EMMA: Adding Sequences into a Constraint Alignment with High Accuracy and Scalability

Chengze Shen
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Baqiao Liu
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Kelly P. Williams
Sandia National Laboratories, Livermore, CA, USA

Tandy Warnow[1]
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Abstract

Multiple sequence alignment (MSA) is a crucial precursor to many downstream biological analyses, such as phylogeny estimation [4], RNA structure prediction [8], protein structure prediction [1], etc. Obtaining an accurate MSA can be challenging, especially when the dataset is large (i.e., more than 1000 sequences). A key technique for large-scale MSA estimation is to add sequences into an existing alignment. For example, biological knowledge can be used to form a reference alignment on a subset of the sequences, and then the remaining sequences can be added to the reference alignment. Another case where adding sequences into an existing alignment occurs is when new sequences or genomes are added to databases, leading to the opportunity to add the new sequences for each gene in the genome into a growing alignment. A third case is for *de novo* multiple sequence alignment, where a subset of the sequences is selected and aligned, and then the remaining sequences are added into this "backbone alignment" [5, 6, 10, 3, 7, 11]. Thus, adding sequences into existing alignments is a natural problem with multiple applications to biological sequence analysis.

A few methods have been developed to add sequences into an existing alignment, with MAFFT--add [2] perhaps the most well-known. However, several multiple sequence alignment methods that operate in two steps (first extract and align the backbone sequences and then add the remaining sequences into this backbone alignment) also provide utilities for adding sequences into a user-provided alignment. We present EMMA, a new approach for adding "query" sequences into an existing "constraint" alignment. By construction, EMMA never changes the constraint alignment, except through the introduction of additional sites to represent homologies between the query sequences. EMMA uses a divide-and-conquer technique combined with MAFFT--add (using the most accurate setting, MAFFT-linsi--add) to add sequences into a user-provided alignment. We evaluate EMMA by comparing it to MAFFT-linsi--add, MAFFT--add (the default setting), and WITCH-ng-add. We include a range of biological and simulated datasets (nucleotides and proteins) ranging in size from 1000 to almost 200,000 sequences and evaluate alignment accuracy and scalability. MAFFT-linsi--add was the slowest and least scalable method, only able to run on datasets with at most 1000 sequences in this study, but had excellent accuracy (often the best) on those datasets. We also see that EMMA has better recall than WITCH-ng-add and MAFFT--add on large datasets, especially when the backbone alignment is small or clade-based.

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases Multiple sequence alignment, constraint alignment, MAFFT

Digital Object Identifier 10.4230/LIPIcs.WABI.2023.2

Category Abstract

Related Version *Full Version*: https://doi.org/10.1101/2023.06.12.544642 [9]

[1] corresponding author

Supplementary Material bioRxiv paper has additional supplementary materials
Software (Source Code): `https://github.com/c5shen/EMMA`
archived at `swh:1:dir:d3da832ddc0eb1adb17fbfee398750b89da20544`
Dataset: `https://doi.org/10.13012/B2IDB-2567453_V1`
Dataset: `https://doi.org/10.13012/B2IDB-3974819_V1`

Funding This work was funded in part by the US NSF grant 2006069 and by the Laboratory Directed Research and Development program at Sandia National Laboratories, which is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

1 John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. Number: 7873 Publisher: Nature Publishing Group. `doi:10.1038/s41586-021-03819-2`.

2 Kazutaka Katoh and Martin C. Frith. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics*, 28(23):3144–3146, December 2012. `doi:10.1093/bioinformatics/bts578`.

3 Baqiao Liu and Tandy Warnow. WITCH-NG: efficient and accurate alignment of datasets with sequence length heterogeneity. *Bioinformatics Advances*, 3(1):vbad024, January 2023. `doi:10.1093/bioadv/vbad024`.

4 David A Morrison. Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*, 19(6):479–539, 2006.

5 Nam-phuong D. Nguyen, Siavash Mirarab, Keerthana Kumar, and Tandy Warnow. Ultra-large alignments using phylogeny-aware profiles. *Genome Biology*, 16(1):124, June 2015. `doi:10.1186/s13059-015-0688-z`.

6 Minhyuk Park, Stefan Ivanovic, Gillian Chu, Chengze Shen, and Tandy Warnow. UPP2: fast and accurate alignment of datasets with fragmentary sequences. *Bioinformatics*, 39(1):btad007, January 2023. `doi:10.1093/bioinformatics/btad007`.

7 Minhyuk Park and Tandy Warnow. HMMerge: an ensemble method for multiple sequence alignment. *Bioinformatics Advances*, page vbad052, 2023.

8 Bruce A Shapiro, Yaroslava G Yingling, Wojciech Kasprzak, and Eckart Bindewald. Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology*, 17(2):157–165, 2007.

9 Chengze Shen, Baqiao Liu, Kelly P Williams, and Tandy Warnow. Computing multiple sequence alignments given a constraint subset alignment using EMMA. *bioRxiv*, 2023. `doi:10.1101/2023.06.12.544642`.

10 Chengze Shen, Minhyuk Park, and Tandy Warnow. WITCH: Improved Multiple Sequence Alignment Through Weighted Consensus Hidden Markov Model Alignment. *Journal of Computational Biology*, May 2022. Publisher: Mary Ann Liebert, Inc., publishers. `doi:10.1089/cmb.2021.0585`.

11 Kazunori D. Yamada, Kentaro Tomii, and Kazutaka Katoh. Application of the MAFFT sequence alignment program to large data?reexamination of the usefulness of chained guide trees. *Bioinformatics*, 32(21):3246–3251, November 2016. `doi:10.1093/bioinformatics/btw412`.