Multi-SpaM: a Maximum-Likelihood approach to Phylogeny Reconstruction based on Multiple Spaced-Word Matches

Md. Nazmul Hasan, Arefin Rahman Niloy

Abstract

Word-based or 'alignment-free' methods for phylogeny reconstruction are much faster than traditional approaches, but they are generally less accurate. Most of these methods calculate pairwise distances for a set of input sequences, for example from word frequencies, from so-called spaced-word matches or from the average length of common substrings. In this paper, we propose the first word-based approach to tree reconstruction that is based on multiple sequence comparison and Maximum Likelihood. Our algorithm first samples small, gap-free alignments involving four taxa each. For each of these alignments, it then calculates a quartet tree and, finally, the program Quartet MaxCut is used to infer a super tree topology for the full set of input taxa from the calculated quartet trees. Experimental results show that trees calculated with our approach are of high quality.

Keywords: elsarticle.cls, LaTeX, Elsevier, template

2010 MSC: 00-01, 99-00

1. The Elsevier article class

. To gain a better understanding of the evolution of genes or species, reconstructing accurate phylogenetic trees is essential. This can be done using stan-

[⋆] Fully documented templates are available in the elsarticle package on CTAN.

Email addresses: 0419052003 (Md. Nazmul Hasan), nazmulcse25@gmail.com (Md. Nazmul Hasan), 108052108 (Arefin Rahman Niloy), arefinniloy@gmail.com (Arefin Rahman Niloy)

[1] Since 1880.

dard methods which rely on sequence alignments, either of entire genomes or of sets of orthologous genes or proteins. Character-based methods such as Maximum Parsimony [14, 20] or Maximum Likelihood [15] infer trees based on evolutionary substitution events that may have happened since the species evolved from a common ancestor. These methods are generally considered to be accurate, as long as the underlying alignments are of high quality, and as long as suitable substitution models are used. However, for the task of multiple alignment no exact polynomial-time algorithm exists, and even heuristic approaches can be time consuming [46]. Moreover, the most popular heuristic for multiple alignment, the progressive alignment [19], has been shown to be relatively unstable: multiple alignments calculated with progressive approaches and trees inferred from these alignments depend on the underlying guide trees and even on the order of the input sequences [9]. In addition to these difficulties, exact algorithms for character-based phylogeny approaches are themselves NP hard [11, 21].

Usage. Once the package is properly installed, you can use the document class *elsarticle* to create a manuscript. Please make sure that your manuscript follows the guidelines in the Guide for Authors of the relevant journal. It is not necessary to typeset your manuscript in exactly the same way as an article, unless you are submitting to a camera-ready copy (CRC) journal.

Functionality. The Elsevier article class is based on the standard article class and supports almost all of the functionality of that class. In addition, it features commands and options to format the

- document style
- baselineskip
- front matter
- keywords and MSC codes
- theorems, definitions and proofs

- lables of enumerations
- citation style and labeling.

2. Front matter

The author names and affiliations could be formatted in two ways:

(1) Group the authors per affiliation.

(2) Use footnotes to indicate the affiliations.

See the front matter of this document for examples. You are recommended to conform your choice to the journal you are submitting to.

3. Bibliography styles

There are various bibliography styles available. You can select the style of your choice in the preamble of this document. These styles are Elsevier styles based on standard styles like Harvard and Vancouver. Please use BibTEX to generate your bibliography and include DOIs whenever available.

Here are two sample references: [1, 2].

References

References

[1] R. Feynman, F. Vernon Jr., The theory of a general quantum system interacting with a linear dissipative system, Annals of Physics 24 (1963) 118–173. `doi:10.1016/0003-4916(63)90068-X`.

[2] P. Dirac, The lorentz transformation and absolute time, Physica 19 (1-–12) (1953) 888–896. `doi:10.1016/S0031-8914(53)80099-6`.