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
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-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

EMMA: Adding Sequences to Constraint Alignments

Supplementary Material bioRxiv paper has additional supplementary materials
Software (Source Code): https://github.com/c5shen/EMMA
archived at swh:1:dir:d3da832ddc0eb1adb17fbee398750b89da20544
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


