Swarm and Evolutionary Computation xxx (2018) 1–13

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

Swarm intelligence for optimizing the parameters of multiple sequence aligners

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

Keywords:
Swarm intelligence
Multiple sequence alignment
Characteristics-based framework
Evolutionary algorithms

ABSTRACT

Different aligner heuristics can be found in the literature to solve the Multiple Sequence Alignment problem. These aligners rely on the parameter configuration proposed by their authors (also known as default parameter configuration), that tried to obtain good results (alignments with high accuracy and conservation) for any input set of unaligned sequences. However, the default parameter configuration is not always the best parameter configuration for every input set; namely, depending on the biological characteristics of the input set, one may be able to find a better parameter configuration that outputs a more accurate and conservative alignment. This work's main contributions include: to study the input set's biological characteristics and to then apply the best parameter configuration found depending on those characteristics. The framework uses a pre-computed file to take the best parameter configuration found for a dataset with similar biological characteristics. In order to create this file, we use a Particle Swarm Optimization (PSO) algorithm, that is, an algorithm based on swarm intelligence. To test the effectiveness of the characteristic-based framework, we employ five well-known aligners: Clustal W, DIALIGN-TX, Kalign2, MAFFT, and MUSCLE. The results of these aligners see clear improvements when using the proposed characteristic-based framework.

1. Introduction

The simultaneous alignment of three or more Nucleotides/Amino-Acids (AA) sequences is known in the literature as the Multiple Sequence Alignment (MSA) problem [1], and is considered an NP-complete optimization problem [10]. The MSA problem can be defined as follows:

Given a set of sequences S: $\{s_1, s_2, \ldots, s_k\}$ of lengths $|s_1|, |s_2|, \ldots, |s_k|$ defined over an alphabet Σ, (for example the AA or the nucleotides alphabets), a MSA of S is defined as the set S': $\{s'_1, s'_2, \ldots, s'_k\}$, where the length of all the k sequences is exactly the same. Note that, S' is defined over the same alphabet as S (Σ) with an additional gap symbol (−); S' is thus defined over the alphabet $\Sigma \cup \{-\}$.

In this way, a multiple alignment is obtained by adding gaps to the sequences of S so that their lengths become the same. It can be seen as a matrix representation where the rows are sequences and the columns represent aligned symbols. Each column of an alignment must contain at least one symbol of Σ, (namely, a column with all gaps is not allowed). According to [10], the complexity of finding an optimal alignment is $O(k2^k L^k)$, where k is the number of sequences and L is the $\max(|s_1|, |s_2|, \ldots, |s_k|)$. In the following, we present an example of MSA:

Unaligned set (S):				Aligned set (S'):		
s_1:	GDNI	(4)		s'_1:	GDNI--	(6)
s_2:	KQLTQD	(6)	→	s'_2:	KQLTQD	(6)
s_3:	ACRKN	(5)		s'_3:	ACRK-N	(6)

A well-conserved alignment leads to extra biological significance [31]; therefore, MSA is frequently employed to produce strong biological facts about proteins. Further, MSA mainly focuses on reflecting biological relationships among different sequences, which is an essential

* 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).

https://doi.org/10.1016/j.swevo.2018.04.003
Received 23 November 2016; Received in revised form 8 March 2018; Accepted 8 April 2018
Available online XXX

step for inferring phylogenetic relationships [11,16]. Another important feature of an accurate MSA is that it allows the determination of genes that are susceptible to mutation.

In the literature, we find a range of approaches to deal with the MSA problem. While almost all of them allow us to modify some specific parameters by using different flags, if no flags are used then a default configuration is used. The default configuration is proposed by the developers of the aligner and refers to the best parameter configuration found for aligning any input set of unaligned sequences with a reasonable level of accuracy and conservation. However, the default configuration may not always be the best choice for every input set. Depending on the biological characteristics of the input set, a better configuration may be used to obtain a more accurate and conservative alignment. This is the idea of the proposed characteristic-based framework: to study the biological characteristics of the input set and, consequently, to apply a certain configuration depending on those characteristics. Therefore, the characteristic-based framework uses three input files: aligner (executable), a set of unaligned sequences, and a *characteristics-configuration file*. Note that the *characteristics-configuration file* depends on the input aligner, and contains the best parameter configuration of the aligner for a number of biological datasets with different characteristics. A swarm intelligence approach is applied to optimize the parameters of an input aligner, thereby obtaining its *characteristics-configuration file*. In this way, the framework will run the aligner with the best parameter configuration found for another dataset with similar biological characteristics, improving the input aligner's accuracy and conservation. In Ref. [37], we present a preliminary version of the framework.

As demonstrated by a series of recent publications [19,20], in compliance with the 5-step rule [7] when developing a really useful sequence-based method for a biological system we should follow these five guidelines: (a) construct or select a valid benchmark dataset to train and test the model; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be identified; (c) introduce or develop a powerful algorithm (or engine) to operate the analysis; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the model; and (e) establish a user-friendly web-server for the analysis method that is accessible to the public. Below, we describe how to deal with these steps individually.

The biggest contributions of this work include the following: a characteristic-based framework for improving the quality alignment of any aligner, diverse biological characteristics for describing a set of unaligned sequences, and a comparative study on the framework's effectiveness when it is applied to five aligners: Clustal W, DIALIGN-TX, Kalign2, MAFFT, and MUSCLE.

The rest of the paper is organized as follows. A description of related works is provided in Section 2. In Section 3, we detail the characteristic-based framework. Section 4 compares the framework's accuracy with other aligners published in the literature. Finally, in Section 5 we summarize the conclusions extracted from the study and describe some avenues for future work.

2. Related work

Traditionally, exact approaches for MSA, such as dynamic programming, have been used. Yet these methods become computationally prohibitive when the number of sequences increases. In the literature, we find different heuristics for MSA that are categorized in three groups: progressive-based methods, consistency-based methods, and iterative refinement methods.

In the first group, we find the progressive-based methods, which are widely used [18]. Basically, given a set of unaligned sequences a progressive-based method computes a distance matrix from every pair of sequences. After that, it employs a hierarchical clustering algorithm, such as the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) or Neighbor-Joining (NJ), in order to build a guide-tree. The last step is to perform the alignment among the given sequences by following the guide-tree. Several progressive-based methods exist, such as Clustal W [42], PRANK [28], Fast Statistical Alignment (FSA) [4], Kalign [24], and DIALIGN-TX [41].

In the second group, we have the consistency-based methods. These approaches build a database with the local and global alignments between every pair of sequences. According to [12], the consistency-based approaches first harness the information contained within regions that are consistently aligned among a set of pairwise superpositions in order to realign pairs of proteins through both global and local refinement methods. Among the most important consistency-based methods are: Tree-based Consistency Objective Function For alignment Evaluation (T-Coffee) [32], PROBabilistic CONSistency-based multiple sequence alignment (ProbCons) [9], ProbAlign [36], and MSAProbs [26].

In the third group, we find the iterative refinement methods. These focus on correcting an erroneous gap inserted at an early stage of a progressive alignment. The first step in these methods is to build an initial MSA by using any progressive-based method. The second step consists of dividing the guide-tree of the initial alignment into two subtrees which are re-aligned with the aim of obtaining an improved new alignment. The second step is iteratively repeated until a certain number of iterations is reached. There exist several iterative refinement methods, such as MUltiple Sequence Comparison by Log-Expectation (MUSCLE) [13], Multiple Alignment using Fast Fourier Transform (MAFFT) [21], ProbCons [9] (it allows the option of a final iterative refinement), and MUMMALS [34]. In this group, we can find some evolutionary and/or genetic algorithms techniques for the MSA problem: VDGA [29], GAPAM [30], MO-SAStrE [33], HMOABC [39], H4MSA [40].

To avoid completely losing the sequence-order information, the concept of PseAA (pseudo amino acid) composition was proposed [6]. In contrast with the conventional amino acid composition that contains 20 components with each reflecting the occurrence frequency for one of the 20 native amino acids in a protein, the PseAA composition contains a set of greater than 20 discrete factors, where the first 20 represent the components of its conventional AA composition while the additional factors incorporate some sequence-order information via various modes. Some very powerful bioinformatics tools for analyzing biological sequences were recently developed, e.g. a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition [17] or an effective formulation for analyzing genomic sequences [5]. When harnessing Pse-in-One 2.0 [5], users only need to input DNA, RNA, or protein sequences as well as their selected or defined features, and can immediately obtain the corresponding feature vectors suitable for any of the existing machine-learning programs to conduct various analyses. All the aforementioned works may be considered a starting point for the characteristic-based framework proposed in this paper, which analyzes the composition of the protein sequences in order to select a proper parameter configuration.

3. Characteristic-based framework

This section is divided into two. The first subsection describes how the *characteristic-configuration* file is obtained, while the second discusses the main properties of the proposed framework.