Letter: TreeAdder: A Tool to Assist the Optimal Positioning of a New Leaf into an Existing Phylogenetic Tree

Derek Gatherer*

MRC Virology Unit, Institute of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, UK

Abstract: TreeAdder is a computer application that adds a leaf in all possible positions on a phylogenetic tree. The resulting set of trees represent a dataset appropriate for maximum likelihood calculation of the optimal tree. TreeAdder therefore provides a utility for what was previously a tedious and error-prone process.

INTRODUCTION

The relationship between the number of operational taxonomic units (OTUs), *n*, and the number of possible bifurcating unrooted trees for that set of OTUs, *N*, is [1, 2]:

$$N = \frac{(2n-5)!}{2^{(n-3)}(n-3)!}$$

Recalculation of the best tree *ab initio* every time a new sequence is added to an alignment is therefore impracticable for large data sets. For instance, for only 7 OTUs (see example below) there are 945 possible unrooted trees. Applications that search for a maximum likelihood tree within a set of candidate trees, such as PAML [3], often recommend that a smaller number of reasonably likely candidate trees be generated by other methods, for instance using tools from PHYLIP [4]. When such preliminary tree selection is used, the subsequent maximum likelihood comparison may not necessarily include the optimal tree [5].

TreeAdder approaches this problem by exhaustively fitting the new OTU leaf onto an existing tree. If the existing tree is well established to the extent that it may be regarded as a 'true' tree, for instance by palaeontology or previous extensive molecular phylogenetics, then the confident assumption may be made that the addition of a further OTU will not alter the topology of that underlying tree. The question then becomes one of finding the maximum likelihood position of the new OTU on the established tree. For a given 'true' tree, TreeAdder rapidly provides the correct set of trees to allow PAML to derive the maximum likelihood position of any new OTU.

AVAILABILITY

The TreeAdder script and the example files used in this note are downloadable by anonymous FTP from: ftp://gamma.vir.gla.ac.uk/pub/ in both zip and gzip formats. Further processing of the TreeAdder output requires a maximum likelihood tool. In the illustrative example given below, PAML (version 3,15) [3] is used (http://abacus.gene.ucl.ac.uk/software/paml.html).

*Address correspondence to this author at the MRC Virology Unit, Institute of Virology, Church Street, Glasgow, G11 5JR, UK; Tel: 0141 330 6268; Fax: 0141 337 2236; E-mail: d.gatherer@mrcvu.gla.ac.uk

The treeadder_files.zip and treeadder_files.tar.gz files contain:

1. **treeadder.pl**: the TreeAdder script.
2. **UL27.trees**: the unrooted well established tree for 6 herpesviruses, based on McGeoch *et al.* [6].
3. **in.tree**: sample TreeAdder output from the above, for input to PAML.
4. **infile**: an alignment of 7 protein sequences in PAML format. 6 of these are from the species in the provided tree UL27.trees. The 7th is the sequence of the OTU to be added to the tree. Along with in.tree, this is the input for the subsequent PAML analysis.
5. **codeml.ctl**: the PAML configuration file for the subsequent PAML analysis.

TreeAdder is a Perl script, and will run on any system where Perl is installed, requiring no special libraries for its operation. Its algorithm is a simple tree traversal with interpolation of new OTU leaves at all valid terminal and internal positions.

PROGRAM OVERVIEW

Operation of TreeAdder is illustrated here by a simple example, using the UL27 (capsid glycoprotein B) proteins of herpesviruses. These were extracted from full genome sequences using Artemis [7, 8], translated using Transeq from EMBOSS [9] and aligned with MAFFT [10, 11]. Alignment columns containing spaces were removed using Compact [12].

The well established phylogeny for the 6 herpesviruses [6] in the initial unrooted tree can be expressed in Newick format as:

((HSV1,HSV2),(EHV1,BHV1),(VZV,SVV));

This is in the provided file UL27.trees. Note that the input tree for TreeAdder should contain only the names of the OTUs as above. Extraneous characters added by any previous tree-building programs (e.g. distances) and carriage returns, should be removed. The task is to place EHV4 at its maximum likelihood position on the above tree. TreeAdder prompts for the tree of the 6 virus species above and the name of the leaf to be added.

The generated file UL27.trees.pamlin contains all 9 possible trees as follows:

1875-0362/07

```
7 9
(((HSV1,EHV4),HSV2),(EHV1,BHV1),(VZV,SVV));
((HSV1,(HSV2,EHV4)),(EHV1,BHV1),(VZV,SVV));
((HSV1,HSV2),((EHV1,EHV4),BHV1),(VZV,SVV));
((HSV1,HSV2),(EHV1,(BHV1,EHV4)),(VZV,SVV));
((HSV1,HSV2),(EHV1,BHV1),((VZV,EHV4),SVV));
((HSV1,HSV2),(EHV1,BHV1),(VZV,(SVV,EHV4)));
(((HSV1,HSV2),EHV4),(EHV1,BHV1),(VZV,SVV));
((HSV1,HSV2),((EHV1,BHV1),EHV4),(VZV,SVV));
((HSV1,HSV2),(EHV1,BHV1),((VZV,SVV),EHV4));
```

Trees 1 to 6 place EHV4 in turn as a sister of the other 6 OTUs. Tree 7 places it as an outgroup to the HSV1-HSV2 clade, tree 8 as an outgroup to the EHV1-BHV1 clade and tree 9 as an outgroup to the VZV-SVV clade.

The alignment of the 7 genes in PAML format (provided as infile) and the TreeAdder output file UL27.trees.pamlin (saved as in.tree) are then used as input for codeml in run-mode 0 (user tree mode). The necessary codeml.ctl file is also provided. Running codeml will then choose the best tree, in this case tree 3 above, demonstrating that EHV4 is a sister clade of EHV1.

The example provided uses protein sequences, but it is unimportant to TreeAdder whether DNA or protein is used in the alignment, provided this is configured in codeml.ctl. Of course, TreeAdder is designed for use with larger and more complex trees than the example above, where the 9 candidate trees could easily be written manually. However, manually adding a leaf to all possible positions on a large tree is difficult, tedious and, worst of all, error-prone. As a result, one may be forced to either generate a sample of candidate trees using another method, or attempt to recalculate the tree *ab initio*. TreeAdder removes the problem of exhaustive leaf addition, thus avoiding both of the above situations. Speed will vary according to the internal complexity of the tree topology as well as the number of OTUs. The largest tree tested, containing 597 OTUs produced an output of 892 candidate trees in 615 seconds.

All the candidate trees produced by TreeAdder are *a priori*. One can therefore be confident that the Kishino-Hasegawa test, implemented in PAML 3.15, is valid [13] and that the best tree is therefore the most likely [5]. The only *caveat* to the method is that one must have good grounds for believing that the TreeAdder input tree is correct. Examples of the application of TreeAdder to real problems can be found in McGeoch and Gatherer [14] and McGeoch *et al*. [15].

FUNDING

The Medical Research Council UK

ACKNOWLEDGEMENTS

The author thanks Duncan J. McGeoch (MRC Virology Unit, Glasgow) for the original suggestion of a requirement for TreeAdder, and several other colleagues for testing the code on their own datasets.

REFERENCES

[1] L. Cavalli-Sforza, and A. Edwards, "Phylogenetic analysis: models and estimation procedures", *Evolution*, vol. 21, pp. 550-570, 1967.

[2] D.L. Quicke, *Principles and Techniques of Contemporary Taxonomy*. London: Blackie, 1993.

[3] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood", *Comp Appl Biosci*, vol. 13, pp. 555-556, 1997.

[4] J. Felsenstein, "PHYLIP - Phylogeny Inference Package (Version 3.2)", *Cladistics*, vol. 5, pp. 164-166, 1989.

[5] N. Goldman, J.P. Anderson, and A.G. Rodrigo, "Likelihood-based tests of topologies in phylogenetics", *Syst Biol*, vol. 49, pp. 652-670, 2000.

[6] D.J. McGeoch, A. Dolan and A.C. Ralph, "Toward a comprehensive phylogeny for mammalian and avian herpesviruses", *J Virol*, vol. 74, pp. 10401-10406, 2000.

[7] K. Rutherford, J. Parkhill, and J. Crook, *et al.* "Artemis: sequence visualization and annotation", *Bioinformatics*, vol. 16, pp. 944-945, 2000.

[8] M. Berriman, and K. Rutherford, "Viewing and annotating sequence data with Artemis", *Brief Bioinform*, vol. 4, pp. 124-132, 2003.

[9] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite", *Trends Genet*, vol. 16, pp. 276-277, 2000.

[10] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform", *Nucleic Acids Res*, vol. 30, pp. 3059-3066, 2002.

[11] K. Katoh, K. Kuma, H. Toh, and T. Miyata, "MAFFT version 5: improvement in accuracy of multiple sequence alignment", *Nucleic Acids Res*, vol. 33, pp. 511-518, 2005.

[12] S. Cook, L. Bell, A.C. Ralph, and D.J. McGeoch (1992-2002, unpublished) Compact: Remove all columns with one or more dots from a given GCG MSF file

[13] H. Kishino, and M. Hasegawa, "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea", *J Mol Evol*, vol. 29, pp. 170-179, 1989.

[14] D.J. McGeoch, and D. Gatherer, "Integrating reptilian herpesviruses into the family herpesviridae", *J Virol*, vol. 79, pp. 725-731, 2005.

[15] D.J. McGeoch, D. Gatherer, and A. Dolan, "On phylogenetic relationships among major lineages of the Gammaherpesvirinae", *J Gen Virol*, vol. 86, pp. 307-316, 2005.

Received: November 12, 2007 Revised: November 26, 2007 Accepted: November 27, 2007