Reconstructing Evolution

# Reconstructing Evolution

## New Mathematical and Computational Advances

*Edited by*

OLIVIER GASCUEL AND MIKE STEEL

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6PD

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 Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea 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 2007

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd., King's Lynn, Norfolk

ISBN 978–0–19–920822–7

1 3 5 7 9 10 8 6 4 2

# ACKNOWLEDGEMENTS

## Many 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: Elizabeth S. Allman, Cécile Ané, Michäel G. B. Blum, Alexei Drummond, Oliver Eulenstein, Gregory Ewing, Joseph Felsenstein, David Fernãndez-Baca, Stefan Grünewald, Stéphane Guindon, Luke J. Harmon, Klaas Hartmann, Stephen B. Heard, Katharina Huber, Daniel Huson, Junhyong Kim, Michelle M. McMahon, Arne Ø. Mooers, Raul Piaggio-Talice, John A. Rhodes, Allen Rodrigo, Michael J. Sanderson, Charles Semple, Dennis H. J. Wong.

A number of distinguished anonymous referees, whose suggestions, recommendations, and corrections greatly helped to improve the contents of this volume.

Dietrich Radel (Christchurch) for his tireless work in reformatting the original submissions; Megan Foster (Christchurch) for proofreading some chapters; Jessica Churchman and Alison Jones (Oxford University Press) for encouragements and helpful advice.

The people from Institut Henri Poincaré (Paris) and LIRMM (Montpellier), who helped in organizing the 'Mathematics of Evolution and Phylogeny' conference in June 2005, from which this book arose: Céline Berger, Denis Bertrand, Samuel Blanquart, Isabelle Duc, Etienne Gouin-Lamourette, Julie Hussin, Sylvie Lhermitte, Corine Mélançon.

Olivier Gascuel and Mike Steel
Montpellier–Christchurch, November 2006.

# INTRODUCTION

*Olivier Gascuel and Mike Steel*

It has become clear that the fundamentals of biology are much more complex than expected in the 1950s and 1960s following the discovery of the DNA double-strand and of the genetic code. The 'one gene, one protein, one function' hypothesis and the 'central dogma of molecular biology' have been profoundly revised and enriched. Now we know that alternative splicing [41] is frequent in eukaryotes and viruses. In this process, a single pre-messenger RNA transcribed from one gene can lead to different mature messenger RNA molecules (mRNA) and therefore to different proteins (up to tens of thousands [58]). Moreover, we understand the central role of post-RNA-translation modifications more clearly; these can extend the range of functions of a protein by attaching to it other biochemical functional groups, by changing the chemical nature of certain residues, or by modifying its sequence and/or structure. The discovery of micro RNAs [1] and interference RNAs [26] which appear to underlie the regulation of numerous biological functions have considerably augmented the repertoire of known non-coding genes. From these discoveries, it appears that one gene may correspond to a non-protein functional unit as well as to a number of proteins and biochemical functions. However, it is also clear that the gene content of an organism is only one factor, and that gene regulation could be at least as important in explaining the differences between species. For example, microarray-based studies [30] have shown that gene regulation in chimps and humans is significantly different, although their gene repertoire is almost identical. Moreover, species cannot be understood without considering their ecological environment and their interactions with other species. For example, we are just starting to explore the relationships between humans and their (bacterial and archaeal) intestinal flora, which involve numerous interactions and regulations between the host and symbiont genes [31]. These few examples show that biology is extraordinarily complex and constitutes a territory that is currently being explored more deeply and rapidly, but still has many uncharted regions.

Our vision of evolution has also changed considerably during the last few years. The mechanisms described above are likely to play an important role (e.g. alternative splicing could play a key role in the evolution of eukaryotic proteins [14, 62]). Moreover, molecular data have demonstrated that tree-like evolution as represented by Darwin (Fig. 1) is often a gross simplification of ancestry. Gene trees and species trees often differ, due to lineage sorting [18], or to lateral gene transfers [47]. Recent works have shown that gene transfers occurred (and

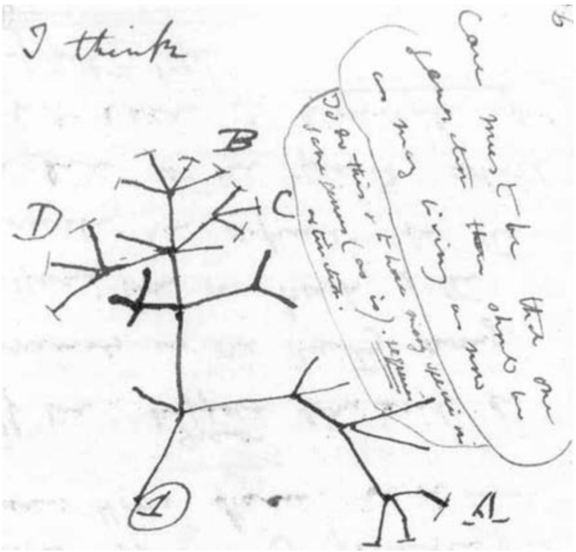
I think
B
C
D
1
A
Case must be that one generation then should be as many living as now
To do this & to have many species in same genus (as is). requires extinction.

FIG. 1. Darwin's first sketch of an evolutionary tree (1837)

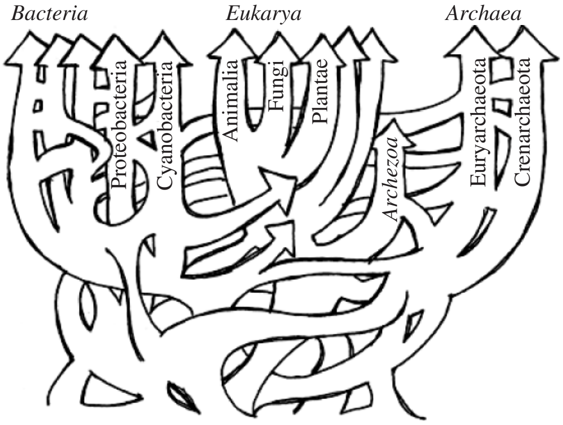

FIG. 2. Doolittle's network of life [20].

still occur) extensively in bacteria [40] and are not rare in eukaryotes [2, 6]. From Darwin's tree of Fig. 1, we are thus moving to a network view. Fig. 2 [20] is an artist's view of such a network, showing the reticulations that occurred in an organism's evolutionary history. It shows how a single species may have multiple ancestor species corresponding to its different parts. It may have one ancestor for its nuclear genome (several in case of major endosymbiotic events, e.g. in *Guillardia theta* [22] or *Plasmodium falciparum* [28]) and others for its organelles such as mitochondria and chloroplasts. In this emerging 'web of life' viruses play a pervasive role as they seem to have a central part in lateral transfer mechanisms [16, 27]. We are now at the point where the notion of species may be hard to define [21, 48], particularly for simple, primitive organisms such as bacteria and archaea, which appear as a genetic puzzle arising from multiple inheritance (transfers, hybridizations, endosymbiosis, etc.) rather than the result of a progressive and continuous evolution of a unique lineage.

Genomes have their own evolutionary dynamics and are subjected to various rearrangements (inversions, translocations, segmental or global duplications, etc., see [57] and Chapters 9 to 13 in [29]) which may be heavy even within a (relatively) short time period [23]. Relationships between these genome rearrangements and phenotypes are still unclear, and the variability of genomic configurations within any given species or along the course of time is just starting to be explored. Computer scientists have intensively investigated genome rearrangements, following the seminal works of D. Sankoff [56] and S. Hannenhali and P. Pevzner [32], but we are still at the early stages of understanding the biological implications of these rearrangements.

Genes may be seen as elementary building blocks, but sometimes they also have complex histories. They are subjected to duplications which tend to modify their function and to create new genes with new functions [49]. But genes are also subject to gene conversion, whereby multiple variable copies of a single gene become partially or fully homogenized. Genes may undergo recombinations and segmental transfers which make them mosaic-like; they are then composed of interspersed blocks of nucleotide sequence which have different evolutionary histories. Such mosaic genes are relatively frequent in bacteria [45], but have also been reported in eukaryotes [38]. All these events may have to be accounted for when reconstructing gene histories, which may be non-tree-like and resemble the scheme shown in Fig. 2. Yet most genes still seem to fit well with standard Darwinian tree scheme, although evolutionary forces are variable through time, and the structure and/or the function of the proteins may change. Detailed reconstruction of evolution thus necessitates the use of models which account for this variability, in order to be able to describe the precise history of the genes at the site level.

All life arises by evolution, via inheritance, mutation and selection. Even though evolutionary mechanisms are complex (as described above), and sometimes result in mosaic-like taxa with network-like histories, they reinforce the much cited assertion by T. Dobzhansky [19]: 'Nothing in biology makes sense, except in the light of evolution'. In particular, phylogenetics and the study of sequence evolution are fundamental for bioinformatics and the deciphering of genomes. One of the central goals in this field is to infer the function of proteins from genomic sequences. To this end, alignment methods are nowadays the most frequently used, based on the fact that homologous proteins most often have similar structure and function. To estimate (through alignment) the similarity between any sequence pair, we rely heavily on Markovian substitution models such as the famous Dayhoff [17] or JTT matrices [37]. Moreover, to obtain reliable functional predictions, we frequently distinguish between paraloguous and orthologous proteins (only the latter are likely to share the same function), which is a complex task requiring phylogenetic analysis of extensive sets of homologous proteins [59]. However, alignment typically gives functional indications for only ~50% of the proteins in a newly sequenced genome. This limit encourages the development of new methods, a number of them being based on evolutionary analyses, such as phylogenomic profiling [24], gene cluster conservation [50], and phylogenetic footprinting [7]. Another non-sequence example of the pervasiveness of evolutionary approaches, is the elucidation and analysis of regulatory networks and metabolic pathways, which has become topical with the flood of microarray gene expression data. A deeper understanding of the structure and function of regulatory networks and metabolic pathways is emerging from comparative studies, phylogenetic analysis [46] and the search for conserved motifs [5].

Phylogenetics is also central to species-level studies. Most notably, several Tree of Life projects [60] are underway worldwide, aiming to establish the phylogenetic relationships between all living species. Massive sequencing approaches such as barcoding [9] and metagenomics [61, 15, 31] are becoming mainstream to the point where an organism's place in the Tree of Life will often become one of the first things we know about it. Phylogenies are becoming a preferred way to represent and measure biodiversity, to survey invasive species, and to assess conservation priorities [42]. Notably, interspecies phylogenies with divergence dates contain information about rates and distributions of species extinctions and about the nature of radiations after previous mass extinctions [8]. Comparative approaches have also been used to model extinction risk as a function of a species' biological characteristics [52], which could then be used as a basis for evaluating the status of species with an unknown extinction risk.

Phylogenetic analysis is also fundamental to modern epidemiology. Understanding how organisms, as well as their genes and gene products, are related to one another has become a powerful tool for identifying and classifying rapidly evolving pathogens, tracing the history of infections, and predicting outbreaks. Phylogenetic studies were crucial in identifying emerging viruses such as SARS [44], and in understanding the relationships between the virulence and the genetic evolution of HIV [53] and influenza [25].

Due to recent progress [43] in sequencing technologies, genomic data continue to grow exponentially. The genomic database Genbank has information on about 265,000 species and contains over 100 billion base pairs. Moreover, a number of species have been completely sequenced, e.g. ~400 bacteria, but also 12 mammals (see Ensembl web site). Consequently, ever increasing numbers of phylogenetic studies are performed, as assessed by the citation numbers of the most famous phylogeny programs (e.g. above 14,000 for NJ and 3,000 for MrBayes, see Web of Science). However, due to the complexity of evolutionary processes, building phylogenetic trees is neither straightforward nor an end in itself, and new concepts and computational tools flourish—for example, for exploring phylogenetic networks, for studying evolution within populations, and for understanding evolution at the molecular level. This quantity of data provides us with extraordinary new possibilities to understand and reconstruct the past. For example, thanks to complete sequencing of both *Human* and *Tetraodon* (a fish), we have been able to reconstruct (in broad terms) the genome of a vertebrate ancestor [36]. As another example, the complete sequencing of *Paramecium tetraurelia* (an unicellular eukaryote) showed that most of the genes arose through at least three successive whole-genome duplications; moreover, phylogenetic analysis indicated that the most recent duplication coincides with an explosion of speciation events that gave rise to a number of sibling species [3].

But reconstructing evolution faces similar challenges to those that arise in other disciplines that deal with events that occurred in the past (e.g. astrophysics or earth history). We have no time machine, as imagined by H.G. Wells, evolution occurred just once, and there are few direct observations or experimental results on evolutionary processes. Most data are contemporary, and we rely on mathematical models to understand the past.

Pioneering work on the mathematical aspects of phylogenetics began during the 1960s and 1970s, and some of these early papers, particularly by D. Sankoff [54, 55] and P. Buneman [11, 12, 13] were enlightened predictors of the field to come in later decades. Statistical approaches, pioneered by A. Edwards and J. Felsenstein began by considering simple models of sequence site evolution. Typically these involved symmetric (and often two-state) Markov models in which each site evolves at a constant rate across the tree. This model is still studied for its mathematical properties (and it has been studied in related fields such as statistical physics and broadcasting theory). More recently, however, models have become increasingly sophisticated to account for the inherent complexity of evolution. They usually involve non-symmetric Markov processes which can vary across sites, and sometimes also across the tree (as with covarion-type processes). This has led to some debate as to what is the 'right' model for a phylogenetic study and an emerging pragmatism that there is no global model, rather each data set has its own characteristics that can suggest (and support) the most appropriate model [51].

Modelling of site substitutions has been primarily a statistical exercise, first studied within a likelihood framework, and more recently from the Bayesian (MCMC) perspective. Site substitution models also harbour a good deal of mathematical structure – for example, the Hadamard representation [33], as well as phylogenetic invariants. These invariants are algebraic identities first described in the mid 1980s, and which have been investigated with sporadic intensity ever since. Recent advances this century have stemmed from algebraic geometers and experts in commutative algebra, particularly B. Sturmfels and colleagues at UC Berkeley, together with E. Allman and J. Rhodes.

Site substitution is just one aspect of genomic evolution, and other genome rearrangement and insertion events are becoming increasingly important as phylogenetic markers. In the case of gene order, computer scientists during the 1990s devoted much effort to finding the smallest number of transformations of given types required to transform one gene sequence into another. At the same time, a group based around D. Sankoff investigated the properties of the more easily-computed breakpoint distance. In contrast to site sequence data, for gene order and for other rare genomic events, such as Short interspersed nuclear elements (SINEs), the state space is potentially very large, and this can be useful for methods that work well on data that exhibits low (or zero) homoplasy. The concept of reconstructing a tree from such compatible characters was investigated mathematically back in the 1970s and 1980s by G. Eastabrook, F. McMorris, C. Meacham, and others; it was resurrected in the early 1990s by T. Warnow and her colleagues as the 'perfect phylogeny problem' and has enjoyed further development due to the rich connection this problem has with chordal graph theory and closure operators. One recent result in this area is the theorem [34] that every fully-resolved phylogenetic tree can be uniquely specified by just *four* homoplasy-free characters, a finding that is surprising to many biologists (and some mathematicians!).

Although the reconstruction of evolutionary trees directly from character data is widespread, distance-based approaches are also popular due to their flexibility (distances can be easily computed and 'corrected'), and the computational efficiency of algorithms such as Neighbor-Joining. Mathematically, the idea of modelling distances on a tree seems to have first appeared in the 1960s in Russia after K. Zaretskii's pioneering work [63], and many of the classic results—the four-point condition, and the uniqueness of a tree representation—have since been rediscovered several times. A unified treatment was provided by A. Dress and H.-J. Bandelt in a series of papers between the late 1980s and early 1990s. One of the outcomes of their collaboration was the development of split decomposition theory [4] which provided, for the first time, a mathematically natural way to construct phylogenetic networks (rather than just trees) from distance data. This method is still used and it is implemented in the software package *SplitsTree* [35]. However the theory has also inspired more effective techniques for network reconstruction, including the now widely-used *Neighbor-Net* algorithm [10]. The turn of this century also saw mathematicians and computer scientists mount a series of attacks on the problem of reconstructing phylogenetic networks from different types of data—trees, characters, and distances. Supertree methods have also enjoyed a recent renaissance, as have methods for using phylogenetic trees to study processes of molecular evolution (such as selection and recombination), and to investigate processes of speciation and extinction.

This book aims to present these recent models, their biological relevance, their mathematical basis, their properties, and the algorithms for applying them to data. In addition, the book highlights some of the ways in which mathematics and computer science have been enriched by their interaction with evolutionary biology. These include results from the emerging field of 'phylogenetic combinatorics' which is developing a detailed theory for studying trees and networks, as well as some recent algebraic advances in the theory of phylogenetic invariants. The range of topics involves mathematics, statistics, and computer science, and in particular the subfields of combinatorics, graph theory, probability theory and Markov models, algebraic geometry, statistical inference, Monte Carlo methods, and continuous and discrete algorithms.

This book contains ten chapters, which are grouped into five main parts:

### I. Evolution within populations

The first two chapters investigate within-species evolution of gene copies, under relatively short time scales, as opposed to standard phylogenetics which considers between-species evolution of genes and much larger time periods. Chapter 1, by J. Felsenstein, shows that the coalescent trees (coalescents for short), first proposed by J. F. C. Kingmann [39], allow us to think about evolution within and between populations, and to make the connection between phylogenies and population genetic analyses. Coalescents are essential in developing methods for making inferences about populations. The chapter reviews the properties of coalescents, and the likelihood-based and Bayesian inference methods which are based on them. Chapter 2, by A. Rodrigo and co-authors, deals with rapidly evolving species, typically viruses such as HIV. Because these species are evolving so rapidly, their sequences accumulate a significant number of substitutions over short time periods (∼1% per year with HIV), and serial sampling gives us useful insights on their evolution. The chapter reviews the methods that have been developed to study these measurably evolving populations, e.g. for estimating the substitution rate and its time variations, the population size, or the migration rates.

### II. Models of sequence evolution

The mathematical and statistical properties of models that describe the evolution of aligned DNA sequences have been intensively studied since the 1970s. Indeed this branch of molecular phylogenetics is arguably the most well-developed theoretically. But many questions still remain, as does the potential for further work. Early models concentrated on simple scenarios in which site substitution was described by a basic (usually symmetric) process running at a constant rate across the sequences. Increasingly sophisticated models have allowed for more complex (and realistic) processes that may vary across the sequence and throughout the tree. In Chapter 3, O. Gascuel and S. Guindon show how standard Markov models of DNA site substitution can be further extended to handle these complexities and to detect selection, and the authors illustrate the use of these models on data sets from plants and HIV-1. In Chapter 4, E. Allman and J. Rhodes describe the current state-of-the-art in phylogenetic invariants. These fundamental algebraic identities arise within site substitution models and they are becoming useful for answering basic questions such as whether one can estimate certain parameters (including the tree) when the models become sufficiently complex. They also look promising for the future development of more efficient ways to undertake maximum-likelihood analysis or the development of new statistical approaches to phylogenetic reconstruction.

### III. Tree shape, speciation, and extinction

Phylogenetic trees relate contemporary species which have arisen from past speciation and extinction events. Depending on periods and places, evolution may be diversifying and induce high speciation levels (up to ‘explosive radiation’), or may tend towards massive extinction, as is the case today due to increasing human impact. Phylogenetic trees retain signatures of the evolutionary conditions and mechanisms that gave rise to them, and are invaluable tools to represent biodiversity. Chapter 5, by A. Mooers and co-authors, reviews a variety of models designed to represent different hypotheses about diversification processes. These models range from the simple Yule model to more complex approaches that treat species as collections of individuals rather than simple lineages. The fit of these models to real data is discussed in the light of two widely-used measures of phylogenetic tree shape, that is, tree imbalance, which measures the variation in subgroup size, and a waiting-time index based on the root-to-tip distribution of speciation events. Chapter 6, by K. Hartmann and M. Steel, discusses ‘phylogenetic diversity’ which measures the biodiversity of a set species as being the length of the phylogenetic tree connecting them. Phylogenetic diversity has been widely used for prioritising taxa for conservation and is the basis of the ‘Noah’s ark problem’ in biodiversity management. The chapter reviews some new and recent algorithmic, mathematical, and stochastic results concerning phylogenetic diversity, ranging from survival probabilities and diversity loss, to tree reconstruction.

### IV. Trees from subtrees and characters

One of the challenges faced by attempts to reconstruct a ‘Tree of Life’ is that typically one has a great deal of partial information–for example, trees for certain collections of taxa may be obtained from different groups or different data, or fundamental partitions of taxa may be made on the basis of the presence or absence of various markers. How to combine these efficiently and effectively into a phylogeny is a complicated task, involving mathematical and computational questions. In Chapter 7, M. Sanderson and colleagues describe some new approaches for studying collections of trees, going beyond the current ‘supertree’ approach. Using graph-theoretic approaches, they describe ways to extract phylogenetic signal, cluster subsets of data, and identify ‘groves’ of phylogenetic trees. In Chapter 8, S. Grünewald and K. Huber use combinatorial techniques to investigate how trees can be reconstructed from multi-state characters (and subtrees). These characters can arise in several ways–either as primary data describing how taxa are partitioned by complex genomic characters, or from existing taxonomic classifications of groups that represents different divisions of life. The results are also relevant to supertree construction where overlapping taxon sets are combined into a larger parent tree.

### V. From trees to networks

As we explained above, evolution is not always tree-like and network representations are required (see Fig. 2). Actually, there are several types of reticulation events (lateral transfer, recombination, hybridization, etc.) and even more types of phylogenetic networks. Chapter 9, by D. Huson, makes a clear distinction between the implicit network methods that aim to display (non-tree-like) phylogenetic signals, and the explicit networks aiming to model reticulate evolution. This chapter looks at split networks as a major class of implicit networks and discusses a number of approaches to produce split networks from sequences, evolutionary distances, and tree collections. This chapter also discusses explicit network methods for analysing hybridization and recombination. Chapter 10, by C. Semple, deals with the combinatorics of hybridisation networks and the problem of finding the smallest number of reticulation events that are required to explain conflicting phylogenetic signals. Here, the signals correspond to rooted phylogenetic trees—for example trees for genes collected within the species under consideration—and the chapter mostly deals with the case where we just have two conflicting trees. A number of mathematical and algorithmic properties are described, and these establish close connections between this problem, the rooted subtree prune and regraft distance, agreement forests, and recombination networks.

## References

[1] Ambros, V. (2001). MicroRNAs: Tiny regulators with great potential. *Cell*, **107**, 823–826.

[2] Andersson, J. O. (2005). Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences*, **62**(11), 1182–1197.

[3] Aury, J. M. *et al.* (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, **444**(7116), 171–178.

[4] 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.

[5] Berg, J. and Lässig, M. (2004). Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Science USA*, **101**(41), 14689–14694.

[6] Bergthorsson, U., Adams, K., Thomason, B., and Palmer, J. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, **424**, 197–201.

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

[8] 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.

[9] Brownlee, C. (2004). DNA Bar Codes: Life under the scanner. *Science News*, **166**(23), 360–361. (see also: http://phe.rockefeller.edu/barcode/)

[10] Bryant, D. and Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**(2), 255–65.

[11] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences* (ed. F. R.

Hodson, D. G. Kendall, and P. Tautu), pp.387–395. Edinburgh University Press, Edinburgh.

[12] Buneman, P. (1974*a*). A characterisation of rigid circuit graphs. *Discrete Mathematics*, **9**, 205–212.

[13] Buneman, P. (1974*b*). A note on the metric property of trees. *Journal of Combinatorial Theory, Series B*, **17**, 48–50.

[14] Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A. (2003). Evolution of the protein repertoire. *Science*, **300**(5626), 1701–1703.

[15] Daniel, R. (2005). The metagenomics of soil. *Nature Reviews Microbiology*, **3**(6), 470–478.

[16] Daubin, V. and Ochman, H. (2004). Start-up entities in the origin of new genes. *Current Opinion in Genetics & Development*, **14**(6), 616–619.

[17] Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (ed. M. Dayhoff), Volume 5, 345–352. National Biomedical Research Foundation, Washington, D. C.

[18] Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, 762–768.

[19] Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, **35**, 125–129.

[20] Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, **284**, 21246–2129.

[21] Doolittle, W. F. and Papke, R. T. (2006). Genomics and the bacterial species problem. *Genome Biology*, **7**(9), 116.

[22] Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L. T., Wu, X., Reith, M., Cavalier-Smith, T., and Maier, U. G. (2001). The highly reduced genome of an enslaved algal nucleus. *Nature*, **410**(6832), 1091–1096.

[23] Eichler, E. E. and Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**(5634), 793–797.

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

[25] Ferguson, N. M., Galvani, A. P., and Bush, R. M. (2003). Ecological and immunological determinants of influenza evolution. *Nature*, **422**(6930), 428–433.

[26] Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.

[27] Forterre, P. (2006) Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proceedings of the National Academy of Science USA*, **103**(10), 3669–3674.

[28] Gardner, M. J. *et al.* (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**(6906), 498–511.

[29] Gascuel, O. (ed) (2005). *Mathematics of Evolution & Phylogeny*, Oxford University Press, Oxford.

[30] Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P., and White K. P. (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, **440**, 242–245.

[31] Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, **312**(5778), 1355–1359.

[32] 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.

[33] Hendy, M. D. (1989). The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, **38**, 310–321.

[34] Huber, K., Moulton, V., and Steel, M. (2005). Four characters suffice to convexly define a phylogenetic tree. *SIAM Journal on Discrete Mathematics*, **18**(4), 835–843.

[35] Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254-267. Software available from www.splitstree.org.

[36] Jaillon, O. *et al.* (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**(7011), 946–957.

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

[38] Keeling, P. J. and Palmer, J. D. (2001). Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. *Proceedings of the National Academy of Science USA*, **98**(19), 10745–10750.

[39] Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, **13**, 235-248.

[40] Lerat, E., Daubin, V., Ochman, H., and Moran N. A. (2005). Evolutionary origins of genomic repertoires in bacteria. *PLoS Biology*, **3**(5), e130.

[41] Lopez, A. J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annual Review of Genetics*, **32**, 279–305.

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

[43] Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–80.

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

[45] Maynard Smith, J., Dowson, C. G., and Spratt, B. G. (1991). Localized sex in bacteria. *Nature*, **349**, 29–31.

[46] Medina, M. (2005). Genomes, phylogeny, and evolutionary systems biology. *Proceedings of the National Academy of Science USA*, **102** (Suppl. 1), 6630–6635.

[47] Ochman, H., Lawrence, J. G., and Groisman E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**(6784), 299–304.

[48] Ochman, H., Lerat, E., and Daubin, V.(2005). Examining bacterial species under the specter of gene transfer and exchange. *Proceedings of the National Academy of Science USA*, **102**(Suppl 1), 6595–6599.

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

[50] 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 Science USA*, **96**(6), 2896–2901.

[51] Posada, D. (2006). ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Research*, **34**, W700-W703.

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

[53] Ross, H. A. and Rodrigo, A. G. (2002). Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *Journal of Virology*, **76**(22), 11715–11720.

[54] Sankoff, D. (1972). Reconstructing the history and geography of an evolutionary tree, *American Mathematical Monthly*, **79**, 596-603 (Correction: American Mathematical Monthly 79, p.1100).

[55] Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, **28**, 35–42.

[56] Sankoff, D. (1992). Edit distances for genome comparison based on non-local operations. In *Proc of 3rd Conference on Combinatorial Pattern Matching (CPM'92)* (ed. A. Apostolico, M. Crochemore, Z. Galil, and U. Manber), Volume 644 in Lecture Notes in Computer Science, 121–135, Springer-Verlag, Berlin.

[57] Sankoff, D. (2003). Rearrangements and chromosomal evolution. *Current Opinion in Genetics & Development*, **13**(6), 583–587.

[58] Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., and Zipursky S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, **101**(6), 671–84.

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

[60] Tree of Life (2003). *Science*, special issue, **300**(5626), 1691–1709.

[61] Venter, J. C. *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**(5667), 66–74.

[62] Xing, Y. and Lee C. (2005). Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proceedings of the National Academy of Science USA*, **102**(38), 13526–135231.

[63] Zarestkii, K. (1965). Reconstructing a tree from the distances between its leaves. *Uspehi Mathematicheskikh Nauk*, **20**, 90–92 (in Russian).

# CONTENTS

# LIST OF CONTRIBUTORS

This looks poor, mostly due to the fact that most of addresses require more than half-line to be readable. Use a single column per page (instead of two).

**Elizabeth S. Allman**
Department of Mathematics and Statistics
University of Alaska Fairbanks, Fairbanks, AK USA
http://www.dms.uaf.edu/∼eallman
e.allman@uaf.edu

**Cécile Ané**
Department of Statistics
University of Wisconsin-Madison, USA
http://www.stat.wisc.edu/∼ane
ane@stat.wisc.edu

**Michaël G. B. Blum**
Laboratoire TIMC
Université Joseph Fourier & CNRS, Grenoble, France
http://sitemaker.umich.edu/michael.blum/home
michael.blum@imag.fr

**Alexei Drummond**
Bioinformatics Institute and Department of Computer Science
University of Auckland, New Zealand
alexei@cs.auckland.ac.nz

**Oliver Eulenstein**
Department of Computer Science
Iowa State University, USA
http://www.cs.iastate.edu/∼oeulenst
oeulenst@cs.iastate.edu

**Gregory Ewing**
Bioinformatics Institute, and
Allan Wilson Centre for Molecular Ecology and Evolution
University of Auckland, New Zealand, and

Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL), Austria
gregory.ewing@univie.ac.at

**Joseph Felsenstein**
Department of Genome Science and Department of Biology
University of Washington Seattle, Washington, U.S.A.
http://www.gs.washington.edu/faculty/felsenstein.htm
joe@gs.washington.edu

**David Fernández-Baca**
Department of Computer Science
Iowa State University, USA
http://www.cs.iastate.edu/∼fernande
fernande@cs.iastate.edu

**Olivier Gascuel**
Centre National de la Recherche Scientifique
LIRMM (CNRS-UM2), Montpellier, France
http://www.lirmm.fr/∼gascuel
gascuel@lirmm.fr

**Stefan Grünewald**
CAS-MPG Partner Institute for Computational Biology
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
http://www.picb.ac.cn
stefan@picb.ac.cn

**Stéphane Guindon**
Centre National de la Recherche Scientifique
LIRMM (CNRS-UM2), Montpellier, France
http://www.lirmm.fr/∼guindon/wordpress
guindon@lirmm.fr

**Luke J. Harmon**
Biodiversity Centre
University of British Columbia, Vancouver, Canada
http://www.zoology.ubc.ca/biodiversity/centre/harmon
harmon@zoology.ubc.ca

**Klaas Hartmann**
Biomathematics Research Centre
University of Canterbury, Christchurch, New Zealand
k.hartmann@math.canterbury.ac.nz

**Stephen B. Heard**
Department of Biology
University of New Brunswick, Fredericton, Canada
http://www.unb.ca/fredericton/science/biology/Faculty/Heard.html
sheard@unb.ca

**Katharina T. Huber**
School of Computing Sciences
University of East Anglia, United Kingdom
http://www.cmp.uea.ac.uk/people/kth
katharina.huber@cmp.uea.ac.uk

**Daniel Huson**
Center for Bioinformatics
University of Tübingen, Germany
http://www-ab.informatik.uni-tuebingen.de
huson@informatik.uni-tuebingen.de

**Junhyong Kim**
Department of Biology
University of Pennsylvania, USA
http://kim.bio.upenn.edu
junhyong@sas.upenn.edu

**Michelle M. McMahon**
Department of Plant Sciences
University of Arizona, USA
http://cals.arizona.edu/∼mcmahonm
mcmahonm@email.arizona.edu

**Arne Ø. Mooers**
Biological Sciences
Simon Fraser University, Burnaby, Canada
http://www.sfu.ca/∼amooers
amooers@sfu.ca

**Raul Piaggio-Talice**
Department of Computer Science
Iowa State University, USA
rpiaggio@iastate.edu

**John A. Rhodes**
Department of Mathematics and Statistics
University of Alaska Fairbanks, Fairbanks, AK USA
http://www.dms.uaf.edu/∼jrhodes
j.rhodes@uaf.edu

**Allen Rodrigo**
Bioinformatics Institute, and
Allan Wilson Centre for Molecular Ecology and Evolution
University of Auckland, New Zealand
a.rodrigo@auckland.ac.nz

**Michael J. Sanderson**
Department of Ecology and Evolutionary Biology
University of Arizona, USA
http://ginger.ucdavis.edu
sanderm@email.arizona.edu

**Charles Semple**
Biomathematics Research Centre
Department of Mathematics and Statistics
University of Canterbury, Christchurch, New Zealand
http://www.math.canterbury.ac.nz/∼cas83
c.semple@math.canterbury.ac.nz

**Mike Steel**
Biomathematics Research Centre
University of Canterbury, Christchurch, New Zealand
http://www.math.canterbury.ac.nz/bio
m.steel@math.canterbury.ac.nz

**Dennis H. J. Wong**
Department of Biology
University of New Brunswick, Fredericton, Canada
dhjwong@gmail.com