Assessment on The Performance of Pairwise Sequence Aligmment Programs

Farhan Tahir and Iskandar Ilyas Tan

[1]Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia

farhanmohdtahir@gmail.com, iskandar@utm.my

Abstract. Aim of this research was to assess the performance of pairwise sequence alignment (PSA) programs in terms of their accuracy, sensitivity and specificity in aligning simulated sample sequence into known porcine sequence. The problem that sparks an idea to carried out this research is because the presence of many PSA programs causing the performance of each PSA program was questioned. In order to identify the performance of each PSA programs, the result of alignment sequence from PSA programs was compared to the reference alignment sequence created from the alignment between duroc sp and simulated sequence. The findings of this research was EMBOSS Stretcher was the most top perform while EMBOSS Needle is the most consistent, in terms of their accuracy, sensitivity and specificity.

Keywords: Sequence alignment, pairwise sequence alignment, needleman-wunsch, global alignment.

1. Introduction

In this fast era, bioinformatics had become the most important platform to solve many biological problem. One of it was sequence alignment method which aligning sequence to identify the similarity between sequences, so, the relationship between the sequences can be determined.Pairwise sequence alignment (PSA) is one of the method under sequence alignment method. Recently, PSA was used to identify the presence of porcine DNA in the ingredient sample by aligning the genomic sequence of the ingredient sample to the known porcine, cytochrome b sequence. But, the problem was there were many PSA programs that available and can be used which causing their performance was questioned in identifying the similarity between the porcine DNA and ingredient sample sequence. Hence, this research was carried out to assess the performance of the PSA programs which had been classified as give high accurate alignment sequence result in terms of their accuracy, sensitivity, specificity and consistency. The objectives of this research was to to develop a C++ program which produce simulated sample's and reference alignment's file of Duroc sp, cyt b, to develop a C++ program which can calculate the accuracy, sensitivity and the specificity of the programs result in respect to reference alignment sequence and To plot the accuracy, sensitivity and the specificity's graphs of each PSA program. This may help the researchers to identify which PSA programs was fit to be implemented in their research and help bioinformatics researchers to develop a methodology on assessing the sequence alignment programs. The scope of this research was it use Duroc sp, cytb as reference sequence and Duroc sp simulated sequence from alignerReference program. PSA programs that was assessed is ALIGN Query [1], NW Blast[2], EMBOSS Needle [3], NW [4] and EMBOSS Stretcher[5].

2. Literature Review

Based on research done by Nuin *et al* [6], the research was carried out to assess the performance of nine MSA programs which are ClustalW, Dialign, T-Coffee, POA, Muscle, MAFFT, ProbCons, Dialign-T and Kalign. The performance of the MSA programs was assessed in terms of its accuracy and the speed of programs execution. This research used SIMPROT to create simulated sequence and reference alignment sequence and BaliBase which act as the benchmark of the MSA. The limitation of this research was the probability of the result to be bias was high. This is because the performance of MSA program result was assessed from accuracy only, but, there is no assessement on its sensitivity and specificity.

Research done by Penn *et al* [7] was carried out to assess the performance of the six MSA performance result, which, ClustalW, Dialign, Muscle, ProbCons, ProDa, T-Coffee. İn order to assess the performance, this research identified the accuracy, sensitivity, specificity and the speed of the MSA programs. The disadvantage of this research was the MSA programs was assessed by using default parameter instead of constant user-define parameters value.

Lastly, development done by Pang *et al* [8] was referred to see on how they develop SIMPROT program which can be used to create simulated sequence and reference alignment file. Pang *et al* develop a program which will included the insertion, deletion, substitution on simulated sequence. From that, a reference alignment sequence was developed which can captured evolutionary relationship between amino acids.

Hence, this research was carried out by implementing the methodology done on assessing the performance of MSA programs to assess the PSA programs with some improvement.

3. Methodology

The research framework consist of four phases which are problem identification and specification, data preparation and factor determination, development of programs and analysis of results. In problem identification and specification, this research define the problem background which cause an idea for this research to be carried out, aim and objectives to be achieved in the end of this research.

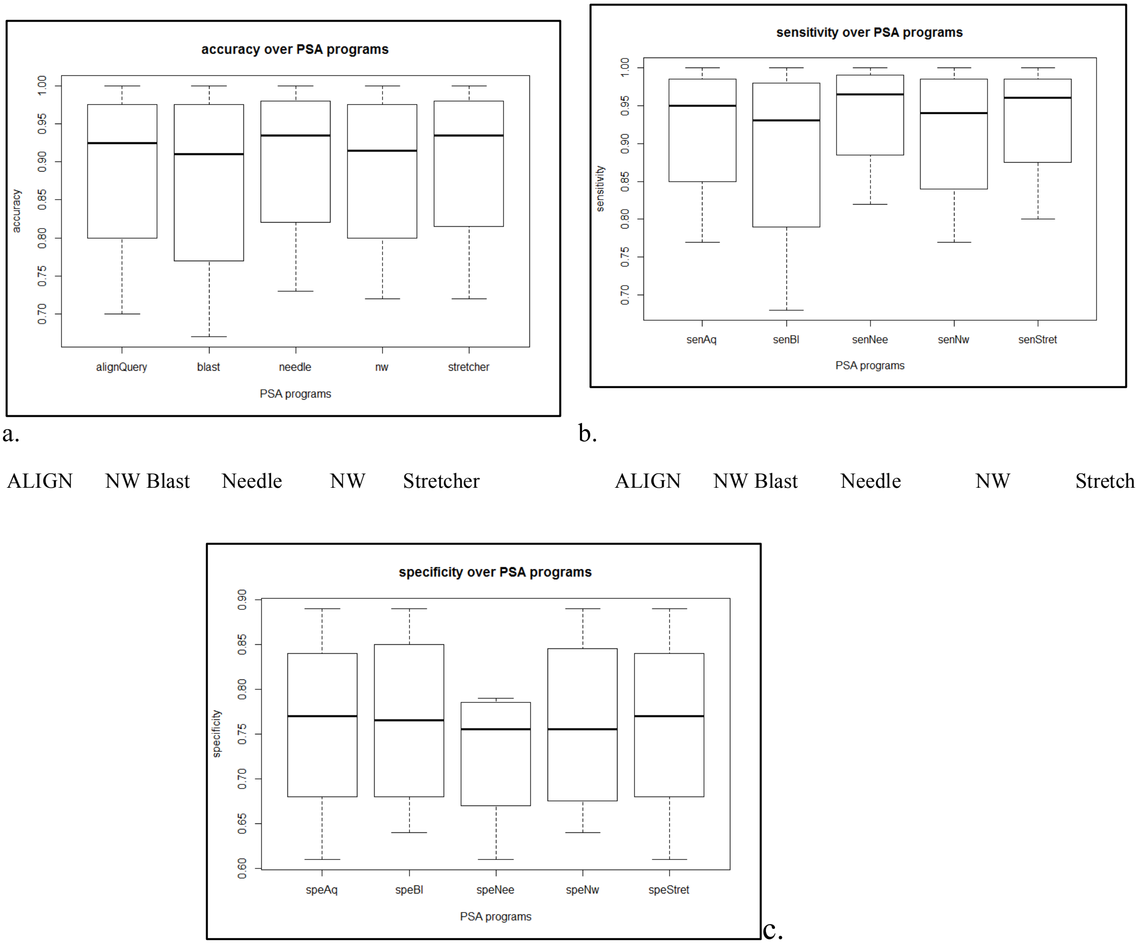
In data preparation and factor determination, since this research was carried out to comply with the method of identifying the presence of porcine in the sample ingredients, Duroc sp, cytochrome b was chosen as reference sequence in this research. Duroc sp, cytb can be get from NCBI website [9]. The five PSA programs was chosen based on their accuracy and how much they had been used to conduct a PSA method. The five PSA programs was ALIGN Query, NW Blast, EMBOSS Needle, NW and EMBOSS Stretcher. The factor which might affect the performance of PSA programs was determined by referring the literature review, by improve the limitation of previous research. The improvement that had been done in research was the performance of PSA programs was assessed by identifying their accuracy, sensitivity, specificity and consistency to avoid bias on the assessment result. Besides that, performance of PSA programs was assessed on different factor which are different mutation percentage occur on simulated sequence of Duroc sp, different parameters value and length of sequence.

In development of programs phase, two programs, named as AlignerReference and AlignerPerf program was developed in order to achieve the objective one and two. AlignerReference program was developed to create simulated sequence and reference alignment sequence of Duroc sp, cytb. The reference alignment sequence can be used to be compared by the alignment sequence from PSA programs. This can help AlignerPerf program to calculate the accuracy, sensitivity the specificity of the PSA program's result. Both of program was developed using C++ language which was the most suitable language to handle complex function, file and arrays.

In result analysis phase, accuracy, sensitivity and specificity of each PSA programs which can be get from the output of AlignerPerf program will be presented in the graphical forms in order to achieve the third objective. Type of graph used to analyze the performance was line and box-plot graph. The graph ease the analysis process. The performance analysis was made for each factor that had been stated before.

4. Result Analysis

As mentioned before, there was three factor that was determined in assessing the performance of PSA programs which are different percentage mutation occur in simulated sequence of Duroc sp, parameters value used and length of sequence. For the factor of different percentage mutation occur in simulated sequence of Duroc sp, refer Figure 1 to see the box-plot graph of accuracy, sensitivity and specificity of PSA programs at each different percentage mutation.



a. b.

ALIGN NW Blast Needle NW Stretcher ALIGN NW Blast Needle NW Stretcher

c.

ALIGN NW Blast Needle NW Stretcher

Figure 1: Accuracy, sensitivity and specificity of PSA programs at different mutation percentage

From the box-plot graph, we can see that EMBOSS Stretchers median line and starting value of box was the top compared to others PSA programs. It was followed by EMBOSS Needle in the second and followed by ALIGN Query in the third. NW Blast shows the least in its accuracy and sensitivity but improved in its specificity.

Next, the performance of PSA programs was assessed at factor of different parameters value used. Analysis from graph shows that only NW Blast accuracy and sensitivity shows an improvement when default value was used instead of used-defined paramaters value. Refer Figure 2 to see the user-define vs default parameters value, and Figure 3 shows the new graph of PSA programs performance used the best parameters value.

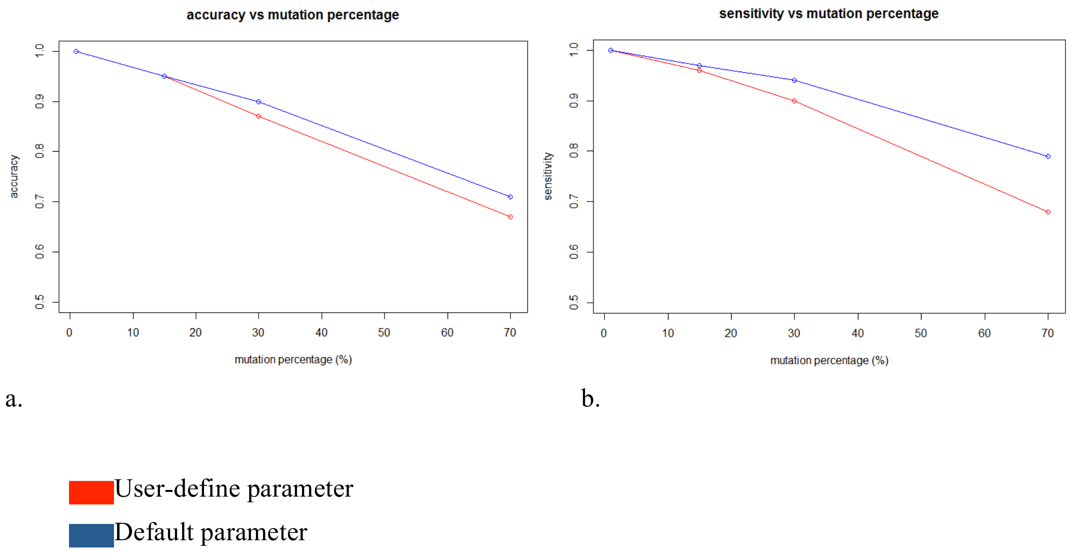


Figure 2: Accuracy and sensitivity graph of NW Blast when user-define and default parameters value was used.

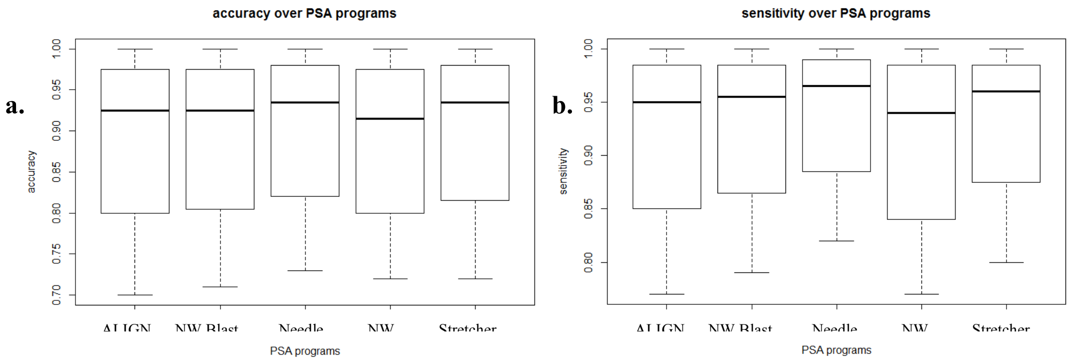


Figure 3: Accuracy and sensitivity graph of PSA programs using the best parameters value (either user-define or default)

From Figure 2(a) and Figure 2(b) shows that accuracy and sensitivity of NW Blast was improved when default value was used instead of used-define parameters value. In Figure 3(a) and 3(b) shows that the performance of NW Blast had been improved compared to its performance in Figure 1(a) and 1 (b). For factor at different length of sequence, refer figure 4 to see the box-plot graph of PSA programs consistency in terms of their accuracy, sensitivity and specificity.

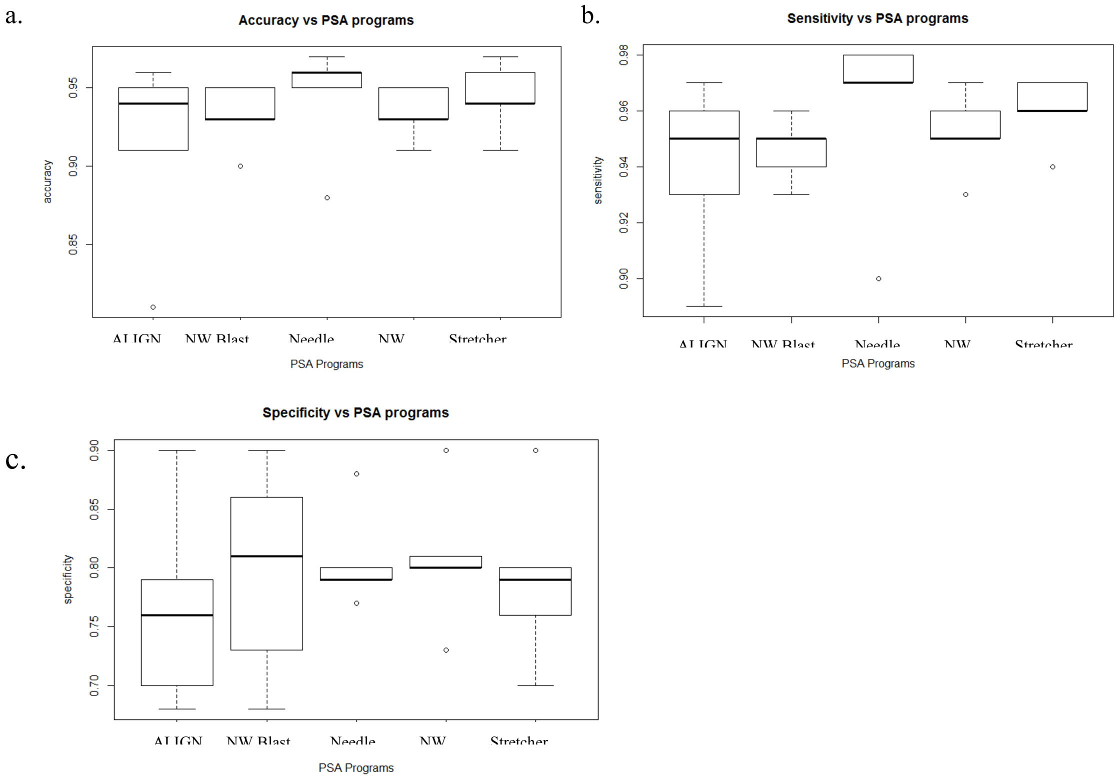


Figure 4: Accuracy, sensitivity and specificity of PSA programs on different length of sequence.

From the graph 4(a), we can see that the accuracy of EMBOSS Stretcher and NW was the most consistent because there was no outliers and the range size was not too big. From graph 4(b), we can see that the sensitivity of NW Blast was the most consistent, while, from Figure 4(c), we can see thatEMBOSS Needle was the most consistent

5.0. Conclusion

As a conclusion, in terms of the performance when mutation percentage was increase, EMBOSS Stretcher, EMBOSS Needle and NW Blast (when using default value as parameter) was the best programs to be used in determining the presence of Duroc sp gene sequence in the ingredients sample. While, in terms of result consistency, NW and EMBOSS Stretcher was the most consistent in terms of their accuracy, while, NW Blast was consistent in terms of its sensitivity and EMBOSS Needle was the most consistent in terms of its specificity.

Acknowledgements. I would like to thanks to my supervisor, Dr Iskandar bin Ilyas Tan and Akzam Saidin because guide and help me to complete this research. Besides that, a big thanks to Yayasan Peneraju Pendidikan Bumiputera for the moral and financial support for me.

References

32. http://www.bioinf.org.uk/software/nw/ NW programs download website
33. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=blastn&BLAST_SPEC=GlobalAln&LINK_LOC=BlastHomeLink BLAST NW online website
34. http://xylian.igh.cnrs.fr/bin/align-guess.cg ALIGN Query online website
35. http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html EMBOSS Needle online website
36. http://www.ebi.ac.uk/Tools/psa/emboss_stretcher/nucleotide.html EMBOSS Stretcher online website
37. Nuin, Paulo AS, Zhouzhi Wang, and Elisabeth RM Tillier. "The accuracy of several multiple sequence alignment programs for proteins." BMC bioinformatics 7.1 (2006): 471.
38. Penn, Osnat, et al. "GUIDANCE: a web server for assessing alignment confidence scores." Nucleic acids research 38.suppl 2 (2010): W23-W28.
39. Pang, Andy, et al. "SIMPROT: using an empirically determined indel distribution in simulations of protein evolution." BMC bioinformatics 6.1 (2005): 236.
40. https://www.ncbi.nlm.nih.gov/ NCBI Homepage