Arturo Chavoya, et al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 11, Issue 2, (Series-VI) February 2021, pp. 59-65

RESEARCH ARTICLE OPEN ACCESS

Characterization of Protein Sequences Aligned with MUSCLE using Guide Trees from SARELI

Arturo Chavoya*, Ricardo Ortega**
**Corresponding Author: Department of Information Systems – CUCEA, Guadalajara University, Zapopan, Jalisco, Mexico, achavoya@cucea.udg.mx*
***Department of Information Systems – CUCEA, Guadalajara University, Zapopan, Jalisco, Mexico*

ABSTRACT
SARELI is a software tool that can generate guide trees for multiple sequence alignments of protein sequences. These guide trees are generated using a metric named Radial Distance. The guide trees producedby SARELI were fed into MUSCLE to proceed with the rest of the alignment procedure. The resulting alignments were compared against the results from MUSCLE (with its original guide trees), Clustal Omega, three variants of MAFFT, and T-Coffee on the BAliBASE 3, PREFAB 4.0, and SABRE protein sequence databases. The sum of pairs score, and the column score were used for scoring the results against the reference alignments of the three protein benchmark databases. SARELI can be used as a specialized tool for generating guide trees that in conjunction with MUSCLE can obtain significantly better multiple sequence alignment scores than the other MSA methods tested when the set to be aligned contains more than 35 protein sequences, and its average sequence length and p-distance are less than 239 and 0.81, respectively. As future work we would like to improve the performance of the SARELI algorithm.

Keywords–SARELI, multiple sequence alignment, guide trees, MUSCLE, protein benchmark databases

---

Date of Submission: 16-02-2021 Date of Acceptance: 02-03-2021

I. INTRODUCTION

The multiple sequence alignment (MSA) of related proteins can help predict both protein structure and function, and can also shed light on the phylogenetic relationship of species. However, despite significant advances in the performance of alignment algorithms, finding consistently accurate alignments can prove elusive [1].

In essence, MSA algorithms start with a series of possibly related biological sequences (proteins or nucleic acids) and try to obtain a set of sequences of the same length that matches as many homologous symbols (representing amino acids or nucleotides) as possible from the initial sequences; gap symbols can be introduced to displace the columns of the sequences in order to obtain a better alignment[2].

MSA algorithms normally initiate by comparing a pair of sequences at a time from the original set, and the aligned sequence resulting from the pair alignment is then compared against other sequence in the set, or other aligned sequence from a pair comparison [2]. The order in which the sequences in the set are to be compared in pairs is determined by a guide tree, which specifies the order of pair comparisons, usually starting with the most similar pair and progressing with the most dissimilar. A robust guide tree is essential for an accurate final alignment and can be obtained by applying a consistent and precise metric at the start of the clustering algorithm; failing to correctly identify the sequences most closely related can have a negative impact on the final alignment. Additional heuristics and metrics can be used to refine the guide tree with intermediate steps in the process in order to generate a better alignment [3]–[5].

SARELI, which stands for Sequence Alignment by Radial Evaluation of Local Interactions, is a software tool that uses a metric named Radial Distance for the production of guide trees that are used in protein MSA algorithms [6]. The Radial Distance metric considers the effect that adjacent symbols within a certain radius can have on the different symbols to be aligned for the construction of the initial distance matrix that leads to the final guide tree. On the other hand, in addition to their own guide trees, MSA tools such as MUSCLE [4] and Clustal Omega [7] can use external guide trees, such as those generated by SARELI, as input to construct a final alignment [6].

In a previous report it was shown that when MUSCLE uses the guide trees from SARELI, it can generate statistically better sum of pairs and column scores of alignments on the SABRE and PREFAB protein benchmark databases, than when MUSCLE

uses its original guide trees. However, when comparing MUSCLE coupled with SARELI against other MSA methods, such as Clustal Omega, MAFFT and T-Coffee, mixed results were obtained [6].

In this paper we further explore the characteristics of the sequences from the three benchmark databases for which MUSCLE coupled with SARELI can produce statistically better scores than the MSA methods mentioned above, with the intent of understanding the strengths and weaknesses of the SARELI algorithm so as to improve its performance in the future.

II. METHODS

This section describes the Radial Distance (RD) metric used for the construction of the initial distance matrix, which in turn is used by the Neighbor Joining algorithm to produce the guide trees fed into MUSCLE. The protein benchmark databases used in the present study are also described, as well as the definition of the scores used to evaluate the alignments, and the generation of guide trees.

2.1 Radial Distance

The Radial Distance is a metric which has been previously reported, and that assesses the distance between two sequences, considering not only the symbol in the column to be aligned, but also the symbols surrounding the column [6]. The Radial Distance takes a radius parameter value that considers a number of symbols around each column when doing the pairwise alignment of two sequences. As the distance from the referenced column increases, the effect on the score is decreased with an asymptotic function. As previously reported, the Radial Distance (RD) between sequences A and B is defined as

$$RD(A,B) = \sum_{i=1}^{M} \sum_{j=i-R>0}^{i+R\leq M} \frac{Score\ (A_i,B_j)}{|i-j|+1}, \qquad (1)$$

where M is the length of the initially aligned sequences, and R is the radius value that indicates how far the adjacent columns will influence the score [6]. The *Score* function used was the BLOSUM62 substitution matrix [8].

Figure 1 shows an example of the calculation of the RD for two sequences for a radius value of 2, and until the sixth step of the comparison process; the partial calculation of the RD values is presented up tothat point, where *nc* represents an RD value not yet calculated. At the end of the process, the Radial Distance between the two sequences is computed as the addition of all the individual values.

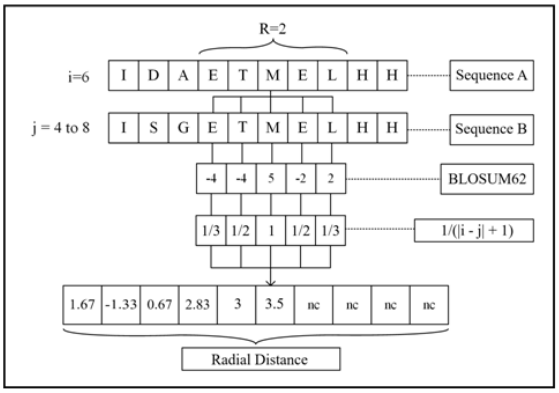

Fig.1.Example of the partial calculation of the Radial Distance for sequences A and B

2.2 Databases

Three different benchmark databases were used in this paper to explore the characteristics of the sequences aligned with the guide trees from SARELI: BAliBASE 3 [9], PREFAB 4.0 [4] andSABRE [10].

The BAliBASE database was manually designed as an evaluation resource for addressing problems that arise when aligning complete sequences [11] and has been widely used for testing and comparison of related protein sequences [4], [12]–[14]. PREFAB is a database designed by applying an automated protocol that uses known methodologies, test data, and statistical methods [4]. Finally,SABRE is a set of sequences derived from the SABmark database [15] that was selected to be used in MSA comparisons [10].

For the characterization of these databases, four main aspects from the point of view of symbols were considered. As a first step, the number of files per database was counted; Table 1 shows that PREFAB has the largest number of sequence sets, followed by SABRE, and with BAliBASE having the smallest number of files.

Table 1. Number of files per database

Database	Number of files
BAliBASE	386
PREFAB	1682
SABRE	423

The number of sequences per file in the databases was the second aspect considered. The distributions presented in Table 2 were generated by sorting in ascending order the number of sequences per file for each database and making a numerical regression to obtain the corresponding equation with its *R*-squared value; Fig. 2 shows a graphical representation of these distributions.

Table 2.Number of sequences per file

Database	Distribution	R^2
BAliBASE	$e^{1.174+0.009x}$	0.9941
PREFAB	First 400:$-7.43138+2.77544\sqrt{x}$ Rest: 50	0.9672
SABRE	$e^{1.01873+0.00000948672\,x^2}$	0.9629

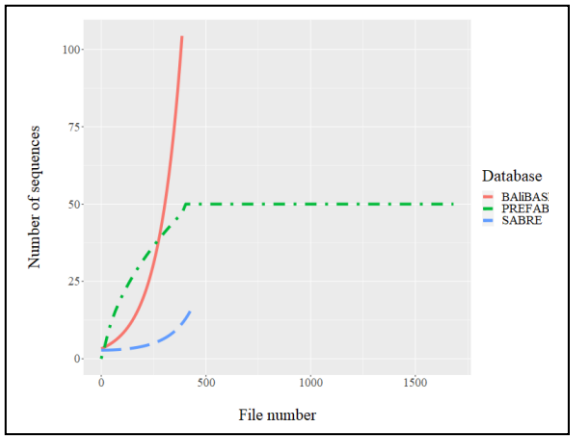

Fig. 2.Distributions of the number of sequences per file

The third aspect considered for the databases was the average length of the sequences per file. As the process described for the previous characteristic, the average length per file for each database was sorted in ascending order and a numerical regression was made to calculate the distributions presented in Table 3. A graphical representation of these distributions is shown in Fig. 3.

Table 3.Average length of sequences per file

Database	Distribution	R^2
BAliBASE	$e^{3.67227+0.148005\sqrt{x}}$	0.9828
PREFAB	$e^{4.38954+0.00109666\,x}$	0.9801
SABRE	$e^{4.0504+0.00444247\,x}$	0.9592

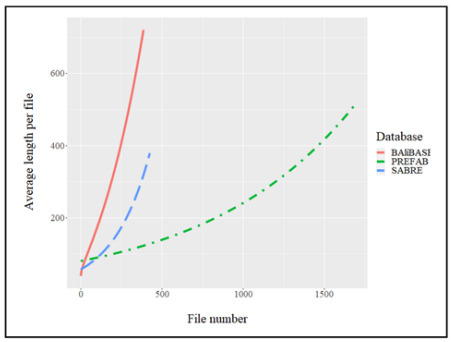

Fig. 3.Distributions of the average length of sequences per file

The last characteristic considered was the p (proportion) distance [16], which measures the degree of sequence divergence, even when sequences have a different length. The p-distance for a pair of sequences was calculated by dividing the number of amino acid differences by the total number of amino acids compared; the MEGA software was used to calculate the p-distances [17]. An average of the p-distances between all pairs of sequences in the set was calculated for every file and the results are presented in Table 4; these distributions are graphically shown in Fig. 4.

Table 4.Average proportion distance between all pairs of sequences per file

Database	Distribution	R^2
BAliBASE	$\sqrt{0.24792+.00114021x}$	0.9584
PREFAB	$-0.4495+0.1856\cdot\ln(x)$	0.9432
SABRE	$\sqrt{0.2646+0.0290\sqrt{x}}$	0.9896

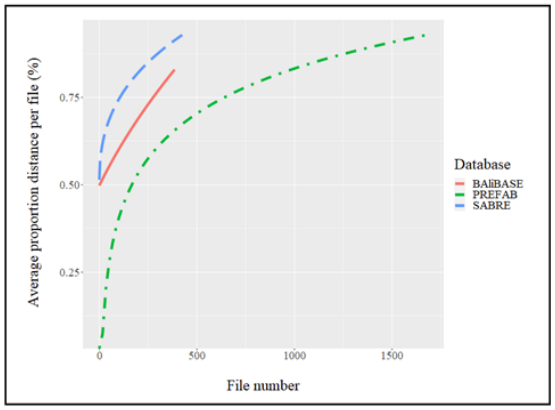

Fig. 4.Distributions of the average proportion distance of sequences per file

2.3Scoring

The multiple sequence alignments were evaluated by the sum of pairs, which is a score commonly used [18]–[23], and wascomputed by adding all the possible pair comparisons from each column of the alignment set, without repetition. For the sake of comparing the quality of an alignment against a benchmark reference, the sum of pairs score (SPS) for the set of sequences *A*was calculated as

$$SPS(A)=\frac{\sum_{i=1}^{M}S_i}{\sum_{i=1}^{Mr}S_{ri}}, \qquad (2)$$

where *M*is the length of the alignment, *Mr* is the number of columns in the reference alignment, and $S_{ri}$is the S_i score in the reference alignment, which was calculated by

$$S_i = \sum_{j=1, j\neq k}^{N} \sum_{k=1}^{N} p_{ijk} \ , \quad (3)$$

where N is the number of sequences in the file and, if A_{i1}, A_{i2}, …, A_{iN} is the i-th column of alignment A, then for each pair of amino acids $A_{ij}$and A_{ik}, p_{ijk} is defined as 1 if A_{ij} and A_{ik} are aligned with each other in the reference alignment file, and 0 otherwise [24].

Other common evaluation measure for MSA is the column score (CS), which measures the ratio of completely aligned columns against the reference alignment. This score is commonly used along with SPS, in order to evaluate the quality of the alignments. The column score (CS) is calculated as

$$CS(A) = \frac{\sum_{i=1}^{M} C_i}{M}, \quad (4)$$

where A is the set of sequences, M is the length of the aligned sequences, and $C_i = 1$ if all the residues in the i-th column are aligned, and 0 otherwise [24].

2.4Generation of guide trees

The process followed by SARELI for generating the guide tree for each file of the three databases (BAliBASE, PREFAB, andSABRE) has been previously described [6]. Briefly, using the Radial Distance, an initial distance matrix of the two sequences to be compared is built, testing with a range of radius values from 3 to 10. Next, the matrix is input to the Neighbor Joining algorithm to generate the corresponding guide tree in PHY format, which is then fed to MUSCLE to be used instead of its default guide tree. A command line example of the use by MUSCLE of an external guide tree in PHY format applied to a sequence set in TFA format to generate an alignment in FASTA format is

muscle -usetree "name.phy" -in "name.tfa" -out "name.fasta"

where *name* is the name of the sequence set file.

For evaluating the quality of the alignments, the sum of pairs score (SPS) and the column score (CS) were calculated—as described in Section 2.3—on the alignments from the BAliBASE, PREFAB, and SABRE benchmark databases; for the calculation of the scores, the QSCORE software was used [25]. Each sequence set from these databases was aligned using the guide trees generated by SARELI and fed into MUSCLE, and their scores were compared against those from MUSCLE [4], Clustal Omega [7], MAFFT [26], and T-Coffee [27]. In the case of MAFFT, three different versions of the algorithm were used: E-INS-i, G-INS-i, and L-INS-I [28], which were named in the present paper as MAFFT GE, MAFFT GL, and MAFFT LO, respectively.

All the runs were performed on a PC computer with one 8-core 3.52-GHz FX-8320 AMD processor, 16 GB of RAM, and an NVIDIA GeForce GT 730 card. The operating system used was Windows 10, whereas the library containing the alignment algorithm for SARELI was coded in C# using the Visual Studio 2013 IDE and compiler. The SARELI source files are freely available and can be found at [29]. The latest SARELI executable file for Windows 10 can be downloaded from [30] and is released under the MIT License. Finally, the statistical analysis was performed using STATGRAPHICS Centurion XVI.

III. RESULTS AND DISCUSSION

The sequence files in the BAliBASE3, PREFAB 4.0, and SABRE databases were aligned using SARELI coupled with MUSCLE (termed SARELI & MUSCLE for the presentation of the results) and the same was done with MUSCLE (with its default guide trees), Clustal Omega, the three variations of MAFFT, and T-Coffee. For each alignment file obtained, the sum of pairs score (SPS) and the column score (CS) were calculated, as defined in Section 2.3, against the reference alignments for each database, using the QSCORE software [25].

In order to explore the characteristics of the sequence sets under which SARELI & MUSCLE could perform better than the other tested methods, different combinations of the features described in Section 2.2 of the protein databases were performed. We found one set of characteristics in which the alignments from SARELI & MUSCLE were statistically better than the rest of the methods for both SPS and CS. This set of characteristics was when the number of sequences in the set was above 35, the average length of the sequences was below 239 symbols, and the average p-distance was below 0.81. After applying the above filter, the number of files obtained was 39 for BAliBASE, 421 for PREFAB, and 0 files from SABRE for a total of 460 sequence sets. None of these datasets showed a normal distribution; thus, we used the Friedman test to verify a statistically difference between the medians of the samples, and the results are presented in Table 5.

Table 5.Filtered dataset medians

Method	Score	
	SPS	CS
SARELI & MUSCLE	0.9250	0.9120
MUSCLE	0.8970	0.8800
Clustal Omega	0.8980	0.8870
MAFFT GE	0.9140	0.9000
MAFFT GL	0.9120	0.8950
MAFFT LO	0.9150	0.9000
T-Coffee	0.9070	0.8930

Since the *p*-values indicated that at least one of the samples was statistically different, we used the Wilcoxon test per pairs to determine which of the medians was different, obtaining the *p*-values presented in Table 6.

Table 6.Wilcoxon test *p*-values for SARELI & MUSCLE with the filtered dataset

Method	Score	
	SPS	CS
MUSCLE	0.0000	0.0000
Clustal Omega	0.0000	0.0000
MAFFT GE	0.0003	0.0018
MAFFT GL	0.0000	0.0001
MAFFT LO	0.0012	0.0149
T-Coffee	0.0000	0.0000

From these two tables, it can be seen that SARELI & MUSCLE showed statistically better results at 99% of confidence on the SPS and CS scores when compared against all of the other MSA methods, except for MAFFT LO with CS, in which case the level of confidence was 95%.

The box-and-whisker plots with median notchfrom the Friedman test for both SPS and CS are presented in Fig. 5 and Fig. 6, respectively.

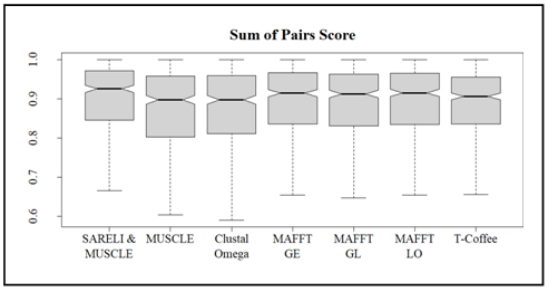

Fig. 5.Box-and-whisker plot from the Friedman test on the sum of pairs score

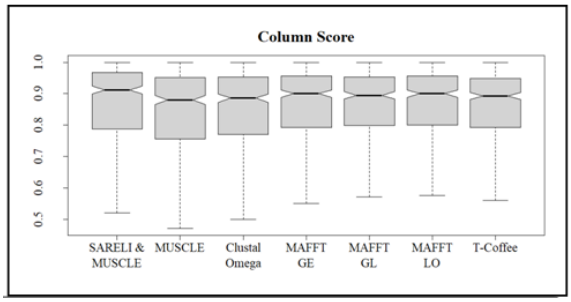

Fig. 6.Box-and-whisker plot from the Friedman test on the column score

IV. CONCLUSION

We compared the alignments produced by the guide trees from SARELI in conjunction with MUSCLE against the well-known MSA programs MUSCLE (with its default guide trees), Clustal Omega, three variants of MAFFT, and T-Coffee, using the BAliBASE 3, PREFAB 4.0, and SABRE protein sequence databases. We compared the resulting alignments using the sum of pairs score and the column score with the reference files provided with these benchmark databases. Even though the guide trees generated with SARELI have been reported to yield mixed results in the comparisons against the other multiple sequence alignment methods, when coupling SARELI with MUSCLE and using sequence sets with more than 35 sequences, an average length below 239 and an average p-distance below 0.81 in BAliBASE and PREFAB, the alignments obtained were statistically better than those from all of the other methods for these databases, both in sum of pairs score and column score. This set of characteristics will in principle help us improve the performance of SARELI by fine-tuning the algorithm. In the meantime, we recommend using SARELI as a specialized tool for generating guide trees fed into MUSCLE to obtain significantly better alignments when dealing with sequence sets that are close to the above restrictions.

As future work, we would like to improve the performance of SARELI by trying different values for the parameters controlling the algorithm. We would also like to further explore the characteristics of additional protein sequence databases in order to confirm or extend the results obtained with our approach. We would also like to compare against other MSA methods to expand the knowledge on the type of sequences for which the combination of SARELI and MUSCLE can provide accurate alignments.

REFERENCES

[1] K. Katoh and H. Toh, "Recent developments in the MAFFT multiple sequence alignment program.," *Brief. Bioinform.*, vol. 9, no. 4, pp. 286–298, Jul. 2008, doi: 10.1093/bib/bbn013.

[2] D.-F. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *J. Mol. Evol.*, vol. 25, no. 4, pp. 351–360, 1987, doi: 10.1007/BF02603120.

[3] M. A. Larkin *et al.*, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[4] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, Mar. 2004.

[5] F. Naznin, R. Sarker, and D. Essam, "Iterative progressive alignment method (IPAM) for multiple sequence alignment," in *2009*

International Conference on Computers Industrial Engineering, 2009, pp. 536–541, doi: 10.1109/ICCIE.2009.5223562.

[6] R. Ortega, A. Chavoya, C. López-Martín, and L. Delaye, "SARELI : Sequence Alignment by Radial Evaluation of Local Interactions," *Curr. Bioinform.*, vol. 13, no. 3, pp. 290–298, 2018, doi: 10.2174/1574893613666180130143055.

[7] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, and W. Li, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Mol Syst Biol*, vol. 7, 2011, doi: 10.1038/msb.2011.75.

[8] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 22, pp. 10915–9, Nov. 1992.

[9] J. D. Thompson, F. Plewniak, and O. Poch, "BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs," *Bioinformatics*, vol. 15, no. 1, pp. 87–88, 1999, doi: 10.1093/bioinformatics/15.1.87.

[10] R. C. Edgar, "Quality measures for protein alignment benchmarks," *Nucleic Acids Res.*, vol. 38, no. 7, pp. 2145–2153, 2010, doi: 10.1093/nar/gkp1196.

[11] A. Bahr, J. D. Thompson, J.-C. Thierry, and O. Poch, "BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations," *Nucleic Acids Res.* , vol. 29, no. 1, pp. 323–326, Jan. 2001.

[12] M. A. Larkin *et al.*, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, Nov. 2007.

[13] K. Karplus and B. Hu, "Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set," *Bioinformatics*, vol. 17, no. 8, pp. 713–720, 2001.

[14] T. Lassmann and E. L. Sonnhammer, "Quality assessment of multiple alignment programs," *FEBS Lett.*, vol. 529, no. 1, pp. 126–130, Oct. 2002, doi: 10.1016/S0014-5793(02)03189-7.

[15] I. Van Walle, I. Lasters, and L. Wyns, "SABmark—a benchmark for sequence alignment that covers the entire known fold space," *Bioinformatics*, vol. 21, no. 7, pp. 1267–1268, Apr. 2005.

[16] R. H. Thomas, "Molecular Evolution and Phylogenetics," *Heredity (Edinb).*, vol. 86, no. 3, p. 385, 2001, doi: 10.1046/j.1365-2540.2001.0923a.x.

[17] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.," *Mol. Biol. Evol.*, vol. 35, no. 6, pp. 1547–1549, Jun. 2018, doi: 10.1093/molbev/msy096.

[18] J. D. Thompson, F. Plewniak, and O. Poch, "A comprehensive comparison of multiple sequence alignment programs," *Nucleic Acids Res.*, vol. 27, no. 13, pp. 2682–2690, 1999.

[19] I. Van Walle, I. Lasters, and L. Wyns, "Align-m-a new algorithm for multiple alignment of highly divergent sequences," *Bioinformatics*, vol. 20, no. 9, pp. 1428–1435, 2004, doi: 10.1093/bioinformatics/bth116.

[20] J. Stoye, V. Moulton, and A. W. M. Dress, "DCA: An efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment," *Bioinformatics*, vol. 13, no. 6, pp. 625–626, Dec. 1997.

[21] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu, "A tool for multiple sequence alignment," *Proc. Natl. Acad. Sci.*, vol. 86, no. 12, pp. 4412–4415, 1989.

[22] C. Lee, C. Grasso, and M. F. Sharlow, "Multiple sequence alignment using partial order graphs," *Bioinformatics*, vol. 18, no. 3, pp. 452–464, 2002.

[23] Q. Zhan, Y. Ye, T.-W. Lam, S.-M. Yiu, Y. Wang, and H.-F. Ting, "Improving multiple sequence alignment by using better guide trees," *BMC Bioinformatics*, vol. 16, no. Suppl 5, pp. S4–S4, Mar. 2015, doi: 10.1186/1471-2105-16-S5-S4.

[24] J. D. Thompson, F. Plewniak, and O. Poch, "A comprehensive comparison of multiple sequence alignment programs," *Nucleic Acids Res*, vol. 27, 1999, doi: 10.1093/nar/27.13.2682.

[25] R. C. Edgar, "QSCORE multiple alignment scoring Software." https://www.drive5.com/qscore/ (accessed Nov. 16, 2020).

[26] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.

[27] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: a novel method for fast and accurate multiple sequence alignment," *J. Mol. Biol.*, vol. 302, no. 1, pp. 205–217, 2000, doi: 10.1006/jmbi.2000.4042.

[28] K. Katoh and D. M. Standley, "MAFFT ver.7 - a multiple sequence alignment program."

https://mafft.cbrc.jp/alignment/software/.
[29] R. Ortega, "SARELI Source code," 2016. https://github.com/icariantk/SARELI (accessed Nov. 18, 2020).
[30] R. Ortega, "SARELI Windows Binary," 2016. https://raw.githubusercontent.com/icariantk/SARELI/master/SARELI/bin/Debug/SARELI.exe (accessed Nov. 18, 2020).

Arturo Chavoya, et. al. "Characterization of Protein Sequences Aligned with MUSCLE using Guide Trees from SARELI."*International Journal of Engineering Research and Applications (IJERA)*, vol.11 (2), 2021, pp 59-65.