Neurocomputing 164 (2015) 123–136

Contents lists available at ScienceDirect

Neurocomputing

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

Comparing different machine learning and mathematical regression models to evaluate multiple sequence alignments

Francisco M. Ortuño [a,*], Olga Valenzuela [b], Beatriz Prieto [a], Maria Jose Saez-Lara [c], Carolina Torres [c], Hector Pomares [a], Ignacio Rojas [a]

[a] Department of Computer Architecture and Computer Technology, CITIC-UGR, University of Granada, Spain
[b] Department of Applied Mathematics, University of Granada, Spain
[c] Department of Biochemistry and Molecular Biology I, University of Granada, Spain

ARTICLE INFO

Article history:
Received 21 February 2014
Received in revised form
10 January 2015
Accepted 12 January 2015
Available online 16 March 2015

Keywords:
Multiple sequence alignments (MSAs)
Alignment quality
Least squares support vector machines (LS-SVM)
Decision trees
Bootstrap aggregation
Gaussian process

ABSTRACT

The evaluation of multiple sequence alignments (MSAs) is still an open task in bioinformatics. Current MSA scores do not agree about how alignments must be accurately evaluated. Consequently, it is not trivial to know the quality of MSAs when reference alignments are not provided. Recent scores tend to use more complex evaluations adding supplementary biological features. In this work, a set of novel regression approaches are proposed for the MSA evaluation, comparing several supervised learning and mathematical methodologies. Therefore, the following models specifically designed for regression are applied: regression trees, a bootstrap aggregation of regression trees (bagging trees), least-squares support vector machines (LS-SVMs) and Gaussian processes. These algorithms consider a heterogeneous set of biological features together with other standard MSA scores in order to predict the quality of alignments. The most relevant features are then applied to build novel score schemes for the evaluation of alignments. The proposed algorithms are validated by using the BAliBASE benchmark. Additionally, an statistical ANOVA test is performed to study the relevance of these scores considering three alignment factors. According to the obtained results, the four regression models provide accurate evaluations, even outperforming other standard scores such as BLOSUM, PAM or STRIKE.

1. Introduction

Multiple sequence alignments (MSAs) provide useful information for several bioinformatics and biomedical tasks. Sequence alignments are defined as comparisons of molecular chains (namely, nucleotides from genes or amino-acids from proteins) to extract their relevant similarities and differences. Four different symbols (A, T, C and G) are used to represent nucleotides whereas the 22 standard amino acids are codified by their 22 corresponding symbols. The main purpose is to match the highest possible number of similar symbols among the compared sequences. In order to maximize these matches, additional symbols, called gaps and represented by '*', can be incorporated in the sequences. These gaps represent the biological processes of deletion (removal of either a single nucleotide or amino acids to the sequence) or insertion (addition of either of them). Finally, a mutation is being modeled when two different symbols are aligned. An example of a protein sequence alignment is shown in Fig. 1.

The study of these similarities and differences is usually required by a huge number of applications such as phylogenetic analyses [1], structural modelling [2], functional predictions [3] or sequence database searching [4]. MSAs are also facilitating the analysis of functional, structural and genomic information provided by novel high-throughput and next-generation sequencing (NGS) experiments [5]. Moreover, these experiments have prompted a great demand of MSAs in the last years. Current MSA applications should therefore be capable of dealing with and efficiently analysing the massive amount of information generated by these former techniques. With this purpose, high-performance computing (HPC) resources [6] and techniques based on graphics processing units (GPUs) [7,8] are being increasingly applied. Also, advanced computational approaches and machine-learning techniques are also widely used to improve the accuracy in MSAs. Thus, genetic algorithms can be implemented to combine different alignment scores in the fitness function and achieve more accurate MSAs [9,10]. Moreover, another well-known technique such as swarm optimization has been recently applied together with hidden Markov models (HMMs) to generate better alignments [11]. Finally, support vector machines (SVMs) are also applied in MSAs, for example, to improve the annotation associated to the alignments [12]. Therefore, MSAs are still being one of the most useful and required tools in bioinformatics [13].

An important challenge which has not already been properly addressed in MSAs is the estimation of the quality in alignments

* Corresponding author.
E-mail address: fortuno@ugr.es (F.M. Ortuño).

http://dx.doi.org/10.1016/j.neucom.2015.01.080
0925-2312/

```
SEQ1 ******MQDRVKRPKYRP***RRKAKMLPK
SEQ2 ********MHIKKPSARDNYGKKKKRKREK
SEQ3 MKKLKKHPDFPKKPREDH***PDLIQNAKK
SEQ4 ***GKGDPKKPPR*IPPK***GE*******
```

Fig. 1. Standard representation of a protein sequence alignment. In this case, four sequences are aligned.

[14,15]. Sequence alignments have been traditionally scored by using weighted matrices, such as Point Accepted Mutation (PAM) [16] or BLOSUM [17]. These matrices are normally associated with the probability of finding specific mutations between each pair of nucleotides or amino acids and they are still widely applied in alignment algorithms. However, since these matrices only take into account the sequence information (nucleotides or amino acids), they do not achieve a sufficiently accurate score, specially with less related sequences. Consequently, evaluation scores are currently being improved by using additional biological features, such as protein structures or homologies. For instance, both Contact Accepted Mutation (CAO) [18] and STRIKE scores [19] include information about molecular contacts in protein structures to estimate the quality of alignments. In the same way, other algorithms propose alternative evaluations by adding several features related to secondary structure, gaps and conservation [20,21]. However, most of the previously described alignment approaches still consider suboptimal scores such as PAM and BLOSUM [10,11] or STRIKE [9]. In this work, novel alternative scoring schemes will be proposed to address the MSA evaluation problem by integrating a wide dataset of both other scores and biological features.

Several benchmarks have been designed to determine if a particular approach can achieve an accurate evaluation. These benchmarks, e.g. BAliBASE [22], Prefab [23] or SABMark [24], provide different groups of sequences together with their reference alignments. The references, which are also known as gold standard, are usually handmade and carefully obtained, being BAliBASE ones among the most popular. They can then be considered as the best possible alignment for these particular sequences. Moreover, benchmarks usually provide accurate scores to determine the similarity between other alignments and their references. Therefore, a proposed quality score can be assessed by comparing with the provided reference alignments.

Since the estimation of the alignment quality proposed in this work aims to be similar to the previously described benchmarks, a regression problem is being addressed. Consequently, several mathematical and supervised learning approaches are proposed and compared here in order to determine an efficient score scheme. Specifically, following regression approaches are compared: regression trees, bootstrap aggregation trees, least-squares support vector machines (LS-SVMs) and Gaussian processes. Similar algorithms have already been proved to be useful for several biological data mining problems [25,26].

Therefore, this work will take advantage of the integration of 22 carefully extracted biological features and additional scores to determine the quality of alignments by using these standard regression algorithms. The main advantage of the proposed scoring schemes is the retrieval of relevant features specially designed for an accurate evaluation of alignments. Features are extracted from different resources and databases related to secondary and tertiary protein structures, domains, homologies and molecular properties. Other score schemes are also integrated in the dataset to complement these features. To the best of our knowledge and after a careful revision of the literature, such a wide feature dataset together with these supervised learning algorithms have not been proposed before for the MSA evaluation. The full procedure is graphically presented in Fig. 2.

2. Construction of the alignment dataset

A complete dataset with 2160 different alignments was applied in this work. These alignments were built by aligning several sets of protein sequences provided by the BAliBASE benchmark [22]. BAliBASE includes a dataset specially prepared to allow the alignments of sequences by standard MSA algorithms. The dataset contains a total of 218 manually extracted sets of sequences, principally retrieved from Protein Data Bank (PDB) [27]. Sequence sets are organized in six groups according to their families and similarities, namely RV11, RV12, RV20, RV30, RV40 and RV50. RV11 subset includes the least similar sequences ($<20\%$ of identities) whereas RV12 considers medium-to-distant sequences with 20–40% identity. RV20 and RV30 provide sequences from families and subfamilies with $>40\%$ of identities. Finally, sequences in RV40 and RV50 contain $>20\%$ of identities with large terminal and internal insertions. From this dataset, two sets of sequences were removed due to the fact that they do not contain enough information about protein structures. Thus, a total of 216 sets of sequences were considered.

BAliBASE also collects a set of reference alignments (*gold standard*) carefully obtained from expert knowledge by applying a wide and complex handmade analysis. These references can then be compared with alignments obtained by running other tools. Thus, BAliBASE provides an accurate score to evaluate alignments, also named BAliscore. This score was applied in the regression approaches as the output variable to train the evaluation models.

Ten representative MSA tools were then run with the sequences provided by BAliBASE (see '*Alignment Dataset*' stage in Fig. 2). Two standard strategies are traditionally applied for sequence alignments: progressive algorithms and consistency-based methods. Among progressive algorithms, ClustalW [28], Muscle [23], Kalign [29], Mafft [30] and RetAlign [31] were selected. Alignments from three additional algorithms based on consistency were also added, namely T-Coffee [32], FSA [33] and ProbCons [34]. Recently, more sophisticated algorithms have also been implemented to add further data in alignment tools, such as domains, secondary/tertiary structures or homologies. 3DCoffee [35] and Promals [36] were chosen among these novel MSA algorithms. Therefore, since 216 sets of sequences were used from BAliBASE, the proposed dataset included a total of 2160 alignments.

3. Databases and feature extraction

The main goal in the proposed automatic alignment score prediction was to provide an accurate evaluation scheme by using a set of relevant biological features. Some features related to the previously defined alignment dataset were then extracted from the well-known biological databases: Pfam [37], PDB [27], Uniprot [38] and Gene Ontology (GO) [39]. Such databases were consulted to obtain useful data which could enrich the sequence information, building an heterogeneous set of features. Depending on the consulted database, features were focused on a particular biological property (see '*Feature Extraction*' stage in Fig. 2). Thus, Pfam repositories provided data related to functional regions in proteins, also called domains (see features with the '*Domain*' type in Table 1). Information related to the secondary structure and location of proteins was collected from Uniprot ('*Secondary Structure*' and '*Location*' types). Moreover, tertiary structure was retrieved from PDB database whereas the Gene Ontology provided some molecular attributes of proteins by means of a controlled