DupliPHY - User Guide

R. M. Ames

Contents

0.1 Change log

Version 2.0. Major update to the newick tree parser to accept a wider variety of newick formatted trees. Added extra exception handling to prevent uninformative program exits. Added a default gain and loss weights matrix which will be used if no matrix file is specified on the command line.

Version 1.0 Initial release.

0.2 License

DupliPHY is licensed under GPL v3 (see gel.txt for more information).

0.3 Citation

If you use DupliPHY please cite [1].

0.4 Overview

Recent large-scale studies of individuals within a population have demonstrated that there is wide-spread variation in copy number in many gene families. In addition, there is increasing evidence that the variation in gene copy number can give rise to substantial phenotypic effects.

In some cases these variations have been shown to be adaptive. These observations show that a full understanding of the evolution of biological function requires an understanding of gene gain and gene loss. Accurate, robust evolutionary models of gain and loss events are, therefore, required.
We have developed weighted parsimony and maximum likelihood methods for inferring gain and loss events[1]. These methods have been tested on a range of simulated data and *Drosophila* data. We have shown that maximum likelihood and weighted parsimony have similar accuracy for reconstructing the ancestral state. For ancestral reconstruction we recommend weighted parsimony because it has similar accuracy to maximum likelihood, but is much faster.

0.5 Installation

DupliPHY has been implemented in Java 1.6 and is distributed as an executable jar file. As such there is no need for any complex installation, if java (version >= 1.6) is installed the program should run on any platform.
To check java is installed on your system type *java -version* at the command prompt. If an error message is displayed you can download and install java (version >= 1.6) from the Oracle website (`http://www.oracle.com/us/technologies/java/overview/index.html`).
Once java is installed simply download the DupliPHY jar file and follow the commands outlined in section 0.8.

0.6 Methods

DupliPHY implements Sankoff's dynamic programming procedure [2], to assign duplication and loss events on a phylogenetic tree. This algorithm uses a post-order tree-traversal and to assign each internal node a cost for each potential character at that node given the characters at the descendants of the node, followed by a pre-order tree-traversal to assign ancestral states.When calculating weighted parsimony with DupliPHY it is possible that multiple gene family sizes have the same parsimony score at the root. In cases of multiple family sizes having the same parsimony score at the root we arbitrarily choose the family with the fewest members. To ensure this choice does not affect the accuracy of DuliPHY we compared the accuracy of choosing the family with the fewest members to choosing a random family; we find there is little difference [1]. The program can use either a user defined matrix or a default matrix (as used in [1]) for the weights of gain and loss events.

0.7 Input data

DupliPHY has 3 mandatory inputs and one optional input.

0.7.1 Family file

The family file is a tab delimited file containing a header line and then a line per family. The header line lists the species in the analysis. Each subsequent line indicates the number of members of that family present in each species. The first column of the file is reserved for a family ID. **NB. All species listed in the family file must be present in the phylogenetic tree.**

FAMILY	dana	dere	dgri	dmel	dmoj	dpse	dsim	dvir	dyak
Fam1	1	1	2	2	1	1	1	1	2

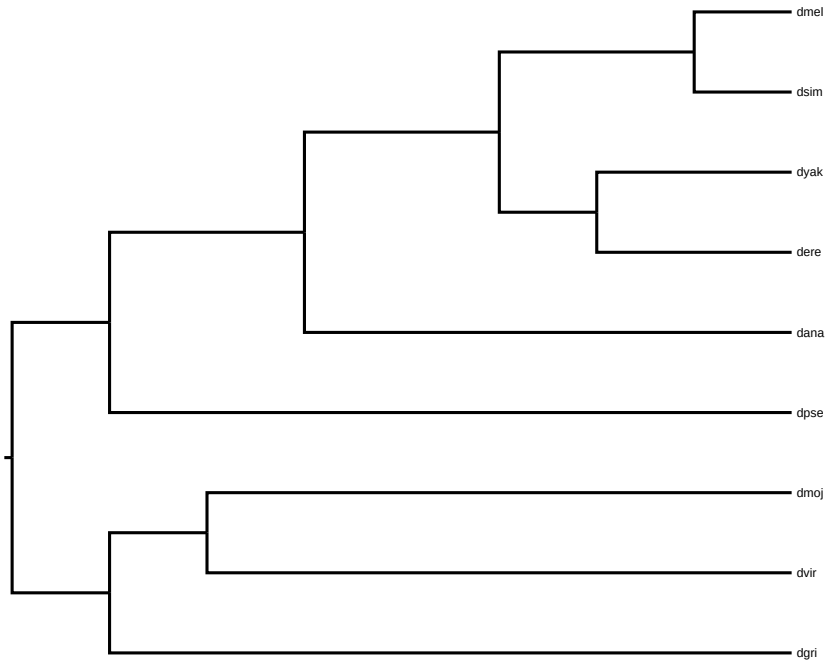

Figure 1: Phylogenetic tree of 9 *Drosophila* species.

Fam2	8	7	6	5	6	7	2	7	8

0.7.2 Tree file

DupliPHY needs a phylogeny in order to infer the ancestral gene family sizes (Figure 1). DupliPHY accepts newick formatted trees. The trees can either have labels on the internal nodes or these labels can be omitted. If no internal nodes labels are present DupliPHY will add these automatically and they will be included in the output. Any of the following trees will be accepted by DupliPHY. **NB. All species in the phylogenetic tree must be present in the gene family file.**

(((((dmel:1.0, dsim:1.0) A:2.0,(dyak:2.0, dere:2.0) B:1.0) C:2.0, dana:5.0) E:2.0, dpse:7.0) F:1.0,((dmoj:6.0, dvir:6.0) G:1.0, dgri:7.0) H:1.0) Root:0.0;

(((((dmel,dsim),(dyak,dere)),dana),dpse),((dmoj,dvir) dgri));

(((((dmel:1.0, dsim:1.0) :1.0,(dyak:1.0, dere:1.0) :1.0) :1.0, dana:1.0) :1.0, dpse:1.0) :1.0,((dmoj:1.0, dvir:1.0) :1.0, dgri:1.0) :1.0);

0.7.3 Results prefix

The final mandatory option that the user must supply is a results prefix. This is simply the name that should be given to the results files.

0.7.4 Matrix file (optional)

By providing a matrix file on the command line DupliPHY will use the user supplied weights to infer gain and loss events. The weights matrix must be a square matrix in the form:

0	1	2	3	4
1	0	1	2	3
2	1	0	1	2
3	2	1	0	1
4	3	2	1	0

For rows (i) and columns (j) the matrix value at i, j defines the cost of a family changing from i members to j members. **NB. The supplied matrix must be at least as large as the largest family in the family file as DupliPHY will not automatically expand a supplied matrix. For example if the largest family size is the dataset is 20 the matrix must be 20x20.**
If the matrix file is omitted from the command line DupliPHY will use the default weights matrix that can be generated from the input family file. This matrix was used in [1] and follows the same pattern as the matrix described above but extended to fit the gene family data.

0.8 Running DupliPHY

DupliPHY is a command line tool where usage follows:

java -jar <family file> <tree file> <results prefix> <matrix file>

family file The tab delimited gene family file (see 0.7.1). e.g. family.txt

tree file The phylogenetic tree in newick format (see 0.7.2). e.g. tree.ph

results prefix The prefix for the generated results files (see 0.7.3). e.g. myResults

matrix file - optional The matrix file with user defined weights for gains and losses (see 0.7.4) e.g. matrix

To run DupliPHY on the supplied example data use:

java -jar family.txt tree.ph myResults matrix
OR
java -jar family.txt tree.ph myResults

This will generate the results files with the prefix myResults.

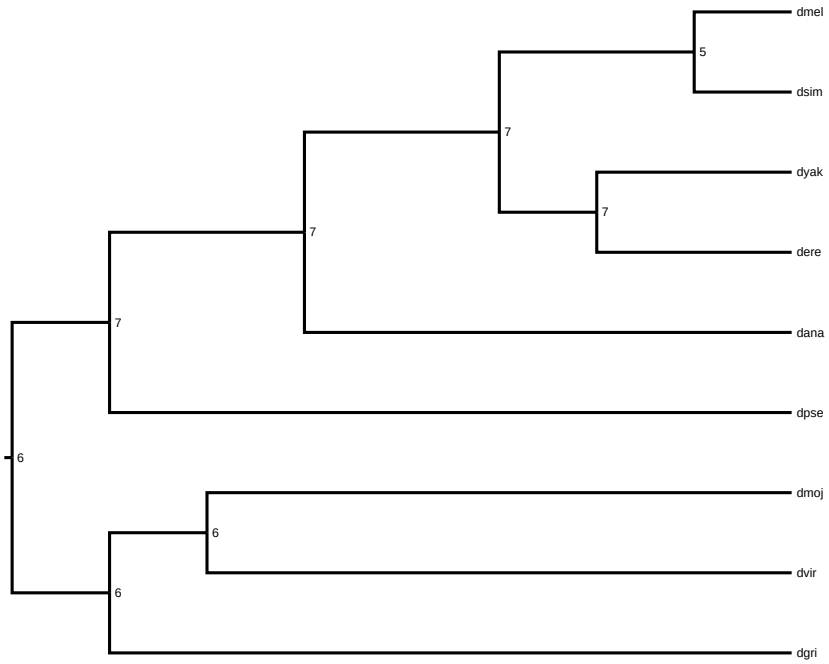

Figure 2: Phylogenetic tree of 9 *Drosophila* species showing the ancestral reconstruction of Fam2.

0.9 Outputs

Running DupliPHY with the command listed in 0.8 will generate two output files:

myResults_desc.txt A tab delimited gene family file showing the extant gene family sizes (as provided by the user) and the ancestral reconstruction at the internal nodes of the tree.

myResults_tree.ph A phylogenetic tree for each input gene family in newick format with the ancestral gene family sizes in the internal node labels. The tree can be viewed in a tree viewer such as fig tree (`http://tree.bio.ed.ac.uk/software/figtree/`) and the internal node values can be displayed to show the evolutionary history of a gene family (Figure 2).

Bibliography

[1] R. Ames, D. Money, V. Ghatge, S. Whelan, and S. Lovell, "Determining the evolutionary history of gene families," *Bioinformatics*, vol. 28, no. 1, pp. 48–55, 2012.

[2] D. Sankoff and P. Rousseau, "Locating the vertices of a steiner tree in an arbitrary metric space," *Mathematical Programming*, vol. 9, no. 1, pp. 240–246, 1975.