Determining the most suitable multiple sequence alignment methodology by using a set of heterogeneous biological features.

Francisco Ortuño[1]*, Olga Valenzuela[2], Hector Pomares[1], and Ignacio Rojas[1]

[1] Department of Computer Architecture and Computer Technology
CITIC-UGR, University of Granada
[2] Department of Applied Mathematics, University of Granada

Abstract. Multiple sequence alignments (MSAs) are well-known procedures which provide useful information to other techniques in bioinformatics such as biological function analyses, structure predictions or next-generation sequencing. Nevertheless, the alignments provided by current MSA methodologies are quite different depending on the particular biological features of the aligned sequences. Thus, current tools do not totally agree on the most suitable way to align a specific set of sequences, overall when sequences are less related. In this work, we propose a novel machine learning technique based on support vector machines (SVMs) to predict "a priori" the MSA tool which will provide a more accurate alignment for a particular set of sequences. A set of heterogeneous biological features retrieved from well-known databases is applied to train the proposed algorithm. Finally, the SVM approach will be assessed by the benchmark BAliBASE v3.0.

Keywords: multiple sequence alignments (MSAs), feature selection, machine learning, support vector machine (SVM).

1 Introduction

Multiple sequence alignments (MSAs) are usually a key issue in order to analyse other biological tasks such as protein structures, mutations and functionality. More recently, MSAs have even become more relevant due to their applicability to novel experimental procedures like high-throughput experiments or next-generation sequencing (NGS). Therefore, MSA tools are currently essential to the development of several bioinformatics researches [10].

However, an important drawback of using MSAs is the huge amount of already existing tools, making hard to decide which one is the most suitable aligner to align a set of sequences. Currently, there is no a fair standard to build alignments and the aligners usually provide quite different alignment according to their criterion and the specific features of sequences. Therefore, the quality of an alignment can be significantly altered depending on the aligner being used.

* Corresponding author: fortuno@ugr.es

For this reason, a novel procedure to determine which aligner is the most promising to obtain a better alignment is proposed in this work. This algorithm is based on the extraction of the most relevant biological features, which enrich the information provided by the aligned sequences. These biological features are retrieved from different databases in order to build a heterogeneous dataset of features. These features are related to the sequence structures, homologies or chemical properties. Subsequently, a support vector machine (SVM) approach is implemented to predict the MSA tools which will align better a particular set of sequences. This approach has been assessed with the BAliBASE benchmark [17] by using a 10-fold cross-validation.

2 Materials and Methods

As commented above, the prediction procedure was evaluated by the BAliBASE benchmark. This dataset provides 218 manually extracted sets of sequences. This dataset was then aligned by using 10 well-know MSA tools (see Table 1). Therefore, a total dataset of 2180 alignments was used to asses the proposed methodology. Additionally, BAliBASE also provides a set of handmade reference alignments (*gold standard*) in order to score the alignments obtained by other tools. This score provided by BAliBASE, called BAliscore, is an accurate measure of the quality of the alignment.

The proposed method firstly extracted a dataset of 20 features in order to enrich the sequence information of multiple sequence alignments. Such features were retrieved from well-known databases such as the Gene Ontology Anotation (GOA) [4], Pfam [7], Uniprot [1] or the Protein Data Bank (PDB) [2]. Additionally, these features were evaluated according to the feature selection procedure called minimal-Redundancy-Maximal-Relevance (mRMR) [14], in order to choose the most relevant and less redundant features. Subsequently, the subset of optimal features were used to train and test the proposed prediction by using the SVM algorithm.

Table 1. Summary of MSA tools which were used to align the BAliBASE sequences. They are gathered in progressive approaches, consistency-based methods or aligners using more sophisticated features.

METHOD	Type	Version
ClustalW [16]	Progressive	2.0.10
Muscle [6]	Progressive	3.8.31
Kalign [9]	Progressive	2.04
Mafft [8]	Progressive	6.85
RetAlign [15]	Progressive	1.0
T-Coffee [11]	Consistency	8.97
FSA [3]	Consistency	1.15.5
ProbCons [5]	Consistency	1.12
3DCoffee [12]	Additional features	8.97
Promals [13]	Additional features	vServer

Table 2. Summary of features extracted from several databases. The relevance ranking provided by the mRMR procedure is also shown. (1) Percentage of amino acids (AA) with that specific feature. (2) Number of occurrences per sequence.

	FEATURE	SOURCE	RANK
f_1	# of sequences	BAliBASE	3
f_2	Average length	BAliBASE	4
f_3	Variance length	BAliBASE	6
f_4	Reference subset	BAliBASE	5
f_5	AA in α-helix(1)	Uniprot	16
f_6	AA in β-strand(1)	Uniprot	7
f_7	Domains(2)	Pfam	1
f_8	Shared Domains(2)	Pfam	15
f_9	GO terms(2)	GOA	11
f_{10}	MF-GO terms(2)	GOA	17
f_{11}	CC-GO terms(2)	GOA	20
f_{12}	BP-GO terms(2)	GOA	19
f_{13}	Shared GO terms(2)	GOA	18
f_{14}	3D-Structures(2)	PDB	14
f_{15}	Polar AA(1)	Biochemistry	9
f_{16}	Non-polar AA(1)	Biochemistry	12
f_{17}	Basic AA(1)	Biochemistry	10
f_{18}	Aromatic AA(1)	Biochemistry	13
f_{19}	Acid AA(1)	Biochemistry	8
f_{20}	MSA Method	—	2

In order to assess the proposed prediction, the full dataset was divided in the training and test subsets. A total of 1960 alignments (196 problems by 10 aligners) were used to train the SVM approach according to the quality provided by BAliscore. This procedure was assessed by using a 10-fold cross-validation. Subsequently, the remaining 220 alignments were used to test the results obtained by the SVM approach against the BAliscore.

3 Results and Discussion

The proposed algorithm aims to predict the most suitable methodology to align a specific set of sequences according to the Baliscore provided by BAliBASE. As described above, ten selected methodologies were firstly run for the 218 sets of BAliBASE in order to obtain a total dataset of 2180 alignments. Subsequently, 20 biological features related to the alignments and their sequences were retrieved. A subset of these features was then selected according to the ranking of features provided by the mRMR procedure (see column 'RANK' in Table 2). Finally, this subset of features was used to perform a support vector machine (SVM) procedure which predicted those methods which provided a good quality of alignment.

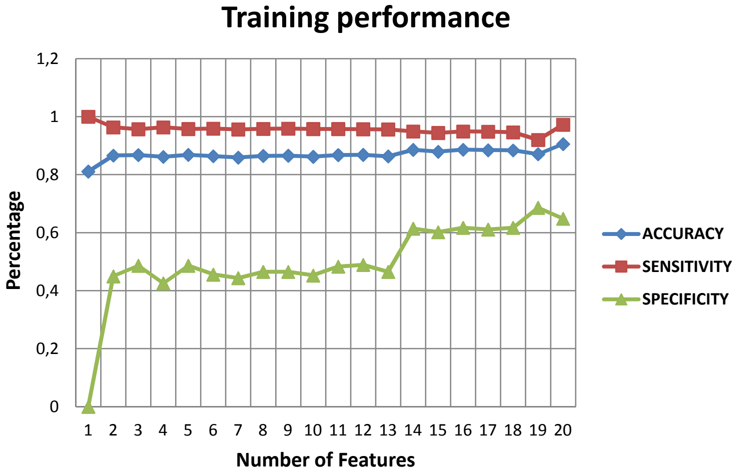

Fig. 1. Sensitivity, specificity and accuracy measures obtained for the training procedure. Alignments were classified as accurate/inaccurate alignments. Values are shown according to the number of features included.

In order to classify the alignments into accurate and inaccurate alignments, a BAliscore limit was chosen. Then, an alignment was defined as accurate when it had a BAliscore higher than 0.6. Thus, the dataset of 2180 alignments was classified according to the quality of their alignments (accurate vs inaccurate). Taking this classification into consideration, the SVM approach was trained by using a 90% of this dataset (1960 alignments). The training procedure was assessed with a 10-fold cross-validation. The predicted classification in this training was measured in terms of sensitivity (true positive rate), specificity (true negative rate) and accuracy. These scores are shown in the Figure 1 against an increasing number of features being added. Additionally, the remaining 220 alignments (10% of total) were applied to test the performance of our proposed prediction. In this case, the obtained sensitivity, specificity and accuracy measures for the test dataset are shown in the Figure 2.

As shown in the previous figures, the proposed SVM approach significantly predicted the accuracy of alignments (accurate or inaccurate) in around 88% for the training and 80% for the test. The maximum value taking the three measures into account was reached with a subset of 14 features. Therefore, it was not necessary to include the full dataset of features. Additionally, it can be appreciated that a higher number of features, e.g. 20 features, usually produced an overfitting in the algorithm, reducing the accuracy and specificity in the test dataset. For the first 14 features (see ranking in Table 2), the obtained results for the training dataset were of 88.56% of accuracy, 94.89% of sensitivity and 61.40% of specificity. Regarding the test dataset, the shown results were: 79.77% of accuracy, 84.54% of sensitivity and 65.45% of specificity.

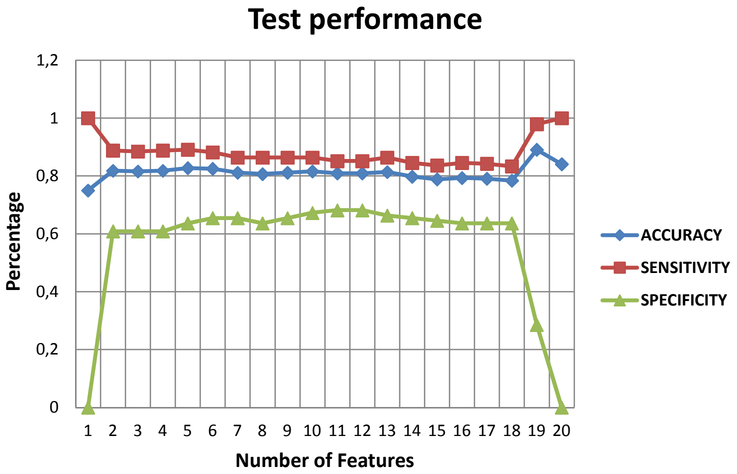

Fig. 2. Sensitivity, specificity and accuracy measures obtained for the test procedure. Alignments were classified as accurate/inaccurate alignments. Values are shown according to the number of features included.

4 Conclusions

In this work, a novel algorithm to predict the quality provided by MSA tools has been proposed. This algorithm takes advantage of several biological sources to build a dataset of heterogeneous features. Such dataset was used to train a support vector machine in order to predict whether a particular set of sequences could be accurately aligned by any of the ten proposed methodologies.

The proposed prediction was then assessed by using the BAliBASE benchmark. In this case, the SVM approach achieved to correctly predict the quality of the alignments (accurate vs inaccurate) in 88.56% of the cases for the training dataset and 79.77% for the test dataset (using the 14 most relevant features). Although these results are still preliminary, the accuracy percentages suggested that this algorithm was effectively predicting the MSA tools providing the most accurate alignments for a particular dataset of sequences.

References

1. Apweiler,R., Bairoch,A., Wu,C.H., Barker, W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H.Z., Lopez,R., Magrane,M., Martin,M.J., Natale,D.A., O'Donovan,C., Redaschi,N. and Yeh,L.S.L. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res.*, 32:D115–D119, 2004.
2. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

3. R. K. Bradley, A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, and L. Pachter. Fast Statistical Alignment. *PLoS Computational Biology*, 5(5), 2009.
4. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res.*, 32:D262–D266, 2004.
5. C. Do, M. Mahabhashyam, M. Brudno, and S. Batzoglou. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340, 2005.
6. R. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.
7. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K., Holm,L., Sonnhammer,E.L.L., Eddy,S.R. and Bateman,A. The pfam protein families database. *Nucleic Acids Res.*, 38:D211–D222, 2010.
8. K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.
9. T. Lassmann and E. Sonnhammer. Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6, 2005.
10. H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.
11. C. Notredame, D. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, 2000.
12. O. O'Sullivan, K. Suhre, C. Abergel, D. Higgins, and C. Notredame. 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology*, 340(2):385–395, 2004.
13. J. Pei and N. V. Grishin. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, 23(7):802–808, 2007.
14. Peng,H., Long,F. and Ding,C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1226–1238, 2005.
15. A. Szabo, A. Novak, I. Miklos, and J. Hein. Reticular alignment: A progressive corner-cutting method for multiple sequence alignment. *BMC Bioinformatics*, 11, 2010.
16. J. Thompson, D. Higgins, and T. Gibson. ClustalW: Improving the sensivity of progressive multiple sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.
17. J. Thompson, P. Koehl, R. Ripp, and O. Poch. BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins-Structure Function and Bioinformatics*, 61(1):127–136, 2005.