Journal of Computer-Aided Molecular Design (2019) 33:497–507
https://doi.org/10.1007/s10822-019-00195-y

ToGo-WF: prediction of RNA tertiary structures and RNA–RNA/protein interactions using the KNIME workflow

Satoshi Yamasaki[1,2] · Takayuki Amemiya[1] · Yukimitsu Yabuki[1,3] · Katsuhisa Horimoto[1] · Kazuhiko Fukui[1]

Received: 1 November 2018 / Accepted: 28 February 2019 / Published online: 6 March 2019

Abstract

Recent progress in molecular biology has revealed that many non-coding RNAs regulate gene expression or catalyze biochemical reactions in tumors, viruses and several other diseases. The tertiary structure of RNA molecules and RNA–RNA/protein interaction sites are of increasing importance as potential targets for new medicines that treat a broad array of human diseases. Current RNA drugs are split into two groups: antisense RNA molecules and aptamers. In this report, we present a novel workflow to predict RNA tertiary structures and RNA–RNA/protein interactions using the KNIME environment, which enabled us to assemble a combination of RNA-related analytical tools and databases. In this study, three analytical workflows for comprehensive structural analysis of RNA are introduced: (1) prediction of the tertiary structure of RNA; (2) prediction of the structure of RNA–RNA complexes and analysis of their interactions; and (3) prediction of the structure of RNA–protein complexes and analysis of their interactions. In an RNA–protein case study, we modeled the tertiary structure of pegaptanib, an aptamer drug, and performed docking calculations of the pegaptanib-vascular endothelial growth factor complex using a fragment of the interaction site of the aptamer. We also present molecular dynamics simulations of the RNA–protein complex to evaluate the affinity of the complex by mutating bases at the interaction interface. The results provide valuable information for designing novel features of aptamer-protein complexes.

Keywords RNA · RNA–protein · Tertiary structure · Workflow · Aptamer · Nucleic acid drug

Introduction

Recent advances in micro-array and sequencing technologies have revealed that large amounts of non-coding RNAs exist in cells. For example, in human cells, tens of thousands of long non-coding RNAs have been identified [1]. The functions of non-coding RNAs have been studied extensively, and relationships between mutations in non-coding RNAs and diseases have been investigated [2–4]. These non-coding RNAs have various functions, including regulation of gene expression and catalysis of biochemical reactions inside cells. Protein–RNA and RNA–RNA interactions are key elements in such functional processes. The RNA-induced silencing complex (RISC), small nuclear ribonucleoproteins (snRNPs) and small nucleolar RNA–protein complexes (snoRNPs) are protein–RNA complexes that regulate gene expression. Additionally, RNA interference can be used readily as a mechanism to decompose target mRNAs [5–7]. Several types of non-coding RNAs are associated with serious diseases. In cancer cells, many types of micro-RNA (miRNA) are expressed and affect tumor progression or

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10822-019-00195-y) contains supplementary material, which is available to authorized users.

✉ Satoshi Yamasaki
s-ymsk@ims.u-tokyo.ac.jp
http://togo.medals.jp/active_local_rna_prediction.eng.html

✉ Kazuhiko Fukui
k-fukui@aist.go.jp
http://togo.medals.jp/active_local_rna_prediction.eng.html

[1] Molecular Profiling for Drug Discovery Research Center (molprof), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

[2] Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo (IMSUT), 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[3] IMSBIO Co., Ltd, 4-21-1-601 Higashi-Ikebukuro, Toshima-ku, Tokyo 170-0013, Japan

metastasis by interacting with various tumor factors [8–12]. Interaction between cellular human miRNA and viral RNA is important for infection or defense processes of pathogenic viruses, including HIV, influenza virus, SARS coronavirus and Ebola virus [13–17]. Moreover, there is a general acceptance that many kinds of structured non-coding RNAs exist in nature and play important roles in living cells; however, details of their functional mechanisms remain largely unknown [18–20].

Non-coding RNAs have therapeutic value as potential drugs and drug targets [21]. There is high expectation that nucleic acid medicines will alleviate or cure disease processes, because specifically designed nucleic acids (oligonucleotides) will directly target DNA, RNA, or proteins that cause diseases. The depletion of drug discovery targets in low molecular weight drug discovery programs, the problems associated with mass production of antibody drugs that show high specificity, and issues of high cost to sales ratios will possibly be solved by nucleic acid medicine. Consequently, the development of nucleic acid medicinal approaches is attracting significant attention. Currently, five nucleic acid-based drugs have been approved and are on the market worldwide. Several companies are considering commercialization based on industry-academia co-operations, and a dramatic increase in this market is anticipated in the near future [5, 22].

Nucleic acids generally play important roles in preserving genetic information and protein synthesis, and these biomolecules form various three-dimensional structures. In an effort to exploit their potential benefit to human health, it is important to carry out detailed analyses of protein–RNA or RNA–RNA interactions. Thus, solving the tertiary structures of RNA molecules, as seen with proteins, should aid nucleic acid drug development. Therefore, a large body of research investigating or predicting the tertiary structures of RNA has been performed [19, 20, 23–33]. Previously-developed software for predicting the secondary (RNAfold [34], Mfold [35], CentroidFold [36]) and tertiary structures (FARNA [23], MC-fold and MC-sym [24], RNA2D3D [27], SimRNA [32], SPQR [33]) of RNA molecules from their nucleotide sequences are available. A recent computational challenge is the prediction of tertiary structures of more difficult targets such as RNA–RNA and RNA–protein complexes. The current difficulties associated with using available tools and software are the requirement of several complex interactive manipulations and a user that has a strong understanding of RNA tertiary structures. To simplify the task of computational RNA analysis, we provide a novel workflow, which facilitates the study of RNA molecules from their sequence to their tertiary structure, and their interactions with various biomolecules. By using this workflow, it is possible to use various analysis tools, software and databases developed for RNA analytics in a single platform. For example, in the case of tertiary structure prediction, the input sequence selected from a database passes to secondary structure prediction software and the calculated secondary structure is used for tertiary structure calculations in a one-stop platform. The intermediate results of each computational step can be visualized.

Workflow systems with a graphical user interface are useful tools for addressing how to treat large-scale data and to efficiently integrate analytical resources [37]. We introduce the ToGo-WF workflow as a bioinformatic analytical resource, which is capable of analyzing, modeling and visualizing biological data by using the KNIME (Konstanz Information Miner) platform [38]. The KNIME workflow platform is freely available via http://www.knme.org. Over the last few years, we have released several bioinformatic nodes, mainly developed by bioinformatics groups at AIST, and have deposited them in the node repository using the KNIME environment. On this platform, users can combine independent analytical resources, called nodes in KNIME, from the repository of bioinformatics tools developed at AIST and external tools/databases by selecting them in an easy drag-and-drop manner. In this study, we have developed a workflow that combines RNA-related analytical nodes to predict RNA tertiary structures and RNA–RNA/protein interactions that provide insights and a starting point for tertiary structural analysis by RNA molecular simulations. Using the structure of an RNA–protein complex obtained from this workflow, we have performed molecular dynamics simulations of this RNA–protein complex to analyze the structural stability and affinity of the interaction.

Methods

KNIME workflow

We have created several novel RNA analytical nodes for predicting RNA tertiary structures and RNA–RNA/protein interactions. The KNIME environment, including the RNA nodes, can be downloaded from the ToGo-WF website (http://togo.medals.jp/active_local_rna_prediction.eng.html) for Microsoft Windows, Linux and Mac OSX operating systems. We have also posted installation instructions and user manuals for the RNA workflow. After installation, three workflow lines are released by selecting RNA structure prediction in the KNIME explorer, as shown in Fig. 1. Nodes in the workflow that were developed by AIST include Sparql, CentroidFold, IPknot, RNA2DChecker, Viewer, FragmentSelector, IntMinMM, RactIP, RASSIE and Rascal. Prediction of RNA secondary structures is performed using the programs CentroidFold [36], IPknot and RactIP [39], and tertiary structure prediction is achieved using the programs RASSIE [30] and Rascal [31]. The RNA2DChecker node

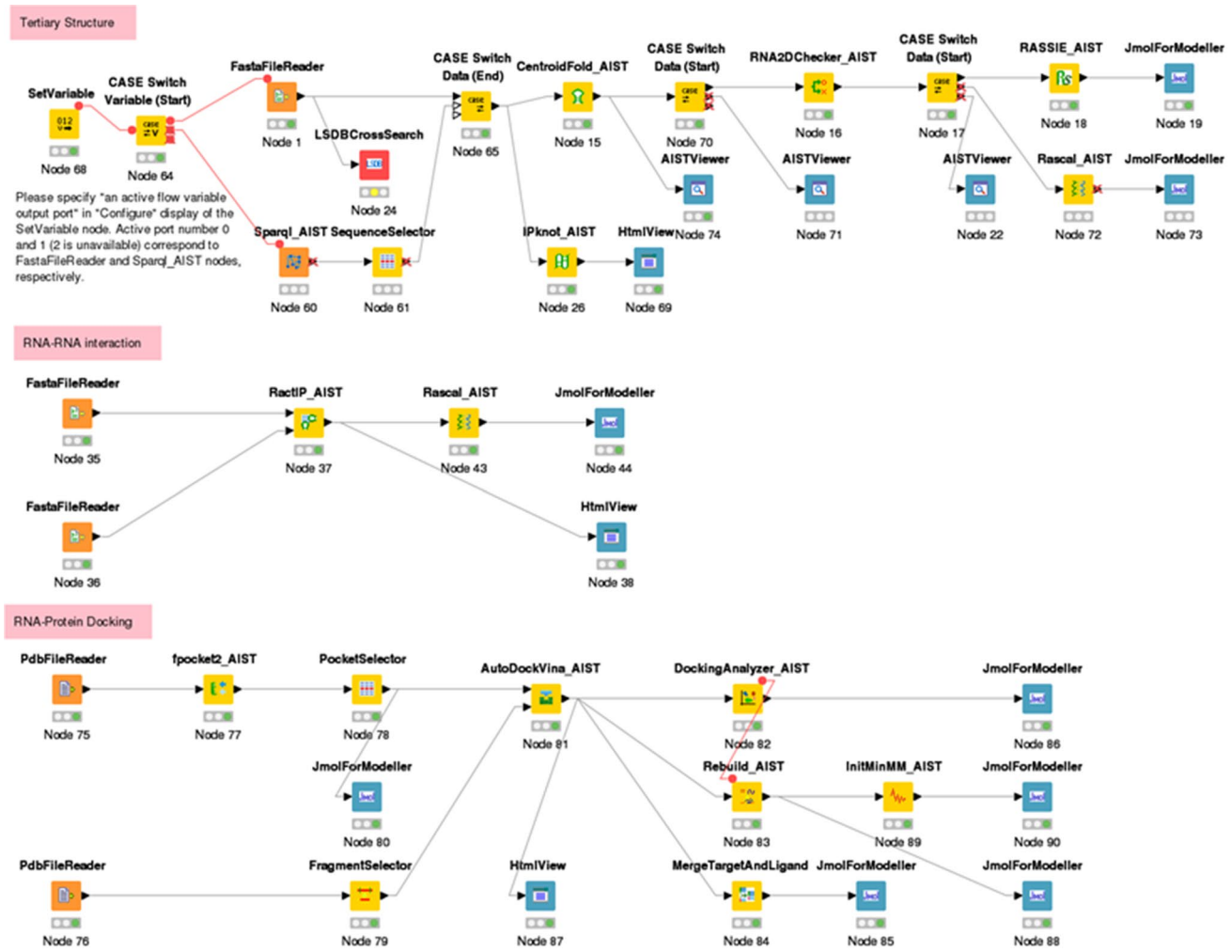

Fig. 1 KNIME workflow for tertiary structure RNA analysis. The top workflow shows the prediction of the RNA 2D/3D structure. The middle workflow shows the prediction of the RNA–RNA interaction. The bottom workflow presents the structure prediction of the RNA–protein complex

is used to judge the difficulty of the tertiary structure prediction based on the secondary structure prediction results and from this analysis either RASSIE or Rascal is selected. The details of RASSIE and Rascal based on fragment-based methods are described in references [30, 31]. In brief, prediction using RASSIE is performed by assembling small tertiary structural elements based on secondary structures (stem(duplex), hairpin-loop, internal-loop, single-strand-loop). Those small structural elements were derived from RNA structures found in the PDB. Targets with long loop structures that are not found in the PDB or targets with high complexity (e.g., targets with too many structural elements) are switched to the Rascal node by RNA2DChecker. Rascal is based on a fragment-assembly method, which uses dihedral angles of single strand three base fragments and randomly replaces them during 10–50 thousands of steps of Metropolis Monte Carlo simulation. Rascal can predict tertiary structures with any secondary structures because the single stranded three-base fragments derived from RNA structures found in the PDB are not based on their secondary structure like RASSIE. Rascal can handle single and double stranded structures as input with a secondary structure and joint secondary structures of RNA–RNA interactions. Rascal also has the feature of predicting tertiary structures by adjusting base-pair potentials without the secondary structure. RASSIE predicts relatively short chain RNAs with simple secondary structure elements in rapid computational time. Long-chain RNAs with complex secondary structures can be predicted by Rascal; however, more computational time is required to obtain good prediction results.

In the workflow for determining RNA–protein complexes, Fpockts2 [40] and AutoDockVina [41] are used as the analytical resources for RNA–protein docking. The node of the fragment selector for RNA can create the fragment structure

of the given RNA and the RNA fragment is used for docking calculations with the target protein whose active sites are specified by the Fpockts node. After docking, the RNA fragment–protein complex structures are obtained using the Merge Target and Ligand node. The original input RNA structure is reconstructed by the Rebuild node based on the docking position of the RNA fragment, and the DockingAnalyzer node is used to cluster the complex structures. By selecting one of the complex structures, the structure is energetically minimized by molecular mechanics, initMinMM.

In the KNIME environment, users can configure each node by clicking the right mouse button and can view a simple annotation in the node description by clicking the left mouse button. The nodes developed by AIST allow users to execute jobs on our AIST remote computer (not on the user's local computer). The nodes, RASSIE, Rascal, AutoDockVine and initMinMM demand high amounts of computing power and memory. These nodes are in our cloud-based workflow system and jobs with these nodes are submitted to the Message Passing Interface (MPI) parallel jobs system of the AIST cluster machine. The manuals on how to install and use ToGo-WF for RNA can be downloaded from our website (https://togo.medals.jp/active_local_rna_prediction.eng.html).

Database of RNA-based drugs

We constructed a database by searching the literature for nucleic acid-based drugs and web services. RNA-based drugs can be classified by the mechanism of activity, and include inhibitors of mRNA translation (antisense) and RNAs that bind proteins and other molecular ligands (aptamers) [42]. Information about the authorization status of nucleic acid drugs, primary structure (sequence information) and previous researchwas primarily collected from the EU Clinical Trials Register (https://www.clinicaltrialsregister.eu), U.S. Food and Drug Administration (https://www.fda.gov/) and Ionis Pharmaceuticals, Inc. (http://www.ionispharma.com/pipeline/). The developed database for medals Nucleic Acid Drug Database (mNADD) provides "ATC Classification", "Product Name", "Synonymous", "Phase", "Description" and "Target protein ID". In the workflow, the Sparql node (Fig. 1) is used to select a sequence from the mNADD database, which contains approved RNA drugs. The database also has short non-coding RNA entries less than 150 nucleotides obtained from fRNAdb [43], which is accessible at the National Bioscience Database Center (https://dbarchive.biosciencedbc.jp/en/frnadb/desc.html).

Molecular dynamics simulationsof RNA tertiary structures

Molecular dynamics (MD) simulations in explicit solvent water were performed to relax the structures derived from either the RASSIE or Rascal module of the workflow. All initial structures were placed at the center of the water box and water molecules were modeled using the TIP3P 3-point charge model potential [44]. K^+ and Cl^- ions were added using the LEaP module in the Amber package to neutralize the net charge of the system and to give the system an ionic concentration of 0.2 M. These systems were energy-minimized. All calculations were performed using Amber 16 [45] with the *protein.ff14SB* (for protein), *RNA.OL3* (for RNA) and *tip3p* (for solvent) force fields [46–49] with a time step of integration set to 1 fs. All bond lengths involving hydrogen atoms were constrained to their respective equilibrium values by the SHAKE method[50]. Electrostatic interactions were treated with the particle mesh Ewald method with a cut-off of 9 Å [51]. Each system was gradually heated to 300 K during the first 60 ps at a heating rate of 5 K/ps in the NVT ensemble. Following a 200 ps heating and equilibrating NVT simulation, for each system, a 6 ns dynamic calculation was performed in the NTP ensemble, with pressure constant at 1 atm and temperature constant at 300 K using the Berendsen's thermostat [52].

Structures of the RNA–protein complex

Molecular Mechanics-Generalized Born Surface Area (MM-GBSA) calculations were performed to evaluate the stability of the docked complex structure. Theoretical free energies were calculated using the MM-GBSA approach [53]. The MD simulations of docked structures were performed using the same method described above. Snapshot structures were sampled every 5 ps (from 0.5 to 1.5 ns of the MD trajectory) and used in estimating the binding free energies. In the analysis of the binding free energies, water molecules were replaced with the implicit solvent GB model.

This evaluation of free energies using the MM-GBSA approach was performed for complex structures, which were selected from using the Rebuild node to confirm that the molecules do not overlap following the rebuilding of the fragments after the docking calculation. The structure of the complex that gave the lower binding free energy with adequate distance was used as the template model for building and carrying out simulations of the RNA mutants.

Results

Workflow for tertiary structures of RNAs

The workflow, RNA Structure Prediction, at the top of Fig. 1 models the tertiary structure of the RNA target according to secondary structure information, which is based on the calculated results from CentroidFold [36]. The RNA sequence input is provided by the user or using the Sparql node, which

is connected to our mNADD database. Figure 2a, b show the predicted secondary and tertiary structures for RNA aptamer [54] calculated from the CentroidFold and RASSIE nodes. The input sequence was obtained from a human IgG-aptamer complex (PDB ID: 3AGV). In the workflow, RNA2DChecker automatically selects RASSIE or Rascal, depending on the complexity of the secondary structure. For the short sequence, RASSIE was selected and succeeded in constructing the tertiary structure based on the secondary structure. Figure 2c, d show the predicted secondary and tertiary structures for the non-coding RNA (FR000373 of fRNAdb) calculated by the CentroidFold and Rascal nodes. The input sequence was selected from mNADD using the Sparql node in the workflow. The tertiary structure of such a long hairpin-loop was predicted using Rascal. The input sequence and the predicted secondary structure in Fig. 2c were more complicated than those in Fig. 2a. Thus, RNA-2DChecker automatically selected Rascal for further tertiary structure calculations in the workflow.

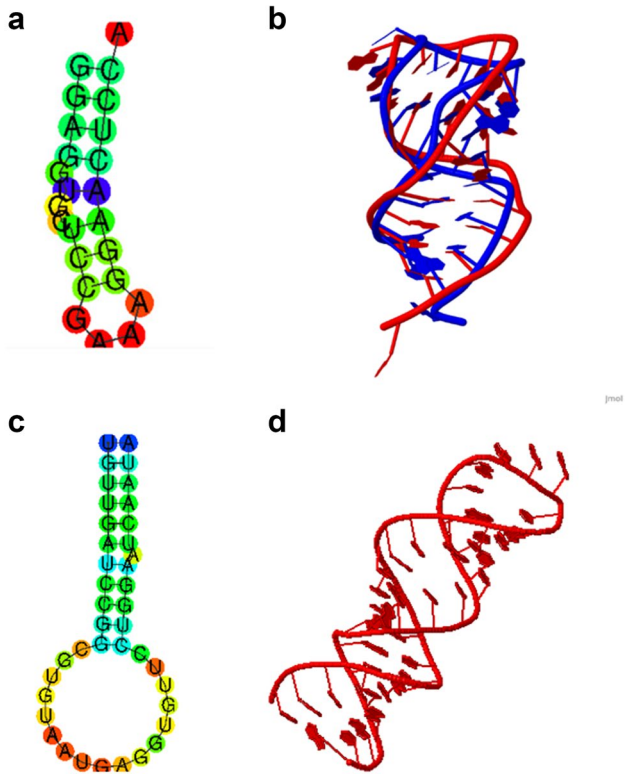

Fig. 2 Prediction of the 2D/3D structure. **a** The input sequence of the RNA aptamer for RASSIE is GGAGGUGCUCCGAAAGGAACU CCA. The secondary structure, ((((((...(((....)))))))), was calculated by the CentroidFold node in the workflow. Gamma was set to 2 ^ 3 using the configuration from CentroidFold. **b** One of the predicted tertiary structures derived with the RASSIE node. The predicted and native (PDB ID: 3AGV) RNA structures are shown in red and blue, respectively. The rmsd (root-mean-squared deviation) between the predicted and native structure (3AGV) is 3.25 Å. **c** The input sequence of the putative conserved noncoding region for Rascal is TGTTGATCCGGCGTGTAATGAGGTGTTCCTGGAATCAATA, which was selected from the mNADD database using the Sparql node (FR000373). The secondary structure, ((((((((((.....)))))).))))), was calculated by the CentroidFold node in the workflow. **d** One of the predicted tertiary structures derived with the Rascal node

To evaluate the performance and capabilities of ToGo-WF, the predicted structures obtained from the workflow were compared with secondary structures predicted by MC-fold, and tertiary structures predicted by MC-sym [24] and SimRNA [32]. The detailed comparison is given in Supplementary Material 1. The workflow predicted the RNA secondary and tertiary structures to a higher accuracy than MC-fold/MC-sym and simRNA. A reason for this higher accuracy is that the fragment-based method with high-resolution potential energy functions in RASSIE and Rascal enables highly accurate tertiary structure predictions. ToGo-WF uses a platform that aims to provide one-stop RNA analysis that can switch the workflow depending on the difficulty of the problem. Additionally, by linking the analysis workflow and database with one platform, large-scale analysis is possible.

The predicted RNA structures can be used easily as initial structures for MD simulations to obtain detailed tertiary structure analysis of conformational changes and the stability of potential chemical modifications. The calculations introduced here for prediction of tertiary structures of RNAs are described in Sect. 5 of the online manual (https://togo.medals.jp/download/RNA_UserManual_E_v1.0.pdf).

Workflow for structures of RNA–RNA complexes

RNA–RNA interactions are important for many biological functions and considered to be potential targets of new medicines. The kissing hairpin loop of the dimerization initiation site (DIS) of HIV-1 is a typical example of such interactions. In the lifecycle of HIV-1, the dimerization of genomic RNA is crucial for viral replication [26]. Formation of a kissing loop hairpin structure in the DIS region is the primary event of dimerization. Therefore, inhibiting formation of the kissing loop is a potential target for treating HIV. The workflow, RNA–RNA Structure Prediction, in the middle of Fig. 1 predicts the tertiary structure of RNA–RNA interacting complexes with Rascal using the results of RactIP, which provides RNA two-dimensional complex structures for RNA–RNA interactions with kissing hairpins. We have previously predicted the tertiary structures of many kissing loops and duplex targets successfully from nucleotide sequences by using the RactIP-Rascal pipeline [31].

Recently, patisiran (onpattro), which is the first siRNA therapeutic agent, was approved by the FDA for treating the rare disease hereditary transthyretin amyloidosis[55]. Using the workflow, we have modeled the tertiary structure of the double-stranded siRNA, patisiran (Fig. 3). This workflow can model tertiary structures of RNA–RNA interactions

and double-stranded RNA structures. The obtained tertiary structures can be used for further molecular simulations to compute the dimerization process from the kissing structure and the interaction of RNA–protein complexes. The results may provide new suggestions or data that cannot be explained without an understanding of the tertiary structures of the complexes.

Database of RNA-based drugs

By collecting information about the authorization status of nucleic acid drugs, primary structures and observations from previous research, we found 35 drugs that have gone through clinical trials. They are classified into two major types of nucleic acid-based drugs: one type targets nucleic acids and represents antisense drugs, whereas the other type targets proteins and are termed aptamers. Five of the 35 drugs tested have been approved. Eteplirsen (EXONDYS 51), Vitravene (Fomivirsen), Kynamro (Mipomersen) and Nusinersen (SPINRAZA) are antisense drugs that bind exon51 of dystrophin pre-messenger ribonucleic acid (pre-mRNA), target mRNA of cytomegalovirus, degrade ApoB-100 mRNA expressed in the liver and target survival motor neuron 2 (SMN2) pre-mRNA to inhibit factor suppressing splicing of exon 7 from binding to pre-mRNA, respectively. The other, pegaptanib (Macugen), is an aptamer that binds to vascular endothelial growth factor (VEGF). Pegaptanib is the first approved aptamer of a new therapy to slow vision loss in people with the eye disease neovascular (wet) age-related macular degeneration (AMD)[56–59].

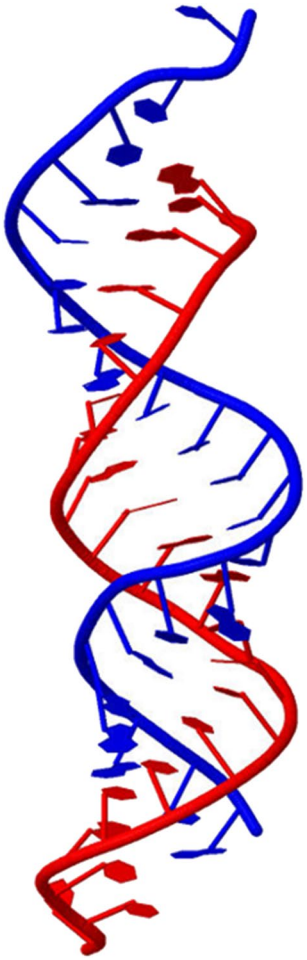

Fig. 3 Prediction of the complex structure of patisiran (Onpattro). The input sequences of RNA1 and RNA2 are GUAACCAAGAGU AUUCCAU (red) and AUGGAAUACUCUUGGUUACUU (blue), respectively. The predicted tertiary structure is derived with the Rascal node using the result obtained with the RactIP node

Antisense drugs represent the majority of nucleic acids in clinical trials with most targeting cancer and viral infections. Consequently, there is increasing interest in antisense technology. Ions Pharma is currently conducting ongoing clinical trials of other antisense drugs for the treatment of common medical conditions, including rheumatoid arthritis, cancers and Crohn's disease, a serious intestinal illness. Figure 4 shows major adaptation diseases that are current targets of nucleic acid medicines. We summarize the information of nucleic acid drugs in terms of "Approval Status of Nucleic Acid Drugs", "Target Proteins and Number of its Nucleic Acid drugs" and "Number of Nucleic Acid drugs participating in Pathway", which are provided at the website (https://medals.jp/enucleicacid.html). The database is readily accessible to the corresponding server's URL (https://medals.jp/edrugnainfo.html) and can be downloaded from https://dbarchive.biosciencedbc.jp/en/nucleic-acid-drug/download.html.

In the workflow, RNA based drugs and non-coding RNAs less than 150 nucleotides obtained from fRNAdb are stored in the RDF-formatted database. Users can access the database with the Sparql node and use the environment in which database and analysis tools are connected seamlessly in the workflow.

Workflow for structures of RNA–protein complexes

The RNA–protein workflow is presented at the bottom of Fig. 1 and is used to predict the structure of a nucleic acid drug–target protein complex by molecular simulations. We applied our workflow to the complex formed between the antagonistic aptamer pegaptanib and VEGF. Figure 5 shows the secondary structure of pegaptanib, (((....(((....))).....))). When we used the workflow for simple prediction of the secondary structure (Fig. 1, top workflow of the CentroidFold with default parameters), the predicted structure is ((((..... .....)))). The secondary structure predicted is very different from the experimentally obtained secondary structure in Fig. 5. It is difficult to estimate the tertiary structure using an inaccurate secondary structure as input. Thus, in this

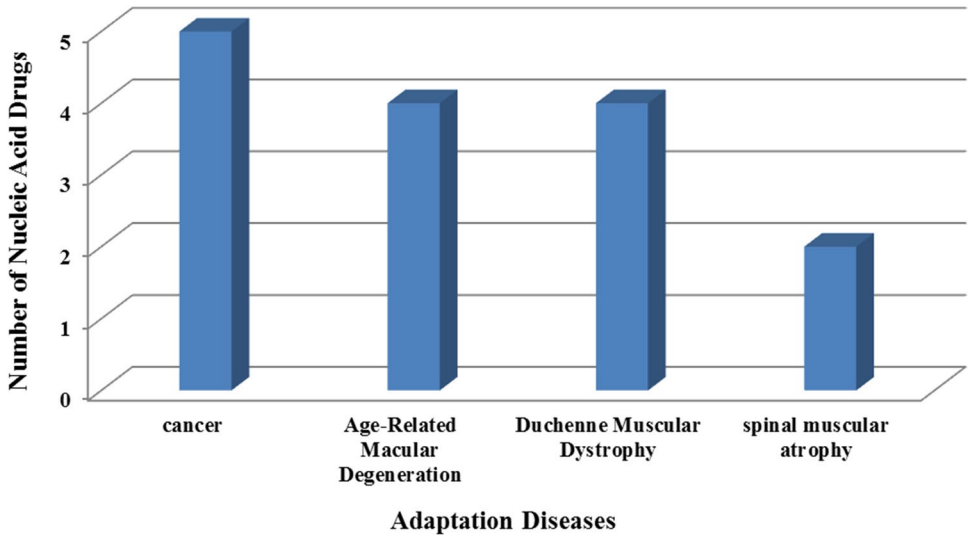

Fig. 4 The number of nucleic acid drugs represented by antisense, small interfering RNAs (siRNAs) and aptamers that target adaptive diseases

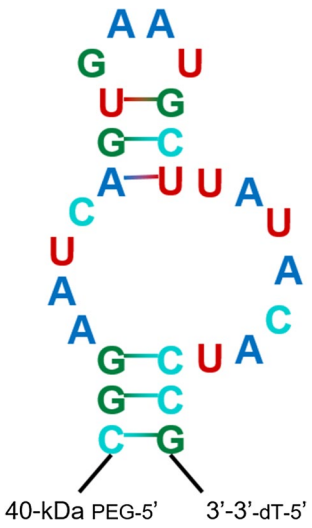

Fig. 5 Secondary structure of pegaptanib, ((((....((((....)))).....))),which is an antagonistic aptamer that targets VEGF

case, the tertiary structure is derived using Rascal, which can model structures with internal loops without the need of a secondary structure as a starting point. In obtaining the candidate tertiary structures by Rascal, the potential function was set as a penalty if base pairs were formed between A4–C7 and U18–U24. A matrix file of dimensions (number of bases × number of bases) defining the weight of base-pair potentials is used as input together with the sequence. A detail explanation of how to use the workflow to obtain ten candidate structures of pegaptanib is given in Supplementary Material 2.

From the ten candidate structures derived from the workflow for tertiary structure determination of RNA by the Rascal module, two candidate structures shown in Figure S2 (A) and (B) have open internal-loop structures for A4–C7 and U18–U24 (Fig. 5). NMR experiments have shown[57] that U14 interacts with the side chain of Cys27 of VEGF. U14 faces outward in the two selected structures. To relax these structures in a water environment, 6 ns MD simulations of each structure immersed in TIP3P water + KCl (0.2 M) were performed.One of the RNA structures that has the U14 base facing outward was selected as the tertiary structure in the RNA–protein docking process.

The relaxed RNA structure and protein structures in PDB format were given as input for the complex structure workflow analysis. Using the FragmentSelector node in the workflow, a three-base fragment A13-U14-G15 from the structure was extracted for docking. Coordinates of residues in the protein that are in the vicinity of the three-base fragment can be easily specified using the Fpocket 2 node. The coordinates near Cys27 in the protein were used as input, and then docking calculations of the pegaptanib fragment and VEGF were performed. The docking result is given in Supplementary Material 3.The structures of the complex were rebuilt based on the docking poses of the RNA fragment. By confirming that the protein and RNA structures do not sterically clash with each other in the process of rebuilding the full-length RNA, we selected three structures of the complex that are reasonable for carrying out MD calculations. Figure 6 shows the selected structures of the complex after the docking and rebuilding processes.

MD simulations of RNA–protein complexes

MD simulations and MM-GBSA calculations were performed after energy minimization with the initMinMM node to evaluate the stability and binding free energy of the docked structures. For the three docked structures (Fig. 6), 1.5 ns MD simulations in TIP3P water + KCl (0.2 M) were performed, and snapshot structures every

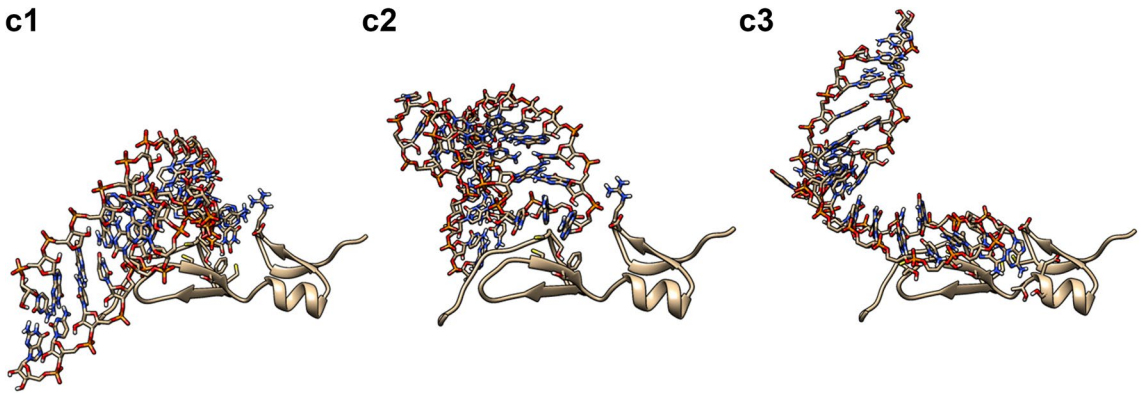

Fig. 6 Structures of the pegaptanib–VEGF complex after rebuilding the full-length RNA

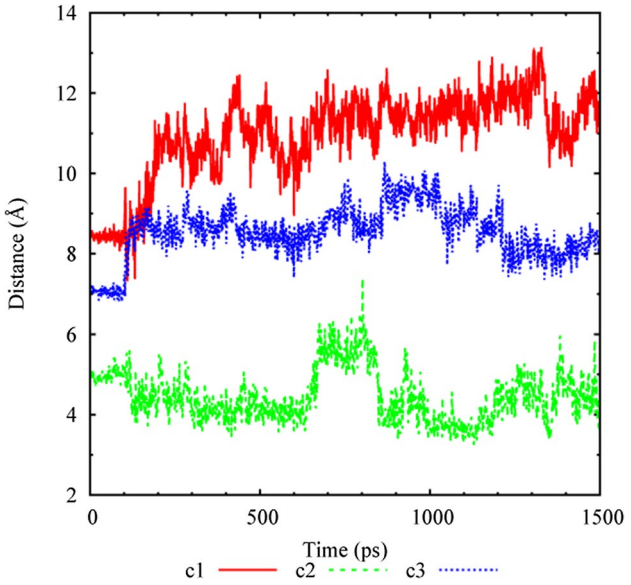

Fig. 7 Changes in the distance between the centroid of base heavy atoms of U14 and the centroid of the Cys10-Cys27 disulfide bond during 1500 ps MD simulations for the three RNA–protein complexes. Complexes c1 (red), c2 (green color) and c3 (blue) are shown. The initial three structures are those presented in Fig. 6

5 ps were taken during the 0.5–1.5 ns simulation period and used for MM-GBSA calculations. The distance between the centroid of base heavy atoms of U14 and the centroid of the Cys10-Cys27 disulfide bond was also assessed to evaluate the protein–RNA distance (Fig. 7). Previous experiments reported that Cys27 displayed the largest chemical shift change upon aptamer binding in NMR spectra and the mechanism for the formation of a photo-crosslink between U14 and the disulfide bond, Cys27:S-Cys10:S [56]. Among the c1–c3 structures, the simulation of c2 showed the closest distance between the aptamer and protein. Table 1 shows the binding free energy, $\Delta\Delta G$, derived from the MM-GBSA calculations

Table 1 Binding free energy, $\Delta\Delta G$, for the structures of the complexes, and the distance between the centroid of base heavy atoms of U14 and Cys10-Cys27

Initial coordinate	MMGBSA (kcal/mol)		Distance (Å)	
	ΔΔG	Std. dev	Average	Std. dev
c1	−6.99	5.34	11.43	0.62
c2	−28.61	5.86	4.45	0.71
c3	−35.42	4.20	8.65	0.57

for the three complexes. Although the calculations showed that the c3 structure had the lowest $\Delta\Delta G$, indicating the highest affinity, the distance is 8 Å. This distance is too large to reconcile the experimental observation of a photo-cross link. Thus, the c2 structure is considered to be consistent with the experimental results and the probable complex between the aptamer pegaptanib and VEGF. Furthermore, the structure was used for simulations of the mutants, which investigated the stability and affinity of the complex. Three simulations were performed, c2_A, c2_G and c2_C, where U14 in the c2 structure was mutated to A, G and C nucleotides, respectively. The simulations were performed using the same conditions as those used for the native sequence. Table 2 shows the results of these simulations. The native c2 structure showed the shortest distance and the c2_G had the highest binding affinity. This indicates that the native and the mutated c2_G structures form the probable complex, and these c2 structures support the experimentally obtained results.

In aptamers that recognize proteins by a three-dimensional structure, there are cases where the base moiety is modified to gain diversity of the three-dimensional structure or enhance affinity with the protein. Development of an artificial base pair for aptamers has also been attempted [60]. Our workflows that integrate RNA analytical tools can easily provide complex structures that may provide new

Table 2 Binding free energy, $\Delta\Delta G$, for the structures of the mutated RNA and protein complexes, and the distance between the centroid of base heavy atoms of U14 and Cys10-Cys27

ID	MMGBSA (kcal/mol)		Distance (Å)	
	ΔΔG	Std. dev	Average	Std. dev
c2	−28.61	5.86	4.45	0.71
c2_A	−14.59	3.83	5.93	0.68
c2_C	−20.18	3.72	5.08	0.58
c2_G	−33.09	5.84	4.67	0.45

suggestions for the design of aptamers by interpreting the recognition mechanism of RNA–protein complexes.

Conclusions

The current cloud-based workflow on our website integrates analytical resources for RNA tertiary structure determination into KNIME workflow systems. We have developed three analytical workflows for comprehensive structural analysis of RNA: (1) prediction of the tertiary structures of RNA molecules; (2) prediction of the structures of RNA–RNA complexes, including analysis of the interface; and (3) prediction of the structures of RNA–protein complexes, including analysis of the interface. Taking advantage of the graphical tools offered by KNIME communities, users visually interpret the structural results, reproduce the calculated data and generate data suitable for fast and reliable preparation of MD simulations. We have also introduced database-linking tools to offer seamless connection between tools and the RNA database in the workflow.

Using the workflow, we have evaluated the RNA–protein complex formed between the RNA aptamer pegaptanib and VEGF. Molecular simulations were used to calculate changes in the affinity of complexes where key bases in the aptamer involved in complex formation were mutated. The results provide valuable information and insights that could aid the design of RNA aptamer-protein complexes with enhanced affinity, which affords an efficacy at a lower concentration and circumvents toxicity issues. Our workflows for the comprehensive structural analysis of RNA based on KNIME offer a powerful tool for structural analysis and design of nucleic acid drugs, and these structural analyses can be directly inputted into MD simulations.

Acknowledgements K.F. thanks Mr. Hiroshi Kouno for searching the literature of nucleic acid-based drugs to construct the database and docking simulations. This research was partially supported by the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP17am0101001. The workflow was initially developed as a part of the Life-Science Database Integration Project: Core Technology Development Program at the Japan Science and Technology Agency (JST).

References

1. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigo R, Johnson R (2018) Towards a complete map of the human long noncoding RNA transcriptome. Nat Rev Genet 19:535–548
2. Cheetham SW, Gruhl F, Mattick JS, Dinger ME (2013) Long noncoding RNAs and the genetics of cancer. Br J Cancer 108:2419–2425
3. Cooper TA, Wan L, Dreyfuss G (2009) RNA and disease. Cell 136:777–793
4. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M (2016) Role of non-coding sequence variants in cancer. Nat Rev Genet 17:93–108
5. Aagaard L, Rossi JJ (2007) RNAi therapeutics: Principles, prospects and challenges. Adv Drug Deliver Rev 59:75–86
6. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806–811
7. Hannon GJ (2002) RNA interference. Nature 418:244–251
8. Ferrarelli LK (2015) Focus issue: noncoding RNAs in cancer. Sci Signal 8:8–10
9. Huang Q, Gumireddy K, Schrier M, Le Sage C, Nage lR, Nair S, Egan Da, Li A, Huang G, Klein-Szanto AJ, Gimotty P, Katsaros D, Coukos G, Zhang L (2008) The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. Nat Cell Biol 10:202–210
10. Lu J, Getz G, Miska Ea, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando Aa, Downing JR, Jacks T, Horvitz HR, Golub TR (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838
11. Png KJ, Halberg N, Yoshida M, Tavazoie SF (2012) A microRNA regulon that mediates endothelial recruitment and metastasis by cancer cells - with comments. Nature 481:190–194
12. Zhu S, Wu H, Wu F, Nie D, Sheng S, Mo Y-Y (2008) MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. Cell Res 18:350–359
13. Liang HW, Zhou Z, Zhang SY, Zen K, Chen X, Zhang CY (2014) Identification of Ebola virus microRNAs and their putative pathological function. Sci China Life Sci 57:973–981
14. Mallick B, Ghosh Z, Chakrabarti J (2009) MicroRNome analysis unravels the molecular basis of SARS infection in bronchoalveolar stem cells. PLoS ONE 4:e7837
15. Pfeffer S, Voinnet O (2006) Viruses, microRNAs and cancer. Oncogene 25:6211–6219
16. Song L, Liu H, Gao S, Jiang W, Huang W (2010) Cellular microRNAs inhibit replication of the H1N1 influenza A virus in infected cells. J Virol 84:8849–8860
17. Triboulet R, Mari B, Lin Y-L, Chable-Bessia C, Bennasser Y, Lebrigand K, Cardinaud B, Maurin T, Barbry P, Baillat V, Reynes J, Corbeau P, Jeang K-T, Benkirane M (2007) Suppression of microRNA-silencing pathway by HIV-1 during virus replication. Science 315:1579–1582
18. Westhof E (2010) The amazing world of bacterial structured RNAs. Genome Biol 11:108
19. Behrouzi R, Roh JH, Kilburn D, Briber RM, Woodson SA (2012) Cooperative tertiary interaction network guides RNA folding. Cell 149:348–357
20. De Leonardis E, Lutz B, Ratz S, Cocco S, Monasson R, Schug A, Weigt M (2015) Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. Nucleic Acids Res 43:10444–10455

21. Ling H (2016) Non-coding RNAs: therapeutic strategies and delivery systems. Adv Exp Med Biol 937:229–237
22. Hayes J, Peruzzi PP, Lawler S (2014) MicroRNAs in cancer: biomarkers, functions and therapy. Trends Mol Med 20:460–469
23. Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci USA 104:14664–14669
24. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452:51–55
25. Cruz JA, Blanchet M-F, Boniecki M, Bujnicki JM, Chen S-J, Cao S, Das R, Ding F, Dokholyan NV, Flores SC, Huang L, Lavender Ca, Lisi V, Major F, Mikolajczak K, Patel DJ, Philips A, Puton T, Santalucia J, Sijenyi F, Hermann T, Rother K, Rother M, Serganov A, Skorupski M, Soltysinski T, Sripakdeevong P, Tuszynska I, Weeks KM, Waldsich C, Wildauer M, Leontis NB, Westhof E (2012) RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. RNA (New York, NY) 18:610–625
26. Ennifar E, Dumas P (2006) Polymorphism of bulged-out residues in HIV-1 RNA DIS kissing complex and structure comparison with solution studies. J Mol Biol 356:771–782
27. Martinez HM, Maizel JV, Shapiro B (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. J Biomol Struct Dynam 25:669–683
28. Reinharz V, Major F, Waldisphl J (2012) Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. Bioinformatics 28:i207–i214
29. Shapiro BA, Yingling YG, Kasprzak W, Bindewald E (2007) Bridging the gap in RNA structure prediction. Curr Opin Struct Biol 17:157–165
30. Yamasaki S, Nakamura S, Fukui K (2012) Prospects for tertiary structure prediction of RNA based on secondary structure information. J Chem Inf Model 52:557–567
31. Yamasaki S, Hirokawa T, Asai K, Fukui K (2014) Tertiary structure prediction of RNA-RNA complexes using a secondary structure and fragment-based method. J Chem Inf Model 54:672–682
32. Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, Rother KM, Bujnicki JM (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Res 44:e63
33. Poblete S, Bottaro S, Bussi G (2018) A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs. Nucleic Acids Res 46:1674–1683
34. Gruber AR, Bernhart SH, Lorenz R (2015) The ViennaRNA web services. Methods Mol Biol 1269:307–326
35. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415
36. Sato K, Hamada M, Asai K, Mituyama T (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. Nucleic Acids Res 37:W277–W280
37. Warr WA (2012) Scientific workflow systems: pipeline pilot and KNIME. J Comput Aid Mol Des 26:801–804
38. Berthold MR, Cebron N, Dill F, Gabriel TR, Ktter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: The Konstanz Information Miner. Data Anal Mach Learn Appl:319–326
39. Kato Y, Sato K, Hamada M, Watanabe Y, Asai K, Akutsu T (2010) RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. Bioinformatics 26:i460–i466
40. Schmidtke P, Bidon-Chanal A, Luque FJ, Barril X (2011) MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics 27:3276–3285
41. Trott O, Olson AJ (2010) Software news and update autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461
42. Burnett JC, Rossi JJ (2012) RNA-based therapeutics: current progress and future prospects. Chem Biol 19:60–71
43. Mituyama T, Yamada K, Hattori E, Okida H, Ono Y, Terai G, Yoshizawa A, Komori T, Asai K (2009) The functional RNA database 3.0: databases to support mining and annotation of functional RNAs. Nucleic Acids Res 37:D89–D92
44. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of Simple potential functions for simulating liquid water. J Chem Phys 79:926–935
45. Case DABR, Cerutti DS, Cheatham TE, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz AW, Homeyer N, Izadi S, Janowski P, Kaus J, Kovalenko A, Lee TS, LeGrand S, Li P, Lin C, Luchko T, Luo R, Madej B, Mermelstein D, Merz KM, Monard G, Nguyen H, Nguyen HT, Omelyan I, Onufriev A, Roe DR, Roitberg A, Sagui C, Simmerling CL, Botello-Smith WM, Swails J, Walker RC, Wang J, Wolf RM, Wu X, Xiao L, Kollman PA (2016) AMBER 2016. University of California, San Francisco
46. Zgarbova M, Otyepka M, Sponer J, Mladek A, Banas P, Cheatham TE III, Jurecka P (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. J Chem Theory Comput 7:2886–2902
47. Sponer J, Bussi G, Krepl M, Banas P, Bottaro S, Cunha RA, Gil-Ley A, Pinamonti G, Poblete S, Jurecka P, Walter NG, Otyepka M (2018) RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. Chem Rev 118:4177–4338
48. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput 11:3696–3713
49. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M (2007) Refinenement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers. Biophys J 92:3817–3829
50. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341
51. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an N \textperiodcentered log(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
52. Berendsen HJC, Postma JPMa (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
53. Bashford D, Case DA (2000) Generalized born models of macromolecular solvation effects. Annu Rev Phys Chem 51:129–152
54. Nomura Y, Sugiyama S, Sakamoto T, Miyakawa S, Adachi H, Takano K, Murakami S, Inoue T, Mori Y, Nakamura Y, Matsumura H (2010) Conformational plasticity of RNA for target recognition as revealed by the 2.15 A crystal structure of a human IgG-aptamer complex. Nucleic Acids Res 38:7822–7829
55. Adams D, Gonzalez-Duarte A, O'Riordan WD, Yang CC, Ueda M, Kristen AV, Tournev I, Schmidt HH, Coelho T, Berk JL, Lin KP, Vita G, Attarian S, Plante-Bordeneuve V, Mezei MM, Campistol JM, Buades J, Brannagan TH, Kim BJ, Oh J, Parman Y, Sekijima Y, Hawkins PN, Solomon SD, Polydefkis M, Dyck PJ, Gandhi PJ, Goyal S, Chen J, Strahs AL, Nochur SV, Sweetser MT, Garg PP, Vaishnaw AK, Gollob JA, Suhr OB (2018) Patisiran, an RNAi therapeutic, for hereditary transthyretin amyloidosis. New Engl J Med 379:11–21
56. Lee JH, Canny MD, De Erkenez A, Krilleke D, Ng YS, Shima DT, Pardi A, Jucker F (2005) A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of VEGF(165). Proc Natl Acad Sci USA 102:18902–18907
57. Ng EWM, Shima DT, Calias P, Cunningham ET, Guyer DR, Adamis AP (2006) Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease. Nat Rev Drug Discov 5:123–132

58. Lee JH, Jucker F, Pardi A (2008) Imino proton exchange rates imply an induced-fit binding mechanism for the VEGF(165)-targeting aptamer, Macugen. Febs Lett 582:1835–1839
59. Ni X, Castanares M, Mukherjee A, Lupold SE (2011) Nucleic acid aptamers: clinical applications and promising new horizons. Curr Med Chem 18:4206–4214
60. Hirao I, Kimoto M (2012) Unnatural base pair systems toward the expansion of the genetic alphabet in the central dogma. P Jpn Acad B 88:345–367