

# An improved mode of running PASTA

Qikai Yang<sup>1</sup> and Tandy Warnow<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Illinois at Urbana-Champaign

**Abstract.** PASTA is a method for estimating alignments and trees that has been able to provide excellent accuracy on large sequence datasets. By design, PASTA operates using iteration, in which the tree from the previous iteration is used to inform a divide-and-conquer strategy during which a new alignment is computed on the sequence dataset, and then a new maximum likelihood tree is estimated on the new alignment. In its default setting, PASTA runs for three iterations and returns that alignment/tree pair from the last iteration. Here we use both biological and simulated nucleotide datasets to show that returning the alignment/tree pair that has the best maximum likelihood score improves on the default usage.

## 1 Introduction

Multiple sequence alignment (MSA) is a fundamental step in many biological studies, and large-scale MSA is particularly difficult. Several methods have been designed for large-scale MSA, including Clustal-Omega (Sievers et al., 2011), KAlign3 (Lassmann, 2019), PASTA (Mirarab et al., 2015), and UPP (Nguyen et al., 2015). Furthermore, sequence length heterogeneity also presents challenges for alignment (Smirnov and Warnow, 2020), a problem that is addressed by UPP. Specifically, when given a dataset with sequence length heterogeneity, UPP extracts a set of sequences that are “full length” and uses PASTA to align them, thus forming the “backbone alignment”. The remaining sequences are then added into the backbone alignment using an ensemble of Hidden Markov Models (Durbin et al., 1998). Thus, UPP is an extension of PASTA to enable the accurate estimation of alignments given sequence length heterogeneity. Hence, here we focus on PASTA, due to its high accuracy on large datasets with high rates of evolution and applicability to both nucleotide alignment and protein alignment.

PASTA uses a combination of divide-and-conquer and iteration, so that each iteration operates by taking the alignment and tree from the previous iteration, dividing the sequence input into subsets using the tree, aligning the subsets using the selected base method, and then merging the alignments together. A maximum likelihood tree is then estimated on the alignment using FastTree2 (Price et al., 2010). This process repeats for several iterations (default 3), and

then the last alignment/tree pair is returned. While this default usage has performed well in many studies, here we explore the potential for additional improvement in accuracy by returning the alignment/tree pair that had the best maximum likelihood score, as computed using FastTree2.

## 2 Methods and materials

### 2.1 Methods

We compare two variants of PASTA, the default version and a variant where we return the alignment/tree pair with the best maximum likelihood score, on a collection of biological and simulated datasets with up to 1000 sequences, each run for four iterations, and we report the impact on alignment error. All datasets are available in public repositories from prior studies.

PASTA-BestML differs from the default settings for PASTA in two ways: it does not mask sites that are highly gappy (as is performed in PASTA-Default, in order to speed up the analysis on large datasets) and it returns the alignment/tree pair that has the best maximum likelihood score across all the reported iterations (instead of the last alignment/tree pair). Since PASTA has some randomness in its execution, we ran several independent experiments on each dataset. Therefore, we ran 100 independent experiments for each biological dataset and 50 independent experiments for each replicate of each simulated dataset.

### 2.2 Datasets

We include both biological nucleotide datasets and simulated nucleotide datasets from prior studies. We use a standard protocol to preprocess these datasets to reduce to a set of sequences that are full-length: we discard all sequences whose length is different from the median sequence length by over 20%. We use two sources for simulated datasets: the 1000- and 500-sequence datasets used in Liu et al. (2009, 2011) to evaluate SATé and SATé-II in comparison to other methods, which were simulated using ROSE (Stoye et al., 1998), and some subsets of the RNASim dataset (studied in Mirarab et al. (2015)) with 1000 sequences; the ROSE datasets are available at Mirarab (2020) and the RNASim datasets are available at Smirnov (2020).

All the simulated datasets evolve with substitutions and indels, but differ from each other in various ways. The ROSE datasets from Mirarab (2020) evolve under a modification of the GTRGAMMA model to allow for indels, and the RNASim datasets evolve with substitutions and indels under a biophysical model that reflects the selective pressure of RNA structure conservation (see Mirarab et al. (2015)). The ROSE datasets include a wider range of sequence evolution and thus include some conditions that are challenging to align, while the RNASim datasets are also challenging due to highly variable rates of evolution across the length of the sequences. Each of the simulated datasets has

20 replicates. We include six biological nucleotide datasets from the Comparative Ribosomal Website (CRW) (Cannone et al., 2002), available from Mirarab (2020). They are 16S.M, 16S.M.aa\_ag, 23S.M, 23S.M.aa\_ag, 23S.E and 23S.E.aa\_ag; these range in size from 96 to 740 sequences, and from 930 nucleotides to 3599 nucleotides in average sequence length. See Tables 1 and 2 for empirical statistics about these datasets.

## 2.3 Criteria

We report alignment error using SPFP and SPFN, calculated by FASTSP (Mirarab and Warnow, 2011), where SPFN is the percentage of the pairwise homologies in the reference alignment that do not appear in the estimated alignment and SPFP represents the percentage of the homologies in the estimated alignment that are missing from the reference alignment.

## 2.4 Dataset Information

Datasets	Average p-distance	Average seq. length	# sequences (after screening)	# sequences (before screening)
16S.M	0.298	947	740	901
16S.M.aa_ag	0.300	930	633	1028
23S.E	0.283	3599	99	117
23S.E.aa_ag	0.284	3531	96	144
23S.M	0.326	1580	211	278
23S.M.aa_ag	0.339	1578	197	263

Table 1: Information on biological nucleotide datasets from Mirarab (2020), and originally from Cannone et al. (2002). The p-distance of a pair of aligned sequences was calculated by dividing the number of sites where the two sequences had different nucleotides by the number of sites in which both sequences had nucleotides.

## 3 Results and Discussion

Datasets	Average p-distance	Average seq. length	Average # sequences (after screening)	Average # sequences (before screening)
ROSE-1000L1	0.695	1015	1000	1000
ROSE-1000L2	0.696	1007	1000	1000
ROSE-1000L3	0.687	1031	1000	1000
ROSE-1000L4	0.500	1007	1000	1000
ROSE-1000L5	0.496	1006	1000	1000
ROSE-1000M1	0.695	1011	1000	1000
ROSE-1000M2	0.684	1014	1000	1000
ROSE-1000M3	0.660	1007	1000	1000
ROSE-1000M4	0.495	1006	1000	1000
ROSE-1000M5	0.499	1003	1000	1000
ROSE-1000S1	0.694	1002	1000	1000
ROSE-1000S2	0.693	1001	1000	1000
ROSE-1000S3	0.686	1002	1000	1000
ROSE-1000S4	0.501	1000	1000	1000
ROSE-1000S5	0.498	1000	1000	1000
ROSE-500L1	0.670	1042	500	500
ROSE-500L2	0.657	1037	500	500
ROSE-500L3	0.658	1023	500	500
ROSE-500L4	0.499	1023	500	500
ROSE-500L5	0.497	1010	500	500
ROSE-500M1	0.674	1017	500	500
ROSE-500M2	0.658	1018	500	500
ROSE-500M3	0.657	1009	500	500
ROSE-500M4	0.491	1011	500	500
ROSE-500M5	0.495	1004	500	500
ROSE-500S1	0.673	1004	500	500
ROSE-500S2	0.655	1004	500	500
ROSE-500S3	0.656	1002	500	500
ROSE-500S4	0.492	1002	500	500
ROSE-500S5	0.498	1000	500	500
RNASim-1000	0.411	1555	1000	1000
RNASim-1000_C_100	0.317	1551	1000	1000
RNASim-1000_C_500	0.378	1555	1000	1000

Table 2: Information on simulated nucleotide ROSE datasets from Mirarab (2020) and RNASim datasets from Smirnov (2020).

	SPFP		SPFN	
	Default	BestML	Default	BestML
16S.M	14.22%	<b>14.01%</b>	12.54%	<b>12.31%</b>
16S.M.aa_ag	13.20%	<b>13.17%</b>	11.46%	<b>11.43%</b>
23S.E	21.04%	<b>20.61%</b>	17.03%	<b>16.58%</b>
23S.E.aa_ag	20.39%	<b>20.14%</b>	16.41%	<b>16.12%</b>
23S.M	22.72%	<b>22.67%</b>	17.80%	<b>17.71%</b>
23S.M.aa_ag	24.12%	<b>24.06%</b>	18.70%	<b>18.64%</b>

Table 3: Comparison between PASTA’s default mode and PASTA’s BestML mode for alignment error (SPFP/SPFN) on biological nucleotide datasets. For each dataset and criterion, the best result is boldfaced.

	SPFP		SPFN	
	Default	BestML	Default	BestML
ROSE-1000L1	8.48%	<b>7.54%</b>	8.68%	<b>7.75%</b>
ROSE-1000L2	3.22%	<b>3.08%</b>	3.29%	<b>3.14%</b>
ROSE-1000L3	16.36%	<b>15.66%</b>	16.84%	<b>16.15%</b>
ROSE-1000L4	0.60%	<b>0.59%</b>	0.60%	<b>0.59%</b>
ROSE-1000L5	<b>0.39%</b>	0.40%	<b>0.39%</b>	0.40%
ROSE-1000M1	18.18%	<b>17.28%</b>	18.85%	<b>17.95%</b>
ROSE-1000M2	13.98%	<b>13.40%</b>	14.34%	<b>13.76%</b>
ROSE-1000M3	4.81%	<b>4.67%</b>	4.82%	<b>4.69%</b>
ROSE-1000M4	<b>1.01%</b>	1.02%	<b>1.00%</b>	1.01%
ROSE-1000M5	0.60%	<b>0.59%</b>	0.60%	<b>0.59%</b>
ROSE-1000S1	14.99%	<b>13.91%</b>	15.32%	<b>14.25%</b>
ROSE-1000S2	7.52%	<b>6.83%</b>	7.58%	<b>6.90%</b>
ROSE-1000S3	5.82%	<b>5.45%</b>	5.85%	<b>5.48%</b>
ROSE-1000S4	0.54%	<b>0.53%</b>	0.54%	<b>0.53%</b>
ROSE-1000S5	0.21%	<b>0.20%</b>	0.21%	<b>0.20%</b>
ROSE-500L1	18.07%	<b>17.64%</b>	18.71%	<b>18.29%</b>
ROSE-500L2	14.70%	<b>14.40%</b>	15.04%	<b>14.73%</b>
ROSE-500L3	6.77%	<b>6.59%</b>	6.79%	<b>6.60%</b>
ROSE-500L4	1.67%	<b>1.65%</b>	1.64%	<b>1.62%</b>
ROSE-500L5	0.74%	<b>0.73%</b>	0.72%	<b>0.71%</b>
ROSE-500M1	17.02%	<b>16.52%</b>	17.73%	<b>17.18%</b>
ROSE-500M2	12.27%	<b>12.05%</b>	12.63%	<b>12.41%</b>
ROSE-500M3	6.20%	<b>6.14%</b>	6.27%	<b>6.22%</b>
ROSE-500M4	1.38%	<b>1.36%</b>	1.37%	<b>1.35%</b>
ROSE-500M5	0.76%	<b>0.73%</b>	0.75%	<b>0.73%</b>
ROSE-500S1	15.89%	<b>15.28%</b>	16.26%	<b>15.62%</b>
ROSE-500S2	11.05%	<b>10.74%</b>	11.21%	<b>10.88%</b>
ROSE-500S3	5.28%	<b>5.07%</b>	5.30%	<b>5.09%</b>
ROSE-500S4	1.70%	<b>1.69%</b>	1.68%	<b>1.67%</b>
ROSE-500S5	0.74%	<b>0.73%</b>	<b>0.73%</b>	<b>0.73%</b>
RNASim-1000	10.08%	<b>9.90%</b>	9.97%	<b>9.78%</b>
RNASim-1000_C_100	5.73%	<b>5.34%</b>	5.66%	<b>5.28%</b>
RNASim-1000_C_500	<b>7.59%</b>	7.73%	<b>7.44%</b>	7.59%

Table 4: Comparison between PASTA’s default mode and PASTA’s BestML mode for alignment error (using SPFP/SPFN) on simulated nucleotide datasets, taken from Mirarab (2020) (simulated using ROSE) and Smirnov (2020) (simulated using RNASim). For each model condition and criterion, the lower error rate is boldfaced at each row.

As seen in Table 3 and Table 4, PASTA-BestML has lower alignment error than PASTA-Default for all biological datasets and nearly all simulated datasets. Thus, running PASTA with BestML typically improves alignment accuracy compared to default PASTA. The difference is generally small, but because it is nearly universal, it suggests that PASTA-BestML may be a more reliable technique for using PASTA than its default mode. This improvement also suggests that the maximum likelihood score may be a valuable criterion to use when selecting an alignment from a set of candidate alignments.

## 4 Conclusions

Multiple sequence alignment is a basic step in many bioinformatics pipelines, with phylogeny estimation one of the applications of interest. Here we have shown that a small change to the PASTA pipeline yields a consistent improvement in alignment accuracy on a collection of biological and simulated datasets. Although the improvement was small, it is suggestive of a trend that could lead to bigger improvements on larger datasets. More generally, it suggests also the possibility of using criteria, such as maximum likelihood scores, to select between competing multiple sequence alignments computed for the same dataset. In general, future work is needed in order to better understand how to estimate multiple sequence alignments.

## Acknowledgments

This work was performed while the first author was an undergraduate, and participating in a Research Experience for Undergraduates program in the Department of Computer Science at the University of Illinois at Urbana-Champaign.

## Funding

This work was supported by the National Science Foundation [1458652 and 1513629 to TW].

## References

- Cannone, J. J., S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D’Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, et al. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics* 3:2.
- Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press.
- Lassmann, T. 2019. Kalign 3: multiple sequence alignment of large datasets. *Bioinformatics* 36:1928–1929.
- Liu, K., S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324:1561–1564.
- Liu, K., T. J. Warnow, M. T. Holder, S. M. Nelesen, J. Yu, A. P. Stamatakis, and C. R. Linder. 2011. SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* 61:90–106.
- Mirarab, S. 2020. Public datasets. <https://sites.google.com/eng.ucsd.edu/datasets/home?authuser=0>, retrieved August 20, 2020.
- Mirarab, S., N. Nguyen, S. Guo, L.-S. Wang, J. Kim, and T. Warnow. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology* 22:377–386 PMID: 25549288.
- Mirarab, S. and T. Warnow. 2011. FASTSP: linear time calculation of alignment accuracy. *Bioinformatics* 27:3250–3258.
- Nguyen, N.-P. D., S. Mirarab, K. Kumar, and T. Warnow. 2015. Ultra-large alignments using phylogeny-aware profiles. *Genome Biology* 16.
- Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE* 5:1–10.
- Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* 7:539.
- Smirnov, V. 2020. Vlad’s Datasets. University of Illinois at Urbana-Champaign <https://databank.illinois.edu/datasets/IDB-6955387>.
- Smirnov, V. and T. Warnow. 2020. Phylogeny estimation given sequence length heterogeneity. *Systematic Biology* Doi: 10.1093/sysbio/syaa058.
- Stoye, J., D. Evers, and F. Meyer. 1998. Rose: generating sequence families. *Bioinformatics* 14:157–163.