*Bioinformatics*, 14/04/2023
Application Note Category

*Application Note*

# MSAProbs Parallelization for High-Performance Sequence Alignment: A Comparative Study on Multithreading and Time Measures

Babih V. De Burnay[1]

[1]UDC student, subject: HPC.

*To whom correspondence should be addressed.

## Abstract

**Motivation:** The alignment of sequences is an essential task in bioinformatics, and its computational complexity can often lead to a time-consuming process. Therefore, the parallelization of sequence alignment algorithms is necessary to improve performance and scalability. This study aims to compare the performance of MSAProbs, a popular software tool for sequence alignment, when parallelized with different numbers of threads, to determine its effectiveness in reducing the time required for sequence alignment.

**Results:** The results show how MSAProbs using parallelization has a great ability to reduce sequence alignment time. It is also observed that MSAProbs has a high scalability.

**Availability:** Virtual Campus.

**Contact:** babih.velazquez@udc.es

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

In the field of bioinformatics, one of the most common problems is comparing protein sequences. Sequence alignment is a key technique for this comparison and involves finding similarities between two or more protein sequences. Accurate alignment is essential for detecting homologies, predicting protein structures and functions, and understanding protein evolution (Zielezinski et al., 2017).

However, the task of aligning protein sequences can be extremely computationally expensive due to the enormous number of alignment possibilities. The computational complexity of the task increases significantly with the size of the sequences and the number of sequences to be aligned (Edgar & Batzoglou, 2006). In this context, process parallelization becomes an essential tool to reduce processing time and improve alignment task efficiency. Parallelization allows simultaneous execution of tasks on multiple cores or processors, which significantly reduces the time required to complete the process (Kleinjung et al., 2002).

MSAProbs is a program for multiple protein sequence alignment, developed by Liu et al. (2010). MSAProbs is based on the hidden Markov model (HMM) and uses a probabilistic approach to alignment. MSAProbs has been shown to be highly effective in identifying homologies, detecting conserved motifs, and predicting protein structures.

Nevertheless, the sequential version of MSAProbs can be very slow for processing large datasets. This is where the multithread version of MSAProbs plays a key role in accelerating the alignment process. The multithread version of MSAProbs uses multiple processing threads to divide the work into several simultaneous processes, resulting in a notable improvement in processing speed. The positive impact of this multithread approach on processing speed will be demonstrated and discussed in detail in the subsequent sections of this article, showing a significant improvement in alignment performance.

## 2 Materials and Methods

We present the use of the MSAProbs tool on the Finis Terrae supercomputer III, which is located in Santiago de Compostela, for multiple sequence alignment of protein sequences. It is necessary to highlight the importance of loading the ‘cesga/2020’ library to compile and run MSAProbs effectively on the supercomputer. The cesga/2020 library provide the necessary dependencies to effectively run tools such as BLAST, Python, R and in our case, MSAProbs.

In this study, we used three datasets of protein sequences obtained from the BAliBASE 3.0 benchmark, which is a widely used database for evaluating multiple sequence alignment tools. The BAliBASE 3.0 database was developed by Thompson et al. (2005) and contains a large collection of reference alignments and corresponding unaligned sequences for evaluating the accuracy of multiple sequence alignment tools. Specifically, the three datasets we used in the study contained protein sequences of different sizes, which we termed small, medium an large respectively (41 kB, 70 kB, and 110 kB). These datasets were used to evaluate the performance of MSAProbs in different alignment scenarios.

To evaluate the performance of MSAProbs on the Finis Terrae supercomputer, we based our assessment on the runtime metrics of the tool for aligning protein sequences using different numbers of threads. In particular, we used one thread to measure the time it takes to align the sequences sequentially, and then used 2, 4, 8, 16, 32, and 64 threads to evaluate the parallel performance of the tool.

In addition to evaluating the runtime metrics for MSAProbs on the Finis Terrae III supercomputer using different numbers of threads, we also analyzed the speedup obtained by the tool for each of the three datasets used in our study. The speedup, which is defined as the ratio of the time taken to run a task sequentially to the time taken to run the same task using parallel computing, is a measure of the improvement in performance achieved by using parallel computing compared to sequential computing. Is an important performance metric in the field of high-performance computing as it provides a quantitative measure of the benefits of using parallel computing.

## 3 Results

The results obtained for the three datasets are very good in terms of computing time. As can be seen in **Table 1.** the computing time decreases significantly as the number of threads used for the alignment is increased.

**Table 1.** Execution time (min:sec). Datasets vs Threads.

| Dataset \ Threads | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|---|
| Small | 17:38 | 8:19 | 4:28 | 2:20 | 1:20 | 0:44 | 0:30 |
| Medium | 71:36 | 34:24 | 17:37 | 9:49 | 4:54 | 2:53 | 1:43 |
| Large | 200:50 | 103:47 | 54:53 | 27:47 | 14:59 | 8:28 | 5:35 |

The table represents the execution time (minutes:seconds) of the different datasets to align the protein sequences as a function of the number of threads used.

We will now review what percentage of time we have reduced by parallelizing using 64 threads in each dataset.

- Small: From 17:38 min to 0:30 min → 97.16 %
- Medium: From 71:36 min to 1:43 min → 97.59 %
- Large: From 200:50 min to 5:35 min → 97.22 %

Parallelizing sequence alignment using 64 threads saves an average of 97.32% time compared to what it would have taken to do it sequentially.

The following graph **Fig. 1** shows the speedup results, the shape of the speedup graph can provide insights into the efficiency of parallel execution and the scalability of the program.

When using 2 threads, the speedup is greater than 2, indicating a superlinear acceleration. However, from 4 threads, it no longer exhibits superlinear behavior, approaching what would be considered perfect linear speedup until 16 threads. Beyond that, the speedup gradually decreases, but still yields satisfactory results.

In this case, the fact that the speedup is superlinear when using 2 threads and not with more than 2 indicates a manifestation of the complex interactions between the algorithm and the hardware used. It doesn't indicate poor parallel performance. However, as more threads are added, communication overhead between them can reduce the efficiency of parallel processing, decreasing the superlinear acceleration.

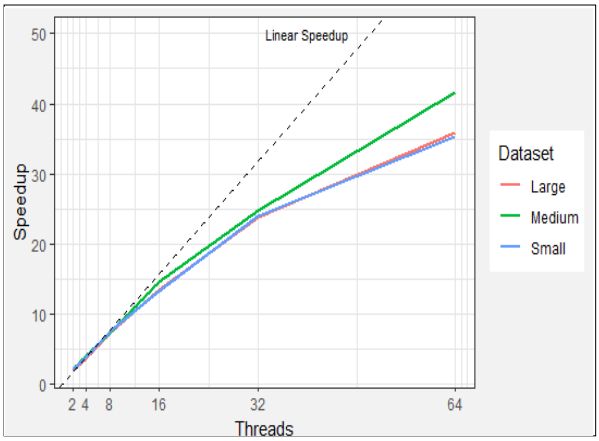


**Fig. 1. Relation between Speedup and Threads.**

The results obtained in **Fig. 1** suggest that MSAProbs has good scalability in terms of its parallel performance. Furthermore, the near-linear speedup observed with 4, 8, 16 and 32 threads indicate that MSAProbs can easily scale to accommodate an increasing number of threads and align even larger datasets. Although there was a slight decrease in the speedup slope with 32 and 64 threads, the achieved speedup values of 35 and 41 for the "large" & "small" and "medium" datasets respectively are still impressive.

## 4 Conclusion

In conclusion, MSAProbs is a software tool that can significantly reduce the time required to align sequences when multiple threads are used. The results of the scalability analysis showed that MSAProbs is capable of efficiently using the available computational resources and scaling well with increasing dataset size and number of threads.

## References

Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. Current Opinion in Structural Biology, 16(3), 368-373. https://doi.org/10.1016/j.sbi.2006.04.004

Kleinjung, J., Douglas, N., & Heringa, J. (2002). Parallelized multiple alignment. Bioinformatics, 18(9), 1270-1271. https://doi.org/10.1093/bioinformatics/18.9.1270

Liu, Y., Schmidt, B., & Maskell, D. L. (2010). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. Bioinformatics, 26(16), 1958-1964. https://doi.org/10.1093/bioinformatics/btq338

Thompson, J. D., Koehl, P., Ripp, R., & Poch, O. (2005). BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. Proteins, 61(1), 127-136. https://doi.org/10.1002/prot.20527

Zielezinski, A., Vinga, S., Almeida, J. S., & Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology, 18(1). https://doi.org/10.1186/s13059-017-1319-7