Quicktree-SD

Stephan Frickenhaus & Bánk Beszteri

August 11, 2009

Bioinformatics, Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570, Bremerhaven, Germany

Abstract

Phylogenetic methods are becoming part of the standard methodologies used in the analyses of biological sequence data, but several of the classical tools of phylogenetic inference were not designed with high throughput applications in mind. During our surveys, we haven't found software for the fast inference of neighbor joining trees from sequence alignments which could use biologically realistic substitution models with protein alignments. Thus, we developed quicktree-sd based on an efficient implementation of the neighbor joining algorithm by [3], by implementing an amino acid distance correction based on Scoredist distances. The tool is available at ftp.awi.de.

High throughput phylogenetic inference is being used to an increasing extent in the context of genomic data. On the one hand, phylogenetic methods are used to improve genome annotation [1],[2]; on the other, genomic data are used in order to understand evolutionary / phylogenetic questions [4].

One of the simpler methods of phylogenetic inference is neighbor joining (NJ; [5]), which provides a good compromise between computational requirements and accuracy, provided that biologically realistic substitution models are used for estimating distances between sequences [7]. Surprisingly, we have not found any available tool for constructing NJ trees from sequence alignments combining performance with a biologically realistic amino acid substitution model.

Quicktree [3] was developed as a tool for the fast inference of NJ trees, with high throughput applications in mind. However, the original implementation of Quicktree did not use an amino acid substitution matrix for calculating the distance matrices. Scoredist was proposed as a simple and generally usable protein sequence distance estimator by [6]. We implemented this distance estimator in quicktree and provide the tool for the scientific community as source code and binaries at ftp.awi.de (also see www.awi.de/en/go/bioinformatics).

References

[1] Jonathan A Eisen and Claire M Fraser. Phylogenomics: intersection of evolution and genomics. *Science*, 300(5626):1706–1707, Jun 2003.

[2] Kristian Hanekamp, Uta Bohnebeck, Bánk Beszteri, and Klaus Valentin. Phylogena–a user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics*, 23(7):793–801, Apr 2007.

[3] Kevin Howe, Alex Bateman, and Richard Durbin. Quicktree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, 18(11):1546–1547, Nov 2002.

[4] Ahmed Moustafa, Bánk Beszteri, Uwe G Maier, Chris Bowler, Klaus Valentin, and Debashish Bhattacharya. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science*, 324(5935):1724–1726, Jun 2009.

[5] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[6] Erik L L Sonnhammer and Volker Hollich. Scoredist: a simple and robust protein sequence distance estimator. *BMC Bioinformatics*, 6:108, 2005.

[7] Koichiro Tamura, Masatoshi Nei, and Sudhir Kumar. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A*, 101(30):11030–11035, Jul 2004.