Genome analysis

MPP: a microarray-to-phylogeny pipeline for analysis of gene and marker content datasets

Robert Davey1,2,* , George Savva2, Jo Dicks2 and Ian N. Roberts1

1National Collection of Yeast Cultures, Institute of Food Research and 2Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK

Received on June 30, 2006; revised on January 29, 2007; accepted on January 30, 2007

Advance Access publication February 21, 2007

ABSTRACT

MPP is a Java application, encompassing both new and established algorithms, for the analysis of gene and marker content datasets arising from high-throughput microarray techniques. MPP analyses flat file output from microarray experiments to determine the probability of the presence or absence of genes or markers within a genome. MPP can construct gene or marker content datasets for a number of genomes and can use the data to estimate an evolutionary tree or network. Results from gene content analyses may be validated by comparing them to known gene contents. MPP was initially developed to analyse data derived from comparative genome hybridization (CGH) microarray experiments in fungi and bacteria. It has recently been adapted to analyse retrotransposition-based insertion polymorphism (RBIP) marker scores derived from tagged microarray marker (TAM) experiments in pea. New analytical procedures may be added easily to MPP as plugins in order to increase the scope of the software.

Availability: MPP source code, executables and online help are available at http://cbr.jic.ac.uk/dicks/software/

Contact: robert.davey@bbsrc.ac.uk

1 INTRODUCTION

The proliferation of genome sequencing projects in the last two decades has led to a wealth of information concerning the gene content of many organisms. Since these data have become publicly available, the analysis of gene content has become a popular tool in evolutionary studies. In order to exploit datasets arising from genome sequencing projects, new high-throughput methods such as comparative genome hybridization (CGH) and tagged microarray marker (TAM) microarrays have been developed to provide gene or marker content datasets for closely related organisms.

Unlike classical expression microarrays, a CGH microarray experiment (Kallioniemi et al., 1992) involves the cohybridization of cDNA probes from differentially labelled genomes. Each experiment involves two genomes, a reference genome (often a sequenced and annotated model organism) and a test genome (often a close relative of the reference), and provides intensity ratios for each gene present in the reference genome. Such data can be analysed computationally to predict whether each gene present in the reference genome is present or absent in the test genome. However, no information is provided on genes present in the test genome that are absent in the reference. We refer to this as an unbalanced gene content dataset. By analysing a series of CGH microarrays, each with the same reference genome, it is possible to ascertain the gene contents of each test genome with respect to the reference genome.

In TAM experiments (Flavell et al., 2003), we wish to predict the presence or absence of a particular RBIP marker in a number of test genomes, where these genomes are often derived from varieties within a species. By analysing a number of experiments, where each involves the same test genomes but for different markers, we can estimate the marker content of each test genome, giving rise to a balanced marker content dataset.

For both CGH and TAM microarrays, each genome is thus characterized by a series of 0s and 1s, where 0 represents the absence of a gene or marker and 1 represents its presence. MPP then allows us to infer the evolutionary history of the group of organisms as a phylogenetic tree or network. MPP thus provides an analytical pipeline to take data from the microarray, to predict gene or marker content, to estimate a phylogenetic tree or network and, in certain cases, to validate the output of the analysis.

2 SYSTEMS AND METHODS

MPP is a fast data analysis application developed in the Java programming language. As such, it is platform-independent and takes advantage of a graphical environment to produce colour plots alongside immediately viewable tabular and text outputs.

Input files are preferred in tab-delimited format to optimize loading times, but MPP can also parse Excel workbooks via the Apache POI Java API (The Apache Software Foundation, 2005). Upon successful parsing of the input data, various parameters can be set to fit the format of the dataset, e.g. to define which columns represent the spot intensities. Many other MPP settings may also be modified. For example, bin widths of intensity histograms can be optimized for datasets of variable quality.

MPP requires a dataset of spot intensity values following separate spot identification, or ‘gridding’. It is usual to preprocess this dataset
prior to analysis by transforming the raw intensities, to calibrate values from different samples and to stabilize the variance of the intensity ratios. The commonly used log2 transform is provided as a standard method. Additionally, MPP enables analysis using the $\text{arsinh}$ transform (Huber et al., 2002), which is known to work better than log2 for many data sets exhibiting several low or negative intensity values for one or more channels. By interfacing with the R statistical package via the TCP/IP Rserve daemon (Urbanek, 2005), MPP provides access to the $\text{arsinh}$ transform via the vsn method, made available in the BioConductor suite (Gentleman et al., 2004). Once the intensity values have been transformed, the data may be analysed to provide gene or marker content by one of MPP’s analysis modules, i.e. mppLM or mppTAMLM for CGH and TAM microarray datasets, respectively.

The mppLM algorithm classifies genes into present and absent categories by fitting normal distributions to an optimized range of the overall frequency distribution of intensity ratio values. For this purpose it uses a modified version of a publicly available Java class (Lewis, 2004) implementing the Levenberg–Marquardt method of nonlinear least squares estimation (Marquardt, 1963). After one or more rounds of curve fitting have been performed, depending on the form of the overall frequency distribution, a probability of presence for each gene is calculated using Bayes Theorem. The user may use these probabilities directly or alternatively set a presence/absence threshold so that a probability above a certain value infers a ‘1’ and below the value infers a ‘0’.

mppTAMLM also makes use of the R interface by implementing routines within the mclust R package (Fraley and Raftery, 2002) to predict marker presence or absence via a model-based clustering method. Again, the probability of a particular RBIP marker being present or absent in a particular plant accession can be calculated and a presence or absence status associated with this marker may be assigned.

Results from a single array are saved in tab-delimited format for gene/marker content and as plain text for summary output. MPP uses the JFreeChart library (Gilbert and Morgner, 2005) to generate scatter and line plots of raw and transformed data, with all data points coloured according to presence or absence. Plots may be exported as image files, e.g. SVG, PNG and JPEG. Figure 1 shows an example MPP session.

Fig. 1. A MPP screenshot depicting a sample plot output from a mppLM analysis, with raw data in red, and normal distributions for divergent genes in green and conserved genes in blue. Textual output of the analysis appears in the lower window.

Post-analysis, MPP can assemble results from a related series of CGH or TAM microarrays into a single output file that holds a matrix of the unique IDs against the corresponding status designation, i.e. 0 or 1, in each of the test genomes. This matrix may then be analysed to generate a pairwise evolutionary distance matrix. Specifically, MPP provides a Java port of the CGHdist software (http://cbr.jic.ac.uk/dicks/software/cghdist/, Savva et al., 2005), originally developed in C. CGHdist uses a death process to model gene loss in a group of closely related organisms and is ideally suited to the unbalanced nature of CGH-derived data. Similarly, by considering the TAM data as the loss of empty marker sites, as opposed to the insertion of retrotransposable elements, the software is also suitable for analysis of RBIP markers. MPP provides an interface to the JSplits program (Huson and Bryant, 2006), which enables construction of both phylogenetic trees and networks from distance matrices.

Validation of MPP-predicted gene content matrices can be achieved by comparison to known gene content data discovered through genome sequencing experiments. The predicted values are compared to the known results in a pairwise fashion to elucidate percentage matches, i.e. true positives and true negatives, false positives and false negatives. This enables the user to train MPP settings to data produced by particular array machinery in order to achieve the best results possible, or, concerning MPP development, to test the efficacy of the analytical algorithms. Results from mppTAMLM can be validated by PCR experiments demonstrating that the predicted marker can be amplified from the relevant accession and visualized by gel electrophoresis.
MPP is also fully extensible in that third parties can write Java data transform and analysis module plugins in addition to the available built-in modules, e.g. an Agilent chip data analysis plugin or a plugin interfacing with R to provide the LOWESS data transform. Details of plugin structure, alongside MPP downloads and other related information, are present on the JIC Computational and Systems Biology web pages (http://cbr.jic.ac.uk/dicks/software/mpp/).

ACKNOWLEDGEMENTS

R.D. was sponsored by a BBSRC PhD Studentship and G.S. by a John Innes Foundation PhD studentship. I.R. and J.D. are supported by the BBSRC.

Conflict of Interest: none declared.

REFERENCES


