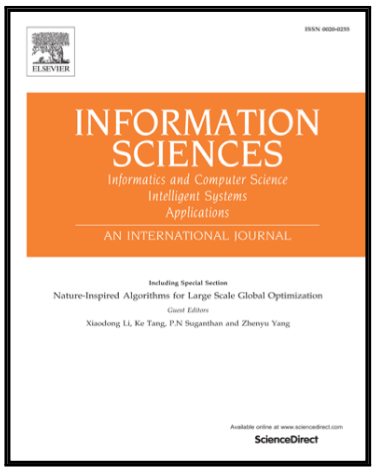

Using Biological Knowledge for Multiple Sequence Aligner Decision Making

Álvaro Rubio-Largo, Leonardo Vanneschi, Mauro Castelli, Miguel A. Vega-Rodríguez

PII: S0020-0255(17)30136-6
DOI: 10.1016/j.ins.2017.08.069
Reference: INS 13074

To appear in: *Information Sciences*

Received date: 10 January 2017
Revised date: 16 May 2017
Accepted date: 20 August 2017

Please cite this article as: Álvaro Rubio-Largo, Leonardo Vanneschi, Mauro Castelli, Miguel A. Vega-Rodríguez, Using Biological Knowledge for Multiple Sequence Aligner Decision Making, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.08.069

Using Biological Knowledge for Multiple Sequence Aligner Decision Making

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*

Abstract

Multiple Sequence Alignment (MSA) is the simultaneous alignment among three or more biological sequences (nucleotides or amino acids). In recent years, important efforts have been assigned to the development of MSA approaches. In this work, we propose a framework that extracts the biological characteristics of an input set of unaligned sequences and uses this knowledge to decide which is the most suitable aligner and parameter configuration. We refer to it as Multiple Aligner Framework (MAF). The selection of the tuple {Aligner, Configuration} is based on searching, in a pre-computed file, the best tuple for a dataset with similar biological characteristics. In order to create this file, we use multiobjective optimization. In fact, three well-known multiobjective evolutionary algorithms (NSGA-II, IBEA and MOEA/D) have been used. To validate the framework, we have used five popular benchmark suites: BAliBASE 3.0, PREFAB 4.0, SABmark 1.65, OX-Bench and CDD 3.14. After comparing with well-known aligners published in the literature, such as Kalign2, MUSCLE, MAFFT, T-Coffee, MSAProbs, ProbCons, Clustal Ω and MUMMALS, we conclude that the multiple aligner framework is, in average, the method with the best balance between alignment accuracy/conservation and required runtime.

Keywords: biological knowledge, aligner decision making, multiple sequence alignment, multiobjective optimization

1. Introduction

In molecular biology, the problem of simultaneously aligning three or more biological sequences is known as Multiple Sequence Alignment (MSA) [1]. MSA is an important step to infer phylogenetics relationships among the different input sequences [6], [8]. An accurate and conservative MSA leads to strong biological significance, which is critical in the study of proteins and nucleotides.

Given a set of k unaligned sequences S: {s1, s2, ..., sk} defined over an alphabet Σ (amino-acids or nucleotides alphabet), the multiple sequence alignment of this set is defined as S': {s1', s2', ..., sk'}, where all the sequences are of equal length. Therefore, the produced alignment (S') will be defined over the alphabet $\Sigma \cup \{-\}$, where $-$ refers to gap symbol.

For example, given the following set of unaligned sequences (S):

```
>s1
GRLIHPASGRSYHKIFNPPKEDMKDDVTGEALVQRSDD
>s2
GRRLDPVTGKIYHLKYSPPENEEIASRLTQR
>s3
GRRICRNCGATYHLIFHPPAKPGVCDKCGGELYQR
```

a possible alignment (S') would be:

Email addresses: arl@unex.es (Álvaro Rubio-Largo), lvanneschi@novaims.unl.pt (Leonardo Vanneschi), mcastelli@novaims.unl.pt (Mauro Castelli), mavega@unex.es (Miguel A. Vega-Rodríguez)