Sequence analysis

Parallelization of MAFFT for large-scale multiple sequence alignments

Tsukasa Nakamura\textsuperscript{1,2}, Kazunori D. Yamada\textsuperscript{2,3}, Kentaro Tomii\textsuperscript{1,2,4,5} and Kazutaka Katoh\textsuperscript{2,6,*}

\textsuperscript{1}Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-8562, Japan, \textsuperscript{2}Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan, \textsuperscript{3}Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan, \textsuperscript{4}Biotechnology Research Institute for Drug Discovery (BRD), AIST, Tokyo 135-0064, Japan, \textsuperscript{5}AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL), Tokyo 152-8550, Japan and \textsuperscript{6}Research Institute for Microbial Diseases, Osaka University, Suita 565-0871, Japan

*To whom correspondence should be addressed.

Received on October 19, 2017; revised on February 7, 2018; editorial decision on February 24, 2018; accepted on February 28, 2018

Abstract

Summary: We report an update for the MAFFT multiple sequence alignment program to enable parallel calculation of large numbers of sequences. The G-INS-1 option of MAFFT was recently reported to have higher accuracy than other methods for large data, but this method has been impractical for most large-scale analyses, due to the requirement of large computational resources. We introduce a scalable variant, G-large-INS-1, which has equivalent accuracy to G-INS-1 and is applicable to 50 000 or more sequences.

Availability and implementation: This feature is available in MAFFT versions 7.355 or later at \url{https://mafft.cbrc.jp/alignment/software/mpi.html}.

Contact: katoh@ifrec.osaka-u.ac.jp

Supplementary information: Supplementary data are available at \textit{Bioinformatics} online.

A large number of biological sequences from widely divergent organisms are becoming available. Accordingly, the need for multiple alignments of large numbers of sequences is increasing for various kinds of sequence analysis. The G-INS-1 option of MAFFT was recently reported to have higher accuracy than other methods for large data, but this method has been impractical for most large-scale analyses, due to the requirement of large computational resources. We introduce a scalable variant, G-large-INS-1, which has equivalent accuracy to G-INS-1 and is applicable to 50 000 or more sequences. Our strategies to reduce computational costs are (i) parallelization across multiple machines and/or processor cores using MPI and Pthreads to increase speed and (ii) the use of a high-speed shared filesystem, which is becoming common for processing big data. An MPI-based parallelization of another high-accuracy MSA method, MSAProbs, was recently released (González-Domínguez et al., 2016), but it cannot be applied to thousands of sequences. The present update of MAFFT is designed to satisfy the need for accurately aligning large numbers of sequences but is not applicable to long genomic sequences since the length dependence of the computational cost is unchanged. The G-large-INS-1 option is available in MAFFT versions 7.355 or later and the online service (Katoh et al., 2017).

Accuracy of G-large-INS-1 was compared with that of conventional G-INS-1 using different benchmarks, QuanTest (Le et al., 2017) (Fig. 1a), HomFam (Sievers et al., 2011), OXFam (Raghava et al., 2003; Yamada et al., 2016) and ContTest (Fox et al., 2016) (Supplementary Table S1). Both methods ran with different input orders and/or minor variations in pairwise alignment and guide tree

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
To assess instability of accuracy scores, QuanTest was used to compare G-INS-1 (version 7.245, blue bold lines) with other popular methods. We used 1940 entries from the SEED alignment in Silva (Glockner et al., 2017) and 50157 sequences from the ‘sdr’ family taken from HomFam, to predict protein secondary structure. Accuracy was calculated based on the number of correct predictions per sequence, and the accuracy improved as the number of sequences increased. For LSU rRNA sequences, the wall-clock time for the all-to-all alignment stage decreased almost linearly with the number of cores used for the calculation. However, for datasets with very short sequences, the efficiency varied depending on the filesystem: high in Lustre (magenta) but low in NFS (green). This variation is due to the balance between calculation and disk operations. As noted earlier, a considerable amount of temporary data is written in parallel into the filesystem, taking a significant amount of time and disk resources.

For LSU rRNA sequences (b, 1521–4102 bases, 1000 sequences randomly selected from the SEED alignment in Silva (Glockner et al., 2017) and 50157 sequences from the ‘sdr’ family taken from HomFam), the wall-clock time for the all-to-all alignment stage decreased almost linearly with the number of cores used for the calculation. However, for a dataset with very short sequences (f, 12–35 amino acids, 88 345 sequences, the ‘zf-CCHH’ family taken from HomFam), the efficiency differed depending on the filesystem: high in Lustre (magenta) but low in NFS (green). This difference is due to the balance between calculation and disk operations. As noted earlier, a considerable amount of temporary data is written in parallel into the filesystem, taking a significant amount of time and disk resources.

Figure 1c, e and g suggest that the wall-clock time of the progressive stage varies for each run and does not linearly decrease, but usually this is not a speed-limiting step. CPU time and wall-clock time for various problems are shown in Supplementary Table S1.
increase in accuracy observed in Figure 1a for more than 200 sequences is due to the prediction phase not due to the alignment phase (see the last section in Supplementary Data and black dashed lines in Supplementary Fig. S1). As a result, it was difficult to know how many sequences should be included in an MSA. With more sequences, the MSA has richer comparative information, but the alignment quality is expected to decrease. The optimal balance between these two factors may differ by case. In contrast, the accuracy of G-large-INS-1 and G-INS-1 (red and blue dashed lines in Supplementary Fig. S1) was robust to data size in this test. The number of sequences to include in the MSA can now be determined simply based on the computational resources available and the requirements for the downstream analysis.

Acknowledgements
The authors thank Daron M. Standley and John Rozewicki, Osaka University, and Shun Sakuraba, the University of Tokyo, for discussion and computational support. The NIG supercomputer at ROIS National Institute of Genetics and the Reedbush System in the Information Technology Center, the University of Tokyo, were used.

Funding
This work was supported by JSPS KAKENHI [grant numbers JP16K07464 (to K.D.Y., K.T. and K.K.) and JP17K06457 (to T.N.)] and Platform Project for Supporting Drug Discovery and Life Science Research [grant numbers JP17am0101110 (to T.N., K.D.Y. and K.T.) and JP17am0101108 (to K.K.)] from AMED, Japan.

Conflict of Interest: none declared.

References
Boyce,K. et al. (2015) Instability in progressive multiple sequence alignment algorithms. Algorithms Mol Biol, 10, 26.
Drozdetskiy,A. et al. (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res., 43, W389–W394.
Fox,G. et al. (2016) Using de novo protein structure predictions to measure the quality of very large multiple sequence alignments. Bioinformatics, 32, 814–820.
Glockner,F.O. et al. (2017) 25 years of serving the community with ribosomal RNA gene reference databases and tools. J. Biotechnol., 261, 169–176.
González-Dominguez,J. et al. (2016) MSAPros-MPI: parallel multiple sequence aligner for distributed-memory systems. Bioinformatics, 32, 3826–3828.
Gudys,A. and Deorowicz,S. (2017) QuickProbs 2: towards rapid construction of high-quality alignments of large protein families. Sci. Rep., 7, 41553.
Katoh,K. et al. (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinformatics, in press.
Le,Q. et al. (2017) Protein multiple sequence alignment benchmarking through secondary structure prediction. Bioinformatics, 33, 1331–1337.
Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., 48, 443–453.
Notredame,C. et al. (1998) COFFEE: an objective function for multiple sequence alignments. Bioinformatics, 14, 407–422.
Raghava,G.P.S. et al. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics, 4, 47.
Sievers,F. et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol., 7, 539.
Sievers,F. et al. (2013) Making automated multiple alignments of very large numbers of protein sequences. Bioinformatics, 29, 989–995.
Yamada,K.D. et al. (2016) Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. Bioinformatics, 32, 3246–3251.