Mathematics of Evolution and Phylogeny

Mathematics of Evolution and Phylogeny

Edited by

OLIVIER GASCUEL

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi Kuala Lumpur
Madrid Melbourne Mexico City Nairobi New Delhi Taipei Toronto
Shangai

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan South Korea Poland Portugal
Singapore Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2005

British Library Cataloguing in Publication Data
(Data available)

Library of Congress Cataloging in Publication Data
(Data available)

ISBN 0 19 856610 7 (Hbk)

10 9 8 7 6 5 4 3 2 1

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by

ACKNOWLEDGEMENTS

thanks to:

All the contributors, who have spent time, energy, and patience in writing and writing again their chapters, and have cross-reviewed other chapters with much care: Andrew Meade, Anne Bergeron, Bernard Moret, David Bryant, David Sankoff, Denis Bertrand, Elchanan Mossel, Jens Stoye, Jijun Tang, Julia Mixtacki, Katharina Huber, Li-San Wang, Marie-Anne Poursat, Mark Pagel, Mike Hendy, Mike Steel, Nadia El-Mabrouk, Nicolas Galtier, Olivier Elemento, Richard Desper, Susan Holmes, Tandy Warnow, Vincent Moulton, and Ziheng Yang.

A number of distinguished "anonymous" referees, whose suggestions, recommendations and corrections greatly helped to improve the content of this volume: Avner Bar-Hen, Gary Benson, Mathieu Blanchette, Emmanuel Douzery, Dan Graur, Xun Gu, Sridhar Hannenhalli, Daniel Huson, Alain Jean-Marie, Hirohisa Kishino, Bret Larget, Nicolas Lartillot, Michal Ozery-Flato, Hervé Philippe, Andrew Roger, Naruya Saitou, Ron Shamir, Edward Susko, Peter Waddell, and Louxin Zhang.

Sèverine Bérard and Denis Bertrand, the Latex specialists who have given this volume its final form, which was a real challenge regarding the extreme diversity of original manuscripts.

The people from Institut Henri Poincaré and elsewhere, who helped in organizing the "Mathematics of Evolution and Phylogeny" conference in June 2003: Etienne Gouin-Lamourette, Stéphane Guindon, Sylvie Lhermitte, and Bruno Torresani.

Olivier Gascuel
Montpellier-Montréal, June 2004

INTRODUCTION

Olivier Gascuel

The subject of this volume is evolution, which is considered at different scales: sequences, genes, gene families, organelles, genomes, and species. The focus is on the mathematical and computational tools and concepts, which form an essential basis of evolutionary studies, indicate their limitations and, inevitably, give them orientation. Recent years have witnessed rapid progress in this area, with models and methods becoming more realistic, powerful, and complex. This expansion has been driven by the phenomenal increase in genomic data available. Databases now contain tens of billions of sequence base pairs. Hundreds of species' genomes, including most notably the human genome, have been completely sequenced. This flood of data demands the development and use of formal mathematical, statistical, and computational methods. Tools derived from an evolutionary perspective are not the only ones, but they play a central part. Indeed, Nature did not explore all physical and chemical possibilities open to her. All components of life (e.g. proteins) have a specific histories, which are of a great help for understanding their functions and mechanisms. Simple comparisons are often enough to obtain deep insight into the structure, function, and role of sequences, while chemical and physical approaches (e.g. energy minimization) are more problematic and can only be applied at a late confirmatory or refinement stage. It is no accident that many of the the most widely used bioinformatics tools, for example, BLAST [2] and neighbour-joining [39], have an evolutionary basis.

Research in evolution and genetics has also been a driving force in mathematics, statistics, and computer science [41]. Recall that R.A. Fisher, the founder of so many central concepts in statistics, was primarily a geneticist. Branching processes were first seen in the field of particle physics, but were also investigated by Yule to model the speciation process [49], and recently have been the subject of much work in the field of evolution, with important results on random trees [1, 30]. The first studies of tree metrics were partly conducted from an evolutionary perspective [9, 10, 20, 50]. Later developments, generally motivated by problems in evolution, have led to fundamental results in combinatorics [3], geometry [4], and probability theory [43]. As a final example, the recent profusion of research into genome rearrangements has undoubtedly promoted a new vision and understanding of permutations of finite sets [23].

This volume follows a conference organized at Institut Henri Poincaré (Paris, June 2003). Following enthusiastic feedback from the participants, we asked the speakers to write survey chapters based on the research they had presented, with the aim of compiling a compact summary of the state-of-art mathematical

techniques and concepts currently used in the field of molecular phylogenetics and evolution. The key to the success of this conference lay in the scientific relevance and timeliness of the subjects presented (e.g. [45]), and their multidisciplinary nature.

Evolutionary patterns, processes, and history

Evolutionary studies most often have multiple aims: determining the rates and patterns of change occurring in DNA sequences, proteins, organelles or genomes, and reconstructing the evolutionary history of those entities and of organisms and species. A general goal is to infer process from pattern: the processes of organism evolution deduced from patterns of DNA or genomic variation, and processes of molecular or genomic evolution inferred from the patterns of variations in the DNA or genome itself. Given patterns observed today, the aim is then to reconstruct the history (typically a phylogenetic tree) and to understand the processes that govern evolution. Consequently, a large part of this volume is devoted to mathematical (mostly Markov) models of sequence and genome evolution. These models are used to reconstruct phylogenetic trees or networks, for example using maximum-likelihood or Bayesian approaches. The aim is not only to obtain accurate reconstructions but also to check the models' fidelity in reflecting the evolution of the sequences or genomes. Model design has therefore been thoroughly researched during recent years, both at the sequence (e.g. [21, 48]) and genome (e.g. [31, 46]) levels, with a subsequent dramatic improvement in accuracy of phylogenetic reconstruction.

Comparative and functional genomics

One of the central goals in bioinformatics is to infer the function of proteins from genomic sequences. To this end, alignment methods are nowadays the most refined and used. Sequence alignment attempts to reconstruct evolution by postulating substitution, insertion, and deletion events that occurred in the past [40]. The mutation process is described by Markov models such as the famous Dayoff [11] and JTT matrices [25]. Related or "homologous" proteins are assumed to share a common ancestor and usually have similar structure and function. We distinguish paralogous proteins (separated by one or more duplication event) from orthologous proteins (derived through speciation only) [18]. Since duplication is one of the major evolutionary processes triggering functional diversification [32], only orthologous proteins are likely to share the same function. Assessing orthology is a complicated task that requires phylogenetic analysis of an extensive set of homologous proteins [44].

When the first genomes were fully sequenced, one of the main surprises was that only about half of the proteins of an organism were considered homologous to proteins already in databases. Alignment therefore gives indications of the function of only 50% of proteins in a genome. This limit has encouraged the development of new methods that exploit the information contained within the full genomic sequence. Phylogenomic profiling [14] is one of the major

non-alignment-based methods. It is designed to infer a likely functional relationship between proteins, and is based on the assumption that proteins involved in a common metabolic pathway, or constituting a molecular complex, are likely to evolve in a correlated manner. Each protein is given a phylogenetic profile denoting the presence or absence of that protein in various genomes with a known phylogeny. Similar or complementary function can then be assigned to proteins if they have a similar phylogenetic profile. A number of other approaches have been proposed. For example, conservation of gene clusters between genomes allows the prediction of functional coupling between genes [26, 33]. Phylogenetic footprinting [5] is a method for the discovery of regulatory elements in a set of homologous regulatory regions, making use of the phylogenetic relationships among those sequences. The detection of lateral gene transfer from multi-gene or genome sequence analysis gives insight on genome adaptation [29]. These methods are examples of the pervasiveness of the feedback loops between genomic analysis and evolutionary studies, and are grouped into the new field of "phylogenomics" [13].

Tree of Life

The genomics database GenBank has information on about 100,000 species. More than 4 million species of organisms have been discovered and described, and it is estimated that tens of millions remain to be discovered. Placing these species on the Tree of Life is among the most complex and important problems facing biology [45]. Since the mid-1980s, there has been an exponential growth in the number of phylogenetic papers published each year. Recently, the Deep Green consortium achieved a first draft of the phylogeny of all green plants [7, 35]. The Tree of Life project therefore promises to be a substantial, international research program involving thousands of biologists, computer scientists, and mathematicians. The scientific aim is to understand the origins of life, the shape of its evolution, the extent of modern biodiversity, and its vulnerability to existing or possible threats. Indeed, phylogenetic analysis is playing a major role in discovering new life forms. For example, many microorganisms cannot be cultivated and studied in the laboratory, thus the principal road to discovery is to isolate their DNA from samples collected from water or soils. The DNA samples are then sequenced and identified using phylogenetic analyses based on sequences of previously described organisms. This has led to the discovery of major microbial lineages, especially in the Archaea group. Phylogenetic analysis is also of primary importance in epidemiology. Understanding how organisms, as well as their genes and gene products, are related to one another has become a powerful tool for identifying disease organisms, for tracing the history of infections, and for predicting outbreaks. Phylogenetic studies have been crucial in identifying emerging viruses such as SARS [28]. Many other examples (e.g. in agriculture) could be given to illustrate the relevance of the Tree of Life project. Most important is the fact that phylogenetic knowledge is increasing invaluable to the effort to mine, organize, and exploit the enormous amount of biological data held in numerous databases worldwide.

Biodiversity, ecology, and comparative biology

In the near future the Tree of Life should become the most natural way to represent biodiversity. With initiatives to sequence all the biota on the horizon [47], the amount of sequence data in public domain is rapidly accumulating, and it could even be that an organism's place in the Tree of Life will often be one of the few things known about it. Moreover, phylogenies provide new ways to measure biodiversity, to survey invasive species and to assess conservation priorities [27]. Notably, dated interspecies phylogenies contain information about rates and distributions of species extinctions and about the nature of radiations after previous mass extinctions [6]. Phylogenetic comparative approaches have also modelled extinction risk as a function of species' biological characteristics [36], which could be used as a basis for evaluating the status of species with unknown extinction risk. Comparative studies in biology also make an extensive use of phylogenetics when investigating adaptative traits and circumstances of adaptation [16, 24]. Indeed, species descended from a common ancestor are expected to resemble each other simply because they are related, and not necessarily because their common traits have common adaptive functions. We thus need phylogenies to infer which species are related; we need to know ancestral traits so that we can figure out what has evolved and when; and we need to know evolutionary dynamics to predict how often we should expect "chance" (i.e. non-adaptive) associations.

The goal of this volume is not to describe the numerous applications of phylogenetics and of other approaches that aim at reconstructing specific aspects of evolution. A large number of textbooks discuss the subjects rapidly surveyed above (e.g. [17, 22, 34]). Here, we concentrate on the fundamental mathematical concepts and research into current reconstruction methods. We describe a number of (probabilistic or combinatorial) models that address evolution at different scales, from segments of DNA sequences to whole genomes. We detail methods and algorithms that exploit such models for reconstructing phylogenetic trees and networks, and other mathematical techniques for various evolutionary inferences, for example, molecular dating. We explain how these reconstructions can be tested in a statistical sense and what are the inherent limits of these reconstructions. Finally, we present a number of mathematical results which give an in-depth understanding of the phylogenetic tools.

This volume is organized in fourteen chapters:

1 The minimum evolution distance-based approach of phylogenetic inference

Distance-based methods such as UPGMA [42] and neighbour joining [39] were among the first techniques used to reconstruct phylogenies. These methods are still widely used as they combine reasonable accuracy and computational speed. This chapter presents the most recent developments of distance-based methods, with a focus on the minimum evolution principle, which forms the basis of neighbour joining and other improved inference algorithms [12].

2 Likelihood calculation in molecular phylogenetics

Likelihood estimation was first introduced in molecular phylogenetics by Felsenstein [15], and is now widely used due to its accuracy and to the fact that it makes explicit the assumptions about the evolutionary model. This chapter outlines the basic probabilistic model and likelihood computation algorithm, as well as extensions to more realistic models and strategies of likelihood optimization. It surveys several of the theoretical underpinnings of the likelihood framework: statistical consistency, identifiability, effect of model misspecification, as well as advantages and limitations of likelihood ratio tests.

3 Bayesian inference in molecular phylogenetics

The Bayesian approach to phylogenetic inference was first introduced by Rannala and Yang [37], and is now widely used, thanks, in part, to the MrBayes software [38]. The main advantage of this approach is its ability to accommodate uncertainty, for example, by inferring several alternative phylogenies (instead of a single one) and estimating their respective posterior probabilities. This chapter introduces Bayesian statistics through comparison with the likelihood method. It discusses Markov chain Monte Carlo algorithms, the major modern computational methods for Bayesian inference, as well as two applications of Bayesian inference in molecular phylogenetics: estimation of species phylogenies and estimation of species divergence times.

4 Statistical approaches to test involving phylogenies

Statistical testing is an important issue in phylogenetics, for example to measure the support of a clade or to decide which evolutionary model is best. This chapter presents both the classical framework with the use of sampling distributions involving the bootstrap and permutation tests, and the Bayesian approach using posterior distributions. It contains a review of literature on parametric tests in phylogenetics and some suggestions for non-parametric tests. A number of open problems are discussed, mainly related to the non-conventional nature of tree space.

5 Mixture models in phylogenetic inference

The standard models of sequence evolution presume that sites evolve according to a common model or allow rates of evolution to vary across sites. This chapter discusses how a general class of approaches known as "mixture models" can be used to accommodate heterogeneity across sites in the patterns of sequence evolution. Mixture models fit more than one model of evolution to the data but do not require *a priori* knowledge of the evolutionary patterns across sites or of any site partitioning. The approach is illustrated on a concatenated alignment of 22 genes used to infer the phylogeny of mammals.

6 Hadamard conjugation: an analytic tool for phylogenetics

Phylogenetic inference is the process of estimating an unknown phylogeny from the evolutionary patterns that are observed in a set of aligned homologous sequences, thus inverting the mechanism which generated these patterns. For most models this inversion cannot be analysed directly. This chapter considers simple models of nucleotide substitution where this inversion is possible, thanks to "Hadamard conjugation" (or "phylogenetic spectral analysis"). Hadamard conjugation provides an analytic tool that gives insight into the general phylogenetic inference process. This chapter describes the basics of Hadamard conjugation, together with illustrations of how it can be applied to analyse a number of related concepts, such as the inconsistency of Maximum Parsimony or the determination of Maximum Likelihood points.

7 Phylogenetic networks

Phylogenetic networks are a generalization of phylogenetic trees that permit the representation of conflicting signal or alternative phylogenetic histories. Networks are clearly useful when the underlying evolutionary history is non-treelike, for example, when there has been recombination, hybridization, or lateral gene transfer. Even in cases where the underlying history is treelike, phenomena such as parallel evolution, model heterogeneity, and sampling error can make it difficult to represent the evolutionary history by a single tree, and networks can then provide a useful tool. This chapter reviews some methods for network reconstruction that are based on the representation of bipartitions or splits of the data set in question. As we shall see, these methods are based on a theoretical foundation that naturally generalizes the theory of phylogenetic trees.

8 Reconstructing the duplication history of tandemly repeated sequences

Tandemly repeated sequences can be found in all of the genomes that have been sequenced so far. However, their evolution is only beginning to be understood. In contrast to previous chapters, which study the evolution of orthologous sequences within a number of distant species, the objective in this chapter is to reconstruct the evolutionary history of paralogous sequences that are tandemly repeated within a single genome. This chapter presents a model, first proposed by Fitch [19], which assumes that duplications are caused by unequal recombination during meiosis. Duplication histories are then constrained by this model and duplication trees constitute a proper subset of phylogenetic trees. This chapter demonstrates strong biological support for this model, provides extensive mathematical and combinatorial characterizations of duplication trees, and describes various algorithms to infer tandem duplication trees from sequences.

9 Conserved segment statistics and rearrangement inferences in comparative genomics

This chapter continues the study of genome evolution, but at a much larger scale. Full genomes are compared in order to study genome rearrangements. It is shown that this field has evolved along with the biological methods for producing pertinent data, with each new type of data suggesting new questions and leading to new analyses. The development of conserved segment statistics is traced, from the mouse linkage/human chromosome assignment data analysed by Nadeau and Taylor in 1984, the comparative gene order information on organelles (late 1980s) and prokaryotes (mid-1990s), to higher eukaryote genome sequences, whose rearrangements have been recently studied without prior gene identification.

10 The reversal distance problem

Among the many genome rearrangement operations, signed inversions stand out for many biological and computational reasons. Inversions, also known as reversals, are widely identified as one of the common rearrangement operations on chromosomes, they are basic to the understanding of more complex operations such as translocations, and they offer many computational challenges. This chapter presents an elementary treatment of the problem of sorting by inversions. It describes the "anatomy" of signed permutations, gives a complete proof of the Hannenhalli–Pevzner duality theorem [23], and details efficient and simple algorithms to compute the inversion distance.

11 Genome rearrangement with gene families

The major focus of the first genome rearrangement approaches has been to infer the most economical scenario of elementary operations transforming one linear order of genes into another. Implicit in most of these studies is that each gene has exactly one copy in each genome. This hypothesis is clearly unsuitable for divergent species containing several copies of highly paralogous genes, such as multigene families. This chapter reviews the different algorithmic methods that have been developed to account for multigene families in the genome rearrangement context, in the phylogenetic context, and when reconstructing ancestral genomes.

12 Reconstructing phylogenies from gene-content and gene-order data

This chapter continues to deal with genome rearrangements, but the focus shifts to phylogenetic reconstruction from gene-content and gene-order data, whereas standard phylogeny methods exploit DNA or protein sequences. Indeed such data offer low error rates, the potential to reach further back in time, and immunity from the so-called gene-tree versus species-tree problem. This chapter surveys

the state-of-the-art techniques that use such data for phylogenetic reconstruction, focusing on recent work that has enabled the analysis of insertions, duplications, and deletions of genes, as well as inversions of gene subsequences. It concludes with a list of research questions that will need to be addressed in order to realize the full potential of this type of data.

13 Distance-based genome rearrangement phylogeny

Evolution operates on whole genomes through mutations, such as inversions, transpositions, and inverted transpositions. This chapter details a Markov model of genome evolution, assuming these three rearrangement operations. The mathematical derivation of various statistically based evolutionary distance estimators is described, and it is shown that the use of these new distance estimators with methods such as neighbour joining [39] and Weighbor [8] can result in improved reconstructions of evolutionary history.

14 How much can evolved characters tell us about the tree that generated them?

This chapter reviews some recent results that shed light on a fundamental question in molecular systematics: how much phylogenetic "signal" can we expect from extant data? Both sequence and gene-order data are examined, and evolution is modelled using Markov processes. Results presented here apply to most of the approaches discussed throughout this volume. They provide upper bounds on the probability of accurate tree reconstruction, depending on the number of species, data, and model parameters. The chapter also discusses transition phase phenomena, which make phylogenetic reconstruction impossible when substitution rates exceed a critical value.

References

[1] Aldous, D.A. (1996). Probability distributions on cladograms. In *Random Discrete Structures* (ed. D.A. Aldous and R. Pemantle), pp. 1–18. Springer-Verlag, New York.

[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

[3] Bandelt, H.-J. and Dress, A.W.M. (1992). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, **92**, 47–105.

[4] Billera, L., Holmes, S., and Vogtmann, K. (2001). The geometry of tree space. *Advances in Applied Mathematics*, **28**, 771–801.

[5] Blanchette, M., Schwikowski, B., and Tompa, M. (2002). Algorithms for phylogenetic footprinting. *Journal of Computational Biology*, **9**(2), 211–223.

[6] Bromham, L., Phillips, M.J., and Penny, D. (1999). Growing up with dinosaurs: Molecular dates and the mammalian radiation. *Trends in Ecology and Evolution*, **14**(3), 113–118.

[7] Brown, K.S. (1999). Deep Green rewrites evolutionary history of plants. *Science*, **285**(5430), 990–991.

[8] Bruno, W.J., Socci, N.D., and Halpern, A.L. (2000). Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, **17**(1), 189–197.

[9] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences* (ed. F.R. Hodson *et al.*), pp. 387–395. Edinburgh University Press, Edinburgh.

[10] Cavalli-Sforza, L.L. and Edwards, A.W. (1967). Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, **19**(3), 233–257.

[11] Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1979). A model for evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, **5** (Suppl. 3), 345–352.

[12] Desper, R. and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, **9**(5), 687–705.

[13] Eisen, J.A. and Fraser, C.M. (2003). Phylogenomics: Intersection of evolution and genomics. *Science*, **300**(5626), 1706–1707.

[14] Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. (2000). Protein function in the post-genomic era. *Nature*, **405**(6788), 823–826.

[15] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**(6), 368–376.

[16] Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, **125**, 1–12.

[17] Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

[18] Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**(2), 99–113.

[19] Fitch, W.M. (1977). Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven-amino-acid repeat in human apolipoprotein A-I. *Genetics*, **86**(3), 623–644.

[20] Fitch, W.M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, **155**(760), 279–284.

[21] Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, **18**(5), 866–873.

[22] Graur, D. and Li, W.-H. (1999). *Fundamentals of Molecular Evolution* (2nd edn). Sinauer, Sunderland, MA.

[23] Hannenhalli, S. and Pevzner, P.A. (1999). Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *Journal of ACM*, **46**(1), 1–27.

[24] Harvey, P.H. and Pagel, M.D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

[25] Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in Biosciences*, **8**(3), 275–282.

[26] Luc, N., Risler, J.L., Bergeron, A., and Raffinot, M. (2003). Gene teams: A new formalization of gene clusters for comparative genomics. *Computational Biology and Chemistry*, **27**(1), 59–67.

[27] Mace, G.M., Gittleman, J.L., and Purvis, A. (2003). Preserving the tree of life. *Science*, **300**(5626), 1707–1709.

[28] Marra, M.A. *et al.* (2003). The Genome sequence of the SARS-associated coronavirus. *Science*, **300**(5624), 1399–1404.

[29] Nelson, K.E. *et al.* (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**(6734), 323–329.

[30] McKenzie, A. and Steel, M. (2000). Distributions of cherries for two models of trees. *Mathematical Biosciences*, **164**(1), 81–92.

[31] Miklos, I. (2003). MCMC genome rearrangement. *Bioinformatics*, **19** (Suppl. 2(3)), II130–II137.

[32] Ohno, S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

[33] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences USA*, **96**(6), 2896–2901.

[34] Page, R.D.M. and Holmes, E.C. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Scientific, Oxford.

[35] Pennisi, E. (2003). Plants find their places on the tree of life. *Science*, **300**(5626), 1696.

[36] Purvis, A., Gittleman, J.L., Cowlishaw, G., and Mace, G.M. (2000). Predicting extinction risk in declining species. *Proceedings of the Royal Society of London, Series B Biological Sciences*, **267**(1456), 1947–1952.

[37] Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, **43**(3), 304–311.

[38] Ronquist, F. and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**(12), 1572–1574.

[39] Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

[40] Sankoff, D. and Kruskal, J.B. (ed.) (1999). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (2nd edn). CSLI Publications, Stanford, CA.

[41] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press, New York.

[42] Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*, pp. 230–234. W.K. Freeman and Company, San Francisco, CA.

[43] Steel, M. (1994). Recovering a tree from the leaf colourations it generates under Markov model. *Applied Mathematics Letters*, **7**, 19–23.

[44] Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. *Science*, **278**(5338), 631–637.

[45] Tree of Life (2003). *Science*, **300**(special issue)(5626).

[46] Wang, L.-S. and Warnow, T. (2001). Estimating true evolutionary distances between genomes. In *Proc. 33th Annual ACM Symposium on Theory of Computing (STOC'01)* (ed. J.S. Vitter, P. Spirakis, and M. Yannakakis), pp. 637–646. ACM Press, New York.

[47] Wilson, E.O. (2003). The encyclopedia of life. *Trends in Ecology and Evolution*, **18**(2), 77–80.

[48] Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**(1), 431–449.

[49] Yule, G.U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. *Philosophical Transactions of the Royal Society of London, Series B*, **213**, 21–87.

[50] Zaretskii, K. (1965). Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspeh Mathematicheskikh Nauk*, **20**, 90–92.

CONTENTS

LIST OF CONTRIBUTORS

Anne Bergeron
LaCIM, Université du Québec à Montréal, Canada
anne@lacim.uqam.ca

Denis Bertrand
Méthodes et algorithmes pour la bioinformatique, LIRMM
CNRS—Université de Montpellier II
France
dbertran@lirmm.fr

David Bryant
McGill Centre for Bioinformatics
Montréal, Canada
bryant@mcb.mcgill.ca

Richard Desper
National Center for Biotechnology Information NLM, NIH,
Bethesda, MD USA
desper@ncbi.nlm.nih.gov

Nadia El-Mabrouk
Département Informatique et Recherche Opérationnelle
Université de Montreal, Canada
mabrouk@iro.umontreal.ca

Olivier Elemento
Lewis-Sigler Institute for Integrative Genomics
Princeton University
NJ, USA
elemento@princeton.edu

Nicolas Galtier
UMR 5171
CNRS—Université de Montpellier II
France
galtier@univ-montp2.fr

Olivier Gascuel
Méthodes et Algorithmes pour la Bioinformatique, LIRMM
CNRS—Université de Montpellier II
France
gascuel@lirmm.fr

Michael D. Hendy
Allan Wilson Centre for Molecular Ecology and Evolution
Massey University
Palmerston North
New Zealand
m.hendy@massey.ac.nz

Susan Holmes
Statistics Department
Stanford University
USA
susan@stat.stanford.edu

Katharina T. Huber
School of Computing Sciences,
University of East Anglia,
Norwich, UK
katharina.huber@cmp.uea.ac.uk

Andrew Meade
School of Animal and Microbial Sciences
University of Reading
England
a.meade@reading.ac.uk

Julia Mixtacki
Fakultät für Mathematik
Universität Bielefeld, Germany
mixtacki@mathematik.
uni-bielefeld.de

Bernard M.E. Moret
Department of Computer Science
University of New Mexico
USA
moret@cs.unm.edu

Elchanan Mossel
Statistics
U.C. Berkeley, USA
mossel@stat.berkeley.edu

Vincent Moulton
School of Computing Sciences,
University of East Anglia,
Norwich, UK
vincent.moulton@cmp.uea.ac.uk

Mark Pagel
School of Animal and
Microbial Sciences
University of Reading
England
m.pagel@reading.ac.uk

Marie-Anne Poursat
Laboratoire de Mathématiques
Université Paris-Sud
Paris, France
Marie-Anne.Poursat@math.
u-psud.fr

David Sankoff
Department of Mathematics and
Statistics
University of Ottawa, Canada
sankoff@uottawa.ca

Mike Steel
Biomathematics Research Centre
University of Canterbury
Christchurch, New Zealand
m.steel@math.canterbury.ac.nz

Jens Stoye
Technische Fakultät
Universität Bielefeld, Germany
stoye@techfak.uni-bielefeld.de

Jijun Tang
Department of Computer Science
and Engineering
University of South Carolina, USA
jtang@cse.sc.edu

Li-San Wang
Department of Biology
University of Pennsylvania
USA
lswang@mail.med.upenn.edu

Tandy Warnow
Department of Computer Sciences
University of Texas at Austin, USA
tandy@cs.utexas.edu

Ziheng Yang
Department of Biology
University College London
London, UK
z.yang@ucl.ac.uk