Opinion

Clanistics: a multi-level perspective for harvesting unrooted gene trees

François-Joseph Lapointe[1], Philippe Lopez[2], Yan Boucher[3], Jeremy Koenig[4] and Eric Bapteste[2]

[1] Département de sciences biologiques, Université de Montréal, Montréal, QC, Canada
[2] Université Pierre et Marie Curie (UPMC), Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche 7138, 75005 Paris, France
[3] Department of Microbiology, Cornell University, Ithaca, NY, USA
[4] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Prokaryotic evolution takes place within and between genomes, when significant amounts of genes are transferred and recombined between interacting genetic partners. These non-tree-like evolutionary processes, intertwined with events of vertical descent, lead to a massive production of unrooted trees in which branches, nodes and groupings have different biological meanings than for the rooted trees usually studied by phylogenetics. Such unrooted gene trees can not only inform us about organismal phylogeny, but also about the variety of evolutionary, genetic, functional and ecological relationships affecting a plurality of evolutionary units, at multiple levels – from genes, groups of genes, organisms and consortia, to communities. Here we introduce new notions designed to analyze unrooted trees with more depth and accuracy. We demonstrate how a clanistic perspective can significantly improve our knowledge of evolutionary processes and relationships for most evolving systems, whether they are mobile genetic elements or cellular genomes.

Harvesting the forest of unrooted gene trees

Our understanding of the evolutionary processes acting on the vast majority of life forms and their component parts has radically changed [1–4]. In addition to vertical descent, the mechanisms creating genetic diversity in many genomes involve a significant proportion of lateral DNA transfer and recombination. This is true for the genomes of mobile elements such as phages [5–7] or plasmids [8–10], and for prokaryotic chromosomes [11,12]. Estimates of these lateral events (LE) differ between lineages, but are often considerable, sometimes far exceeding the amount of variation introduced by internal sources, such as spontaneous mutations within genomes [11]. Importantly, these LE affect genetic partners (DNA donors and hosts) that are not necessarily directly phylogenetically related [3].

Consequently, no genome is an island, isolated on its phylogenetic branch, only diverging from other evolving entities by drift, natural selection and vertical descent. Rather, genomes of many prokaryotes and mobile elements are genetic chimeras, interacting with multiple partners in an integrated fashion. Their evolution is affected by the overall genetic diversity of the milieu in which these interacting partners find themselves [1,9,13]. Moreover, genes are not merely organismal parts, but are also parts of other evolutionary units [14,15]. Thus, gene evolution and organism evolution are partly decoupled, and this multiplies the number of evolutionary units required to describe evolution accurately. In particular, genetic modules (or sets of co-evolving genes) can result from this mix-and-matching of genes [16] and be sustained (or disrupted) by specific selective (environmental) pressures. Such modules comprise a variable number of (functionally) integrated genes, each with their own original phylogenetic histories.

Glossary

Adjacent groups: groups of OTUs in an unrooted tree that correspond to sister groups in a rooted tree.
Bipartition (split): separation of OTUs in two groups by the removal of one edge of a tree.
Clade (monophyletic group): a group of OTUs comprising all the descendants of a common ancestor. Clades are identified by every bipartition of a rooted tree.
Clan: the unrooted analogue of clade. Clans are identified by every bipartition of an unrooted tree.
Clip: a group of OTUs that are each other's closest members in terms of path-length distances.
Complete group: a group of OTUs (i.e. clan, clip, slice, or clade) including all members of a given category.
Equitability index (E): a local measure of diversity (or equitability) of the different categories present in a subtree (i.e. clan, clip, slice, or clade).
Homogeneous group: a group of OTUs (i.e. clan, clip, slice, or clade) with members pertaining to a single category.
Heterogeneous group: a group of OTUs (i.e. clan, clip, slice, or clade) with members pertaining to more than one category.
Incomplete group: a group of OTUs (i.e. clan, clip, slice, or clade) including some but not all members of a given category.
Intruder: the OTU of a category included in a subtree grouping (i.e. clan, clip, slice, or clade) of a different category.
Operational taxonomical unit (OTU): synonymous with terminal taxon; a group of organisms used in a taxonomic study without designation of taxonomic rank.
Perfect group: a group of OTUs (i.e. clan, clip, slice, or clades) that is complete and homogeneous at the same time.
Root: a node of a tree that imposes an ancestor–descendant direction away from the root on the other nodes.
Shannon diversity index (H): overall measure of diversity of the different categories of OTUs present in the tree.
Sister groups: groups of OTUs that are each other's closest relatives in a rooted tree.
Slice: identified by tripartitions of an unrooted tree, thus defining an internal segment of an unrooted tree.
Tripartition: separation of OTUs in three groups by the removal of two edges of a tree.

Corresponding author: Lapointe, F.-J. (francois-joseph.lapointe@umontreal.ca).

0966-842X/$ – see front matter doi:10.1016/j.tim.2010.03.009 Trends in Microbiology 18 (2010) 341–347

Box 1. Different possible rootings of unrooted gene trees

Genes are mobilized between interacting genetic partners, and are affected by multiple selective pressures, because they belong to multiple evolutionary units (e.g. genetic modules, organisms, and communities in a given environment). Inevitably, this interplay of vertical inheritance and fundamentally non-tree-like processes result in unrooted gene trees harboring a complex mix of relationships including orthology (resulting from speciation and reflecting genealogical relationships), paralogy (resulting from duplication events), and xenology (resulting from LE and reflecting genetic partnership relationships) (Figure Ia). These trees are unrooted because they cannot be polarized from ancestors to descendants, and searching for a root among them is pointless, because by definition their branching patterns do not simply reflect the species and lineages genealogy, but inform us on the peculiar history of the genes. Typically, phylogenetically unrelated genetic partners will branch together in these gene trees, even though they did not share one last common ancestor. In practice, such gene trees can be rooted with an outgroup to impose a direction upon the branches of the tree (Figure Ib). In the absence of an outgroup, midpoint-rooting can also be used to place the root halfway between the two most distant leaves. When ancient paralogous sequences can be clearly identified, the tree can be rooted at the node representing the ancestral duplication event.

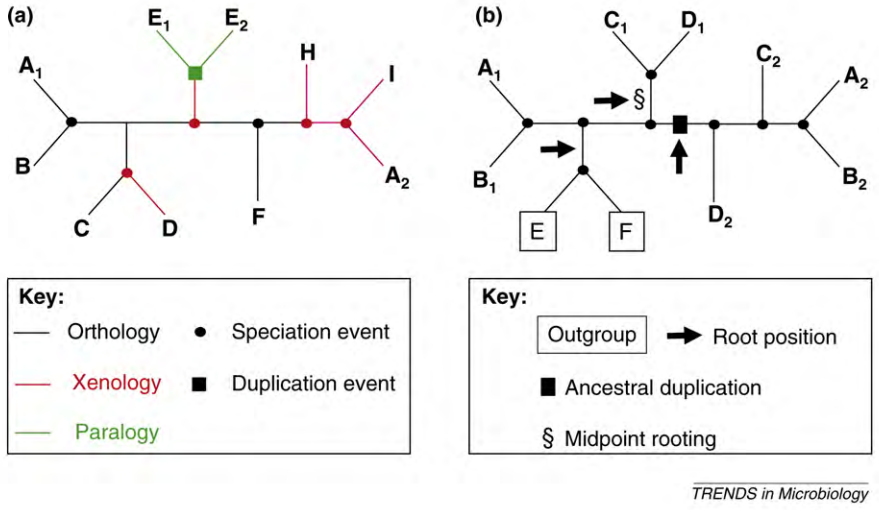


Figure I. Rooting a theoretical gene tree. **(a)** Example of an unrooted phylogenetic tree based on a hypothetical gene family containing duplication and speciation events and a mix of relationships including orthology, paralogy and xenology. **(b)** The gene tree could be rooted at outgroups or with an ancestral duplication event, or by midpoint rooting.

In addition to this modular aspect of gene evolution, LE define broader evolutionary units: the communities of interacting genetic partners thriving on a shared gene pool. Understanding the rules of genetic exchanges (the nature and frequency of LE) between members of such communities is also compelling [14].

Just as spontaneous mutations leave their historical fingerprints in DNA molecules, traces of integrative evolution are also recorded in the genes. Consequently, gene trees are much more informative than initially conceived [17]. Yet there is a huge catch. For historical reasons, current phylogenetic concepts and methods are not suited to exploit the signal of integrative evolution. Molecular phylogenetics was initially developed to retrace the history of biological species and lineages, by providing their genealogical relationships as reflected by their branching order on a furcating topology, globally polarized in time [17]. Accordingly, evolutionary history is usually analyzed with concepts stemming from ideal gene trees (i.e. identifying the root, hypothetical ancestors, monophyletic groups, or sister groups) [18].

It has often been tempting, although inaccurate, to use these concepts also when studying unrooted gene trees. However, not all gene trees are genealogies of species and lineages due to well-known problems of hidden paralogy [19], coalescence [20,21], lineage sorting [22], lateral gene transfer [1], and integrative evolution, affecting molecular evolution at multiple levels. Consistent with this, Wilkinson *et al.* [23] pointed out the misapplication of terminology initially defined to characterize relationships on rooted trees. They notably introduced two notions to describe unrooted phylogenies: the terms 'clans' and 'adjacent groups' that should be used as respective analogs of 'clades' and 'sister groups' employed to interpret rooted topologies (Glossary). Clans are not clades and adjacent groups are not sister-groups, because, in the absence of a root, it is impossible to decide which node is ancestral and which node is derived. Moreover, there are up to three times as many adjacent groups as there are sister groups, and thus many adjacent groups do not come from one last common ancestor. These conceptual distinctions to analyze rooted versus unrooted trees are more than merely semantic progresses. They raise a fundamental issue: rooted trees and unrooted trees cannot be – and thus should not be – interpreted with identical concepts.

Although gene trees can be rooted by various approaches (Box 1), phylogenomic studies by definition produce unrooted trees, of which an overwhelming number might not be polarized in time, because the vast majority of prokaryotes and mobile elements sequenced within the context of genomic and metagenomic projects [24] have probably undergone some LE. Although various methods have been proposed to estimate lateral gene transfer (LGT) [25], the exploitation of such unrooted trees would be limited if adopting only the current 'organismal' phylogenetic thinking. Typically, methods used to deal

with gene trees derived from coalescent theory [22,26] would be inappropriate when not all the genes under study are expected to come from a last common ancestral genome. Analyses of unrooted trees with supernetworks summarizing the phylogenetic signals contained in the phylogenetic forest [27], or supertree and supermatrices reconciling or averaging these signals, can also be problematic [28,29]. Specifically, phylogenetic networks remain difficult to interpret because their edges still reflect three different relationships (xenology, paralogy and orthology), and not all their nodes correspond to ancestors [30]. The use of a supertree or supermatrices, as a baseline against which LGTs could be mapped [31], and some polarization attempted, faces at least three difficulties. First, unrooted trees with different samplings of operational taxonomical units (OTUs) are not exploited in a balanced way. Pruning steps increase the correspondance between gene trees [14], and dramatically underestimate the conflict in the data. Thus, these analyses seem little suited for sets of genomes with a minimal overlap in gene content (e.g. microbial genomes at large evolutionary scale and mobile elements). Second, when multiple copies of a given OTU are distantly located on an unrooted tree, it must be decided what copy (if any) reflects the organismal history. This decision is always difficult, if not impossible for microbial data and for mobile elements. Third, if lateral signals obfuscate the vertical signal in molecules, the baseline tree proposed might not be representative of the organismal history, and further inferences on evolution can be misleading [32].

There is thus major room for improvement in addressing the growing number of challenges created by unrooted gene trees. Here, we extend the efforts of Wilkinson *et al.* [23] to avoid a misapplication of the phylogenetic concepts developed for rooted trees to unrooted trees. Moreover, we propose an additional perspective and new phylogenetic concepts to gain information on the variety of relationships affecting the different evolutionary units (genes, modules, organisms and communities) found in unrooted gene trees. We demonstrate how this multi-level perspective called clanistics could increase our knowledge of evolutionary processes, notably in regards to LE between genetic partners, that affect most evolving systems.

A new terminology for clanistics

When acknowledging the fact that branches, nodes and groups have different biological meanings in rooted and unrooted trees (Box 2), one needs to define a new terminology for unrooted trees.

Slices of unrooted trees

One interesting property of the clades defined on rooted trees is that they can be intersected. Given two clades A and B on a same tree, either clade A is included in B, clade B is included in A, or A and B are disjoint clades [33]. Let us now focus on the example of clans A and B on an unrooted tree. It is easy to show that either clan A is included in B, clan B is included in A, the two clans are disjoint, or that their intersection gives rise to another relationship (C). For example, if the clans {1, 2, 3, 4, 5, 6} and {5, 6, 7, 8, 9, 0} in Figure 1a are intersected, the resulting group is {5, 6}. In this particular instance, {5, 6} is also a clan. However, the intersection of the clans {1, 2, 3, 4, 5, 6, 7, 8} and {5, 6, 7, 8, 9, 0} is {5, 6, 7, 8}; this group is not a clan and it cannot be converted into a clade by any rooting of the tree. We call this new type of relationship a 'slice'. Whereas clans are identified by splits (bipartitions) of an unrooted tree, slices are defined by pairs of splits, creating tripartitions of OTUs. Indeed, the split 56 | 123437890 uniquely defines the group {5, 6}, but two splits are required to define the group {5, 6, 7, 8}. They are the bipartitions 1234 | 567890 and 12345678 | 90 that produce the tripartition 1234 | 5678 | 90, where {1, 2, 3, 4} and {9, 0} are clans, whereas {5, 6, 7, 8} is a slice.

Box 2. Fundamental differences in the interpretation of rooted and unrooted trees

Phylogenetic trees are usually depicted as rooted trees with terminally labeled nodes. Such a representation graphically depicts the ancestor–descendant relationships among operational taxonomic units (OTUs, such as species, populations, or any other taxa), represented at the tips (the leaves) of the phylogeny. Internal nodes are seldom labeled in phylogenies, if only to represent hypothetical ancestors. Phylogenetic trees can be weighted to represent the divergence of OTUs in time, shorter branches representing less divergent OTUs than longer branches. The root of phylogenetic trees (either explicit or implicit) is unique and imposes a strict framework to identify any types of relationships among the OTUs.

Strictly speaking, a root is not different from any other node, except that it also implies an ordering (temporal) of the internal nodes, from the root to the leaves. In rooted gene trees, internal nodes represent speciation events between two lineages, or duplication events. Furthermore, because branches are oriented, mutations only accumulate along each branch away from the root, with increasing divergence. Consequently, the only way to travel along the path between any pair of leaves is downward towards the last common ancestor (i.e. back in time) and then upwards towards the leaves. A clade is a monophyletic group of OTUs that includes a common ancestor and all of its descendants, whereas a paraphyletic group includes OTUs that share a common ancestor but only some of its descendants. Sister groups are taxa that are each other's closest relatives; that is, they share a most recent common ancestor in a phylogeny.

The interpretation of relationships such as monophyletic, paraphyletic and sister groups, as well as of a node, a branch, or a path, is greatly influenced by the presence (and the position) of a root. Unrooted trees do not have an ordering of internal and terminal nodes. All nodes are equivalent. In the absence of a root, some internal nodes represent the multilateral union of different lineages by lateral gene transfer or by recombination, and not only speciation and duplication events. Because branches are undirected, they not only depict the divergence between a pair of nodes, but can also represent the bilateral fusion of DNA of unrelated taxa. Paths are time-reversible in unrooted trees. For any internal branch, there exists a nontrivial split (or bipartition) of OTUs in two complementary clans (*sensu* Wilkinson *et al.* [23]), whereas terminal branches define trivial splits associated with a clan comprising only one OTU, and another clan comprising all the others. Similarly, adjacent groups represent the unrooted analog of sister groups, but although there is only one sister group for any clade, there is up to three distinct adjacent groups for any clan. Thus, many adjacent groups do not share one last common ancestor. Being 'related' in rooted and unrooted trees thus has a distinct meaning, and whether adjacent groups are more closely related to each other than to other groups is ambiguous. To resolve this difficulty, additional criteria based on path-length distances could be used to determine adjacent-group relationships (e.g. the shortest path or sum of path-lengths between groups could identify closest adjacent groups, either based on the global similarity of their members or on their most basal, and presumably most ancient, divergence).

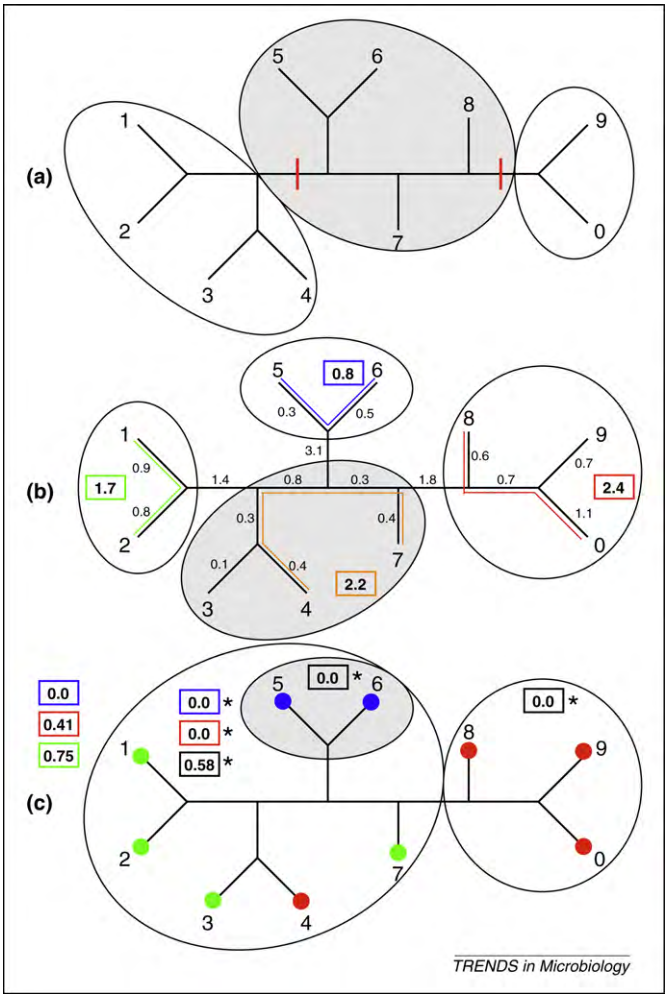


Figure 1. Illustration of clanistic concepts for unrooted trees. **(a)** A single bipartition of the tree defines a clan (encircled in white), whereas pairs of bipartitions (in red) are used to define a slice (encircled in grey). In this example, the groups {1, 2, 3, 4} and {9, 0} are clans, whereas {5, 6, 7, 8} is a slice. **(b)** Branch lengths are used to define clips based on phenetic criteria. Here, all clips with diameter smaller than s = 2.5 are presented. The different colors illustrate the longest path within each clip, with corresponding values in a square. Three of these clips are clans (encircled in white) and one clip {3, 4, 7} (encircled in grey) is neither a clan nor a slice. **(c)** Three colors are used to depict different categories of OTUs. The clan {8, 9, 0} is homogeneous because it contains elements of a single category. The clan {1, 2, 3, 4, 5, 6, 7} is complete because it contains all of the elements of a given category (here green); the clan {5, 6} is a perfect clan because it is both complete and homogeneous. The Equitability (E) and Intruders-Equitability (E*) values are presented in squares of the corresponding colors. The E* value in the black square is for all categories of intruders.

Clips of unrooted trees

One wishful assumption of phylogenetics is that if evolutionary rates are similar in different taxa, topological proximity could be correlated with phenetic distances, or that within-group distances could be smaller that between-group distances [34]. In other words, clans (or clades) could not only identify groups of OTUs that are related in terms of splits, but also in terms of path-length distances. Numerous studies comparing gene trees have shown that this assumption is often violated (because of heterotachy or departure from the molecular clock [19]). Nevertheless, groups of OTUs that exhibit similar divergence could be characterized to obtain some additional information on top of the topological relationships. To capture this type of information, we define a 'clip' as a group of OTUs for which all pairwise path-length distances are smaller than a given threshold value s. Figure 1b depicts a weighted tree with OTUs evolving at different rates. Fixing s at 2.5 allows us to define four clips on this tree, three of which are clans. However, the last clip {3, 4, 7} is neither a clan, nor a slice. In fact, clips group together OTUs that share similar divergence, not common splits.

Diversity indices for unrooted trees

A most interesting application of clanistics is the detection of relationships between genetic partners caused by LE. For the sake of the demonstration, let us consider that the leaves of a tree represent three categories of OTUs depicted by different colors in Figure 1c. Such categories can represent different taxonomical groups, samples from different environments, different functional or ecological aspects, distinct types of mobile genetic elements, and so on. In a world structured around these categories, one would expect these three types of OTUs to be clustered in distinct homogeneous clans. Namely, a clan (e.g. {8, 9, 0}) is said to be homogeneous when it contains OTUs of a single type (with no loss of generality, this definition also applies to slices or clips). Otherwise the clan is heterogeneous (e.g. {1, 2, 3, 4}): all OTUs of different types distinct from the dominant type within this clan are called intruders. A clan is also said to be complete if it includes all of the OTUs of a given type present in the tree. For example, the clan {5, 6} is complete for the blue color, whereas the clan {8, 9, 0} is incomplete for the red color. We define a smallest complete clan as a clan of minimum size that includes all OTUs of a given type (e.g. {1, 2, 3, 4, 5, 6, 7}); and a largest homogeneous clan as a clan of maximum size that only includes OTUs of the same type (e.g. {1, 2}). Optimal separation among OTUs is obtained when all the clans are complete and homogeneous (no intruders) at the same time. We call such groups 'perfect' clans.

To account for the distribution of OTUs of different categories among the clans (slices or clips) of an unrooted tree, a measure of dispersion is necessary. For instance, the Shannon diversity index (H) [35] could be applied to the different categories depicted on the tree. A diversity of 0 indicates that all OTUs of a given type are in the same clan, whereas positive values indicate a fragmented dispersion of OTUs from the same type in different clans (Figure 1c). To allow for the comparison of trees of various sizes, the corresponding Equitability index (E) is computed by dividing H by $\log(n)$, its maximum possible value for a tree with n OTUs. Maximum equitability (E = 1) thus corresponds to trees in which all OTUs of a given type are in separate clans.

As defined, these indices are computed on the tree as a whole, but they can also apply to parts of the tree (i.e. subtrees defined by clans, slices and clips). Furthermore, it is particularly telling to compute the topological distribution of intruders of different types (e.g. red and blue) in the smallest complete clan of a given type (e.g. green; Figure 1c). We call these restricted indices Intruders-Diversity (H*) and Intruders-Equitability (E*). A null value of E* indicates that all intruders are inserted as a single clan (slice or clip) within the smallest complete clan (slice or clip) of another type. The larger the values of E*, the less clustered the intruders. E* can be computed for all

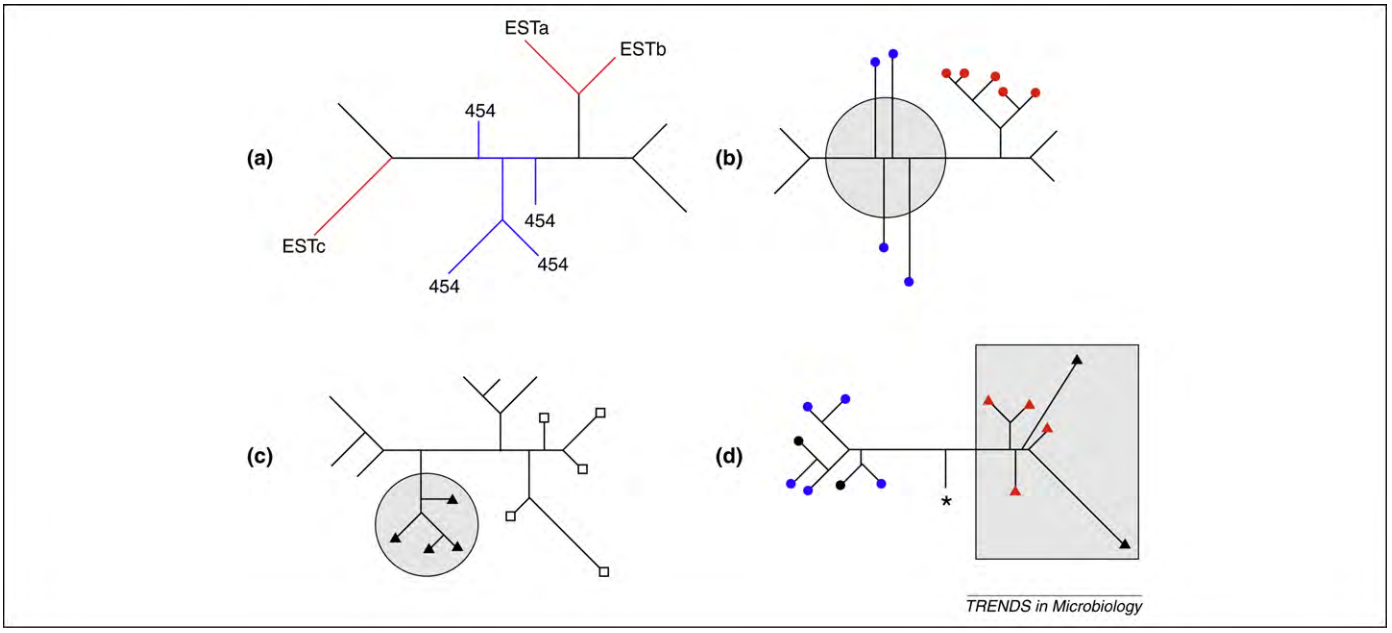


Figure 2. Hypothetical unrooted trees with meaningful patterns. **(a)** Branches in red or blue identify ESTs or 454 sequences from a given sequencing project. ESTa and ESTb fall in the same clan, and all the 454 sequences fall in the same slice. **(b)** Red nodes indicate OTUs violating the evolutionary model of the tree reconstruction method, suspiciously grouping in a same clan. Blue nodes correspond to fast-evolving taxa, falling in the same slice but a different clip (whose limits are represented by a grey circle), suggesting possible issue of Long Branch Attraction between these OTUs. **(c)** Triangles indicate taxa thriving in gut microflora, falling in a homogeneous clan and a clip (grey circle), suggesting that these taxa undergo a comparable selective pressure for this gene. Squares correspond to taxa of the gut microflora, for which closely related genes, comprising a clan, evolve at distinct rates, and thus do not fall in a clip. **(d)** Blue circles correspond to bacteria from a lineage *i*, black circles correspond to bacteria from a lineage *j*. All belong to a heterogeneous clan, and to a clip with a very small diameter *s*, with a positive E* value. This pattern suggests that these OTUs repeatedly exchanged conserved copies of an adaptive gene. Red triangles correspond to taxa of lineage *k*, and black triangles to taxa of lineage *l*. These form an heterogeneous clan (grey square) in which a long branch separates OTUs of a different kind, suggesting the exchange of a gene over long taxonomical distance or an accelerated evolutionary rate of the exchanged gene.

intruders at once, or separately for each category of intruders. All these concepts of clanistics can be used to extract informative trees (i.e. with meaningful patterns) from the gigantic forest of unrooted trees.

Meaningful patterns in unrooted trees

Meaningful (and statistics on recurring) clans, slices or clips identified in an unrooted gene tree (or a forest of such gene trees) offer precious insights for phylogenetic inferences from the most trivial to the least expected lessons.

First, phylogenetic inferences could benefit from the identification of homogeneous clans (slices) of OTUs with the same taxonomic affinities. Trees could be quickly sorted based on the topological distribution of such sets of OTUs: when they all fall in a perfect clan or slice ('homoclany' and 'homoslicy' as indicated by a null E value), the grouping of these taxa is unlikely to be affected by ancient paralogy. A converse situation (positive E value) would, by contrast, indicate that the OTUs of interest are evolving in multiple independent groups (clans or slices) rather than branching closely with one another. In practice, these estimates could help to improve the analysis of short fragments of DNA by classifying trees in which, for instance, short DNA fragments from a given 454 pyrosequencing [24,36] or expressed-sequence tag (EST) project branch together (Figure 2a), thus rapidly identifying which of these sequences could be assembled in contigs without affecting the overall phylogenetic relationships.

Second, homogeneous clans, slices or clips could help by detecting trees suffering from possible phylogenetic artifacts. OTUs that similarly violate the evolutionary model [37] used in the tree-reconstruction method (e.g. OTUs with significantly biased GC contents or odd substitution matrices [19,38]) will cluster in unusual clans or slices (as reported by a null E value for these taxa) (Figure 2b). Potential Long Branch Attraction artifacts [39] could be detected because they tend to group OTUs in a same clan (or slice) but in different clips.

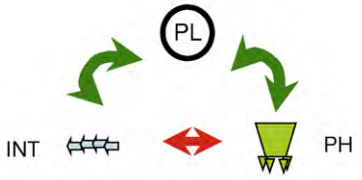
Third, clans, slices and clips grouping phylogenetically diverse OTUs can have biological causes; hence they can help in identifying candidate groups of genetic partners (Figure 2c). In this case, clanistics can be informative on questions of integrative biology – clarifying the rules of DNA transfer and recombination between these partners. Typically, homogeneous clans (slices or clips) explain the grouping between these partners by a shared biology, for example when microbes thrive in the same environment or community, have a close ecology or phenotype, or share some genomic or structural features (e.g. the presence or absence of CRISPRs [40] or of mobile elements [7,9,41]). Homogeneous clips also indicate what OTUs diverge in comparable proportion for a given gene, therefore suggesting that these genes evolve under shared selective pressures, and this could reflect a functional demand posed by the environment.

On the other hand, heterogeneous clans (slices or clips) that include intruders will be informative about the frequency of LE between OTUs of different types (Figure 2d). Heterogeneous clans with a high E* suggest frequent, independent, transfers of a given gene between the partners. As these exchanges are repeated, these genes are potentially adaptive for these OTUs. By contrast, heterogeneous clans

Box 3. Application of clanistics to mobile elements

A clanistic analysis of the genes from three types of mobile elements – phages and plasmids from the NCBI database and integron gene cassettes from the ACID (annotation of cassette and integron data) database [41] – produced 2177 unrooted gene trees (analyzed with Phyml [42], with a Γ law, 4 categories, and 100 bootstrap replicates) with at least two types of mobile elements on their leaves. These trees correspond to the evolutionary histories of the genes carried by one or more of these mobile elements known to date; to explain such trees at least one gene transfer had to take place between the different types of mobile elements. Clanistic analysis was used to investigate the relative frequencies of gene transfers between all these types of mobile elements, because rare versus frequent genetic transitions between them leave distinct topological signatures in unrooted trees. Rare transfers of genes between two types of mobile elements generate unrooted trees with low (or null) equitability values (E) for each type of mobile element because the OTUs of each type are all found in separated perfect clans. Intermediate rates of gene exchange between two types of mobile elements create unrooted trees, with heterogeneous clans and smallest complete clans with low E*, because the intruders (the OTUs of type *i* included in the smallest complete clan of type *j*) branch together, or almost so, indicative of a single or a few events of transition. Frequent gene exchanges between types of mobile elements also produce unrooted trees, with heterogeneous clans and smallest complete clans with high E*, because the intruders of type *i* repeatedly and independently branch in many places within the smallest complete clan of type *j*. From these patterns, knowledge of the evolutionary processes can be obtained. For these data (Figure I), transition between phages (PH) and integrons (INT) were rare (red arrow), but both transitions between phages and plasmids (PL), and between integrons and plasmids, were frequent (green arrows). Thus, plasmids likely play a central role in the spreading of homologous genes between multiple types of mobile elements.

For unrooted trees with:	INT-PL-PH	INT-PH	INT-PL	PL-PH
% E = 0	79	100	71,6	55
% E > 0	21	0	28,4	45
% **E * = 0** // *%E * > 0*	**5,3** // *94,7*	-	**21,7** // *78,2*	**20** // *80*
% of transitions between types of mobile elements	41,1 (INT-PL-PH) 58,8 (INT-PL)	-	100 (INT-PL)	100 (PL-PH)



Frequencies of transitions between the three types of mobile elements

TRENDS in Microbiology

Figure I. Simple clanistic analysis of genes from the mobile elements PH, PL and INT.

with a null E* indicate rare transfers between partners of different types. Some barriers to LE are likely to exist between these occasional partners, questioning what mechanism limited their genetic interaction. When transfers involve changes in the evolutionary rates of the transferred or recombined molecules only, members of heterogeneous clans will also likely belong to different clips (Figure 2d). By contrast, OTUs falling in the same heterogeneous clan and clip tend to exchange conserved versions of the gene (Figure 2d). Thus, two cases of special interest could be distinguished in unrooted trees: (i) radiative – direct – adaptation by exchange of a gene between genetic partners, when the inner branches of the heterogeneous clan (or slice) are short, and (ii) gene recruitment from a distinct lineage or significant tinkering of the transferred sequence(s) when the branch(es) leading to the clan (or the slice or to their intruders) are long. Groups of genes with similar patterns, possibly belonging to modules, can thus be easily identified in a forest of unrooted trees. Such notions could notably help in exploiting the evolution of mobile elements. Because plasmids, phages and integrons present a partly overlapping gene pool, their gene trees do not follow an 'organismal' phylogeny. Even so, these complex unrooted trees are informative (Box 3).

Conclusion

Clanistics promotes a multi-level perspective to analyse the mass production of incongruent, unrooted, or possibly unrootable gene trees, the typical output of microbial genomic and metagenomic projects [1–4]. Its ambition is to provide a conceptual and practical framework to infer a greater number of relationships, and to gain knowledge on a greater number of processes, than the usual exploitation of rooted trees, centered on the quest of organismal relationships. This perspective assumes that multiple evolutionary units are suitable and needed for a proper understanding of the evolution of most microbial genomes. These units, from smallest to largest: genes, groups of genes, organisms, and communities, can be harvested from unrooted gene trees, thanks to clanistic concepts highlighting their origin and their evolution. In particular, the notions of intruders and intruder diversity should help in the identification of genetic partners (OTUs that exchange genetic material and sometimes genetic modules). Likewise, the notion of clips will capture sets of OTUs subjected to similar selective pressures. Ultimately, clanistic analyses should teach us what genes are transferred between what genetic partners, or environments, and how often. Consequently, clanistics has the potential to offer a fine-scale description of the impact of the life style and ecology of OTUs on their genomic composition. Notably, it could determine what proportions of the microbial genomes (and of the genomes of mobile elements) depend more on their genetic interactions with unrelated partners than they reflect the 'organismal' genealogy. Should clanistics show that in some instances integrative evolution better explains biological diversity than common descent does, evolutionary biology will have to go through one more conceptual revolution.

References

1 Bapteste, E. *et al.* (2009) Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* 4, 34

2 Dagan, T. and Martin, W. (2009) Getting a better picture of microbial evolution *en route* to a network of genomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2187–2196
3 Doolittle, W.F. (2009) The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2221–2228
4 Ragan, M.A. *et al.* (2009) The network of life: genome beginnings and evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2169–2175
5 Pedulla, M.L. *et al.* (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* 113, 171–182
6 Brussow, H. (2009) The not so universal tree of life or the place of viruses in the living world. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2263–2274
7 Lima-Mendez, G. *et al.* (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* 25, 762–777
8 Brilli, M. *et al.* (2008) Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics* 9, 551
9 Norman, A. *et al.* (2009) Conjugative plasmids: vessels of the communal gene pool. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2275–2289
10 Fricke, W.F. *et al.* (2008) Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. *J. Bacteriol.* 190, 6779–6794
11 Hanage, W.P. *et al.* (2005) Fuzzy species among recombinogenic bacteria. *BMC Biol.* 3, 6
12 Dagan, T. *et al.* (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10039–10044
13 Rohwer, F. and Thurber, R.V. (2009) Viruses manipulate the marine environment. *Nature* 459, 207–212
14 Bapteste, E. and Burian, R.M. (2010) On the need for integrative phylogenomics – and some steps toward its creation. *Biol. Philos.* doi:10.1007/s10539-010-9218-2 (http://www.springerlink.com/content/102856/)
15 Morgan, G.J., (2010) Evaluating Maclaurin and Sterelny's conception of biodiversity in cases of frequent, promiscuous lateral gene transfer. *Biol. Philos.* doi:10.1007/s10539-010-9221-7 (http://www.springerlink.com/content/102856/)
16 Charlebois, R.L. and Doolittle, W.F. (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* 14, 2469–2477
17 Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366
18 Felsenstein, J. (2004) *Inferring Phylogenies*, Sinauer Associates
19 Gribaldo, S. and Philippe, H. (2002) Ancient phylogenetic relationships. *Theor. Popul. Biol.* 61, 391–408
20 Rosenberg, N.A. and Nordborg, M. (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3, 380–390
21 Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340
22 Degnan, J.H. and Rosenberg, N.A. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2, e68
23 Wilkinson, M. *et al.* (2007) Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol. Evol.* 22, 114–115
24 Hugenholtz, P. and Tyson, G.W. (2008) Microbiology: metagenomics. *Nature* 455, 481–483
25 Beiko, R.G. and Ragan, M.A. (2008) Detecting lateral genetic transfer: a phylogenetic approach. *Methods Mol. Biol.* 452, 457–469
26 Degnan, J.H. *et al.* (2009) Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.* 58, 35–54
27 Holland, B.R. *et al.* (2008) Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol. Biol.* 8, 202
28 de Queiroz, A. and Gatesy, J. (2007) The supermatrix approach to systematics. *Trends Ecol. Evol.* 22, 34–41
29 Ren, F. *et al.* (2009) A likelihood look at the supermatrix–supertree controversy. *Gene* 441, 119–125
30 Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267
31 Beiko, R.G. and Ragan, M.A. (2009) Untangling hybrid phylogenetic signals: horizontal gene transfer and artifacts of phylogenetic reconstruction. *Methods Mol. Biol.* 532, 241–256
32 Boucher, Y. and Bapteste, E. (2009) Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective. *Bioessays* 31, 526–536
33 Bosibud, H.M. and Bosibud, L.E. (1972) A metric for classifications. *Taxon* 21, 607–613
34 Estabrook, G.F. (1986) Evolutionary classification using convex phenetics. *Syst. Zool.* 35, 560–570
35 Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423 and 623–656
36 Pennisi, E. (2007) Metagenomics. Massive microbial sequence project proposed. *Science* 315, 1781
37 Keeling, P.J. *et al.* (2000) Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol. Biol. Evol.* 17, 23–31
38 Lockhart, P.J. *et al.* (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.* 93, 1930–1934
39 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410
40 Barrangou, R. *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712
41 Joss, M.J. *et al.* (2009) ACID: annotation of cassette and integron data. *BMC Bioinformatics* 10, 118
42 Guindon, S. *et al.* (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* 537, 113–137