Phylogenetic Networks

Concepts, Algorithms and Applications

The evolutionary history of species is traditionally represented using a rooted phylogenetic tree. However, when reticulate events such as hybridization, horizontal gene transfer or recombination are believed to be involved, phylogenetic networks that can accommodate non-treelike evolution have an important role to play.

This book provides the first interdisciplinary overview of phylogenetic networks. Beginning with a concise introduction to both phylogenetic trees and phylogenetic networks, the fundamental concepts and results are then presented for both rooted and unrooted phylogenetic networks. Current approaches and algorithms available for computing phylogenetic networks from different types of datasets are then discussed, accompanied by examples of their application to real biological datasets. The book also summarizes the algorithms used for drawing phylogenetic networks, along with the existing software for their computation and evaluation.

All datasets, examples and other additional information and links are available from the book's companion website at: www.phylogenetic-networks.org.

Daniel H. Huson is Professor of Algorithms in Bioinformatics at Tübingen University. He has authored numerous papers in bioinformatics, biology and mathematics, and is the main author of the widely used computer programs Dendroscope, MEGAN and SplitsTree.

Regula Rupp received her Ph.D. in Mathematics from Bern University in 2006. Between 2007 and 2009 she held a postdoctoral research position at Tübingen University, working with Daniel H. Huson in developing robust methods for computing phylogenetic networks from real biological data.

Celine Scornavacca is a postdoctoral researcher working on algorithms for phylogenetic networks with Daniel H. Huson at Tübingen University. She received her Ph.D. in Computer Science from Montpellier University in 2009.

Phylogenetic Networks

Concepts, Algorithms and Applications

Daniel H. Huson

Eberhard-Karls-Universität Tübingen, Germany

Regula Rupp

Eberhard-Karls-Universität Tübingen, Germany

Celine Scornavacca

Eberhard-Karls-Universität Tübingen, Germany

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521755962

First published 2010

Printed in the United Kingdom at the University Press, Cambridge

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

Library of Congress Cataloging in Publication data
Huson, Daniel H.
Phylogenetic networks : concepts, algorithms and applications / Daniel H. Huson, Regula Rupp,
Celine Scornavacca.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-75596-2 (hardback)
1. Cladistic analysis – Data processing. 2. Cladistic analysis – Mathematics. 3. Phylogeny.
I. Rupp, Regula. II. Scornavacca, Celine. III. Title.
QH83.H87 2011
578.01′2 – dc22 2010037669

ISBN 978-0-521-75596-2 Hardback

Contents

Cambridge University Press
978-0-521-75596-2 - Phylogenetic Networks: Concepts, Algorithms and Applications
Daniel H. Huson, Regula Rupp and Celine Scornavacca
Frontmatter
More information

Preface

The evolutionary history of a set of species is usually described by a rooted phylogenetic tree. The concept of a rooted tree is very simple and has proved to be extremely useful in many application domains. However, *the truth is rarely pure and never simple.*[1]

By definition, phylogenetic trees are well suited to represent evolutionary histories in which the main events are speciations (at the internal nodes of the tree) and descent with modification (along the edges of the tree). But such trees are less suited to model mechanisms of *reticulate evolution* [219], such as horizontal gene transfer, hybridization, recombination or reassortment. Moreover, mechanisms such as incomplete lineage sorting, or complicated patterns of gene duplication and loss, can lead to incompatibilities that cannot be represented on a tree. Although the analysis of individual genes or short stretches of genomic sequence often gives strong support to a phylogenetic tree, different genes or sequence segments usually support different trees.

While it is generally undisputed that bifurcating speciation events and descent with modifications are major forces of evolution, there is also a growing belief that reticulate events play an important role in the shaping of evolutionary histories, too [55, 61, 111, 173].

Horizontal gene transfer (HGT), the direct transfer of genes from one organism to another, is known to occur very frequently in the prokaryotic world, the main mechanisms being transformation, conjugation and transduction [13, 28, 189, 231]. Because of horizontal gene transfer, even between quite distantly related species, phylogenetic trees on the same taxa based on different genes may be very incongruent and the questions arise of how to define the concept of a species tree for a set of prokaryotic taxa, and then how to infer it? One answer may be to represent the evolutionary history of a set of prokaryotes by an appropriate rooted phylogenetic *network* that encompasses the different gene histories [54, 153, 157].

[1] Oscar Wilde, The Importance of Being Earnest, 1895.

In a more general setting, horizontal gene transfers are considered together with gene duplication and loss events [56, 197]. Here the goal is to reconcile incongruent gene trees with a given species tree under a model of *duplication, loss and transfer* (DLT) [100].

Speciation by hybridization is a widespread phenomenon in plants [201, 202], but also occurs in some other types of organisms [153, 165]. In allopolyploidization, two individuals from distinct species hybridize and merge their sets of chromosomes. In rare cases this produces a new fertile species that is reproductively isolated from the parent species. In diploid hybridization, two parents from different species each supply a gamete and produce a diploid hybrid. In very rare cases the hybrid may become reproductively isolated from the parents and then evolve as a distinct species. Phylogenies involving hybrid species are more informative when they explicitly include postulated hybridization events.

Recombination and gene conversion produce new combinations of genetic material through pairing and shuffling of very similar DNA sequences [190]. It is usually considered a mechanism that belongs within the realm of population genetics, which deals with the statistical analysis of the inheritance and prevalence of genes in populations [118]. In this context, the evolution of sequences under the *coalescent-with-recombination* model gives rise to an *ancestral recombination graph* (ARG) [89, 108], which is used for statistical inference and is beyond the scope of this book. However, as sequencing technologies advance and more projects aim at the full (re)sequencing of many individuals, strains and species [242], the fields of phylogenetic analysis and population genetics are drawing closer together. Being able to explicitly represent recombination in a network is of value to both fields [98, 128, 169, 194].

When interspecific recombination occurs, it may result in different histories for different segments of an individual gene and thus impact the performance of phylogenetic tree reconstruction methods [198, 207, 211]. In such cases, a network reconstruction method may be more suitable.

Reassortment is akin to recombination in that it involves the swapping of genetic material between individual organisms. Many viruses, such as influenza A, have segmented genomes. The evolution of such viruses involves mutations, of course. Moreover, when the viruses co-infect a host cell, then segments of their genomes can be swapped in a process called reassortment [47]. Hence, a phylogenetic tree will not always suffice to correctly represent the evolutionary history of a population of such viruses in a host, and sometimes a network representation will be more appropriate [21].

In an essay entitled “Mathematics is biology’s next microscope, only better; biology is mathematics’ next physics, only better” [53], the author poses “five biological challenges that could stimulate, and benefit from, major innovations in

Cambridge University Press
978-0-521-75596-2 - Phylogenetic Networks: Concepts, Algorithms and Applications
Daniel H. Huson, Regula Rupp and Celine Scornavacca
Frontmatter
More information

mathematics." The third challenge is: "Replace the tree of life with a network or tapestry to represent lateral transfers of heritable features such as genes, genomes, and prions."

While there is a great need for practical and reliable computational methods for inferring rooted phylogenetic networks to *explicitly* describe evolutionary scenarios involving reticulate events, generally speaking, such methods do not yet exist, or have not yet matured enough to become standard tools.

In contrast, there exist a number of established computational methods for inferring *unrooted* phylogenetic networks, which are used to *abstractly* describe reticulate evolution by providing a visualization of incompatible evolutionary pathways. Among the most widely used are methods for computing split networks [9], median networks [11] quasi-median networks [10], and other types of haplotype networks [52]. Such methods are not only used in phylogenetic analysis, but also in phylogeography and population genetics, as well.

The phylogenetic analysis of molecular sequences using phylogenetic trees is an established field and there are a number of books that describe the different approaches in detail, such as [77, 163, 217]. Population genetics is a similarly mature discipline that has been treated in a number of books, including [102, 108]. Both phylogeny and population genetics remain very active areas of research that are developing further.

Although there has been a number of book chapters published on the subject of phylogenetic networks (for example [123, 181, 182, 184, 214]), to the best of our knowledge, this is the first book that is solely dedicated to the topic.

The overall aim of our book is to give an introduction to the field of phylogenetic networks. As bioinformaticans, we sit between the mathematicians who develop theories and concepts for modeling and calculating phylogenetic networks, and the biologists who are focused on understanding the evolution of the organisms that they are interested in, using the concepts, algorithms and tools that we help to provide for them. Hence, while the content of this book is complementary to *Phylogenetics* [217] by Semple and Steel and *Inferring Phylogenies* [77] by Felsenstein, we have aimed at making the exposition in our book less mathematical than the former, while being more formal and algorithmic than the latter.

The book has three parts. In Part I, Introduction, we first describe some basic concepts, mainly from elementary graph theory. We then give introductions to sequence alignment and to phylogenetic analysis using trees. Our hope is that this will make the book a self-contained introduction both to phylogenetic trees and networks, to a degree. In the last chapter of Part I we give an introduction to phylogenetic networks, providing a high-level overview of the area. The details are then given in the remaining two parts of the book.

In Part II, Theory, we develop the theoretical underpinning first for splits and unrooted networks, and then for clusters and rooted networks. Here we attempt to develop a unified treatment of a number of different aspects of both types of networks.

In Part III, Algorithms and Applications, we systematically describe many of the existing algorithms for computing phylogenetic networks. The chapters here are organized by type of input that the algorithms work on. For many of the algorithms we briefly summarize their applications that have been reported in the literature. The last chapter gives an overview of some of the software that is available for computing phylogenetic networks from biological data.

We would like to acknowledge the advice and support that we have received from a number of colleagues. First, and foremost, we would like to thank Andreas Dress in Shanghai, who introduced D. H. to the topic in the early nineties, who has been a source of great inspiration and is the “father” of a whole generation of bio-mathematicians and bio-informaticians in Germany and abroad. Second, D. H. would like to thank Tandy Warnow in Austin, Texas, for two very valuable post-doc years working with her on phylogenetics. Third, we would like to thank Mike Steel, Pete Lockhart and David Bryant in New Zealand, with whom D. H. has worked closely on phylogenetic networks over a number of years.

We are also very grateful to the following colleagues for helpful discussions and for comments on different parts of the manuscript: Elizabeth S. Allman, Thomas Bonfert, Magnus Bordewich, David Bryant, Sydney Cameron, Tobias Dezulian, Johannes Fischer, Olivier Gascuel, Stefan Grünewald, Dan Gusfield, Jotun Hein, Mike Hendy, Leo van Iersel, Tobias Klöpper, Pete Lockhart, Bill Martin, David Morrison, Luay Nakhleh, Kay Nieselt, David Penny, Christian Rausch, John A. Rhodes, Stephan C. Schuster, Charles Semple, Yun Song, Mike Steel, Ali Tofigh, Gabriel Valiente, Detlef Weigel and Jim Whitfield.

We thank the Newton Institute of Cambridge University for hosting the *Phylogenetics* research program in 2007, which was jointly organized by Vincent Moulton, Mike Steel and D. H. The program gave us an excellent opportunity to draft and discuss a first outline of this book. We used early versions of our manuscript as the basis of two courses on *Phylogenetic Networks* at Tübingen University in 2008 and 2009 and would like to thank the participating students for their useful feedback. We are grateful to the Deutsche Forschungsgemeinschaft for financial support of our research on phylogenetic networks.

Finally, D. H. would like to thank his wife, Elke Grieswelle, and his two sons, Marlon and Moritz, for their love and support. R. R. would to like thank all her friends for their encouragement and especially her husband, Bernhard Nemec, for his love and patience. C. S. would like to thank her friends and family for their encouragement, support and affection.