IVisTMSA: Interactive Visual Tools for Multiple Sequence Alignments

Muhammad Tariq Pervez1,2, Masroor Ellahi Babar3, Asif Nadeem2, Naeem Aslam2,4, Nasir Naveed5, Sarfraz Ahmad1, Shah Muhammad1, Salman Qadri6, Muhammad Shahid7, Tanveer Hussain3 and Maryam Javed2

1Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan. 2Institute of Biochemistry and Biotechnology, University of Veterinary and Animal Sciences, Lahore, Pakistan. 3Department of Bioinformatics, Virtual University of Pakistan, Lahore, Pakistan. 4Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan. 5Institute for Web Science and Technologies, University of Koblenz-Landau, Germany. 6Department of Computer Science, The Islamia University of Bahawalpur, Pakistan. 7Department of Physics, Government Sadiq Egerton College Bahawalpur (S.E. College Bahawalpur), Pakistan.

ABSTRACT: IVisTMSA is a software package of seven graphical tools for multiple sequence alignments. MSApad is an editing and analysis tool. It can load 40% more data than Jalview, STRAP, CINEMA, and Base-by-Base. MSA comparator allows the user to visualize consistent and inconsistent regions of reference and test alignments of more than 21-MB size in less than 12 seconds. MSA comparator is 5,200% efficient and more than 40% efficient as compared to BALibase c program and FastSP, respectively. MSA reconstruction tool provides graphical user interfaces for four popular aligners and allows the user to load several sequence files at a time. FASTA generator converts seven formats of alignments of unlimited size into FASTA format in a few seconds. MSA ID calculator calculates identity matrix of more than 11,000 sequences with a sequence length of 2,696 base pairs in less than 100 seconds. Tree and Distance Matrix calculation tools generate phylogenetic tree and distance matrix, respectively, using neighbor joining identity and BLOSUM 62 matrix.

KEYWORDS: MSA editing tool, MSA comparator, FASTA, MSA Identity Matrix

INTRODUCTION

Multiple sequence alignments (MSAs) are an essential first step for a number of computational approaches such as protein secondary structure/function prediction, phylogeny inference, and many other common tasks in sequence analysis.1–3 Several software tools for generating multiple sequence alignments are available,4 but none of them is suitable for all types of data sets.5 Consequently, in order to generate a true alignment, a need arises for inspection and adjusting alignments by hand, which is a very laborious job.5 Furthermore, handling hundreds of thousands of alignments is another problem in the domain of bioinformatics.

Many popular alignment editors such as Jalview,5 STRAP,6 SeaView,7 MEGA,8 CINEMA,9 and Base-By-Base10 are available. All these tools either do not support tens of thousands of sequences (where length of each sequence is more than 2,500 base pairs) or they do not have user-friendly editing features. Jalview and STRAP do not work when the size of an alignment exceeds 30.01 MB. CINEMA and Base-By-Base are also good for small alignments. They hang up if the size of an alignment increases 3 MB. SeaView and MEGA can load big alignments, but the editing features provided by them involve multiple steps. Many software tools provide graphical user interfaces to reconstruct MSAs, but they do not allow loading multiple sequence files at the same time. MSA comparison tools such as SuiteMSA11 and SinicView12 permit to compare MSAs, but they support alignments comprising less than 1,000 sequences. Multiple formats of MSAs exist, but FASTA is the most popular format. Presently, there is no tool that can convert the format of an alignment of unlimited size into FASTA format. Many tools, such as MatGAT13 and Sequence Identity and Similarity (SIAS),14 are available, to calculate the identity matrix but they do not support more than 1,000 sequences.

This article describes IVisTMSA, which is a software package of seven graphical tools for multiple sequence alignments. All tools allow the user to load tens of thousands of sequences to edit/analyze and compare them. IVisTMSA implemented a divide-and-conquer (DnC) approach to calculate efficiently consensus/conserved sequences, identity matrix, Sum-of-Pairs Score (SPS), Colum Score (CS),
and conserved regions of two alignments. SPS assesses the capability of MSA methods for aligning the input sequences accurately. SPS of a column is calculated as the ratio of the sum of scores of all pairs of residues in every column of test alignment by sum of scores of all pairs of residues in the similar column of reference alignment. To calculate SPS of the entire alignment, sum of sum of scores of each column of the test alignment is divided by sum of sum of scores of each column of reference alignment. Consequently, SPS raises with the number of sequences accurately aligned. CS investigates the ability of MSA programs to correctly align all sequences. CS is computed as the ratio of matched columns (between test and reference alignments) by the number of considered columns in the test alignment. The sequence file with 807 sequences named as BBA0039.tfa available in “RV100” folder enclosed in a zipped file named as ‘msa_reference.tar.gz’ in BALiBASE was used for testing all tools IVisTMSA. This data set was replicated to generate an alignment with 1,614 sequences, which was replicated to generate 3,228 sequences, and so on. Replication was done in order to save time and avoid from constructing big alignments from an MSA method.

Design and Implementation
All tools of IVisTMSA (Fig. 1) are written in Java programming language. NetBeans IDE (Integrative Development Environment) 7.4 was used to write IVisTMSA. XML was used for saving the state of the work performed in MSApad and MSA comparator. BioJava was used to embed Jmol in IVisTMSA. A computing machine having Core i7 3.34-GHz processor and 8-GB RAM was used to write and analyze IVisTMSA.

Figure 1. Main interface of IVisTMSA. This provides links to start the seven tools of IVisTMSA. The specific tool can be run by clicking the respective link.
DnC Approach
Most of the tools of IVisTMSA use DnC approach for performing efficient computations on MSAs. The DnC approach has been implemented using a power feature of multithreading provided by Java programming language. DnC approach divides an alignment horizontally into sub-alignments, and Java threads are generated for each fragment. All Java threads return results to the main thread, which computes the final value. MSAPad uses DnC approach to compute consensus, conserved sequence(s), and distance matrix to construct a phylogenetic tree. Figure 2A shows the algorithm that divides an alignment into subalignments. Step 1 calculates number of subalignments for a given MSA. By default, each alignment is divided into two subalignments. Step 2 creates a list/array for holding size of each sub-MSA. Step 3 calculates the size of the first subalignment and stores it into the first position of the list. Step 4 stores the size of the first sub-MSA into a variable 'totalSeqsCounted' to track the size of each sub-MSA. Step 5 subtracts one from the “subMSAs” variable because up to now, the size of the first sub-MSA has been calculated. Step 6 creates a new variable “counter” and initializes with 1. The “counter” variable is used to store the size of subalignments at correct positions in the list. Steps 8–11 are repeated until the size of each sub-MSA is calculated. Figure 2B shows the overall process of dividing MSA into sub-MSAs, computations performed by each thread, and the final computation by the main thread.

Alignment-Rendering Model
The alignment-rendering model has played a very important role to display, edit, and analyze very big alignments efficiently. The model sits between the alignment and its viewer and manages efficiently various types of manipulations on

![Diagram A](image1)

1. Set subMSAs to (number-of-sequences/100 +2)
2. Set subMSAsSizeList[] = new int [subMSAs]
3. Set subMSAsSizeList[0] to (number-of-sequences/subMSAs)
4. Set totalSeqsCounted to subMSAsSizeList[0]
5. Set subMSAs to subMSAs - 1
6. Set counter to 1
7. While subMSAs > 0
   Start
   8. Set subMSAsSizeList[counter] to (number-of-sequences-totalSeqsCounted)/subMSAs
   9. Set totalSeqsCounted to totalSeqsCounted+subMSAsSizeList[counter]
10. Set counter to counter+1
11. Set subMSAs to subMSAs-1
   End

![Diagram B](image2)

Figure 2. DnC approach used in various tools of IVisTMSA. In the first step an alignment is divided into subalignments (A). In the second step, Java threads are generated to perform various types of computations on the sub-MSAs and at the end the main thread performs the final computations (B).
the alignments. These manipulations include finding a residue, all types of editing features provided by IVisTMSA, changing color schemes, and loading the alignment itself. We developed an alignment-rendering model by extending “AbstractTableModel” class provided as part of the swing library of the Java programming language. Figure 3 shows the alignment-rendering model that is implemented by IVisTMSA.

**Results**

MSApad is an editing and analysis environment for multiple sequence alignments. Results showed that, in contrast to all other alignment editors written in java programming language such as STRAP, Jalview, and Base-by-Base, it can load more than 400% big alignments. It also provides several unique editing features. MSApad allows the user to save the alignment, find a single residue or sequence name, add the sequence at the start or end of an alignment, add a sequence before or after a selected sequence, and move a sequence up or down (Fig. 4A). MSApad provides several analysis features as well (Fig. 4B). It can find consensus and conserved sequence/regions of an alignment of 120-MB size in just 27 seconds. It allows the user to calculate a phylogenetic tree using neighbor joining% identity and BLOSUM 62 matrix. The tree is drawn using Archaeopteryx, which is an open-source software tool for displaying and analyzing a phylogenetic tree. The user can also calculate the identity of a sequence with other selected sequences. Archaeopteryx and Jmol are embedded to view and analyze phylogenetic trees and protein 3D chemical structures, respectively (Fig. 4C).

Table 1 shows a comparison summary of MSApad with other popular alignment editors. The highlighted region shows the unique features provided by MSApad. MSA comparator is actually MQAT version 2.0.1. Its version 1.0.2 is already published in *Life Science Journal*. Data structures used in MQATv1.0.2 were improved, and it is now more efficient than qscore program (http://www.drive5.com/qscore/) and FastSP as well. The new version of MQAT allows the user to visualize the conserved and nonconserved regions of the selected alignments from the main window of MQAT. Figure 5A displays the main window of MQAT, which allows the user to select an alignment. Figure 5B shows the conserved regions of reference and test alignments. Conserved regions of two alignments can be calculated by selecting “Alignment Consistency” option from the “View” menu of the main window. Results showed that the MSA comparator can display conserved and nonconserved regions of two alignments comprising more than 8,000 sequences with a sequence length of 2,696 base pairs in less than 12 seconds.

MSA reconstruction tool provides graphical interfaces for Clustal Omega, ClustaW2, MAFFT, MUSCLE, and BioJava implementation for MSAs. Its unique feature is that the user can load multiple sequence files at the same time. Now the user does not need to sit before the system and wait for the completion of the process so that the next sequence file may be uploaded. Using MSA reconstruction tool of IVisTMSA, the user can provide multiple sequence files simultaneously and may perform other tasks. IVisTMSA will generate MSAs of the provided sequence files one by one. Interfaces for Clustal Omega, ClustaW2, MAFFT, MUSCLE, and BioJava can be loaded by clicking the MSA Reconstruction drop-down list (Fig. 1) provided on the main interface of IVisTMSA.

FATSA generator can convert ClustalW, MSF, Phylip, PIR, GDE, MEGA, and Nexus formats of alignments of unlimited size into FASTA format. We converted successfully
an alignment of MSF format comprising 102,102 sequences into FASTA format in less than 1 second. The other important feature of this tool is that the user can load alignments of different formats. FASTA generator recognizes the format of an alignment automatically and converts it into FASTA format.

MSA ID calculator allows the user to calculate the identity matrix of an alignment. Results showed that it can calculate the identity matrix of more than 11,000 sequences with a sequence length of 2,696 base pairs in less than 100 seconds. It also allows the user to save and print the identity matrix.

Tree calculation tool calculates a phylogenetic tree using neighbor joining% identity and BLOSUM 62. The generated distance matrix can be used for further statistical and phylogenetic analysis.

Discussion
IVIS-TMSA is a suite of seven interactive visual tools to generate, view, edit, and analyze MSAs. Presently, a lot of MSA editing and analyzing tools are available. The popular and widely used tools include Jalview, SeaView, MEGA, STRAP, PFAAT, Base-by-Base, and CINEMA. Jalview, STRAP, PFAAT, and Base-by-Base are Java programs, whereas SeaView, MEGA, and CINEMA are written in C++ language. Results showed that MSApad loaded more than 50,000 sequences with a sequence length of 2,696 base pairs, whereas other alignment editors written in Java programming language could load less than 12,000 sequences only. Alignment editors written in C++ language loaded more sequences than MSApad, but their editing features are not user friendly. MSApad allows the user to save an alignment, find a residue/sequence name, add a sequence at any position in the alignment, move sequence up or down, calculate phylogenetic tree, conserved or consensus sequence, identity (with gaps or without gaps), set color of an alignment using six color schemes, and view and analyze protein 3D chemical structure and phylogenetic tree.
Table 1. Comparison of MSApad with other MSA editing tools.

| PARAMETERS/SOFTWARE | JALVIEW | STRAP | MEGA | SeaView | CINEMA | BASE BY BASE | IVisTMSA |
|---------------------|---------|-------|------|---------|--------|-------------|---------|
| Language            | Java    | Java  | C++  | C++     | C++    | Java        | Java    |
| Number of sequences | < = 13000 | < = 13000 | > = 50000 | >50000 | < = 5000 | < = 5000 | >50000 |
| Editing a single cell | x       | x     | x    | x       | x      | x           | √       |
| Inserting sequence at the desired place | x       | x     | x    | x       | x      | x           | √       |
| Moving column       | x       | x     | x    | x       | x      | x           | √       |
| Sorting by column   | x       | x     | x    | x       | x      | x           | √       |
| Identity (gaps/no gaps) | x       | x     | x    | x       | x      | x           | √       |
| Sorting by taxa name| √       | x     | √    | x       | x      | √           | √       |
| View of 3D chemical structure | √       | √     | √    | √       | x      | x           | √       |
| View/Analysis of phylogenetic tree | √       | √     | √    | √       | x      | x           | √       |
| Conserved sequences (selected/all) | √       | x     | √    | x       | x      | x           | √       |
| Consensus sequences (selected/all) | √       | x     | √    | √       | √      | √           | √       |
| Sequence formats    | 8       | 8     | 8    | 6       | 6      | 4           | 8       |
| Find sequence name  | √       | √     | √    | √       | √      | √           | √       |
| Moving Sequence     | √       | √     | √    | x       | x      | √           | √       |

Figure 5. Consistent and inconsistent regions of the reference and test alignment. When the user selects a test file from the window displaying SPS and CS (A), a new interface showing consistency of test and reference alignment (B) is loaded. The consistent residues are shown in blue color, whereas inconsistent elements are displayed in grey color.
Interactive tools for MSAs

to edit a residue at its own position without opening a new interface, whereas all other tools allow editing of an alignment in a new interface. MSApad provides the feature to insert a sequence at any location of an alignment, whereas most of the available editing tools allow the user to insert a sequence at the end of an alignment. Similarly moving a column and sorting an alignment by column are also the unique features of MSApad. Some of the alignment editors such as HOMED58 and MALIGNED39 are no longer maintained by their authors. Several tools such as SuiteMSA,4 SinicView,13 AItAvist,7 and BALiBASE c program20 are available to compute SPS and CS but they cannot process more than 1,000 sequences. Results showed that the MSA comparator was 5,200% efficient as compared to BALiBASE c program. It calculated conserved regions of two alignments comprising more than 8,000 sequences with a sequence length of 2,696 base pairs in less than 12 seconds, whereas SuiteMSA allowed the user to compute conserved regions of two alignments consisting of less than 1,000 sequences. It also did not allow loading multiple test alignments simultaneously. Many software tools such as STRAP, SuiteMSA, PFAAT, and SeaView provide graphical user interfaces for generating MSAs, but they do not allow the user to load several sequence files simultaneously. MSA reconstruction tool allows the user to load several sequence files at a time to align them one by one using Clustal Omega, ClustalW2, MAFFT, MUSCLE, or BioJava. Lot of tools and web servers such as ALTER23 and Readseq22 are available for converting formats of MSAs, but they process alignments comprising a few thousand sequences. FASTA generator allows the user to convert seven MSA formats (ClustalW, MSF, Phylip, PIR, GDE, and Nexus) into FASTA format. Its unique feature is that it can process alignments of unlimited size. Several tools such as SIAS15 and MatGAT23 are available to calculate the identity matrix. MatGAT first reconstructs alignment and then computes the identity matrix. It can generate an identity matrix of more than 500 sequences with sequence length of 2,696 base pairs, which was very time consuming. Since, SIAS is a web application, it was also not a good tool for calculating identity matrix of more than 1,000 sequences with a sequence length of 2,696 base pairs. MSA ID calculator is an efficient tool, which calculated the identity matrix of more than 11,000 sequences with a sequence length of 2,696 base pairs in less than 100 seconds. Comparison indicates that MSApad, MSA comparator, FASTA generator, and MSA ID calculator are more efficient than other similar tools available in the market for multiple sequence alignments.

Conclusion

IVisTMSA is a suite of seven interactive tools for multiple sequence alignments. MSApad allows the user to edit and analyze 40% more data than Jalview, STRAP, CINEMA, and Base–by–Base. MSA comparator (MQAT version 2.0.1) allows the user to visualize consistent and inconsistent regions of reference and test alignments of more than 21-MB size in less than 12 seconds. The MSA comparator is 5,200% efficient as compared to BALiBASE c program. MSA reconstruction tool allows a user to upload several sequence files through the graphical user interfaces of Clustal Omega, ClustalW2, MAFFT, and MUSCLE, and then align them one by one. FASTA generator converts the other seven formats of alignments of unlimited size into FASTA format in a few seconds. MSA ID calculator is a tool that allows a user to calculate the identity matrix of more than 11,000 sequences with a sequence length of 2,696 base pairs in less than 100 seconds. Tree and Distance Matrix calculation tools generate phylogenetic tree and distance matrix, respectively, using the neighbor joining% identity and BLOSUM 62 matrix. IVisTMSA allows scientists to view, edit, interpret, and analyze very big alignments.

Availability and Requirements

• Project name: IVisTMSA
• Project home page: http://ivistmsa.com/
• Operating system(s): Tested on windows but it should run on other platforms as well
• Programming language: JAVA 1.7
• Execution requirements: JDK1.7 or higher
• License: none
• Any restrictions to use by nonacademics: none

Acknowledgments

A synopsis of this work was published in the Daily Messenger, a newspaper in Karachi, Pakistan, in August 2014, and in The Nation, a newspaper in Lahore, Pakistan, in December 2014. Publication of the synopsis in the newspapers was in accordance with PhD requirements for Dr MT Pervez at the University of Veterinary and Animal Sciences, Lahore.

Author Contributions

Conceived: MTP, MEB, AN. Designed and performed the experiments and tested and analyzed IVisTMSA: MJ, AN, MS, SQ, TH. Developed the software: MTP, SA, SM, NA, NN. Drafted the article: MTP, MEB, SQ.

REFERENCES

1. Kim J, Ma J. PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucleic Acids Res. 2011;39(15):6359–68.
2. Sullivan OO, Zehnder M, Higgins D, Bucher P, Grosdidier A, Notredame C. APDB: a novel measure for benchmarking sequence alignment methods without reference. Bioinformatics. 2003;19:1.
3. 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.
4. Thompson JD, Linard B, Lecompte O, Poch O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One. 2011;6(3):e18093.
5. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview version 2: a multiple sequence alignment and analysis workbench. Bioinformatics. 2009;25(9):1189–91.
6. Catherine LA, Cory LS, Etsuko NM. SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation. BMC Bioinformatics. 2011;12:184.
7. Morogenreti R, Goel S, Scarya A, Dress A, AltAviT: comparing alternative multiple sequence alignments. Bioinformatics. 2003;19:425–6.
8. Gille C, Fröhmel C. STRAP: editor for structural alignments of proteins. Bioinformatics. 2001;17(4):377–8.
9. Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci. 1996;12:543–8.
10. Kumar S, Tamura K, Jakobsen IB, Nei M. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics. 2001;17:1244–5.
11. Lord PW, Selley JN, Attwood TK. CINEMA-MX: a modular multiple alignment editor. Bioinformatics. 2002;18:1402–3.
12. Brodie R, Smith AJ, Roper RL, Tcherepanov V, Upton C. Base-by-base: single nucleotide-level analysis of whole viral genome alignments. BMC Bioinformatics. 2004;5:96.
13. Shih AC, Lee DT, Lin L, et al. SinicView: a visualization environment for comparisons of multiple nucleotide sequence alignment tools. BMC Bioinformatics. 2006;7:103.
14. Campanella JJ, Bitincka L, Smalley J. MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. BMC Bioinformatics. 2003;4:29.
15. Rech P. SLAS: Sequence identities and similarities. 2008. Available at: http://imed.ucm.es/Tools/sias.html
16. Pervez MT, Babar ME, Nadeem A, et al. MQAT: an efficient quality assessment tool for large multiple sequence alignments. Life Sci J. 2013;10(9s):9–16.
17. Caffrey DR, Dana PH, Marbur V, et al. PFAAT version 2.0: a tool for editing, annotating, and analyzing multiple sequence alignments. BMC Bioinformatics. 2007;8:381.
18. Stockwell PA, Petersen GB. HOMED: a homologous sequence editor. Comput Appl Biosci. 1987;3:37–43.
19. Clark SP. MALIGNED: a multiple sequence alignment editor. Comput Appl Biosci. 1992;8:535–8.
20. Thompson J, Plewniak F, Poch O. BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics. 1999;15:87–8.
21. Glez-Peña D, Gómez-Blanco D, Reboiro-Jato M, Fdez-Riverola F, Posada D. ALTER: program-oriented conversion of DNA and protein alignments. Nucleic Acids Res. 2010;38:W14–8.
22. Gilbert D. Sequence file format conversion with command-line readseq. Curr Protoc Bioinformatics. 2003;Appendix 1:Appendix1E.
23. Campanella JJ, Bitincka L, Smalley J. MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. BMC Bioinformatics. 2003;4(1):29.