Article title: Software for predicting the 3D structure of RNA molecules.

First author: David Dufour
Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG) and Gene Regulacion, Stem Cells and Cancer Program, Centre de Regulació Genòmica (CRG), Barcelona, Spain.

Second author: Marc A. Marti-Renom
Institució Catalana de Recerca i Estudis Avançats (ICREA), Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG) and Gene Regulacion, Stem Cells and Cancer Program, Centre de Regulació Genòmica (CRG), Barcelona, Spain.
mmarti@pcb.ub.cat

Abstract

RNA is not regarded anymore as a simple transfer molecule between DNA and proteins. Indeed, over the last decades a plethora of new functional roles have been assigned to RNA molecules. Such functions are carried-out either by RNA molecules alone or through interactions with DNA, other RNA molecules, or proteins. In all cases, the structure that the RNA molecule adopts will impact its function, as it happens with proteins. Therefore, to fully characterize the function of an RNA molecule, its structure needs to be either determined by experiments or predicted by computation. Unfortunately, our knowledge of the atomic mechanism by which RNA molecules adopt their biological active structures is still limited. Such hurdle is now being addressed by the development of new computational methods for RNA structure prediction, which complement experimental methods such as X-Ray crystallography, Nuclear Magnetic Resonance, Small-angle X-ray scattering or Cryo-Electron Microscopy. This software focus is not dedicated to a single computational method but aims at outlining the most used methods for computational RNA structure prediction.

Introduction

The central dogma of molecular biology stated the RNA role as a mere information transfer molecule between DNA and proteins. However, soon other RNA molecules were discovered proving to be essential pieces of the translation machinery (*eg*. tRNA and rRNA molecules) or enzymes (ribozymes). For example, the 23S rRNA catalyses the transpeptydil reaction, while surrounding proteins offered structural support, not the reaction mechanism[1]. More recently, the discovery of other non-codding RNAs with regulatory functions has further challenged this "simple" information transfer function for RNAs[2]. For example, small nuclear RNAs (snRNAs) complex with proteins to form small nuclear ribonucleoproteins (snRNPs), which play fundamental roles in RNA splicing. Small nucleolar RNAs (snoRNAs) are involved in rRNA maturation through methylation and pseudouridylation. Micro RNAs (miRNAs), which are 22 to 26 nucleotides long, regulate gene expression through imperfect annealing to the target mRNA. Small interfering RNAs (siRNAs), which are 20 to 25 nucleotide long double-stranded RNAs, modulate gene expression through transcription inhibition of an RNA molecule complementary to the siRNA. They also participate in the RNA interference (RNAi) pathways with a range of effects from defence against dsRNA viruses or chromatin remodelling[3]. Finally, the recently-expanding list of long non-coding RNA (lncRNAs) is increasing the repertoire of functions that RNA molecules can have as well as their relation to disease[4]. Given the variety of regulatory functions that RNA can perform, there is an increased interest in using RNA molecules as targets for drug discovery or antisense therapy[5].

Many of these RNAs have a precise three-dimensional (3D) structure that allows them to perform their functions. RNA is a versatile molecule that can adopt different conformations, which allows its adaptation to different ligands, temperatures or conditions. If RNA is to be used as a therapeutic target, its 3D structure must be characterized. There are four main biophysical methods used to determine macromolecular structures: X-ray crystallography, NMR, SAXS and Cryo-EM. They respond to different needs, as the two first methods provide atomic resolution structures of a limited molecular size, while SAXS or Cryo-EM can provide low-resolution structures for macromolecular complexes. Unfortunately, such experimental methods are not always applicable or can be too expensive and time-consuming. Thus, computational approaches are becoming widely adopted[6]. This software focus outlines a limited number of the widely used computational methods for predicting the 3D structure of RNAs. An exhaustive list of methods and databases for RNA structure analysis and prediction can be found here: http://marciuslab.org/www/software/?rna_resources.

COMPUTATIONAL RNA STRUCTURE PREDICTION

The increase over the last decade of the number of available structures deposited in the PDB, including X-ray and NMR models (Figure 1), has stimulated the structural biology community to develop computational tools for analysing the RNA structural space and predicting the structure of known RNA sequences. The field of RNA modelling has been greatly influenced (and benefited) by many decades of development in the field of protein structure prediction. Although both kinds of molecules are in nature very different, many of the successful modern methods for RNA structure prediction use now a days "protein-like" approaches[7]. For example, and similarly to proteins, RNA 3D modelling can be improved by predicting its secondary structure[8]. This is because RNA folding is a hierarchical process where the tertiary structure is mainly determined by the secondary structure, which is in turn determined by the nucleotides in the RNA primary sequence. It is also important to note that RNA molecules act in many cases in conjunction with proteins. Therefore, the development of computational methods for predicting RNA-protein interactions is gaining rapid interest in the recent years. These methods, which are based on the RNA/protein sequences/structures or on RNA-protein docking, predict which proteins bind to RNA and how they bind[9, 10]. Similarly to RNA structure prediction, RNA-protein interaction methods benefit from the knowledge of the RNA secondary structure, which can now be obtained at genomic level[11].

Overall, the existing methods for RNA structure prediction can be divided into three main approaches, (i) *ab-initio* modelling where the structure of the query RNA sequence is predicted based on first principles (physics), (ii) comparative modelling where the query RNA structure is predicted based on homology to a known structure, and (iii) knowledge-based modelling, which uses statistical potentials or machine-learning methods for predicting RNA structures. Besides to these three categories, another orthogonal classification could be established in two categories, depending if the method uses spatial restrains or not for modelling. Spatial restraints are a set of conditions to be met by the model, so actually they restrain the number of possible valid models. These restraints can be extracted from a template or from experiments. Next, we outline some of the existing methods for each of the three mentioned categories for RNA structure analysis and prediction.

Ab-initio modelling

The iFoldRNA program

The iFoldRNA program (http://troll.med.unc.edu/ifoldrna)[12] uses multi-scale molecular dynamic simulations to obtain coarse-grained structural models of RNA. As an input, it requires only an RNA sequence, so there is no need for a previous knowledge of the secondary structure. It uses the DMD engine[13] and the Medusa force field[14] to simulate RNA folding mechanics. Three beads representing the phosphate atom and the sugar and aromatic bases of the nucleotide represent the coarse-grained models. iFoldRNA samples the conformational space of RNA at different temperatures by performing several coarse-grained simulations. Representative structures are then selected from the coarse-grained simulations based on energies and/or additional filters such as the radius of gyration or other experimentally known parameters. iFoldRNA can rapidly predict structures for RNAs smaller than 50 nucleotides at atomic resolution (2 to 5 Å RMSD to its native structure). Its back draw is that tertiary long-range contacts are predicted with less accuracy as the size of the RNA molecule to predict increases. This method has been improved by incorporating SHAPE-derived secondary and tertiary restraints[15]. However, such data is not necessary for iFoldRNA to predict a structure. Finally, iFoldRNA provides additional information to the end used such as specific heat, contact maps, simulation trajectories, radius of gyration, RMSDs from native state and fraction of native-like contacts.

Comparative modelling

The ModeRNA program

The ModeRNA program (http://genesilico.pl/moderna)[16], similarly to all comparative modelling programs, requires a known three-dimensional RNA structure as a template and a sequence alignment between the target sequence and the template. With such input, ModeRNA will build a 3D structure of the template by copying the coordinates of invariant residues in the alignment, refining substitutions and base modifications, and modelling insertions and deletions (indels) using structural fragments from a database of known structures. Specifically, the coordinates of all atoms of the identical positions between the template and the target are copied from the template residue to the model, while substitutions are superimposed onto three atoms of the template base adjacent to the glycosidic bond to allow backbone continuity. ModeRNA allows the insertion of fragments up to 17 residues long. To do so, it uses a library of 131,316 known fragments of 2 to 19 residues long, called RNADB2005[17], derived from a set of 172 non-redundant RNA structures from different families. A larger library covering fragments up to 100nt long derived from the same database was created for modelling longer inserts[16]. For each indel, ModeRNA identifies a backbone fragment of appropriate length and superimposes its flanking residues onto the corresponding anchoring residues in the template structure. If the fragment insertion generates a gap that cannot be closed ModeRNA will generate a model with an unsealed gap. Importantly, ModeRNA is able to recognize modified nucleosides and add or remove them accordingly in the model building process. To date, 115 different nucleotide modifications have been characterized in the MODOMICS database[18, 19]. ModeRNA recognizes one-character modification symbols[20] as well as a MODOMICS numbering scheme.

The MMB program

The MMB program (https://simtk.org/home/rnatoolbox)[21] is a software package for modelling DNA, RNA or protein structures that uses the internal coordinate space of dihedral angles and thus has time requirements proportional to the number of moving parts rather than the number of atoms. It provides accurate physics-based response to applied forces, but also allows user-specified restraints for incorporating experimental information. Another feature of MMB is that all Leontis-Westhof base pairs can be specified be satisfied during model construction. Additionally, it does not rely on fragment libraries, so the possible models to be constructed are not limited by the known RNA structure space.

The RNAComposer program

The RNAComposer program (http://rnacomposer.ibch.poznan.pl)[22], a fragment-based approach, uses the RNA sequence and its secondary structure as inputs, being the last one predicted or determined by experimental methods. It fragments the secondary structure into basic components such as stems, loops (apical, internal, bulge and n-way junctions) and single strands. The derived fragments are then used as query against a dictionary derived from the RNA FRABASE[23]. The search against the database is based on secondary structure similarity, sequence similarity, purine/pyrimidines compatibility, source energy resolution and energy. Residues not identical in the alignment are replaced by inserting the appropriate base from the NAB residues library[23]. After this, RNAComposer merges the different fragments for assembling the final structure. Each new fragment is superimposed to the old one by their common terminal canonical base pairs, and the new fragment base pair is deleted to avoid coordinate duplication. This initial 3D structure is subjected to two energy minimizations, one for the torsion angles and another for the Cartesian atom coordinate using the CHARMM force field[24].

Knowledge-based **modelling**

FARNA, FARFAR and SWA suite of programs

The Fragment Assembly of RNA server (FARNA, http://rosettaserver.graylab.jhu.edu)[25] is an energy-based program that predicts RNA 3D structures from its sequence alone. It was inspired by the Rosetta low-resolution protein structure prediction method. In FARNA each base is represented by a coarse-grained model of one centroid bead at the base origin. To reduce the sampling of the conformational space, the program builds a 3D structure library consisting of three-nucleotide fragments taken from a large rRNA subunit, from which torsion angles and sugar puckering parameters are stored, which allows capturing local conformations in such fragments. Then the fragments are selected based on their composition of purines and pyrimidines, and a simulation using Monte Carlo methods is used to assemble fragments into native-like structures. The folding simulation guided by a knowledge-based energy function takes into account both the backbone conformational preferences and side-chain interaction preferences observed in experimentally determined RNA structures. FARNA's energy function is a sum of five terms: (i) radius-of-gyration, (ii) a penalizing term for steric clashes between several representative atoms, (iii) a base-pairing potential, (iv) a term enforcing co-planarity of pairing bases, and (v) term enforcing base stacking. FARNA can incorporate tertiary-interaction restraints too. Importantly, a multiplexed hydroxyl radical (·OH) cleavage analysis (MOHCA)[26] infers the tertiary helical arrangements of large RNA molecules. It enables the detection of numerous pairs of interacting residues via random incorporation of radical cleavage agents followed by two-dimensional gel electrophoresis.

An update to FARNA, termed FARFAR (Fragment Assembly of RNA with Full Atom Refinement), added a refinement phase for atomic-level interactions[27]. This improvement was motivated by the need to predict the non-canonical interactions in RNA, which are responsible for the final 3D conformation of the molecule. The poor discrimination of native states by low-resolution energy functions of the FARNA program was thus addressed by introducing a high-resolution refinement phase driven by an accurate force field for atom-atom interaction. The method combines the previous FARNA protocol for low-resolution conformational sampling with optimization in the full-atom Rosetta energy function[28]. The FARNA score function was improved to model base- backbone and backbone-backbone interactions at a coarse-grained level.

Additionally, and in comparison to FARNA, in FARFAR the three-residue fragments were reduced to two and one in successive stages of Monte Carlo fragment assembly. This method is called the stepwise assembly (SWA)[29] and is based in a stepwise ansatz, which constructs atomic-detailed RNA models in small steps by exploring several million conformations for each monomer, and covering all build-up paths. This is an *ab-initio* method for recursively constructing small RNAs (up to 15 nucleotides) in small building steps in a polynomial computing time. It has been used to model single-stranded regions as loops which are difficult to model because of their non-canonical base pairs and unusual torsion angles, but it could be extended to bigger molecules. The drawback of this method is that the Rosetta all-atom energy function in which it is based is not able to model some very important details[30] as metal ions, long-range effects, higher-order dispersion effects[31] or hydrogen bond cooperativity.

BARNACLE

The BAyesian network model of RNA using Circular distributions and maximum Likelihood Estimation program (BARNACLE, http://sourceforge.net/projects/barnacle-rna)[32] describes RNA structure in a continuous space. This program makes it possible to sample 3D conformations that are RNA-like on a short length scale. Such a model can be used purely as a proposal distribution, but also as an energy term enforcing local conformations. BARNACLE combines a dynamic Bayesian network (DBN) with directional statistics. This approach is conceptually related to the probabilistic models of protein structure, although the RNA backbone has many more degrees of freedom than proteins, which makes it much more complicated to implement this approach, so it can't be used for long-range interactions.

MC-Fold/MC-Sym

The MC-Fold/MC-Sym pipeline (http://www.major.iric.ca/MC-Pipeline)[33] builds RNA 3D structures using the coordinates and relations between bases from known RNA structures. First, MC-Fold predicts the secondary structure of the RNA molecule, which is used as an input into MC-Sym, which predicts its 3D structure. Additional constraints can be applied to the model during the building procedure to ensure the conservation of particular structural features. MC-Fold/MC-Sym uses molecular dynamic simulations to minimize the energy of the predicted structure.

The Nucleic Acid Simulation Tool (NAST) program

The NAST program (https://simtk.org/home/nast)[34] predicts coarse-grained 3D structures based in knowledge-based statistical potentials. It requires secondary structure constraints and admits experimentally or phylogenetically-derived 3D constraints for predicting the 3D RNA structure. Within NAST, each nucleotide is represented as a bead by its C3' atom, which are then used by the knowledge-based function to calculate the geometric distances, angles and dihedrals from available structures between two, three, and four sequential residues, respectively. Moreover, the structures are then refined by molecular dynamics to satisfy the secondary and tertiary restrains. The NAST energy function is composed by four types of information: (i) geometries from solved ribosome structures, (ii) repulsive interactions between bases not farther than two positions away, (iii) ideal helical geometry for nucleotides participating in secondary structures, and (iv) long-range interactions between nucleotides participating in tertiary contacts. NAST assumes that the regions that have no secondary structure follow similar geometrical distributions to those observed in solved RNA structures. In summary, NAST is capable of modelling large (>150 nt) RNA molecules with a variety of accuracies depending on the provided external restraints.

Conclusion

Similar to what happened in the protein structure prediction field in the 80s, the RNA structural prediction field is currently observing a rapid increase of new computational methods, which are now reaching good accuracies for short RNA molecules. However, the field still have many limitations as well as several challenges ahead. Both the nature of the RNA sequence to model as well as the method used for modelling will have a large impact on the final accuracy of the model. Fortunately, both users and developers have now access to the results of the first collective blind test experiment in RNA 3D structure prediction. The RNA-Puzzles experiment (http://paradise-ibmc.u-strasbg.fr/rnapuzzles)[35] is a CASP-like experiment that aims at evaluating the accuracy of both manual and automatic methods for RNA structure prediction. The results from the RNA-Puzzles experiment provided deeper insights into the accuracy of available methods for different applications at the same time that stimulated the RNA structure prediction community for its on going efforts to improve its tools. In the future, one can expect the field of RNA structure prediction focusing in properly treating indels, correctly modelling tertiary long-range contacts or combining additional experimental data into the modelling of assemblies.

Given the limitation of space for this review, we have here very briefly introduced a limited number of software packages for each of the categories for RNA structure prediction. The readers are encouraged to read the references to those programs or acquire further knowledge in several reviews that cover each category in greater deepness[36-39].

References

1. Cech TR. Structural biology. The ribosome is a ribozyme. *Science* 2000, 289:878-879.

2. Mendes Soares LM, Valcarcel J. The expanding transcriptome: the genome as the 'Book of Sand'. *EMBO J* 2006, 25:923-931.
3. Bernstein E, Allis CD. RNA meets chromatin. *Genes Dev* 2005, 19:1635-1655.
4. Yang L, Froberg JE, Lee JT. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem Sci* 2014, 39:35-43.
5. Burnett JC, Rossi JJ. RNA-based therapeutics: current progress and future prospects. *Chem Biol* 2012, 19:60-71.
6. Capriotti E, Marti-Renom MA. Computational RNA structure prediction. *Current Bioinformatics* 2008, 3:32-45.
7. Rother K, Rother M, Boniecki M, Puton T, Bujnicki JM. RNA and protein 3D structure modeling: similarities and differences. *J Mol Model* 2011, 17:2325-2336.
8. Mathews DH, Moss WN, Turner DH. Folding and finding RNA secondary structure. *Cold Spring Harb Perspect Biol* 2010, 2:a003665.
9. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein-RNA interactions. *J Struct Biol* 2012, 179:261-268.
10. Zhao H, Yang Y, Zhou Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol Biosyst* 2013, 9:2417-2425.
11. Li X, Kazan H, Lipshitz HD, Morris QD. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA* 2014, 5:111-130.
12. Sharma S, Ding F, Dokholyan NV. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* 2008, 24:1951-1952.
13. Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA* 2008, 14:1164-1173.
14. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. *PLoS Comput Biol* 2006, 2:e85.
15. Gherghe CM, Leonard CW, Ding F, Dokholyan NV, Weeks KM. Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *J Am Chem Soc* 2009, 131:2541-2546.
16. Rother M, Rother K, Puton T, Bujnicki JM. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res* 2011, 39:4007-4022.
17. Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, Hershkovits E, Williams LD, et al. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* 2008, 14:465-481.
18. Dunin-Horkawicz S, Czerwoniec A, Gajda MJ, Feder M, Grosjean H, Bujnicki JM. MODOMICS: a database of RNA modification pathways. *Nucleic Acids Res* 2006, 34:D145-149.
19. Czerwoniec A, Dunin-Horkawicz S, Purta E, Kaminska KH, Kasprzak JM, Bujnicki JM, Grosjean H, Rother K. MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res* 2009, 37:D118-121.
20. Juhling F, Morl M, Hartmann RK, Sprinzl M, Stadler PF, Putz J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* 2009, 37:D159-162.
21. Flores SC, Sherman MA, Bruns CM, Eastman P, Altman RB. Fast flexible modeling of RNA structure using internal coordinates. *IEEE/ACM Trans Comput Biol Bioinform* 2011, 8:1247-1257.
22. Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J, Adamiak RW. Automated 3D structure composition for large RNAs. *Nucleic Acids Res* 2012, 40:e112.
23. Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak RW. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics* 2010, 11:231.

24. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al. CHARMM: the biomolecular simulation program. *J Comput Chem* 2009, 30:1545-1614.

25. Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. *Proc Natl Acad Sci U S A* 2007, 104:14664-14669.

26. Das R, Kudaravalli M, Jonikas M, Laederach A, Fong R, Schwans JP, Baker D, Piccirilli JA, Altman RB, Herschlag D. Structural inference of native and partially folded RNA by high-throughput contact mapping. *Proc Natl Acad Sci U S A* 2008, 105:4144-4149.

27. Das R, Karanicolas J, Baker D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* 2010, 7:291-294.

28. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* 2002, 322:65-78.

29. Sripakdeevong P, Kladwang W, Das R. An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc Natl Acad Sci U S A* 2011, 108:20573-20578.

30. Das R, Baker D. Macromolecular modeling with rosetta. *Annu Rev Biochem* 2008, 77:363-382.

31. Sato T, Tsuneda T, Hirao K. A density-functional study on pi-aromatic interaction: benzene dimer and naphthalene dimer. *J Chem Phys* 2005, 123:104307.

32. Frellsen J, Moltke I, Thiim M, Mardia KV, Ferkinghoff-Borg J, Hamelryck T. A probabilistic model of RNA conformational space. *PLoS Comput Biol* 2009, 5:e1000406.

33. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 2008, 452:51-55.

34. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* 2009, 15:189-199.

35. Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cao S, Das R, Ding F, Dokholyan NV, Flores SC, et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* 2012, 18:610-625.

36. Magnus M, Matelska D, Lach G, Chojnowski G, Boniecki MJ, Purta E, Dawson W, Dunin-Horkawicz S, Bujnicki JM. Computational modeling of RNA 3D structures, with the aid of experimental restraints. *RNA Biol* 2014, 11.

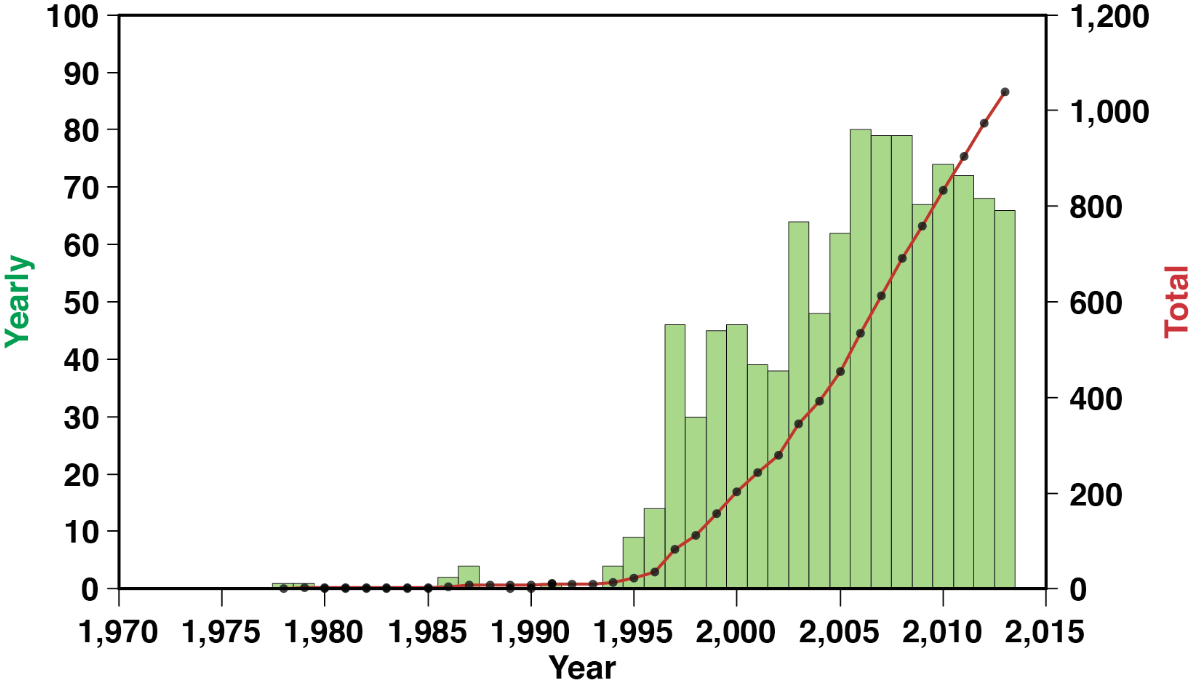
37. Laing C, Schlick T. Computational approaches to RNA structure prediction, analysis, and design. *Curr Opin Struct Biol* 2011, 21:306-318.

38. Westhof E, Masquida B, Jossinet F. Predicting and modeling RNA architecture. *Cold Spring Harb Perspect Biol* 2011, 3.

39. Masquida B, Beckert B, Jossinet F. Exploring RNA structure by integrative molecular modelling. *N Biotechnol* 2010, 27:170-183.

Figure captions

Figure 1. RNA structure deposition in the PDB database. The green bars (left y axis) indicate yearly new PDB entries and the red line (right y axis) represents the total number of RNA structures in the PDB database. The data ends in December 2013.


Yearly
Total
Year