

## 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.

Cambridge University Press  
978-1-107-01965-2 - Bayesian Evolutionary Analysis with BEAST  
Alexei J. Drummond and Remco R. Bouckaert  
Frontmatter  
[More information](#)

---

Cambridge University Press  
978-1-107-01965-2 - Bayesian Evolutionary Analysis with BEAST  
Alexei J. Drummond and Remco R. Bouckaert  
Frontmatter  
[More information](#)

# Bayesian Evolutionary Analysis with BEAST

ALEXEI J. DRUMMOND

University of Auckland, New Zealand

REMCO R. BOUCKAERT

University of Auckland, New Zealand



Cambridge University Press  
978-1-107-01965-2 - Bayesian Evolutionary Analysis with BEAST  
Alexei J. Drummond and Remco R. Bouckaert  
Frontmatter  
[More information](#)

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](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9781107019652](http://www.cambridge.org/9781107019652)

© Alexei J. Drummond and Remco R. Bouckaert 2015

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

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>

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

	<i>Preface</i>	<i>page</i> ix
	<i>Acknowledgements</i>	x
	<i>Summary of most significant capabilities of BEAST 2</i>	xi
<b>Part I</b>	<b>Theory</b>	<b>1</b>
<b>1</b>	<b>Introduction</b>	<b>3</b>
	1.1 Molecular phylogenetics	4
	1.2 Coalescent theory	6
	1.3 Virus evolution and phylodynamics	8
	1.4 Before and beyond trees	8
	1.5 Probability and Bayesian inference	10
<b>2</b>	<b>Evolutionary trees</b>	<b>21</b>
	2.1 Types of trees	21
	2.2 Counting trees	24
	2.3 The coalescent	27
	2.4 Birth–death models	36
	2.5 Trees within trees	40
	2.6 Exercise	43
<b>3</b>	<b>Substitution and site models</b>	<b>44</b>
	3.1 Continuous-time Markov process	45
	3.2 DNA models	46
	3.3 Codon models	51
	3.4 Microsatellite models	52
	3.5 Felsenstein’s likelihood	52
	3.6 Rate variation across sites	54
	3.7 Felsenstein’s pruning algorithm	55
	3.8 Miscellanea	57

vi	<b>Contents</b>	
<b>4</b>	<b>The molecular clock</b>	<b>58</b>
4.1	Time-trees and evolutionary rates	58
4.2	The molecular clock	58
4.3	Relaxing the molecular clock	60
4.4	Calibrating the molecular clock	65
<b>5</b>	<b>Structured trees and phylogeography</b>	<b>68</b>
5.1	Statistical phylogeography	68
5.2	Multi-type trees	69
5.3	Mugration models	71
5.4	The structured coalescent	71
5.5	Structured birth–death models	73
5.6	Phylogeography in a spatial continuum	73
5.7	Phylogenetics with structured trees	74
5.8	Conclusion	76
	<b>Part II Practice</b>	<b>77</b>
<b>6</b>	<b>Bayesian evolutionary analysis by sampling trees</b>	<b>79</b>
6.1	BEAUti	80
6.2	Running BEAST	86
6.3	Analysing the results	89
6.4	Marginal posterior estimates	90
6.5	Obtaining an estimate of the phylogenetic tree	91
6.6	Visualising the tree estimate	94
6.7	Comparing your results to the prior	94
<b>7</b>	<b>Setting up and running a phylogenetic analysis</b>	<b>97</b>
7.1	Preparing alignments	97
7.2	Choosing priors/model set-up	100
7.3	Miscellanea	112
7.4	Running BEAST	114
<b>8</b>	<b>Estimating species trees from multilocus data</b>	<b>116</b>
8.1	Darwin’s finches	116
8.2	Bayesian multispecies coalescent model from sequence data	119
8.3	*BEAST	119
8.4	SNAPP	123
<b>9</b>	<b>Advanced analysis</b>	<b>127</b>
9.1	Sampling from the prior	127
9.2	Serially sampled data	128

	Contents	vii
	9.3 Demographic reconstruction	129
	9.4 Ancestral reconstruction and phylogeography	134
	9.5 Bayesian model comparison	135
	9.6 Simulation studies	138
10	<b>Posterior analysis and post-processing</b>	139
	10.1 Trace log file interpretation	140
	10.2 Model selection	145
	10.3 Troubleshooting	148
11	<b>Exploring phylogenetic tree space</b>	154
	11.1 Tree space	154
	11.2 Methods of exploring tree space	156
	11.3 Tree set analysis methods	157
	11.4 Summary trees	159
	11.5 DensiTree	163
	<b>Part III Programming</b>	167
12	<b>Getting started with BEAST 2</b>	169
	12.1 A quick tour of BEAST 2	170
	12.2 BEAST core: BEAST-objects and inputs	173
	12.3 MCMC library	174
	12.4 The evolution library	180
	12.5 Other bits and pieces	182
	12.6 Exercise	183
13	<b>BEAST XML</b>	184
	13.1 What is XML?	184
	13.2 BEAST file format and the parser processing model	186
	13.3 An annotated example	190
	13.4 Exercise	193
14	<b>Coding and design patterns</b>	195
	14.1 Basic patterns	195
	14.2 Input patterns	198
	14.3 <code>initAndValidate</code> patterns	200
	14.4 <code>CalculationNode</code> patterns	201
	14.5 Common extensions	203
	14.6 Tips	204
	14.7 Known ways to get into trouble	205
	14.8 Exercise	206

viii	<b>Contents</b>	
<b>15</b>	<b>Putting it all together</b>	<b>207</b>
	15.1 Introduction	207
	15.2 What is a package?	208
	15.3 BEAUti	209
	15.4 Variable selection-based substitution model package example	214
	15.5 Exercise	219
	<i>References</i>	220
	<i>Index of authors</i>	240
	<i>Index of subjects</i>	244



## 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](http://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	