Phylogenetics

Mtreemix: a software package for learning and using mixture models of mutagenetic trees

Niko Beerenwinkel\textsuperscript{1,*}, Jörg Rahnenführer\textsuperscript{1}, Rolf Kaiser\textsuperscript{2}, Daniel Hoffmann\textsuperscript{3}, Joachim Selbig\textsuperscript{4} and Thomas Lengauer\textsuperscript{1}

\textsuperscript{1}Max Planck Institute for Informatics, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, \textsuperscript{2}Institute of Virology, University of Cologne, Fürst-Pückler-Strasse 56, 50935 Köln, Germany, \textsuperscript{3}Center of Advanced European Studies and Research, Ludwig-Erhard-Allee 2, 53175 Bonn, Germany and \textsuperscript{4}Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany

Received on August 18, 2004; revised on December 21, 2004; accepted on January 11, 2005

Advance Access publication January 18, 2005

ABSTRACT

Summary: Mixture models of mutagenetic trees constitute a class of probabilistic models for describing evolutionary processes that are characterized by the accumulation of permanent genetic changes. They have been applied to model the accumulation of chromosomal gains and losses in tumor development and the development of drug resistance-associated mutations in the HIV genome.

Mtreemix is a software package for estimating mutagenetic trees mixture models from observed cross-sectional data and for using these models for predictions. We provide programs for model fitting, model selection, simulation, likelihood computation and waiting time estimation.

Availability: Mtreemix, including source code, documentation, sample data files and precompiled Solaris and Linux binaries, is freely available for non-commercial users at http://mtreemix.bioinf.mpi-sb.mpg.de/

Contact: niko@math.berkeley.edu

1 INTRODUCTION

Several evolutionary processes are adequately described as the accumulation of non-reversible genetic changes. For example, the progression of tumor development of colorectal cancer can be regarded as the accumulation of chromosomal gains and losses (Vogelstein et al., 1988). These alterations can be detected experimentally by comparative genomic hybridization. Similarly, the development of meningioma has been described as a clonal evolutionary process starting from the set of complete chromosomes and characterized by subsequent chromosomal gains and losses (Zang, 2001).

Amino acid substitutions may also be modeled as permanent under certain conditions. For example, human immunodeficiency virus (HIV) is exposed to strong selective pressure under antiviral drug therapy. In the rapidly evolving viral population advantageous mutations that confer resistance to the drugs are fixed and virtually never lost under continuous drug pressure (Shafer et al., 2000). Thus, the development of drug resistance in the HIV genome can be regarded as the accumulation of resistance-conferring mutations.

More generally, we consider clonal evolutionary processes on a finite set of events. An event can be a genetic alteration, such as the loss of a chromosome arm or an amino acid substitution, but could also be any phenotypic change. We assume that the occurrence of events is non-reversible in the considered time period. Such cumulative evolutionary processes have been modeled by weighted branchings (Desper et al., 1999). These directed tree models define a probability distribution on the set of all patterns of events. Branchings provide an intuitive model of directed dependencies between events and their time of occurrence. Model estimation relies on cross-sectional rather than longitudinal (time series) data. The single tree model (also known as oncogenetic tree) has been extended to mixtures of trees (so-called mutagenetic trees mixture models) in order to capture more complex evolutionary scenarios (Beerenwinkel et al., 2004).

Mtreemix is a software package for analyzing observed event occurrences by means of mutagenetic trees mixture models. We provide efficient C code for a variety of tasks, including model fitting, model selection and comparison, likelihood computation, simulation and computation of the waiting time distribution for all patterns of events (Table 1).

Table 1. Programs in the mtreemix package

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fit</td>
<td>Estimate $K$-mutagenetic trees mixture model from data</td>
</tr>
<tr>
<td>loglike</td>
<td>Log-likelihood computation</td>
</tr>
<tr>
<td>select</td>
<td>Model selection: estimate out-of-sample likelihood</td>
</tr>
<tr>
<td>bootstrap</td>
<td>Analyze model variance</td>
</tr>
<tr>
<td>distr</td>
<td>Compute entire distribution encoded by model</td>
</tr>
<tr>
<td>compare</td>
<td>Compare density estimates</td>
</tr>
<tr>
<td>time</td>
<td>Estimate pattern waiting times</td>
</tr>
<tr>
<td>sim</td>
<td>Draw patterns from model</td>
</tr>
<tr>
<td>wait</td>
<td>Draw patterns and their waiting times from model</td>
</tr>
<tr>
<td>transprob</td>
<td>Transition probabilities between all patterns</td>
</tr>
</tbody>
</table>

*To whom correspondence should be addressed at Department of Mathematics, University of California, 898 Evans Hall, Berkeley, CA 94720, USA.
2 PROGRAMS

Observed patterns of events are encoded in a $N \times \ell$ matrix, where each row corresponds to an observation and each column to an event. Matrix entries 0 and 1 denote the absence and the presence of an event, respectively. Missing data is encoded by $-1$. The program `fit` takes the data matrix and the number $K$ of tree components as input and computes a $K$-mutagenetic trees mixture model. Estimation of a single tree is based on solving an instance of the maximum likelihood in a mixture model as the weighted sum of the single tree transition probabilities, are based on variations of tree traversals in the single trees. Since these operations are linear in the number of trees $K$, their time complexity is $O(K \ell)$. The EM algorithm for model fitting takes $O(K \ell^3 + K \ell N)$ time in each step of the iteration.

Program options are invoked via command line switches. For example, `fit -f filestem -K 3` generates a 3-mutagenetic trees mixture model from the data in the file `filestem.dat`. For each program, the `-h` switch informs about usage, options, input and output files.

4 CONCLUSION

Mtreemix provides efficient C programs for learning mutagenetic trees mixture models from cross-sectional data. The package allows for estimating the expected time to occurrence of specific patterns of events associated, for example, with disease progression or poor treatment outcome.

ACKNOWLEDGEMENT

N.B. acknowledges funding by DFG under grant No. HO 1582/1-3.

REFERENCES


