A parallel approach of COFFEE objective function to multiple sequence alignment

Journal of Physics: Conference Series **633** (2015) 012084 doi:10.1088/1742-6596/633/1/012084

A parallel approach of COFFEE objective function to multiple sequence alignment

G F D Zafalon[1], J M V Visotaky[1], A R Amorim[1], C R Valêncio[1], L A Neves[1], R C G de Souza[1], J M Machado[1]

[1] Department of Computer Science and Statistics (DCCE), São Paulo State University (UNESP), São José do Rio Preto, Brazil.

E-mail: zafalon@sjrp.unesp.br

Abstract. The computational tools to assist genomic analyzes show even more necessary due to fast increasing of data amount available. With high computational costs of deterministic algorithms for sequence alignments, many works concentrate their efforts in the development of heuristic approaches to multiple sequence alignments. However, the selection of an approach, which offers solutions with good biological significance and feasible execution time, is a great challenge. Thus, this work aims to show the parallelization of the processing steps of MSA-GA tool using multithread paradigm in the execution of COFFEE objective function. The standard objective function implemented in the tool is the Weighted Sum of Pairs (WSP), which produces some distortions in the final alignments when sequences sets with low similarity are aligned. Then, in studies previously performed we implemented the COFFEE objective function in the tool to smooth these distortions. Although the nature of COFFEE objective function implies in the increasing of execution time, this approach presents points, which can be executed in parallel. With the improvements implemented in this work, we can verify the execution time of new approach is 24% faster than the sequential approach with COFFEE. Moreover, the COFFEE multithreaded approach is more efficient than WSP, because besides it is slightly fast, its biological results are better.

1. Introduction

The biology has always been a study of the human object, trying to understand the functioning of living organisms and their possible relationships. So with the technological advances brought in recent decades, several studies have enabled a better understanding of living beings of all kinds, from the simplest to the most complex [1].

In this context, it is necessary to conduct experiments to prove the studies, which produce a large amount of information that need to be organized and arranged to become intelligible. This organization is not feasible using conventional means, ie a manual organization, emerging as important field of genetic computing resources, which led to Bioinformatics [2].

In this work, we proposed the implementation of parallel version of COFFEE objective function using multithread paradigm in the multiple sequence alignment tool MSA-GA. Thus, it is possible to reach improvements in execution time of the objective function without lost the biological significance of the alignments.

This work is organized as follows: in the section 2 a brief review about parallel sequence alignment is provided. In the section 3 are described the materials and methods with the special attention to the implementation of the objective function in the MSA-GA tool. Some analysis and results are presented in the section 4. Finally, in the section 5, the conclusions and future perspectives are presented.

2. Parallel Multiple Sequence Alignments

Bioinformatics has as primary goal to provide computational solutions to problems that are important to biology, assisting in the analysis of experimental data in search of biological identities between

sequences [1]. However, the difficulty lies in the fact that, given the huge amount of scheduled sequences, using deterministic techniques to treat them is highly unfeasible due to the computational cost [2]. Thus, as an alternative, we started to use heuristic techniques, especially for multiple sequence alignment.

It is interesting the development of tools that can mitigate the high cost in terms of execution time in bioinformatics problems**,** using the processing provided by parallelism [3**].** As an example, we can cite parallel versions of CLUSTALW [4]**,** the MAFFT [5]**,** T-COFFEE [6], among others.

3. Materials and Methods

The COFFEE is an objective function that works on a pairwise alignment reference library. The alignment of the evaluation is performed by positions. For each estimated alignment position, a scoring matrix is built, which is filled with the weights assigned to each pairwise alignment available in the library. In position (column) analyzed, each cell of the matrix corresponds to the alignment between two residues of the position. If the alignment between the two residues are found in the library, then is allocated to the cell the weight of the library, otherwise it is assigned the value 0. The score of the position is given by the sum of all values in the scoring matrix divided by the sum of weights of the alignments involved. Then the total score of alignment is calculated by the sum of the scores of all columns divided by the alignment length [7]. In the Figure 1 is presented the schematic calculation of COFFEE score.

To achieve greater gains in terms of execution time and use any idle computing resources in tournaments phase of the genetic algorithm, the implementation of COFFEE objective function was slightly altered to allow its calculations are performed simultaneously by the processing elements processing that are idle or incomplete recovery. As shown in Figure 1, the function is able to calculate multiple column scores simultaneously for alignment being evaluated, rather than just one at a time as it was in the original version. All obtained column scores are used to calculate the final score of the alignment.

4. Tests and Results

The tests were performed in a computer with Intel Xeon X3430@2.4 GHz Quad-Core processor, 16 GB of RAM memory and Windows 7 Home Premium 64-bits. Due the stochastic approach of genetic algorithms, we have executed each test case five times, in order to compare the execution time of the new approach with multithreaded COFFEE and the version with sequential COFFEE objective function. The result used in each case test, for comparison purposes, was the average of the five alignments execution times.

We have used the well-known test cases from BAliBase[1]. The sets we have chosen are divided into two main categories: the less similar ones (< 25% of identity and 20% ~ 40% of identity) and the more similar ones (> 35% of identity). Thus, in the Table 1 are shown the obtained execution times, given in seconds, of COFFEE Multithread (COFFEE-M) comparing with the sequential COFFEE (COFFEE-S) approach.

Thus, it can be noticed that multithread COFFEE approach was able to improve the execution time of alignments of the less similar test cases, when compared with the sequential COFFEE. The results obtained in the present work for these cases have achieved at least 7% of improvements and 28% at most, as showed in Figure 2 (a) and Figure 2 (b). Moreover, it can be noticed that the improvements here presented were also obtained for alignments of more similar sequences sets, as can be seen in Figure 2 (c). In these cases, the multithread COFFEE was at least 5% faster than sequential COFFEE and 29% at most.

We have calculated the maximum, minimum and the average standard deviation for COFFEE multithreaded and COFFEE sequential. For the multithreaded approach, the maximum, the minimum and the average were 1.61, 0.01 and 0.40, respectively.

[1] http://www.lbgi.fr/balibase/.

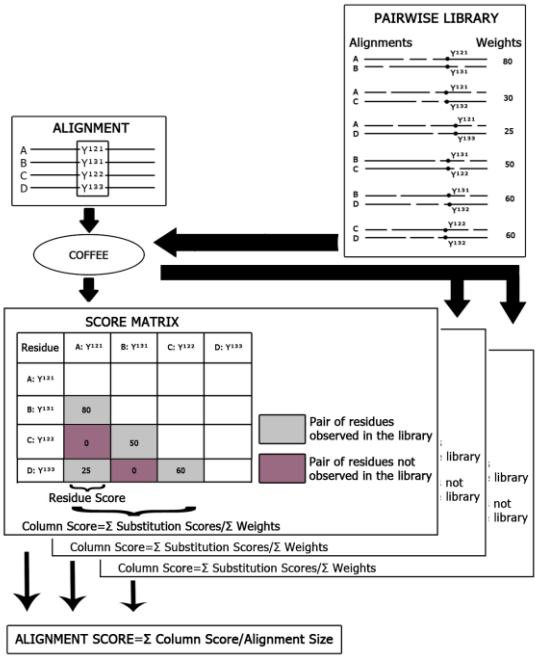

Figure 1. Operation mode of COFFEE objective function.

Table 1. Execution time of each test case.

Sequence Set	COFFEE-M (s)	COFFEE-S (s)	Sequence Set	COFFEE-M (s)	COFFEE-S (s)	Sequence Set	COFFEE-M (s)	COFFEE-S (s)
<25% identity			**20% ~ 40% identity**			**>35% identity**		
1idy	008,7	010,0	1ycc	012,1	013,2	1amk	034,4	046,8
1tvxA	008,7	009,4	1ad2	022,0	030,8	1aho	009,7	010,2
Kinase	052,4	067,6	1aym3	026,3	033,8	1csp	009,8	011,0
1r69	008,7	010,0	1fieA	091,4	121,4	1ar5A	019,8	023,6
1ubi	010,4	012,0	1ldg	034,6	044,4	1ad3	065,4	090,0
1wit	014,5	017,6	1sesA	093,7	125,4	1gpb	201,4	282,2
1ped	031,1	039,8	3cyr	011,9	013,2	1krn	011,0	013,2
2myr	070,4	089,2						

For the sequential approach, the maximum, the minimum and the average were 5.35, 0.44 and 1.29, respectively. This indicates that the variation of multithreaded approach is smaller than sequential one, which shows the stability of the multithread strategy. Finally, when comparing the total execution time of all test cases, it can be noticed that the overall multithread COFFEE execution time was 24% faster than the sequential COFFEE approach. Moreover, the quality of the results was kept without lost the biological significance of the final alignments.

5. Conclusion

Through extensive evaluation of the final version of the multithreaded COFFEE objective function developed in this work, we proved to be able the reduction of execution time of COFFEE objective function, improving the efficiency in relation to the initial version and using all avaliable computational resources. Moreover, we did not depreciate the quality of final alignments. We expect to implement the stratategy using Graphics Processing Units (GPUs).

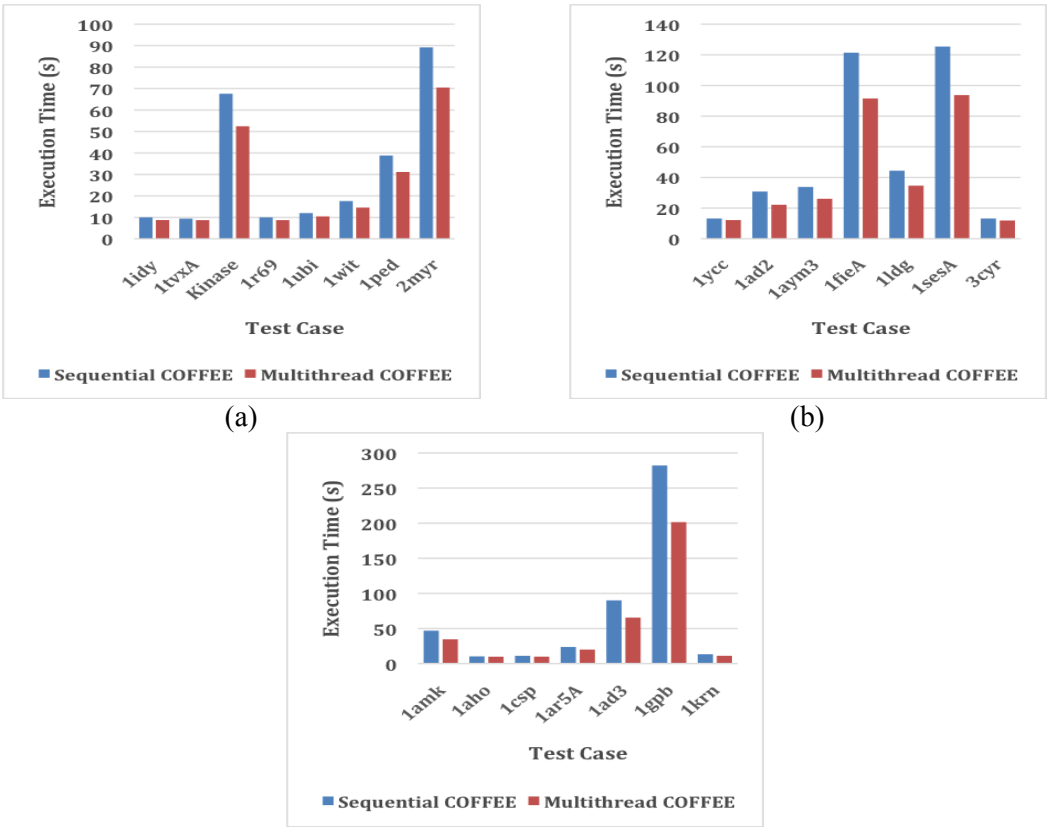

Figure 2. Execution time given in seconds: (a) sets with < 25% of similarity, (b) sets with similarity between 20% and 40% and (c) sets with > 35% of similarity.

6. Acknowledgments

This work was partially supported by São Paulo Research Foundation (FAPESP), under grant 2013/08289-0 and São José do Rio Preto Extension and Research Foundation (FAPERP).

References

[1] Khuri S 2008 A bioinformatics track in computer science. *ACM SIGCSE Bulletin* **40** n. 1 p. 508–512
[2] Amorim A R et al. 2015 Improvements in the sensibility of MSA-GA tool using COFFEE objective function *Journal of Physics: Conference Series* **574** n. 1 p. 12104
[3] Marucci A M et al. 2014 An Efficient Parallel Algorithm for Multiple Sequence Similarities Calculation Using a Low Complexity Method. *Biomed Research International* **2014** p. 1-6
[4] Li, K-B 2003 ClustalW-MPI: ClustalW analysis using distributed and parallel computing *Bioinformatics* **19** n. 12 p. 1585–1586
[5] Katoh, K and Toh H 2010 Parallelization of the MAFFT multiple sequence alignment program *Bioinformatics* **26** n. 15 p. 1899–900
[6] Zola J et al. 2007 PARALLEL-TCOFFEE: A parallel multiple sequence aligner. *ISCA PDCS* p. 248–253
[7] Notredame C and Holm L and Higgins D 1998 COFFEE: an objective function for multiple sequence alignments *Bioinformatics* **14** n. 5 p. 407–422