Sequence analysis

PASTA for proteins

Kodi Collins¹ and Tandy Warnow²,*

¹Department of Computer Science, University of California, Los Angeles, CA 90025, USA and ²Department of Computer Science, University of Illinois, Urbana, IL 61866, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on April 2, 2018; revised on June 13, 2018; editorial decision on June 14, 2018; accepted on June 15, 2018

Abstract

Summary: PASTA is a multiple sequence method that uses divide-and-conquer plus iteration to enable base alignment methods to scale with high accuracy to large sequence datasets. By default, PASTA included MAFFT L-INS-i; our new extension of PASTA enables the use of MAFFT G-INS-i, MAFFT Homologs, CONTRAlign and ProbCons. We analyzed the performance of each base method and PASTA using these base methods on 224 datasets from BAliBASE 4 with at least 50 sequences. We show that PASTA enables the most accurate base methods to scale to larger datasets at reduced computational effort, and generally improves alignment and tree accuracy on the largest BAliBASE datasets.

Availability and implementation: PASTA is available at https://github.com/kodicollins/pasta and has also been integrated into the original PASTA repository at https://github.com/smirarab/pasta.

Contact: warnow@illinois.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Protein multiple sequence alignment is a key first step in much biological research, including protein structure and function prediction, domain identification, inference of ancestral proteins and construction of protein–protein interaction networks. However, alignment error can have a substantial impact on the downstream analyses and large datasets can be particularly difficult to align with high accuracy. For these reasons, among others, there is a great interest in the development of new protein sequence alignment methods that can provide good accuracy on large datasets (Iantorno et al., 2014; Le et al., 2017).

2 Materials and methods

PASTA (Mirarab et al., 2015) is a method that was designed to improve the accuracy and scalability of a base method for multiple sequence alignment (Abuìn et al., 2017). PASTA computes an initial tree, and then iterates between alignment estimation and tree estimation, typically performing three iterations. Each iteration uses the selected base method to compute alignments on small subsets with at most 200 sequences, and then merges those alignments into an alignment on the full dataset. Once the full alignment is computed, a maximum likelihood tree is computed using FastTree-2 (Price et al., 2010). The standard version of PASTA enables only a few base methods; here, we explore the impact of including other base methods for protein multiple sequence alignment. In addition to the use of MAFFT L-INS-i as the subset aligner in PASTA, we include two ways of running MAFFT version 7.149b: G-INS-i and Homologs (Katoh and Standley, 2013), CONTRAlign version 1.04 (Do et al., 2006) and ProbCons 1.12 (Do et al., 2005). The public distribution of MAFFT Homologs is limited to 99 sequences, and we turned off the flag restricting its analysis to small datasets to enable it to analyze larger datasets.

3 Results

We explored accuracy and running time on BAliBASE (Thompson et al., 1999), a collection of protein sequences with reference alignments based on structural features, restricted to datasets with at least 50 or more sequences. These 224 datasets have between 50 and 807 sequences. When the input to PASTA is at most 200 sequences, it decomposes the dataset into two subsets; otherwise, PASTA decomposes into subsets with at most 200 sequences. We compare base alignment methods (several variants of MAFFT, ProbCons and
CONTRAlign (variant of MAFFT and then by MAFFT L-INS-i). PASTA improves the TC scores for most methods and slightly decreases the TC scores for the two best-performing methods. When the PID is low, however, TC scores drop and the relative performance between methods changes. Here, the best average TC scores are obtained using PASTA+MAFFT G-INS-i, followed closely by MAFFT G-INS-i; CONTRAlign is in third place and PASTA+ProbCons and PASTA+CONTRAlign are nearly tied and in fourth place. The lowest average TC scores are obtained by ProbCons. For these harder datasets (PID < 25%), the impact of PASTA is variable—sometimes improving scores and sometimes reducing scores, but when it reduces scores the reductions are small.

We also compared methods with respect to the accuracy of maximum likelihood trees computed on their alignments, using the eight large RV10 BAliBASE datasets. CONTRAlign and ProbCons had the lowest accuracy of all methods, but using PASTA improved the accuracy substantially; all other methods had similar average accuracy on these datasets (Fig. 1D).

4 Conclusions

This study shows that PASTA can be used to improve the scalability of several protein alignment methods. The optimal choice of PASTA variant (i.e. sub-aligner) depends on the properties of the dataset, but PASTA reduces the running time of ProbCons and CONTRAlign and improves the TC scores and tree accuracy these methods obtain on large datasets.

Acknowledgement

The authors thank Mike Nute for helpful discussions during this project.

Funding

This work was supported by the National Science Foundation [ABI-1458652 to T.W.]. K.C. was also supported by the Eugene V. Cota-Robles Fellowship at UCLA. This research was performed on the National Center of Supercomputing Applications Blue Waters supercomputer and also on the Campus Cluster platform at UIUC.

Conflict of Interest: none declared.

References

Abuin, J.M. et al. (2017) PASTAspark: multiple sequence alignment meets Big Data. Bioinformatics, 33, 2948–2950.

Do, C.B. et al. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res., 15, 330–340.

Do, C.B. et al. (2006) CONTRAlign: discriminative training for protein sequence alignment. In: Proceedings of 10th Annual International Conference on Research in Computational Molecular Biology, RECOMB, Venice, Italy. Springer, Berlin, Heidelberg, pp. 160–174.
Iantorno, S. et al. (2014) Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. In: *Multiple Sequence Alignment Methods*. Humana Press, Totowa, NJ, pp. 59–73.

Katoh, K. and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30, 772–780.

Le, Q. et al. (2017) Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics*, 33, 1331–1337.

Liu, K. et al. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324, 1561–1564.

Mirarab, S. et al. (2015) PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.*, 22, 377–386.

Nguyen, N. et al. (2015) Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.*, 16, 124.

Price, M. N. et al. (2010) Fasttree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, 5, e9490.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22, 2688–2690.

Thompson, J. et al. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15, 87–88.