAs (and by extension—all RNAs) is limited by availability and correct choice of a template in an appropriate physiological state. However, we must emphasize that as with proteins, the problem of template selection is to a large extent independent from the actual process of model building.

In order not to take into account the most flexible regions in the all-atom RMSD calculation, five residues from the anticodon loop were excluded from our calculations (the residues with numbers 33–37 or 38–42 in case of 2V0G_F, 2BTE_B, 2BTE_E, 2V0G_B, 2BYT_E and 2BUT_B). For the 3′-end of tRNA the terminal CCA residues were excluded. If these regions are excluded, the average RMSD between the tRNA model and the native structure in our benchmark drops down to 4.9 Å (Figure 2).

The analysis of results presented in Figure 2 reveals that, expectedly, high sequence similarity is correlated with lower RMSD. Some tRNA models exhibit high accuracy despite low sequence similarity to the template used, for example 2HGI_C based on the template 2B64_V (Figure 3). There is, however, a group of models that exhibit a high RMSD value to the native structure despite high sequence similarity to the template used. This is mostly caused by different conformations of the same molecule observed in different RNA–protein complexes. However, we also found a few cases such as the model for 2DXI_C built on the template 2DET_C (Figure 3), sequence identity of 72%), whose RMSD to the native structure is 8.0 Å. The deviation of the model is mainly caused by six residues from the 3′-end that had to be modeled '*de novo*' because they were missing from the template. ModeRNA inserted a helical fragment there, while the native structure folds in a different way.

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

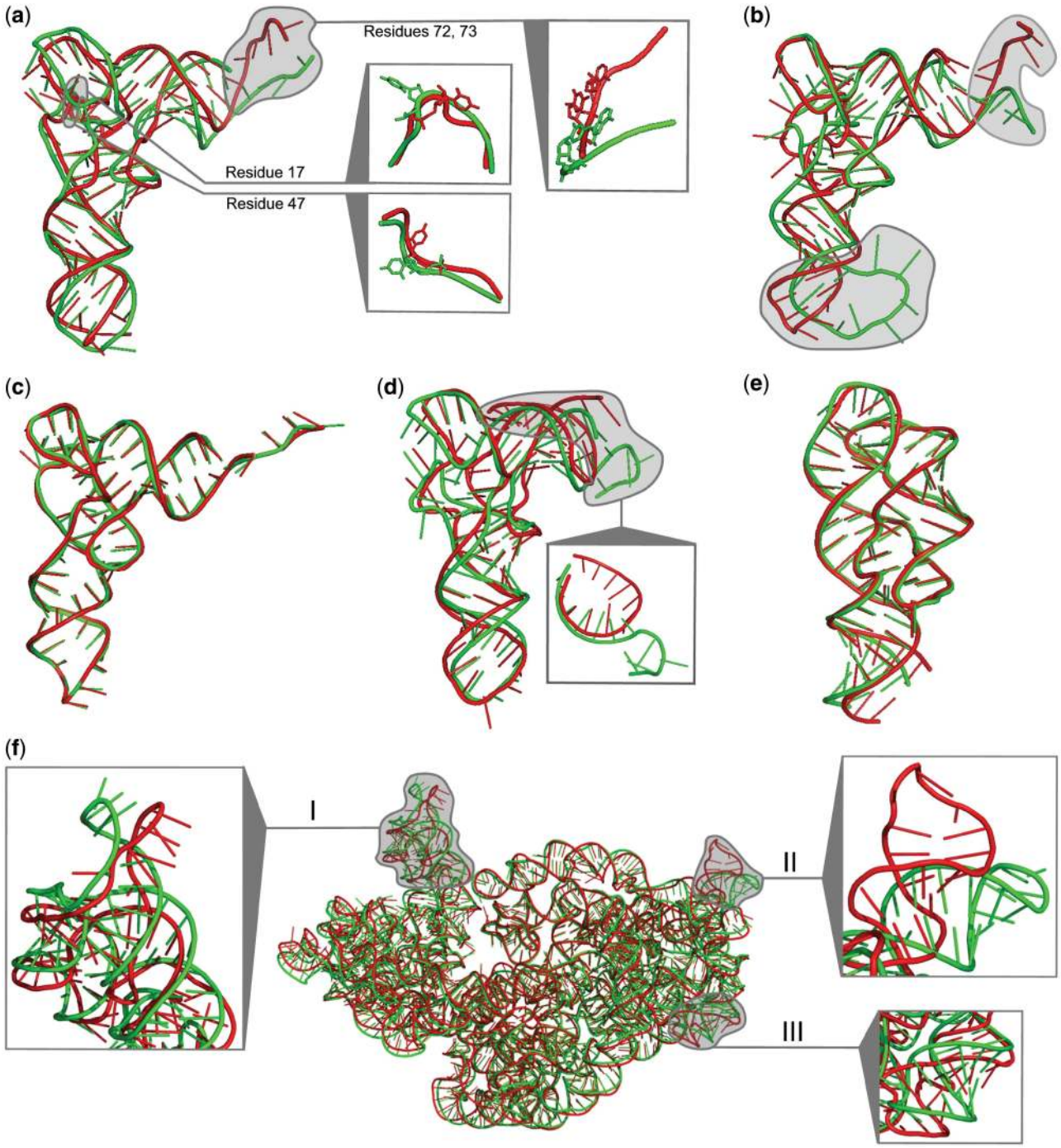

Figure 3. Examples of models built by ModeRNA. Models are shown in red, the native structure is shown in green. (**a**) Model of *E. coli* $tRNA_{Phe}$ (native structure 2J00_W) built on the target 2HGP_D (*E. coli* $tRNA_{Phe}$). The sequences of both molecules are 100% identical, the RMSD value is relatively high—3.61. The residues that contribute the most in the high RMSD are marked with gray clouds and their conformation is shown in separate boxes. (**b**) Model of *E. coli* $tRNA_{Thr}$ (the native structure 1QF6_B—PDB-ID 1QF6, chain B) built on the template 1B23_R *E. coli* $tRNA_{Cys}$. The native structure is interacting with threonyl-tRNA synthetase, while the template structure is in contact with the translation elongation factor EF-Tu-shifting the conformation of the acceptor stem and anticodon loop by several Å, both regions are marked by gray clouds. (**c**) Model of *E. coli* $tRNA_{fMet}$ (native structure 2HGI_C) built on the template 2B64_V (*E. coli* $tRNA_{Phe}$). Model structure has low RMSD—1.38 Å despite medium sequence similarity (47%) between the target and the template molecule. (**d**) Model of $tRNA_{Glu}$ (native structure 2DXI_C) built on 2DET. Both structures have a high-sequence similarity (72%). Yet, the RMSD amounts 8.05 Å. The reason is the 6 nt long fragment that is missing in the template on the 3′ end. In the model it has a completely different conformation than the native one. (**e**) Adenine-binding riboswitch (1Y26) modeled using a guanine-binding riboswitch (1Y27). (**f**) 30S ribosomal subunits from *T. thermophilus* (1J5E_A) modeled using 30S from *E. coli* (2AVY_A). Three regions where the model did not match the native PDB structure well are highlighted: I—two hairpins connected by a junction (residues 970–1022), II—stem loop (residues 65–89), III—stem loop (residues 173–196).

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

In our evaluation we also included the deformation index (DI) and the deformation profile (DP) measures recently introduced by Parisien *et al*. (48). The DI evaluates the conservation of base–base interactions (including stacking interactions and both canonical and non-canonical base pairing). The DP is a matrix of average distances, calculated by superimposing all nucleotides from model and the reference structure and then computing the average distance between each base from model and the corresponding base from the reference structure. For the set of tRNA models we obtained an average DI of 0.62 and an average DP of 13.82 (Figure 4). According to the histogram, the majority of models achieve a DI score in the range of 0.5–0.8. The low average is caused mainly by changes in intermolecular interaction patterns between tRNAs in different functional states (e.g. splaying out bases upon complex formation with protein partners). A more in-depth analysis of contact patterns in tRNA would be beneficial to improve this particular aspect. The distortion profile average and maximum values are higher than for the examples given in the article by Parisien *et al*. (48). The main reason for this are: (i) the DP score is size-dependent and tRNA molecules are larger than the mentioned examples and (ii) the conformation and interactions of residues in the anticodon loop and in the acceptor stem exhibit large variations (depending on the functional state and interactions with other macromolecules). When different structures are superimposed based on these residues, other residues can be displaced to generate global RMSD values up to 50 Å. Nevertheless, the majority of residues generate superpositions with RMSD values around 5 Å.

Other modeling examples

In order not to restrict tests of ModeRNA to one family of RNAs, tests on other families were also carried out. As mentioned earlier, in each modeling exercise we used a variant of the fragment library, from which the native structure of the target has been removed. As an example for a straightforward modeling case, we have built two models for the riboswitch structures 1Y26 (adenine specific) and 1Y27 (guanine specific). The all-atom RMSDs for 1Y26 built on 1Y27 is 1.6 Å, while for 1Y27 built on 1Y26 is 1.2 Å, and the GDT_TS scores are 0.8 and 0.9, respectively. The specificity-determining differences in ligand-binding sites of both riboswitches were modeled successfully (Figure 3).

Two more complicated example models were built for 23S rRNA sequences extracted from the 30S ribosomal subunit structures of *T. thermophilus* (1J5E) and *E. coli* (2AVY) using each other as templates. For aligning their sequences, the size of the structures prevented the usage of a structural alignment program [e.g. Locarna (49) failed to align them], and no exact sequence match in a rRNA database could be found. Instead, a manual alignment was generated with RNAView (50), aided by comparison of secondary structures. The model of *T. thermophilus* 23S rRNA based on the 2AVY structure, and the model of the *E. coli* 23S rRNA based on the 1J5E structure exhibited an all-atom RMSD of 5.5 and 5.2 Å, respectively. The structures are shown in Figure 3. The modeling of residues 180–228 in the *T. thermophilus* structure turned out to be challenging. It was already recognizable during the alignment that the secondary structures of the target and the template differ considerably in this region. It was therefore not surprising that in the region between residues 179–191 the per-residue deviation rises above 10 Å. A few other divergent regions are the main contributors to the high RMSD, while most of the structure exhibits almost perfect matches between the model and the template.

We also constructed a model of the *Azoarcus* group I intron, to enable comparison with a model of the same molecule generated by the RNAbuilder program described in a recent publication by the Altman group (51). A pairwise alignment of the *Twort* group I intron template structure (1Y0Q_A) and the target sequence (corresponding to 1U6B) was created according to the description in the mentioned article. A first model with an overall RMSD of 6.9 could be built using ModeRNA standard,

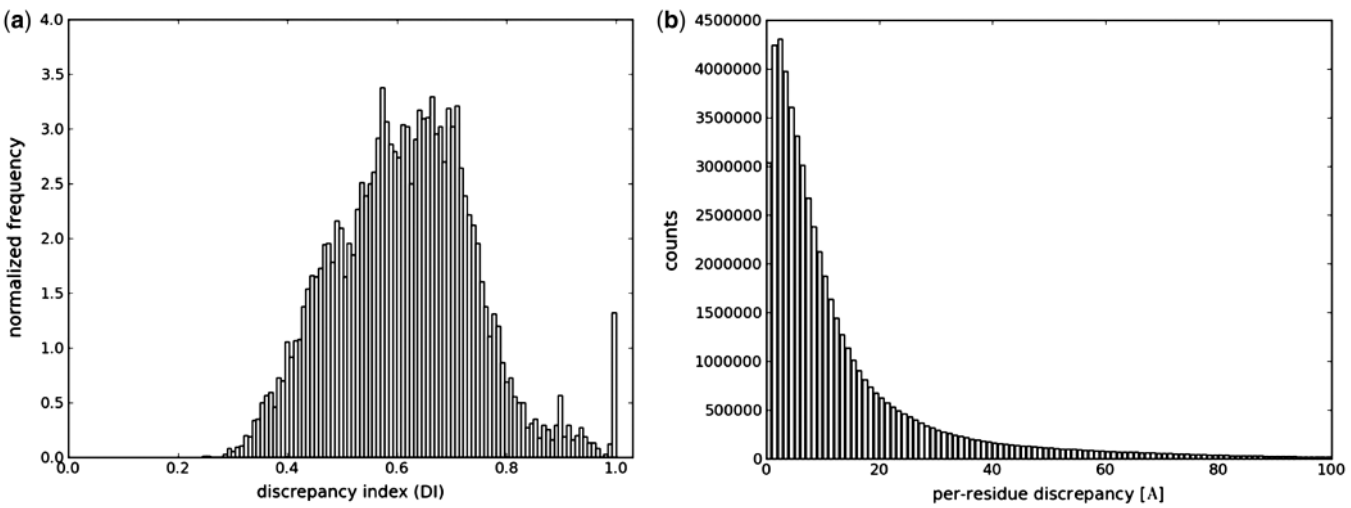

Figure 4. Evaluation of tRNA models generated from templates and alignments with (**a**) DI and (**b**) DP measures.

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

naïve mode of operation (one template + target–template alignment → model). Subsequently, we used the scripting interface to model the L9/P5 and L2/P8 tertiary loop interactions using the *Tetrahymena* intron as an additional template, according to the description of the advanced modeling protocol used for modeling by the Altman group. With these additional commands, the accuracy of the model generated with ModeRNA improves significantly, with the RMSD reaching 4.3 Å for the whole structure, and 2.0 Å for the core. These values compare favorably with the RMSD of 4.4 and 2.7 (entire model and the core alone) calculated for the model created by RNABuilder. The model and the experimentally solved structure are presented in Figure 5 and the detailed RMSD values comparison between the ModeRNA and the RNABuilder models are shown in Supplementary Table S1. The RMSD values for both models were obtained with the same script, which is available on request. We conclude that the lack of a sophisticated 'folding simulation' procedure for model refinement in ModeRNA does not prevent it from building accurate models. On the other hand, using multiple templates and utilizing the advanced mode of ModeRNA can significantly improve the accuracy, compared to modeling based on a single template.

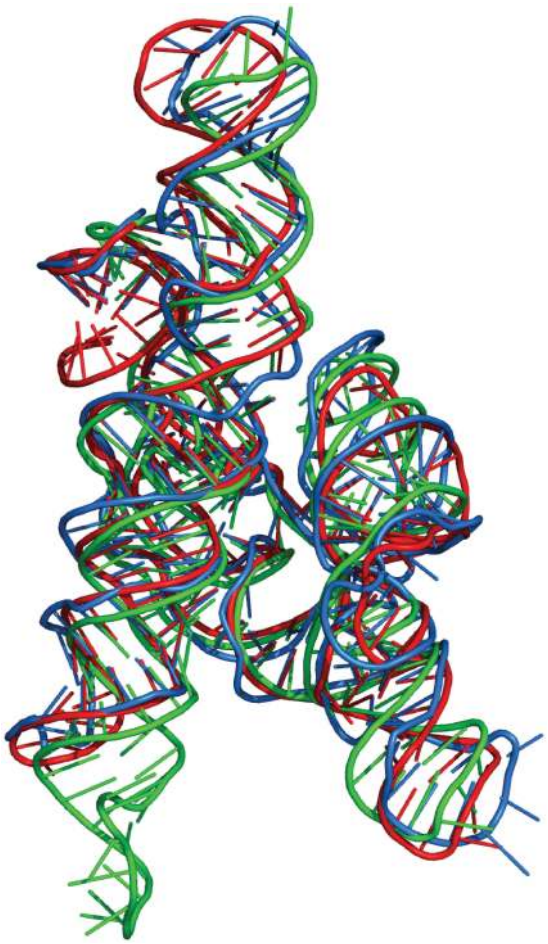

Figure 5. The model of *Azoarcus* group I intron built with ModeRNA (in red) and with RNABuilder (51) (in blue) compared with the experimentally solved structure, PDB code 1U6B (in green).

To illustrate the influence of alignment quality on comparative modeling, we have generated additional models of the *Azoarcus* intron based on the *Twort* intron structure alone, using the expert alignment (51), alignments automatically calculated with the Infernal method (52), or with ClustalW (53). Detailed comparisons between the models and the experimentally determined structure are shown in Supplementary Table S1. It is evident that the use of a single template (without any additional restraints) decreases the quality of the model (RMSD for the whole structure 6.9 and 12.4 Å for models built with ModeRNA and RNAbuilder, respectively), and that the expert alignment is superior to automatic alignments. In this particular case study, both ClustalW and Infernal failed to produce functionally relevant alignments, as judged by the poor quality of the resulting models (RMSD >20 Å, even if the region of comparison is limited to the active site). It is therefore evident that ModeRNA, as all methods for comparative modeling, fails to build correct models with incorrect target–template alignments as an input, and at this moment there seems to be no perfect 'purely computational' solution to this problem. On the other hand, our test demonstrates that ModeRNA can build reasonable 3D models of complex RNA molecules, using a 'protein-like' approach based on single templates, if an approximately correct (e.g. expert-guided) alignment is provided.

DISCUSSION

Comparison of ModeRNA with existing software

The currently available modeling programs are characterized by different features and can be used in different modeling scenarios. We have summarized their main features in Table 1 and we discuss some of the comparative and knowledge-based modeling programs below. The field of RNA prediction is rapidly developing and thus we do not claim this list to be complete. A review of *ab initio* and *de novo* folding methods is beyond the scope of this article; such methods have different purpose than ModeRNA and have been referred to only in the context of possible (re)modeling of regions that lack the template for comparative modeling.

RNABuilder (51) is a new method for comparative modeling of RNA structures, in which RNA molecules are simulated in parallel at multiple levels of detail, ranging from coarse-grained resolution to atomic scale (including hydrogen atoms). It uses internal coordinate multibody dynamics to satisfy constraints on all these levels. The coarse grained force field consisting of forces and torques acts to bring together bases into a base pair geometry indicated by the user. The user can also set flexibility or rigidity of the molecule. As mentioned in the Results section, RNABuilder has been tested on the *Azoarcus* group I intron. As shown in this article, ModeRNA can build a model of the same molecule (with the same template and the same target–template alignment) with comparable accuracy, despite using a

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

Table 1. Comparison of RNA 3D structure prediction methods

Features	Modeling program											
	ModeRNA (this work)	RNABuilder (51)	PARADISE/ Assemble (15)	RNA2D3D (18)	ERNA-3D (17)	MC-Fold/ MC-Sym (30)	FARNA/ FARFAR (56)	NAST (58)	DMD/ iFoldRNA (57)	YUP (72)	PyMOL (54)	AmiraMol (55)
Prediction from sequence (*de novo*)	○	○	●	○	○	●	●	○	●	○	○	○
Prediction from secondary structure	○	●	●	●	●	●	○	●	○	●	○	○
Prediction from template (*comparative*)	●	●	●	●	●	○	○	○	○	○	○	○
Uses alignments	●	●	●	○	○	○	○	○	○	○	○	●
Secondary structure constraints	●	●	●	●	●	●	●	●	○	●	○	○
Distance/tertiary constraints	○	●	●	●	●	●	○	●	○	●	○	○
Modeling of modified nucleotides	●	○	○	○	○	○	○	○	○	○	○	○
Fragment library	●	○	●	●	○	●	●	●	○	○	○	○
All-atom representation	●	●	●	●	●	●	●	●	○	○	●	●
Manipulating individual residues/coordinates	●	●	●	●	●	○	○	○	○	●	●	●
Molecular dynamics/Monte Carlo/replica exchange	○	●	○	○	○	○	●	●	●	●	○	○
Energy minimization	○	○	●	●	○	●	●	○	●	●	○	●
Geometrical analysis/report	●	○	●	○	○	●	●	○	●	●	○	○
Graphical interface	○	○	●	●	●	○	○	○	○	○	●	●
Scripting interface	●	●	○	○	○	●	●	●	○	●	●	●
Web interface	○	○	○	○	○	●	○	○	●	○	○	○
Freely available	●	●	●	●	○	●	●	●	●	●	●	○
Open source	●	○	○*	○	○	○*	○	●	○	○	●	○
Platform	Win, Lin, Mac	Win, Mac	Lin, Mac	Lin	Win, SGI	Lin, Mac	Win, Lin	Win, Lin, Mac	All	Lin, Mac	All	Win, Lin, Mac

A black dot denotes a presence of a particular feature, an open circle denotes lack of the feature. A star indicates that the source code is available upon request from the authors. In the case of PARADISE/Assemble, MC-Fold/MC-Sym, NAST, YUP and AmiraMol the list of features was kindly verified by the authors.

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

completely different approach. ModeRNA and RNABuilder are based on different principles and in remote analogy to the protein modeling software may be regarded as counterparts of SwissModel (33) and Modeller (32), respectively.

S2S/Assemble (available via the PARADISE platform) (15,16) is a graphical system that combines various tools and web services into a powerful environment to edit sequences and structures of RNA. It contains explicit annotation of base pairing and stacking interactions, multiple sequence alignments, a motif library and an automatic procedure to generate 3D models from the annotation or *de novo*. The drawback of the system is that all interactions have to be annotated manually, and it is therefore difficult to perform a high-throughput analysis, such as the one presented here. As the user has control over every single interaction and can force tertiary interactions from the sequence level, the modeling process relies strongly on the user's experience. Thus, the shape of the final model will largely depend on the user's knowledge and expertise. We envisage that PARADISE (and perhaps other programs for 'graphics-based interactive expert modeling' mentioned below) may be used to manipulate and combine models generated with ModeRNA, e.g. to build complex structures that require spatial assembly of individually modeled elements.

RNA2D3D (18), as the name implies, generates approximated 3D models starting from a sequence and secondary structure. The program procedure does not include simulation of the folding process and for this reason it cannot automatically generate a satisfactory 3D model from base paring information alone. Manual manipulation is required, and RNA2D3D provides a graphical interface, where individual residues and base pairs can be manipulated individually. For example, for modeling of tRNA, at least two important manual changes need to be done to obtain the typical L shape (stem-stacking operation and rigid-body rotation of a stacked pair).

ERNA-3D (17) is another program for manual modeling of RNA 3D structure. ERNA3D provides a graphical interface, where the user can manipulate various structural elements at different levels of abstraction, while the rotations of molecular groups around single bonds along the backbone chain of the molecule are simulated in the real time.

It should be mentioned that programs developed primarily for molecular visualization are also sometimes used to model RNA structures. In such cases, the modeling is performed almost completely manually. Some of these programs have extensive scripting interfaces that make it easy to customize them for very specific modeling tasks. For a comparison of features, we have included PyMOL (54) and AmiraMol (55) in the table.

In the publication of the RNABuilder tool, Altman and coworkers (51) defined four criteria that a comparative modeling program should fulfill: (i) allowing the user to specify correspondences between residues in the target sequence and residues in the template structure, (ii) allowing to use more than one template, (iii) being able to model connecting regions with no template (e.g. indels) and (iv) being accessible to experimentalists who are not knowledgeable in computer science. Here, we examine ModeRNA with respect to these criteria:

(i) The target–template correspondences are specified by a sequence alignment provided by the user (this is a typical feature of all comparative modeling methods, including ModeRNA, PARADISE, RNABuilder and with limitations also AmiraMol). ModeRNA can model post-transcriptional modifications, which can be specified at the level of the alignment. Constructing an explicit Watson–Crick pair at a given nucleotide is a key feature for building RNA structures that is shared by most RNA modeling programs, including ModeRNA.

(ii) The combination of multiple templates can be achieved with ModeRNA, RNABuilder and PARADISE. The use of multiple templates can greatly improve the accuracy of comparative models, as has been shown for the *Azoaracus* group I intron (see Supplementary Table S1). The possibility to explicitly change the length of helical segments and to add fragments with a custom structure allows modeling of complex structural changes. In a wider sense, MC-Sym, FARNA and other methods based on the fragment assembly approach always combine multiple templates, but not for homology-based modeling.

(iii) ModeRNA is able to model regions with no template by using a fragment insertion approach, similarly to the RLooM method (40). Our choice of the fragment library is discussed in more detail in the following section of this article. For short indels (up to 15 nt), a potentially native-like fragment can be inserted from the fragment library, or a set of candidates can be generated for the user to chose from. This situation is similar in both PARADISE and RNABuilder that can also construct short regions without a template. ModeRNA allows for inserting large user-defined fragments and to restrict the search of the fragment database according to the user-defined criterion of secondary structure. None of the comparative modeling methods are capable of folding large fragments that lack correspondencies to the templates, which is the domain of *ab initio* or *de novo* folding methods, e.g. FARNA/FARFAR (29,56), DMD/iFoldRNA (57), the MC-Fold/MC-Sym pipeline (30), or NAST (58). There is clearly a need to develop specialized modeling software that would be able to refold large fragments of RNA structure in the context of a correctly modeled core.

(iv) The capabilities of the existing programs for RNA modeling differ in the amount of user commitment, from e.g. web interfaces that require the user to paste a target sequence, to complex graphical user interfaces with many options. The current version of ModeRNA does not include a graphical user interface, but instead offers a powerful command-line interface, which enables both straightforward modeling, and expert modeling with complex operations with sequences and structures defined by scripts. In order to

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

facilitate expert-guided modeling, the ModeRNA web page presents a collection of modeling examples and situation-specific scripts. Users can write their own scripts, and interface ModeRNA with other programs for data processing and visualization.

Discussion of the fragment library

The primary purpose of the fragment library described in this article is to provide fragments connecting two residues, not to search for particular motifs. To include a defined secondary-structural element, ModeRNA supports the explicit combination of fragments taken from different templates. Several databases with specialized RNA structural motifs facilitate the search for an appropriate fragment. For example RNAJunction provides over 12 000 3D fragments of junction and kissing-loop structures (59), and the structural classification of RNA (SCOR) (60) supports queries for RNA motifs like kink turns and GNRA loops. Also FRABASE (61) references secondary-structural elements in PDB structures. Fragments referenced by these databases can be extracted and added by the ModeRNA user. Important features that allow customized modeling of substructures without corresponding regions in the template are the possibility to search for fragments with a user-defined secondary structure or to insert user-defined structural elements. This goes beyond modeling of simple loops, and allows to assemble RNA models from the mentioned library of 3D fragments. The users must be aware that ModeRNA cannot infer by itself base pairs between the inserted fragment and the rest of the molecule, unless they are explicitly specified in the input file. Automated assembly of secondary-structure fragments would allow for building models *de novo*, similarly to RNA2D3D (18) or FARNA (29), but this type of modeling is clearly distinct from the concept of comparative modeling and is out of scope of the ModeRNA project at its current stage.

The rnaDB2005 structure set used to generate the fragment library for ModeRNA was originally developed for the calculation of backbone conformers (45). It covers a wide range of conformations, and, as the tests on the tRNA family have shown, it can be used to generate fragments with sufficient conformational variability to enable modeling of insertions and deletions in the context of comparative modeling. The fragments are derived from different structural families so no particular kind of structure is favored. It should be emphasized that the rRNA structure alone has been considered as a source of fragments sufficient for model building (29). Further support for the validity of modeling by the fragment insertion approach comes from a publication that appeared while this article was under review (40). It was found that structural similarity is maintained even in large loops, and that it is correlated with sequence similarity, which has reinforced our decision to include the sequence similarity criterion in the fragment selection algorithm. However, in the study carried out by Schudoma *et al*. (40) only loops and single-stranded (i.e. unpaired) segments have been considered. Here, we demonstrate that the fragment insertion approach works well for homology modeling of entire RNA molecules.

Solving the template search problem

An obvious limitation of ModeRNA (as of any program for comparative modeling) is that in order to build a model, a template structure must be provided. Although the PDB database covers many important families, it may be difficult to find a proper template molecule for a particular target. And if a template structure is available, a critical issue is to create an accurate, biologically relevant target–template sequence alignment. However, we must emphasize that searching for solutions to these problems is not a part of a comparative modeling program per se. In the protein structure prediction field, separate methods exist for template identification (called 'fold recognition') and for the refinement of target–template alignments (62) and we believe that such separation of efforts is also justified in the case of RNA comparative modeling, especially that many programs for sequence alignment and sequence-structure alignment already exist.

Pre-calculated alignments are already available for many RNA families e.g. in the Rfam database (4). As our analysis of tRNA structures shows, their usage may require manual refinement. R-Coffee, a multiple RNA alignment package, can be used to refine alignments when 2D structure data is available (63). For cases where no precalculated alignment exists, but a template structure is known (or a particular template is arbitrarily selected by the modeler), a number of tools exist for RNA sequence alignment. There are programs utilizing either sequence information alone like Muscle (64) or ClustalW (53), as well as methods that combine sequence and secondary-structure information like Consan (65) for pairwise alignments, or LocARNA (66), FoldalignM (67) or Stemloc (68) for multiple alignments. A more comprehensive list of available tools is discussed in the article describing the R-Coffee method (63). When no suitable template is known, a database search must be carried out [using homology search tools like nucleotide BLAST on a set of RNA sequences extracted from PDB files (69)]. It is also possible to build RNA secondary-structure profiles or covariance models, if a 2D structural alignment is available for a given family, and then use the covariance models for searching a database for putative homologs, as in the Infernal method (52). However, as demonstrated in this article for the difficult example of group I intron modeling, automatically generated alignments with single templates usually produce models of lower quality than alignments refined by hand based on additional information and utilizing multiple templates. The development of a good alignment (required for successful modeling) usually requires previous expertise of the respective RNA family.

To facilitate finding templates, aligning sequences and identifying secondary-structure elements, the ModeRNA website provides sequences and secondary structures for a representative selection of RNA structures from the PDB that can serve as potential templates. This does not solve the template search problem a priori, but provides a

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

pragmatic tool that can be used under various conditions and does not restrict future use.

Model refinement

All modeling programs generate imperfect structures that contain various errors and inaccuracies. Even in 'globally correct' comparative models (as in the examples discussed in this article) various local features require improvement. ModeRNA by default performs local optimization of backbone geometry to generate physically reasonable conformations if the backbone is distorted. However, for more extensive remodeling and searching for structures that are close to the global energy minimum (which is outside the domain of comparative modeling), users are expected to use other specialized software. ModeRNA contains a function to use the external molecular dynamics package MMTK (70) for model optimization. It can be applied to perform conjugate gradient energy minimization using the AMBER force field to refine the model. The optimization may be restricted to particular regions of the model, in order to lower calculation time. Alternatively, any other molecular dynamics program or statistical potential for RNA can be used. We are currently working with the developers of the Adun package (71) to integrate ModeRNA into that framework, and enable multiscale modeling that would combine comparative and *de novo* modeling. To create a fast and user-friendly method for refinement of RNA models (obtained with ModeRNA or other modeling methods) is on the top of our list for future developments.

CONCLUSIONS

ModeRNA software and detailed descriptions of commands, examples, as well as a tutorial can be found on http://iimcb.genesilico.pl/moderna. ModeRNA is free for all users and released under an Open Source license, which means that it can be customized and integrated with other software.

The main advantages of ModeRNA that distinguish it from other modeling programs are that it provides a flexible scripting framework that can build RNA structures using various strategies (including fast automated modeling based on template structure and target–template alignment without additional data) and that it can handle modified nucleotides. There is no restriction on size of the molecules to be modeled. The automated modeling mode is appropriate for modeling of RNAs that are expected to exhibit high structural similarity to the template, and for which a correct sequence alignments can be easily generated. When more challenging modeling is required, ModeRNA supports advanced operations such as the assembly of structure from fragments and modeling of base pairs specified by the user. We believe ModeRNA will be helpful for researchers interested in studying RNA structure and function and will stimulate the development of other methods for RNA structure prediction. Our future efforts will focus on development of a method for prediction of global and local model quality (potential deviation from the unknown native structure) and on refinement of poorly modeled regions, in analogy to software that exists for proteins.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Russ Altman and his coworkers, who provided us with the model of *Azoarcus* intron as well as with detailed description of the alignment of *Azoarcus* and *Twort* introns, which was essential for the modeling exercise to compare RNABuilder with ModeRNA. We also thank Francois Major, Fabrice Jossinet, Magdalena Jonikas, Stephen Harvey and Daniel Baum for exchanging information about their methods, which allowed us to improve the discussion section of the manuscript. We would also like to acknowledge our colleagues involved in testing preliminary version of ModeRNA, in particular Ewa Wywiał, Paweł Skiba, Piotr Byzia, Irina Tuszyńska, Joanna Kasprzak, Jerzy Orłowski, Paweł Łukasz, Tomasz Osiński, Marcin Domagalski, Anna Czerwoniec, Stanisław Dunin-Horkawicz, Marcin Skorupski and Marcin Feder. We also would like to express our gratitude for recommendations from members of the RNA Ontology Consortium.

FUNDING

Polish Ministry of Science (HISZPANIA/152/2006 grant to J.M.B.); Faculty of Biology, Adam Mickiewicz University (PBWB-03/2009 grant to M.R.). German Academic Exchange Service (grant D/09/42768 to K.R.); 6th EU Framework Program (LSHG-CT-2005-518238 'EURASNET' to J.M.B., M.R. and T.P.). Funding for open access charge: Waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Gesteland,R.F., Cech,T.R. and Atkins,J.F. (2005) *The RNA World*, 3 edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
3. Berman,H.M., Westbrook,J., Feng,Z., Iype,L., Schneider,B. and Zardecki,C. (2002) The Nucleic Acid Database. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 889–898.
4. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
5. Moult,J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, **15**, 285–289.
6. Laederach,A. (2007) Informatics challenges in structured RNA. *Brief Bioinform.*, **8**, 294–303.
7. Krieger,E., Nabuurs,S.B. and Vriend,G. (2003) Homology modeling. *Methods Biochem. Anal.*, **44**, 509–523.
8. Chothia,C. and Gerstein,M. (1997) Protein evolution. How far can sequences diverge? *Nature*, **385**, 579, 581.

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

9. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *Embo J.*, **5**, 823–826.
10. Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
11. Fiser,A., Feig,M., Brooks,C.L. 3rd and Sali,A. (2002) Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.*, **35**, 413–421.
12. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21(Suppl 2)**, ii47–ii53.
13. Krasilnikov,A.S., Xiao,Y., Pan,T. and Mondragon,A. (2004) Basis for structural diversity in homologous RNAs. *Science*, **306**, 104–107.
14. Leontis,N.B., Lescoute,A. and Westhof,E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.
15. Jossinet,F. and Westhof,E. (2005) Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics*, **21**, 3320–3321.
16. Jossinet,F., Ludwig,T.E. and Westhof,E. (2010) Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics.*, **26**, 2057–2059.
17. Zwieb,C. and Muller,F. (1997) Three-dimensional comparative modeling of RNA. *Nucleic Acids Symp. Ser.*, **36**, 69–71.
18. Martinez,H.M., Maizel,J.V. Jr and Shapiro,B.A. (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *J. Biomol. Struct. Dyn.*, **25**, 669–683.
19. Westhof,E., Romby,P., Romaniuk,P.J., Ebel,J.P., Ehresmann,C. and Ehresmann,B. (1989) Computer modeling from solution data of spinach chloroplast and of Xenopus laevis somatic and oocyte 5 S rRNAs. *J. Mol. Biol.*, **207**, 417–431.
20. Krol,A., Westhof,E., Bach,M., Luhrmann,R., Ebel,J.P. and Carbon,P. (1990) Solution structure of human U1 snRNA. Derivation of a possible three-dimensional model. *Nucleic Acids Res.*, **18**, 3803–3811.
21. Michel,F. and Westhof,E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.*, **216**, 585–610.
22. Tung,C.S. and Sanbonmatsu,K.Y. (2004) Atomic model of the Thermus thermophilus 70S ribosome developed in silico. *Biophys. J.*, **87**, 2714–2722.
23. Bergman,N.H., Lau,N.C., Lehnert,V., Westhof,E. and Bartel,D.P. (2004) The three-dimensional architecture of the class I ligase ribozyme. *RNA*, **10**, 176–184.
24. Burks,J., Zwieb,C., Muller,F., Wower,I. and Wower,J. (2005) Comparative 3-D modeling of tmRNA. *BMC Mol. Biol.*, **6**, 14.
25. Kosinski,J., Cymerman,I.A., Feder,M., Kurowski,M.A., Sasin,J.M. and Bujnicki,J.M. (2003) A 'FRankenstein's monster' approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins*, **53(Suppl 6)**, 369–379.
26. Bujnicki,J.M. (2006) Protein-structure prediction by recombination of fragments. *Chembiochem*, **7**, 19–27.
27. Hardin,C., Pogorelov,T.V. and Luthey-Schulten,Z. (2002) Ab initio protein structure prediction. *Curr. Opin. Struct. Biol.*, **12**, 176–181.
28. Tozzini,V. (2009) Multiscale modeling of proteins. *Acc. Chem. Res.*, **43**, 220–230.
29. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci. USA*, **104**, 14664–14669.
30. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
31. Moult,J. (2006) Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **361**, 453–458.
32. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
33. Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
34. Grosjean,H. (2009) *DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution*. Landes Bioscience, Austin.
35. Poehlsgaard,J. and Douthwaite,S. (2005) The bacterial ribosome as a target for antibiotics. *Nat. Rev. Microbiol.*, **3**, 870–881.
36. Czerwoniec,A., Dunin-Horkawicz,S., Purta,E., Kaminska,K.H., Kasprzak,J.M., Bujnicki,J.M., Grosjean,H. and Rother,K. (2009) MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res.*, **37**, D118–D121.
37. Juhling,F., Morl,M., Hartmann,R.K., Sprinzl,M., Stadler,P.F. and Putz,J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.
38. Michalsky,E., Goede,A. and Preissner,R. (2003) Loops In Proteins (LIP)—a comprehensive loop database for homology modelling. *Protein Eng.*, **16**, 979–985.
39. Richardson,J.S., Schneider,B., Murray,L.W., Kapral,G.J., Immormino,R.M., Headd,J.J., Richardson,D.C., Ham,D., Hershkovits,E., Williams,L.D. *et al.* (2008) RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, **14**, 465–481.
40. Schudoma,C., May,P., Nikiforova,V. and Walther,D. (2010) Sequence-structure relationships in RNA loops: establishing the basis for loop homology modeling. *Nucleic Acids Res.*, **38**, 970–980.
41. Boomsma,W. and Hamelryck,T. (2005) Full cyclic coordinate descent: solving the protein loop closure problem in Calpha space. *BMC Bioinformatics*, **6**, 159.
42. Parsons,J., Holmes,J.B., Rojas,J.M., Tsai,J. and Strauss,C.E. (2005) Practical conversion from torsion space to Cartesian space for in silico protein synthesis. *J. Comput. Chem.*, **26**, 1063–1068.
43. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
44. Knight,R., Maxwell,P., Birmingham,A., Carnes,J., Caporaso,J.G., Easton,B.C., Eaton,M., Hamady,M., Lindsay,H., Liu,Z. *et al.* (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol.*, **8**, R171.
45. Murray,J.M. and Carr,A.M. (2008) Smc5/6: a link between DNA repair and unidirectional replication? *Nat. Rev. Mol. Cell Biol.*, **9**, 177–182.
46. Cozzetto,D., Giorgetti,A., Raimondo,D. and Tramontano,A. (2007) The Evaluation of Protein Structure Prediction Results. *Mol. Biotechnol.*, **39**, 1–8.
47. Yang,X.L., Otero,F.J., Ewalt,K.L., Liu,J., Swairjo,M.A., Kohrer,C., Rajbhandary,U.L., Skene,R.J., McRee,D.E. and Schimmel,P. (2006) Two conformations of a crystalline human tRNA synthetase-tRNA complex: implications for protein synthesis. *EMBO J.*, **25**, 2919–2929.
48. Parisien,M., Cruz,J.A., Westhof,E. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
49. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res.*, **36**, W70–W74.
50. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
51. Flores,S.C., Wan,Y., Russell,R. and Altman,R.B. (2010) Predicting RNA structure by multiple template homology modeling. *Pac. Symp. Biocomput.*, 216–227.
52. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
53. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
54. DeLano,W.L. (2002) The PyMOL molecular graphics system. *DeLano Scientific*. San Carlos, CA, USA.
55. Biegeleisen,K. (2006) The probable structure of the protamine-DNA complex. *J. Theor. Biol.*, **241**, 533–540.

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022

56. Das,R., Karanicolas,J. and Baker,D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.
57. Sharma,S., Ding,F. and Dokholyan,N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.
58. Jonikas,M.A., Radmer,R.J., Laederach,A., Das,R., Pearlman,S., Herschlag,D. and Altman,R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
59. Bindewald,E., Hayes,R., Yingling,Y.G., Kasprzak,W. and Shapiro,B.A. (2008) RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. *Nucleic Acids Res.*, **36**, D392–D397.
60. Klosterman,P.S., Tamura,M., Holbrook,S.R. and Brenner,S.E. (2002) SCOR: a Structural Classification of RNA database. *Nucleic Acids Res.*, **30**, 392–394.
61. Popenda,M., Blazewicz,M., Szachniuk,M. and Adamiak,R.W. (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res.*, **36**, D386–D391.
62. Godzik,A. (2003) Fold recognition methods. *Methods Biochem. Anal.*, **44**, 525–546.
63. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52.
64. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
65. Dowell,R.D. and Eddy,S.R. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, **7**, 400.
66. Smith,C., Heyne,S., Richter,A.S., Will,S. and Backofen,R. (2010) Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LosARNA. *Nucleic Acids Res.*, **38**, 373–377.
67. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.
68. Holmes,I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, **6**, 73.
69. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
70. Hinsen,K. (2000) The molecular modeling toolkit: a new approach to molecular simulations. *J. Comp. Chem.*, **21**, 79–85.
71. Johnston,M.A., Galvan,I.F. and Villa-Freixa,J. (2005) Framework-based design of a new all-purpose molecular simulation application: the Adun simulator. *J. Comput. Chem.*, **26**, 1647–1659.
72. Tan,R.K.Z., Petrov,A.S. and Harvey,S.C. (2006) YUP: A molecular simulation program for coarse-grained and multiscaled models. *J Chem. Theory Comput.*, **2**, 529–540.

Downloaded from https://academic.oup.com/nar/article/39/10/4007/1294922 by guest on 20 August 2022