Example Title

Rex Kwok

University of New South Wales, Australia
rkwok@cse.unsw.edu.au

Technical Report
UNSW-CSE-TR-0823
December 2008

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

Three main codes currently govern biological nomenclature: (i) The International Code of Botanical Nomenclature (ICBN), (ii) The International Code of Zoological Nomenclature (ICZN), and (iii) The International Code of Nomenclature of Bacteria (ICNB). Recently, the PhyloCode – a code based on phylogenetic nomenclature, has been presented as an alternative. To facilitate a comparison between the various codes, this paper presents a formal study into the properties of phylogenetic nomenclature – as presented in the PhyloCode. While much of the PhyloCode necessarily deals with the procedures for publishing and registering names, an important component deals with phylogenetic definitions. It is this component that will be studied in detail here. The various types of phylogenetic definition will be formalised in a mathematical setting. Results will be presented showing that under phylogenetic trees that much of the intuition surrounding phylogenetic definitions match up with the formalisation. However, ambiguity in the meaning of such definitions arises under the more general case when a phylogenetic hypothesis is allowed to be a rooted directed acyclic graph – a situation expressly allowed by the PhyloCode. Solutions to such problems will be presented. The issue of semantic stability – an often stated desirable property of a nomenclatural system – will also be examined. Conditions will be presented showing how stability can be improved for phylogenetic definitions. Two new types phylogenetic definition, the minimality–based definition and the maximality–based definition, will be presented as generalisations of the PhyloCode definitions. How the PhyloCode definitions relate to each other will be shown from this new perspective.

Keywords: phylogenetic nomenclature, biological nomenclature, PhyloCode, clade

1 Introduction

Since the work of Linnaeus in the mid 18^{th} century, biological organisms have been classified by placing them in a balanced taxonomic tree. In *Origin of Species*, Darwin [8] argued that biological classification should, and to some extent does, reflect the recency of common ancestry. Under the Linnaean taxonomy, this means that organisms which share a recent common ancestor are grouped closely while organisms which are distantly related are grouped far apart. Biological nomenclature is currently governed by: (i) The International Code of Botanical Nomenclature (ICBN), (ii) The International Code of Zoological Nomenclature (ICZN), and (iii) The International Code of Nomenclature of Bacteria (ICNB). These codes specify the correct procedures for naming and recognising the correct name of a taxon (a collection of organisms). Whether a taxon is interesting, important, or even what a taxon contains is not specified by the codes; genetic relatedness is not necessarily reflected in taxonomy. Recently, the *PhyloCode*[6] has been proposed as an alternative – to be used concurrently, or instead of, the preexisting nomenclatural codes. The PhyloCode seeks to implement Darwin's views by restricting the contents of taxa to *clades.* Under Article 2.1 of the PhyloCode, a clade is:

> "an ancestor (an organism, population, or species) and all of its descendants." [6]

In recent years there has been a lively debate [1, 5, 9, 20, 2, 4, 12] about the relative merits of the PhyloCode and the current systems of nomenclature. This debate is far from being resolved. Two highly related questions in this debate are: (i) what should be the basis for biological classification?, and (ii) what properties should a classification scheme possess? The first issue is difficult to resolve because it involves a degree of personal preference. For instance, Brummitt [3] is in favour of a classification system which places "like with like" and measures diversity. Those in favour of the PhyloCode wish to use relatedness through descent as the basis for classification. A choice of basis does, however, imply that a classification system will possess a number of properties. An often stated desirable property is called *stability.* The exact meaning of this has been the subject of some debate [1, 3, 4] because the notion has never been formalised. The concept can refer to the stability of the name for referring to some set of organisms or it can refer to the stability of the organisms referred to by a name. A thorough formal and mathematical analysis of nomenclatural systems would alleviate much of the misunderstanding about definitions and properties. This paper will contribute by formalising and proving a number of the properties of the PhyloCode.

The PhyloCode is divided into a series a rules, recommendations, and notes. The division reflects the differing importance of different parts of the PhyloCode. Rules are mandatory while notes are for "clarification". As might me expected, recommendations lie between rules and notes in importance. The are a number of ways in which clades can be specified using *phylogenetic definitions.* Interestingly, some aspects of clade definitions – notably, the definition of a clade and the exemplar organisms used in definitions – are governed by rules while the various kinds of phylogenetic definitions are listed in the extensive Note 9.4.1. It is exactly these kinds of phylogenetic definitions which are most suitable for formal logical analysis. This paper will present a formalisation of the framework for phylogenetic nomenclature. From this, the various types of phylogenetic definition will be studied showing whether the properties informally ascribed to them are reflected in a formal setting. Stability – in terms of the organisms which a name refers to – will be one property which will be given special attention. Where applicable, some criticism will be presented showing how phylogenetic definitions, and the PhyloCode itself, can be improved. The next section will present the details of the framework. Since virtually all examples of phylogenetic hypotheses from the literature are trees, this will be initial assumption carried into the analysis of phylogenetic definitions. In a series of 5 sections, one for each main type of phylogenetic definition, results will demonstrate a close match between intuition and formalism. It turns out that all the PhyloCode definitions are instances of two new simple and elegant phylogenetic definitions. Along with the presentation of these two new definitions will be an illustration of how the different types of PhyloCode definition are related to each other. The PhyloCode allows phylogenetic hypotheses to be rooted directed acyclic graphs to cater for hybridization and endosymbiosis. In Section 9 it will be shown that this generalisation introduces ambiguity into the meaning of phylogenetic definitions. Solutions to such problems will be proposed. The paper ends with a summary and a discussion.

2 Preliminaries

The PhyloCode is ambiguous on the issue of what objects the code covers. More specifically, the ambiguity is over granularity. In some parts there are indications that the PhyloCode is applicable over the space of individual organisms while in others, the space of species. In the preamble to Version2b, the PhyloCode is "applicable to the names of all clades" while species will only be applicable in future versions. In the meantime, the PhyloCode will rely on existing nomenclatural codes for the naming of species. This suggests that species are the most atomic object under consideration. However, the definition of a clade under Article 2.1 mentions organisms and populations. This would imply that currently extant species are clades and their naming governed by the current PhyloCode. Moreover, since an ancestor can be an individual organism, any currently living organism without any descendants would constitute a clade. The intention of the PhyloCode authors seems to be that species are the objects to be classified; certainly all examples contained in the PhyloCode only refer to species. The assumption adopted in this paper is that species are the atomic objects under consideration.

The assumption that species are atomic solves another problem with the PhyloCode. Recall that a clade consists of an ancestor and all of its descendants. Article 2.1 of the PhyloCode allows an ancestor to be either an organism, a population, or a species. This freedom of choice is problematic. A clade is meant to be monophyletic, *ie.*, a set "consisting of an ancestor and all of its descendants" by the glossary in the PhyloCode. Since the smallest object for classification mentioned in the PhyloCode is the *organism*, it stands to reason that the set of specifiers S should at least refer to all organisms; populations and species would then be represented by sets of organisms. If, however, the ancestor of a clade is a *species*, then a clade is not monophyletic because the *source* of a clade consists of many organisms. The problem here is an abstraction problem.

What is actually being classified? All organisms? Only species? Either answer is acceptable. However, if all organisms are being classified and a species is an acceptable ancestor, then the definition of "monophyletic" will have to be weakened to allow multiple sources – provided they are from the same species.

All phylogenetic definitions in the PhyloCode use examples to help determine what is or is not contained in a clade. Governed by Article 11, these examples are called *specifiers* in the PhyloCode. Generally, specifiers either refer to "species, specimens, or apomorphies" (derived character states that arise through evolution). As a minor departure from the PhyloCode, apomorphies will be defined separately since they refer to character states and not to organisms.

Definition 1 (Classification Frame) *A* classification frame C *is a pair* (S, O) *where* S *is called the set of* specifiers *and* O *the* omnigenus *such that* $S \subseteq O$.

Basically, the classification frame codifies what is to be classified. The set of specifiers contains all known species while the omnigenus contains all species that exist and have existed in the past – a finite set. In time, elements of O may be added to S as fossils and extant specimens are found and become available to science. In the meantime, phylogenetic definitions may only refer to elements in S. Any element of S will be called a *specifier* while any element of O will be called a *name*.

One of the main aims of the PhyloCode is to make all taxa monophyletic, *ie.*, the members of a taxon can be defined by an ancestor and all descendants of that ancestor. In order to test whether a taxon is monophyletic it is necessary to adopt a *phylogenetic hypothesis* that asserts how species are related to each other through descent. There are a number of methods for generating such hypotheses [17, 14, 13]. Even though the PhyloCode allows phylogenetic hypotheses to relate organisms in a directed acyclic graph, most published methods [17, 14, 13] organise a group into a tree. So, even though the generality of the PhyloCode allows for hybrids and endosymbiosis, it will be instructive to initally assume assume that a phylogenetic hypothesis is a tree.

Definition 2 (Phylogenetic hypothesis) *A* phylogenetic hypothesis H *on a set* O *is a binary relation on* O *such that the pair* (O, H) *constitute a tree.*

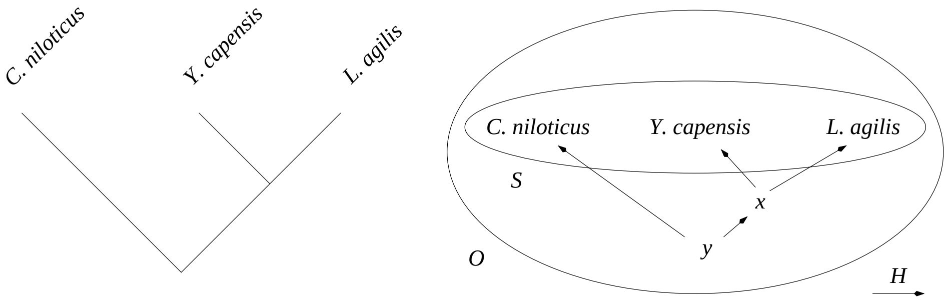

Figure 2.1: An example of a phylogenetic hypothesis.

As an example, consider the left hand side of Figure 2.1 which has been obtained from Example 2 in Article 11.9 of the PhyloCode. The PhyloCode example places three species – *Crocodylus niloticus*, *Youngina capensis*, and *Lacerta* *agilis* – in a tree as seen in Figure 2.1. This will be reflected in a classification scheme $C = (S, O)$ and a phylogenetic hypothesis H on O where S contains the three species and O contains S plus two theoretical species x and y. These theoretical species are names for the internal nodes in the phylogenetic tree. They also make explicit the implicit claims that are shown in the phylogenetic tree. Placing *Y. capensis* and *L. agilis* in a sub–tree without *C. niloticus* makes the implicit claim that the most recent common ancestor of *Y. capensis* and *L. agilis* – here named x – is a descendant of the most recent common ancestor of *C. niloticus* and *Y. capensis* (alternately *L. agilis*) – here named y. The phylogenetic hypothesis H on O encodes this ancestor–descendant information. Therefore, pairs such as (y, x) and (x, *Y. capensis*) are included in H. All this is depicted on the right hand side of Figure 2.1. It is important to note that there may be more theoretical species than just x and y. For instance, in the linear phyletic segment between x and *Y. capensis*, there can be any number of species.

A clade, consisting of an ancestor and all of its descendants, is defined as follows:

Definition 3 (Clade) *Given a classification frame $C = (S, O)$ and a phylogenetic hypothesis H on O, a* clade *X with respect of C and H is a set where:*

1. *$X \subseteq O$.*
2. *X is closed under descent, ie., if $x \in X$ and there is a path[1], according to H from x to any $y \in O$, then $y \in X$.*
3. *X has a source, ie., there exists a $z \in X$ such that for every $x \in X$, if $x \neq z$ then x is a descendant[2] of z.*

Consider the set $X = \{x$, *Y. capensis*$\}$ and the classification frame depicted in Figure 2.1. This set would *not* be a clade since it violates condition 2 in Definition 3 – not all descendants are included; *L. agilis* is a descendant of x which is not an element of X. Now consider the set $Y = \{x$, *C. niloticus*, *Y. capensis*, *L. agilis* $\}$. This is not a clade because condition 3 is violated. There is no source for Y – both x and *C. niloticus* are elements of X, yet no common ancestor of x and *C. niloticus* is in X.

Phylogenetic definitions impose constraints on clades and the meaning of a definition is given by the clades that satisfy the constraints. It is possible that no clade will satisfy the constraints of a definition – a situation noted in Article 11.10 of the PhyloCode. In this case, the definition will be called *meaningless.* When a definition is satisfied by at least one clade it will be called *meaningful* and if this clade is unique, the definition will be called *unambiguous.* In the following sections, as each type of phylogenetic definition is formalised, the conditions under which each of these various cases arise will be examined.

3 Node-Based Definitions

A node–based phylogenetic definition specifies a clade by providing examples of specifiers that are definitely contained in the clade. Only the smallest clade

[1] Technically, a path from nodes x to y in a tree H is a sequence of edges $(x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n)$ where $x = x_1$ and $y = x_n$ and each edge $(x_i, x_{i+1}) \in H$.

[2] x is a descendant of z if and only if there is a path from z to x.

containing all the examples satisfy the node–based definition. The relevant portion of Note 9.4.1 in the PhyloCode says:

> A node-based definition may take the form "the clade stemming from the most recent common ancestor of A and B" (and C, D, etc., as needed) or "the least inclusive clade containing A and B" (and C, D, etc.), where A-D are specifiers (see Art. 11.1). A node-based definition may be abbreviated "clade (A and B)".

This is formalised as follows:

Definition 4 (Node-based definition) *Consider a classification frame* $C = (S, O)$*. A* node–based *clade definition is a two–placed function* $\mathrm{node\_clade}(A, H)$ *taking a set of specifiers* $A \subseteq S$ *and a phylogenetic hypothesis* H *on* O *to a set of clades such that a clade* $X \in \mathrm{node\_clade}(A, H)$ *if and only if*

1. $a \in X$ *for every* $a \in A$*.*
2. X *is minimal, ie., if for every other clade* Y *where* $A \subseteq Y$ *and* $Y \subseteq X$*, then* $X \subseteq Y$*.*

This definition contains several features which are tacit in the PhyloCode description. Firstly, it makes the role of the phylogenetic hypothesis explicit in the meaning of a clade definition. For instance, the phrase "most recent common ancestor of A and B" as used in the PhyloCode does not have a definite meaning until A and B are placed in the context of a particular phylogenetic hypothesis. Secondly, the PhyloCode preempts the fact that a node–based definition will have one, and only one, clade associated with the definition. As will be shown later, this is in fact the case. However, that is a consequence of the definition and not a component of it. Thus, a node–based definition is a two–placed function which maps a set of specifiers A and a phylogenetic hypothesis H to the smallest clades X which contain all specifiers in A. Such a clade X is said to *satisfy* the definition.

Node–based definitions can be thought of in the following way. Consider a set of specifiers A. Each specifier a in A will have a phylogenetic history which stretches from a to the root of the phylogenetic tree. The root of the tree is certainly a common ancestor of all specifiers in A. However, what is important for a node–based definition is the *most recent common ancestor* of all pairs of specifiers in A (in what follows $MRCA(a_1, a_2)$ will denote the most recent common ancestor of specifiers a_1 and a_2). This is because node–based clades are defined as the least inclusive clade containing all the specifiers. Designate this particular common ancestor node by b. Then the node–based clade defined by A is simply b and all the descendants of b. This is depicted in Figure 3.1.

A node–based definition has the affirming property that it is always meaningful. More than that, the meaning of a node–based definition is unambiguous; one and only one clade satisfies a node–based definition.

Proposition 1 *Consider a classification frame* $C = (S, O)$ *and a node-based clade definition* $\mathrm{node\_clade}(A, H)$ *where* $A \subseteq S$*. Under any phylogenetic hypothesis* H *on* O*,* $\mathrm{node\_clade}(A, H)$ *is meaningful. Moreover, a unique clade satisfies this definition, ie.,* $\mathrm{clade}(A, H) = \{X\}$ *for some clade* X *with respect to* C *and* H*.*

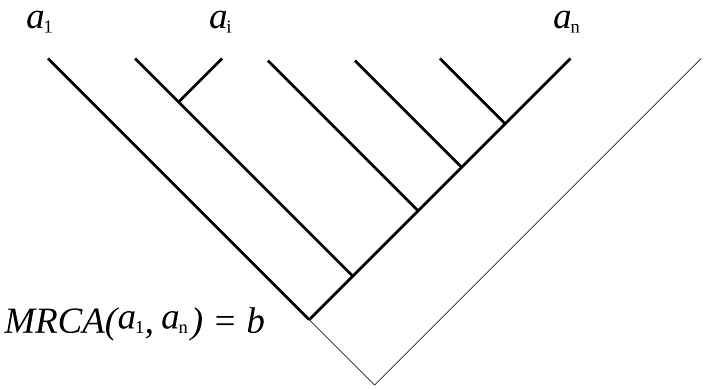

Figure 3.1: The semantics of a node–based clade definition. Given a set of specifiers A, a specifier b is identified as the most recent common ancestor of all specifiers in A. The clade then contains b and all descendants of b.

Since a node-based clade definition is alway meaningful and unique, it will be harmless and convenient to identify $\text{node\_clade}(A, H)$ with the clade that satisfies the definition. In what follows, this will be done for other clade definitions exactly when a definition is meaningful and unambiguous.

The specifiers used in a node–based definition $\text{node\_clade}(A, H)$ are called *internal* specifiers in the PhyloCode, *ie.*, a specifier "explicitly included in the clade". This is shown formally in the next result which shows that A is always included in the clade – irrespective of the phylogenetic hypothesis. Moreover, the commonality of a node–based definition across all phylogenetic hypotheses is exactly A.

Observation 1 *Consider a classification frame $C = (S, O)$ and a set of specifiers A. The intersection of X_H where* $\text{node\_clade}(A, H) = \{X_H\}$ *across all phylogenetic hypotheses H is A.*

Since the specifiers in a node–based definition are internal, more specifiers imply more inclusive clades.

Observation 2 *Consider a classification frame $C = (S, O)$, a phylogenetic hypothesis H on O, and two sets of specifiers A_1 and A_2. If $A_1 \subseteq A_2$, then*

$$\text{node\_clade}(A_1, H) \subseteq \text{node\_clade}(A_2, H).$$

While adding specifiers produces more inclusive clades, no change is effected if the specifiers come from the current clade in question.

Observation 3 *If A be a set of specifiers and $B \subseteq \text{clade}(A, H)$, then*

$$\text{node\_clade}(A, H) = \text{node\_clade}(A \cup B, H).$$

It may seem odd to add specifiers to a node–based definition when the clade remains the same and the complexity of the definition increases. However, such information actually increases semantic stability – in terms of the content of a clade. Recommendation 11D of the PhyloCode suggests that node–based definitions should include representatives from all subclades for which there is "credible evidence". The reason for this is that the meaning of a phylogenetic definition is dependant upon the phylogenetic hypothesis. Since the phylogenetic hypothesis is liable to change, it is favourable to have definitions which have the same semantics across more phylogenetic hypotheses. This is shown by the next result which shows that the addition of seemingly redundant specifiers makes the meaning of a node–based definition more stable.

Proposition 2 *Consider a classification frame* $C = (S, O)$*, a phylogenetic hypothesis* H *on* O*, and* node_clade(A, H) *for some set of specifiers* A*. Let* stable(node_clade(A, H)) *denote the set of phylogenetic hypotheses* $H^{'}$ *such that* node_clade(A, H) $=$ node_clade$(A, H^{'})$*.*

If $B_1 \subset B_2 \subseteq$ node_clade(A, H) *and* $B_1, B_2 \nsubseteq A$*, then*

stable(node_clade$(A \cup B_1, H)$) $\subset$ stable(node_clade$(A \cup B_2, H)$)*.*

As an illustration of semantic stability, consider the trees depicting the relationships between basal birds and therapod dinosaurs in Figure 3.2. The trees depict the same taxa but differ in the relative relatedness of certain taxa. Firstly, consider Figure 3.2(A). The node–based definition node_clade$(\{b, g\}, H)$ (lower case italic characters have been used as abbreviations for the various taxa) gives the clade that contains every taxa – from a to m; b belongs to one sub–tree at the lowest branch point while g belongs to the other sub–tree. Adding more specifiers to the definition, therefore, will have no effect on the meaning of the clade definition under the phylogenetic hypothesis in Figure 3.2(A). For instance, node_clade$(\{b, g, d\}, H)$ also includes all the taxa. However, when the phylogenetic hypothesis changes, to the one depicted in Figure 3.2(B) for example, the extra specifier d gives added semantic stability. In Figure 3.2(B), b and g are each other's closest relative – the opposite of the situation in Figure 3.2(A). Thus, the only specifiers included in node_clade$(\{b, g\}, H^{'})$ are b and g. However, the meaning of node_clade$(\{b, g, d\}, H^{'})$ is the same as node_clade$(\{b, g, d\}, H)$ since the most recent common ancestor of b and d resides at the lowest branch point.

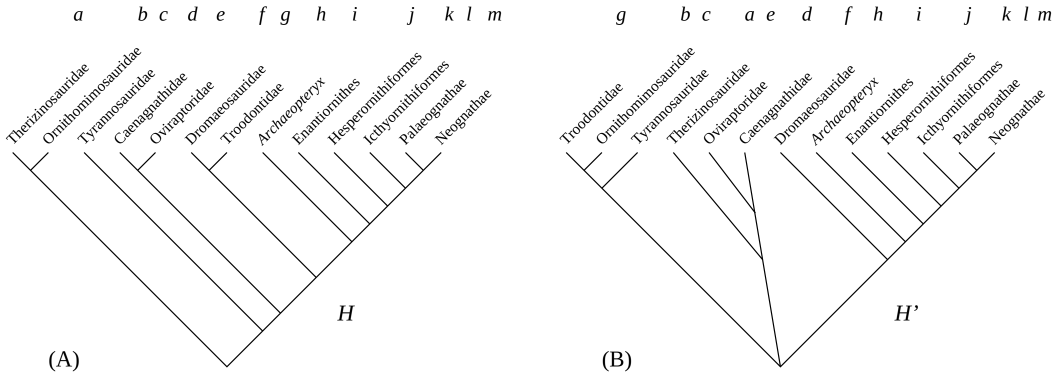

Figure 3.2: Cladograms of basal birds and dinosaur relatives based on the publications of Sereno [19] (A), Padian [18] (B), and Benton [1]. Italic characters are abbreviations for the various taxa.

4 Stem–based definitions

A stem–based phylogenetic definition specifies a clade by providing an internal specifier and a set of external specifiers; examples which are explicitly excluded from a clade definition. They are described in the PhyloCode as follows:

> A stem-based definition may take the form "the clade consisting of A and all organisms or species that share a more recent common ancestor with A than with Z" (and Y and X, etc., as needed) or "the most inclusive clade containing A but not Z" (and Y and X, etc.). A stem-based definition may be abbreviated "clade (A not Z)".

This is formalised as:

Definition 5 (Stem-based definition) *Consider a classification frame $C = (S, O)$. A* stem–based *clade definition is a function* $\mathrm{stem\_clade}(a, Z, H)$ *taking a specifier $a \in S$, a set of specifiers $Z \subseteq S$ and a phylogenetic hypothesis H on O to a set of clades such that a clade $X \in \mathrm{stem\_clade}(a, Z, H)$ if and only if:*

1. *$a \in X$.*
2. *$z \notin X$ for every $z \in Z$.*
3. *X is maximal, ie., for every other clade Y satisfying the above two conditions, if $X \subseteq Y$, then $Y \subseteq X$.*

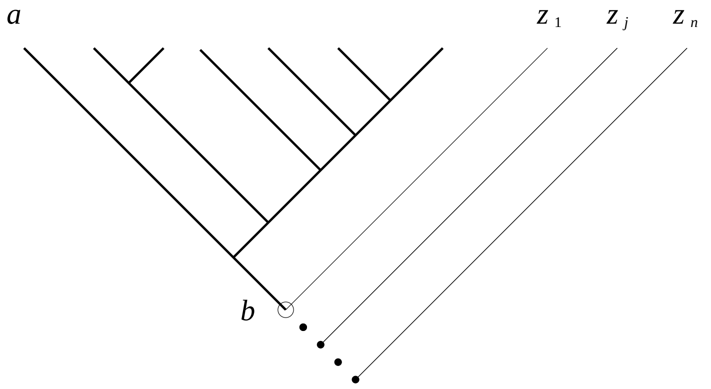

Figure 4.1: The semantics of a stem–based clade definition. Given a specifier a and a set of specifiers Z, a specifier b is identified as the most recent common ancestor of a and some $z_1 \in Z$. The clade then contains the subtree of b containing a. Note that this excludes b.

A stem–based definition has the form $\mathrm{stem\_clade}(a, Z, H)$ and is equivalent to the expression "clade(a not z and y and x, etc.)" in the PhyloCode where Z contains z, y, and x. The meaning of a stem–based definition can be understood by considering the MRCA of a and z for each $z \in Z$. These are all ancestors of a and one will be the youngest. The clade defined is the subtree containing a from this youngest ancestor. This is illustrated in Figure 4.1. In this figure it can be seen that a and each $z \in Z$ has a MRCA. One of these, b, the MRCA of a and z_1, is the youngest. From b the subtree containing a is what is contained in the clade. Note that b itself is not part of the clade. A clade is closed under descent and z_1 must not be contained in the stem–based clade.

Since a clade, by definition, is closed under descent, a stem–based definition is not meaningful when an external specifier is a descendant of the internal specifier. The converse is not true however. A stem–based definition can still be meaningful if the internal specifier is a descendant of an external specifier. This is shown formally in the following result.

Proposition 3 *A stem–based clade definition* $\text{stem\_clade}(a, Z, H)$ *where* $a \in S$ *and* $Z \subseteq S$ *is meaningful if and only if* $z \notin \text{node\_clade}(\{a\}, H)$ *for every* $z \in Z$. *Moreover, if a stem–based definition is meaningful, then it is unambiguous.*

Given, a stem–based definition $\text{stem\_clade}(a, Z, H)$, the set Z is intuitively called the set of external specifiers. This is because all elements of Z are not members of a stem–based clade; the larger Z is the more is excluded from the clade and the smaller the clade.

Observation 4 *Let* Z_1 *and* Z_2 *be sets of specifiers and* $a \in S$ *a specifier. If* $Z_1 \supseteq Z_2$, *then* $\text{stem\_clade}(a, Z_1, H) \subseteq \text{stem\_clade}(a, Z_2, H)$.

Stem–based clade definitions naturally contain a lot of redundancy for any fixed phylogenetic hypothesis.

Observation 5 *For any meaningful stem–based clade definition* $\text{stem\_clade}(a, Z, H)$ *there exists a* $y \in Z$ *such that* $\text{stem\_clade}(a, Z, H) = \text{stem\_clade}(a, \{y\}, H)$.

This is an immediate result following the definition of a stem-based clade. One simply looks for the youngest MRCA of a and some specifier from the exclusion set Z. With the example in Figure 4.1 this specifier is z_1.

In parallel with node–based definitions, seemingly redundant components of stem–based definitions increase semantic stability. Here, semantic stability is a relative concept. It refers to how invariable the meaning of a clade definition is in the face of changing phylogenetic hypotheses. Many authors have argued that this as a favourable property of a nomenclatural system [11, 1, 12]. The authors of the PhyloCode express this in Recommendation 11E:

> "In a stem–based definition, it is best to use a set of external specifiers that includes representatives of all clades that credible evidence suggests may be the sister group of the clade being named. Constructing a stem–based definition in this way will reduce the chance that, under a new phylogenetic hypothesis, the name will refer to a more inclusive clade than originally intended."

This intuition is proven formally in the following result.

Proposition 4 *Consider a classification frame* $C = (S, O)$, *a phylogenetic hypothesis* H *on* S, *and* $\text{stem\_clade}(a, Z, H)$ *for some specifier* a *and set of specifiers* Z. *Let* $\text{stable}(\text{stem\_clade}(a, Z, H))$ *denote the set of phylogenetic hypotheses* $H^{'}$ *such that* $\text{stem\_clade}(a, Z, H) = \text{stem\_clade}(a, Z, H^{'})$.

1. *If* $Z_1 \subset Z_2 \subseteq Z$ *and* $\text{stem\_clade}(a, Z \setminus Z_2, H) = \text{stem\_clade}(a, Z, H)$*, then* $\text{stable}(\text{stem\_clade}(a, Z \setminus Z_1, H) \supset \text{stable}(\text{stem\_clade}(a, Z \setminus Z_2, H)$*.*

2. *If* $Z_1 \subset Z_2 \subseteq S$ *and* $Z_2 \cap \text{stem\_clade}(a, Z \cup Z_2, H) = \emptyset$*, then* $\text{stable}(\text{stem\_clade}(a, Z \cup Z_1, H) \subset \text{stable}(\text{stem\_clade}(a, Z \cup Z_2, H)$*.*

As an example, consider once again the phylogenetic hypotheses depicted in Figure 3.2. Observation 5 shows that only one particular external specifier is needed to fix the contents of a stem–based clade under one phylogenetic hypothesis. However, under different phylogenetic hypotheses, the first part of Proposition 4 shows that removing specifiers from the exclusion set decreases semantic stability. For instance, $\text{stem\_clade}(h, \{g, f, c, a\}, H)$ in Figure 3.2(A) contains h=*Archaeopteryx* and everything to the right of it, *viz.*, $X = \{h, i, j, k, l, m\}$. Note that in this phylogenetic hypothesis that g is the closest relative of h in the exclusion set. This is not the case for Figure 3.2(B). So, if we consider $\text{stem\_clade}(h, \{g, f\}, H)$ (removing c and a from the exclusion set), this definition has the same meaning under both phylogenetic hypotheses. However, if one more specifier is removed, $\text{stem\_clade}(h, \{g\}, H)$ includes X in Figure 3.2(A) but generates a much more inclusive clade in Figure 3.2(B), including a,d, e, and f. The second part of Proposition 4 shows that adding external specifiers increases semantic stability. As an example, it is easy to see that $\text{stem\_clade}(h, \{g\}, H)$ and $\text{stem\_clade}(h, \{g, f\}, H)$ both have the same meaning. However, with a switch of phylogenetic hypotheses to $H^{'}$, the first definition $\text{stem\_clade}(h, \{g, f\}, H^{'})$ retains the same meaning while $\text{stem\_clade}(h, \{g\}, H)$, with fewer external specifiers, is a less inclusive clade than $\text{stem\_clade}(h, \{g\}, H^{'})$.

5 Apomorphy–based definitions

An apomorphy–based definition defines a clade through the first appearance of a character trait through evolution. These definitions are described in the PhyloCode as follows:

> An apomorphy-based definition may take the form "the clade stemming from the first organism or species to possess apomorphy M as inherited by A" or "the most inclusive clade exhibiting character (state) M synapomorphic with that in A." An apomorphy-based definition may be abbreviated "clade (M in A)".

Given a character or *apomorphy* m, as possessed by some specifier a, a clade is defined containing the first organism to possess apomorphy m and all of its descendants (see Figure 5.1). Descendants which subsequently lose apomorphy m are still included in the clade. For instance, the clade tetrapoda might be defined as the possession of digits by vertebrates as exemplified by a dog. The source of this clade would be a vertebrate with digits. However, descendants of this ancestor – such as whales and snakes – which do not possess digits are still members of the clade.

To make sense of an apomorphy–based definition it is necessary to have an *apomorphy hypothesis* which ascribes apomorphies to specifiers. This relation serves a parallel function to the phylogenetic hypothesis which defines the ancestor–descendant relationship.

Definition 6 (Apomorphy hypothesis) *Consider a classification frame $C = (S, O)$. An* apomorphy hypothesis *on S is a pair (M, P) where M is the set of* apomorphies *and P is a binary relation on $M \times S$ called the* apomorphy relation.

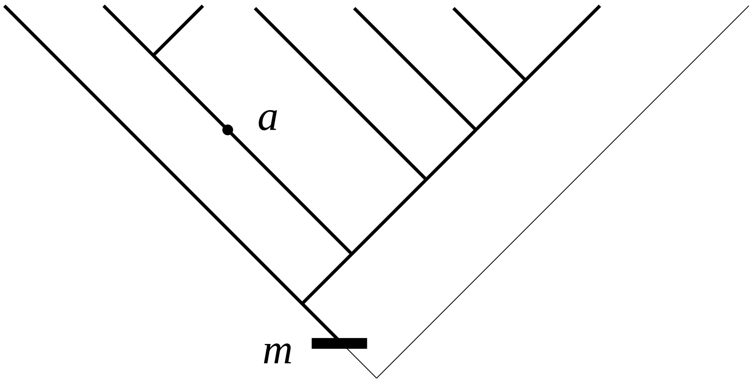

Figure 5.1: The semantics of a apomorphy–based clade definition. Given an apomorphy m as possessed by specifier a, the clade contains the first organism that possesses character m and all it's descendants.

The meaning of an apomorphy–based definition is most often taken as the clade stemming from the first organism to possess an apomorphy [10, 11, 12, 1]. However, as noted by Lee [16], this assumes that the first organism to possess an apomorphy has been identified. Moreover, there is the implicit assumption that all organisms that have been ascribed this particular apomorphy are descendants of this "first organism". The PhyloCode and Lee [16] present alternative interpretations. Lee suggests that an apomorphy–based clade is one which is "diagnosed by trait X" and the PhyloCode ' "the most inclusive clade exhibiting character (state) M synapomorphic with that in A." '. This alternative reading is formalised as follows:

Definition 7 (Apomorphy-based definition) *Consider a classification frame $C = (S, O)$, a phylogenetic hypothesis H on S, and a apomorphy hypothesis (M, P). An* apomorphy–based *clade definition is a function of the form* apomorphy_clade(m, a, H, M, P)*, where $a \in S$ is a specifier and $m \in M$ is an apomorphy, to clades X that satisfy:*

1. *$a \in X$.*
2. *$b \in X$ for every $(m, b) \in P$.*
3. *X is minimal, ie., for every other clade Y satisfying the above two conditions, $X \subseteq Y$.*

A clade satisfies an apomorphy–based definition apomorphy_clade(m, a, H, M, P) when it is the smallest set containing all organisms possessing m. The conditions for such a definition to be meaningful are thus trivial, *viz.*, that a should possess apomorphy m. Furthermore, a minimality condition fixes the semantics of the definition. Formalising apomorphy–based clades in this way obviates the need to identify the earliest organism to possess an apomorphy.

6 Stem–modified node–based definitions

Since most biologists study extant organisms [11, 7], many taxa important to biology have been studied based only on living representatives. In part, this is due to the fact that much more can be ascertained from extant organisms than extinct organisms. Catering for this kind of taxa in the PhyloCode is the concept of a *crown clade* – a clade where "both of the basal branches have extant representatives" Note 9.4.1. [6]. The PhyloCode presents two methods for defining crown clades: the stem–modified node–based definition and the apomorphy–modified node–based definition. The stem–modified node–based definition has the form:

$$\text{crown\_clade}(a, Z, H)$$

where a is an internal specifier and Z is a set of external specifiers. The PhyloCode describes the semantics of such a definition as the clade stemming from a and all extant organisms that more closely related to a than with any organism in Z. This can be seen graphically in Figure 6.1. Here z_1 is the organism most closely related to a in the exclusion set Z. From the most recent common ancestor of a and z_1, consider the stem leading towards a. This stem will eventually reach a point where both basal branches have an extant representative. In this case b. The most recent common ancestor of a and b, *viz.* y is the source for this crown clade.

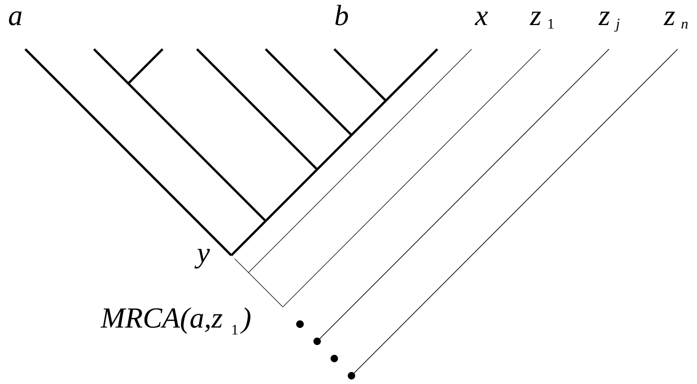

Figure 6.1: The semantics of a stem–modified node–based clade definition.

Definition 8 (Stem-modified node–based definition) *Consider a classification frame $C = (S, O)$ and a phylogenetic hypothesis H on O. A* stem–modified node–based *clade definition is a function* crown_clade(a, Z, H) *taking a specifier $a \in S$, a set of specifiers $Z \subseteq S$ and a phylogenetic hypothesis H on O to a set of clades such that a clade $X \in$* crown_clade(a, Z, H) *if and only if:*

1. $a \in X$.

2. $z \notin X$ *for every* $z \in Z$.

3. *if b is extant and $MRCA(a, b)$ is a descendant of $MRCA(a, z)$ (for any $z \in Z$), then $b \in X$.*

4. X is minimal, ie., for every other clade Y satisfying the above three conditions, $X \subseteq Y$.

Since a stem–modified node–based definition contains a stem–based component, the definition is meaningful if no external specifier is a descendant of the internal specifier.

Proposition 5 *A stem–modified clade definition* $\text{crown\_clade}(a, Z, H)$ *where* $a \in S$ *and* $Z \subseteq S$ *is meaningful if and only if* $z \notin \text{node\_clade}(\{a\}, H)$ *for every* $z \in Z$*. Moreover, if a stem–modified node–based definition is meaningful, then it is unambiguous.*

Parallelling stem–based definitions, a stem–modified node–based clade definition contains a lot redundancy for any fixed phylogenetic hypothesis. However, this redundancy increases the semantic stability of such definitions.

Observation 6 *A stem–modified node–based clade definition* $\text{crown\_clade}(a, Z, H)$ *where* $a \in S$ *and* $Z \subseteq S$*. For some* $z_i \in Z$*,* $\text{crown\_clade}(a, \{z_i\}, H) = \text{crown\_clade}(a, Z, H)$.

Consider a stem–modified node–based definition of a clade $\text{crown\_clade}(a, Z, H)$. Firstly, this is equivalent to the node–based definition $\text{node\_clade}(\{a\} \cup X_{a,Z}, H)$ where $X_{a,Z}$ is the set of all extant organisms which share a more recent common ancestor with a than with any member of Z. From Figure 6.1 it can be seen that $X_{a,Z}$ contains all extant organisms in the subtree containing a from $MRCA(a, z_1)$. This is not necessarily a clade as there may be extinct members of $\text{stem\_clade}(a, Z, H)$. Thus $X_{a,Z} \subseteq \text{stem\_clade}(a, Z, H)$.

Observation 7 $\text{crown\_clade}(a, Z, H) \subseteq \text{stem\_clade}(a, Z, H)$.

Observation 8 *All extant members of* $\text{stem\_clade}(a, Z, H)$ *are contained in* $\text{crown\_clade}(a, Z, H)$*.*

7 Apomorphy–modified node–based definitions

Some crown clades are defined by the possession of a certain apomorphy. The apomorphy–modified node–based definition captures such a clade by providing an extant example which possesses the apomorphy. Such a definition has the following form: $\text{crown\_clade}(a, m)$, for some specifier a and apomorphy m. There are two intuitive readings for the semantics of such a definition in the PhyloCode:

1. the clade stemming from the most recent common ancestor of A[a] and all extant organisms or species that possess apomorphy M[m] as inherited by A[a]
2. the most inclusive crown clade exhibiting character (state) M[m] synapomorphic with that in A[a]

Also by note 9.4.1 of the PhyloCode, a crown clade is a clade for which “both of the basal branches have extant representatives”. This implies that reading 1 will only be guaranteed to generate a crown clade if a is extant. As examples, consider the phylogenetic trees in Figure 7.1. Here, a, b, and c are all specifiers with apomorphy m. Suppose that b and c are extant while a is extinct. The tree on the right will not give a crown clade since one basal branch (the one leading to a) has no extant representatives. This is not the case with the tree on the left hand side even though a is not extant. When a is extant, reading 1 makes the clade definition equivalent to a node–based definition containing only extant specifiers. This will always give a crown clade.

By reading 2, crown clade(a, m) may not contain a; for similar reasons why reading 1 may not produce a crown clade. Thus given the phylogenetic tree on the right of Figure 7.1, reading 2 would produce a clade consisting of b and c but not a since a is not extant. Like reading 1, however, if a is assumed to be extant, then the problem disappears and a becomes a member of the clade. By reading 1, a is always a member of the clade.

Another difference between the two readings depends on the exact meaning of a "crown clade". The first reading essentially gives a node–based definition; the second an apomorphy–based definition. An apomorphy can originate from a bifurcation point in the phylogenetic tree. However, it may also originate in the middle of a linear segment of the tree. Node–based definitions typically (when there are two or more specifiers that are not in an ancestor–descendant relationship) define clades consisting of a sub–tree starting at a bifurcation point. This means that reading 2 can be more inclusive than reading 1. For instance, consider extant specifiers a and b possessing apomorphy m. A non extant specifier c also possesses m and is an ancestor of a and b. Suppose further that the apomorphy m originated with a specifier d which is an ancestor of c. By reading 1, the apomorphy–modified node–based clade would contain a, b, and c. By reading 2, it would also contain d.

Proposition 6 *If a is extant then* crown clade(a, m) *by reading 2 is a superset of* crown clade(a, m) *by reading 1.*

Given the ambiguities of apomorphy–modified node–based definitions, a thorough analysis of such definitions will not be presented here.

8 Generalising Phylogenetic Definitions

Phylogenetic definitions generate clades via three main constructors: internal specifiers, external specifiers, and a minimality–maximality switch. Apomorphy based clades and crown clades may appear to be generated by different means, but the concepts of *apomorphy* and *extant* are in fact means of generating internal specifiers. Two new types of phylogenetic definition will be presented which generalise existing phylogenetic definitions. The reasons for doing so are two–fold. Firstly, the new definitions are generalisations and will allow more clades

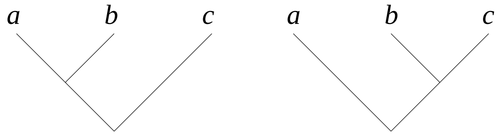

Figure 7.1: Two phylogenetic trees where a, b and c all possess apomorphy m. However, a is not extant while b and c are extant.

to be defined. Secondly, the relationship between existing types of definition will be made clear.

A node–based definition maps to the smallest clade containing a set of specifiers. Generalising this by adding a set of external specifiers results in the following minimality–based clade.

Definition 9 (Minimality-based definition) *Consider a classification frame* $C = (S, O)$. *A* minimality–based *clade definition is a function* $\text{min\_clade}(A, Z, H)$ *taking sets of specifiers* $A, Z \subseteq S$*, and a phylogenetic hypothesis* H *on* O *to a set of clades such that a clade* $X \in \text{min\_clade}(A, Z, H)$ *if and only if:*

1. $A \subseteq X$.
2. $z \notin X$ *for every* $z \in Z$.
3. X *is minimal, ie., for every other clade* Y *satisfying the above two conditions, if* $Y \subseteq X$*, then* $X \subseteq Y$.

The minimality–based clade definition covers node–based, apomorphy–based, and crown clade definitions. An immediate and obvious consequence of the above definition is that a node–based definition is generated when the set of external specifiers is empty. Thus, for any set of specifiers A, the node–based definition $\text{node\_clade}(A, H)$ is equivalent to $\text{min\_clade}(A, \emptyset, H)$. The point of adding external specifiers to a node–based definition, as with all other phylogenetic definitions, is that the meaning of the clade definition remains the same under more phylogenetic hypotheses.

An apomorphy-based definition is really a node–based definition in disguise. For any given specifier a exhibiting apomorphy m, the two interpretations of an apomorphy–based $\text{apomorphy\_clade}(m, a, H, M, P)$ are given in the PhyloCode. The first isolates a sole specifier b – the first organism to possess m. The second generates the set of specifiers A that contains all organisms possessing apomorphy m. These specifiers are the internal specifiers to a minimal clade and will generate the same clade provided it is possible to isolate b. Thus, $\text{apomorphy\_clade}(m, a, H, M, P)$ is the same as either $\text{node\_clade}(\{b\}, H)$ or $\text{node\_clade}(A, H)$. Either way, the apomorphy-based definition will be an instance of the minimality–based clade.

The PhyloCode lists two types of crown clade: the stem–modified node–based definition and the apomorphy–modified node–based definition. As their names suggest, they are variants of the node–based definition. In the first case, the notion of what is *extant* is used to generate a set of internal specifiers. With a stem–modified node–based definition $\text{crown\_clade}(a, Z, H)$, a set of internal specifiers A is generated which contains a and all extant specifiers b that share a more recent common ancestry with a than any specifier in Z. It can then be shown that $\text{crown\_clade}(a, Z, H)$ is the same as $\text{min\_clade}(A, Z, H)$. As shown in Section 7, there slight differences in the two PhyloCode interpretations of an apomorphy–modified node–based definition. Both cases, however, are covered by the minimality–based clade. Given a specifier a that possesses apomorphy m, the first interpretation generates a single specifier b that is the most recent common ancestor of a and all extant specifiers that possess m. The specifier b is the single internal specifier to a node–based definition. This would be equivalent to the minimality–based definition $\text{min\_clade}(\{b\}, \emptyset, H)$. The second interpretation generates a set of internal specifiers A – the set containing all extant specifiers that exhibit apomorphy m. In this case, the clade would be equivalent to min_clade$(A, \emptyset, H)$.

A stem–based definition stem_clade(a, Z, H) maps to the largest clade containing a but not any element of Z. Allowing for multiple internal specifiers generates the following maximality–based clade.

Definition 10 (Maximality-based definition) *Consider a classification frame $C = (S, O)$. A* maximality–based *clade definition is a function* max_clade(A, Z, H) *taking sets of specifiers $A, Z \subseteq S$, and a phylogenetic hypothesis H on O to a set of clades such that a clade $X \in$* max_clade(A, Z, H) *if and only if:*

1. *$A \subseteq X$.*
2. *$z \notin X$ for every $z \in Z$.*
3. *X is maximal, ie., for every other clade Y satisfying the above two conditions, if $X \subseteq Y$, then $Y \subseteq X$.*

The maximality–based definition is a simple extension of the stem–based definition. It is a straightforward observation that a stem–based definition stem_clade(a, Z, H) is equal to max_clade$(\{a\}, Z, H)$.

All phylogenetic definitions presented in the PhyloCode are instances of the minimality and maximality based definitions. Both contain a set of internal specifiers and a set of external specifiers. They only differ in generating a minimal or a maximal clade. Analogous with the results shown earlier, both definitions are meaningful if, and only if, no external specifier is a descendant of an internal specifier. Moreover, if a minimality or maximality based definition is meaningful, then it is unambiguous. Apomorphy and crown based definitions rely on added structure to generate internal specifiers; structure that defines what is extant and what organisms possess which apomorphy. There are most certainly other concepts which biologists may like to use which generate internal specifiers. There may even be concepts for generating external specifiers.

9 Phylogenetic Hypotheses as Directed Acyclic Graphs

The PhyloCode (Note 2.1.1 and Note 2.1.3) allows a phylogenetic hypothesis to be a rooted directed acyclic graph (DAG); catering for hybridization and endosymbiosis. In a tree everything has only one parent, whereas a DAG allows multiple parents. Moreover, since a DAG is a more general structure than a tree, some of the properties of a phylogenetic tree will not hold in this more general case. Importantly, it will be shown that the meaning of node–based and stem–based clades can become ambiguous. However, since it is beneficial to be able to represent hybrids and endosymbionts in a phylogenetic hypothesis, options to remedy the ambiguity will be presented.

One important difference between a tree and a DAG phylogenetic hypothesis is the way in which a clade partitions a classification frame $C = (S, O)$. This is illustrated in Figure 9.1 which shows a clades X and a name $o \in O$ that lies outside X. There are two possibilities with a phylogenetic tree. Either o is an

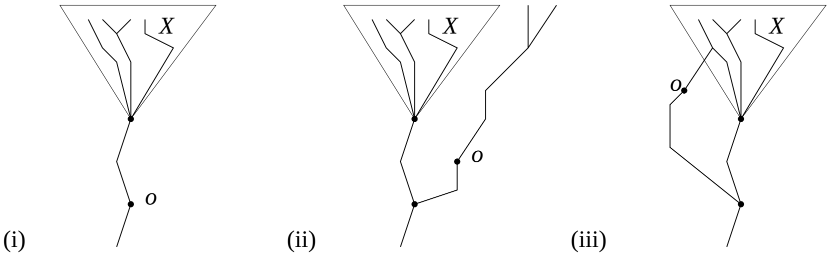

Figure 9.1: A difference between phylogenetic tree and a phylogenetic DAG. Given a clade X and a name $o \in O$ that is outside of X, there are two possibilities with a phylogenetic tree: either (i) o is an ancestor of every member of X, or (ii) o is not an ancestor of any element of X. With a phylogenetic DAG, there is also the possibility, depicted in (iii), that o is an ancestor of a subclade of X.

ancestor of every member of X or o is not an ancestor of any member of X. This is depicted by Figure 9.1 (i) and Figure 9.1 (ii). However, a phylogenetic DAG admits a third possibility as shown in Figure 9.1 (iii). Here o is the ancestor of some subclade of X. It is exactly this case which can make phylogenetic definitions ambiguous.

A node–based definition becomes ambiguous when a set of internal specifiers has more than one '*most recent common ancestor*'. In the case of a phylogenetic tree this can never happen because any set of specifiers has exactly one most recent common ancestor. However, as shown in Figure 9.2, this is not always the case with a phylogenetic DAG. In the figure, both a and b are hybrids produced

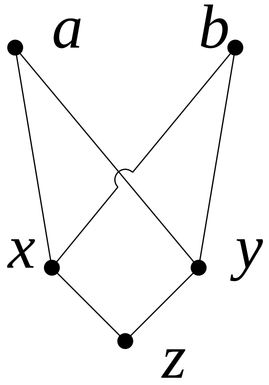

Figure 9.2: An example of a phylogenetic DAG which makes the meaning of the node–based definition containing a and b ambiguous.

from x and y. Recall that the node–based definition node_clade($\{a, b\}, H$) is the set of all minimal clades that contain a and b. Two clades satisfy this condition, *viz.*, $X = \{a, b, x\}$ and $Y = \{a, b, y\}$. This is an example of the situation depicted in Figure 9.1 (iii). From the perspective of clade Y, x is something which lies outside Y and but is also an ancestor to a and b which are contained in Y. The ambiguity in the node–based definition arises exactly because a and b are the internal specifiers to the clade definition.

Recall that one reading of a node–based definition gives the clade stemming from the most recent common ancestor of a and b. Adopting these semantics may seem to resolve the ambiguity, however, a closer inspection will show that it is no solution at all. Since both a and b are hybrids of x and y, it can only be the case that x and y are contemporaneous. Thus it should not be possible to assert whether x or y is the more recent.

One method for resolving this ambiguity involves changing the meaning of a node–based definition to generate a possibly more inclusive clade. Consider assigning to a node–based definition node_clade(A, H) the smallest clade X containing A such that for every name $o \in O$, if o is not an element of X then either o is an ancestor of everything or nothing in X. This alteration serves to eliminate the situation depicted in Figure 9.1 (iii). With respect to the example depicted in Figure 9.2, the node–based definition node_clade$(\{a, b\}, H)$ contains the single clade $\{a, b, x, y, z\}$. It can be shown that this alteration to the meaning of node–based definitions preserves the properties of the old meaning shown in Section 3. Moreover, when the phylogenetic hypothesis is a tree, the new definition generates exactly the same clades as the old definition.

A phylogenetic hypothesis which is a DAG can also cause ambiguities in stem–based definitions. Consider the phylogenetic DAG depicted in Figure 9.3. Denote the stem–based clades stem_clade$(\{a\}, \{z_1\}, H)$ and stem_clade$(\{a\}, \{z_2\}, H)$ by X_1 and X_2 respectively. These clade definitions are unambiguous. However, the stem–based definition stem_clade$(\{a\}, \{z_1, z_2\}, H)$ will be ambiguous since both X_1 and X_2 are both maximal clades that contain a but not z_1 or z_2. Once again, this is an example of the situation depicted in Figure 9.1 (iii). Consider clade X_1 and $MRCA(a, z_2)$. $MRCA(a, z_2)$ lies outside X_1 but is the ancestor of some subset of X_1. Eliminating ambiguity revolves around preventing this situation.

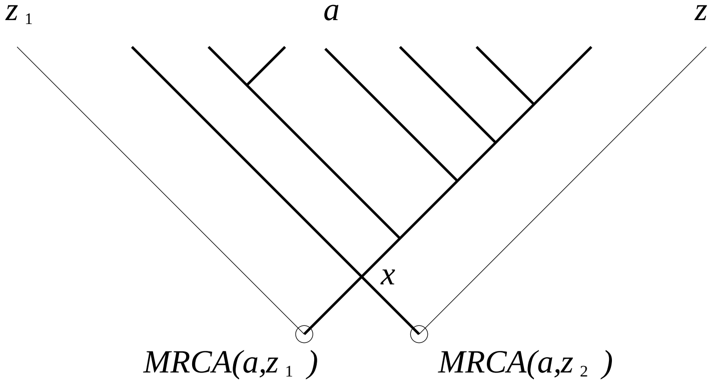

Figure 9.3: An example of a phylogenetic DAG which makes the meaning of the node–based definition containing a and b ambiguous.

A resolution to the ambiguity of stem–based definitions is dual to the resolution for node–based definitions. First of all, consider what the meaning of stem_clade$(\{a\}, \{z_1, z_2\}, H)$ should be. Recall that one property of a stem–based definition on phylogenetic trees was that the larger the set of external specifiers, the smaller the clade. If this property is to be preserved under phylogenetic DAGs (combined with the idea of a maximal clade), then there is only one way to resolve the ambiguity. Recall that clade X_1 in Figure 9.3 is generated by the internal specifier a and the external specifier z_1. Adding another external specifier z_2 should produced something smaller than X_1. By a symmetric argument, stem_clade($\{a\}, \{z_1, z_2\}, H$) should also produce something smaller than X_2. The largest clade which satisfies these constraints is the clade stemming from x, *viz.*, the intersection of X_1 and X_2. In general, the meaning of a stem–based definition stem_clade(A, Z, H) can be set to clades X that contain all of A and no elements of Z such that for any name $o \in O$ that lies outside of X, either o is an ancestor of all elements of X or no elements of X.

The generalisation from phylogenetic trees to phylogenetic DAGs has advantages and disadvantages. A phylogenetic DAG allows hybridization and endosymbiosis to be represented. However, given previous formulations of the semantics of node–based and stem–based definitions, ambiguities may arise in the meanings of such definitions. This section has shown that a minor modification to the meaning of such phylogenetic definitions will not only resolve the ambiguities but leave the meaning of such definitions unchanged under phylogenetic trees and preserve the properties of those definitions under phylogenetic trees.

10 Discussion and Future Work

One of the main themes in the debate about biological nomenclature concerns the purposes and aims of a nomenclatural system. Before such an issue can be resolved, it is necessary to closely scrutinise the properties of each contending system of nomenclature. This paper has contributed to this for phylogenetic nomenclature. A formalisation of the framework and definitions has been presented which sets out the assumptions and structures present when defining clades. Furthermore, a results have been presented showing the properties possessed by phylogenetic definitions. The most significant of these are the results on the semantic stability of node–based and stem–based definitions; information that is redundant for a particular phylogenetic hypothesis increases the semantic stability for different phylogenetic hypotheses. Two new phylogenetic definitions have also been presented which generalise existing definitions. Phylogenetic definitions have been shown to be ambiguous under certain conditions when the phylogenetic hypothesis is a rooted directed acyclic graphs. Solutions to such problems have also been presented.

Formalisation facilitates the comparison of nomenclatural codes through the properties which the codes possess. If there is general agreement in the biological community about what properties a nomenclatural system should possess, then it would be a relatively simple task to see which system possesses the most properties. However, this is not likely to be the case. Proponents of phylogenetic nomenclature seek to make all taxa monophyletic clades. Others [15, 3] have argued that taxonomy should reflect overall similarity. These two properties are not mutually exclusive. In fact, it is precisely the case that these two properties are highly similar [3] that causes much of the debate. A formal study will be able to shed light on where these two properties coincide and where they differ.

Bibliography

[1] M. J. Benton. Stems, nodes, crown clades, and rank–free lists: is Linnaeus dead? *Biological Reviews*, 75:633–648, 2000.

[2] P. E. Berry. Biological inventories and the PhyloCode. *Taxon*, 51:27–29, 2002.

[3] R. K. Brummitt. How to chop up a tree. *Taxon*, 51:31–41, 2002.

[4] H. N. Bryant and P. D. Cantino. A review of criticisms of phylogenetic nomenclature: is taxonomic freedom the fundamental issue. *Biological Reviews*, 77:39–55, 2002.

[5] P. D. Cantino. Phylogenetic nomenclature: addressing some concerns. *Taxon*, 49:85–93, 2000.

[6] P. D. Cantino and K. de Queiroz. Phylocode: A phylogenetic code of biological nomenclature. http://www.ohiou.edu/phylocode, June 2004. Version 2b.

[7] R. A. Crowson. *Classification and Biology.* Atherton, New York, 1970.

[8] C. Darwin. *On the Origin of Species by Means of Natural Selection.* John Murray, London. Reprinted 1998. Modern Library Paperback Edition, 1859.

[9] K. de Queiroz and P. D. Cantino. Taxon names, not taxa, are defined. *Taxon*, 50:821–826, 2001.

[10] K. de Queiroz and J. Gauthier. Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. *Systematic Zoology*, 39:307–322, 1990.

[11] K. de Queiroz and J. Gauthier. Phylogenetic taxonomy. *Annual Review of Ecology and Systematics*, 23:449–480, 1992.

[12] P. L. Forey. *PhyloCode* – pain, no gain. *Taxon*, 51:43–54, 2002.

[13] M. Holder and P. O. Lewis. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, 4:275–284, 2003.

[14] J. P. Huelsenbeck and F. Ronquist. Mr Bayes:Bayesian inference of phylogeny. *Bioinformatics*, 17:754–755, 2001.

[15] C. Jeffrey. *Biological Nomenclature.* London, Edward Arnold; New York, Crane Russak, 1973.

[16] M. S. Y. Lee. Ancestors and taxonomy. *Trends in Ecology and Evolution*, 11:26, 1998.

[17] P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50:913–925, 2001.

[18] K. Padian, J. R. Hutchinson, and T. R. Holtz Jr. Phylogenetic definitions and nomenclature of the major taxonomic categories of the carnivorous Dinosauria (Theropoda). *Journal of Vertebrate Paleontology*, 19:69–80, 1999.

[19] P. C. Sereno. A rationale for phylogenetic definitions, with application to the ghigher-level taxonomy of Dinosauria. *Neues Jahrbuch für Geologie und Paläontologie, Abhandlungen*, 210:41–83, 1998.

[20] T. F. Stuessy. Taxon names are *still* not defined. *Taxon*, 50:185–186, 2001.