# FAQMAA: A Framework for Automatic Quality Assessment of Multiple Sequence Alignments

Muhammad Tariq Pervez[1] Naeem Aslam[2], Sarfraz Ahmad[1], Syed Shah Muhammad[1], Salman Qadri[3], Sajid Ali[4] and Usman Waheed[1].

[1.] Department of Computer Science, Virtual University of Pakistan
[2.] Institute of Biochemistry and Biotechnology, University of Veterinary and Animal Sciences, Lahore, Pakistan
[3.] Department of Computer Science, Islamia University of Bahawalpur
[4.] Institute of Virtual Reality and Visual Technology, Beijing Normal University, Beijing, China
tariq_cp@hotmail.com

**Abstract:** Multiple Sequence Alignments (MSAs) have significant role in the downstream analysis which includes identifying conserved patterns through evolution, functionally important residues, protein secondary and tertiary structure etc. MSAs, thus, have become an active area of research in the domain of bioinformatics. A large number of MSA methods are available but none of them is capable of producing a correct alignment for all situations. Therefore, knowledge of the most accurate MSA method in the initial stage of a biological research work may help in choosing the right MSA method for the right situation. Traditional technique to assess quality of MSA requires a lot of prior work to be completed such as calculation of guide tree, indel parameters, the best protein evolution model and reference alignment. Currently, no bioinformatic tool is available that performs all the prior work on its own. In this article, we present a framework titled 'FAQMAA' and its implementation in Java programming language that automatically assesses quality of a MSA method. FAQMAA has embedded interrelated open source software applications such as lamba.pl which is used to calculate the indel parameters, PortTest for extracting the best protein evolution model, amino acid frequencies and guide tree, INDELible for generating true alignment and finally SuiteMSA for calculating sum of pairs score and column score. FAQMAA does not require from user a guide tree file, indel parameters, the best protein evolution model, amino acid frequencies or even a reference alignment in order to perform its job. All the prior task is performed by the FAQMAA on its own and helps user save time, tiredness and cumbersome. FAQMAA has expedited the process of measuring quality of a MSA method and helped in selecting the right MSA method in the right situation.
[Muhammad Tariq Pervez Naeem Aslam, Sarfraz Ahmad, Syed Shah Muhammad, Salman Qadri, Sajid Ali and Usman Waheed. **FAQMAA: A Framework for Automatic Quality Assessment of Multiple Sequence Alignments.** *Life Sci J* 2013;10(9s):41-45] (ISSN:1097-8135). http://www.lifesciencesite.com. 5

**Keywords:** Multiple Sequence Alignment, MSA Quality, Comparison

## 1. Introduction

Multiple sequence alignment (MSA) is an approach which aligns two or more DNA, RNA or protein sequences (Kim and Ma, 2011) into a matrix form with the objective that characters in a specific column are homologous or have same function. MSAs have significant role in the downstream analysis which includes identifying conserved patterns through evolution, functionally important residues, protein secondary and tertiary structure and the nsSNPs (non synonymous Single Nucleotide Polymorphisms) that have a basic role for altering a protein function (Waterhouse et al. 2009; Thompson et al. 2011; Sullivan et al. 2003). Almost all areas of bioinformatics and evolutionary biology are based on correct MSA, which is, thus, one of the most active and highly scrutinized areas of research in bioinformatics (Catherine et al. 2011; Morgenstern et al. 2003). The more correct MSA the more correct are results of downstream analysis.

A large number of MSA methods are available (Notredame, 2011) such as MAFFT (Katoh et al. 2005), MUSCLE (Edgar, 2004), Kalign2 (Lassmann et al. 2009) and ClustalW (Larkin et al. 2007) etc., but none of them is capable of producing a correct alignment for all situations. All MSA methods have some deficiencies in one or the other way. Therefore, knowledge of the most accurate MSA method in the initial stage of a biological research work is very essential and important which may help in choosing the right MSA method for the right situation. Measuring quality of a MSA method involves calculating some type of score of a test alignment against a reference or true alignment. Sum of pairs score (SPS) and column score (CS) (Catherine et al. 2011) are two most popular scores used for measuring quality of MSAs.

One of the best ways of testing accuracy of multiple alignments is to construct reference alignment (standard of truth) and then comparing it with the test alignment generated by some MSA method whose accuracy is to tested (Morgenstern et al. 2003). Reference alignment may be constructed either based on actual data by a MSA method and

then enhance its accuracy by making some editing using some MSA editing tool such as Jalview or through a simulated tool such as indel-Seq-Gen (iSG) (Strope et al. 2009), Rose (Stoye et al. 1998), Simprot (Pang et al. 2005) etc. Alignment generated by a MSA method for the purpose of testing its accuracy is called test alignment and that is generated a simulated tool is called true or reference alignment.

Traditional technique to assess quality of MSA requires a lot of prior work to be completed such as calculation of guide tree, indel parameters, the best protein evolution model, amino acid frequencies and reference alignment. In addition to it, performing the prior task needs the usage of various software applications. Currently, no bioinformatic tool is available that performs all the prior work for assessing quality of a MSA method. A biological sequence simulation software tool such as indel-seq-gen is required for generating true or reference alignment. Getting indel parameters needs another software tool. Calculation of the best protein evolution model, amino acid frequencies and a guide tree requires another software application. Thus, the prior task makes the job of MSA quality assessment very laborious and time consuming which may be very difficult task especially for a novice user.

In this article, we present a framework titled as 'A Framework for Automatic Quality Assessment of Multiple Sequence Alignments' (FAQMAA) that automatically assesses quality of a MSA method. Implementation of FAQMAA is provided in Java programming language. FAQMAA has brought all bioinformatic tools at one place that are needed to perform the job of measuring quality of a MSA method. FAQMAA is a suite of interrelated open source software applications such as lamba.pl which is a Perl script that comes with DAWG (DNA Assembly with Gaps: a software tool to simulate phylogenetic evolution of recombinant DNA sequences) to calculate the indel parameters, PortTest for extracting the best protein evolution model, amino acid frequencies and guide tree, indel-seq-gen for generating true alignment and finally SuiteMSA for calculating SPS and CS. FAQMAA has integrated all the required software tools by modifying them as per its need. FAQMAA takes an alignment file as an input and displays its quality in the form of statistics of SPS and CS. FAQMAA does not require from user a guide tree file, indel parameters, the best protein evolution model, amino acid frequencies or even a reference alignment in order to perform its job. All the prior task is performed by the FAQMAA on its own and helps user save time, tiredness and cumbersome. FAQMAA has expedited the process of measuring quality of a MSA method and helped in selecting the right MSA method in the right situation.

## 2. Material and Methods

### Framework

Figure 2 represents the proposed framework. FAQMAA starts its working by building a command for ProtTest3.2 and running it without any intervention of the user (Figure 2). Output of ProtTest3.2 is parsed and FAQMAA extracts the best protein model, amino acid frequencies, sequence length and tree according to the best model for subsequent software like lambda.pl and INDELible. The best protein model, amino acid frequencies, sequence length are used in building control file for INDELible and tree is used as input for INDELible as well as for lambda program. FAQMA then builds a command for lambda program, runs it, parses its output and filters the indel parameters. Indel parameters are used in building control file for INDELible. At third stage, FAQMAA runs INDELible and gets the true alignment file. Finally, FAQMAA runs SuiteMSA by providing test and true alignments and displays quality of the given MSA.

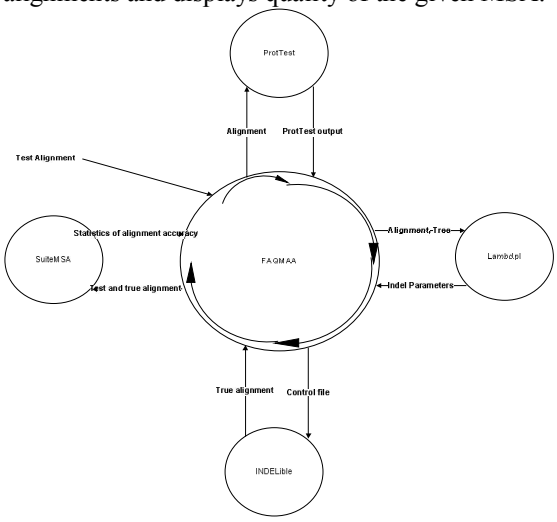


Figure 1. FAQMAA starts its working with the provided test alignment. At first, FAQMAA builds a command comprising of the provided alignment and all required parameters to run prottest. FAQMAA then parses output of the protest and extracts the best protein model, amino acid frequencies, sequence length and tree according to the best model. As a second step, FAQMAA builds a command for lambda.pl program, runs it, parses output of the lambda program and extracts indel parameters. At third stage, FAQMAA builds a command to run INDELible and gets the true alignment. At the end FAQMAA runs SuiteMSA to display statistics of alignment accuracy.

## 3. Implementation

FAQMAA (Figure 1) has been developed in Java programming language using Netbeans7.2 as an

integrated development environment. Main window of FAQMAA is divided into two parts which can be interactively resized by the user. Upper part provides options to be selected by the user and below part shows output and progress of each software. As figure 1 shows, FAQMAA now has made measuring quality of an MSA a few clicks task. A user selects the alignment to be tested by clicking the 'Select Test MSA' button, information criterion, indel size and then clicks the 'start' button to let the FAQMAA show quality of the given MSA.

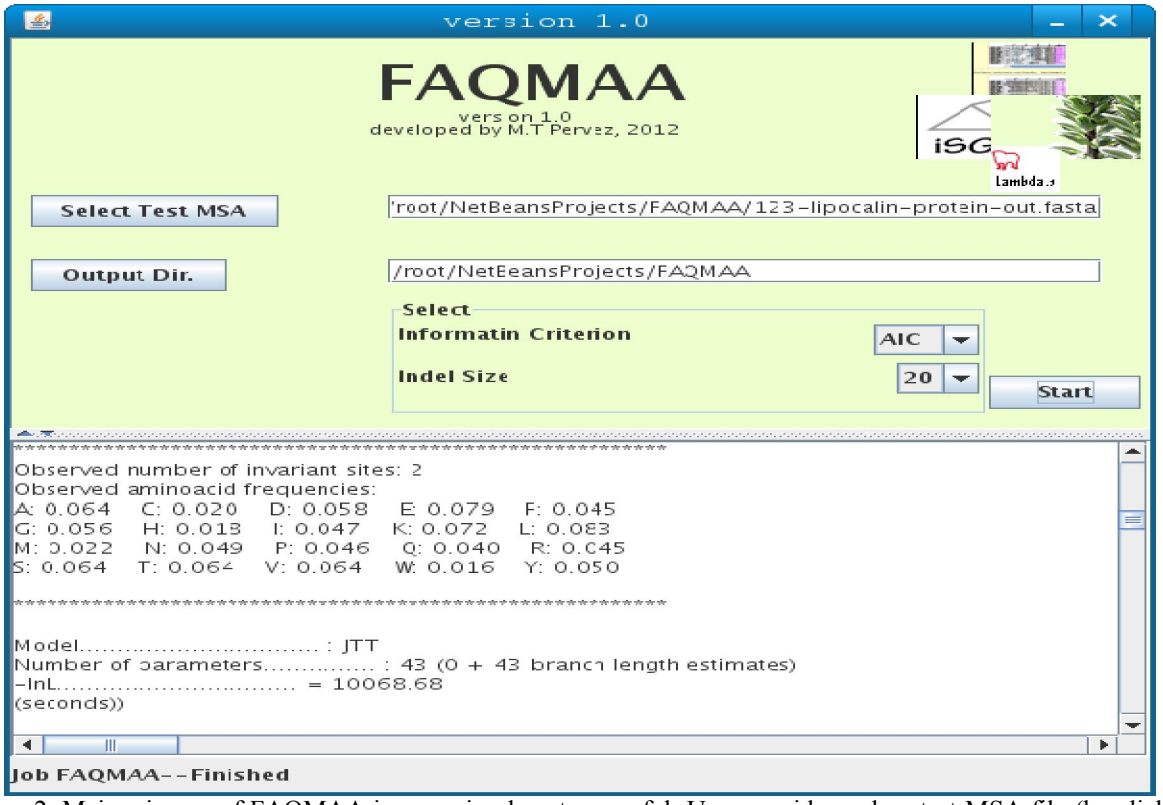


Figure 2. Main winnow of FAQMAA is very simple yet powerful. User provides only a test MSA file (by clicking 'Select Test MSA' link) generated by any MSA method in FASTA format and an output directory. FAQMAA then runs in behind all required software applications to complete the prior task of estimating indel parameters, obtaining guide tree, amino acid frequencies, generating known alignment and finally displays the quality of the test MSA both in the statistical and graphical form. FAQMAA saves/writes output of all software tools in the output directory.

**Multiple alignment tools evaluated**

We generated alignments of the lipocalin protein family from the web interface of the latest versions of 8 multiple alignment methods (Table 1) All the MSA tools were run with the default options. FAQMAA was run on and tested on All programs were run on a Sun Enterprise V40z server (4 Opteron processors with 4616 Gb memory) under RedHat Enterprise Linux.

Table 1. Multiple sequence alignment programs evaluated in this study

| MSA Tool | Version | Availability |
|---|---|---|
| T-Coffee | 9.02.r1228 | http://www.tcoffee.org/Projects/tcoffee/ |
| ProbCons | 1.12 | http://probcons.stanford.edu |
| Dialign-TX | 1.0.2 | http://dialign-tx.gobics.de/download |
| Kalign | 1.04 | http://msa.sbc.su.se/cgi-bin/msa.cgi |
| MultAlign | 5.4.1 | http://multalin.toulouse.inra.fr/multalin/ |
| Mafft | 6.903b-LINS-I | http://mafft.cbrc.jp/alignment/software/ |
| Clustal omega | 1.1.0 | http://www.clustal.org/omega/ |
| Muscle | 3.8.31 | http://www.drive5.com/muscle/ |

## 4. Results and Discussion

One of the famous techniques to measure quality of MSA is to compare test alignment against a reference alignment. There are a number of simulated bioinformatic tools available to generate reference alignments. Almost all famous MSA quality measuring tools such as SuiteMSA, AltaVist (Morgenstern et al. 2003), SinicView (Shih et al. 2006) etc requires reference alignment to measure accuracy of the test alignment. Calculating reference alignment demands a lot of work to be performed which is very laborious and time consuming task. FAQMAA, however, does not require a reference alignment. It calculates reference alignment on its own through INDELible based on the provided test alignment.

To test performance and accuracy of FAQMAA, we selected lipocalin proteins as test alignment (as Catherine et al. did) to be provided to FAQMAA. Lipocalin is a family of small globular proteins involved in allergic reactions. Lipocalin super family has low sequence identity, but share a common anti parallel beta-barrel conformation comprising of eight beta-strands, and a small highly-conserved motif near the first beta-strand (Flower et al. 1993).

### MSA method quality using column score

FAQMAA uses SuiteMSA to calculate statistics for measuring quality of the MSA methods. Figure 2 shows alignment quality of each MSA method using CS, SPS and average of CS and SPS.

FAQMAA runs ProtTest for finding the best protein model, amino acid frequencies and calculating tree. Lambda.pl then is run by FAQMAA by providing it guide trees (generated by ProtTest) and the alignment file. As a third step FAQMAA executes iSG by providing it the indel parameters (generated by lambada.pl), guide tree (we used tree generated by ProtTest as a guide tree), the best protein evolution model etc. to generate the true alignment. Finally, FAQMAA runs SuiteMSA to show the comparison result. Results show that FAQMAA is very efficient and accurate for small and medium data sets (sequences less than 500 taxa) and low insertion or deletion rates. For large data sets (sequences greater than 500 taxa) and high insertion and deletion rates, performance of FAQMAA is significantly slow. Especially high insertion rates affect the performance of iSG significantly.

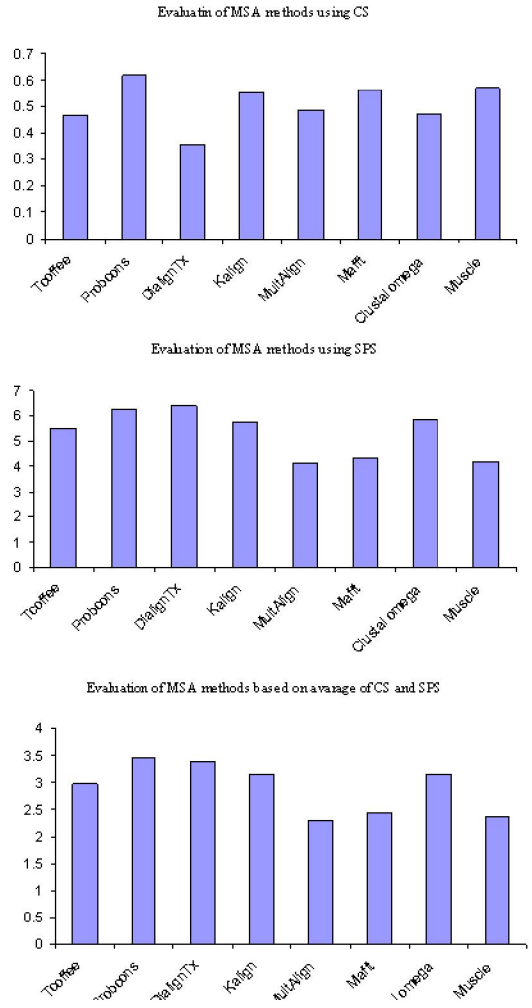


Figure 3. Three graphs showing comparison of selected MSA methods.

## 5. Conclusion

Performance of all tools involved in the downstream analysis in the domain of biology and bioinformatic depends upon the correctness of multiple sequence alignments. Measuring accuracy of MSAs based on a reference alignment is not an easy task. FAQMAA has made this task very easy by integrating all the required software applications for generating reference alignments and calculating accuracy of the test alignments. FAQMAA uses ProtTest2.4 for calculating the best protein evolution model and amino acid frequencies, lambda.pl program for estimating indel parameters, iSG for obtaining the corresponding reference alignment and finally SuiteMSA for displaying the comparison results both in graphical form and various statistics. FAQMAA has helped bioinformatic researchers a lot by saving their time and avoiding them from being

bore while building the reference alignment and calculating accuracy of the test alignment. A user does not need to generate reference alignment because now it is the job of FAQMAA. In future, we have a plan to enhance the ability of FAQMAA so that it may also get generate reference alignments based on an alignment that has highly conserved motifs.

**Acknowledgements**
We thank Higher Education Commission of Pakistan for her generous support for completing this project successfully. We also thank Dr. Shahzad Ahmad Faizi, lecturer, Department of Mathematics and Mr. Imran Ali, lecturer, Department of English, Virtual University of Pakistan for valuable discussions and help in drafting the paper.

**Corresponding Authro**
Muhammad Tariq Pervez
Department of Computer Science
Virtual University of Pakistan
E-mail: tariq-cp@hotmail.com

**References**

1. Kim J, Ma J. 2011. PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucl. Acids Res. 39 (15):6359-6368.doi: 10.1093/nar/gkr334.
2. Catherine LA, Cory LS, Etsuko NM. 2011. SuiteMSA: Visual Tools for Multiple Sequence Alignment Comparison and Molecular Sequence Simulation. BMC Bioinformatics 12:184.
3. Waterhouse AM., Procter,JB., Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench, Bioinformatics doi: 10.1093/bioinformatics/btp033.
4. Thompson JD, Linard B, Lecompte O, Poch O. 2011. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS ONE 6(3): e18093. doi:10.1371/journal.pone.0018093
5. Morgenstern B, Goel S, Sczyrba A, Dress, A. 2003. AltAVisT: comparing alternative multiple sequence alignments. Bioinformatics, 19, 425–426.
6. Sullivan OO, Zehnder M, Higgins D, Bucher P, Grosdidier A, Notredame C. 2003. APDB: a novel measure for benchmarking sequence alignment methods without reference. Bioinformatics. 19:1. i1215-i1221.
7. Notredame C. 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput. Biol., 3, e123.
8. Katoh K, Kuma K, Toh H. 2005. MAFFT v. 5: improvement in accuracy of multiple sequence alignment. Nucleic. Acids Res. 33, 511–518.
9. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-1797.
10. Lassmann T, Frings OS, Sonnhammer, EL (2009) Kalign 2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 37: 858–865.
11. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948.
12. Strope CL, Abel K, Scott SD, Moriyama EN. 2009. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Mol Biol Evol 2009, 26:2581-2593.
13. Stoye J, Evers D, Meyer F. 1998. Rose: generating sequence families. Bioinformatics. 14:157–163.
14. Pang A, Smith AD, Nuin PAS, Tillier ERM. 2005. SIMPROT: using an empirically determined indel distribution in simulations of protein evolution. BMC Bioinformatics. 6:236.
15. Flower DR, North ACT, Attwood TK: Structure and sequence relationships in the lipocalins and related proteins. Protein Sci 1993, 2:753-761.
16. Shih AC, Lee DT, Lin L, Peng CL, Chen SH, Wu YW, Wong CY, Chou MY, Shiao TC, Hsieh MF. 2006. SinicView: a visualization environment for comparisons of multiple nucleotide sequence alignment tools. BMC Bioinformatics, 7:103.

7/20/2013