TRANSLATIONAL RESEARCH IN VETERINARY SCIENCE
Vol 3, No 2, 2020
DOI: http://dx.doi.org/10.12775/TRVS.2020.009
ISSN 2657-4489 (online)

Benchmark of algorithms for multiple DNA sequence alignment across livestock species

Artur Bąk[1], Grzegorz Migdałek[2], Chandra S. Pareek[3], Kacper Żukowski[§,1,3]

[1]Department of Cattle Breeding and Genetics, National Research Institute of Animal Production, Balice, Poland; artur.bak@izoo.krakow.pl, kacper.zukowski@umk.pl, https://orcid.org/0000-0002-5690-3634
[2]Institute of Biology, Pedagogical University of Cracow, Poland; grzegorz.migdalek@up.krakow.pl, https://orcid.org/0000-0003-1458-2673
[3]Department of Fundamental and Preclinical Sciences, Faculty of Biology and Veterinary Sciences, Nicolaus Copernicus University in Toruń, Poland. pareekcs@umk.pl, https://orcid.org/0000-0002-0329-787X, kacper.zukowski@umk.pl, https://orcid.org/0000-0002-5690-3634

[§]**Corresponding author**: dr Kacper Żukowski, Adjunct, Department of Fundamental and Preclinical Sciences, Faculty of Biology and Veterinary Sciences, Nicolaus Copernicus University, ul. Gagarina 9, 87-100 Toruń, Poland. kacper.zukowski@umk.pl

Abstract

Background: Due to the growing amount of biological data, it is often necessary to select the most optimal estimation method for DNA sequence alignment across livestock spe-

cies. One of the most important benches of genomics is to modelling homology between considered DNA sequences. A multiple sequence alignment is a potent tool for molecular and evolutionary biology, and there are several programs and algorithms applicable for this purpose. The purpose of this paper was to study the most commonly used DNA alignment algorithms to select the optimal tool dedicated for short sequences.

Methods: Four steps of bioinformatics pipelines were considered to benchmark the algorithms for multiple DNA sequence alignment across livestock species: 1) selection of reference genome sequences of ARS1.2 for cattle, EquCab3.0 for horse and vicPac2 for alpaca with a low E-value using TBLASTn 2) removing gaps for these sequences 3) alignment of obtained sequences using examined algorithms 4) matching the quality of aligned sequences with sequences of reference genomes by more software. The time of computation was archived for the whole analysis. The seven programs were utilized, each based on different alignment algorithms, namely: ClustalO, ClustalW, Kalign, MAFFT, MUSCLE, Probcons and T-Coffee.

Results: The result obtained in this study showed that the fastest is progressive algorithms such as Kalign or MUSCLE-FAST. Moreover, the iterative algorithms like MAFFT and MUSCLE revealed a higher quality of the alignment. The T-Coffee and Probcons programs were computational cost-effective; simultaneously, they were generating a medium-quality calculation in a relatively long time. The best quality of alignment was shown by iterative variants of the MAFFT program; however, the speed of the calculations was relatively low. The fastest algorithm was Kalign, making alignment much faster than the competitors, but achieving average results in the quality of the alignment. The average speed ratio concerning the quality of the analyzed algorithms was obtained by the progressive version of MAFFT, NS1.

Conclusions: We conclude that the results of this study can be used to re-alignment of variant primers in new livestock genome releases.

Keywords: multiple sequence alignment; ClustalO; ClustalW; Kalign; MAFFT; MUSCLE; Probcons and T-Coffee; bioinformatics pipeline; livestock.

Introduction

Advances in genome sequencing have progressed at a rapid pace, with increased throughput accompanied by plunging costs. However, these advances go far beyond faster and cheaper [**1**]. The high-throughput (HT) next-generation genome sequencing (NGS) technologies are now routinely being applied to a wide range of important topics in biological, medical and, veterinary sciences [**2**]. The big challenges of these rapidly developed of NGS technology are the proper selection and utilization of

advanced bioinformatics pipelines and NGS related tools [**3**]. Advent of these NGS technologies have enabled the mapping of high throughput sequence that were previously not possible on a genomic scale and resolution. In our previous paper, we have analyzed BWA, Bowtie2 and SMALT mapping tools using three different NGS sequencing platform [**4**]. In this paper, we have the analyzed and calculated the performance of the most popular algorithms that compare multiple DNA sequences in order to assess their suitability for use in alignment programs.

Methodology

A special workstation has been created for the calculation to provide the same environment for each algorithm. The following alignment programs: Clustal [**5–6**], ClustaiW [**7-9**], Kalign [**10**], MAFFT [**11**], MUSCLE [**12**], Probcons [**13**] and T Coffee [**14**] were utilized to analyze the alignment algorithm. The reference sequences were utilized from the available MDS database [**15**] and from the prepared sequences available in the genetic maps of the tested animals. Finally, comparison of seven alignment programs were performed using more than 100 selected sequences has the gold standard (**Figure 1**). For each alignment program, the sequences were compared to the standard ones and their quality was calculated. The resulting equations were then compared with the reference sequences using the software program (**Figure 2**). The CS factor is the main criterion determining the quality of the equation. The second, equally important criterion, is the central processing unit (CPU) time, which determines how long the program executed the sequence.

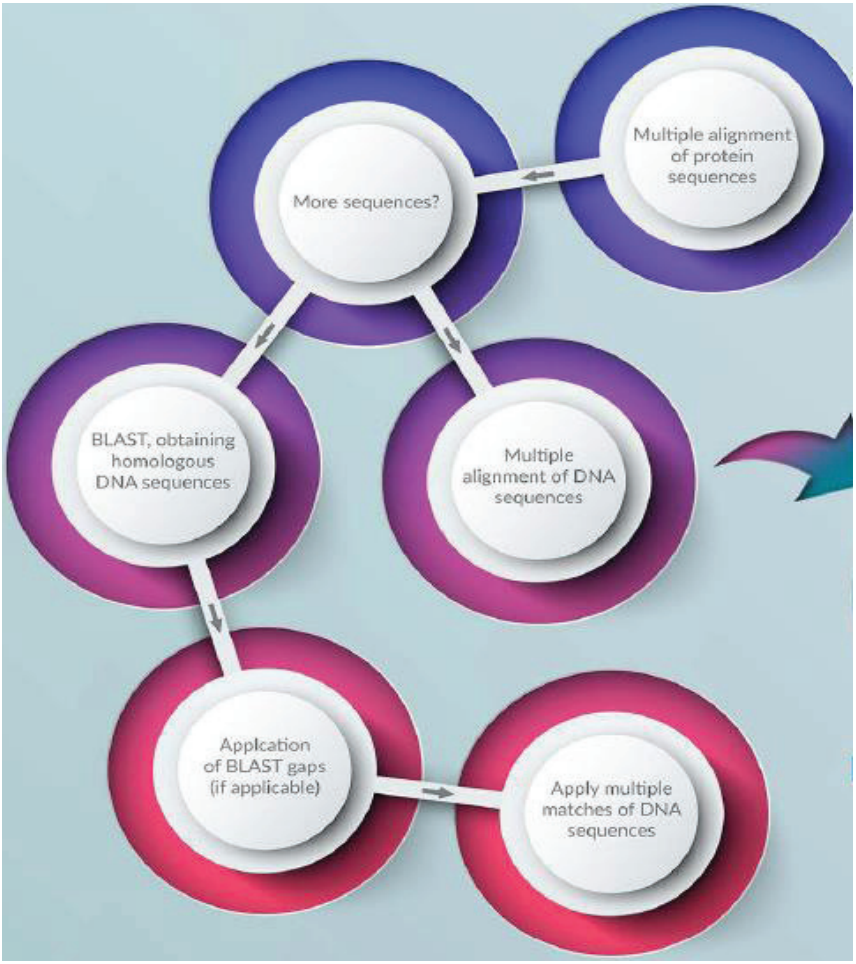

Figure 1. Sequence preparation processing events for multiple DNA sequence alignment across livestock species.

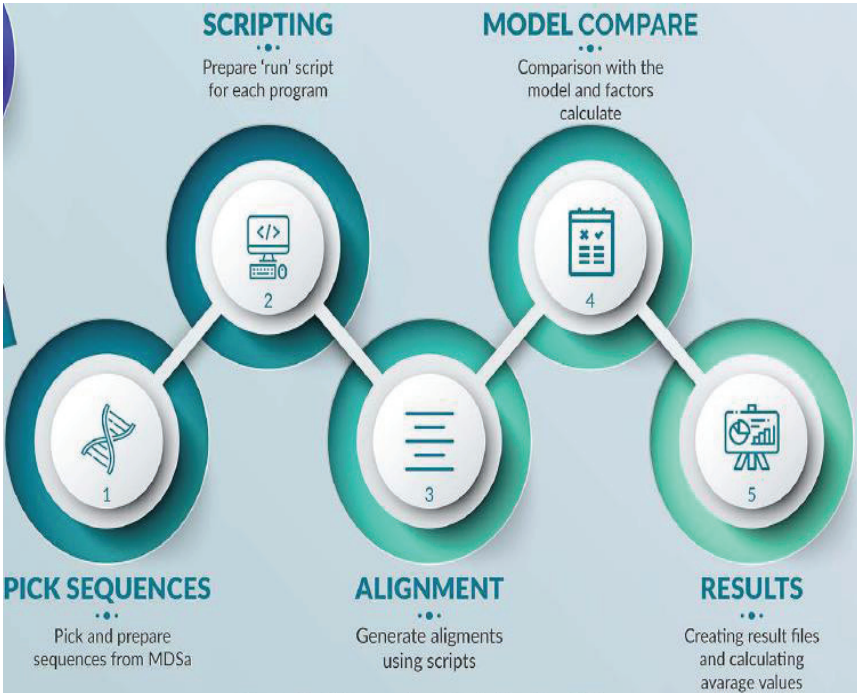

Figure 2. Scheme for calculating the benchmarks of alignment score for multiple DNA sequence alignment across livestock species.

Results and discussions

The result of the calculations are the arrays containing the average CS coefficient and the average time of execution of each of the programs on the basis of which the graph was obtained (**Figure 3**). The result obtained in this study show that the fastest are progressive algorithms of Kalign or MUSCLE-FAST. Moreover, the iterative algorithms like MAFFT and MUSCLE have a higher quality of alignment score. The best quality/time ratio was found in the NS1 variant of the MAFFT algorithm. Probcons was the worst performer, which despite the good quality of the equation turned out to be ineffective due to the time of performing calculations.

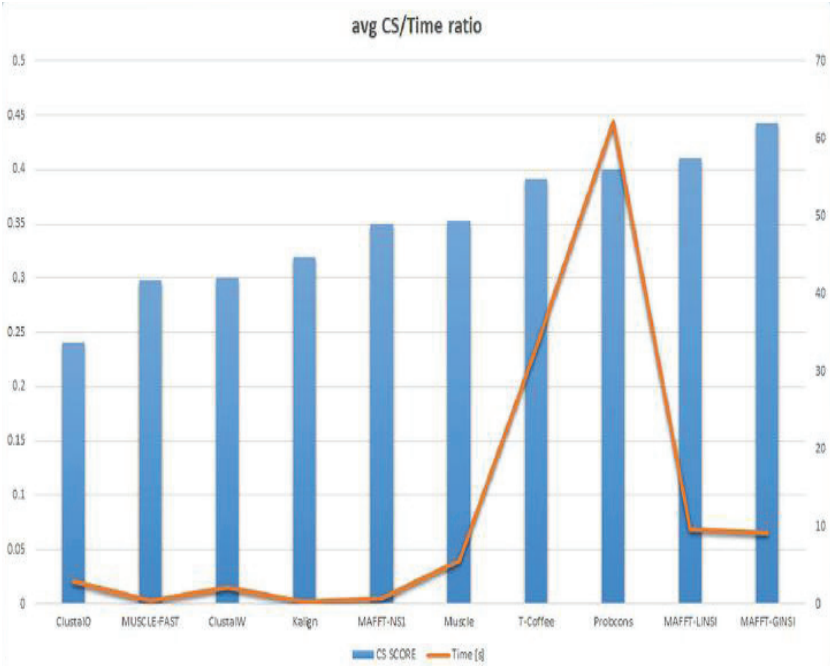

Figure 3. Estimates of the average CS/CPU time ratio for multiple DNA sequence alignment across livestock species.

Conclusions: Study conclude that the results of this study can be used to re-alignment of variant primers in new livestock genome releases.

References

[**1**] Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. Mol Syst Biol. 2013;9:640.
[**2**] Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52:413–35.
[**3**] Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J. The next-generation sequencing technology and application. Protein Cell. 2010;1:520–36.

[**4**] Bąk A, Bodziony D, Migdałek G, Pareek CS, Żukowski K. Evaluation of analytical protocols of alignment mapping tools using high throughput next-generation genome sequencing data. Transl Res Vet Sci. 2020;3:62–65.
[**5**] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.
[**6**] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 1997;25:4876–82.
[**7**] Higgins DG, Bleasby AJ, Fuchs R. CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci. 1992;8:189–91.
[**8**] Sievers F, Higgins DG. Clustal omega. Curr Protoc Bioinformatics. 2014;48:3.13.
[**9**] Sievers F, Higgins DG. The Clustal Omega Multiple Alignment Package. Methods Mol Biol. 2021;2231:3–16.
[**10**] Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005;6:298.
[**11**] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–3066.
[**12**] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–1797.
[**13**] Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 2005;15:330–340.
[**14**] Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang JM, Taly JF, Notredame C. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 2011;39:W13–7.
[**15**] Carroll H, Beckstead W, O'Connor T, Ebbert M, Clement M, Snell Q, McClellan D. DNA reference alignment benchmarks based on tertiary structure of encoded proteins. Bioinformatics. 2007;23:2648–9.