Journal of Theoretical Biology 265 (2010) 535–542

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Species, clusters and the 'Tree of life': A graph-theoretic perspective

Andreas Dress [a], Vincent Moulton [b], Mike Steel [c,*], Taoyang Wu [a,b]

[a] CAS-MPG Partner Institute for Computational Biology, Shanghai, China
[b] School of Computing Sciences, University of East Anglia, Norwich, UK
[c] Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 23 December 2009
Received in revised form
25 May 2010
Accepted 26 May 2010
Available online 2 June 2010

Keywords:
Species
Ancestry
Hierarchy
Cluster
Digraph

ABSTRACT

A hierarchical structure describing the inter-relationships of species has long been a fundamental concept in systematic biology, from Linnean classification through to the more recent quest for a 'Tree of Life'. In this paper we use an approach based on discrete mathematics to address a basic question: could one delineate this hierarchical structure in nature purely by reference to the 'genealogy' of present-day individuals, which describes how they are related with one another by ancestry through a continuous line of descent? We describe several mathematically precise ways by which one can naturally define collections of subsets of present day individuals so that these subsets are nested (and so form a tree) based purely on the directed graph that describes the ancestry of these individuals. We also explore the relationship between these and related clustering constructions.

1. Introduction

In this paper, we apply discrete mathematical arguments to study how hierarchical structures arise naturally from a very basic graph in systematic biology.

Consider the collection of all organisms that ever lived on earth—this includes not just the set X of organism alive at present, and other organisms we can directly observe (e.g. fossil specimens), but a much larger set V consisting of all organisms (or vertebrates or dicots or …) that ever lived on this planet. There is a very natural directed graph structure on V: place a directed arc from $u \in V$ to $v \in V$ if u was a 'parent' of v. Here, the word 'parent' means that u contributed directly to the genetic make-up of v; in a sexually-reproducing population, this is the usual meaning of the word (the two parents of v are the contributors of the sperm and egg), while in an asexually reproducing (haploid) population, each individual typically has one parent (e.g. the prokaryote cell whose division led to the new cell) though, occasionally, v may be regarded as having additional 'parents' beyond those described, as a result of processes such as lateral gene transfer (LGT) or other forms of reticulate evolution (e.g. a hybrid taxa).

This graph—let us call it G—can thus be regarded as a 'history of life' network, that describes how different past and present individual organisms are related to one another by ancestry Steel (2007). The graph G cannot be directly observed—we have access only to a subset X of V of 'observable' individuals along with some clues as to the gross structure of the rest of the graph gleaned from the genomic data of individuals in X, and other observable information (morphology, biochemistry, behavior, fossils, etc.). Nevertheless, the graph G is a well-defined entity, based on the premise that each organism has at least one parent, back to the earliest forms of life that existed on earth.

Such a huge graph would not be of much interest were it not for Darwinian evolution. The idea that all life traces back to one common ancestor suggests that G is a connected graph, with the lines of descent of populations that we call 'species' merging (coalescing) as we trace their ancestry, from child to parent, backward in time. Thus, rather than being an isolated set of component graphs—one for each 'species'—the graph G is more like a very large, diffuse 'tree of populations' (see Fig. 1), where the populations occasionally split when a 'speciation event' occurs, for example when a population becomes separated into two reproductively isolated groups (a process referred to as allopatric speciation), though occasionally these lineages may later intersect, for example if hybrid species arise from two lineages. At the microbial level, with extensive LGT, and occasional endosymbiotic events, this picture may appear more like a 'net of life' (Kunin et al., 2005).

The history of populations is usually represented in systematic biology as a rooted *phylogenetic tree*—that is a rooted tree where

* Corresponding author. Tel.: +64 3 3667001x7688; fax: +64 4 4642587.
E-mail addresses: andreas@picb.ac.cn (A. Dress), v.moulton@uea.ac.uk (V. Moulton), m.steel@math.canterbury.ac.nz (M. Steel), taoyang.wu@gmail.com (T. Wu).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.05.031

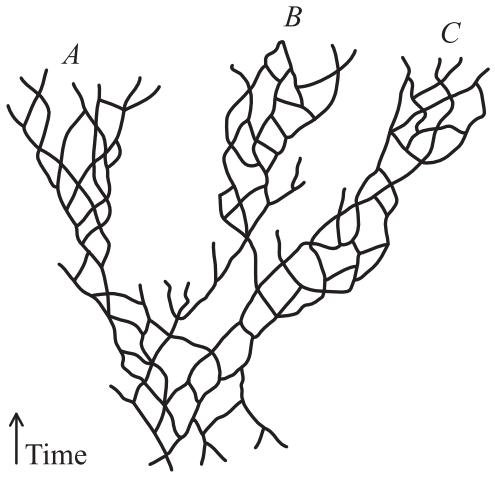

Fig. 1. A simplified picture of a history of populations. In this example *A*, *B* and *C* form tight clusters.

the leaves are labeled by the extant 'species', and which has edges and interior vertices that correspond to ancestral 'species' and 'speciation events', respectively (Semple and Steel, 2003; Felsenstein, 2004). In this representation, the fine detail of the descent of a population through time is lost, creating an unfortunate separation between phylogenetics and population genetics.

This high level picture of evolution via phylogenetic trees is problematic for two further reasons. Firstly, it requires one to address the much-debated notion of the nature and definition of 'species', a concept that is particularly ambiguous at the microorganism level (Doolittle, 1999; Wheeler and Meier, 2000). Secondly, it is increasingly being argued that processes of reticulate evolution such as LGT require that the evolution of 'species' should really be described by a network rather than a tree (Doolittle, 1999; Kunin et al., 2005; Dagnan and Martin, 2006; Lawton, 2009).

In this paper, we take a simple if somewhat novel approach to this issue by asking whether we can simply use G directly to define a tree (or tree-like structure) that reflects the bifurcating history of life studied in evolutionary theory, and which (i) does not require the prior identification or definition of 'species' and (ii) is robust to the many processes that can complicate a tree-like history, such as LGT. Viewing an evolutionary tree in this direct way is perhaps in the spirit of Darwin's suggestion to 'discover and trace the many diverging lines of descent in our natural genealogies' (Darwin, 1872). Of course, the notion that there is a hierarchical structure to the life we see today is a concept that came well before Darwinian evolution; for example, Linnean classification (Linneaus, 1735) dates back more than 100 years before Charles Darwin's *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* appeared. Moreover, the nature of 'species' has been discussed much earlier—from Plato through to the 17th Century English naturalist John Ray.

In this paper, we do not provide any general procedure for constructing hierarchies from genomic data; our interest here is purely in addressing the more fundamental questions:

- Can we construct from G systems of clusters (subsets of X) that reflect complex ancestral relationships and yet behave in a nested (tree-like) fashion?
- What are the properties of, and relationships between, different possible constructions?
- What assumptions, if any, concerning evolution are required so that the clusters derived from G are guaranteed to form a tree?

Fortunately, for this last question, we can be confident about one very helpful property: G has no directed cycles, simply because a 'parent' is always born before its child. We ask then whether any acyclic digraph G with a distinguished subset X of its vertex set induces a natural rooted tree structure on X (described in terms of a hierarchy, i.e. a system of pairwise nested or disjoint subsets of X) that reflects the process of populations splitting and separating through time. We describe several ways to define such hierarchies, and we explore their properties and the connections between them.

The use of discrete mathematics to investigate possible tree-like systems of classification arising in evolutionary biology more systematically has been explored by a number of authors from different perspectives. For example, Aldous et al. (2008) recently considered three formal ways whereby genera could be defined in terms of species, based on a phylogenetic tree, obtaining an elegant characterization of these three classifications (Theorem 1 of Aldous et al., 2008). A number of authors in the edited volume (Mirkin et al., 1997) deal with the mathematical aspects of defining hierarchies and related structures in biology. However, all these approaches to date have worked at a level that is 'higher' than G.

Our approach combines two themes developed in our earlier (independent) investigations into processes whereby trees arise by general connectivity considerations in two situations: (i) a general setting of locally connected topological spaces (Dress et al., 2009), and (ii) a particular metric space associated with ancestry within populations (Steel, 2007).

The structure of the paper is as follows. We begin by introducing some further definitions, followed by some comments concerning a purely 'genetic' variant of the graph G. We will define five general ways of obtaining a collection of subsets of X from G based on notions of ancestry. Our main result (Theorem 2) asserts that these all lead to hierarchies (or a related structure, a weak hierarchy), and describes some connections between them. In the final section, we explore some properties of these constructions further.

2. Notation

Consider a finite, directed, and cycle-free graph (i.e. an acyclic digraph) $G=(V,E\subseteq V\times V)$, with vertex set V and arc set E. Consider the associated partial order '$\preceq$' $=$ '$\preceq_G$' of V defined, for all $u,v\in V$, by $u\preceq v$ if and only if there exists a (directed) path from u to v in G, i.e., a sequence $u_0:=u,u_1,\ldots,u_k:=v$ of some length $k\geq 0$ of elements in V with $(u_{i-1},u_i)\in E$ for all $i=1,\ldots,k$ in which case u will also be called an *ancestor* of v, and v a *descendant* of u. Note that we also write $u\prec v$ in the case where $u\preceq v$ and $u\neq v$ holds.

We will sometimes refer to the elements of V as *individuals* and, given any arc (u,v) in E, the individual u will be called a *parent* of v and v will be a *child* of u. Clearly, given any two elements u,v in V, we have $(u,v)\in E$ if and only if $|\{w\in V: u\preceq w\preceq v\}|=2$ holds.

Let X denote a distinguished subset of V, which we will regard as a set of 'observable individuals' in G (e.g. present-day individuals, and perhaps some fossil specimens). We will refer to a subset of X as a *cluster*. While no specific conditions need to be placed on X in what follows, it may be natural to assume that every v in $V-X$ has a descendant in X (implying in particular that X contains all elements $v\in V$ that do not (yet) have any children), as eliminating all elements from $V-X$ that do not have a

descendant in X will not change the clusters in X we are going to consider below.

For any $v \in V$, let $\overrightarrow{v} = \overrightarrow{v_X}$ denote the set of individuals in X that are descendants of v, and for any subset U of V, put $\overrightarrow{U} := \bigcup_{v \in U} \overrightarrow{v}$.

2.1. Organismal history versus genetic history

The graph G we have defined in the introduction describes the detailed genealogical history of individual organisms. However it may also be of interest to consider a subgraph of this graph that reflects just those lines of descent that carry genetic material that survives in at least one of the organisms in our observed set X. Clearly it is possible for an individual organism that lived long ago in a diploid population to have many descendants today, and yet have no surviving genetic material (gene, homologous nucleotide, etc) today due to the processes of population genetics (a gene is inherited from one parent, not both). This distinction between genetic ancestry and organismic ancestry has been noted by many authors over the years, and has been discussed recently by Baum (2009) and more theoretically, by Matsen and Evans (2008).

We can formalize this distinction as follows: let us say that an arc (u,v) of G is (genetically) *trivial* if none of the genome of v that is inherited from u is present in any of the descendants of v in X. Let G_g be the graph obtained from G by deleting all the genetically trivial arcs. Thus, in G_g we only retain those parent-child arcs for which the child inherits from that parent genetic material that survives in at least one of the observed individuals.

Many of our results (including our main result, Theorem 2) remain true for both types of graphs, since they are stated in the generality of a finite, directed, cycle-free graph that contains X within its vertex set, and G_g clearly inherits these properties from G. However, some examples (e.g. the example of a tight cluster involving humans), and some discussion depends more crucially on which type of graph we are considering, and so, for the sake of simplicity, we will regard G as the genealogical rather than ancestral genetic graph from now on.

2.2. Hierarchies and weak hierarchies

We say that a collection $\mathcal{H}$ of subsets of X forms a (generalized) *hierarchy* on X if $\mathcal{H}$ satisfies the nesting property:

$$A,B \in \mathcal{H} \Rightarrow A \cap B \in \{\emptyset, A, B\}.$$

Note that this condition is also referred to in the hypergraph literature as a *laminar family*, and the word 'hierarchy' often also requires further conditions such as $X \in \mathcal{H}, \emptyset \notin \mathcal{H}$, or $\{x\} \in \mathcal{H}$ for all $x \in X$. Here, however, we will insist on the nesting property, only. Hierarchies are closely related to rooted X-forests and rooted X-trees, whose definition we briefly recall here (for further details, see Semple and Steel, 2003). A *rooted X-forest* is a finite, acyclic directed graph, together with a labeling map from X to the vertices of the graph for which (i) each vertex has in-degree at most 1 and (ii) each vertex having out-degree less than two is the image of the labeling map. A rooted X-tree is rooted X-forest with exactly one vertex of in-degree 0, also called the *root* of that tree.

A natural bijection exists between (isomorphism classes of) rooted X-forests and hierarchies on X that do not contain the empty set (see, for example, Edmonds and Giles, 1977, Section 8) which restricts to a bijection between the set of (isomorphism classes of) rooted X-trees and the set of hierarchies on X that contain X but not $\emptyset$. Note also that if $\mathcal{H}$ is a hierarchy on X, then so is any subset of $\mathcal{H}$, and also that, for any set $Y \subset X$, the collection $\{A \cap Y : A \in \mathcal{H}\}$ is a hierarchy on Y. Given any collection $\mathcal{P}$ of subsets of X, there is a simple way to define an associated hierarchy $\mathcal{H}_\mathcal{P}$ by setting:

$$\mathcal{H}_\mathcal{P} := \{C \in \mathcal{P} : \forall C' \in \mathcal{P}, C \cap C' \in \{C, C', \emptyset\}\}. \quad (1)$$

A weaker condition than that satisfied by a hierarchy is the condition:

$$A,B,C \in \mathcal{H} \Rightarrow A \cap B \cap C \in \{A \cap B, B \cap C, A \cap C\}.$$

If $\mathcal{H}$ satisfies this condition, it is said to form a *weak hierarchy*. One way to generate a weak hierarchy is to take the union of any two hierarchies (though not all weak hierarchies arise in this way); moreover, often (but not necessarily) the union of several hierarchies can form a weak hierarchy. In this way, weak hierarchies have provided a useful computational tool for representing reticulate evolution where conflicting tree-like histories are present in data (Dress et al., 1996). From a mathematical perspective, weak hierarchies share some of properties with 'proper' hierarchies. For example, clusters from $\mathcal{H}$ can be identified using at most two elements from X. This in turn implies that a weak hierarchy $\mathcal{H}$ can never be very large—we have: $|\mathcal{H}| \leq \binom{|X|+1}{2}$ for any weak hierarchy $\mathcal{H}$ that does not contain the empty set. While this upper bound is larger than the bound $2|X|$ that applies to hierarchies, it is still much less than the total number of subsets of X.

2.3. Connectivity through evolution

Evolution suggests that all organisms we can observe today descended from a small group of common ancestors and this suggests that the graph G is connected in various possible ways. These are summarized by the following, increasingly liberal connectivity requirements:

(C1) G contains a vertex v with $\overrightarrow{v} = X$.
(C2) For all $x,y \in X$, there exists $v \in V$ with $v \preceq x,y$.
(C3) The (undirected) graph $\Gamma_G(X) := (X, \{\{x,y\} \in \binom{X}{2} : \exists v \in V : x,y \in \overrightarrow{v}\})$ is connected (i.e. any two elements of X can be connected by a path in this graph).

In the biological context, Condition (C1) is merely the statement that all living organisms today have (at least) one common ancestor some time in the past. Condition (C2) says that every pair of individuals in X has a common ancestor, while Condition (C3) says any two individuals in X are related through a chain of relatives in X. Mathematically, (C2) implies that $\Gamma_G(X)$ is a complete graph (i.e. all possible edges are present); moreover, we have (C1) $\Rightarrow$ (C2) $\Rightarrow$ (C3). Although (C1) is usually held to be biologically reasonable (Crick, 1968; Futuyma, 1998; Woese, 2000; Sober and Steel, 2002; Theobald, 2010), we do not necessarily assume this condition here; the choice of any particular condition (C1)–(C3) is relevant only for two reasons: (i) It can determine whether or not X is an element of some of the hierarchies we construct and (ii) Condition (C2) can be helpful to ensure the existence of clusters defined by pairwise ancestral relationships. Note that the graph $\Gamma_G(X)$ in Condition (C3) has undirected edges, in contrast to G which has (directed) arcs.

3. X-Clusters from G

We now describe a variety of ways whereby an acyclic digraph G with $X \subseteq V$ can naturally give rise to specific collections of subsets of X based on concepts of ancestry. In Section 4, we will show how these constructions lead to (weak) hierarchies.

3.1. Tight clusters

We begin with an intuitively simple way to generate clusters on X from any acyclic digraph $G=(V,E)$ with $X \subseteq V$. Although the conditions a cluster must satisfy in this first definition are more severe than those we consider later, we will describe in the remark below how results in population genetics provide some justification for the existence of such tightly-constrained clusters.

For a non-empty subset C of X, let $D(\supseteq C)$ denote the set of all individuals $v \in V$ whose descendants contain every individual in C, let $D(\subseteq C)$ denote the set of individuals in V all of whose descendants in X are contained in C, and let $D(=C) := D(\supseteq C) \cap D(\subseteq C)$ denote the set of all individuals in V whose descendants in X coincides exactly with C. That is, we put

$$D(\supseteq C) := \{v \in V : \overrightarrow{v} \supseteq C\}, \quad D(\subseteq C) := \{v \in V : \overrightarrow{v} \subseteq C\},$$

and put

$$D(=C) := \{v \in V : \overrightarrow{v} = C\}.$$

So, $D(=C)$ consists of all individuals in V that are ancestral exactly to *every* element in C, but no other elements in X. Recall that a subset D of V is said to *separate* two other subsets C_1 and C_2 of V if every (undirected) path in G from a vertex in C_1 to a vertex in C_2 passes through some vertex from D. With this in hand, we define a subset C of X to be a *tight* cluster (in X relative to G) if and only if it is non-empty and $D(=C)$ separates C from $X-C$. Note that for any non-empty subset C of X, and any non-empty subset V' of $D(\supseteq C)$, we have $C \subseteq \bigcap_{v \in V'} \overrightarrow{v} \subseteq V' = \bigcup_{v \in V'} \overrightarrow{v}$ as well as

$$\overrightarrow{V'} = C \iff V' \subseteq D(=C). \tag{2}$$

Notice that a subset C of X is a tight cluster if there exists a subset V_C of $D(=C)$ that separates C from $X-C$.

As an example, the non-singleton tight X-clusters of the graph G shown in Fig. 2 are $\{a,b\}$ and X, as $D(=\{a,b\}) = \{v_1,v_2,v\}$ holds where v is the left-hand parent of v_1 and v_2, and this set clearly separates $\{a,b\}$ from $\{c,d,e\}$; yet the subset $\{v_1, v_2\}$ of $D(=\{a,b\})$ also separates $\{a,b\}$ from $\{c,d,e\}$.

Notice that X itself is (trivially) a tight cluster. Notice also that the set of tight X-clusters of G is always a subset of the hierarchy $\mathcal{H}_{\mathcal{P}}$ defined in (1) for $\mathcal{P} = \{\overrightarrow{v} : v \in V\}$, though, in general, the latter set can be strictly larger than the set of tight X-clusters of G.

The concept of a tight cluster is a relaxation of the notion of 'organismic exclusivity' described recently by Baum (2009), which requires that there is an element in $D(=C)$ that separates C from $X-C$. In case G is itself a rooted tree with leaf set X, the tight clusters correspond precisely to the clusters of this tree (i.e. the sets $\{\overrightarrow{v} : v \in V\}$).

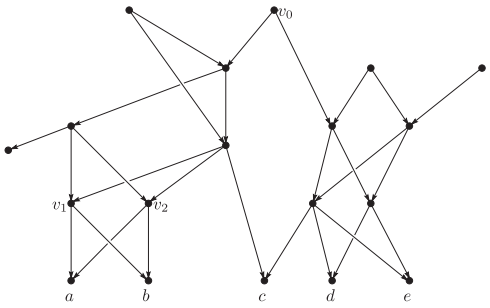

Fig. 2. An illustrative example of an acyclic digraph G, with vertex set V and $X = \{a,b,c,d,e\} \subset V$.

3.2. An example of a tight cluster in recent evolution

The conditions for a tight cluster are strong. However, results in population genetics suggest that for diploid (sexually-reproducing) populations, it may sometimes be reasonable. This is because, under a neutral model of random diploid mating, Chang (1999) showed that if we trace back the ancestry of a set of n extant individuals by (at least) $1.77\log_2(n)$ generations, the population extant at this earlier time is likely to have the property that each individual in this ancestral population either has no extant descendants, or has all n extant individuals as descendants. This sharp $\log_2(n)$ behavior was shown to extend to more realistic models of human mating behaviour, including migration, at the price of a constant larger than 1.77 by Rohde et al. (2004).

The significance of this finding can be illustrated by considering, for example, the entire extant human population P_{hom} as a subset of the set X of all extant organisms on earth today. The work of Rohde et al. (2004), along with recent evidence that the radiation of modern humans from Africa occurred within the last 150,000 years (Liu et al., 2006) suggests that—excluding the existence of a *Homo erectus* type Yeti or Bigfoot—every individual v that was living, say, 200,000 years ago satisfies either $P_{\text{hom}} \subseteq \overrightarrow{v}$ or $\overrightarrow{v} \cap P_{\text{hom}} = \emptyset$. Moreover, as we can presumably be confident that no currently living organism $x \in X$ that is not in P_{hom} is a descendant of any such individual v whose current descendants include individuals from P_{hom}, the collection V_{hom}^{200K} of all individuals living 200,000 years ago whose current descendants include at least one individual from P_{hom}, must be contained in $D(=P_{\text{hom}})$ and, therefore, V_{hom}^{200K} separates P_{hom} from all other currently living organisms.

Thus, we may assume that P_{hom} is, formally, a tight cluster in the set X of all extant organisms alive today, provided we consider organismal ancestry (i.e. we are ignoring the possible transfer of human genes into other organisms (e.g. by viruses)).

The example also underlines that, because of our specific choice of X, side lines with no descendants today (like, presumably, the late European Neanderthals Green, 2010) are of no direct interest in this context. Indeed, we may probably (that is, unless the Yeti or Bigfoot exists and belongs to the *Homo erectus* group) also replace V_{hom}^{200K} by V_{hom}^{2M}, the group of all individuals that lived two million years ago and that also have current descendants in P_{hom}.

In the case of haploid reproduction, coalescence times are much longer, being of order n rather than $\log(n)$ (Hein et al., 2005). Nevertheless, consider a current population of n individuals with haploid reproduction. Suppose the ancestors of this population dating back as far as N generations into the past constituted a homogeneous population was of approximately constant size, and was genetically isolated (i.e. if there were LGT events involving this ancestral population then they were restricted to exchanges between members of that population) and which left no other descendants today. Then, provided $N \gg n$, this current population would be a likely candidate for a tight cluster in the set X of all extant organisms.

3.3. Strict clusters

We now describe a second class of clusters; we will see in Theorem 2 that these include the tight clusters, yet they are still guaranteed to form a hierarchy.

Define a subset C to be a *strict X-cluster* (relative to V and $\preceq$) provided that

- $v \in V$ and $C \cap \overrightarrow{v} \neq \emptyset$ implies that either $C \subseteq \overrightarrow{v}$ or $\overrightarrow{v} \subseteq C$—or, equivalently, $v \in D(\supseteq C)$ or $v \in D(\subseteq C)$ holds, and

- the *cousinship graph*

$$\Gamma_G(C) := \left(C, \left\{\{x,y\} \in \binom{C}{2} : \exists v \in D(\subseteq C) : x,y \in \overrightarrow{v}\right\}\right)$$

of C is connected.

As an example, the non-singleton strict X-clusters of the graph G shown in Fig. 2 are $\{a,b\}$, $\{d,e\}$ and X. To see that $\{a,b\}$ is a strict X-cluster, notice that the vertices v in V with $C \cap \overrightarrow{v} \neq \emptyset$ are v_1, v_2 and the five other vertices of the graph that have v_1 (or v_2) as a descendant, and for each of these vertices v, we have $\{a,b\} \subseteq \overrightarrow{v}$, so the first condition in the definition is fulfilled. The cousinship graph $\Gamma_G(\{a,b\})$ consists of the two vertices a,b connected by an edge, so it is connected as required for the second condition. Notice in this example that $\{d,e\}$ is a strict X-cluster that fails to be a tight cluster.

Note that, in general, the set X is itself a strict cluster if and only if the weakest connectivity condition (C3) holds.

3.4. Clusters based on ancestry

We begin this sub-section with some further definitions.

For any pair of elements $\{a,b\}$ in V, let

$$\mathrm{ca}(a,b) := \{v \in V : v \preceq a \text{ and } v \preceq b\}$$

be the set of *common ancestors* of a and b. Provided that $\mathrm{ca}(a,b)$ is non-empty, let $\mathrm{mrca}(a,b)$ be the maximal elements in $\mathrm{ca}(a,b)$; this is often referred to as the set of the *most recent common ancestors* of a and b. For $a,b,c \in X$, let us write $ab\|c$ if $\mathrm{ca}(a,b)$ is non-empty, and for each $v \in \mathrm{mrca}(a,b)$ there exists $v' \in \mathrm{mrca}(a,c)$ and $v'' \in \mathrm{mrca}(b,c)$ such that $v' \prec v$ and $v'' \prec v$ hold.

As an example, for the graph G in Fig. 2, we have $ab\|x$ for each $x \in \{c,d,e\}$, and we have $de\|y$ precisely when $y \in \{a,b\}$.

We will write $ab|c$ under the strictly weaker condition that $\mathrm{ca}(a,b)$ is non-empty, and there exists, for each $v \in \mathrm{mrca}(a,b)$, some $v' \in \mathrm{mrca}(a,c) \cup \mathrm{mrca}(b,c)$ with $v' \prec v$.

A dual notion to the ancestral relation $\|$ is the following: For $a,b,c \in X$, let us write $ab \perp c$ if $\mathrm{ca}(a,c)$ and $\mathrm{ca}(b,c)$ are both non-empty and there exist, for all $v \in \mathrm{mrca}(a,c)$ and $v' \in \mathrm{mrca}(b,c)$, some $u,u' \in \mathrm{mrca}(a,b)$ (where u need not necessarily be different from u') such that $v \prec u$ and $v' \prec u'$ holds. Note that $\|$ is neither stronger nor weaker than $\perp$, that is, there are examples for which $xx' \perp y$ holds but $xx'\|y$ fails (Fig. 3(a)) and also for which $xx'\|y$ holds but $xx' \perp y$ fails (Fig. 3(b)).

The following result summarizes a basic property of these relations, and will be useful in the next section.

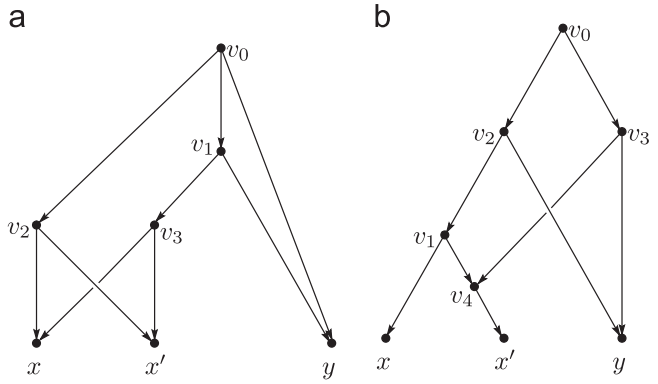

Fig. 3. (a) An acyclic digraph G on $X=\{x,x',y\}$ for which $\{x,x'\}$ is a tight cluster and a co-ancestral cluster but is not an ancestral cluster. (b) An acyclic digraph G on $X=\{x,x',y\}$ for which $\{x,x'\}$ is an ancestral cluster but not a co-ancestral cluster.

Lemma 1. *Suppose that G is any finite, directed, cycle-free graph, with $X \subseteq V$. Given three distinct elements $a,b,c \in X$:*

(i) *At most one of $ab\|c$, $ac\|b$ and $bc\|a$ holds;*
(ii) *At most two of $ab|c$,$ac|b$,$bc|a$ hold;*
(iii) *At most one of $ab \perp c$,$ac \perp b$,$bc \perp a$ holds.*

Proof. For part (i), assume that both $ab\|c$ and $ac\|b$ hold. Let v be any element in $\mathrm{mrca}(a,b)$; then there exists $v' \in \mathrm{mrca}(a,c)$ with $v' \prec v$. On the other hand, there also exists an element $u \in \mathrm{mrca}(a,b)$ such that $u \prec v'$ in view of $ac\|b$. Therefore, we have $u \prec v$ and $u,v \in \mathrm{mrca}(a,b)$, a contradiction to the definition of $\mathrm{mrca}(a,b)$. The second and third parts follow by a similar proof by contradiction. This completes the proof of the Lemma. □

With these definitions, we say that C is an *ancestral X-cluster* (respectively, *relaxed ancestral X-cluster* and *co-ancestral cluster*) if for all $x,x' \in C$ and $y \in X-C$, we have $xx'\|y$ (respectively, $xx'|y$ and $xx' \perp y$). Notice that the entire set X is both an ancestral cluster and a co-ancestral cluster under the intermediate connectivity condition (C2).

Note that, even for a digraph G that has a vertex v_0 with $\overrightarrow{v_0} = X$, there may exist a tight X-cluster that is not an ancestral cluster, as Fig. 3(a) shows for $C=\{x,x'\}$. In this example, $D(=C)=\{v_2, v_3\}$, from which it is easily seen that C is a tight cluster. Note that $v_2 \in \mathrm{mrca}(x,x')$ yet v_2 is not a descendant of any vertex in either $\mathrm{mrca}(x,y)=\{v_1\}$ or $\mathrm{mrca}(x',y)=\{v_1\}$.

3.5. Clusters relative to a 'time scale'

In this section, we exploit an additional aspect of evolution—the fact that the vertices of G have an associated 'date' (e.g. time when they were born) and this provides a further avenue to define a system of clusters.

Suppose that, in addition to the digraph $G = (V,A)$, with $X \subseteq V$, we have a map $T : V \to \mathbb{R}$ that strictly preserves the partial order $\preceq$, i.e.

$$u \prec v \Longrightarrow T(u) < T(v).$$

We refer to the pair (G,T) as a *valuated digraph on X*. Of course, the condition that such a map T exists is equivalent to the condition that G has no directed cycles (Bang-Jensen and Gutin, 2008), but we think of T as being a specific map, where, in the biological context, $T(v)$ would denote the time when the individual v was born (we may regard the present as time 0 and so T is a map from V to the non-positive reals).

Following Steel (2007) we say that $C \subseteq X$ is an *Apresjan X-cluster relative to T* if there exists $t \in \mathbb{R}$ such that:

(T1) For all $x,y \in C$, there exists $v \in V : v \preceq x,y, T(v) \geq t$; and
(T2) For all $x \in C$, $y \in X-C$, if $v \in V$ satisfies $v \preceq x,y$ then $T(v) < t$.

In words, C is an Apresjan X-cluster relative to T if every two individuals in C have at least one common ancestor after time t, but each individual in C and each individual in $X-C$ have all their common ancestors earlier than t.

We say that $C \subseteq X$ is a *strong Apresjan X-cluster relative to T* if (T1) is strengthened to:

(T1′) For all $x,y \in C$, and *every* $v \in \mathrm{mrca}(x,y)$, $T(v) \geq t$.

Thus, C is a strong Apresjan X-cluster relative to T if every two individuals in C have *all* their most recent common ancestors after time t, but any individual in C and individual in $X-C$ have all their common ancestors earlier than t.

4. Main result

We have described a variety of ways to construct a set of X-clusters from G. We now show that they all lead to hierarchies (in one case a weak hierarchy), and describe some relationships between them, in the following main result of this paper.

Theorem 2. *Suppose that G is any finite, directed, cycle-free graph, with $X \subseteq V$.*

(1) *The following sets form a hierarchy*:
  (a) *The set of tight X-clusters of G*;
  (b) *The set of strict X-clusters of G*;
  (c) *The set of ancestral X-clusters of G*;
  (d) *The set of co-ancestral X-clusters of G.*
(2) *The set of relaxed ancestral X-clusters of G forms a weak hierarchy.*
(3) *Suppose that (G,T) is a valuated digraph on X. Then the set of Apresjan X-clusters relative to T forms a hierarchy (as does the subset of strong Apresjan X-clusters relative to T).*
(4) *Every tight X-cluster C of G is also a strict X-cluster and, under connectivity condition* (C2)*, a co-ancestral cluster. If G has a valuation map T, C is also an Apresjan X-cluster relative to T.*

Proof of Part 1(a). Suppose that for two tight clusters C_1 and C_2 we have $C_1 \cap C_2 \neq \emptyset$ and that C_2 is not a subset of C_1. We will show that $C_1 \subseteq C_2$. Let $V_1 = D(=C_1)$ and $V_2 = D(= C_2)$. By assumption, there exists $x \in C_1 \cap C_2$, $y \in C_2 - C_1$. First observe that if $V_2 \subseteq V_1$ then $\overrightarrow{V_2} \subseteq \overrightarrow{V_1}$, which implies that $C_2 \subseteq C_1$ in violation of our assumption. Thus, there exists $v \in V_2 - V_1$. Now, since $x,y \in C_2$, and $v \in V_2$ there exists a directed path from v to x and a directed path from v to y. In particular these provide an undirected path P in G connecting x and y. But now, since $x \in C_1$ while $y \in X - C_1$, and since C_1 is tight (so V_1 separates C_1 from $X - C_1$) at least one vertex, say w, in P must lie in V_1. Regardless of where w lies on P we have $v \preceq w$ (since every vertex v' on P satisfies $v \preceq v'$) and so $\overrightarrow{w} \subseteq \overrightarrow{v}$. Therefore, since $\overrightarrow{w} = C_1$ and $\overrightarrow{v} = C_2$, we have $C_1 \subseteq C_2$, as required. This completes the proof of *Part* 1(a).

To establish *Part* 1(b), suppose that C,C' are strict X-clusters, and that $C \cap C'$ and $C - C'$ are both non-empty. We will show $C' \subseteq C$. Take $x \in C \cap C', y \in C - C'$. By the connectivity of the cousinship graph $\Gamma_G(C)$ there is a path in this graph from x to y, say $x = x_1, x_2, \ldots, x_k = y$. Let x_i, x_{i+1} be the first pair of adjacent vertices in this path for which $x_i \in C \cap C'$ and $x_{i+1} \in C - C'$. Since x_i and x_{i+1} are adjacent there is a vertex $v \in V$ for which $x_i, x_{i+1} \in \overrightarrow{v} \subseteq C$. Moreover, we have $\overrightarrow{v} \cap C' \neq \emptyset$ (since $x_i \in C' \cap \overrightarrow{v}$) and so the first condition in the definition of a strict cluster implies that either $C' \subseteq \overrightarrow{v}$ or $\overrightarrow{v} \subseteq C'$. But the second of these two inclusions is impossible, since $x_{i+1} \in \overrightarrow{v} - C'$. Thus $C' \subseteq \overrightarrow{v}$ and since $\overrightarrow{v} \subseteq C$, this implies that $C' \subseteq C$, as required to establish *Part* 1(b).

For *Part* 1(c), assume, for the sake of contradiction, that C,C' are ancestral clusters, and there exist three elements a,b,c with $a \in C - C'$, $b \in C' - C$ and $c \in C \cap C'$. Then, by definition, we have $ac \| b$ and $bc \| a$, a contradiction to Lemma 1(i). A similar argument applies for *Part* 1(d). This completes the proof of *Part* 1. □

Proof of Part 2. Suppose that A,B,C are three relaxed ancestral clusters which violate the condition $A \cap B \cap C \notin \{A \cap B, A \cap C, B \cap C\}$. Then we can select $x \in A \cap B - C, y \in A \cap C - B, z \in B \cap C - A$. We have $xy|z$ (since x and y but not z are in A), and $xz|y$ (since x and z but not y are in B), and $yz|x$ (since y and z but not x are in C), in violation of Lemma 1(ii). □

Proof of Part 3. This result is from Steel (2007), based on earlier related results from Apresjan (1966), Bryant and Berry (2001), Devauchelle et al. (2004). Since the proof is short, we provide it here for completeness. Suppose C_1, C_2 are Apresjan X-clusters relative to T and there exists $x \in C_1 \cap C_2, y \in C_1 - C_2, z \in C_2 - C_1$; we will show that this leads to a contradiction. For $i \in \{1,2\}$, let t_i be a value of t for which (T1), (T2) applies for $C = C_i$. If $t_1 \geq t_2$ then, by condition (T1) on C_1, there exists v with $v \preceq x,y$ with $T(v) \geq t_1 \geq t_2$. But applying (T2) to C_2 gives $T(v) < t_2$ (since $y \in X - C_2$), a contradiction. A similar argument applies if $t_1 \leq t_2$.

Proof of Part 4. Suppose that C is a tight X-cluster. We first show that C is a strict X-cluster. Select any $w \in D(=C)$. Then $\overrightarrow{w} = C$, and so the cousinship graph $\Gamma_G(C)$ is a clique (and hence a connected graph). Now, suppose that $C \cap \overrightarrow{v} \neq \emptyset$, and that $\overrightarrow{v}$ is not a subset of C. We will show that $C \subseteq \overrightarrow{v}$. Select $x \in C \cap \overrightarrow{v}, y \in \overrightarrow{v} - C$. There exists a directed path in G from v to x and a directed path from v to y. In particular, these provide an undirected path P in G connecting x and y. Since $x \in C$ but y lies outside of C, path P must contain at least one vertex $v' \in D(=C)$ (since $D(=C)$ separates C from $X - C$). Then $v \preceq v'$ and so $\overrightarrow{v'} \subseteq \overrightarrow{v}$. But $\overrightarrow{v'} = C$ (since $v' \in V_C$) so that $C \subseteq \overrightarrow{v}$, as required to establish that C is strict X-cluster.

Next we show that C is a co-ancestral cluster, i.e. for any $x,x' \in C, y \in X - C$ we have $xx' \perp y$. Let v be a vertex in $\mathrm{mrca}(x,y)$ (such a vertex exists by (C2)) and consider the (undirected) path P from x to v to y. Since $x \in C$ to $y \in X - C$, the fact that $D(=C)$ separates C from $X - C$ (because C is a tight cluster) implies that one vertex, say w, in P must lie in $D(= C)$. The vertex w does not lie on the path from v to y, otherwise we have $y \in \overrightarrow{v} = C$, so w is in the path from v to x. Since $x' \in \overrightarrow{w}$, it follows that w is, or has as a descendant, a vertex in $\mathrm{mrca}(x,x')$. A similar argument applies to any vertex in $\mathrm{mrca}(x',y)$ and so $xx' \perp y$. Since this holds for all $x,x' \in C$ and $y \in X - C$, C is a co-ancestral cluster of X.

For the final claim in *Part* 4, suppose that C is a tight X-cluster of G. We will show that (T1) and (T2) hold for $t = t_C$ where: $t_C := \max\{t(v) : v \in D(=C)\}$. First select $v_0 \in V_C$ with $T(v_0) = t_C$. Observe that for all $x,x' \in C$, we have $v_0 \preceq x,x'$ and since $T(v_0) \geq t_C$ we see that condition (T1) is satisfied for $t = t_C$, and $v = v_0$. To verify condition (T2), suppose that $x \in C, y \in X - C$ and there exists $v \preceq x,y$ with $T(v) \geq t_C$. Consider the (undirected) path P in G from x to v and then to y. If $v \in V_C$ then $\overrightarrow{v} = C$ which is impossible since $v \preceq y \Rightarrow y \in \overrightarrow{v}$ yet y is not an element of C. Thus v is not an element of V_C. Moreover, for any vertex w in P that is different from v, we have $T(w) > t_C$ (since $v \prec w$ and T is strictly monotone) and so w is also not an element of V_C (since all the vertices w' in V_C satisfy $T(w') \leq t_C$). In summary, none of the vertices in P belongs to V_C, thus deleting V_C fails to disconnect x from y, violating the assumption that $D(=C)$ separates C from $X - C$. This establishes property (T2), as required, and thereby completes the proof. □

5. Discussion

Our paper is motivated partly as a response to a currently promoted viewpoint that processes of reticulate evolution, such as extensive LGT implies that no sensible or well-defined 'tree of life' can be constructed (Doolittle, 1999; Kunin et al., 2005; Lawton, 2009). However, this statement depends on how one views such a tree, and where the transfer events occurred in it. For example, even if each gene has been transferred once during its history (Dagan and Martin, 2007), provided that these transfer events all occurred before the separation of certain populations then we may still expect to find Apresjan or stronger (e.g. tight) clusters, which will therefore form a tree. Consider, for example, the collection C of

all extant mammals. The most recent common ancestors of mammals most likely occurred within the last 120 million years (Eizirik et al., 2001). Thus if those genes that are found in mammals and which underwent a gene transfer event some time in their past did so at a much earlier stage of evolution (i.e. well before 120 million years ago) then the concept of a 'mammal tree' composed of clusters of a type described above seems reasonable.

Neither are recent LGT events necessarily problematic. In particular, such events will not destroy even a tight cluster C provided they occur amongst those ancestors of C that are descendants of $D(=C)$. Of course it is possible to explicitly include or exclude different types or levels of LGT in the definition of G and this may lead to different collections of tight clusters for G. For example, in the discussion of modern humans as a tight cluster we have explicitly ignored any transfer of human genes into other organisms (e.g. by viruses). The same comment applies to the other cluster constructions in this paper.

For prokaryotes, where a tree structure is most vigorously called into question, the concept of a tree is still well defined, but it may indeed be poorly resolved (depending on the type of cluster considered, and the extent to which a LGT event from individual x to y might be counted as an arc in G from x to y—for example, one could indicate all such instances or just those for which the gene transfers survives to a present copy). In cases where LGT (and other types of reticulate evolution) are extensive and on-going, then set systems such as weak hierarchies may give a more informative picture of evolution than a tree. We have described one way to generate such a hierarchy above, but it may be useful to explore other approaches.

In this paper, we have concentrated instead on ways by which a hierarchy on X can be constructed from G based on concepts of ancestry and separation. Of course, the possibilities we have outlined are by no means exhaustive, as there will surely be other combinations of conditions that will allow for a hierarchy or related set system from G.

However, we would like any procedure for constructing a hierarchy to have some reasonable biological motivation and also, if possible, to satisfy some desirable properties. One such desirable property is that when G is itself a rooted tree with leaf set X then the set of clusters derived from G should be precisely the clusters of this tree (the sets $\overrightarrow{v} : v \in V$). This property is enjoyed by all of the constructions described in this paper.

A second desirable property is that the construction should be *equivariant* with respect to any re-labelling (permutation) of the elements of X. More precisely, if $\mathcal{H}$ is the hierarchy constructed from G, and if σ is a permutation of the elements of X, then the hierarchy constructed from the graph obtained from G by permuting the elements in X according to σ should be the hierarchy $\{\sigma(A) : A \in \mathcal{H}\}$ where $\sigma(A) = \{\sigma(a) : a \in A\}$. This property recognizes that the construction should be mathematically canonical in that the 'names' we attach to individuals should not play any special role in the construction of the hierarchy; rather, all that matters is how the individuals are related to each other through the ancestral graph G. All of the constructions described in this paper also satisfy this equivariance property.

A third desirable property is that the procedure be 'robust' with respect to the possibility that we have not sampled or observed all individuals in X. We can make this precise as follows.

Suppose that G is any finite, directed, cycle-free graph with $X \subseteq V$, and that Y is a subset of X. Let $G|Y$ be the directed graph obtained from G by regarding the vertices in $X-Y$ as unlabeled vertices. Now suppose we have a function ϕ that associates to each such pair (X,G) a collection of subsets of X. We say that ϕ satisfies *sampling consistency* if it satisfies the condition:

$$C \in \phi(X,G) \Rightarrow C \cap Y \in \phi(Y,G|Y).$$

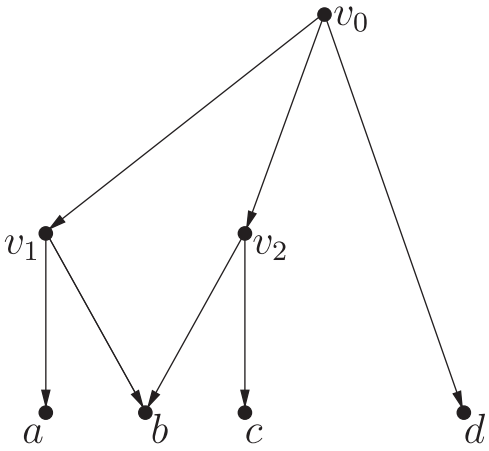

Fig. 4. An example to illustrate a violation of sampling consistency for strict clusters.

We can extend this concept to valuated digraphs in the obvious way (namely, $C \in \phi(X,G,T) \Rightarrow C \cap Y \in \phi(Y,G|Y,T)$).

It can be checked that the following constructions satisfy sampling consistency: tight clusters, ancestral clusters, and Apresjan clusters (with respect to a time scale). However, the strict cluster construction can violate this condition—for example, consider the graph G in Fig. 4. Then $C=\{a,b,c\}$ is a strict X-cluster where $X=\{a,b,c,d\}$. But if we select $Y=\{a,c,d\}$ then $C \cap Y = \{a,c\}$ is not a strict Y cluster in the graph $G|Y$, since the cousinship graph $\Gamma_{G|Y}(C \cap Y)$ is not connected.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments. A.D. and T.W. thank the CAS, the BMBF, and the MPG for financial support. V.M. and T.W. thank the Engineering and Physical Sciences Research Council (EPSRC) for its support [Grant EP/D068800/1]. M.S. thanks the Royal Society of NZ under its James Cook Fellowship scheme.

References

Aldous, D., Krikun, M., Popovic, L., 2008. Stochastic models for phylogenetic trees on higher-order taxa. J. Math. Biol. 56, 525–557.
Apresjan, J.D., 1966. An algorithm for constructing clusters from a distance matrix. Mashinnyi perevod: prikladnaja lingvistika 9, 3–18.
Bang-Jensen, J., Gutin, G.Z., 2008. Digraphs: Theory, Algorithms and Applications, second ed. Springer.
Baum, D.A., 2009. Species as ranked taxa. Syst. Biol. 58 (1), 74–86.
Bryant, D., Berry, V., 2001. A structures family of clustering and tree reconstruction methods. Adv. Appl. Math. 27, 705–732.
Chang, J., 1999. Recent common ancestors of all present-day individuals. Adv. Appl. Prob. 31, 1002–1026.
Crick, F., 1968. The origin of the genetic code. J. Mol. Biol. 38, 367–379.
Dagan, T., Martin, W., 2007. Ancestral genome sizes specify the minimum rate of lgt during prokaryote evolution. Proc. Natl. Acad. Sci. USA 104, 870–875.
Dagnan, T., Martin, W., 2006. The tree of one percent. Genome Biol. 7, 118.
Darwin, C., 1872. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, sixth ed. John Murray, London.
Devauchelle, C., Dress, A.W.M., Grossmann, A., Grünwald, S., Henaut, A., 2004. Constructing hierarchical set systems. Ann. Combin. 8, 441–456.
Doolittle, W.F., 1999. Phylogenetic classification and the universal tree. Science 284, 2124–2128.
Dress, A.W.M., Moulton, V., Wu, T., 2009. A topological approach to tree (re-)construction. Submitted for publication.
Dress, A.W.M., Huson, D., Moulton, V., 1996. Analysing and visualising sequence and distance data using splitstree. Discr. Appl. Math. 71, 95–109.
Edmonds, J., Giles, R., 1977. A min-max relation for submodular functions on graphs. In: Hammer, P.L., Johnson, E.L., Korte, B.H., Nembauser, G.L. (Eds.), Studies in Integer Programming (Proceedings of Workshop, Bonn, 1975)

Annals of Discrete Mathematics, vol. 1. Elsevier Science Ltd., North-Holland, Amsterdam, pp. 185–204.
Eizirik, E., Murphy, W.J., O'Brien, S.J., 2001. Molecular dating and biogeography of the early placental mammal radiation. J. Hered. 92 (2), 212–219.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.
Futuyma, D.J., 1998. Evolutionary Biology. Sinauer Associates, Sunderland, MA.
Green, R.E., et al., 2010. A draft sequence of the neandertal genome. Science 328 (5979), 710–720.
Hein, J., Schierup, M.H., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press.
Kunin, V., Ogldovsky, L., Darzentas, N., Ouzounis, A., 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res. 15, 954–959.
Lawton, G., 2009. Why Darwin was wrong about the tree of life. New Sci. 2692, 34–39.
Linneaus, C., 1735. Systema Naturae.
Liu, H., Prugnolle, F., Manica, A., Balloux, F., 2006. A geographically explicit genetic model of worldwide human-settlement history? Am. J. Hum. Genet. 79 (2) 230–237.
Matsen, F.A., Evans, S.N., 2008. To what extent does genealogical ancestry imply genetic ancestry? Theor. Popul. Biol. 74 182–190.
Mirkin, B., McMorris, F.R., Roberts, F.S., Rzhetsky, A., 1997. Mathematical Hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 37. American Mathematical Society, Providence RI.
Rohde, D., Olson, S., Chang, J., 2004. Modelling the recent common ancestry of all living humans. Nature 431, 562–566.
Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press.
Sober, E., Steel, M., 2002. Testing the hypothesis of common ancestry. J. Theor. Biol. 218, 395–408.
Steel, M., 2007. Tools to construct and study big trees: a mathematical perspective. In: Hodkinson, T., Parnell, J., Waldren, S. (Eds.), Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa. CRC Press (Taylor and Francis), pp. 97–112.
Theobald, D.L., 2010. A formal test of the theory of universal common ancestry. Nature 465, 219–222.
Wheeler, Q.D., Meier, R., 2000. Species Concepts and Phylogenetic Theory: A Debate. Colombia University Press, New York.
Woese, C.R., 2000. Interpreting the universal evolutionary tree. Proc. Natl. Acad. Sci. USA 97, 8392–8396.