Published online xx xxxx 2012

Nucleic Acids Research, 2012, Vol. xx, No. xx ***1–5***
doi:10.1093/nar/gkn000

CARNA - Alignment of RNA Structure Ensembles

Dragoș Alexandru Sorescu [1,*], Mathias Möhl [1,*], Martin Mann [1,2,*], Rolf Backofen [1,3], Sebastian Will [1†]

[1]Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany and [2]Theoretical Biochemistry, University of Vienna, Waehringerstrasse 17, 1090 Vienna, Austria and [3]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Albertstasse 19, 79104 Freiburg, Germany

Received February 28, 2012;

ABSTRACT

Due to recent algorithmic progress, tools for the gold standard of comparative RNA analysis, namely Sankoff-style simultaneous alignment and folding, are now readily applicable. Such approaches, however, compare RNAs with respect to a simultaneously predicted, single, nested consensus structure. To make multiple alignment of RNAs available in cases, where this limitation of the standard approach is critical, we introduce a web server that provides a complete and convenient interface to the RNA structure alignment tool **CARNA**. This tool uniquely supports RNAs with multiple conserved structures per RNA and aligns pseudoknots intrinsically; these features are highly desirable for aligning riboswitches, RNAs with conserved folding pathways, or pseudoknots. We represent structural input and output information as base pair probability dot plots; this provides large flexibility in the input, ranging from fixed structures to structure ensembles, and enables immediate visual analysis of the results. In contrast to conventional Sankoff-style approaches, **CARNA** optimizes all structural similarities in the input simultaneously, for example across an entire RNA structure ensemble. Even compared to already costly Sankoff-style alignment, **CARNA** solves an intrinsically much harder problem by applying advanced, constraint-based, algorithmic techniques. Although **CARNA** is specialized to the alignment of RNAs with several conserved structures, its performance on RNAs in general is on par with state-of-the-art general-purpose RNA alignment tools, as we show in a Bralibase 2.1 benchmark. The web server is freely available at **http://rna.informatik.uni-freiburg.de/CARNA**.

INTRODUCTION

With the discovery of numerous regulatory and catalytic RNAs that act *per se* without requiring translation to proteins, non-coding RNA has moved into the focus of biological research. Computational methods for RNA analysis have become indispensable tools, due to the vast amount of genomic data and the high cost of experimental analysis.

Because the function of non-coding RNAs is often stronger related to their structure than to their plain sequence, their comparison requires methods based on sequence and structure similarity. Commonly, those approaches consider canonical secondary structures, namely the set of base pairs between the organic bases A·U, C·G, and G·U, which are stabilized by H-bonds. This is usually sufficient, since the tertiary structure of RNAs largely depends on their secondary structure.

The secondary structure can be predicted using thermodynamic methods like `RNAfold` (1) or `Mfold` (2). Moreover, such tools compute base pair probabilities in thermodynamically equilibrated RNA structure ensembles (3).

Simultaneous alignment and folding, as originally proposed by Sankoff (4), is the acknowledged gold-standard to predict the consensus structure and alignment of a set of related RNA sequences. Since the computational complexity of the original Sankoff algorithm is too high for most practical applications, several faster variants of the approach have been developed.

The Sankoff-simplification introduced by `PMcomp` (5) significantly reduces the run-time by using a simplified energy model based on base pair probability matrices. Due to this idea, an alignment is obtained in two steps. First, one computes a base pair probability matrix, a.k.a. dot plot, for each sequence separately using `RNAfold`. The dot plot contains the probabilities for all base pairs under the assumption of a Boltzmann distributed ensemble of structures. Subsequently, one computes an alignment that scores, in addition to sequence similarity, the similarity of matched base pairs in a nested consensus structure according to their probabilities. This approach has been extended in `LocARNA` (6), which significantly improves the space and time complexity further by employing the sparsity of RNA dot plots; other such tools are `Lara` (7), `FoldalignM` (8), and `RAF` (9).

All these `PMcomp`-like alignment variants have the limitation that they score only a subset of the matching base pairs in the input dot plots. While this works in scenarios where only a single nested consensus structure is conserved, there are situations where this is not the case. First, many RNA molecules do not form nested structures, but crossing structures which are called pseudoknots. Second, riboswitches have more than one stable structure and switch between these structures to perform certain functions. Furthermore it is unknown to what extend intermediate structures in folding pathways are conserved.

*Joint first authors.
†To whom correspondence should be addressed. *will@informatik.uni-freiburg.de*

© 2012 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

To overcome this limitation of the simultaneous alignment and folding approach, we have developed the RNA alignment tool CARNA (10). This tool aligns RNAs with multiple structures per RNA or entire structure ensembles without committing to a single consensus structure. Instead of scoring the alignment of only a nested subset of the base pairs, it scores the matches of all base pairs in the base pair probability dot plots. Effectively, this allows aligning the entire Boltzmann distributed ensemble of structures.

In CARNA, all input structure information is encoded as dot plots, eventually. Like all PMcomp-style approaches, CARNA does by design align base pairs only if they occur with non-zero probability in the input dot plots. Understanding this relation and specifying the input dot plots plays a crucial role in the practical use of the CARNA tool. For this purpose the web server offers various ways to specify the structure input. While providing full low-level access, the server makes CARNAs reliance on input dot plots largely transparent. For example, input structure ensembles with and without pseudoknots can be generated automatically. Given pseudoknotted input, pseudoknots are aligned by CARNA naturally. For example, dot plots with pseudoknots can be computed using the approach of (11), which is directly supported by the server. Optionally, the server computes structure ensembles under structure constraints and allows to align RNAs according to single structures with and without pseudoknots.

The web server provides convenient and complete access to CARNA 1.0. Since the introduction of CARNA in (10), we have extended the tool significantly. First of all, we support multiple alignments based on a progressive alignment scheme. Furthermore, we support anchor constraints, which allow the user to guide CARNA's alignment based on prior knowledge. Finally, the scoring scheme of CARNA has been improved and supports now affine gap costs.

The web server provides informative graphical output, particularly showing a novel variant of dot plots, which we call *consensus conservation dot plots*. This allows the user to easily identify the conserved structural elements due to an intuitive color highlighting. Furthermore, the dot plots support the comparison of the individual sequences to the average dot plot obtained from the computed alignment.

Thus, the CARNA web server combines several unique features compared to general-purpose RNA alignment servers, e.g., our own LocARNA server (12) or the WAR web server (13) that allows comparing a variety of RNA alignment tools. Users who are already familiar with the LocARNA web server or command line tool, will appreciate that the scoring parameters, as well as the syntax for anchor or structure constraints for both tools are identical. Hence, in any work-flow, LocARNA can be easily exchanged with CARNA to account for the special requirements of pseudoknots or riboswitches.

WEB SERVER

Input and Output

The input to the CARNA tool consists of a set of sequences and one dot plot per sequence. The associated dot plot encodes a set of potential structures by base pair probabilities. CARNA computes an optimal alignment of the sequences with respect to a sequence and structure similarity based scoring. This scoring is identical to the one of LocARNA, except that it scores *all* structural matches of base pairs according to their probability in the input dot plots. In contrast, LocARNA scores only the structural matches of base pairs in a single nested consensus structure. In addition to the alignment, the server outputs a consensus dot plot that is an average over the input dot plots according to the alignment. We show two copies of this dot plot, one in the upper right triangle and one in the lower left triangle. The plot of the lower left triangle is annotated with conservation information of each base pair, resulting in a *conservation consensus dot plot*. More precisely, the conservation of a consensus base pair is measured as "inverse deviation" $1-2sd$, where sd is the standard deviation of the base pair's probability across all sequences in the alignment. In this way, an inverse deviation of one corresponds to perfect conservation, whereas zero corresponds to maximum variance. To facilitate the interpretation further, we also provide a conservation dot plot for each single RNA. For these dot plots, we project the input dot plots to the alignment and complement them with consensus and conservation information in the lower left triangle. Whereas the upper right triangle shows the probabilities of base pairs in the single sequence, the lower left triangle shows the corresponding averaged probabilities. In the upper right triangle, we optionally highlight all base pairs that are highly probable in the consensus, where the user can specify a threshold probability. Figure 1 shows the output from an alignment of three tRNA sequences. In the upper right triangle, all base pairs with average probability above 0.5 are highlighted. The color markup in the lower left triangle visualizes that the outermost stem is more conserved (red) than the other stems (blue/green).

Usage

In the simplest case, CARNA requires only a set of sequences in fasta format. Then, CARNA associates each sequence with a base pair probability matrix generated by RNAfold of the Vienna RNA package (14) or pairs (11) of NUPACK. Optionally, the former computation is controlled by structure constraints. Alternatively, the user can specify fixed structures or upload custom probability matrices for some or all sequences. Fixed structures can be provided in dot-bracket representation (using different symbols for encoding pseudoknots). Structure and anchor constraints are specified as annotation of the fasta input. Finally, one can customize the alignment scoring parameters; we support the same comprehensive parameter set as our web server for LocARNA (12).

Once a job is completed the user is forwarded to the output page. For long-running jobs, the user can request email notification even after submitting the job. Typical jobs with five to ten sequences of lengths up to 200 are usually completed within seconds to minutes. Moreover, the run time depends on the number of base pairs in the input structures, e.g. fixed input structures are typically aligned faster than entire ensembles. In general, due to the high complexity of the problem, the run-time of CARNA can vary strongly. Even

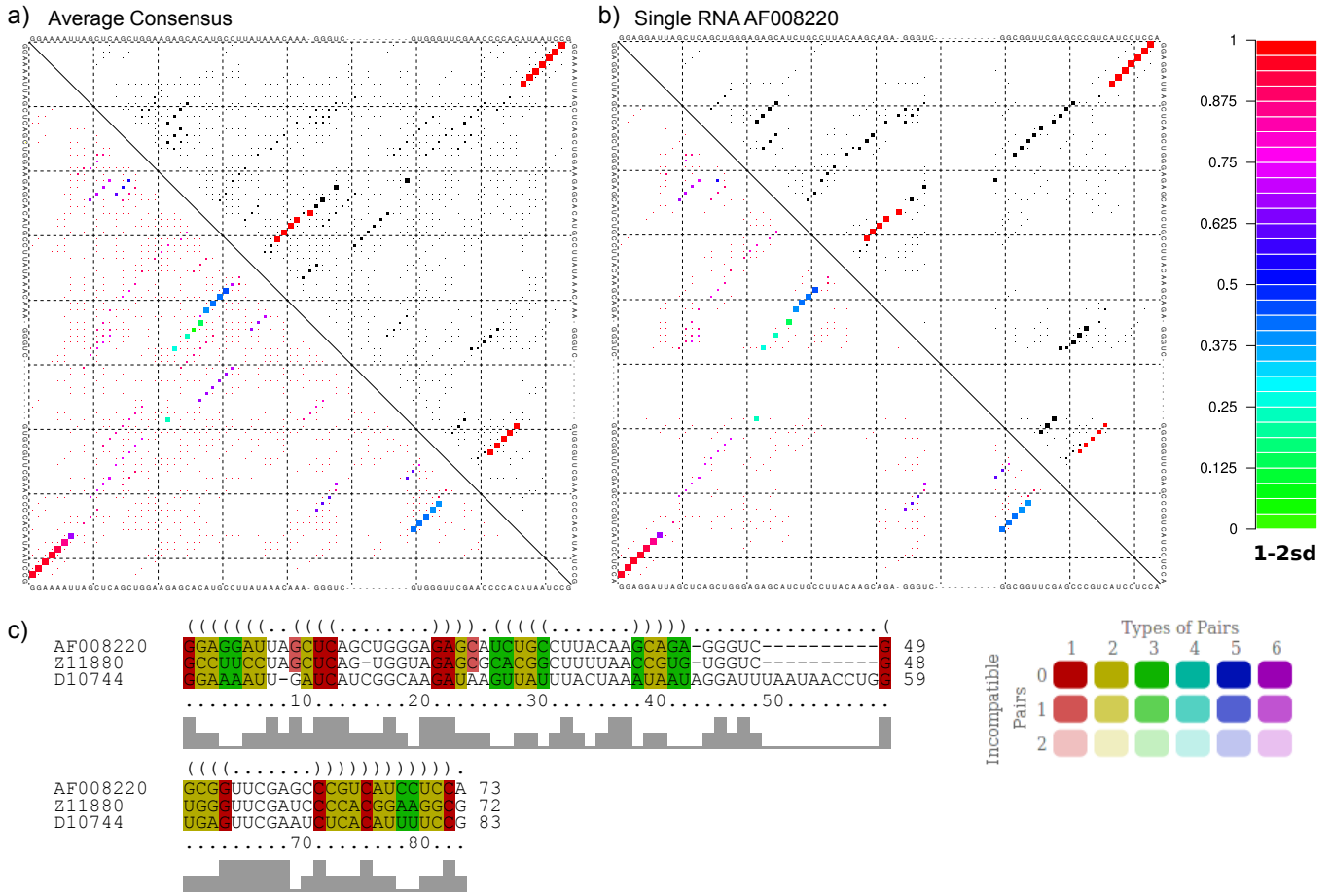

Figure 1. Example figures from an alignment of tRNA sequences with automatically generated dot plots by the CARNA web server. a) Consensus conservation dot plot. b) Conservation dot plot for the first sequence AF008220. The colors in the lower left triangles of the dot plots show the "inverse deviation" $1-2sd$ (cf. legend); in the upper right triangle, base pairs with average probability above the user-defined threshold 0.5 are shown in red. c) Computed CARNA alignment. The alignment figure and the consensus structure are generated by RNAalifold; the color markup indicates the number of base pair types and number of mismatches in a consensus column pair (cf. legend).

for same input sizes it depends on the complexity of the input dot plots. In some cases CARNA would run significantly longer than the typical time to find the best solution. For this reason, the server limits the search time by default (cf. Section Algorithm). Consequently, the run-time is limited to grow on average quadratically with the number of input sequences. The server does not limit the input size, however cancels jobs that run for more than one week or use more than 2GB of memory. Results are stored on the server for 30 days. Details of the visualization of the conservation consensus dot plots can be interactively controlled on the output page. Furthermore, one can download the individual dot plots and the alignment as postscript or pdf files or an archive containing the complete results.

Implementation

The web server is based on a generic framework that simplifies the setup of new frontends for arbitrary bioinformatics command line tools. The framework has been previously applied to the Freiburg RNA Tools Web Server (12) and was significantly extended for CARNA. Since many routine tasks of a web server, like input consistency checking, error handling, and job scheduling are handled by the framework and shared by all tools, this contributes to a robust and consistent user experience.

Internally, the framework implements a generic wrapper around arbitrary command line tools. The tool-specific input parameters are described in a simple XML dialect. Based on this, the framework can automatically check the input for consistency, perform preprocessing, and pass it to the command line tool. For example, the parameter corresponding to the input sequences is annotated with the constraint "fasta file" in the XML specification. Consequently, the framework checks the sequence input for valid fasta syntax and automates the error handling.

When the user submits a valid request from the input page, a script is submitted to our compute cluster managed by a Sun Grid Engine. This allows to handle many requests in parallel and adapt to varying workloads. Once the script finishes its execution, the user is redirected to the result page.

The framework runs as a web application in an Apache Tomcat container. The main functionality is controlled by a Java servlet, whereas JavaServer Pages technology allows to

easily provide dynamic content. The interactive elements of the input and output pages are implemented using JavaScript.

ALGORITHM

The core algorithm of CARNA performs a pairwise alignment of sequences with dot plots. Although the alignment variant solved by CARNA is computationally complex (MAX-SNP-hard), we make this variant available for practical applications using advanced constraint programming techniques. Applying a branch and bound scheme, the algorithm finds step by step better solutions until the best solution is found. Usually CARNA finds an optimal solution within seconds to minutes. To find solutions of hard instances in reasonable time, the user can specify a time limit for each pairwise alignment. When this limit is exceeded, the tool returns the so far best solution. The constraint algorithm is described in full detail in (10).

Based on the pairwise algorithm, we construct multiple alignments using a progressive alignment scheme. This step is analogous to the multiple alignment method of LocARNA (6). Thus, we follow the protocol of (6), only replacing LocARNA by CARNA. The procedure starts by constructing a guide tree out of all-against-all pairwise comparisons. Then, beginning with the most closely related RNAs, one progressively aligns the RNAs following the guide tree from the leaves to the root. This scheme requires the ability to align partial multiple alignments to each other; therefore, after each progressive alignment step, we compute an alignment profile as well as a consensus dot plot that can be used as input for the next progressive step.

RESULTS

In (10), we have studied the performance of CARNA on pseudoknotted RNAs and riboswitches. Here, we additionally present a Bralibase 2.1 benchmark (15), comparing CARNA to the state-of-the-art general-purpose RNA alignment tools LocARNA, Lara, RAF, and the purely sequence-based alignment tool MAFFT (16). Bralibase 2.1 contains instances of pairwise alignments and multiple alignments of up to 15 sequences. Figure 2 shows LOWESS curves (17) based on all instances for two important alignment quality measures versus the average pairwise sequence identity (APSI). Figure 2a) shows the similarity of the generated alignment to the reference alignment; this is measured by the sum-of-pairs score (SPS) introduced with Bralibase 2.1. For Figure 2b), we predict structures out of the generated alignments using RNAalifold and measure its similarity to the reference structure as Matthews correlation coefficient (MCC) (18). The motivation for plotting these measures against APSI is that RNA alignment is particularly hard for low sequence similarity instances.

Notably, the Bralibase 2.1 benchmark is tailored to the standard case of RNA comparative analysis, where only a single RNA structure is conserved. Although this benchmark does therefore not require CARNA's special capabilities, CARNA's performance is on par with the state-of-the-art RNA alignment tools.

DISCUSSION

We have presented the web server CARNA for multiple alignment of RNA structure ensembles. In contrast to previous approaches for RNA alignment, CARNA is tailored for aligning RNAs with pseudoknots or several conserved structures. The latter is particularly useful for the alignment of riboswitches. The web server offers a convenient and intuitive way to interactively explore the results. For this purpose, we developed the concept of conservation dot plots, which visualizes the conservation of base pairs by average probabilities and color-encodes variance information.

FUNDING

This work was partially supported by the German Research Foundation (BA 2168/3-1, MO 2402/1-1, and WI 3628/1-1) and by the German Federal Ministry of Education and Research (S.H., BMBF grant 0313921 FRISYS to R.B.).

ACKNOWLEDGEMENTS

We thank Benjamin Schulz for running the Bralibase 2.1 benchmarks.

Conflict of interest statement. None declared.

REFERENCES

1. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie,* **125**, 167–188. [doi:10.1007/BF00818163].
2. Zuker, M. (1994) Prediction of RNA secondary structure by energy minimization. *Methods in Molecular Biology,* **25**, 267–94. [PubMed:7516239] [doi:10.1385/0-89603-276-0:267].
3. McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers,* **29**(6-7), 1105–19. [PubMed:1695107] [doi:10.1002/bip.360290621].
4. Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.,* **45**(5), 810–825. [doi:10.1137/0145048].
5. Hofacker, I. L., Bernhart, S. H., and Stadler, P. F. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics,* **20**(14), 2222–7. [PubMed:15073017] [doi:10.1093/bioinformatics/bth229].
6. Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology,* **3**(4), e65. [PubMed:17432929] [doi:10.1371/journal.pcbi.0030065].
7. Bauer, M., Klau, G. W., and Reinert, K. (2007) Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics,* **8**, 271. [PubMed:17662141] [PubMed Central:PMC1955456] [doi:10.1186/1471-2105-8-271].
8. Torarinsson, E., Havgaard, J. H., and Gorodkin, J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics,* **23**(8), 926–32. [PubMed:17324941] [doi:10.1093/bioinformatics/btm049].
9. Do, C. B., Foo, C.-S., and Batzoglou, S. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics,* **24**(13), i68–76. [PubMed:18586747] [PubMed Central:PMC2718655] [doi:10.1093/bioinformatics/btn177].
10. Palu, A. D., Möhl, M., and Will, S. (2010) A propagator for maximum weight string alignment with arbitrary pairwise dependencies. In *Proceedings of the 16th International Conference on Principles and Practice of Constraint Programming (CP-2010)* p. 8. [doi:10.1007/978-3-642-15396-9_16].
11. Dirks, R. M. and Pierce, N. A. (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J Comput Chem,* **25**(10), 1295–304. [PubMed:15139042] [doi:10.1002/jcc.20057].
12. Smith, C., Heyne, S., Richter, A. S., Will, S., and Backofen, R. (2010) Freiburg RNA Tools: a web server integrating

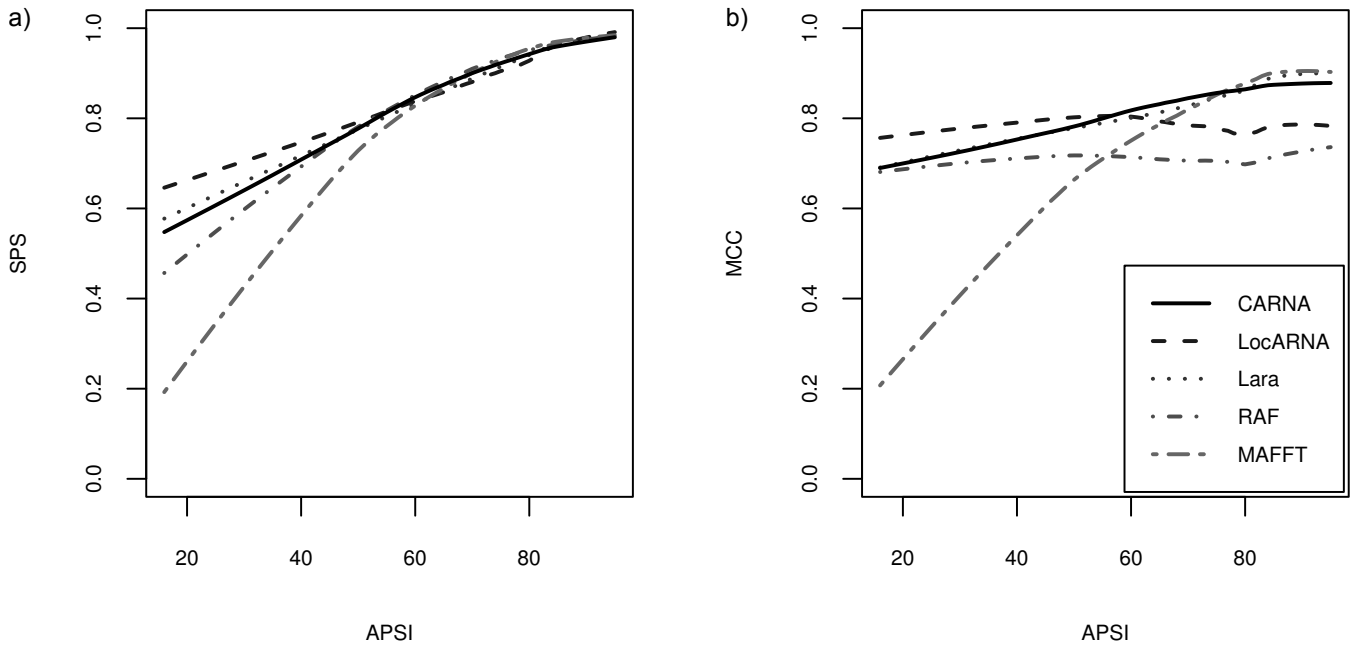

Figure 2. Bralibase 2.1 benchmark of CARNA compared to state-of-the-art, general-purpose RNA alignment tools. a) Average pairwise sequence identity (APSI) vs. similarity to reference alignment measured by the sum-of-pairs score (SPS) introduced with Bralibase 2.1. b) APSI vs. similarity to reference structure measured by Matthews correlation coefficient (MCC).

IntaRNA, ExpaRNA and LocARNA. *Nucleic Acids Research,* **38 Suppl**, W373–7. [PubMed:20444875] [PubMed Central:PMC2896085] [doi:10.1093/nar/gkq316].

13. Torarinsson, E. and Lindgreen, S. (2008) WAR: Webserver for aligning structural RNAs. *Nucleic Acids Research,* **36**(Web Server issue), W79–84. [PubMed:18492721] [doi:10.1093/nar/gkn275].
14. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., and Hofacker, I. L. (2008) The Vienna RNA websuite. *Nucleic Acids Research,* **36**(Web Server issue), W70–4. [PubMed:18424795] [PubMed Central:PMC2447809] [doi:10.1093/nar/gkn188].
15. Wilm, A., Mainz, I., and Steger, G. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol,* **1**, 19. [PubMed:17062125] [PubMed Central:PMC1635699] [doi:10.1186/1748-7188-1-19].
16. Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research,* **30**(14), 3059–66. [PubMed:12136088] [PubMed Central:PMC135756].
17. Cleveland, W. S. (1981) Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician,* **35**(54). [doi:10.2307/2683591].
18. Matthews, B. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochem. Biophys. Acta,* **405**, 442–451. [PubMed:1180967].