Sequence Similarity Search, Multiple Sequence Alignment, Model Selection, Distance Matrix and Phylogeny Reconstruction

Felix Bast (felix.bast@gmail.com)
Molecular Genetics Laboratory, Central University of Punjab

Method Article

Keywords: Sequence similarity search, Multiple Sequence Alignment, Model Selection, Distance Matrix, Phylogeny Reconstruction

Posted Date: July 11th, 2013

DOI: https://doi.org/10.1038/protex.2013.065

Abstract

This is a generic sequence analysis protocol suitable for plant and algal phylogeographic studies. Generated sequences from bidirectional Sanger sequencers are first assembled using Geneious. Sequence assembly is then trimmed and similarity search is conducted using BLASTn within Geneious. BLAST hits and other target taxa are selected and multiple sequence alignment is constructed. The alignment is then refined by checking using eye and exported as .fasta. Using MEGA, best-fitting nucleic acid substitution models will be calculated in MLModelTest. Model with lowest BIC score is selected and used for further phylogenetic analysis using MEGA or Geneious, which include distance matrix construction, phylogeny reconstruction using ML and BI.

Introduction

Nucleic acid sequence analysis is an integral part of phylogenetic and phylogeographic analyses. This protocol is for the DNA sequence \(s) analysis for phylogeny reconstruction using freeware MEGA and paid version of GeneiousPro- a combination that works best in my experience. For ModelTest, Distance Matrix construction and ML analysis, MEGA is preferred while for Sequence assembly, similarity search, alignment and BI, Geneious is preferred.

Equipment

Windows-based computer with good processing power \(intel i7 with 4GB RAM or equivalent)

Procedure

Sequence Trimming and Assembly 1. Drag sequences from bidirectional Sanger sequence output \(.ab1 files containing electropherograms) into "GeneiousPro"http://www.geneious.com/web/geneious/download-geneious main window. 2. Select both sequences, right click and choose "trim ends" and choose default values. 3. Select trimmed sequences and click deNovo assembly option under "assembly" menu. 4. Right click assembled sequence and choose "Generate Consensus Sequence", name it and describe with NCBI Taxon ID-if any. **BLASTn Sequence Similarity Search** 5. Select the sequence generated in step 4 and choose "sequence search" option on the top menu. 6. In the BLAST options, choose BLASTn \(Nucleotide BLAST) as the database, opt for "Fully annotate hits" and choose 100 as number of hits. 7. Once the search is complete, drag these results to the folder in which you are working. 8. In the left panel, choose 'NCBI > Nucleotide" and get additional sequences of interest. This include target taxonomic representatives and suitable outgroups.Drag all pertinent results to the folder. **Multiple Sequence Alignment \(MSA)** 9. Select all sequences that need to be aligned and choose Ctrl+Shift+A 10. Align first by Geneious alignment, with default parameters. Make sure "Automatically determine sequence's direction" is selected. 11. Align once again \(Ctrl_Shift+A) using MUSCLE alignment with 8 iterations. 12. Once sequence is aligned, zoom to check accuracy of the alignment. Obviously un-alignable sequences should be removed and realigned.

Ends of the alignment may be trimmed to match ends of query sequence. Alignment should be carefully edited by eye and introduce gaps wherever necessary. **Importing MSA in MEGA and performing analysis there** 13. Select the final trimmed alignment and choose Ctrl+Shift+E 14. Choose FASTA and all the default options, and save it to a folder of your data 15. Open this folder, right click on the fasta file and choose “Open with MEGA” 16. In the alignment Explorer window of MEGA, choose “Phylogenetic analysis” in the main menu. 17. Choose appropriate option. If sequence is Introns or other non-coding regions, choose no. If sequence is a CDS/Gene, choose Yes. 18. In the MEGA main menu, choose “Find best DNA/Protein Models” under _Models_. 19. Choose an appropriate options. For sequences with many gaps, “use all sites” may be appropriate. For general, good quality alignments, “Complete deletion” option is better. Perform the ModelTest. 20. In the result table, choose the first model and note its BIC score to quote in paper. 21. In the MEGA main menu, choose Distance > Compute Pairwise Distance 22. In the options, choose appropriate. Choose the best model found in step 20. Choose same options selected for step 19. Perform the analysis. 23. Result of distance matrix will be presented. Choose “export/print distances” from file menu and choose lower-left matrix with excel as output format. 24. In the MEGA main menu, choose Phylogeny >Construct/Test ML Phylogeny 25. Choose appropriate options. Choose the best model found in step 20. Choose same options selected for step 19. Choose 1000 bootstrap replicates. Perform the analysis. 26. Save this tree in a vector format and export as .nexus for uploading to TreeBASE. **Performing Bayesian Inference Phylogeny** 27. Go back to Geneious and choose the same alignment. 28. In the Tree option, choose MrBayes \(with MrBayes add-in installed). Choose pertinent options. Choose the best model found in step 20. Choose same options selected for step 19. If the best model is not available, choose the model with lowest BIC score from available options. Perform the analysis. 29. Save this tree in a vector format and export as .nexus for uploading to TreeBASE. 30. Use appropriate vector image editor \(Adobe Illustrator) to combine these two trees and make the final tree.

Timing

2-5 hours depending on number of taxa and sequence length.

Anticipated Results

1. Assembled sequence 2. Multiple Sequence Alignment 3. Results of ML ModelTest as a table 4. results of pairwise distance analysis as an excel matrix 5. ML Phylogram, with Bootstrap Proportions 6. BI Phylogram with Posterior Probabilities

Acknowledgements

I thank DST-INSPIRE, Government of India to fund this work. I also thank Vice Chancellor, Central University of Punjab for his support.