SAliBASE: A Database of Simulated Protein Alignments

Muhammad Tariq Pervez[1], Hayat Ali Shah[2], Masroor Ellahi Babar[3], Nasir Naveed[2] and Muhammad Shoaib[4]

[1]Department of Bioinformatics and Computational Biology, Virtual University of Pakistan, Lahore, Pakistan. [2]Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan. [3]Department of Biotechnology, Virtual University of Pakistan, Lahore, Pakistan. [4]Department of Computer Science and Engineering, UET, Lahore, Pakistan.

Evolutionary Bioinformatics
Volume 15: 1–4

DOI: 10.1177/1176934318821080

ABSTRACT: Simulated alignments are alternatives to manually constructed multiple sequence alignments for evaluating performance of multiple sequence alignment tools. The importance of simulated sequences is recognized because their true evolutionary history is known, which is very helpful for reconstructing accurate phylogenetic trees and alignments. However, generating simulated alignments require expertise to use bioinformatics tools and consume several hours for reconstructing even a few hundreds of simulated sequences. It becomes a tedious job for an end user who needs a few datasets of variety of simulated sequences. Currently, there is no databank available which may help researchers to download simulated sequences/alignments for their study. Major focus of our study was to develop a database of simulated protein sequences (SAliBASE) based on different varying parameters such as insertion rate, deletion rate, sequence length, number of sequences, and indel size. Each dataset has corresponding alignment as well. This repository is very useful for evaluating multiple alignment methods.

KEYWORDS: SAliBASE, simulated alignment, true alignment

RECEIVED: October 25, 2018. **ACCEPTED:** November 26, 2018.

TYPE: Software or Database Review

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Muhammad Tariq Pervez, Department of Bioinformatics and Computational Biology, Virtual University of Pakistan, 54 Lawrence Road, Lahore 54000, Punjab, Pakistan. Email: tariq_cp@hotmail.com

Background

Multiple sequence alignment (MSA) is an essential basic step required for various bioinformatics approaches such as protein secondary structure prediction, phylogeny inference, and identifying significant functional residues.[1,2] It helps for studying hidden patterns in sequence-structure-function associations of DNA or protein sequence families. The role of MSA for predicting protein structures by homology modeling has been explained very well by the results from the CASP2 and CASP3.[3] Manually curated sequences/alignments for evaluating performance of MSA methods can be downloaded from various online databases such as BAliBASE,[4] PREFAB,[5] and SABmark.[6] Alignments in these repositories have some limitations such as their small size, uncertain positional homology, and lack of evolutionary history among the sequences. Small size does not allow the user to cover a complete range of scenarios of protein evolution, whereas uncertain positional homology makes assessing accuracy of the alignments difficult. Lack of evolutionary history among the sequences hinders testing of phylogenetic software applications. Furthermore, the writers of MSA tools may be misguided to design algorithms for solving issues which are reported only in the manually curated datasets, and high-level skill sets are required for reconstructing accurate alignments.[5]

Simulated alignments are alternatives to manually constructed MSAs, and their importance is recognized because of prior knowledge of their true evolutionary history, which is very helpful for reconstructing accurate phylogenetic trees and alignments.[7] Second, as compared with manually curated alignments, it is easy for the end user to generate simulated alignments; however, it involves a number of steps and consumes a lot of time for generating even a few hundred of sequences. A number of sequence simulators such as ROSE,[8] SIMPROT,[9] MySSP,[10] and Indel-Seq-Gen–2.1.03 (iSG)[11] are available. Each of these software tools has its own strengths and weaknesses. The iSG is famous for generating greatly divergent DNA sequences and protein sequences by integrating several indel models, modeling coding and noncoding DNA evolutions. It has many other features such as addition of motif conservation, indel tracking, lineage-specific evolution, subsequence length constraints, and PROSITE-like regular expressions.

Generating simulated alignments requires expertise to use bioinformatics tools and consume several hours for reconstructing even a few hundreds of simulated sequences. It becomes a tedious job for an end user who needs a few datasets of variety of simulated sequences. A comprehensive study of MSA methods without using simulated sequences as test cases is hard to perform. Currently, there is no databank available which may help researchers to download simulated sequences/alignments for their study. For this reason, we have developed SAliBASE (1.0), first version of simulated protein alignments database. Major focus of our study was to develop a database of simulated sequences (SAliBASE) based on different varying parameters such as insertion rate, deletion rate, sequence length, and indel size. Each dataset has corresponding alignment as well. The deletion and insertion rates represent the occurring of deletions and insertions at the specified intervals which indicates that how much genetic material has been discarded. Indel size indicates the number of deletions and

insertions occurring in protein/DNA sequences.[12] This repository is very useful for evaluating multiple alignment methods.

Construction and Content

SAliBASE 1.0 includes 5 simulated alignment sets. Alignments in "Varying Deletion Rate" dataset were generated using deletion rate ranging from 0.000002 to 0.1. Sequence length was 1000 bp, indel size was 20, number of sequences was 100, and the insertion rate was 0.000002. Dataset namely "Varying Insertion Rate" consists of 100 alignments with insertion rate ranging from 0.000002 to 0.1. Other parameters were kept constant, that is, the number of sequences was 100, indel size was 20, sequence length was 1000 bp, and deletion rate was 0.000002. Dataset entitled "Varying Indel Size" includes alignments with indel size ranging from 100 to 5000. Sequence length was 15 000 bp, insertion and deletion rate were 0.000002, and number of sequences was 100. Dataset entitled "Varying Sequence Length" contains 100 alignments with sequence length ranging from 1000 to 20 800. Other parameters such as number of sequences, insertion rate, deletion rate, and indel size were constant, that is, 100, 0.000002, 0.000002, and 20, respectively. Alignments in "Varying Number of Sequences" dataset were constructed using "number of sequence" parameter ranging from 100 to 100 000. Sequence length was 500, indel size was 20, and deletion rate and insertion were 0.000002. Table 1 shows summary of the 5 sets of simulated alignments.

Materials and Methods

Figure 1 describes all steps of the methodology adapted to generate simulated datasets.

Construction of Simulated Trees

TreeSim package of R was used to generate a total of 104 simulated trees under the birth-death model; 100 for "Varying Number of Sequences" dataset and 1 for each of the other datasets. Figure 2B shows commands to generate simulated sequences in iSGv2.1.03.

Construction of Simulated Alignments

iSGv2.1.03 was used to construct the 5 datasets. Each of the 5 datasets consists of 100 alignments; 100 with varying deletion rate, 100 with varying insertion rate, 100 with varying indel size, 100 with varying number of sequences, and 100 with varying sequence length. Thus, a total of 500 known alignments were generated. Figure 2A shows commands to generate simulated tree in R.

Utility and Discussion

SAliBASE (Figure 3) is a repository of 5 datasets of simulated sequences. Each dataset stores 100 sequence and corresponding alignment text files of varying sizes. The user can download all datasets using a link numbered as "1." Other links provide options for downloading individual sequence files. For example, by selecting the link numbered as "2," a list of 100 sequence

Table 1. Parameters used in 5 sets of simulated alignments.

VARYING DELETION RATE					VARYING INSERTION RATE				
Sequence length	Indel size	Insertion rate	**Deletion rate**	Number of sequences in each alignment	Sequence length	Indel size	**Insertion rate**	Deletion rate	Number of sequences in each alignment
1000	20	0.000002	0.000002-0.1	100	1000	20	0.000002-0.1	0.000002	100
VARYING INDEL SIZES					VARYING SEQUENCE LENGTHS				
Sequence length	**Indel size**	Insertion rate	Deletion rate	Number of sequences in each alignment	**Sequence length**	Indel size	Insertion rate	Deletion rate	Number of sequences in each alignment
15000	100-5000	0.000002	0.000002	100	1000-20800	20	0.000002	0.000002	100
VARYING NUMBER OF SEQUENCES									
Sequence length	Indel size	Insertion rate	Deletion rate	**Number of sequences**					
500	20	0.000002	0.000002	100-100000					

In each of 5 sets, 4 parameters were kept constant and 1 was varying (given in boldface).

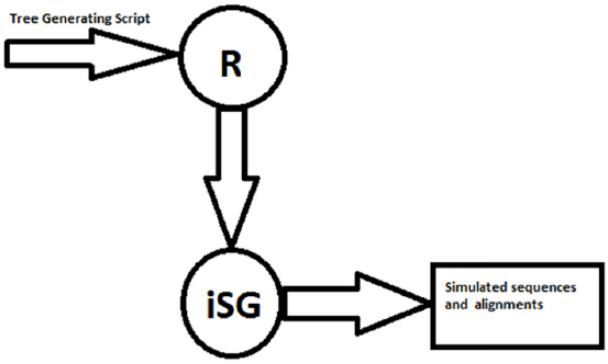

Figure 1. The steps to generate simulated datasets.

A)

# Simulate birth-death trees using various parameters
initTree = sim.bdtree(n=100, extinct=FALSE) // Initial tree with 100 taxa
prunedTree = drop.extinct(initTree) // The tree is pruned based on criteria. Here default criteria have been used. For detail please see R document for this function
write.tree(prunedTree, "100.tre") // Saving tree in "100.tre" file
cat("Done") // completing the proces

B)

./indel-seq-gen -m JTT --outfile [name of the output files] < [simulated tree generated by R]

Figure 2. (A) Command used to generate tree in R and (B) the command used to generate simulated sequences in indel-seq-gen.

files of varying length will be displayed and the user can download the required data. This repository will be very useful for carrying out comparative study of MSA methods which is, currently, one of the important research areas of bioinformatics domain. It will save a lot of time of end user because generating simulated alignment with few hundred sequences needs several hours and multiple steps, and it becomes a frustrating job when a user requires several simulated alignments of varying sizes.

Demonstration of Application of Simulated Alignments

In this section, we describe steps of using simulated alignments for assessing accuracy of an alignment method. The datasets on our website have 2 files in FASTA format. Type of 1 file is "SEQ" and type of the other file is "MA." These 2 files are sequence file and corresponding true alignment file, respectively. After downloading the desired datasets, the user will generate test alignment by providing sequence file to an alignment tool (eg, ClustalO). Now, the user will compare test alignment generated by the selected tool and the true alignment downloaded from our website using sum-of-pairs and total column scores. Figure 2A and B demonstrate the

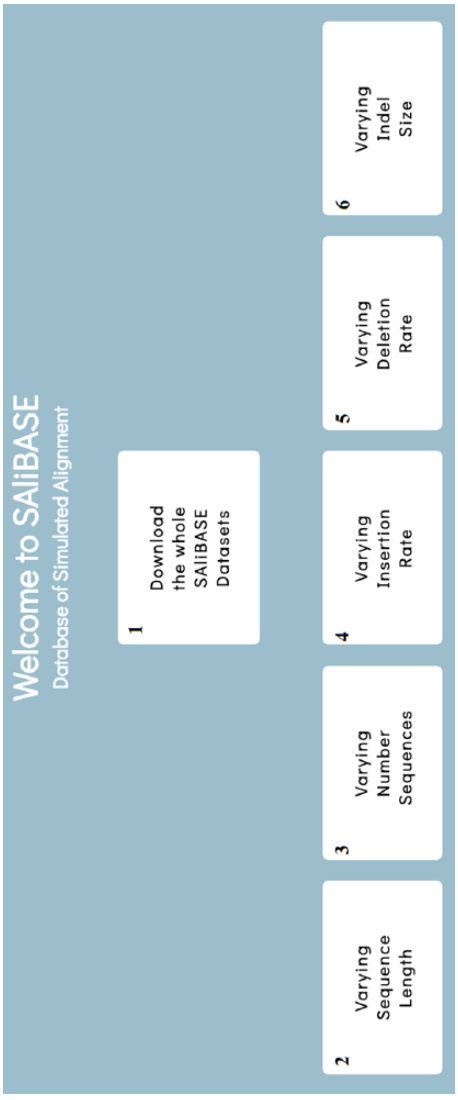

Figure 3. Online interface of SAliBASE which shows links for downloading various datasets.

commands to generate simulated tree in R and simulated sequences in indel-seq-gen, respectively.

Conclusions

Generating simulated alignments requires expert-level skills to use various bioinformatics tools and consumes several hours for reconstructing few hundreds of simulated sequences. It becomes a tedious job for an end user who requires several simulated alignments of varying sizes. A comprehensive study of MSA methods without using simulated sequences as test cases is hard to perform. SAliBASE (1.0 is a database of simulated sequences which were generated based on different varying parameters such as insertion rate, deletion rate, sequence length, and sequence length and indel size). Each dataset has corresponding alignment as well. This repository is very useful for evaluating multiple alignment methods.

Acknowledgements

We are very thankful to Virtual University of Pakistan for her generous support to conduct this research work.

Author Contributions

MTP conceived the idea. MEB and MTP wrote the article. MEB and MS designed datasets. NN and HS generated simulated alignments and developed the website.

Data Availability

Database is available at http://www.salibasepak.com

REFERENCES

1. Pervez MT, Babar ME, Nadeem A, et al. IVisTMSA: interactive visual tools for multiple sequence alignments. *Evol Bioinform*. 2015;11:35.
2. Pervez MT, Babar ME, Nadeem A, et al. MQAT: an efficient quality assessment tool for large multiple sequence alignments. *Life Sci J*. 2013;10:9–16.
3. Jennings AJ, Edge CM, Sternberg MJ. An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein Eng*. 2001;14:227–231.
4. Thompson JD, Koehl P, Ripp R, Poch O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*. 2005;61:127–136.
5. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–1797.
6. Walle IV, Lasters I, Wyns L. SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*. 2005;21:1267–1268.
7. Pervez MT, Babar ME, Nadeem A, et al. Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evol Bioinform Online*. 2014;10:205–217.
8. Stoye J, Evers D, Meyer F. Rose: generating sequence families. *Bioinformatics*. 1998;14:157–163.
9. Pang A, Smith AD, Nuin PAS, Tillier ERM. SIMPROT: using an empirically determined indel distribution in simulations of protein evolution. *BMC Bioinformatics*. 2005;6:236.
10. Rosenberg MS. MySSP: non-stationary evolutionary sequence simulation, including indels. *Evol Bioinform Online*. 2005;1:81–83.
11. Strope CL, Abel K, Scott SD, Moriyama EN. Biological sequence simulation for testing complex evolutionary hypotheses: indel-seq-gen version 2.0. *Mol Biol Evol*. 2009;26:2581–2593.
12. Dang UJ, Devault AM, Mortimer TD, Pepperell CS, Poinar HN, Golding GB. Estimation of gene insertion/deletion rates with missing data. *Genetics*. 2016;204:513–529.