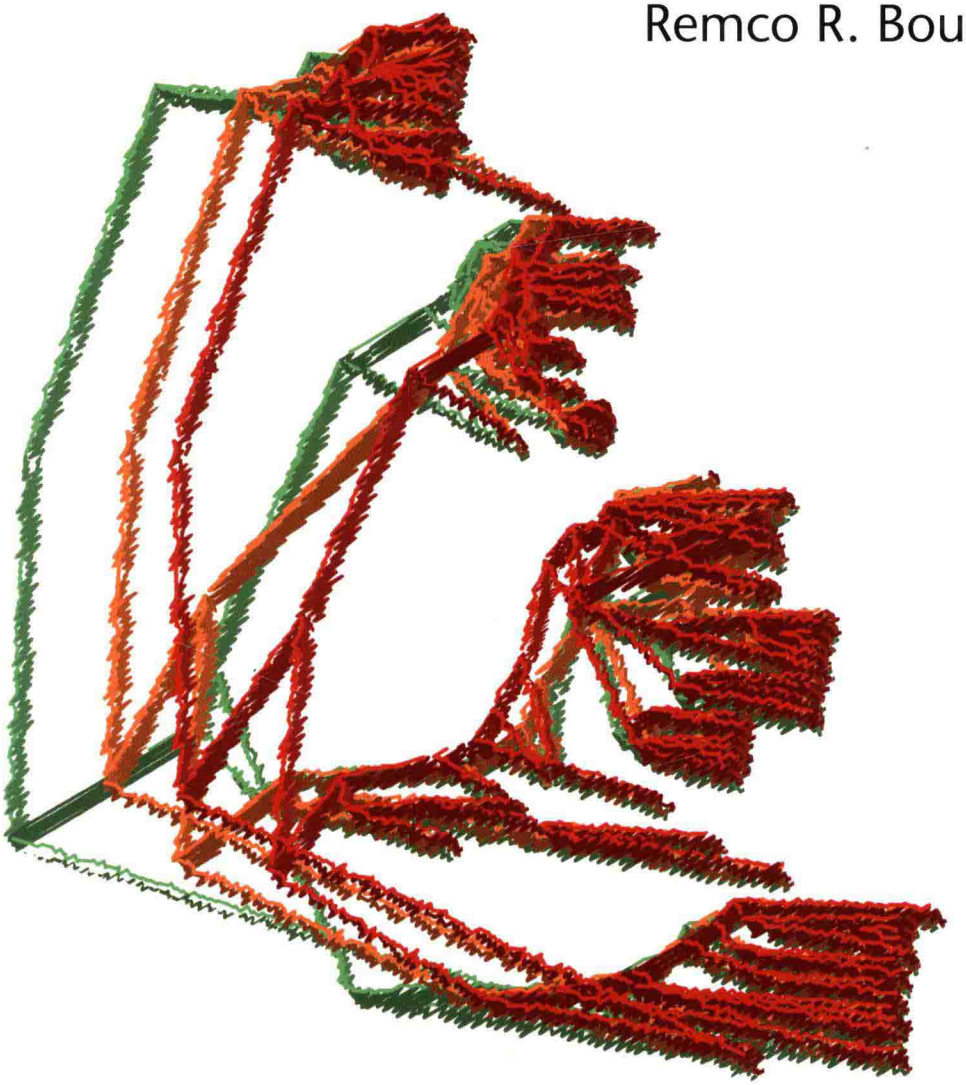
Bayesian Evolutionary Analysis with BEAST

Alexei J. Drummond and
Remco R. Bouckaert



Bayesian Evolutionary Analysis with BEAST

ALEXEI J. DRUMMOND
University of Auckland, New Zealand

REMCO R. BOUCKAERT
University of Auckland, New Zealand

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107019652

First published 2015

Printed in the United Kingdom by Bell and Bain Ltd

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Drummond, Alexei J.
Bayesian evolutionary analysis ; with BEAST 2 / Alexei J. Drummond, University of Auckland, New Zealand, Remco R. Bouckaert, University of Auckland, New Zealand.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-01965-2 (Hardback)
1. Cladistic analysis–Data processing. 2. Bayesian statistical decision theory. I. Bouckaert, Remco R. II. Title.
QH83.D78 2015
578.01′2–dc23 2014044867

ISBN 978-1-107-01965-2 Hardback

Additional resources for this publication at http://beast2.org/book.html

Bayesian Evolutionary Analysis with BEAST

What are the models used in phylogenetic analysis and what exactly is involved in Bayesian evolutionary analysis using Markov chain Monte Carlo (MCMC) methods? How can you choose and apply these models, which parameterisations and priors make sense, and how can you diagnose Bayesian MCMC when things go wrong?

These are just a few of the questions answered in this comprehensive overview of Bayesian approaches to phylogenetics. From addressing theoretical aspects of the field to providing pragmatic advice on how to prepare and perform phylogenetic analysis, this practical guide also includes coverage of the interpretation of analyses and visualisation of phylogenies. The software architecture is described and a guide to developing BEAST 2.2 extensions is provided to allow these models to be extended further.

With an accompanying website (http://beast2.org/) providing example files and tutorials, this one-stop reference to applying the latest phylogenetic models in BEAST 2 will provide essential guidance for all users – from those using phylogenetic tools, to computational biologists and Bayesian statisticians.

Alexei J. Drummond is Professor of Computational Biology and Principal Investigator at the Allan Wilson Centre of Molecular Ecology and Evolution. He is the lead author of the BEAST software package and has gained a reputation in the field as one of the most knowledgeable experts for Bayesian evolutionary analyses.

Remco R. Bouckaert is a computer scientist with a strong background in Bayesian methods. He is the main architect of version 2 of BEAST and has been working on extensions to the BEAST software and other phylogenetics projects in Alexei Drummond's group at the University of Auckland.

Preface

This book consists of three parts: theory, practice and programming. The *theory part* covers theoretical background, which you need to get some insight in the various components of a phylogenetic analysis. This includes trees, substitution models, clock models and, of course, the machinery used in Bayesian analysis such as Markov chain Monte Carlo (MCMC) and Bayes factors.

In the *practice part* we start with a hands-on phylogenetic analysis and explain how to set up, run and interpret such an analysis. We examine various choices of prior, where each is appropriate, and how to use software such as BEAUti, FigTree and DensiTree to assist in a BEAST analysis. Special attention is paid to advanced analysis such as sampling from the prior, demographic reconstruction, phylogeography and inferring species trees from multilocus data. Interpreting the results of an analysis requires some care, as explained in the post-processing chapter, which has a section on troubleshooting with tips on detecting and preventing failures in MCMC analysis. A separate chapter is dedicated to visualising phylogenies.

BEAST 2.2 uses XML as a file format to specify various kinds of analysis. In the third part, the XML format and its design philosophy are described. BEAST 2.2 was developed as a platform for creating new Bayesian phylogenetic analysis methods, by a modular mechanism for extending the software. In the *programming part* we describe the software architecture and guide you through developing BEAST 2.2 extensions.

We recommend that everyone reads Part I for background information, especially introductory Chapter 1. Part II and Part III can be read independently. Users of BEAST should find much practical information in Part II, and may want to read about the XML format in Part III. Developers of new methods should read Part III, but will also find useful information about various methods in Part II.

The BEAST software can be downloaded from http://beast2.org and for developers, source code is available from https://github.com/CompEvol/beast2/. There is a lot of practical information available at the BEAST 2 wiki (http://beast2.org), including links to useful software such as Tracer and FigTree, a list of the latest packages and links to tutorials. The wiki is updated frequently. A BEAST users' group is accessible at http://groups.google.com/group/beast-users.

Acknowledgements

Many people made BEAST what it is today. Andrew Rambaut brought the first version of BEAST to fruition with AJD in the 'Oxford years' and has been one of the leaders of development ever since. Marc Suchard arrived on the scene a few years later, precipitating great advances in the software and methods, and continues to have a tremendous impact. All of the members of the core BEAST development team have been critical to the software's success.

Draft chapters of this book were greatly improved already by feedback from a large number of colleagues, including in alphabetical order, Richard Brown, David Bryant, Rampal Etienne, Alex Gavryushkin, Sasha Gavryushkina, Russell Gray, Simon Greenhill, Denise Kühnert, Tim Vaughan, David Welch, Walter Xie.

Paul O. Lewis created the idea for Figure 1.6. Section 2.3 is derived from work by Joseph Heled. Tanja Stadler co-wrote Chapter 2. Some material for Chapters 7 and 10 is derived from messages on the BEAST mailing list and the FAQ of the BEAST wiki. Section 8.4 is partly derived from 'A rough guide to SNAPP' (Bouckaert and Bryant 2012). Walter Xie was helpful in quality assurance of the software, in particular regression testing of BEAST ensuring that the analyses are valid. Parts of Chapter 4 derive from previous published work by AJD and co-authors.

The development of BEAST 2 was supported by four meetings funded by the National Evolutionary Synthesis Center (www.nescent.org). AJD was funded to write this book by a Rutherford Discovery Fellowship from the Royal Society of New Zealand.

Summary of most significant capabilities of BEAST 2

Analysis	Estimate phylogenies from alignments	
	Estimate dates of most recent common ancestors	
	Estimate gene and species trees	
	Infer population histories	
	Epidemic reconstruction	
	Estimate substitution rates	
	Phylogeography	
	Path sampling	
	Simulation studies	
Models	Trees	Gene trees, species trees, structured coalescent, serially sampled trees
	Tree-likelihood	Felsenstein, Threaded, BEAGLE
		Continuous, ancestral reconstruction
		SNAPP
		Auto partition
	Substitution models	JC96, HKY, TN93, GTR
		Covarion, stochastic Dollo
		RB, subst-BMA
		BLOSUM62, CPREV, Dayhoff, JTT, MTREV, WAG
	Frequency models	Fixed, estimated, empirical
	Site models	Gamma site model, mixture site model
	Tree priors	Coalescent constant, exponential, skyline
		Birth–death Yule, birth–death sampling skyline
		Yule with calibration correction
		Multispecies coalescent
	Clock models	Strict, relaxed, random local clock
	Prior distributions	Uniform, 1/X, normal, gamma, beta, etc.
Tools	BEAUti	GUI for specifying models
		Support for hierarchical models
		Flexible partition and parameter linking
		Read and write models
		Extensible through templates
		Manage BEAST packages
	BEAST	Run analysis specified by BEAUti
	ModelBuilder	GUI for visualising models
	LogCombiner	Tool for manipulating log files

	EBSPAnalyser	Reconstruct population history from EBSP analysis
	DensiTree	Tool for analysing tree distributions
	TreeAnnotator	Tool for creating summary trees from tree sets
	TreeSetAnalyser	Tool for calculating statistics on tree sets
	SequenceSimulator	Generate alignments for simulation studies
Check pointing	Resuming runs when ESS is not satisfactory Exchange partial states to reduce burn-in	
Documentation	Tutorials, Wiki, User forum This book	
Package support	Facilitate fast bug fixes and release cycles independent of core release cycle Package development independent of core releases	

Contents

Part I

Theory

1 Introduction

This book is part science, part technical, and all about the computational analysis of heritable traits: things like genes, languages, behaviours and morphology. This book is centred around the description of the theory and practice of a particular open source software package called BEAST (Bayesian evolutionary analysis by sampling trees). The BEAST software package started life as a small science project in New Zealand but it has since grown tremendously through the contributions of many scientists from around the world, chief among them the research groups of Alexei Drummond, Andrew Rambaut and Marc Suchard. A full list of contributors to the BEAST software package can be found on the BEAST GitHub page or printed to the screen when running the software.

Very few things challenge the imagination as much as does evolution. Every living thing is the result of the unfolding of this patient process. While the basic concepts of Darwinian evolution and natural selection are second nature to many of us, it is the detail of life's tapestry which still inspires an awe of the natural world. The scientific community has spent a couple of centuries trying to understand the intricacies of the evolutionary process, producing thousands of scientific articles on the subject. Despite this Herculean effort, it is tempting to say that we have only just scratched the surface.

As with many other fields of science, the study of biology has rapidly become dominated by the use of computers in recent years. Computers are the only way that biologists can effectively organise and analyse the vast amounts of genomic data that are now being collected by modern sequencing technologies. Although this revolution of data has really only just begun, it has already resulted in a flourishing industry of computer modelling of molecular evolution.

This book has the modest aim of describing this still new computational science of evolution, at least from the corner we are sitting in. In writing this book we have not aimed for it to be comprehensive and gladly admit that we mostly focus on the models that the BEAST software currently supports. Dealing, as we do, with computer models of evolution, there is a healthy dose of mathematics and statistics. However, we have made a great effort to describe in plain language, as clearly as we can, the essential concepts behind each of the models described in this book. We have also endeavoured to provide interesting examples to illustrate and introduce each of the models. We hope you enjoy it.

1.1 Molecular phylogenetics

The informational molecules central to all biology are deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and protein sequences. These three classes of molecules are commonly referred to in the molecular evolutionary field as *molecular sequences*, and from a mathematical and computational point of view an informational molecule can often be treated simply as a linear sequence of symbols on a defined alphabet (see Figure 1.1). The individual building blocks of DNA and RNA are known as *nucleotides*, while proteins are composed of 20 different *amino acids*. For most life forms it is the DNA double-helix that stores the essential information underpinning the biological function of the organism and it is the (error-prone) replication of this DNA that transmits this information from one generation to the next. Given that replication is a binary reaction that starts with one genome and ends with two similar if not identical genomes, it is unsurprising that the natural and appropriate structure for visualising the replication process over multiple generations is a bifurcating tree. At the broadest scale of consideration the structure of this tree represents the relationships between species and higher-order taxonomic groups. But even when considering a single gene within a single species, the ancestral relationships among genes sampled from that species will be represented by a tree. Such trees have come to be referred to as *phylogenies* and it is becoming clear that the field of *molecular phylogenetics* is relevant to almost every scientific question that deals with the informational molecules of biology. Furthermore, many of the concepts developed to understand molecular evolution have turned out to transfer with little modification to the analysis of other types of heritable information in natural systems, including language and culture. It is unsurprising then that a book on computational evolutionary analysis would start with phylogenetics.

The study of phylogenetics is principally concerned with reconstructing the evolutionary history (phylogenetic tree) of related species, individuals or genes. Although algorithmic approaches to phylogenetics pre-date genetic data, it was the availability of genetic data, first allozymes and protein sequences, and then later DNA sequences, that provided the impetus for development in the area.

A phylogenetic tree is estimated from some data, typically a multiple sequence alignment (see Figure 1.2), representing a set of homologous (derived from a common ancestor) genes or genomic sequences that have been aligned, so that their comparable regions are matched up. The process of aligning a set of homologous sequences is itself a hard computational problem, and is in fact entangled with that of estimating a phylogenetic tree (Lunter et al. 2005; Redelings and Suchard 2005). Nevertheless, following convention we will – for the most part – assume that a multiple sequence alignment is known and predicate phylogenetic reconstruction on it.

DNA	{A,C,G,T}
RNA	{A,C,G,U}
Proteins	{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}

Figure 1.1 The alphabets of the three informational molecular classes.

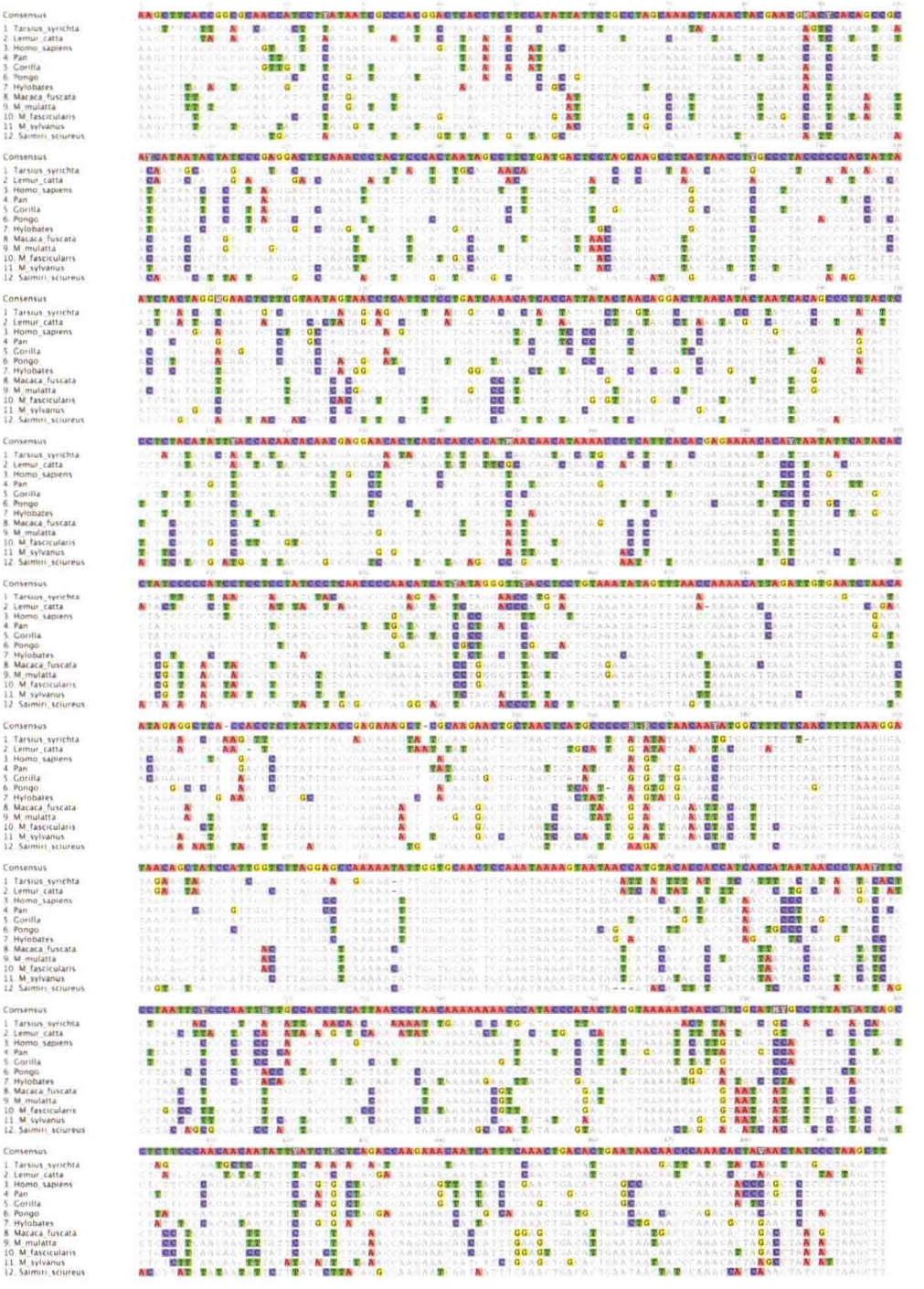


Figure 1.2 A small multiple sequence alignment of mitochondrial sequence fragments isolated from 12 species of primate. The alignment has 898 alignment columns and the individual sequences vary in length from 893 to 896 nucleotides long. Individual differences from the consensus sequence are highlighted. 373/898 (41.5%) sites are identical across all 12 species and the average pairwise identity is 75.7%. The data matrix size is 10 776 (898 × 12) with only 30 gap states. This represents a case in which obtaining an accurate multiple sequence alignment from unaligned sequences is quite easy and taking account of alignment uncertainty is probably unnecessary for most questions.