Phylogeny—discrete and random processes in evolution[1] by Mike Steel, CBMS-NSF Regional Conference Series in Applied Mathematics, 89, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016. xvi+293 pp. ISBN: 978-1-611974-47-8, List Price $64.00, SIAM Member Price $44.80, Order Code: CB89

1 Phylogenetic Trees: The What and The Why

Mathematical phylogenetics is concerned primarily with the study of phylogenetic trees, a class of semi-labeled trees (defined formally below) which depict evolutionary relationships between organisms. Roughly, the branchings of the trees indicate past speciation events, while the labels assigned to the leaves correspond to the names of current species. In addition to providing an important visual representation of the history of life, phylogenetic trees play a key role in most evolutionary analyses. Hence their reconstruction using morphological or genetic data, typically from extant species, is a fundamental—and difficult—problem in computational biology. A rich mathematical theory of phylogenetic trees, as well as various generalizations, has been developed to facilitate this reconstruction and to provide a theoretical basis for downstream biological analyses. Before saying more about this, we briefly review a few basic graph-theoretic definitions and formalize the notion of a phylogenetic tree.

Recall that a *graph* $G = (V, E)$ consists of a set, V, of vertices and a set, $E \subseteq \{\{x, y\} : x, y \in V,\ x \neq y\}$, of edges. We also write $V(G)$ and $E(G)$ for the vertex and edge sets of G respectively. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are *isomorphic* if there is a bijection between the vertex sets $\Psi : V_1 \to V_2$ such that $\{u, v\} \in E_1$ exactly when $\{\Psi(u), \Psi(v)\} \in E_2$. If $e = \{u, v\} \in E$ then u and v are *adjacent* or *neighbors*, and the *degree* $d(v)$ of $v \in V$ is the number of neighbors of v. A graph H is a *subgraph* of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. We say that a subgraph H of G is *induced* if $E(H)$ contains all edges of $E(G)$ between pairs of vertices in $V(H)$. A *path* in G is a sequence of distinct vertices $v_1, \ldots, v_k$ such that, for all $i \in \{1, \ldots, k-1\}$, $\{v_i, v_{i+1}\} \in E$ is an edge. If further $\{v_k, v_1\} \in E$, then the subgraph C with vertices $V(C) = \{v_1, \ldots, v_k\}$ and edges $E(C) = \{v_1, v_2\} \cup \cdots \cup \{v_k, v_1\}$ is a *cycle*. A graph $G = (V, E)$ is *connected* if, for all $u, v \in V$, there is a path between u and v. The maximal connected subgraphs of G are called its *connected components*.

[1] 2010 MSC: Primary 05C90, 92D10, 92D15, Secondary 92E99

We will be concerned here with a particularly simple class of graphs: a *tree* $T = (V, E)$ is a connected, cycle-free graph. Finally, we come to the key definition. Throughout, X is a finite set. Think of it as the names of the species of interest.

Definition 1 (X-tree) *An* X-tree $\mathcal{T} = (T, \phi)$ *is a pair where* T *is a tree and* $\phi : X \to V$ *is a map such that* $\phi(X)$ *contains all vertices with degree at most* 2*. Two* X*-trees* $\mathcal{T}_1 = (T_1, \phi_1)$ *and* $\mathcal{T}_2 = (T_2, \phi_2)$ *are* isomorphic *if there is a graph isomorphism* Ψ *between* T_1 *and* T_2 *such that* $\phi_2 = \Psi \circ \phi_1$*, which we denote by* $\mathcal{T}_1 \cong \mathcal{T}_2$*. We also write* $V(\mathcal{T})$ *and* $E(\mathcal{T})$ *for the vertex and edge sets of* $\mathcal{T}$ *respectively.*

A vertex of degree 1 in a tree T is called a *leaf*. All other vertices of T are *interior* vertices. An edge of T is *interior* if both its end vertices are interior. Typically, one is interested in X-trees where only the leaves are labeled. These usually represent living organisms. Formally, a *phylogenetic tree* $\mathcal{T} = (T, \phi)$ on X is an X-tree whose labeling map ϕ is a bijection into the leaves of its underlying tree T. We say that $\mathcal{T}$ is *binary* if all interior vertices of T have degree 3. We denote by $B(n)$ the set of all binary phylogenetic trees where $|X| = n$. For computational reasons, it matters that the number of phylogenetic trees grows rapidly with the number of leaves n.

Theorem 2 (Counting binary phylogenetic trees) *For all* $n \geq 3$*,*

$$|B(n)| = 1 \times 3 \times \cdots \times (2n-5) \equiv (2n-5)!! \sim \frac{1}{2\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-2}.$$

(See e.g. Section 2.1 of the book under review for a proof.) The edges of a phylogenetic tree $\mathcal{T}$ are often associated with weights $\{w_e\}_{e \in E(\mathcal{T})}$ which may represent either the time elapsed or the expected amount of evolution (e.g., in number of mutations in a segment of the genome) along that edge.

As mentioned above, phylogenetic trees play a critical role in many evolutionary biology analyses. To illustrate, we briefly describe one application in biodiversity conservation. Imagine that, as a conservationist, you have a fixed budget to help shield a number of species from extinction. A natural goal might be to pick a set of species to protect that is "as diverse as possible" from an evolutionary point of view. But how to define phylogenetic diversity precisely?

The *restriction of an* X*-tree* $\mathcal{T}$ *to* $X' \subseteq X$, denoted $\mathcal{T}|X'$, is the X'-tree obtained from $\mathcal{T} = (T, \phi)$ by taking the minimal subtree of T including $\phi(X')$

and suppressing degree-two vertices not in $\phi(X')$. In particular, each edge f of $\mathcal{T}|X'$ corresponds to a set $E_{X'}(f)$ of edges of $\mathcal{T}$. If further the X-tree $\mathcal{T}$ has edge weights $\{w_e\}_{e \in E(\mathcal{T})}$, we associate to each edge f of $\mathcal{T}|X'$ a weight as follows

$$w_f = \sum_{e \in E_{X'}(f)} w_e.$$

The *phylogenetic diversity* of $X' \subseteq X$ under $\mathcal{T}$ is then defined as

$$\mathrm{PD}_{\mathcal{T}}(X') = \sum_{f \in E(\mathcal{T}|X')} w_f.$$

The biodiversity conservation problem is then to identify, for a given k, a set X' which maximizes the phylogenetic diversity $\mathrm{PD}_{\mathcal{T}}(X')$ over subsets of X of size k. Interestingly, an analysis of the properties of the phylogenetic diversity reveals that the biodiversity conservation problem can be solved exactly using a simple greedy strategy. That is, iteratively add a leaf to X' that maximizes the phylogenetic diversity of the set chosen so far, until k is reached. (See e.g. Section 6.4 of the book under review for a proof.)

2 Phylogenetic Trees: How to Reconstruct Them

The reconstruction of phylogenetic trees using data collected from extant species is a fundamental problem in evolutionary biology. Key to the development of effective reconstruction methods has been the derivation of alternate mathematical charaterizations of X-trees.

2.1 A first approach: characters and splits

Biological data can be formalized as follows. Let C be a set of *character states*. For instance, letting $C = \{0, 1\}$, the value 1 might indicate that a species can fly or has four limbs. A *character* χ on X is a function from X to C. In particular, a character is *binary* if $|C| = 2$. Typically, one might collect many such characters over the species of interest and seek to reconstruct a phylogenetic tree that is consistent with them. More formally, this leads us to the following definition.

Definition 3 (Character Convexity) *A character χ is* convex *on an X-tree $\mathcal{T} = (T, \phi)$ if there is an extension $\bar{\chi} : V(T) \to C$ such that*

1. $\bar{\chi} \circ \phi = \chi$;

2. *for each* $\alpha \in C$*, the subgraph of* T *induced by* $\{v \in V : \bar{\chi}(v) = \alpha\}$ *is connected.*

A collection of characters on X *is* compatible *if there is an* X*-tree on which all of them are convex. Finding such a tree is known as the* perfect phylogeny problem.

In words, character convexity corresponds to evolutionary innovations occurring only once in the tree of life, that is, in the absence of reverse transition, i.e. a new state arising but later reverting to its earlier state, and convergent transition, i.e. a new state occurring in two different parts of the tree.

Binary characters are closely related to the notion of split: an X*-split* $A|B$ is a nontrivial bipartition of X. To each edge e of an X-tree $\mathcal{T} = (T, \phi)$ corresponds an X-split as follows: $T \backslash e$ consists of two connected components with vertex sets V_1, V_2; $\phi^{-1}(V_1)|\phi^{-1}(V_2)$ is the X-split corresponding to e. We denote by $\Sigma(\mathcal{T})$ the collection of splits obtained from $\mathcal{T}$ in this manner.

Definition 4 (Split Compatibility) *The* X*-splits* $A_1|B_2$ *and* $A_2|B_2$ *are* compatible *if at least one of the sets* $C_1 = A_1 \cap A_2$*,* $C_2 = A_1 \cap B_2$*,* $C_3 = B_1 \cap A_2$ *and* $C_4 = B_1 \cap B_2$ *is empty.*

It is straightforward to check that the splits of an X-tree are pairwise compatible. There is also a converse:

Theorem 5 (Splits-Equivalence Theorem) *Let* Σ *be a collection of* X*-splits. It holds that* $\Sigma = \Sigma(\mathcal{T})$ *for some* X*-tree* $\mathcal{T}$ *if and only if the splits in* Σ *are pairwise compatible. Such tree is unique up to isomorphism.*

(See e.g. Section 2.4 of the book under review for more details.) The nontrivial direction can be proved via a computationally efficient algorithm for reconstructing X-trees from X-splits, known as *Tree Popping*, which iteratively adds an edge to separate the two sides of each X-split. Viewing binary characters as X-splits, the Splits-Equivalence Theorem and the Tree Popping procedure then provide an algorithmic solution to the problem of checking whether a collection of binary characters are compatible and, if so, of constructing an X-tree on which they are convex. One can then always transform the output into a binary phylogenetic tree by adding some X-splits.

Unfortunately, data is often not be compatible—indeed, reverse and convergent transitions are common in evolution. A more flexible approach is to minimize the number of convergent or reverse transitions needed to "explain the data."

Formally, let $\mathcal{T} = (T, \phi)$ be an X-tree and let $\bar{\chi}$ be an extension of a character χ. The *changing number* of $\bar{\chi}$ on $\mathcal{T}$ is

$$\mathrm{ch}(\bar{\chi}, \mathcal{T}) = |\{\{u, v\} \in E(T) \,:\, \bar{\chi}(u) \neq \bar{\chi}(v)\}|\,.$$

The *parsimony score* $\ell(\chi, \mathcal{T})$ of χ on $\mathcal{T}$ is the minimum value of $\mathrm{ch}(\bar{\chi})$ over all extensions of χ on $\mathcal{T}$ and, for a collection $\mathcal{C} = \{\chi_1, \ldots, \chi_k\}$ of characters, the parsimony score of $\mathcal{C}$ on $\mathcal{T}$ is

$$\ell(\mathcal{C}, \mathcal{T}) = \sum_{i=1}^{k} \ell(\chi_i, \mathcal{T}).$$

Finally, a *maximum parsimony tree* $\mathcal{T}^*$ for $\mathcal{C}$ minimizes $\ell(\mathcal{C}, \mathcal{T})$ over all X-trees.

Given a character χ on X and an X-tree $\mathcal{T}$, one can compute efficiently the parsimony score $\ell(\chi, \mathcal{T})$ using a technique known as *dynamic programming*. However, as Theorem 2 suggests, the space of trees is large and it turns out that constructing a maximum parsimony tree $\mathcal{T}^*$ is in fact computationally intractable. (See e.g. Section 5.3 of the book under review for more details.) Nevertheless a natural heuristic for minimizing the parsimony score of $\mathcal{C}$, which has proved useful in practice, is to perform a local search on tree space. Several notions of "local moves" on this space have been considered. A typical example is the following.

Definition 6 (Nearest-Neighbour Interchange) *Let $\mathcal{T} = (T, \phi) \in B(n)$. A nearest-neighbour interchange (NNI) operation is obtained by choosing an interior edge $e = \{u, v\} \in E(T)$ and two vertices $u_0 \neq v$, $v_0 \neq u$ adjacent respectively to u, v and interchanging the two subtrees rooted at u_0, v_0.*

Theorem 7 (Tree Space is Connected under NNI) *Let $\mathcal{T} \neq \mathcal{T}' \in B(n)$. Then $\mathcal{T}$ can be transformed into $\mathcal{T}'$ by a sequence of NNI operations.*

(See e.g. Section 2.5 of the book under review for more details.)

2.2 A second approach: metrics and quartets

Another natural reconstruction approach comes from thinking of a weighted phylogenetic tree as a metric on the set of species. For instance, by counting the number of differences in the sequence of a protein inherited from a common ancestor, one can estimate "how far apart" two species are in the tree of life. Formally, a *dissimilarity map* on X is a function $\delta : X \times X \to \mathbb{R}$ such that $\delta(x, x) = 0$ and $\delta(x, y) = \delta(y, x)$ for all $x, y \in X$.

Definition 8 (Tree metric) *Let $\mathcal{T} = (T, \phi)$ be an X-tree with edge weights $\{w_e\}_e$. For two vertices $u, v \in V(T)$, we let $\mathrm{Path}(u, v)$ be the set of edges on the unique path between u and v. The* path metric *corresponding to $(\mathcal{T}, w)$ is then defined as*

$$d_{\mathcal{T},w}(x, y) = \sum_{e \in \mathrm{Path}(\phi(x),\phi(y))} w_e, \qquad \forall x, y \in X.$$

A dissimilarity map δ is a tree metric *if there exists an X-tree $\mathcal{T} = (T, \phi)$ and a positive edge weight function $w : E \to \mathbb{R}_{++}$ such that, for all $x, y \in X$, $\delta(x, y) = d_{\mathcal{T},w}(x, y)$. We then say that $(\mathcal{T}, w)$ is a* tree representation *of δ.*

Theorem 9 (Uniqueness of tree representation) *Let δ be a tree metric on X. Up to isomorphism, there is a unique tree representation of δ.*

(See e.g. Section 6.1 of the book under review for more details.) The proof of this theorem uses an important characterization of X-trees in terms of their restriction to four-tuples of X.

Theorem 10 (Quartet theorem) *Let $\mathcal{T}_1, \mathcal{T}_2$ be X-trees. Then, $\mathcal{T}_1 \cong \mathcal{T}_2$ if and only if $\mathcal{T}_1|X' \cong \mathcal{T}_2|X'$ for all $X' \subseteq X$ with $|X'| \leq 4$.*

(See e.g. Section 4.1 of the book under review for more details.) To see the connection with Theorem 9, one needs to establish that, for all $X' = \{x, y, u, v\} \subseteq X$ of size at most four, the tree metric δ determines $\mathcal{T}|X'$. Note for instance that the expression

$$\frac{1}{2}(\delta(x, u) + \delta(y, v) - \delta(x, y) - \delta(u, v))$$

is:

- the weight of the interior edge of $\mathcal{T}|X'$, if $\{x, y\}|\{u, v\} \in \Sigma(\mathcal{T}|X')$;
- $-1\times$ the weight of the interior edge of $\mathcal{T}|X'$, if $\{x, u\}|\{y, v\} \in \Sigma(\mathcal{T}|X')$;
- 0 otherwise.

Furthermore, there is an efficient algorithm for computing the X-splits of $\mathcal{T}$ from the collection $\{\mathcal{T}|X' : X' \subseteq X, |X'| \leq 4\}$, which together with the observation above leads to a metric-based reconstruction approach for phylogenetic trees.

In fact, this "quartet perspective" also provides a useful characterization of tree metrics.

Definition 11 (Four-point condition) *A dissimilarity map δ satisifies that* four-point condition *if for all $x, y, w, z \in X$ (not necessarily distinct)*

$$\delta(w, x) + \delta(y, z) \leq \max\{\delta(w, y) + \delta(x, z), \delta(w, z) + \delta(x, y)\}.$$

Theorem 12 (Tree-Metric Theorem) *Let δ be a nonnegative dissimilarity map. Then, δ is a tree metric if and only if δ satisfies the four-point condition.*

(See e.g. Section 6.1 in the book under review for more details.) Unfortunately, similarly to the case of the split-based approaches, dissimilarities obtained from data never quite satisfy the four-point condition. However, one can establish that standard metric-based methods have a "safety radius," i.e. they return the correct phylogenetic tree $\mathcal{T}$ as long as the input dissimilarity δ is close enough to the tree metric $d_{\mathcal{T},w}$, say in ℓ_∞ norm. (See e.g. Section 6.2 in the book under review for more details.)

3 Beyond Trees

Modern molecular sequencing technology has made it possible to access full genomes from large (and rapidly increasing) numbers of species, leading to new challenges in modeling, analyzing and reconstructing the evolutionary history of life—and an abundance of new related mathematical questions. For one, genomes are naturally subdivided into smaller regions. For example, genomes are comprised of hundreds or thousands of genes that encode proteins. As it turns out, these smaller regions each have their own evolutionary history, which for a number of biological reasons, can differ from one another. Hence it is common in current phylogenetic practice to reconstruct a separate tree for each gene. Information about the overall speciation history must then be inferred from this collection of gene trees.

Here is for instance an issue that arises frequently. Some genes may have been lost in certain species lineages, while still being present in others. The question then arises whether a phylogenetic tree can be reconstructed from a collection of its restrictions to particular subsets of species. Formally, for $X' \subseteq X$, we say that X-tree $\mathcal{T}$ *displays* X'-tree $\mathcal{T}'$ if $\Sigma(\mathcal{T}') \subseteq \Sigma(\mathcal{T}|X')$. For an X-tree $\mathcal{T}$ and a collection $\chi = \{X_1, \ldots, X_k\}$ of subsets of X, define $\mathcal{T}|\chi = \{\mathcal{T}|X_1, \ldots, \mathcal{T}|X_k\}$.

Definition 13 (Decisiveness) *We say that $\chi = \{X_1, \ldots, X_k\}$ is* decisive *for a binary phylogenetic tree $\mathcal{T}$ on X if $\mathcal{T}|\chi$* defines *$\mathcal{T}$ in the sense that $\mathcal{T}$ is the*

unique phylogenetic tree that displays all phylogenetic trees in $\mathcal{T}|\chi$. Further we say that χ is decisive for unrooted phylogenies on X *if χ is decisive for every binary phylogenetic tree on X.*

The following characterization of decisiveness has been established.

Theorem 14 (Criterion for decisiveness) *A collection χ of subsets of X is decisive for unrooted phylogenies on X if and only if for every partition Π of X into four blocks, there exist elements $x_1, \ldots, x_4$, one from each of the four blocks of Π, for which $\{x_1, \ldots, x_4\} \subseteq X'$ for some $X' \in \chi$.*

(See e.g. Section 4.5 of the book under review for more details.)

In the previous example, it was assumed that all gene trees, while partially known, are consistent with a single phylogenetic tree. But that is not always the case. In the presence of hybridization for instance, separate genes may have entirely different evolutionary histories. It is then significantly more challenging to recover the speciation history. Considerable recent work in mathematical and computational phylogenetics has been dedicated to modeling mechanisms that produce incongruence, such as hybridization, lateral genetic transfer, incomplete lineage sorting or duplications and losses, as well as to establishing rigorous properties of these models that allow reconstruction and analysis. In fact, a deeper issue arises in this context: a tree may not be an appropriate representation of the more complex histories produced by these mechanisms. By relaxing some of the conditions in the definition and characterizations of X-trees above, a large number of different notions of phylogenetic networks have been obtained and are actively being investigated.

4 About the Book

Mike Steel's recent monograph *Phylogeny—discrete and random processes in evolution* provides a thorough introduction to the mathematical aspects of phylogenetics. It covers both the more elementary background as well as areas of current interest, such as those alluded to in the previous paragraph. In addition to the combinatorial perspective emphasized here, the author extensively covers probabilistic, algebraic and geometrical aspects which play a key role in modern mathematical phylogenetics. The book is in some sense a follow-up to Steel's previous 2003 monograph with Semple [SS03], which addresses some of the same basic material (often covered differently though and with less coverage of the

probabilitstic aspects than the new book) but does not include the recent developments following its publication of this fast-moving area. The new book has close to 400 references, put in context and covering a large number of sub-areas, that will form an excellent entry point to the recent literature and thusly make it an invaluable resource to those interested in research in mathematical phylogenetics. As a complement, for more background on the computational side of phylogenetics, one may also want to consult the recently published textbook by Warnow [War17].

References

[SS03] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[War17] Tandy Warnow. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge University Press, 2017.

Sebastien Roch
Department of Mathematics
University of Wisconsin–Madison
e-mail roch@math.wisc.edu