Mike Steel, Mathematical and Applied Phylogenetics

By Mathias Cronjager, David Emms, Luca Feretti and Jotun Hein

Review of Mike Steel (2016) PHYLOGENY. Discrete and Random Processes in Evolution

Mike Steel has contributed enormously to mathematical aspects of phylogenetics and is doing the field a major favour in summarising key results in the field. Phylogeneticists would do themselves a major favour by becoming more familiar with these results.

This is a high quality book from an author by one of the leaders in the field. It covers combinatorial properties of trees, different representations, their enumeration and metrics on tree space. The possible composition and decomposition of trees into subtrees and the compatibility of trees. Tree inference from characters and distances, stochastic models on trees and models for evolution or generation of trees are discussed. It is a mathematical book but one with clear applications. The presentation throughout is generally succinct and as such it covers a large amount in its 350 pages including many results since Phylogenetics published with Charles Semple 12 years ago. As an introductory book it is not an easy read, there can often be a number of new concepts on a single page. However, plentiful figures provide an additional, and often easier, way to understand the material. There are few numerical examples or examples taken from real applications. Such examples could have provided stronger insight into the applications as well as the concepts themselves.

There is a lot of new material on stochastic models which underlie many current biological applications. There is a final chapter on networks, but these are not covered in depth. The ancestral recombination graph is only mentioned in passing which is a significant omission since it describes the relationship between a set of sequences sampled from a population. Important algorithmic results are highlighted although the issues are rarely discussed. This seems a sensible choice for maintaining the focus of the book. The discussion of tree inference methods presents a number of important results which are of significance to current disagreements in the literature over conflicting, seemingly well-supported phylogenies. Overall, it combines being a very readable introduction to the subject while giving an overview of many important recent results and current research topics. Some parts of the book could be hard reading for the average biologist, but many in the field have a strong background in statistics or algorithms, and for them it would be well worth the effort.

Historical background

The graphical concept of a tree has been used to describe the relationship much prior to Darwin and the illustration is by Augier from 1801 (Tassy, 2011 #821), but also Linnaeus and others made similar illustrations and it is hard for us today to view these without imposing a historical
evolutionary interpretation. But after the general acceptance of the theory of evolution [Darwin, 1859 #583] the concept of phylogeny has increased in importance steadily for 150 years. The theory of evolution makes it valid to search for a “tree of life”. For the first many decades this was pursued with incomplete fossils and very noisy heterogeneous data and the inferences were correspondingly qualitative. The growth of data led to a need to be explicit about the underlying principles for tree inference or reconstruction. Willi Hennig, starting in the early 50s, attempted to unify concepts relating to trees, traits and evolutionary change. Although fossils and extant species can be summarised quantitatively, the ability to sequence first proteins in the mid 50s, and DNA starting in the late 60s, heralded a revolution in the field. Sequence data have the wonderful properties that they can be determined with very low error rate, have a very well defined mode of inheritance and can now be obtained cheaply in enormous quantities.

Of course sequence data were not central in the beginning but the researchers who saw their potential wrote papers that would subsequently become citation classics.

Sokal and Mitchener {Sokal, 1958 #587} presented the Unweighted Pair Group Method with Arithmetic Mean (UGPMA) that is still widely used, although superior methods have been invented. It was simple, fast and only used a distance matrix as input as did most methods in the following decade.

In {Fitch, 1967 #590} Fitch and Margoliash made a splash with a distance based method and an interesting application based on sequence data in the journal Science.

In the late 60s Margaret Dayhoff made the first sequence database (proteins) and investigated both distance- and character-based methods for analyzing the growing data base. Almost simultaneously, Jukes and Cantor {Jukes, 1969 #593} and Neyman {Neyman, 1971 #594} analyzed the simplest stochastic models of sequence evolution, where a character (nucleotide or amino acid) could mutate to other characters with equal rates. Motoo Kimura published a series of papers from 1979 and also his highly readable “The neutral theory of molecular evolution” {Kimura, 1983 #596} investigating more realistic models of sequence change and how to measure the strength of selection on a protein coding sequence.

Hartigan {Hartigan, 1973 #602} and Fitch {Fitch, 1971 #609} proposed methods of assigning states to the internal nodes of a tree, where the states on leaves are known; this is the core of any parsimony program today.

The rise of sequences lead to the problem of alignment, for which a series of papers were published within a few years, but the one that is famous within biology is Needleman-Wunsch {Needleman, 1970 #610} that used dynamical programming to minimize a similarity score. Sankoff and Cedergren {Sankoff, 1973 #656} realised that alignment and phylogeny was part of one larger phylogenetic alignment problem and proposed an algorithm solving this.
In 1981 Joseph Felsenstein published a simple method that allowed the extension of models from pairwise sequence to sequences at the leaves of a tree. At first this seem like a hopelessly slow method despite its conceptual advantages, but eventually this method became the core of almost all statistical methods of statistical phylogenetics. Following a paper by Bishop and Thompson (Bishop, 1986 #658), Thorne, Kishino and Felsenstein (Thorne, 1991 #657) took the step of making a stochastic version of alignment that was later generalized to statistical phylogenetic alignment by a series of authors about a decade later.

The start of the new millennium coincided with multiple genomes being determined. In 2001 the human genome, in 2003 the mouse genome and after that a virtual explosion of the available data that needed phylogenetic analysis. This had several implications. Firstly, one seems to be approaching the absolute maximal amount of data one could obtain. One will now have maximal phylogenetic knowledge of a given species but also of the tree of life as sequencing spreads through its breadth. Secondly, it led to the rise of comparative genomics where genome annotation of features such as genes, RNAs and regulatory features was strongly enhanced by simultaneous analysis of many related genomes. Thirdly, the simultaneous increase in population genomics led to interesting problems about the relationship between within-species and between-species variation.

The increase in sequence data and computational power has led to an increase in the popularity of Bayesian approaches that can investigate complex models where exact analytical results play a lesser role.

After the first decade of genome sequences, sequence data was about to become trivial and higher level data like structures and networks was getting increased attention. But then Next Generation Sequencing (NGS) methods lowered the cost of sequencing so much that it started to be used in many other areas such a viral dynamics within a patient (Arinaminpathy, 2014 #662), expression level studies (Wang, 2009 #823), pedigree inference (Sheehan, 2014 #663) and more.

The author

Mike Steel (1960-) is today the leader of mathematical phylogenetics, and a major figure within phylogenetics in general. Steel studied mathematics and was offered a PhD scholarship at Cambridge University but dropped out and couldn’t see there would ever be an academic future for him. Most strange for any of his present readers to hear. It was the rest of us that should have dropped out!! Steel was then a journalist for 3 years and wrote newspaper articles on a variety of topics (450+ articles in the period, but many was not peer reviewed). On one memorable occasion he was asked to investigate the truth of the rumour that a city had a high level of crime. The city was eager was improve its reputation and the mayor invited him to his office and assured him of the peacefulness of his city. Unfortunately, Steel was mugged as he was leaving the city. In 1985 he saw an ad for PhD scholarship at Massey University. He wrote to David Penny asking for some examples of the problems he would be working on and Penny sent him a sample. Steel solved them all and was accepted as a PhD student and finished in 1989.

Steel has contributed immensely to the field since then and below we mention a few – actually many – subtopics he has contributed to, each with an example publication:

Tree combinatorics - a range of new mathematical results that are extremely important in understanding the underlying structure of the space that describes the relationship of, for

**Metrics on Trees** - are key to characterising similarities and differences in postulated relationships in biology. [Steel, M.A. and Penny, D. (1993). Distributions of tree comparison metrics - some new results.] {Steel, 1993 #665}

**Trees, Supertrees, Maximum Agreement Trees, Quartets and Phylogenetic Coverage** - important for applications in systematic biology. Supertrees and phylogenetic coverage are of especially practical importance in making very large trees such as "The Tree of Life" (TOL). [Steel, M. and Sanderson, M.J. (2010). Characterizing phylogenetically decisive taxon coverage.] {Steel, 2010 #666}

**Invariants** - was a field that was of great interest in the years around 1990 and searched for functions that was constant on a topology and thus in principle could be used to distinguishing between topologies. [Steel, M.A. and Fu, Y.X. (1995). Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model.] {Steel, 1995 #824}

**Spectral Analysis and the Hadamard Conjugation** - is an explicit bijection between pattern probabilities and phylogenies with an evolutionary process. This research was very stimulating and allowed automated search for properties of estimators such as when parsimony fails under restriction on the phylogenies. [Steel, M.A. Hendy, M.D. and Penny, D. (1998). Reconstructing phylogenies from nucleotide pattern probabilities - a survey and some new results.] {Steel, 1998 #668}

**Parsimony and Likelihood Trees and their relationship** - Mike Steel has produced many interesting results concerning the likelihood of phylogenetic trees and the connection between likelihood estimates and parsimony solutions. [Steel, M. and Penny, D., (2000). Parsimony, likelihood and the role of models in molecular phylogenetics.] {Steel, 2000 #669}

**Identifiability of Phylogenies with different Markovian Evolutionary Processes** - Finding the correct tree/phylogeny given data is the absolutely essential question in evolutionary problems, so studying this relationship: (phylogeny, process)--(data) is fundamental and a long series of startling results were established. [Matsen, F. A., Mossel, E. and Steel, M. (2008). Mixed-up trees: The structure of phylogenetic mixtures.] {Matsen, 2008 #670}

**Data needed to Recover Trees** - very pertinent as sequence data increasingly covers the tree of life and several surprising results here. [Erdös, P.L., Steel, M.A., Székely, L.A. and Warnow, T. (1999). A few logs suffice to build (almost) all trees (Part 1)] {Erdos, 1999 #672}

**Combinatorics of Pedigrees** - in these papers reconstruction problems were addressed analogously to similar questions addressed for phylogenies, such as tree-supertree problems. [Steel, M. and Hein, J. (2006). Reconstructing pedigrees: a combinatorial perspective.] {Steel, 2006 #674}

**Statistical Alignment** - Mike Steel produced the idea that made it possible to align more than 2 sequences using a statistical model. This was only one paper, but it is the founding paper to the whole field. [Steel, M. and Hein, J. (2001). Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star tree.] {Steel, 2001 #675}
**Origin of Life Models** - this research is a major service to "Origins Research" that has been dominated by loose speculations for decades. Now there is a framework within which one can test, prove, simulate and eventually inch toward reality. [Steel, M. (2000). The emergence of a self-catalysing structure in abstract origin-of-life models.] {Steel, 2000 #676}

**Complexity Results and Algorithms** - such results are very important in formulating efficient algorithms and having realistic expectations as to what can be calculated and what not. [Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtrees.] {Steel, 1992 #677}

**Mathematical Results** - there are a series of papers that are clearly motivated by phylogenetics, but where the problem has been generalized so much that it no longer is biological. [Semple, C. and Steel, M. (2006). Unicyclic networks: compatibility and enumeration.] {Semple, 2006 #678}


**Empirical Work** - Mike Steel has occasionally been involved in data analysis (collaboratively!!) due to his many useful results. [Lockhart, P.J., Huson, D., Maier, U. Fraunholz, M. J., Peer, Y. v., Barbrook, A.C., Howe, C. and Steel M.A. (2000). How molecules evolve in Eubacteria.] {Lockhart, 2000 #682}

It is a remarkable range and at a rate of about 9 papers per year for 28 years since the first publication.

**Content of the book**

The book is very logically structured and the preface contains advice on different paths through the book and each chapter has a few exercises embedded in the text. The presentation style is very different from the earlier PHYLOGENETICS in that it gives more space to intuitive explanations and less to theorems followed by rigorous proofs, but the level of mathematics is about the same.

**Chapter 1 Phylogeny**

Chapter 1 covers the prerequisites: graphs, unrooted and rooted trees as well as some structures that will be put to use later on in the book—the intersection graph and chordal graphs. A phylogenetic tree is defined, this is the central object of interest for the book in contrast to the X-tree which was used in his earlier book, Phylogenetics. Rooted and unrooted trees are treated in parallel throughout the book as properties of one frequently have direct equivalents in the other.

When reading the prerequisites, there are clearly differences between Phylogeny and his earlier book Phylogenetics and it is fun to try to guess what these differences says about the content of the rest of the book.
**Phylogenetic X-trees**

\[ X = \{a, b, c, d, e\} \] - set of taxa/species

\[
\begin{align*}
\text{unrooted, binary} & \quad \text{unrooted} \\
\{a, b, c\} \cup \{d, e\} & \quad \{a, b, c, d\} \\
\{b, c\} \cup \{a, d, e\} & \quad \text{Hierarchies} \\
\{a, b\} & \quad \{c, d\}
\end{align*}
\]

**Chapter 2 Basic Combinatorics of Discrete Phylogenies**

Despite being only 30 pages, chapter 2 covers a lot of ground—a feat repeated frequently throughout the book! It begins with the enumeration of trees. We see Cayley’s formula for the number of trees with all nodes labelled. The formula, \( n^4(n-2) \), is simple, but deceptively so. Bifurcating trees and spanning trees are then considered in a similar manner. The results are obtained in a maddeningly compact page using formal power series that should be expanded into at least four pages.

Steel then presents two key equivalences that are central to much of the book and the subject itself: a rooted tree can be represented as a hierarchy of sets on the leaf-set and an unrooted tree as a system of splits (bipartitions) on the leaf-set. He then explores a series of alternative characterisations of hierarchies, and their equivalence to rooted trees. In a hierarchy of sets, any two sets must satisfy: one of the sets is a subset of the other, or they are disjoint. An unrooted tree is shown to be encoded by a set of quartet trees and a rooted tree by a set of rooted triples.

The set representation of phylogenetic trees leads naturally on to the concept of the refinement of trees. Loosely speaking, a phylogenetic tree may not be fully-resolved and a tree, \( T \), is a refinement of a second tree, \( T' \), if it is equivalent to \( T' \) but resolves more of the edges. A set of trees is compatible if there is a tree that is a refinement of all the trees in the set.

The Buneman Graph makes an appearance here and is revisited later in the book when multistate characters are discussed. The Buneman Graph is a graph which may be constructed from an arbitrary set of splits, which is a phylogenetic tree precisely when the splits are pairwise compatible. It will also reappear when Steel considers data that cannot be represented by a tree, but we are getting ahead of ourselves.

Three more ways to encode an unrooted tree are discussed (besides as an unrooted tree itself or as a system of splits) where the most fun was circular orderings, which is natural when you don’t have a root and can let the leaves hang down and be placed on a line. Next is a fun section on the size of neighborhoods and the diameter of tree-space with respect to rearrangement operations (NNI, SPR, TBR) on trees. One could imagine it would be real fun to sit and puzzle with and we discussed which other operations are imaginable. Then move on the metric space of trees with edges (edit operations). Finally, consensus functions are discussed. Important here are the strict consensus, the majority consensus and the Adams consensus. The Adams consensus is only
defined for rooted trees and satisfies some desirable properties not satisfied by the other consensus methods. What is more, an impossibility result is given showing that no method is possible that satisfies a more relaxed, seemingly desirable property for a consensus function!

Chapter 3 Tree Shape and Random [Discrete] Phylogenies

Chapter 3 discusses the shape of trees generated according to a number of models. First we see how to enumerate the number of phylogenetic trees using the orbiter-stabiliser theorem to handle symmetries that do not change the tree such as the symmetry of flipping the children of an inner node. The Yule-Harding model for evolving trees and uniformly sampled trees are discussed. The ‘big picture’ is presented (with a wonderful figure) showing how you can link and count trees with different kinds of labelling. Understanding the details of this was very rewarding. There are 6 classes of trees and there are 7 arrows between them, each showing a simple combinatorial conversion.

Measures of tree imbalance are defined (Colless and Sackin). Yule-Harding trees are more balanced than uniformly sampled trees and observed trees tend to be somewhere between the two. The Aldous beta-splitting model for generating trees is presented. The beta parameter varies the amount of imbalance and the Yule-Harding and uniform model trees come out as special cases of this model. An interesting question is what biologically realistic models could lead to Aldous beta-splitting trees since naïve assumptions of independence will lead to Yule-Harding. There is a very nice generalised urn model with cherry, non-cherry and pendant edges that can be used to calculate the distribution of cherries in a Yule-Harding tree showing how using a Polya urn model formulation can make certain theorems easier. We all loved this section.

Chapter 4 Pulling Trees Apart and Putting Trees Together

Chapter 4 considers how trees can be analysed and described in terms of trees restricted to subsets of labels (taxa). Conversely, how can trees be combined into a ‘super-tree’, for example, in the construction of a ‘tree of life’. The chapter discusses tree restriction to a subset of labels and its converse, the display of a tree by a tree on a superset of labels. Compatibility and consensus trees are discussed, as is the question of when a set of trees on subsets uniquely define a binary tree. Also, when are the restriction of trees to a collection of subsets decisive in distinguishing between the two trees.

Results on when a set of trees are compatible (can all be viewed as restrictions of some common larger tree) are explored. The problem is typically harder for unrooted trees, where you need to go via a display graph that has all labelled and unlabelled nodes of the subtrees with edges such that the subtrees can all be found in the display graph. After triangulation of this graph there are simple criteria using the tree width to determine if they are compatible. The somewhat confusing concept of tree width was only scantly treated earlier in the book, so the implications of these results are hard to interpret without a better explanation.

Quartets are a special class of unrooted trees that have received a lot attention. Besides whether a collection of trees is compatible, it is of interest to know if a set of compatible trees uniquely (in some sense) defines a single tree of which all trees in the collection are a restriction. A natural concept pertaining to collections of trees – excess – is defined as the difference between the number of internal edges in the full tree on all vertices and the number of edges of the subtrees in the collection. Ideally you would want each subtree to give you one internal edge in the tree on
all leaves. It is interesting to see differences and similarities to the previous Semple & Steel book. Even topics present in both books are treated from a different perspective and it is definitely worth reading both books.

The chapter continues with two pages on (rooted & unrooted) maximal agreement subtrees [MAST] and discusses some conjectures on the how small a MAST can be and what their expectation is for randomly picked binary trees. This is followed by analogous questions for the Quartet Metric. There is a practically very relevant section on the situation if you have trees on subsets of a full set. Then random coverage is discussed – the subsets of species for which you have trees are chosen randomly – when can you infer the total tree? The chapter ends with another interesting situation where you have a set of subtrees or quartets and want to reconstruct more than one tree, which very well could be the case if you have horizontal transfer or recombination. As is often the case the chapter is very compact often with several new definitions per 10 lines.

Chapter 5 Phylogenies Based on (Discrete) Characters

We are given characters (i.e. traits) on selected nodes—typically on the leaves of a tree—and the first section discusses when you can have a perfect phylogeny for these characters. A perfect phylogeny is a tree explaining the character states for the observed taxa allowing each trait to have evolved only once and no reversals to have occurred, i.e. there are r-1 mutations for a r-state character. This is easy for full binary characters, but much harder for multi-state characters or partial characters. For multi-state characters, three alternative avenues are given: i) reduce the characters to quartets and find their span (trees that can display them), ii) construct their intersection graph and see if it has a “restricted chordal completion”, or iii) create all binary splits and check if they are compatible. Approach ii) implies a restriction on the tree width of the intersection graph.

Two Binary Characters

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Perfect Phylogeny explaining x & y

A: 00
B: 10
C: 10
D: 01

x & y mutate on indicated branches

Four interesting extensions of the basic problem of having full binary characters are highlighted: i) Persistent Perfect Phylogeny (PPP) - where a single reversal is allowed in a phylogeny. ii) Partial binary characters - the character states for some leaves are unknown for some characters. iii) Perfect Haplotyping Problem (PPH) - an extension of PP to diploid individuals. iv) Incomplete Directed Perfect Phylogeny (IDPP) – a rooted tree must be found so that each particular state can evolve multiple times but no reversals are allowed, and the state is not known for all leaves.

There is a nice result presented showing an extremely clever way to reconstruct ANY tree from only four characters (each of which may however have a large number of states)! So why on earth do we sequence all these genomes? The answer is simple the moment you have seen the trick behind the surprising result. I find this section very important since many biologists have good
intuition about binary characters, but not more general characters such as for four nucleotides! They should read this section.

Some simple combinatorial questions are addressed next: how many binary trees can explain a given character perfectly? And the reverse: given a tree how many characters can you put on the leaves without forcing homoplasy?

**Parsimony** is discussed and for binary characters and connections with Menger’s theorem are drawn—namely how many characters will give a particular parsimony score (i.e. the number of edges in the tree for which the character must mutate). Then some fun connections are made with edit operations on trees which are not so surprising when you have been told them.

A very interesting section follows on **ancestral state reconstruction**. Parsimony assignment to the root is a kind of voting system with special cases such as the star tree (majority rule), balanced tree and caterpillar. Again there are some interesting results here. Majority rule is trivial. In a caterpillar tree the two outliers can determine the state at the root. In balanced trees an extreme unfairness in outcome can also be obtained. Further interesting results are presented in Chapter 9. Then finding **maximum parsimony trees** on multiple characters instead of single characters is discussed. And finally a very brief section on **super-trees** and **short encodings**.

**Chapter 6 Continuous Phylogenies and Distance-based Tree Reconstruction**

The important concept of **distances** between the objects on the leaves is considered. A central question is if it is possible to reconstruct a tree with **edge lengths** assigned to all edges, given only the distances between leaves. Key to this is the **four-point condition**: a metric has a tree representation if and only if it satisfies the four-point condition. On rooted trees the corresponding object is the **ultrametric**. The **Gromov-Farris transform** takes a tree metric and gives an ultrametric for any arbitrary rooting of the tree. **Symbolic ultrametrics** are also discussed although it is not clear what they are for. An obvious but important observation is made: If there is a perfect tree, the distance between strings of characters will be a tree metric, but the reverse might not be true. This is obvious since distance is a major reduction in data so you can clearly arrange characters that conflict, but whose effect cancels out in the distance function.

There follows a slightly discursive section on distances on genomes that should be extended to a fuller description. **Neighbour Joining (NJ)** and **Balanced Minimum Evolution (BME)** are discussed as two **distance-based tree-reconstruction methods**. The generalisation to reconstruction from **partial distances** is also covered.

There is a short section on **indexed pyramids** and **Kalmanson metrics**; a very strange generalisation of the 4-point condition and a way of representing such a distance function using a weighted sum of splits. The section on the **geometry of the space of tree-metrics** was most exciting and way too short. Here trees are embedded in the positive quadrant of a high dimensional Euclidian space and a series of interesting questions are addressed such as the connectedness of the space and the distance between two trees. Mike Steel should put in some sentences in like “[...]T-X is a compact connected simplicial complex of dimension n-4” to give the book some appeal to the average rubber boot owning botanist.

Next follows an excellent review of **phylogenetic diversity**. Ecological Diversity (Pielou, 1975 #777) and *Mathematical Ecology* (Pielou, 1977 #779) by Pielou discussed the topic but the word phylogeny was entirely absent. Everyone interested in biodiversity should read this chapter. The
first measure of phylogenetic diversity (PD) given is the total branch length of the tree relating the species set. Assume you only are allowed to keep k of the original n species. How do we find the subset of size k with largest PD? A simple greedy algorithms is presented. Then some alternative questions are presented: i) $\Delta \text{PD}(Y)$ – find Y where the PD-loss is the largest if you extinguish Y. ii) **Maximum Minimum Distance** (MMD) – find the set so that the longest shortest edge from the lost species to the surviving ones is minimised. iii) A weighted version of PD iv) PD based on splits instead of trees. Optimisation of PD and diversity indices for rooted trees are considered.

Some differences between the rooted and unrooted case are discussed and it is proven that PD is unimodular – the sum of PD on two sets is larger than the sum of PD on their union and their intersection. An interesting problem discussed is when you have some nature reservations that cost something to maintain and you have a budget. Which reservations should you keep to maximise PD while staying within budget? Then there is a subsection on decomposing the PD for a set to contributions from single species. **Fair proportion (FP)** and **Shapley value** are discussed and it is then proven that the expected Shapley value is identical to FP for the complete tree.

The two extension of PD (PD over Abelian Groups & Abstract Diversity Theory) should have either been extended or turned into footnotes since it is hard to see their relevance. Section 9.2.3 is interesting. We have extinctions that kill off species randomly and the effect on PD is considered. This leads to two special problems: **Heightened evolutionary distinctiveness** and the **Noah’s Ark Problem**. In the last section extinctions are applied to trees generated by the Yule process or the Birth-Death process.

**Chapter 7 Evolution on a Tree [I]**

In Chapter 7 processes on a phylogeny are investigated. The chapter discusses different models for the evolution of character states along a phylogeny, how the models are related, how they can be combined and under what conditions the phylogeny can be recovered from the character distributions (i.e. from sufficiently many sampled characters). A good review of **Markov chains** is given and then the theory is extended to **Markov processes on a tree**. It starts out very general, with non-homogenous Markov chains on trees where each branch has its own transition probability. In some sense this still seems overly-restrictive since the internal nodes are arbitrary relative to the process so why let them define different evolutionary regimes? If the determinant of the transition matrix is non-zero then a distance measure can be defined – **logDet** – that obeys the four point condition that we know characterises the tree, so a tree is identifiable from its leaf distributions.

There are a number of natural questions about irreversible rate matrices (Q) that are not discussed. Reversible Q are easy to characterise: give each state an equilibrium rate and each edge a rate of flow. Can one have an intuitive decomposition of the non-reversibility? There is a very interesting discussion of commuting and **non-commuting rate matrices**. Then we get into the basic models starting with the **equal input model** that for nucleotides was introduced by Felsenstein in 1981 but similar models had been used for proteins by Margeret Dayhoff around 1970. Their purpose was to get the equilibrium distribution right. Nick Goldman (published) and Bjarne Knudsen (I am not sure he published it) devised equal output models and you could make a weighted sum of equal input and equal output matrices to created intermediates, which might have been worth mentioning.
There follow some very short sections. First a theorem saying that in the Jukes-Cantor model on r states the Maximum Likelihood (ML) lengths are extreme: paths between different states are maximal [infinite] and between identical states minimal [zero]. This is interesting but is for a single state; for multiple states the ML estimates hopefully are more reasonable.

This is followed by a section on G-equivariant and group based models. G-equivariant is a new concept, there the model doesn’t change under group-permutation of the state space. The most important new models in this class are models that do not change if you swap strand. Group based means the state space is a group and the Kimura 3-parameter model is the most famous. There is a very nice overview figure relating different models. A fun observation is that the most general model here is “General Time Reversible with variable Q” that is not time reversible if the Qs don’t commute. The section on Phylogenetic Mixture Models is very important and it is well described. Here the identifiability of the underlying tree can be a problem in contrast with models where all positions evolve by the same process. The chapter ends with the Hadamard story and Felsenstein Zone which were covered a bit too lightly given how frequently they are discussed.

Chapter 8 Evolution on a Tree [II]

The chapter continues the theory and application of Markov processes on trees. Statistical methods for estimating phylogenies and ancestral character states are discussed. This leads to questions about how much data is required to reconstruct a tree.

Some preliminaries on metrics on distributions are covered then a list of attractive properties of Maximum Likelihood (ML) estimation in general and a set of identifiability definitions and results. Now comes some fun stuff about convergence of logDet, ML and Maximum Parsimony (MP). LogDet can have large variance for finite data and MP will converge with great speed towards the right result (faster than ML) provided we have parameters where it converges towards the right result! Then there are some important sections on information-theoretic results on how much data is needed to infer the correct placement of a long branch and on the reconstruction of an ancestral state. Some of these results are very surprising at first, for example, the amount of data needed to infer a tree grows as the logarithm of the number of taxa.

The section on so-called “Phylogenetic Oranges” describes phylogenies as Euclidian flakes embedded in higher dimensional space that are glued together in even lower dimensional flakes (corresponding to multifurcating phylogenies) and these have certain exact mathematical properties. A fun bit of geometry, but its implications eluded us. Similarly, the phylogenetic invariants (Algebraic Analysis) are charming but the intuition behind them is hard to see.

The Infinite Cluster Random Model is interesting and an extension of what has been discussed before under the equal input model. Going to infinite states gives you the Crow and Kimura 1964 Infinite Allele model. There are two extensions of this: The Ladder Model (investigated by Kingman among others) used to describe protein electrophoresis data and infinite site model by Kimura and Otha in 1971 used for DNA data. These clearly needed to be incorporated into Steel’s algebraic invariant dual homotopic centralising framework!

Chapter 9 Evolution of Trees
The chapter returns to the topic of the evolution of trees themselves and their relation with underlying evolutionary processes ("phylodynamics"). A section on the pure birth process - the simplest tree model - conveys much intuition and interesting paradoxes. The distribution of the number of descendants of a lineage after time $t$ is geometric with parameter $\exp(\lambda t)$, where $\lambda$ is the speciation rate. This simple result could be explained by a simple argument, while the author derives it from a more general point process construction later in the chapter. The subtle issues of conditioning on elapsed time or final number of species are then discussed. Again some interesting paradoxes appear, arising from the properties of the simple Poisson process. For example, the distribution of a random internal edge and an external edge are the same, which is surprising since the latter will continue to grow in the future before becoming the former. The reconstruction of ancestral states is discussed for models with only two states and fixed mutation rates. (The extension to a large number of states is quite relevant for applications but unfortunately not discussed.) The probability that any ancestral-state reconstruction method could be correct for large trees depends on a new relevant parameter, namely the ratio of speciation and substitution rates. In general, if the speciation rate is less than four times the mutation rate, then random guessing is as good as any method. Surprisingly, above this threshold Majority Rule is able to recover the ancestral state, while Maximum Parsimony loses information about the ancestral state faster than Majority Rule, and even above its threshold of 6 it shows poorer performances.

The more interesting case of birth-death processes is discussed. These models are widely applied since most real histories have both types of events. This leads to the interesting addition of the “reconstructed” tree or process, where only the lineages that have living descendants at time $t$ are considered. Important effects like the "pull of the present" and "push of the past" deriving from this conditioning are discussed in a few pages. More discussion would have been nice, since this is very relevant for phylodynamics and the behaviour of these processes is not always intuitive. The next section introduces the Coalescent Point Process. This is an important process since it provides a simple characterisation of birth-death processes conditioned on time and extant lineages, allowing even for time and age dependence of the parameters. The construction is clear, but there are finer connections that eluded us. The classical Kingman coalescent is briefly discussed, since it can be found in great detail in any book on population genetics. Finally, there is an interesting and important discussion on the loss of Phylogenetic Diversity and the difference between Kingman coalescent and birth-death models with respect to branch lengths.

The rest of the chapter is devoted to the relationship between gene trees and species trees. This includes some very interesting results on the Anomalous Gene Trees of Rosenberg and Degnan, where the most likely gene tree disagrees with the topology of the species tree. This is a challenging result reminiscent of the Felsenstein Zone. As in other places in the book, it seems that most of the combinatorial nitty-gritty has been left out. The Degnan-Rosenberg anomalies raise the question as to whether methods can be developed that can give the right species trees. The author discusses these issues briefly, as well as other relevant issues for recent research like concatenation and Maximum Likelihood on gene trees. Coalescent models to embed gene trees within species trees are then discussed. Lateral Gene Transfer is also considered, where the coalescent event is substituted by a species-jump and then coalescent. These problems are extremely important, since you can observe genes but are keen to make statements about the species tree, which is extremely hard to observe directly. Note that most of the chapter focuses
on Incomplete Lineage Sorting, that dominates recent research but it is not the only source of discordance between genes and species trees.

Chapter 10 Introduction to Phylogenetic Networks

The book ends with a chapter on phylogenetic networks. These phylogenetic structures include several deviations from tree-like inheritance and play an important role in prokaryotes due to the high rates of Lateral Gene Transfer, but they appear also in different contexts (e.g. hybridisation in eukaryotes). The topic is clearly worth studying. In the literature on networks there is a trend towards a high ratio of concepts to real biological use; this trend is apparent also in the large number of definitions presented in this chapter, even if the author makes clear that they are needed because of the increased complexity of phylogenetic networks compared to trees. In this respect, this chapter is a very useful guide to this complexity. Most of the results apply to binary trees or networks.

Implicit (unrooted) networks are discussed first: networks with single cycles (unicyclic networks), galled networks (that can have cycles, but no node can be a member of more than one cycle), split networks including the widely used Neighbor-net method, and median networks built from sequences and related to Maximum Parsimony.

The rest of the chapter is devoted to the more interesting case of explicit (directed) networks, which represent evolutionary histories more closely. The important difference between tree vertices (with a single ancestor) and reticulation vertices (with multiple ancestors) is discussed. For mathematical purposes, it is useful to consider subclasses of networks with bounded complexity of reticulation, which can be realised in different ways described here: level-k networks, tree-child networks, tree-sibling networks, reticulation-visible networks. For example, in tree-child networks, reticulation is limited by the fact that every internal vertex has a child that is a tree vertex. Temporal networks can be defined by extra conditions on the order of splits/reticulations, which however make sense only if all species survived and were sampled. More classes of networks with nice mathematical properties are presented (networks without redundant arcs, normal networks, regular networks). The chapter discusses the relations between these networks, as well as the larger class of tree-based network (networks obtained adding links to a tree), whose characterization is less intuitive than expected.

Finally, the relation between trees and networks is unveiled. Removing reticulations, how many trees can be displayed by a network? Can a specific tree be displayed? And conversely, is it possible to reconstruct a network from the trees displayed, or from subnetworks, distances or characters? What is the network that minimises reticulation from a set of trees? These questions find their answers here.

There are many good things in this chapter, but it still feels a pity to ignore the Ancestral Recombination Graph, that is the structure that describes the relation of a set of genomes from a population or viruses. However, much of what is explained is extremely close to the ARG, which is already discussed in other books.

The Role of Mathematical Phylogenetics within Phylogenetics
Although the present book and the earlier Phylogenetics (Semple, 2003 #822) written with Charles Semple are both excellent, they will eventually only be read by a very small minority of biologists. So why are Mike Steel’s contributions still very important for biologists?

1. Heightened the level of the whole field. There are many that have contributed to this and one could mention Charles Semple, Jeff Thorne, Ziheng Yang, Nick Goldman, Mike Hendy, Joe Felsenstein, David Sankoff, John Huelsenbeck and many many more. And this list doesn’t include researchers from population genetics that has a much longer history of mathematical modelling.

But Mike Steel’s contributions are impressive. The field of phylogenetics is today being contributed to by many computer scientists, statisticians and combinatorialists. Much of this might be invisible for the practical biologists and many contributions might remain theoretical, but researchers from these fields are better equipped to accelerate algorithms, design test and explore the structure of phylogenetics spaces.

2. Intuition about the relationship between data and the phylogenetic tree and processes. Although there is little data analysis in the book (close to none), there are many practical lessons to be learnt from it. All biologists have ideas and concepts about the relationship between data and the tree/evolutionary process that could have generated the data. If the ratio of mutations to sequence length is low, one expects the “signal” to be large and internal edges can be directly observed in data as bipartitions in columns. But how does that intuition generalize to more character states? Getting an idea about the shape of a tree as a function of the speciation and extinction process is important for interpreting the resulting phylogeny. How much data is necessary to distinguish different phylogenetic hypotheses? Can different evolutionary models lead to the same probability distribution on data? (they can if there is positional heterogeneity)

Having an intuitive grasp of these questions and their answers are important for practical biologists and is purely within mathematical phylogenetics.

Comparison to other books

Although phylogenetics is an older field, it has for the last 50 years been strongly tied to sequence data and the field of molecular evolution. Prior to this era Willi Hennig’s Basic outline of a theory of phylogenetic systematics (Hennig, 1950 #685) was the major work and motivating in its attempt to formulate rigorous principles, despite no statistics or algorithmics, for phylogenetic inference. It was expanded in 1966 into Phylogenetic Systematics (Hennig, 1966 #693). Sokal and Sneath’s Numerical Taxonomy (Sokal, 1963 #703) was both more exact and statistical. Sheila Embleton’s Statistics in historical linguistics (Embleton, 1986 #704) was clearly focused on language and distance methods but could also be applied to sequence data. The first textbook fully focused on molecular evolution and phylogenetics was the undergraduate text by Graur and Li (Li, 1991 #706), Fundamentals of Molecular Evolution, that later was expanded into W-S Li’s Molecular Evolution (Li, 1997 #709). In 2004 Felsenstein published his large Inferring Phylogenies (Felsenstein, 2004 #710) that is highly readable and also covers the history of phylogenetics. The same year Steel and Semple published Phylogenetics (Semple, 2003 #711) which is a very appealing read for mathematicians, statisticians and computer scientists. The following year Ziheng Yang published Computational Molecular Evolution (Yang, 2006 #712) that in 2013 was expanded into Molecular Evolution: A Statistical Approach (Yang, 2014 #714).
Besides these, there are a series of books with a very hands on approach, instructing readers on how to navigate existing programs such as *Molecular Systematics* (Hillis, 1996 #727) by David M. Hillis and Craig Moritz and *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (Salemi, 2009 #741) by Philippe Lemey

How should these books be prioritized by a researcher who wants to get into phylogenetics? Well, it depends on the interests and background of the researcher. If the researcher is a mathematician, statistician or computer scientist, Felsenstein (Felsenstein, 2004 #710), Yang (Yang, 2014 #714), Steel (2016) and possibly Semple & Steel (Semple, 2003 #711) would provide a sound basis. If the researcher is a biologist, Felsenstein (Felsenstein, 2004 #710), Yang (Yang, 2014 #714) and Lemey (Salemi, 2009 #741) would be good choices. Felsenstein and Yang are on both lists, since Felsenstein provides an excellent background and Yang is closer to data analysis, which after all is the motivation for phylogenetics. The reason for discounting Steel’s books for the biologist is that they are simply too mathematical. However, it would be useful if the insights from Steel’s books diffused as much as possible to the biological community.

**The field of phylogenetics in wider use**

For most of the biological community phylogenetic principles are encountered through key programs and packages that implement the methods such as Rev Bayes {Hothna, 2014 #742}, PAUP {Swofford, 1993 #745}, PHYLIP {Felsenstein, 1989 #754}, PAML (Yang, 1997 #755), RAxML {Stamatakis, 2005 #764} and BEAST {Drummond, 2007 #767}. Such programs are clearly invaluable and responsible for most published data analysis, but relying on them also has negative side effects: Firstly, they allow users to be ignorant of phylogenetic principles and, secondly, what is not implemented in these programs slips below the radar of researchers that cannot program. Many data sets will generate hypotheses that cannot be tested using existing programs and data analysis will need tailored programs.

It would be interesting to know what fraction of published analysis is done via packages and what fraction using new programs. This is most likely not known but it is almost certain that an overwhelming fraction is via the major packages. How much theory is never used? This is not a well defined question since theory can be useful even if it is not implemented.

In either case, if there is no theory there is nothing to implement.

**What is missing?**

Mike Steel covers most of what you could consider relevant but there are some important omissions. Some of these could have been included by adding 40-60 pages and some topics could not have been included without extensively altering the scope of the book.

In the former category we find the bootstrap, statistical alignment, recombination, and phylogenetic regression.

*The Bootstrap* is much used in phylogenetics after its introduction by Felsenstein (Felsenstein, 1985 #783) but its properties and interpretations are unclear to most biologists. The bootstrap is used immensely and biologists love a high bootstrap associated some grouping that defines a clade with some meaning, but the meaning and convergence properties are generally unknown. It is unfortunate that a method of such widespread use is so little understood and this runs against
the intent of making users more theory-literate. Some characterizations of its properties can be found Holmes {Holmes, 2003 #808} and Susko {Susko, 2009 #810}, but given the ubiquity of bootstrap methods in phylogenetics, further research would be help in making one of the most used methods better understood.

**An alignment** is the prerequisite for any phylogenetic analysis if it is not treated as an integral part of the analysis. Although alignment methods have been studied for more than 45 years there is still plenty to do. The mere need for alignment undermines the assumption of independence among sites. The whole sequence becomes the character with a corresponding explosion in the size of the state space. Luckily, an independence assumption about events associated with different positions allow a dynamical programming algorithm to be formulated. The existence of insertion-deletions as part of the evolutionary process is a reality that both theoreticians and users prefer to ignore. What needs to be done is quite clear and can be reliably described by looking at the history of algorithms using some optimisation criteria for defining the best alignment: efficient algorithms, insertion-deletions of several nucleotides/amino acids, heterogeneity among positions, neighbor dependence and more. Most of this has not been done and would have major practical benefits. After the pioneering model by Thorne, Kishino and Felsenstein in 1991, partly motivated by a paper by Bishop and Thompson (1986), that could only be applied to sequence pairs, Mike Steel presented the first method handling multiple sequences, albeit constrained to being related by a star tree. This was quickly followed by Holmes and Bruno {Holmes, 2001 #807} and Hein {Hein, 2000 #805}. Longer insertions were introduced in somewhat artificial model by Thorne, Kishino and Felsenstein in 1992 {Thorne, 1992 #819} and in Miklos, Lunter and Holmes a slow (n^4) method was investigated in 2004 {Miklos, 2004 #813} that could only handle sequence pairs. A series of authors, including Redelings and Suchard {Redelings, 2005 #816}, have used MCMC and a Bayesian framework for the phylogenetic statistical alignment problem.

**The Comparative Method or Phylogenetic Regression** is bound to be of increasing importance. This was again introduced by Felsenstein in 1985 {Felsenstein, 1985 #789} and followed by extension and formalisation by Pagel and Harvey {Pagel, 1989 #795}, and Grafen {Grafen, 1989 #793}. It becomes relevant when multiple data types are observed at the leaves and have to be modelled simultaneously. This kind of data is exploding presently and represent a major opportunity for biologists to test explanatory or causal hypotheses. A simple example could be genome versus expression levels and investigating the causal factors for the expression levels.

Most sequences are subject to recombination and the appropriate structure for describing their relationship is the Ancestral Recombination Graph (ARG) that gets minor mention in chapter 10. In 2014 Dan Gusfield published the almost 600 page **RECOMBINATORICS** {Gusfield, 2014 #768} but there wasn’t a single word on probability or estimation. The ARG seems like an ugly combinatorial structure from which it is hard to disentangle any underlying simplicity. It would have been of interest to see what Mike Steel would have made of this, but it is not in this book.

In the category of topics that would have seriously changed the scope of the book, we find the evolution of Complex Characters (Structures, Networks, Shapes, Phenotypes,..), selection, annotation and MCMC.

Many evolutionary characters can be encoded as real numbers, integers or sets but there are Complex Characters such as structures, networks and phenotypes that cannot easily be so encoded and they are of major interest in biology. Examples of such models would be Challis and
Schmidler [Challis, 2012 #769] for structures and Wiuf and Stumpf [Wiuf, 2006 #770] for one kind of network. Phenotypes represent a special problem since it an encoding might be a high dimensional vector that can be defined in many ways. These characters have very high dimensional state spaces and must pose new and interesting theoretical problems about, for instance, identifiability.

**Selection** is a central feature of evolution and most evolutionary models ignore the population aspect. Also, for instance, the very useful dN/dS ratio is used to measure selection but has neither selection nor fitness in its definition. One exception is Halpern and Bruno [Halpern, 1998 #771] that is widely ignored.

**Annotation** and especially comparative annotation is a tremendous success story that, since the publication of the mouse genome, is known by all molecular biologists. It was introduced by Yang in 1995 [Yang, 1995 #820], by Churchill and Felsenstein in 1996 [Felsenstein, 1996 #772] and also by Thorne, Jones and Goldman [Thorne, 1996 #773] and has since become ubiquitous in genome analysis. Hidden Markov Models are the main tool, but more general stochastic grammars are also used such as stochastic context free grammars. This area might be the part of phylogenetics of greatest practical importance today since these hidden structures are the real objects of interests in sequence analysis. Satija et al. [Satija, 2008 #818] combined annotation and statistical in the very simplest cases, but even there the computational challenges were daunting.

With the large growth in data, **Computational Statistics** and **Bayesian Statistics**, have experienced a growth in importance that is not reflected in this book. Models have grown both in complexity and in the size of investigated phylogenies making computational approaches attractive. This will only grow in the future—understanding the principles and open problems in designing methods for sampling complex spaces is a prerequisite for doing this well.

Does the book fail on some accounts? Or have biases due to Mike Steel being a mathematician? It is easy to ask for the impossibility, like listing a series of topics and computational experiments that would enlarge the book from 300 to 450 pages and would need an extra 6 months from Mike Steel and potentially some additional computational assistance. But it is a review’s obligation to be critical and ask an excellent book to be even more excellent. Since Steel does such an excellent job of extracting the essence of algorithms and mathematical results, it is a pity that certain topics have been ignored.

Getting into mathematical phylogenetics by reading this book is probably 10-20 times faster than tracking down the articles that Steel has digested for us. Thus the topics left out are seriously disadvantaged. However, the book is already 60% longer than it was supposed to be in that book series.

**What is ahead?**

In the next few decades the data will grow by enormous amounts. At the sequence level we could well have hundres of millions of full genomes spread over millions of species approaching maximal knowledge in this area. It will be supplemented by non-sequence data like structures, full mathematical models of cells of many organisms, networks, expression levels, proteomics.

Since such data will be spread over many species, phylogenetic analysis will be central. What will be the key problems in this development?
Dealing with **very large trees** will be a possibility and many existing inference methods will be insufficient and visualisation will be necessary.

Modern sequencing techniques - **Next Generation Sequencing** - brought a massive increase in the number of sequenced species. Most often, the new sequences represent full genomes. However, the quality of these sequences was often relatively low, since they contain a non-negligible amount of sequencing errors. NGS generates data that are noisy and incomplete and sophisticated statistical models are required for their analysis. Therefore, modern phylogenetic analyses are under the pressure of (1) a very large number of species, and therefore the need to infer large trees, which are computationally expensive; (2) the need to infer the species tree by integrating a large number of different (often incompatible) gene trees, which are only partially resolved; (3) errors and biases in the sequence data due to the sequencing process. The book discusses some of the issues related with (2) at the end of chapter 9, mostly in the context of concatenation and Incomplete Lineage Sorting. However, the effect of uncertainties in gene trees is still not well understood. The author does not discuss point (1) explicitly, but he is keen to present results on the algorithmic complexity of most of the algorithms cited in the book. For this reason the book can be used as a reference for the computational aspect as well.

As data grows it will be possible to distinguish more refined models and for instance **temporal non-heterogeneity** will be easier to detect. The same will be true for non-reversible processes, where for smaller data sets there has been a tendency to assume reversibility for purely computational convenience.

Sequence data is the largest data type, but others are of larger biological importance and thus there will be a drive towards **combination with other data** types.

Molecular Evolution is Population Genetics observed on a longer time scale and visa versa and **phylogenetics versus population** genetics will be increasingly studied together.

Recently, phylogenetics found a surprising series of applications of **time-sampled trees** in the study of the rapid evolution of pathogens (viruses and bacteria). The fields of "phyldynamics" (used here with a slightly different meaning to the one in the book) and "genetic epidemiology" are based on phylogenetic reconstruction from bacterial or viral sequences. New methods are being developed to reconstruct transmission chains of infections and infer epidemic dynamics from genetic data. The time since the Most Recent Common Ancestor of such sequences is often of the order of years, instead of millions of years, and data are sampled at different points in time. Strong selection pressure by the host immune system leads to unbalanced trees quite different from the one discussed in the book. Moreover, the geographical structure of the sequences plays a relevant role in many infections and influences their genealogical trees, hence the development of new methods for phylogeography. These time-sampled trees and the related methods and models present new challenges for modelling and inference, that are unfortunately left out of this book due to its focus on classical phylogenetic analysis.

**Summary**

The area of phylogenetics is of tremendous importance within the biosciences and “**PHYLOGENY: Discrete and Random Process Process**” by Mike Steel will hopefully be widely studied.
provide startling insight, an up-to-date review of the field and a solid grounding for further research. It is not an easy read but unquestionably worth the effort of serious study.

References


