PASTIS: Phylogenetic Assembly with Soft Taxonomic InferenceS

Klaas Hartmann, Gavin Thomas, Aki Mimoto, Arne Mooers, Jeffrey Joy, Walter Jetz

July 24, 2013

A bright motmot was acting quite rowdy-
weaving and squawking quite loudly
"Pastis is delise"
he burped with a sneeze
"but it seems to make everything cloudy"
–Arne Mooers

Contents

1 Introduction

PASTIS provides a method for producing fully resolved phylogenetic tree distributions by combining sequence data, taxonomic information and tree constraints. The position of some taxa (particularly those with limited or no sequence information) may be highly uncertain. The PASTIS philosophy is to reflect this uncertainty in the position of these taxa rather than simply omitting these taxa or representing them as unresolved polytomies. For taxa without sequence data placement in the tree is achieved using available taxonomic information (e.g. clade membership). Entire clades may also be included in this manner.

PASTIS assimilates the available information for an analysis into a single mrBayes input file. Beyond the features usually found in a mrBayes input file, an analysis created by PASTIS contains a (potentially) extensive set of generated constraints based on taxonomic information and the provided constraint tree.

After generating the mrBayes input file with PASTIS, the user runs mrBayes (>=3.2) as per usual (this permits use of the standard high performance computing interfaces). After mrBayes finishes execution the user can analyse the generated distribution of trees using their preferred method. PASTIS also provides a method for verifying that taxa have been placed appropriately in the tree.

Throughout PASTIS and this guide the following terminology is used:

Patch: The part of the tree of life being investigated.

Clade: Groups within the patch are referred to as clades. This analysis assumes that clade membership for all taxa is specified.

Taxa: The organisational unit corresponding to the leaves, usually species, or sub-species.

This document provides an overview of the mechanism for using PASTIS. It **does not** provide a motivation or description of the algorithms used by PASTIS. Readers are referred to the paper for an in depth discussion of the algorithms and choices.

2 Quick overview

This section provides a very brief overview and may be most suitable for readers comfortable with R and mrBayes. Subsequent sections provide a more detailed description.

The main function of interest to users is pastis_simple. If the correct files are in your working directory in R, and PASTIS is installed, then running PASTIS can be as simple as:

```
library(pastis)
pastis_simple('PatchName')
```

Where PatchName is the name for your analysis. This will generate a file called PatchName.nexus which you would then use as input for mrBayes (version >=3.2).

To get started quickly example input files are available from figshare here. These provide a good reference point for the range of possible PASTIS inputs. Files beginning pastis_data are simple examples of the most basic PASTIS analyses using only an incomplete phylogeny and a species list. The files beginning Accipitridae provide a more comprehensive example. All of these datasets are provided as part of the pastis package.

The files required in your working directory for a PASTIS analysis are:

3 File formats

For a given analysis PASTIS requires up to four files as previously outlined.

3.1 .tree

The .tree file should be in Newick format and contain a phylogenetic tree for a subset of the taxa. This tree represents the information about this patch that has been well resolved by other studies. We refer to this as the constraint tree. PASTIS will create constraints that force the relationships in this tree to be present in all trees sampled by mrBayes. The constraint tree will typically not include all taxa of interest and need not be fully resolved.

3.2 .taxa

The .taxa file contains a list of all taxa along with their clade membership. The file contains a header "taxon,clade" followed by one line for each taxon. As suggested by the header the first column is the name of the taxon, followed by a comma and then the name of the clade to which that taxon belongs.

For example, consider the .taxa file from the example, the first few lines are:

Table 1: File formats

Filename	Required	Description
PatchName.tree	Yes	Constraint tree containing a subset of all the taxa in your patch. This structure will be included in all output trees and should be basedon the best analysis available for those taxa with good information
PatchName.taxa	Yes	List of all the taxa in your tree and their membership of the various clades in the tree
PatchName.missingclades	Optional	List of clades that are not represented in the constraint tree and where the may be placed in the tree
PatchName.sequences	Optional	Sequence data in FASTA format. This is optional, but will be present in most typical analyses!
PatchName.template	Optional	This is a template file for the mrBayes output file. It outlines options such as the number of iterations, burn in period etc.

taxon,clade
Tyto_alba,Outgroup
Cathartes_aura,Outgroup
Accipiter_albogularis,Accipiter
Accipiter_badius,Accipiter
Accipiter_bicolor,Accipiter
Accipiter_brachyurus,Accipiter
Accipiter_brevipes,Accipiter
Accipiter_butleri,Accipiter
Accipiter_castanilius,Accipiter
...

This defines two taxa ("Tyto_alba" and "Cathartes_aura") as members of the clade "Outgroup" and lists several taxa belonging to "Accipiter".

3.3 .missingclades

Some clades may be completely missing from the constraint tree (.tree), i.e. not a single member of the clade is present in .tree. In this circumstance the default is that the missing clade may attach anywhere (relative to the constraint tree) without entering into any of the other clades (see the later section on the monophyly_constraints and paraphyly_constraints options).

In many cases further information may be available regarding the placement of missing clades relative to the constraint tree. This information can be specified in the .missingclades file.

For each .missingclade it is possible to specify sister clades with which a

missing clade must group (include constraint) and groups of clades which that clade may not enter (exclude constraint). These constraints are specified one line at a time. The first column contains the name of the missing clade, the second column contains the type of constraint (include or exclude), subsequent columns contain names of the clades which define that constraint. All of these clades should be represented in the constraint tree. PASTIS will provide a warning if this is not the case.

Consider the following hypothetical example .missingclades file where A is a missing clade and the other clades are represented in the constraint tree:

A,include,B,C,D,E
A,exclude,B,C

The include line specifies that A attaches to that section of the tree defined by the MRCA of B, C, D and E. The exclude line then specifies that cannot enter that section of the tree defined by the MRCA of B and C.

3.4 .sequences

This contains the aligned sequences for the analysis. This file must be in fasta format.

3.5 .template

PASTIS uses a template for the Nexus file. This defines all the mrBayes options that PASTIS cannot deduce from the input files. This includes information such as the number of chains, burn in period, sampling frequency, model priors etc. If the template file does not exist (or NA is specified to pastis_full), the default template is used. The default template is provided as a starting point and can be examined using:

default_output_template()

You will note that there are four unusual terms in here: "<ntax>", "<nchar>", "<sequences>", "<constraints>" and "<outputfile>". These are placeholders that are automatically replaced by PASTIS with the content generated from input files.

To provide your own template file, simply generate a mrBayes input file of the desired type containing the placeholders as illustrated in the default template.

4 Diagnostics

After executing mrBayes on the .nexus file produced by PASTIS analyses can be conducted on the mrBayes output as per usual. PASTIS provides one additional function, conch, that can be used to verify where clades and species missing from the constraint tree have been placed relative to the constraint tree. To use it simple do:

conch('PatchName.tree','PatchName.nexus.t')

(where PatchName.nexus.t was created by the mrBayes execution).

For each taxon, i, not in the constraint tree this will create a file called 'taxonposition_i.tree'. This tree contains the original constraint tree with the edge lengths equal to the number of sampled trees in which i was descendant from that edge.

The pastis_simple can be run without sequence data to generate a mrBayes .nexus file that will sample from the prior only. When sampling from the prior the topology is determined by the constraints, not by the sequence data. This is an efficient way to check that constraints have been defined correctly prior to a full analysis.

5 Constraints

The PASTIS functions pastis_simple and pastis_main contain two parameters that control the extent to which missing clades and taxa can move around the tree. These parameters are monophyly_constrains and paraphyly_constrains.

5.1 monophyly_constrains

If set to TRUE (default), missing clades and taxa may not enter into monophyletic groups of taxa. This should only be set to FALSE in unusual circumstances, so make sure you know what you are doing!

5.2 paraphyly_constrains

When set to TRUE (default) missing clades and taxa may not enter into clades that are paraphyletic (i.e. mixed). This should only be set to FALSE in unusual circumstances, make sure you know what you are doing!