Applied Soft Computing xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

Multiobjective characteristic-based framework for very-large multiple sequence alignment

Álvaro Rubio-Largo [a,*], Leonardo Vanneschi [a], Mauro Castelli [a], Miguel A. Vega-Rodríguez [b]

[a] NOVA IMS, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal
[b] Depto. of Computer and Communications Technologies, University of Extremadura, 10003 Cáceres, Spain

ARTICLE INFO

Article history:
Received 29 April 2016
Received in revised form 28 April 2017
Accepted 13 June 2017
Available online xxx

Keywords:
Multiobjective optimization
Multiple sequence alignment
Characteristic-based
Evolutionary algorithms

ABSTRACT

In the literature, we can find several heuristics for solving the multiple sequence alignment problem. The vast majority of them makes use of flags in order to modify certain alignment parameters; however, if no flags are used, the aligner will run with the default parameter configuration, which, often, is not the optimal one. In this work, we propose a framework that, depending on the biological characteristics of the input dataset, runs the aligner with the best parameter configuration found for another dataset that has similar biological characteristics, improving the accuracy and conservation of the obtained alignment. To train the framework, we use three well-known multiobjective evolutionary algorithms: NSGA-II, IBEA, and MOEA/D. Then, we perform a comparative study between several aligners proposed in the literature and the characteristic-based version of Kalign, MAFFT, and MUSCLE, when solving widely-used benchmarks (PREFAB v4.0 and SABmark v1.65) and very-large benchmarks with thousands of unaligned sequences (HomFam).

1. Introduction

Multiple sequence alignment (MSA) [2,3,39], is an NP-hard optimization problem in molecular biology [11]. Basically, it is defined as the alignment of three or more nucleotides/amino acids sequences simultaneously.

Given a set of k unaligned sequences S: $\{s_1, s_2, \ldots, s_k\}$ defined over an alphabet Σ (aminoacids or nucleotides alphabet), a multiple sequence alignment of S is defined as S': $\{s'_1, s'_2, \ldots, s'_k\}$, where the length of the k sequences is exactly the same. The MSA (S') is defined over the alphabet $\Sigma \cup \{-\}$, that is, the same alphabet as S with an additional gap symbol ($-$).

Therefore, a MSA is achieved by inserting a number of gap symbols to the different sequences of S to obtain k sequences with the same length. The MSA is commonly represented by means of matrix, where the rows are the sequences and the columns are the aligned symbols. Each column contains at least one symbol of the alphabet Σ (i.e., a column containing only gap symbols is not permitted). Finding an optimal alignment is an NP-hard optimization problem with a complexity equals to $O(k2^k L^k)$ [48,47], where L is the maximum length of the k unaligned sequences of S.

An illustrative example of multiple sequence alignment will help to understand the problem:

- Unaligned Sequences (input):
  s_1: `THEELEPHANTSAREGRAY (19)`
  s_2: `MICEANDELEPHANTSAREFRIENDS (26)`
  s_3: `THEELEPHANTCANFLY (17)`

- Aligned Sequences (output):
  s'_1: `----THEELEPHANTSAREGRAY--- (26)`
  s'_2: `MICEANDELEPHANTSAREFRIENDS (26)`
  s'_3: `----THEELEPHANTCANFLY----- (26)`

According to [30], conservation is crucial to increase the biological significance of an alignment; so, an accurate MSA is critical for finding strong biological facts about proteins. The multiple sequence alignment is also an important step to infer phylogenetics relationships among the different input sequences [12,17]. Finally, a well-formed alignment helps us to determine which genes may be susceptible to suffer mutation.

As we mentioned before, the MSA problem is an NP-hard optimization problem where the complexity becomes prohibitive when the number of input sequences increases. In the literature,

* Corresponding author.
E-mail addresses: arl@unex.es (Á. Rubio-Largo), lvanneschi@novaims.unl.pt (L. Vanneschi), mcastelli@novaims.unl.pt (M. Castelli), mavega@unex.es (M.A. Vega-Rodríguez).

http://dx.doi.org/10.1016/j.asoc.2017.06.022
1568-4946/

we find exact methods (such as dynamic programming) that cannot handle the MSA problem in a reasonable amount of time when the number of sequences is greater than a few sequences [23]. Therefore, approximate methods have been proposed to find pseudo-optimal alignments [8]. These methods may be included in three different groups: *progressive-based* methods, *consistency-based* methods, and *iterative refinement* methods.

The first group contains the well-known *progressive-based* methods [18]. These approaches computes a distance matrix for every pair of unaligned sequences; then, they make use of any hierarchical clustering algorithm with the aim of building a guide-tree. The last step consists in using the guide-tree to construct the alignment. They have been considered the main gear in new aligners. Among the main *progressive-based* aligners, we find: Clustal W [44], Clustal Ω [38], PRANK [25], Fast Statistical Alignment [7], Kalign [22], DIALIGN-TX [42].

The *consistency-based* methods focus on building a database with local and global alignments between every pair of input sequences. These approaches start by harnessing the information contained within regions that are consistently aligned among a set of pairwise superpositions. The objective is to realign pairs of proteins through both global and local refinement methods, [13]. In the *consistency-based* group, we can find very well-known aligners, such as Tree-based Consistency Objective Function For alignment Evaluation (T-Coffee) [31], PROBabilistic CONSistency-based multiple sequence alignment (ProbCons) [10], ProbAlign [33], MSAProbs [24].

The third and last group includes the *iterative refinement* aligners. These aligners make use of a progressive methodology in order to build a preliminary alignment. Then, they perform a number of iterations for correcting any gap error produced during the progressive alignment. Basically, at each iteration, the approach divides the guide-tree into two subtrees that will be re-aligned in order to obtain an improved alignment. In the literature, we can find several *iterative refinement* aligners, among the most important ones are: MUltiple Sequence Comparison by Log-Expectation (MUSCLE) [14] and Multiple Alignment using Fast Fourier Transform (MAFFT) [20]. In this group, we can find some evolutionary and/or genetic algorithms techniques for the MSA problem: VDGA [28], GAPAM [29], MO-SAStrE [32], HMOABC [35], H4MSA [36].

The vast majority of the aforementioned methods makes use of flags to modify certain alignment parameters. The use of different values for these parameters leads to different alignments; therefore, a proper parameter configuration of the aligner is critical to obtain an accurate output. In case of using no flag, the aligner will use a default parameter configuration, which is proposed by the developers of the aligner.

Unfortunately, the alignment produced by considering the default parameter configuration is not always the best choice, the main reason lies in the fact that the default parameters are those that gave the developers best average accuracy in their training sets (different for each aligner). In this work, we propose a framework that, depending on the biological characteristics of the input set of sequences, runs the aligner with the best parameter configuration found for a different set of sequences with similar biological characteristics, improving the accuracy and conservation of the final alignment. The framework requires a *characteristics-configuration file*, that is, the best parameter configuration found for several sets of unaligned sequences with different biological characteristics.

According to [19], the problem of finding an optimal parameter configuration of an aligner is commonly treated as an optimization problem. In this work, we tackle this problem by using multiobjective optimization: given a set of unaligned sequences, we need to find the best parameter configuration for an aligner that simultaneously maximizes the accuracy and conservation of the alignment obtained. Multiobjective optimization has been applied in a wide variety of real-world application domains with successful results [34,37].

The main contributions of the manuscript are:

- A characteristic-based framework for improving the accuracy and conservation of any aligner.
- A set of biological characteristics that describes any input set of unaligned sequences.
- The use of three well-known multiobjective evolutionary algorithms when optimizing the parameters of Kalign, MAFFT, and MUSCLE. The selected algorithms are: the dominance-based *Fast Non-Dominated Sorting Genetic Algorithm* (NSGA-II) [9], the *Indicator-based Evolutionary Algorithm* (IBEA) [51], and the *Multiobjective Evolutionary Algorithm based on Decomposition* (MOEA/D) [49].
- A comparative study between the characteristic-based version of three well-known aligners (Kalign, MAFFT, and MUSCLE) and several aligners proposed in the literature. In the comparative study, we study the advantages of the proposed framework when dealing with well-known benchmarks, such as PREFAB v4.0 and SABmark v1.65; and very-large benchmarks with thousands of unaligned sequences (HomFam).

The rest of the paper is organized as follows. In Section 2, we explain the multiobjective parameter optimization problem and detail the characteristic-based framework. Section 3 contains the comparative study on the effectiveness of the framework, comparing the accuracy of the framework with other aligners published in the literature. Finally, in Section 4, we summarize the conclusions extracted from the study and describe some lines of future work.

2. Methodology

This section is divided into two parts. On the one hand, we describe the multiobjective parameter optimization problem and the multiobjective approaches used. On the other hand, we explain how the characteristic-based framework works, including an example for a better understanding.

2.1. Multiobjective parameter optimization problem

Multiple sequence aligners commonly use parameters that determine the behaviour of the aligner; therefore, an optimal selection of values for these parameters is crucial for obtaining alignments with a higher level of biological significance. In the literature, the Q-score and TC-score [14] have been employed to measure the biological significance of an alignment:

- **Q-score** (quality score, f_1). It indicates the number of correctly aligned residue pairs divided by the number of residue pairs in the reference alignment (true alignment); it is also known as Sum-of-Pairs (SP) score.
- **TC-score** (total column score, f_2). It is the number of correctly aligned columns divided by the number of columns in the reference alignment; it is also known as Column Score (CS).

In this work, given a multiple sequence aligner, the problem of finding the best configuration for its parameters has been formulated as a multiobjective optimization problem (MOP), where the final goal is to find an optimal parameter configuration that simultaneously optimize the Q-score (f_1) and TC-score (f_2).

In the following, we state the problem in a more formal way:

$$\text{maximize} \quad F(x) = (f_1(x), f_2(x))$$

$$\text{subject to} \quad x \in \Omega$$

Please cite this article in press as: Á. Rubio-Largo, et al., Multiobjective characteristic-based framework for very-large multiple sequence alignment, Appl. Soft Comput. J. (2017), http://dx.doi.org/10.1016/j.asoc.2017.06.022