Parallel Computing 42 (2015) 18–34

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

High Performance computing improvements on bioinformatics consistency-based multiple sequence alignment tools

Miquel Orobitg [a,*], Fernando Guirado [a], Fernando Cores [a], Jordi Llados [a], Cedric Notredame [b]

[a] Universitat de Lleida, EPS Building, Avda. Jaume II, n.69, 25001 Lleida, Spain
[b] Universitat Pompeu Fabra – Centre de Regulacio Genomica (CRG), PRBB Building, Dr. Aiguader, n.88, 08003 Barcelona, Spain

ARTICLE INFO

Article history:
Available online 8 October 2014

Keywords:
Multiple Sequence Alignments
Consistency
T-Coffee
High Performance Computing
Scalability
Accuracy

ABSTRACT

Multiple Sequence Alignment (MSA) is essential for a wide range of applications in Bioinformatics. Traditionally, the alignment accuracy was the main metric used to evaluate the goodness of MSA tools. However, with the growth of sequencing data, other features, such as performance and the capacity to align larger datasets, are gaining strength. To achieve these new requirements, without affecting accuracy, the use of high-performance computing (HPC) resources and techniques is crucial. In this paper, we apply HPC techniques in T-Coffee, one of the more accurate but less scalable MSA tools. We integrate three innovative solutions into T-Coffee: the Balanced Guide Tree to increase the parallelism/performance, the Optimized Library Method with the aim of enhancing the scalability and the Multiple Tree Alignment, which explores different alignments in parallel to improve the accuracy. The results obtained show that the resulting tool, MTA-TCoffee, is able to improve the scalability in both the execution time and also the number of sequences to be aligned. Furthermore, not only is the alignment accuracy not affected by these improvements, as would be expected, but it improves significantly. Finally, we emphasize that the presented methods are not just restricted to T-Coffee, but may be implemented in any other alignment tools that use similar algorithms (progressive alignment, consistency or guide trees).

1. Introduction

Multiple Sequence Alignment (MSA) plays a key role in many domains in bioinformatics and computational biology. These tools are used in a wide range of applications, like phylogenetic analysis, homology detection and 2D/3D structures prediction.

In the post-genomic era, the growing complexity of the multiple alignment problem has led to further scalability and performance requirements from MSA tools. Nowadays, the number and length of sequences to be aligned are greater, and stress the capabilities of current alignment tools. Moreover, the quality of the alignment is negatively affected by the number of sequences. The accuracy obtained by traditional methods tends to deteriorate quickly as the number of sequences increases. This degradation in the quality of the alignment may have a significant impact on the correctness of many biological applications whose results depend strongly in turn on the accuracy of the multiple sequence alignments. For example, in phylogeny estimation, inaccurate alignments tend to produce inaccurate trees [27]. Therefore there is a pressing need for tools capable of managing large-scale alignments efficiently without losing accuracy.

[*] Corresponding author.
E-mail addresses: miquel.orobitg@crg.eu (M. Orobitg), f.guirado@diei.udl.cat (F. Guirado), fcores@diei.udl.cat (F. Cores), jordi.llados@udl.cat (J. Llados), cedric.notredame@crg.eu (C. Notredame).

http://dx.doi.org/10.1016/j.parco.2014.09.010
0167-8191/© 2014 Elsevier B.V. All rights reserved.

In recent years, a wide range of MSA tools have been developed with the aim of improving the accuracy and performance of the overall alignment [2,9,11,17,21,23]. These methods range from the very fast and less accurate progressive and iterative methods, such as Clustal-W [23] and Muscle [17], to the highly accurate but slow consistency-based ones, T-Coffee [11] and Probcons [2]. However, no method has been shown to be able to align hundreds of sequences with the quality level required by some biological applications. Clearly, new methods, or further improvements in existing ones, have to be developed to enable highly accurate alignment estimation for large datasets [6].

Our approach addresses this challenge through the utilization of HPC resources and techniques to improve the scalability and performance of the consistency-based methods, and maintaining its intrinsic accuracy. To reach this objective, we need to overcome certain limitations imposed by the original algorithm:

- **Increase the parallelism** of progressive alignment methods. The multiple sequence alignment process follows the order given by a guide tree that identifies the closely related sequences and their alignment precedences. Traditional guide tree building methods, like the neighbor-joining (NJ) method [19], create very large and unbalanced trees that lead to large dependence critical paths with poor parallelism. We proposed a new guide-tree building method, called BGT (Balanced Guide Tree), capable of overcoming the degree of parallelism but maintaining the distance tree structure.
- **Improve the scalability**. Consistency-based MSA tools store the sequence relationship in a dedicated library. The more sequences that are aligned, the more data is stored. The calculation and storage of the library is the main limiting factor for the scalability of these methods. To allow bigger sets of sequences to be processed, we propose a new library-building method, which is called Optimized Library Method (OLM), capable of optimizing the consistency library and reducing the execution time.
- **More accurate alignments**. It is proven that the most accurate alignments are obtained when the sequences are aligned following a phylogenetic order. However, this order cannot be determined a priori. It can only be approximated by a guide tree. As the guide tree is heuristic, a slight variation in it could produce a better alignment. This is the idea from which the MTA (Multiple Tree Alignment) method derives. The MTA provides different variations of any guide tree, processes all of them in parallel and finally selects the most accurate alignment.

The three proposed solutions (BGT, OLM and MTA) have been integrated into T-Coffee package. We have chosen T-Coffee package, the first consistency-based method, because it is capable of aligning any type of sequences (Protein, DNA and RNA). Furthermore, it is able to combine different types of biological information (other alignments, structural and profile information) in order to generate more accurate results. These features make T-Coffee one of the most suitable methods for aligning small datasets of sequences. However, T-Coffee has high memory and computation requirements that limit its scalability to a few hundred sequences. Therefore, improving the scalability of T-Coffee is crucial to increasing its utility for the bioinformatics community and allowing it to undertake a wider range of biological applications.

The present work extends a previous paper [16] where we presented the MTA approach. Thus, we have extended both the description and the validation of the individual methods that comprise MTA. We introduce the two main OML algorithms: the essential and threshold library optimization methods. We have also analyzed the influence of the guide tree on the alignment quality, confirming the potentiality of MTA and the significance of the guide-tree selection algorithm. In the experimentation, we have extended the MTA performance and scalability study. Furthermore, we have analyzed the impact on the alignment accuracy of the two main parameters of MTA: the size of the library (consistency) and the number of guide trees. We look for the most cost-effective combination to trade-of alignment accuracy and performance. Finally, we have extended the accuracy study, including one additional benchmark: BAliBASE.

The rest of the paper is organized as follows: Section 2 presents a brief state of the art of MSA tools. In Section 3 some of the scalability problems of T-Coffee MSA tool are analyzed. In Section 4, we present our approaches to solving the previously presented problems. The performance, scalability and accuracy evaluation are shown in Section 5 and finally the main conclusions are presented in Section 6.

2. State of art

The computation of an optimal mathematical alignment is an NP-Complete problem [26]. For this reason, current implementations of the MSA algorithms are heuristic and none of them guarantees full optimization. The progressive alignment is one of the most widely used heuristic. It assembles a multiple alignment by making a series of pairwise alignments of sequences, which are added one by one following the order established by a guide tree. The most popular progressive alignment implementation is the Clustal family [23].

Although this heuristic provides a great advantage in speed and simplicity, progressive methods are very dependent on the initial alignments, and several studies have shown that the alignment may be sensitive to errors in the guide tree. To correct or minimize errors made in progressive alignment steps, two techniques are frequently used: iterative refinement and consistency scoring.

Iterative refinement is based on performing a progressive alignment and then refining the result by repeatedly dividing the aligned sequences into sub-alignments and realigning the sub-alignments. The most relevant iterative aligners are MAFFT [4], Muscle [17] and ClustalΩ [21].