Simulated sequence alignment software: An alternative to MSA benchmarks

Author : Muniba Faiza

Categories : Algorithms, Databases, Sequence Analysis, Softwares, Tools

Date : March 21, 2018

In our previous article, we discussed different multiple sequence alignment (MSA) benchmarks to compare and assess the available MSA programs. However, since last decade, several sequence simulation software have been introduced and are gaining more interest. In this article, we will be discussing various sequence simulating software being used as alternatives to MSA benchmarks.

The basic motivation of using simulated sequence alignments as the reference sets for assessing the quality of MSAs generated by MSA programs is that they help in creating accurate alignments as their evolutionary history is known and can be easily generated by inserting, deleting, or substituting the residues, by changing the sequence length and number of sequences, which is not the case in MSA benchmark alignments. MSA benchmarks are semi-automatedly generated and generate a specific set of reference alignments such as different sets of BAliBase [1-3]. One of the benchmark databases has been generated using a sequence simulator, i.e., ROSE [4]. There is various sequence simulating software available which have been used in assessing the performance of different MSA programs.

ROSE software can be used for DNA, RNA, and protein sequences incorporating indels in accordance

with the evolutionary distance guided by an evolutionary tree [4]. SIMPROT is one of the most widely used sequence simulators, which can be applied to protein sequences only [5]. Indel-Seq-Gen 2.1.03 creates highly divergent DNA sequences and protein families and incorporates various indel models [6]. MySSP also incorporates different models of evolution such as Jukes-Cantor [7], Hasegawa-Kishino-Yano [8], and Kimura-two parameter [9]. Another software DAWG is used to simulate evolution by incorporating the general time reversible model, gamma, and invariant rate heterogeneity [10]. Recently, some other software has been introduced such as NetRecodon [11], PhyloSim [12], ProteinEvolver [13], and ?-BUSS [14].

These sequence simulator software has been applied in different studies on evaluation of MSA programs and found quite helpful providing sets of reference alignments [15,16]. Apart from the advantages of using the sequence simulators as the reference alignments for comparing the performance and quality of MSA programs, it has some pitfalls also: using the simulation settings more close to an MSA program may provide it an excessive advantage [15]. Another drawback is that the simulated sequences cannot explain the evolutionary aspects because of the dependency of all observations obtained from the true alignments on assumptions of the model used to reconstruct the simulated alignments.

References

1. Thompson, J., Plewniak, F., & Poch, O. (1999). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics, 15(1), 87–88. https://doi.org/10.1093/bioinformatics/15.1.87
2. Bahr, A., Thompson, J. D., Thierry, J.-C., & Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Research, 29(1), 323–326. https://doi.org/10.1093/nar/29.1.323
3. Thompson, J. D., Koehl, P., Ripp, R., & Poch, O. (2005). BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. Proteins: Structure, Function, and Bioinformatics, 61(1), 127–136. https://doi.org/10.1002/prot.20527
4. Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. Bioinformatics, 14(2), 157–163. https://doi.org/10.1093/bioinformatics/14.2.157
5. Pang, A., Smith, A. D., Nuin, P. A., & Tillier, E. R. (2005). SIMPROT: using an empirically determined indel distribution in simulations of protein evolution. *BMC bioinformatics*, *6*(1), 236.
6. Strope, C. L., Abel, K., Scott, S. D., & Moriyama, E. N. (2009). Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Molecular biology and evolution*, *26*(11), 2581-2593.
7. Jukes, T. H., & Cantor, C. R. (1969). Evolution of Protein Molecules. Mammalian Protein Metabolism, 3, 21–132. Retrieved from https://books.google.com/books?hl=en&lr=&id=FDHLBAAAQBAJ&oi=fnd&pg=PA21&dq=Evolution+of+protein+molecules+jukes+cantor&ots=blcsZIY2gB&sig=TuCtkRMRPIk0aXXOkOkGAvegaM0#v=onepage&q=Evolution of protein molecules jukes cantor&f=false
8. Hasegawa, M., Kishino, H., & Yano, T. aki. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22(2), 160–174. https://doi.org/10.1007/BF02101694
9. Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions

through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16(2), 111–120. https://doi.org/10.1007/BF01731581
10. Cartwright, R. A. (2005). DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics, 21(Suppl 3), iii31-iii38. https://doi.org/10.1093/bioinformatics/bti1200
11. Arenas, M. (2012). Simulation of molecular data under diverse evolutionary scenarios. PLoS Computational Biology, 8(5). https://doi.org/10.1371/journal.pcbi.1002495
12. Sipos, B., Massingham, T., Jordan, G., & Goldman, N. (2011). PhyloSim – Monte Carlo simulation of sequence evolution in the R statistical computing environment. BMC Bioinformatics, 12(1), 104+. https://doi.org/10.1186/1471-2105-12-104
13. Arenas, M., Dos Santos, H. G., Posada, D., & Bastolla, U. (2013). Protein evolution along phylogenetic histories under structurally constrained substitution models. Bioinformatics, 29(23), 3020–3028. https://doi.org/10.1093/bioinformatics/btt530
14. Bielejec, F., Lemey, P., Carvalho, L., Baele, G., Rambaut, A., & Suchard, M. A. (2014). ?BUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. BMC Bioinformatics, 15(1), 133. https://doi.org/10.1186/1471-2105-15-133
15. Nuin, P. A., Wang, Z., & Tillier, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC bioinformatics*, *7*(1), 471.
16. Pervez, M. T., Babar, M. E., Nadeem, A., Aslam, M., Awan, A. R., Aslam, N., … & Shoaib, M. (2014). Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evolutionary Bioinformatics*, *10*, EBO-S19199.
17. Iantorno, S., Gori, K., Goldman, N., Gil, M., & Dessimoz, C. (2014). Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. In *Multiple Sequence Alignment Methods* (pp. 59-73). Humana Press, Totowa, NJ.

Sharing is caring. Spread the love!