

T-BAPS

(T-RFLP Bayesian Analysis of Population Structures)

Ver 1.0

Introduction

T-BAPS is a free package for performing clustering analysis using T-RFLP data. T-RFLP is a newly developed molecular fingerprinting technique mainly used to investigate population structures in microbial communities. [1]

T-BAPS provides functionality for building Bayesian models and model selection via BICM. Inference of model parameters combining with the estimation of BICM is realized simultaneously by Markov Chain Monte Carlo simulation in the OpenBUGS environment.

Briefly, the method assumes that the abundance of a particular terminal-restriction fragment length in any population can be modelled as a mixture of Bernoulli and Gaussian distributions. The shape of the mixture densities makes populations characteristically different from each other. T-BAPS aims to find a partition from a sample of T-RFLP data, by first proposing a number of candidate models and then choose the best model with the highest integrated likelihood of the data. For details, see [2].

The inference procedure is automated in T-BAPS according to the following steps:

1. Specifying a range of number of clusters that are plausible for the data.
2. For each of the proposed cluster cardinalities, build the corresponding clustering model.
3. Select the model with the highest integrated likelihoods of the data, using the BICM criterion.
4. Find an optimal partition with the best model.

Getting started

Installation

1. Download and install Matlab Component Runtime from <http://www.abo.fi/fak/mnf/mate/jc/software/t-baps.html>
2. Download and install OpenBUGS from <http://www.mathstat.helsinki.fi/openbugs/>
3. Download and unzip T-BAPS from <http://www.abo.fi/fak/mnf/mate/jc/software/t-baps.html>

Prepare the Data

1. The data must be in a .txt format (plain ASCII).
2. Rows correspond to observations (samples) and columns correspond to variables (fragment lengths)
3. The last column is an index starting from 1 to n, where n is the number of observations.
4. Delimiter could be either comma or blank.
5. The entry values must be non-negative.

An example below shows a data with 4 samples and 3 variables. Note that the index column must be the last column in the data.

```
0, 0, 0.1, 1
0, 0, 0.2, 2
0, 0, 0, 3
0.2,0, 0, 4
```

Run the program

1. Double click on TBAPS.exe. The main GUI will be displayed together with a back window, as shown in Figure 1 and Figure 2.

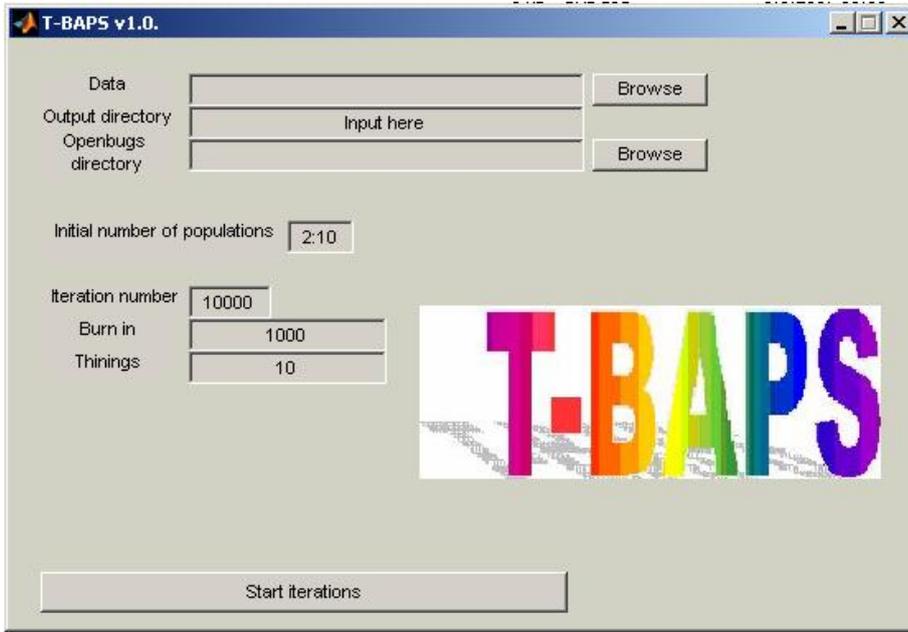


Figure 1 the main GUI of T-BAPS

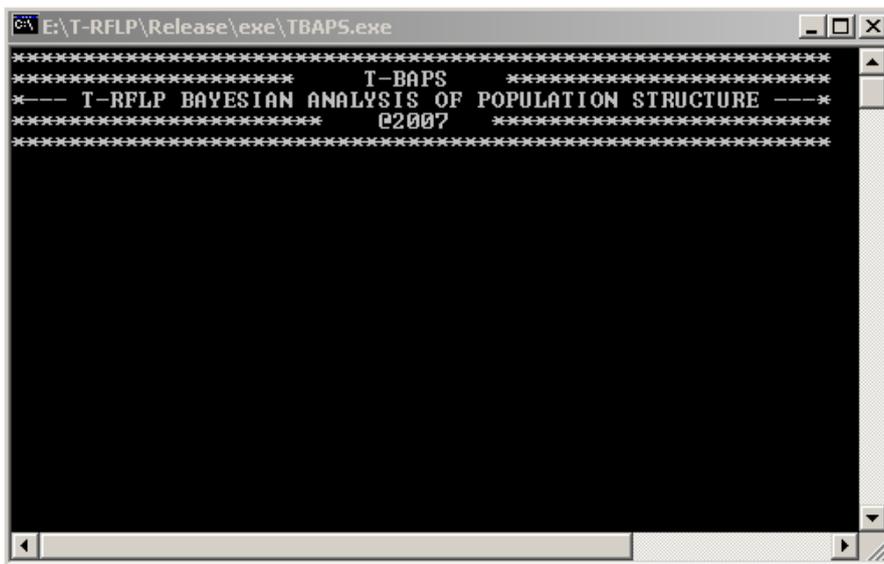


Figure 2 the back window of T-BAPS

Program input

1. Data. Select the T-RFLP data by using the 'Browse' button to the left of the 'Data' text field.
2. Output directory. Input directly the FULL PATH of the directory where all the model and result files are outputted.

3. Openbugs directory. Select the directory where Openbugs has been installed.
4. Initial number of populations. Select the range of plausible number of populations, based on expert knowledge. The default input is 2:10, meaning that T-BAPS will analyze models with initial number of populations starting from 2 to 10. Alternatively one can input a vector explicitly as '2,3,4,5,6,7,8,9,10' without the quotation marks.
5. Iteration number. The number of MCMC iterations. The default is 10,000.
6. Burn in. The number of MCMC iterations which are excluded before the sampling starts. The default is 1000.
7. Thinning. The number of iterations that are jumped over in the MCMC sampling. The default is 10.

After all of the above inputs are specified correctly, clicking on the 'Start iterations' button in the bottom of the GUI will start T-BAPS analysis. The Openbugs software will be called by T-BAPS for MCMC simulation.

NB! Since the running time of MCMC simulation by Openbugs depends on many factors, such as the size of data and the number of iterations, the Openbugs window might appear non-responding for a long time. Please be patient and wait calmly until the MCMC simulation is finished.

Program output

1. In the back window it shows all the relevant information on the analysis.
2. T-BAPS will calculate the BICM score for each initial value of K and determine the best partition with the best model.
3. All the models and results will be saved in the output directory for further analysis.

Reference

- [1] Marsh, T. (1999). Terminal restriction fragment length polymorphism (T-RFLP): an emerging method for characterizing diversity among homologous populations of amplification products. *Curr. Opin. Microbiol.*, 2(3):323-7.
- [2] Tang, J., Tao J., Urakawa, H. and Corander, J. T-BAPS: a Bayesian statistical tool for comparison of microbial communities using terminal-restriction fragment length polymorphism (T-RFLP) data. Submitted for publication