OXFORD

Briefings in Bioinformatics, 18(2), 2017, 306–311

doi: 10.1093/bib/bbw022
Advance Access Publication Date: 16 March 2016
Paper

The BRaliBase dent—a tale of benchmark design and interpretation

Benedikt Löwes, Cedric Chauve, Yann Ponty and Robert Giegerich

Corresponding author. Robert Giegerich, Faculty of Technology, Bielefeld University, Bielefeld, Germany. Tel.: +49 521 1062913; Fax: +49 521 1066411.
E-mail: robert@techfak.uni-bielefeld.de

Abstract

BRaliBase is a widely used benchmark for assessing the accuracy of RNA secondary structure alignment methods. In most case studies based on the BRaliBase benchmark, one can observe a puzzling drop in accuracy in the 40–60% sequence identity range, the so-called ‘BRaliBase Dent’. In this article, we show this dent is owing to a bias in the composition of the BRaliBase benchmark, namely the inclusion of a disproportionate number of transfer RNAs, which exhibit a conserved secondary structure. Our analysis, aside of its interest regarding the specific case of the BRaliBase benchmark, also raises important questions regarding the design and use of benchmarks in computational biology.

Key words: RNA structural alignment; benchmark; RNA family database

Introduction

The role of benchmark data sets

As much as biosequence databases are an invaluable resource in molecular biology research, bioinformatics benchmarks are a crucial aid in the continued development and evaluation of bioinformatics tools and algorithms (see [1] for multiple sequences alignment). They allow us to compare old and new approaches with the same problem with respect to their use of computational resources, as well as with respect to their qualitative performance. When tool developers can make use of established benchmarks, reproducibility of results benefits greatly.

Good benchmark data sets must satisfy a number of criteria. They should contain the best, curated data sets available at their time, should reflect the diversity in their problem domain, and should not be biased towards problem subdomains of a specific nature. Naturally, these criteria are not easy to fulfil, especially when the problem domain itself is a new or difficult research topic, and available data at a given time may only partially reflect its diversity and complexity. For practical use, a benchmark should be subdivided, such that users can extract subsets with characteristic properties, such as sequence length, degree of conservation or phylogenetic classification.

Note that benchmarks do not need to encompass the complete data available in their domain, contrasting in that respect with databases used in applications. However, if domain knowledge grows, benchmark data sets need to be updated once in a while to remain representative of their problem domain.

Two successful benchmarks: BAliBASE and BRaliBase

Widely known and used in sequence analysis is the BAliBASE benchmark, first compiled in 1999 by Thompson *et al.* [2]. It

Benedikt Löwes is a PhD student in the International Research Training Group ‘Computational Methods for the Analysis of the Diversity and Dynamics of Genomes’ at Bielefeld University in Germany in cooperation with Simon Fraser University in Burnaby, Canada. His current research focuses on convergent evolution of viruses and RNA bioinformatics.
Cedric Chauve is a professor of Mathematics at Simon Fraser University. He holds a PhD in Computer Science from Bordeaux University. His main research interest is in the development of comparative genomics algorithms.
Yann Ponty is a CNRS researcher at Ecole Polytechnique, France. He holds a PhD in Computer Science from Université Paris Sud. His main research interests are in RNA bioinformatics, with a special emphasis on dynamic programming.
Robert Giegerich is a professor emeritus of Computer Science at Bielefeld University. His main research interest is in (algabraic) dynamic programming and algorithms for RNA structure analysis.
Submitted: 31 December 2015; **Received (in revised form):** 28 January 2016

provides hand-curated multiple sequence alignments categorized by core blocks of conservation sequence length, similarity and the presence of insertions and N/C-terminal extensions. It has seen several updates, the most recent being version 3.0 compiled in 2005 [3]. It has been used in countless sequence analysis studies (the Google Scholar reference count for the above two publications comes close to 800 at the time of this writing), and how to use and interpret the scores provided by BAliBASE has become a research topic of its own [4].

In the first comparative evaluation of structural alignment algorithms [5] in 2004, the lack of an established benchmark was strongly felt. Motivated by the success of BAliBASE, but targeting a smaller research community, the BRaliBase benchmark was first compiled in 2005 [6] and enhanced in 2006 [7]. It makes available a test-set of 'structural' alignments of RNA sequences that has been an important resource to many researchers, including ourselves, in comparative RNA bioinformatics (again according to Google Scholar, both BRaliBase publications have been cited above 300 times).

This story is dedicated to a peculiarity of the BRaliBase benchmark, which we call the 'BRaliBase dent'.

The BRaliBase dent

The BRaliBase dent refers to the puzzling phenomenon that even the best programs for structural RNA alignment seem to exhibit a weakness of performance in a range of moderate sequence similarity. At least, this is what we used to think. Let us take a closer look.

Originally, BRaliBase was used to demonstrate that enhanced sequence alignment approaches that in some way take account of conserved RNA structure yield more realistic alignments than the best (structure-unaware) sequence alignment programs. The study of Gardner *et al.* [5] focused on comparative RNA structure prediction and classified structural RNA alignment algorithms in three categories:

- Plan A aligns related sequences by pure sequence alignment. Then, it folds the alignment as a whole. Thus, although the second phase assigns the best possible structure to the given alignment, this alignment is determined solely in the first phase, and its quality is expected to degrade with decreasing sequence similarity.
- Plan B implements some form of the Sankoff algorithm [8], simultaneously optimizing sequence similarity and a folding score. Such algorithms have much higher resource requirements than Plan A, but are expected to give better results for diverged sequences.
- Plan C first suggests a set of alternative structures for each sequence separately, aligns the structures (possibly ignoring their sequence content) and then derives a sequence alignment from the structure alignment. At the time of the study [5], no Plan C approach was known.

Besides the above, there are a number of approaches that do not fit these categories, avoiding dynamic programming over the whole search space and using heuristic or probabilistic methods. However, many of the most used methods follow one of the three plans outlined above, and provide a large enough corpus for our study of the BRaliBase dent.

The BRaliBase evaluation of 2005 (Figure 1A) showed that the expectations indeed hold. Pure sequence alignment methods quickly loose quality when sequence identity drops under 75%, and folding the alignment cannot compensate the loss. As we know today, post-processing Plan A with a Plan C-type approach can locally improve the alignment and provide a partial compensation [9]. Plan B-type algorithms as of 2005 perform better. With decreasing sequence identity, their performance degrades more gracefully overall. Figure 1B shows the effect of smoothing on two of the curves. The green (Foldalign), purple (PMcomp) and brown (Dynalign) curves represent the three best performing Plan B approaches. To a varying degree, they show a decline in the area between 60–40% identity, whereas their performance improves again when sequence conservation becomes even lower. This phenomenon is what we call the BRaliBase dent.

To this day, the dent persists. Figure 2A shows a recent re-evaluation of the extended 2006 BRaliBase data with currently available algorithms (The ability to run a re-evaluation a decade later is a benefit of having persistent benchmarks in the first place. Note, [7] only tested their extended data set on pure sequence alignment methods. To our knowledge, [10] was the first publication to use the 2006 BRaliBase data set for benchmarking their structural aligner Lara). The shape of the dent has slightly changed, but the position in the range between 60% and 40% identity is consistent. This can be accounted to the 76-fold

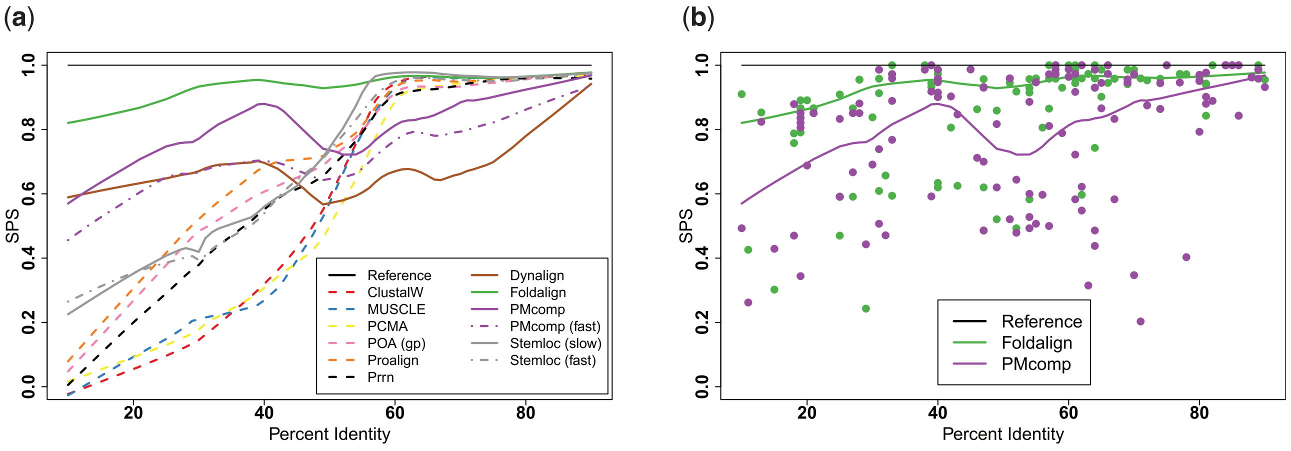

Figure 1. (**A**) Original BRaliBase evaluation of 2005 [6]. Dashed lines show pure sequence aligners, solid lines show structural aligners and dotted-solid lines show structural aligners with varying parameters. (**B**) Extended evaluation for Foldalign and PMcomp that shows all results for the 118 pairwise alignments for both tools using the original data. SPS (Sum of Pairs Scores) is a measure of alignment accuracy compared with a reference data set introduced in [2]. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

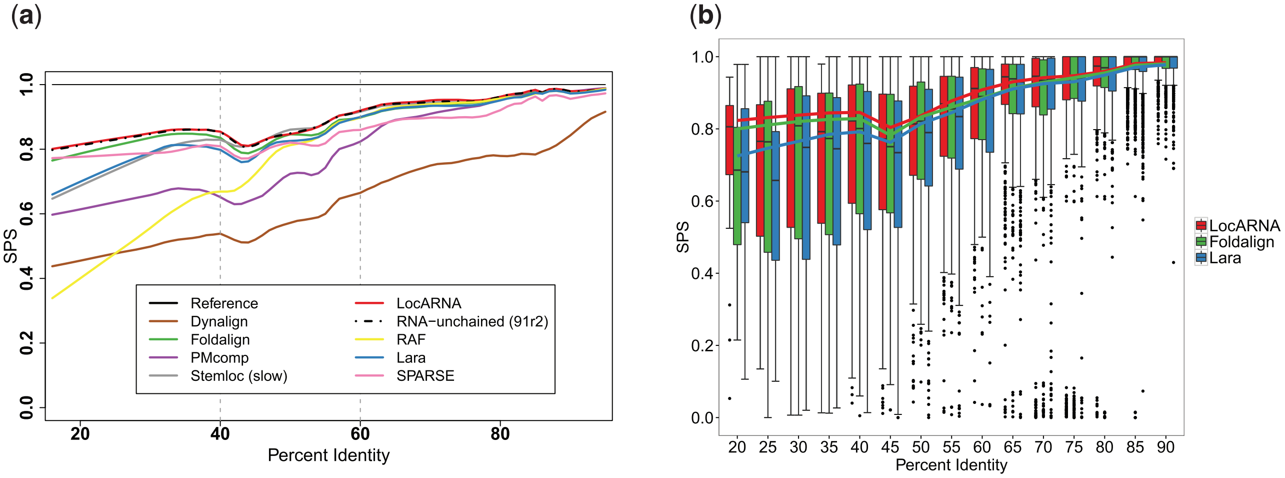

Figure 2. (**A**) Re-evaluation of the 2006 BRaliBase data from [7] with currently available structural aligners. (**B**) The same re-evaluation with only three tools and box plots showing the detailed distribution of SPS. Here, we have chosen to add LocARNA as the best performing tool and substituted PMcomp by Lara, because some of PMcomp's alignment computations resulted in errors and Lara represents an interesting alternative not fitting into the previously mentioned categories. (See Supplementary Data Table S1 for details.) A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

increase in data points for which the effect of smoothing can be seen in Figure 2B.

In 2007, we designed a Plan C approach. It first computed a set of abstract consensus shapes [11, 12], aligned their shape representative structures with RNAforester [13] and picked the best structure alignment to derive the final sequence alignment from it. Aside from better speed, we hoped for a performance improvement in the dent zone. Testing the new approach, it perfectly fell between Foldalign [14] and PMcomp [15], exactly reproducing the dent in the twilight area. This was puzzling, as the approach is quite different in nature from the Sankoff-type approaches. We believed that human curators do something special in the twilight zone of 60–40% identity that our algorithms cannot do. Anyway, the approach was never published. (For details see Supplementary Data 2.)

Researchers kept working on the structural alignment problem, using BRaliBase to evaluate their ideas [10, 14–19]. And, in fact, there was some indication that the dent had been mastered. For example, publications [20, 21] performed benchmarks of their tools using the 2006 version of BRaliBase, indicating a dent-less performance of the algorithms. However, a closer look shows that this impression is because of the use of higher smoothing factors for creating the regression curves. In Figure 2B, we find a clear collapse of the SPS around 45%, which indicates that the use of a lower smoothing factor is key in order not to artificially mask the dent.

In 2015, one of our authors (C.C.) contributed to an algorithm not fitting the above categories, but mimicking the way a human curator might work [22]. The algorithm RNA-unchained first picks significant sequence-structure matches, takes them as alignment seeds and then uses LocARNA to align the rest. Evaluations in [22] also suggested a dent-less performance. However, as the algorithm is a hybrid of seed recognition and Plan B-type alignment by LocARNA, it happens that no seeds are found with data of low similarity. In such a case, the result is purely determined by LocARNA, and these points were dropped from the diagrams, as they do not indicate a contribution of the new algorithm tested. If we include these points of measurement (Figure 2A), the dent returns even with this approach. (For details see Supplementary Data S3.)

Understanding the reason for the BRaliBase dent is a natural question. In particular, one can wonder whether it is owing to an intrinsic weakness of the algorithms applied on RNA sequences with average identity, in which case addressing this weakness can be seen as a methodological goal. However, with so many algorithms based on independent ideas showing the same performance characteristics, we were led to suspect that the dent was a feature of the data rather than the algorithms.

Explaining the BRaliBase dent

For our exploration of the origin of the dent, we will use a single Plan B tool, LocARNA, and the k2 benchmark data set of the 2006 version of BRaliBase that consists of 8,976 pairwise alignments from 36 different RNA families, based on seed alignments from *Rfam* 7.0 [23].

The consensus structure: an abstraction

'Structural' alignment uses sequence comparison and structure prediction to elucidate a consensus structure for a group of functionally related RNA molecules. Such a consensus structure may provide hints regarding the mechanisms underlying the molecule's biological function. However, it must be kept in mind that the consensus structure is only a theoretical construct. Each molecule performs its function as an individual, by a structure it folds into all alone. Not all parts of the sequence are equally determined by the function, and hence might fold into a conserved structure to a varying degree.

Let us take two extreme examples. On the one hand, transfer RNAs (tRNAs) must attain their typical L-shaped 3D structure, to fit in the active site in the ribosome as a whole. Hence, this structure and the underlying 2D cloverleaf structure are strongly preserved, even when the sequence is diverged. On the other hand, let us consider the self-splicing introns. They have a strongly conserved catalytic site, embedded in a pseudo-knotted structure, but may also include large introns (even including nested self-splicing introns) that are irrelevant for self-splicing, and thus cannot be expected to admit a good structural alignment overall. When a benchmarked tool achieves a lower alignment score for RNA family A compared with family B, this does not necessarily mean that it performs worse—it may simply mean that the structure is less conserved in family B, and the tool reflects this fact.

Our best test data: tRNAs

Next to ribosomal RNA, the tRNA may be the most well-studied class of functional RNA molecules. As structure conservation extends over the whole molecule, and as tRNAs are part of the universally conserved translational machinery, we have an enormous amount of well-curated structural alignments, where structure conservation is high even with low sequence identity. As tRNAs are so short that all tools can handle them in reasonable time, they constitute a favourite test case for algorithm developers. This is a strong argument to include a large number of tRNA alignments in a benchmarking data set. In fact, tRNA is the most highly represented family in BRaliBase, contributing 2039 data points. Can too much of the good stuff be hurting in the (bitter) end?

Performance variation on different RNA families

Before including all measurements into a single benchmark, it may be wise to look at the results for individual families. We do so in Figure 3A.

It is apparent that the tRNAs have a special position within the data set. They are represented by the red line. Compared with other families, two observations stand out: first, the computed tRNA alignments perform much better over the whole range of sequence identities than those in other families. This is surprising, as the performances of most other families seem to decline noticeably with decreasing sequence identity. Even though the curves for some families like THI or Cobalamin behave in a rather erratic manner, all of them are decreasing more strongly in principle. (See Supplementary Data Figure S3 for the plots of all 36 families.)

Secondly, as these curves are generated using local regression, the volatility is an indicator of strongly fluctuating values as well as a small amount of data points for various ranges of identities. Figure 3B confirms this assumption by showing that the number of alignments of similar identities fluctuates strongly even within the same RNA families. The most important point is, however, that the tRNA alignments have many data points in the range between 20% and 60% identity, whereas the other families have the highest number of data points in the high identity region. This is taken to the extreme for alignments with sequence identities <40%, where there are almost no data points other than the tRNAs. We can separate the data set into different groups: first, the tRNAs with a rapidly shrinking share in the high identity area and the overwhelming majority in the low identity region. Secondly, the 5S_rRNAs resemble the remaining families, but their share has slightly different peaks for identities >55%. Lastly, all other families follow the same trend, but with different peak heights, and can be considered as background. (The proportion of alignments in all families are individually reported in Supplementary Data Figure S4.)

The combination of all these observations suggests that the dent is mainly caused by the conserved alignments of the tRNA family as well as the unbalanced composition of the data set: with tRNAs aligning especially well, and few other data points in the area of low similarity, when combining all performance curves, the result ends up rising again in the low similarity area.

To verify this hypothesis, we have considered in details the performance of LocARNA as representative for all tools for tRNA and non-tRNA alignments separately (Figure 4A). Both show a decline with decreasing similarity, but the decline for non-tRNAs (yellow) is much steeper. Where non-tRNA data become sparse (around 40% identity), the tRNA curve (red) almost entirely determines the shape of the overall curve (blue). Additionally, the non-5S_rRNAs curve (green) shows a behaviour similar to the overall curve, indicating that the second most abundant family does not influence the presence of the dent. Therefore, we can summarize that the dent in the 2006 version of BRaliBase is caused by the interaction of two data set-specific factors: the exceptionally good performance of tools over tRNAs, as well as the lack of non-tRNA alignments, which would probably yield lower performances, in the identity region below 40%.

To eradicate the imbalance, we tested a sampling approach in which no single RNA family is allowed a bigger representation than 20% within each sequence identity region. The resulting curve (purple) shows a clear decreasing SPS tendency for decreasing identities with some smaller dents in the high identity regions that cannot be avoided.

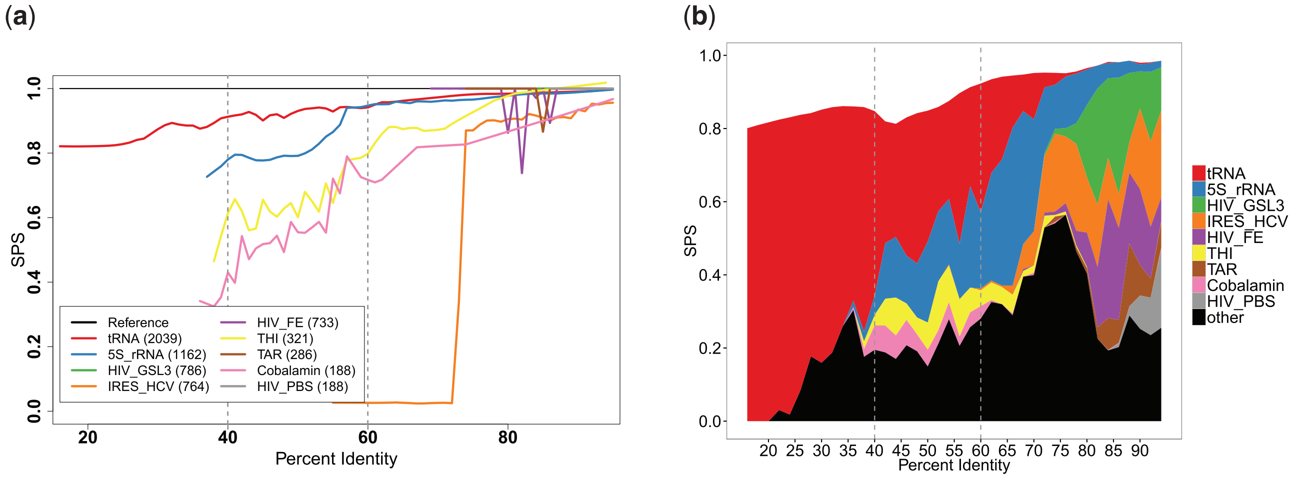

Figure 3. The two plots show 9 of 36 RNA families with at least 180 alignments. (**A**) Familywise performance of LocARNA. The family names in the legend are further accompanied by the total number of alignments for each family in brackets. (**B**) Each family's share of LocARNA's SPS (after local regression) per sequence identity. The remaining families with <180 alignments are grouped into 'other'. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

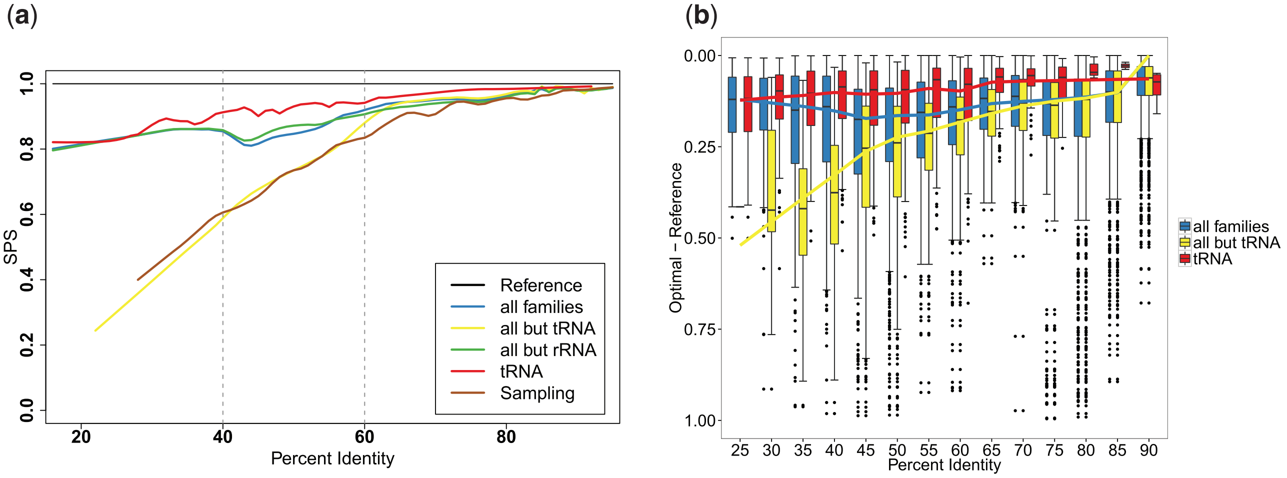

Figure 4. Separate evaluation of tRNA alignments, non-tRNA alignments and the complete data set. Comparison (**A**) by BRaliBase SPS and (**B**) as length normalized score differences between the optimal version and the reference using PMcomp 's scoring scheme (LocARNA was substituted by PMcomp for the ease of a scoring scheme that is easy to reverse engineer and implement). Additionally, (**A**) shows a curve for a sample approach in which no families share is bigger than 20% per sequence identity and the non-5S_rRNA alignments. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

The BRaliBase SPS measures the agreement of computed alignments to a reference. But during the computation of the predicted alignment, different alignments have to be computed and scored internally to choose the optimum. This allows us to reverse our view point. Namely, we score the reference alignment using the tool's internal scoring scheme, and compare it with the prediction, under the rationale that the special status of tRNAs in the benchmark should also be apparent in this comparison.

Figure 4B compares the PMcomp score of the optimal alignment and the reference alignment by representing the length normalized differences for various sequence identities. This evaluation confirms the observations of the SPS-based analysis. Here, the scores of the tRNA reference alignments (red) are close to the ones of the optimal alignments for the whole range of sequence identities. The score differences of the optimal alignments and the non-tRNA reference alignments (yellow) increase with decreasing sequence identities. Together, the score differences of all RNA alignments show the same properties as the SPS-based analysis, with the biggest difference being around 45% identity. This further supports our conclusion that the BRaliBase dent is in fact not an artefact of the SPS or the curve smoothing factor, but rather arises from the biased composition of the overall data set.

Conclusion

The BRaliBase dent, which for a decade has puzzled the authors of this article as well as many colleagues in the field of comparative RNA structure prediction, can be explained by a bias in the benchmark data set, which holds 'too much of the good' data in some conservation regions. The lesson learned from this experience is simple:

- When different data subsets in the benchmark show deviating overall characteristics, one should refrain from merging them into an overall performance diagram.
- When a benchmark holds data with diverging characteristics and a potential representation bias, it should provide means for sampling unbiased subsets.

In the case of BRaliBase, the bias was caused by too much of the best data dominating the performance analysis in low similarity regions, leading to a perceived decay of tools performance, the dent, in the consecutive 40–60% identity region. Quite understandably, this observation has attracted our attention to this region, and this feature of the benchmark data ended up being misdiagnosed as a localized weakness in our algorithms. This conclusion was seemingly confirmed by the apparent demonstration, in recent publications, that the dent could be overcome by innovative approaches. However, as explained above, this evidence turned out to be deceptive. As a word of consolation: a good amount of creative and beautiful algorithmics in the field of comparative structure prediction has been initiated by the misinterpretation of the BRaliBase dent as a challenge to algorithm designers.

A new version of the BRaliBase benchmark, which draws from the increased knowledge about RNA families collected in the past decade, is clearly desirable. Although our study can be taken as advice, marking a pitfall to avoid, the construction of a new benchmark is a substantial contribution of its own, and must be left here as a topic for future research.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- In most case studies that used the BRaliBase benchmark, the accuracy of the tools and algorithms seems to drop within the 40–60% sequence identity range.
- The BRaliBase dent, as the graphical representation of this phenomenon, is caused by the exceptionally good performance of the tools for tRNA alignments, combined with their overrepresentation in the identity region below 40%.
- To prevent biased data sets, one should refrain from merging different data subsets that show deviating overall characteristics or provide sampling capabilities for the users to gain unbiased subsets.

Funding

BL is funded by the International DFG Research Training Group GRK 1906/1 YP is supported by the RNALands program (ANR-14-CE34-0011), jointly funded by the French/ Agence Nationale de la Recherche/ and the/Austrian Science Fund/ (FWF).

References

1. Chatzou M, Magis C, Chang J-M, *et al*. Multiple sequence alignment modeling: methods and applications. *Brief Bioinform* 2015, doi: 10.1093/bib/bbv099.
2. Thompson JD, Plewniak F, Poch O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* 1999;**15**(1):87–8.
3. Thompson JD, Koehl P, Ripp R, *et al*. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* 2005;**61**(1):127–36.
4. Błażewicz J, Formanowicz P, Wojciechowski P. Some remarks on evaluating the quality of the multiple sequence alignment based on the BAliBASE benchmark. *Intl J Appl Math Comput Sci* 2009;**19**(4):675–8.
5. Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 2004;**5**:140.
6. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* 2005;**33**(8):2433–9.
7. Wilm A, Mainz I, Steger G. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol* 2006;**1**:19.
8. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* 1985;**45**(5):810–25.
9. Bremges A, Schirmer S, Giegerich R. Fine-tuning structural RNA alignments in the twilight zone. *BMC Bioinformatics* 2010;**11**(1):222.
10. Bauer M, Klau GW, Reinert K. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics* 2007;**8**:271.
11. Reeder J, Giegerich R. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics* 2005;**21**(17):3516–23.
12. Voß B, Giegerich R, Rehmsmeier M. Complete probabilistic analysis of RNA shapes. *BMC Biol* 2006;**4**(1):5.
13. Höchsmann M, Voß B, Giegerich R. Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans Comput Biol Bioinform* 2004;**1**(1):53–62.
14. Hull Havgaard J, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput Biol* 2007;**3**(10):e193.
15. Hofacker IL, Bernhart SHF, Stadler PF. Alignment of RNA base pairing probability matrices. *Bioinformatics* 2004;**20**(14): 2222–7.
16. Bradley RK, Pachter L, Holmes I. Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics* 2008;**24**(23):2677–83.
17. Do CB, Foo C-SS, Batzoglou S. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics* 2008;**24**(13):i68–76.
18. Sato K, Kato Y, Akutsu T, *et al*. DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. *Bioinformatics* 2012;**28**(24):3218–24.
19. Will S, Reiche K, Hofacker IL, *et al*. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* 2007;**3**(4):e65.
20. Will S, Otto C, Miladi M, *et al*. SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics* 2015;**31**(15): 2489–96.
21. Otto C, Möhl M, Heyne S, *et al*. ExpaRNA-P: simultaneous exact pattern matching and folding of RNAs. *BMC Bioinformatics* 2014;**15**:404.
22. Bourgeade L, Chauve C, Allali J. Chaining sequence/structure seeds for computing RNA similarity. *J Comput Biol* 2015;**22**(3):205–17.
23. Griffiths-Jones S, Bateman A, Marshall M, *et al*. Rfam: an RNA family database. *Nucleic Acids Res* 2003;**31**(1):439–41.