Corroboration and the Popper Debate in Phylogenetic Systematics

by

Justin Bzovy

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF ARTS

Department of Philosophy

University of Manitoba

Winnipeg

Abstract

I evaluate the methods of cladistic parsimony and maximum likelihood in phylogenetic systematics by their affinity to Popper's degree of corroboration. My work analyzes an important recent exchange between the proponents of the two methods. Until this exchange, only advocates of cladistic parsimony have claimed a basis for their method on epistemological grounds through corroboration. Advocates of maximum likelihood, on the other hand, have based the rational justification for their method largely on statistical grounds. In Part One I outline corroboration in terms of content, severity of test and explanatory power. In Part Two I introduce the two methods. In Part Three I analyze three important debates. The first involves the appropriate probability interpretation for phylogenetics. The second is about severity of test. The third concerns explanatory power. In Part Four I conclude that corroboration can decide none of these debates, and therefore cannot decide the debate between cladistic parsimony and maximum likelihood.

Acknowledgements

First and foremost, I would like give my deepest thanks my supervisor, Dr. Rhonda Martens, without her patience I would never have been able to accomplish this thesis. I would like to thank Dr. Carl Matheson, my second reader, for believing in my abilities, and for giving me the tough love when I needed it. Thank you to Dr. Jeffrey Marcus, my external from Biology, for making the most technical aspects of phylogenetics easily understandable. I thank Professor Patrick Walsh for introducing me to twentieth century philosophy of science, and for first putting one of Karl Popper's books in my hands. Thanks, of course, to all the professors I have had the pleasure to work with here at the University of Manitoba: Dr. Dimitrios Dentsoras, Dr. Michael Stack, Dr. Chris Tillman, Dr. Ken Warmbrod, Dr. Esa Diaz-Leon, and Dr. Rob Shaver. Thanks also to all the graduate students who have made my stay here so enjoyable: especially José Larios, John Dyck, Jamie Hebert, Mitchell Kredentser, Max Wolinsky, Si Chen, Chelsey Booth, Curtis Kehler, Daniel Rabinoff, Adam Gingera, Patrick Kaczmarek, Shawn Jordan, Thomas Crosby, and many others that I have forgot to mention. Special thanks go to Laurie Morris and Rosa Di Noto for keeping our department working, and helping me keep it all in order. Last but not least, I thank Dr. Jack Bailey for introducing me to philosophy

so many years ago, and for convincing me that the discipline and the questions it raises are of the utmost importance.

I would also like to thank my parents, John and Lynne, my brother, Calvin, my friends, and all of my other loved ones, for supporting me, and dealing with my behavior during this intense period of my life. I hope that I am much easier to deal with in the next chapter.

Table of Contents

Introduction

Biologists are having a “Popper debate” about method in phylogenetic systematics (Kearney 2008: 220). The main concept involved in this debate is Popper’s degree of corroboration. The Popper debate encompasses many different methods of phylogenetic inference. In this thesis I will restrict my focus to a recent exchange between advocates of maximum likelihood (ML) and cladistic parsimony (CP). In Part One I will first clarify Popper’s key concepts that are involved in corroboration, namely, content severity of test and explanatory power, while explicating how corroboration relates to a more general theory of probability. The last point will be further elaborated upon in Part Three. In Part Two I will introduce the relevant terminology from phylogenetics. In Part Three I will then assess the benefits and faults of ML and CP in terms of corroboration and these key concepts. In Part Four I will assess the results of applying corroboration to phylogenetics. It is not my aim in this thesis to adjudicate between ML and CP, or to suggest a multiple-methods approach, but to discover the limits of Popper’s philosophy, specifically his degree of corroboration, when it is applied to phylogenetics.

Part One: Corroboration

A degree of corroboration is a formalization of the degree to which an hypothesis has *so far* stood up to empirical tests (Popper 1983: 220). Formally corroboration looks similar to any other formulas built out of probabilities. However, in terms of its interpretation corroboration is much different. Probability theory usually gives us an idea about how a hypothesis will perform in the future, whereas a degree of corroboration tells us only of past tests. Corroboration in phylogenetics is often understood as an "optimality criterion", a way of comparing and ranking hypotheses in terms of their relationships to a given body of evidence. When we have two competing hypotheses we choose the more corroborated hypothesis, the hypothesis that has passed more severe tests than the other. This choice is not based on which hypothesis is more probable, because corroboration does not satisfy the laws of the probability calculus (Popper 1959: 387). I will now further explain these fundamental remarks in terms of the relationship between probability and content, severity of test, and explanatory power.

Chapter One: Probability and Content

In developing corroboration, Popper (1959: §8; §80) used a distinction between logical or absolute probability and subjective probabilities,[1] (the latter of which are usually understood as degrees of belief). Popper developed his degree of corroboration on considerations involving the former. Formally, to understand what

[1] Popper is not the only writer to make such a distinction. We could make a similar distinction between probabilities of events occurring and probabilities of hypotheses being true or false. R. B. Braithwaite (1953: 118-122; especially his footnote on page 119), for example, discusses the history of this distinction, noting the differences between Popper's and others' views on the matter.

a degree of corroboration is meant to represent, we can start with the absolute probability, *P*, of a statement, *a*, given some value, *r*, the latter of which is a probability between 0 and 1, (where 0 is a contradiction and 1 a tautology):

Absolute Probability

$P(a) = r$

For example, the absolute probability that a 6 will be rolled, given a normal, fair 6 sided-die, is 1/6. But, when discussing theories of natural science, we will always take into account, implicitly or explicitly, our background beliefs before assigning probabilities. Even in the case of the die, our background belief is that it is fair. Popper notes the absolute probability of statements merely in order to draw a connection between logical or empirical content and probability, which we will now turn to.

Popper maintains that logical or empirical content increases as probability decreases. The logical content (or 'consequence class') of a statement, *x*, is the class of all non-tautological statements logically entailed by *x* (Popper 1959: 120; 1965: 218). From what Popper (1959: 319; *cf*. 1965: 218) called "the monotony law of the probability calculus," the probability of the conjunction of any two statements, $P(xy)$, will always be less than or equal to that of either of its components, $P(x)$ and $P(y)$:

The Monotony Law

$P(x) \geq P(xy) \leq P(y)$

In contrast to the monotony law, the logical content of the conjunction of two statements, Ct(xy), will always be greater than, or at least equal to, that of either of its components, Ct(x) and Ct(y):

The Law of Logical Content

$\mathrm{Ct}(x) \leq \mathrm{Ct}(xy) \geq \mathrm{Ct}(y)$

The empirical content of a statement, x, is the class of all of x's potential falsifiers. Empirical content follows the same rule as logical content, if the empirical statements compared contain no "metaphysical elements" (Popper 1959: 120). For example, "It is raining now" vs. "It is raining now and the World Spirit exists." The conjunction would have greater logical content than the first statement, but would only have the same empirical content. A recent description interprets Popper's notion of empirical content as the breadth and scope of existence that a hypothesis purports to explain (Lienau & DeSalle 2009: 187). With these two laws in mind, we see that the more precise and bolder the statement, the lower will its probability be. Consider the following two statements, x and y, and their conjunction:

x = "Tomorrow will be cloudy"
y = "Tomorrow will be hot"
xy = "Tomorrow will be cloudy and hot"

The conjunction (xy) entails more about a particular event, what tomorrow will be like, thus having greater logical content. It also has more potential falsifiers—if either one of the conjuncts, x or y, are false (if either "Tomorrow will be cloudy" or "Tomorrow will be hot" is false), then the conjunction, xy, is false (then "Tomorrow will be cloudy and hot" will be false). Furthermore, the conjunction simply contains more information than either of its conjuncts alone: it says more about what the world will be like tomorrow. However, the conjoined statement is also less probable than either of its conjuncts.

As content increases so does improbability. Thus, if the goal of science is to progressively explain more and more about the world, then making statements with high probability must not be a scientific goal. If this claim is too strong, we may at least make the weaker claim that increasing the content of our theories *seems* incompatible with increasing the probability of our theories. This second claim I think is more apt, because if we take the first we may think that the goal of science is to come up with improbable, yet absurd statements. For example, the statement: "I will turn into a unicorn at the end of this sentence," would then be a great candidate for a scientific statement.[2]

Chapter Two: Severity of Test

Now that I have discussed the notion of logical and empirical content, I will speak a bit more about Popper's "severity of test." Again, if we aim at a high degree of informative content, then we aim at low probability. If we aim at a low probability of a theory's verification, then we aim at a high probability of a theory's falsification. Testability (falsifiability or refutability) is thus the converse of logical probability. The more testable a theory is, the more it is corroborable.

A theory's degree of corroboration increases with the number of its corroborating instances. The first corroborating instances are more important than

[2] However, a Bayesian who starts testing hypotheses with low prior probabilities may later assign high posterior probabilities after the hypothesis tests successfully. This distinction between prior and posterior probabilities captures something analogous to Popper's idea of the inverse relationship between logical content and logical probability.

later ones. If, for some theory, *h*, we keep performing the very same test over and over again we do not keep increasing *h*'s degree of corroboration. Corroborating instances in a new field of application considerably increase a theory's degree of corroboration.[3] Some examples illustrative of this last point are Johann Gottfried Galle's discovery of Neptune, Heinrich Rudolf Hertz's discovery of electromagnetic waves, Arthur Eddington's eclipse observations, Walter M. Elsasser's interpretation of Clinton J. Davisson's maxima as interference fringes of de Broglie waves, and Cecil Frank Powell's observations of the first Yukawa mesons. In all of these cases, the new field of application was the result, according to Popper's interpretation, of highly improbable predictions, in light of our previous background knowledge. Popper also wished to try to formalize the idea of a "sincere and ingenious attempt" at refuting a scientific hypothesis, although he admits this may not be entirely possible (Popper 1959: 402).

We can now unpack the notion of a severe test[4], and then work towards a formal definition of corroboration. With severity of test, general idea that Popper

[3] This point bears some resemblance to the motivations behind William Whewell's (1847) "consilience of inductions." Whewell, however, believed that surprising results in new fields of applications would indicate that the theory producing the results had discovered a *vera causa*. But all that Popper thought we could say about the truth or falsity of scientific theories, when we compare two theories in terms of corroboration, is that the one with the higher degree of corroboration would correspond better to the truth. The truth about the empirical world is in principle unknowable, but that there is an objective truth to the empirical world we may posit. If a theory has succeeded when tested up against the empirical world relative to its competitors, this theory is closer to the objective truth. Typically we also want the theory with the higher degree of corroboration to explain why its competitors have failed. Verisimilitude plays an important role in avoiding certain tenants of standard verificationist epistemology, but we will neglect to explore this notion further.

[4] Among other things, a severe test first requires a properly prepared theory, i.e., an axiomatized theory that contains no superfluous assumptions (Popper 1959: 71-72). Within any given theoretical system we can discern statements at different levels of universality. The axioms of a system are at the highest level of universality (Popper 1959:75). Popper gives the following four rules for the axioms of a system:

wanted to capture is that if some evidence, e, is a test of an hypothesis, h, then e will be more severe if it is less probable given b, our background knowledge, alone. Continuing from our discussion of logical (and empirical content), we may first start, (1), with the logical content (Ct), of some statement x, as the compliment of its probability (P).[5]

(1) $\mathrm{Ct}(x) = 1 - \mathrm{P}(x)$

Alternatively, we may start, (2), by taking logical content as the multiplicative inverse of probability.

(2) $\mathrm{Ct}(x) = \frac{1}{\mathrm{P(x)}}$

If we start with (1), then we may get a preliminary definition, (3), of severity of test, $\mathrm{S}(e,b)$, where e is interpreted as evidence and b as background knowledge.

(3) $\mathrm{S}(e,b) = 1 - \mathrm{P}(e,b)$

1. The system of axioms must be free from contradiction (or consistent).
    - This is equivalent to demanding that not every arbitrarily chosen statement is deducible from it. (Popper 1959: *24)
2. The system of axioms must be independent.
    - It must not contain an axiom deducible from the other axioms (this requirement is, in a way, a parsimony constraint)
3. The system of axioms must be sufficient for the deduction of all statements belonging to the theory
    - It must be complete.
4. The system of axioms must be necessary for the deduction of all statements belonging to the theory (or indispensable).
    - It must not contain superfluous assumptions.

Under the influence of his student Joseph Agassi, Popper somewhat changed his view on axiomatized deductive systems. Deductive systems are still required, in order that one may test more and more remote consequences, but they are mere “stepping stones” to better testable theories. (Popper 1965: 221) This change is targeted against those who believe that the goal of science is indeed one giant axiomatized deductive system, if we only could figure it out. Giant axiomatized deductive systems are best regarded as a means of science, not as the end.

[5] See Popper (1965: 390-391; 1983: 244-255).

When we 'normalize'[6] Ct(e,b) (using the factor $\frac{1}{1+\mathrm{P}(e,b)}$) we are lead to:

(4) S(e,b) = $\frac{1-\mathrm{P}(e,b)}{1+\mathrm{P}(e,b)}$

Alternatively, if we start with (2) then we would get (5).

(5) S(e,b) = $\frac{1}{\mathrm{P(e,b)}}$

Now we assume that there is some probability, for a hypothesis, h, that may, or may not be equal to 1. From this assumption, we decide to substitute for 1, in either (4) or (5), P(e,hb). If we do this, then we get (6) and (7).

(6) S(e,h,b) = $\frac{\mathrm{P}(e,hb)-\mathrm{P}(e,b)}{\mathrm{P}(e,hb)+\mathrm{P}(e,b)}$

(7) S(e,b) = $\frac{\mathrm{P}(e,hb).}{\mathrm{P(e,b)}}$

For the remainder, we will adopt (6), as Popper himself usually does[7].

In a similar respect to severity of test, confirmationists, like some Bayesians, put forth the principle of *conditionalization* to capture something like Popper's severity of test. This principle captures the following idea: the more unlikely a prediction, the more evidence it gives to a theory.

Conditionalization:

> When degree of belief in e goes to 1, but no stronger proposition also acquires probability 1, set P`(a) = P(a/e) for all a in the domain of P, where P is your probability function immediately prior to the change (Howson & Urbach, 1993: 99).

[6] Popper (1965: 490) suggests that (4), the 'normalized' version of (3), is better because it is mathematically tidier. He suggests the same thing, when he adds a further 'normalization' factor to the full-blown corroboration factor. In my mind, Popper had a set of desiderata that he developed first, and then worked out various formalizations of these desiderata.

[7] Popper, for reasons which we will not evaluate here, believes that there is little to help us decide between (6) or (7) (Popper 1965: 391).

Successful tests of theories with low prior-probabilities will increase their probability. Since the probability, here a subjective probability (belief), changes after each test, a theory's probability won't keep increasing in virtue of a repeated application of the very same test. Still, here again, a confirmationist has the goal of assigning a new probability to a theory. That is, we have reason to believe that a theory with a higher probability, given conditionalization, will perform better than one with a lower probability in the future. Corroboration, on the other hand, will only give us a backward looking report on the success or failure of past tests, and not a reason to believe that the hypothesis will survive future tests.

Chapter Three: Explanatory Power

Popper uses the same intuitive steps as to define explanatory power as he used to define severity of test. The explanatory power of a statement increases as logical content increases and probability decreases. Popper, like many other philosophers of science, follows the deductive-nomological model of explanation, where "to explain" roughly means "to derive from a general law" (Popper 2009: 93; *cf*. Hempel 1965: 249). Formally, we want to figure out the explanatory power of an hypothesis, h, to explain evidence, e, given our background knowledge, b. Again, the more severe the test by evidence, e, of a hypothesis, h, the greater the explanatory power of h, if h passes the test:

(8) $\mathrm{E}(h,e,b)= \frac{\mathrm{P}(e,hb)\,-\,\mathrm{P}(e,b)}{\mathrm{P}(e,hb)\,+\,\mathrm{P}(e,b)}$

Popper's general idea is to show that both the explanatory power of a theory and the severity of a test upon a theory depend upon the theory's content (Popper 1965: 391).

There is one other point that cannot be formalized. Both explanatory power and severity of test can be hampered by what Popper referred to as the adoption of *ad hoc* auxiliary hypotheses: "As regards *auxiliary hypotheses* we decide to lay down the rule that only those are acceptable whose introduction does not diminish the degree of falsifiability or testability of the system in question, but, on the contrary, increases it" (Popper 1959: 82-83). This is what he calls "the principle of parsimony in the use of hypotheses" (Popper 1959: 145). Specifically in relation to explanatory power Popper describes an *ad hoc* hypothesis as one "which goes as little beyond the facts it is expected to explain" (Popper 1983: 232)[8].

We are now in a position to define degree of corroboration [C(h,e,b)] in terms of probability statements. Here we will interpret C(h,e,b) as the support or corroboration provided to an hypothesis, h, by evidence, e, given background knowledge, b. Although most at least agree that $C(h,e,b) \neq P(h,e,b)$, others have recently interpreted Popper's degree of corroboration differently than we will here. For instance, de Quieroz & Poe (2001) argue that corroboration, as used in standard phylogenetic analysis is equivalent to likelihood, that is $C(h,e,b) = P(e,hb)$ in at least this particular case. I will discuss their view in Part Two. For the time being I will

[8] In some ways the convention to dismiss *ad hoc* manoeuvres shows a tension in Popper's overall view which Rowbottom (2010) addresses in detail by considering how Popper can deal with the Duhem-Quine thesis (*cf*. Duhem 1954; Quine 1953).

finish explicating corroboration, by briefly discussing how its normalization factor works.

Chapter Four: Normalizing Corroboration

In most cases we see that degree of corroboration is equal to explanatory power or severity of test, as above:

$$(9)\ \mathrm{C}(h,e,b) = \frac{\mathrm{P}(e,hb) - \mathrm{P}(e,b)}{\mathrm{P}(e,hb) + \mathrm{P}(e,b)}$$

Popper argued that degree of corroboration [C(h,e,b)] approaches explanatory power [E(h,e,b)] in most important cases. But corroboration required normalization to step around certain "blemishes". The bolded normalizing factor, as seen below in definition (10), was added to the denominator because he thought it was the simplest way to remove the blemishes, although he had experimented with other alternatives (*cf*. Popper 1959: 402, n. 8).

$$(10)\ \mathrm{C}(h,e,b) = \frac{\mathrm{P}(e,hb) - \mathrm{P}(e,b)}{\mathrm{P}(e,hb) - \mathbf{P}(\boldsymbol{eh},\boldsymbol{b}) + \mathrm{P}(e,b)}$$

One such blemish is as follows. Consider a situation where we have e falsifying h in the presence of b, that is, where P(e,hb) = 0. But if this is a severe test e will be also very improbable relative to b alone, that is, P (e,b) ≈ 0. If we were to define C(h,e,b) simply as P(e,hb) – P(e,b), we would get a value close to 0. But, the value 0 is reserved for evidence that is irrelevant to the hypothesis. Tautologies, for example, do not count as evidence for or against hypotheses. Tautologies are always true (empirical hypotheses do not explain tautologies). But if we adopt definition (10) above instead of (9), then we will always have C(h,e,b) = – 1 when e falsifies h in the presence of b (Popper 1983: 242).

Corroboration, as given in (10), is most importantly designed to capture the following requirements. Any e supports h when $P(e,hb) > P(e,b)$, i.e., when $C(h,e,b) > 0$. The maximum support, $C(h,e,b) = 1$, will occur when $P(e,hb) = 1$ and $P(e,b) = 0$. Any e undermines h when $P(e,hb) < P(e,b)$, i.e., when $C(h,e,b) < 0$. $C(h,e,b) = -1$ only when e contradicts h in light of b. If some e neither supports nor undermines h, then $C(h,e,b) = 0$. This e would be irrelevant to h, that is, h could not explain this e. One difference between $C(h,e,b)$ and $E(h,e,b)$ is that if e entails h, then for any given h, $C(h,e,b)$ and $P(e)$ increase together, which does not follow from the definition of $E(h,e,b)$ (Popper 1959: 401).

This concludes my discussion of the formalism and interpretation of corroboration. In Part Two, I will introduce the relevant terminology from phylogenetics. In Part Three, as I go through the recent exchange between advocates of ML and CP in phylogenentics along the lines of corroboration, I will use the remarks made here as a starting point, but will elaborate on them as required.

Part Two: Phylogenetic Systematics

Phylogenetics, like almost every science, utilizes technical jargon. I will forgo a detailed introduction to phylogenetics here, since we are only interested in the philosophical issues involved in the Popper debate. In Chapter Five I introduce only the concepts relevant to understanding the Popper debate. In Chapter Six I introduce ML and CP using these concepts, while speaking briefly about the type of probability interpretation each method conforms to.

Chapter Five: The Aim and Content of Phylogenetic Systematics

Phylogeny is the branch in biology that deals with the evolutionary history of a species or higher taxonomic[9] groups of organisms. Phylogeny attempts to classify the evolutionary relationships among organisms, by representing the patterns of lineage-branching produced by the true evolutionary history of life. However most biologists believe that: "The ideal system of phylogeny is unattainable. What we provisionally call a phylogenetic system is more probable than other systems (and that it more closely approximates the ideal system)" (Hennig 1966: 29). Phylogenetics[10], originating in the work of Willi Hennig (1966), attempts to

[9] The major taxonomic ranks in descending order of generality are typically as follows: Life, Domain, Kingdom, Phylum (Zoology) and Division (Botany), Class, Order, Family, Genus, and Species.

[10] Phylogenetic systematics (PS) is not the only approach to reconstructing the evolutionary relationships amongst species. For example, there is another approach known as evolutionary taxonomy (ET). The general difference between them is that ET is less informative about genealogy so it can be more informative about ecology. ET employs non-genealogical and genealogical standards. ET recognizes the possibility that a line of birds might evolve into non-birds: "Though it could not become a line of insects, its members could form a new higher-level taxon" (Laporte 2004: 77). The important difference between ET and PS is that: "Any taxon recognized by systematists of the cladistic school includes every descendant belonging to any organism in the taxon. Taxa recognized by evolutionary taxonomists, on the other hand, sometimes exclude some descendants of organisms belonging to the taxa" (Laporte 2004: 20). Monophyletic groups or clades, are the technical names for the taxa that are recognized by PS.

reconstruct evolutionary history on the basis of observed patterns of sameness and difference in the characteristics of taxa using agreed upon, and rationally justified principles of inference. The two principles of inference under consideration belong to what is called the cladistic school of analysis[11] and are known as maximum likelihood (ML) and cladistic parsimony (CP).

Originally phylogeneticists used morphological data, but now predominantly use molecular data to infer genealogy. Scientists following Hennig (1966) discovered reasons for believing that not all shared characteristics are useful for describing evolutionary relationships between organisms. Wiley (1981: 117-119; the examples are his) contrasts three general types of characters: structural, functional, and phylogenetic, which are displayed in the following table:

Structural	Functional	Phylogenetic
• Characters that look the same way with respect to physical structure • e.g., *Amia calva* (a type of fish commonly called the bowfin) and higher teleosts (*Teleostei* is one of three infraclasses in class *Actinopterygii*, the ray-finned fishes) have cycloid scales covering the body. But, these two groups do not have the same phylogenetic character.	• Characters that act the same way, that perform the same function • e.g., the wings of birds and butterflies. These do not look the same in terms of structural characters, and they do not have the same phylogenetic character.	• Characters that are similar due to inheritance from a common ancestor • These characters do not always (clearly) exhibit their similarity • e.g., the lower jaws of gnathostomes (the group of vertebrates with jaws) is a character that is (hypothesized to be) present in the most recent common ancestor of the group.

[11] The cladistic school of analysis as contrasted with the phenetic school. The latter is more permissive in the patterns of sameness and difference that it accepts as evidence of relationship.

Phylogenetic characters, according to CP, are the only characters that can tell us about evolutionary history. These characters are not observable, and must be inferred from raw data, whether morphological or molecular, using methods of phylogenetic inference. I will now describe the types of character that CP uses in more detail.

5.1 The Logic of Characters

A primitive character (plesiomorophy) is one that, through evolution, gives rise to a derived character (apomorphy). When a plesiomorphy is shared by two or more taxa it is called a "symplesiomorphy" (shared-primitive character). When a derived character is shared by two or more taxa it is called a "synapomorphy" (shared-derived character). Primitive and derived characters are relative to a particular phylogenetic tree (a representation of the evolutionary relationships of a group of organisms). When derived characters are shared amongst members of a group, then the most likely (in a non-technical sense) explanation is that they are shared from a more recent common ancestor. [12]

The following diagram (Fig 1 and 2) shows the distinction between symplesiomorphy (shared primitive character), synapomorphy (shared derived character), and a third type of character, called "homoplasy", which we will discuss further below:

[12] Pheneticists, on the other hand, hold that ancestral states shared by two or more taxa (simplesiomorphies) also provide evidence of phylogenetic relationships (*cf*. Sober 1983: 337-338 for discussion).

Fig. 1

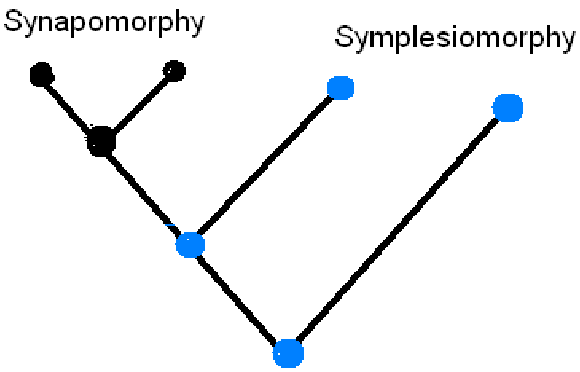
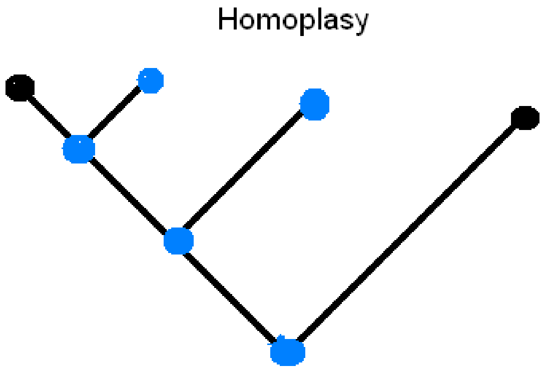


Fig. 2



The branches, represented by lines, in Figure 1 and 2 depict how the different taxa of each particular group, represented by circles, are related by the different traits they bear. Blue represents a primitive trait (plesiomorphy), and black represents a derived trait (apomorphy). In Fig. 1 the two blue circles near the top right are a symplesiomorphy because the ancestor, the blue circle at the bottom of the diagram, has this trait. The two black circles on the top left of Fig. 1 indicate a synapomorphy because these taxa have evolved this trait from their most recent common ancestor (the black circle that branches into them). The difference between synapomorphy and symplesiomorphy is relative to the taxa considered. That rats and apes are more closely related to each other than either are to lizards, cannot be shown on the basis of rats and apes having five toes on their hind legs, because lizards also have this character. If we want to show why cats and dogs are more closely related to each other than they are to kangaroos we can do this on the basis of cats and dogs both having a pair of teeth adapted for tearing apart flesh (called the carnassial pair), that kangaroos lack. The carnassial pair is derived from a more recent common ancestor (a synapomorphy) of cats and dogs, than from the most recent common ancestor of cats, dogs, and kangaroos combined. That is, since cats and dogs share the carnassial pair and kangaroos lack it, kangaroos are more distantly related to cats and dogs, than cats and dogs are to each other. In Fig. 2 the two black circles represent identical traits that evolved in separate lineages (homoplasy). "Wings" are found in both bats and butterflies, but not on the basis of a shared common ancestor. These "wings" have each evolved from different primitive states.

By simply observing two taxa that share a trait we do not know whether the trait that they share is a symplesiomorphy, a synapomorphy, or a homoplasy. The way that CP deals with homoplasy is most problematic, as I will discuss further in Chapter Nine.

5.2 The Logic of Trees

Hypothesized phylogenetic relationships between species are represented by phylogenetic trees (or cladograms).[13] The over-simplified diagram (Fig. 3 below) reflects the following terminology. A phylogenetic tree has *nodes*, *branches*, *tips* and

[13] The distinction between cladograms and phylogenetic trees is not at issue in the Popper debate. But, for some remarks on this distinction please see Sober (1983: 336-337).

a *root*. When the diagram depicts a monophyletic group, the root is often called a "stem species." A monophyletic group is a group of species that encompasses a stem species and all of its descendants. Currently existing species are located at the tips of the branches (cat, dogs, and kangaroos). Nodes are often occupied by merely hypothetical common ancestors (the "cat-dog"), but sometimes these can be known extinct species, as inferred, for example, from fossil records. The branch length often indicates some form of evolutionary distance (ML), as in how much time has passed, but under CP the branch length is not considered to be important to inferring phylogeny.

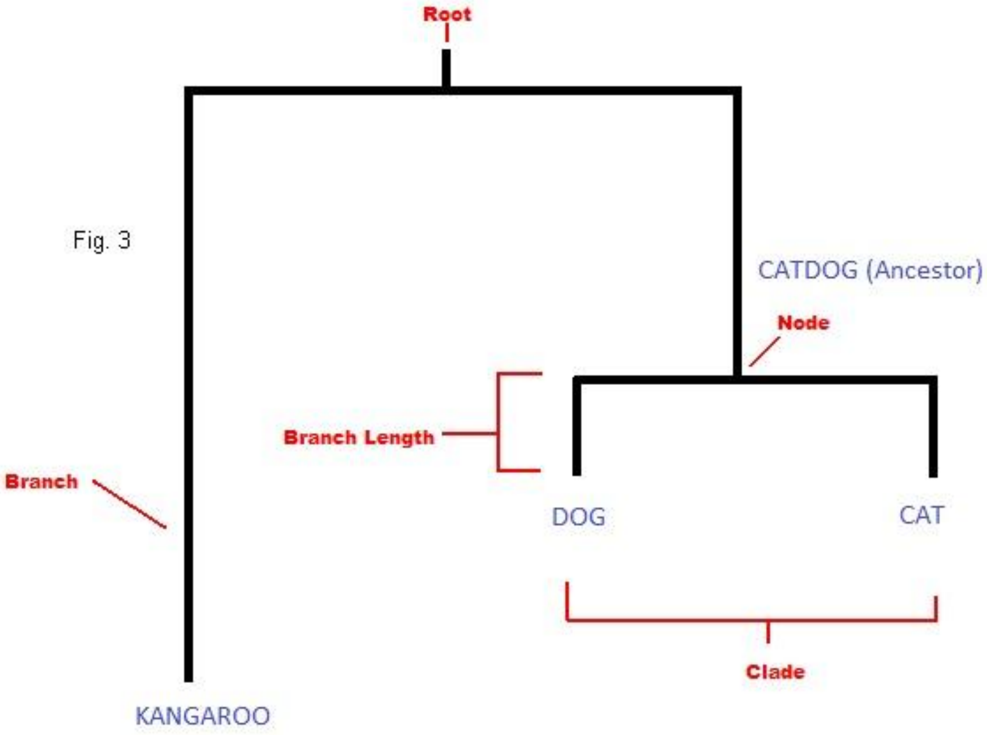


Fig. 3

This concludes the discussion of the required phylogenetic terminology for addressing the Popper debate. In Chapter Six I will briefly outline the differences, using the terminology just introduced, between ML and CP.

Chapter Six: Cladistic Parsimony and Maximum Likelihood

The main difference between the ML and CP, in relation to the terminology presented in Chapter Five, is that ML considers differences in branch length to be relevant to inferring phylogeney, whereas CP does not. In this chapter I will also speak about which probability interpretation each method pays heed to, because this will become relevant in Chapter Seven.

6.1 Cladistic Parsimony and Logical Probability

CP is often called maximum parsimony because it selects the most parsimonious tree. A most parsimonious tree is a tree that requires the smallest number of evolutionary changes in order to explain the phylogeny of the species under study and to explain the observed shared character states these species have. CP is also an ordering criterion, besides choosing a best tree, a most parsimonious tree, it also ranks competing hypotheses from best to worst. CP compares different trees and chooses the tree with the least number of independent evolutionary steps (in molecular data this would be the least substitutions of nucleotides: Cs, As, Gs, and Ts). A homoplasy would give us two different evolutionary steps for one character trait, so CP chooses the tree that has the least amount of homoplasies. In both CP and ML we now consider only molecular data, however, when it comes to fossil records, there is no molecular data to deal with. CP can deal with the morphological data from fossil records, but ML has not been developed to deal with this data.

CP, in the form in which we will consider it, conforms to a logical interpretation of probability. The logical interpretation of probability shares much with the classical interpretation of probabilities. Both classical and logical approaches believe that probability is determined *a priori* by examining the space of possibilities. The logical interpretation is different because it may assign unequal weights to these possibilities, and that it may be computed whether the evidence is symmetrically balanced or not. For example, we examine the space of possibilities for a fair coin and see that there are only heads or tails. Now assuming that everything else is also fair, we get a logical probability of exactly 0.5 for either the occurrence of heads or tails on a particular toss. Popper, again, argued that "the logical probability of a statement is complementary to its degree of falsifiability: it increases with decreasing degree of falsifiability" (Popper 1959: 119). For example, a tautology would have the logical probability of 1, which, in turn would have 0 as its degree of falsifiability. The degree of falsifiability is also referred to as a statement's logical content, or empirical content or testability, as we have discussed. Popper's own understanding of logical probability does not extend much further than the relationship between logical probability and content. However, Popper developed a propensity interpretation of probability that we will introduce later on.

The use of logical probability in CP stems from the space of possibilities that are implied by the assumption of bifurcating evolution[14], although this has not always been made clear by its advocates. That is, "for any *N* taxa there are exactly $\frac{(2N-3)!}{2^{N-2}(N-2)}$ possible bifurcating cladograms, all capable of explaining observed state distributions" (Siddall & Kluge 1997: 313-314). So, for example, if we have any three taxa, A, B, and C, then we have three, and only three, possible cladograms: (AB)C, (AC)B, and (BC)A, which can be tested against possible character evidence. Given the available possibilities, CP chooses the one which requires the fewest evolutionary transformations to give the observed characteristics of A, B and C.[15] At any rate, it cannot merely be shown from some accordance with corroboration that CP conforms to a logical interpretation of probability. I will make this last point clear in Chapter Seven.

6.2 Maximum Likelihood and Frequency Probability

ML is a qualitatively different method of phylogenetic inference that was originally developed to deal with problems with CP. An important difference between CP and ML is that ML requires an explicit model of the evolutionary process. A preferred tree is a tree that makes it most likely that one would observe the characters that we actually have observed, given some model of the evolutionary

[14] One may argue that the assumption of bifurcating evolution forbids the possibility of polytomy (speciation events that produce more than two species) and the possibility of reticulation (speciation events where two or more species become one), and these two events have known occurrences. Unfortunately a satisfactory treatment of this worry, which is not at issue in the aspects of the Popper debate we are considering, would lead us too far astray.

[15] This example is of course an oversimplification. As we see from the above formula, enumerating the number of possible trees for say 10 taxa would give us 34, 459, 425 different trees. This makes it next to impossible to search through all possible trees for large sets of taxa and various approaches, called "heuristic" approaches, have been developed to whittle the options down. These approaches are more concerned with practical issues and should play no role in the logic of phylogenetics.

process. The data, the characters we have observed, do not change, they have already happened, but we can change the model.

Let's consider a brief example. In some ways models are very similar to background knowledge. (Consider the discussion of a die in Chapter One.) If we were to model a die with six sides, and we modeled it as being fair, then the probability that we would roll a 6 based on the model that it is fair, would be 1/6. If we were to model it as having 6 identical sides, say that each side had the number 6 on them, then the probability that we would roll a 6 would be 1. So even if we had only one toss to compare with the two models, landing a 6, we would still see big differences in probability values depending on what model we used.

ML considers aligned genetic sequences (of DNA, RNA, protein etc.,) as data. Sequences are aligned using distance methods first before ML operates on them. For ML in phylogenetics, a tree that relates the sequences and a model of the mechanisms at work in producing the observed patterns of sameness and difference in the sequences are what is being tested.

A model contains two parts. The first is the frequency of its components. So, for DNA, there are only four options for a particular space in an alignment (C, A, T, or G). The simplest model is the Jukes-Cantor model. The Jukes-Cantor assigns the probability of 1/4 to each nucleotide occurring (C, A, T, or G). So if we had a very short sequence, containing only a *T*, then the Jukes-Cantor model would assign this the probability of 1/4. If we sum the likelihood of all the data possibilities, for Jukes-Cantor, or any other model, we get the probability of 1. So with our die, and our model that it is fair, each side, 1 through 6, gets a probability of 1/6. If we some the probability of each side landing, then we get 1/6 x 6 = 1. So different models may assign different probabilities to C, A, T, or G, occurring, but this is the basic idea.

The second part of the model considers multiple sequences, and how the nucleotides in one (C, A, T, or G) may change into the nucleotides in the other. This part is more complicated. We have four different states, and four different possibilities of change (here we count staying the same as a type of change). That gives us 16 different scenarios that must be assigned probabilities by a model. The Jukes-Cantor model assigns equal probabilities for each type of transformation event, but other models vary the probabilities for different types of transformation events. The Kimura model says that transitions and transversions are allowed to have different probabilities. (Purines are A and G, and pyrimidines are C and T. A transversion refers to the substitution of a purine for a pyrimidine or vice versa. A transition is a change from A to G or vice versa, or from C to T or vice versa.)

The last point one might consider in a model is branch length. We might want to say that if more evolutionary time has passed, then the probability of certain changes will increase relative to the amount of evolutionary time that has passed. So if more time has passed, then the probability that a certain nucleotide has remained the same as it was initially would drop. Mutation rates are somewhat random, when compared with morphological data, but they seem to be effected by things like increased birth rate, or, more rarely, environmental factors like radiation or intense heat, for example, making certain compounds more stable.

These sorts of things influence the parameter of branch length. But, considering the probability values, what happens is we get a value for branch length, and multiply it by the 16 values that a model assigns to the different possibilities of change.

As Sober (2004) points out, parameters for models are chosen so that they maximize the likelihood of the various hypotheses under consideration. However, as we choose more and more complex models, we have more and more parameters. A larger number of parameters permit a greater fit to data. And once we have a process model that is complex enough, the likelihoods of each genealogical hypothesis are identical (having achieved the maximum value of unity), independently of what the data is (Sober 2004: 648). Because of this mathematical result, often ML accords with a principle of parsimony, not to be confused with CP, in that it chooses the simplest model of the evolutionary process to account for the evolutionary steps that led to our observed character states.

Sober's (2004) point can be understood in Popperian terms. The simpler a model is, the more restrictive it becomes, that is, the more empirical content it has, the more testable it is, and the more logically improbable it is (as we saw in Part One testability is the converse of logical probability). The more complex a model is the less restrictive it becomes relative to the competing genealogical hypotheses. That is, a sufficiently complex model will not be able to discriminate between the competing hypotheses. Thus, one must use some statistical procedure like the likelihood ratio test to determine whether the likelihood of a given model is sufficiently greater than a less complex model, given some notion of statistical significance, a notion which determines what level of sufficiency is required. Frequency comes in because ML hopes to converge on the correct parameter as more and more character data is taken into account (Siddall & Kluge 1997: 318). Before moving on, I will briefly introduce the frequency interpretation of probability with a coin-toss example.

The frequency interpretation of probability states that the probability of some event occurring is the limit that its relative frequency approaches. This is what Hacking (1965) and others call 'the long run', (which he takes to be a physical property, albeit a property that is not well-defined).[16] For example, if we toss a coin 100 times, and get the result that 30 of those tosses come up heads, we get the relative frequency of 30/100, i.e., a probability of 0.3. But, if the coin is fair and the events of each toss are independent, over an arbitrarily long run of tosses, we will expect that the probability will settle down to approximately 0.5. The probability of getting heads, for the frequentist approach, is identical to the frequency of heads landing over this long run of coin tosses.

More needs to be said about the differences and similarities between ML and CP, but this will come about as I consider the different philosophical arguments, based on corroboration, to prefer one method over another. I will turn to these arguments next in Part Three.

[16] Hacking's (1965) view on probability is closer to Popper's propensity interpretation, which I will discuss further below. This view of probability is very close to the frequency view, but it attributes propensities to the entire experimental set up, which then yields observable frequencies.

Part Three: The Popper Debate

We have already seen how Popper's degree of corroboration [C(h,e,b)] is closely related to the degree to which a test is severe [S(e,h,b)] and to the explanatory power of an hypothesis [E(h,e,b)] by way of the probability values in the shared numerators [P(e,hb) – P(e,b)] of their respective formal expressions. With this in mind, we will see how ML and CP conform to Popper's understanding of severity of test and explanatory power. But before I get to these two points I will consider a more confused debate between ML and CP about probability, that appears to be based on corroboration, but in a way to be determined, is really based on other considerations.

Chapter Seven: The Probability Dispute

I will now consider the debate between ML and CP on this issue of whether or not frequency probabilities are appropriate for phylogenetics, and whether or not corroboration has any stake in this matter. I will argue that a decision in this dispute about probability interpretations requires motivation from sources other than from Popper's general theory of corroboration. That is, in a sense to be determined in this chapter, the application of corroboration in this Popper debate is independent of the choice of an adequate interpretation of probability. Despite corroboration being neutral to a choice between probability interpretations, I will still make some comments to show what Popper's thoughts might be on the matter, and to where the dialectic between ML and CP may turn.

Siddall & Kluge (1997) start their assault on ML by delineating ML and CP by the different probability interpretation each pays heed to (see Chapter Six above). ML rests on a frequency probability approach, whereas CP rests on a logical probability approach. Arguing against ML, Siddall & Kluge contend that a frequency interpretation is inappropriate for historical sciences such as phylogenetics: “Our criticisms of frequency probability theory derive largely from the historical context in which that form of probabilism is employed” (Siddall & Kluge 1997: 314). Briefly their argument is that a frequency interpretation would require a series of identical trials to apply to phylogenetics, and there is no relevant sequence to consider because all the events that phylogeneticists consider are unique. Yet before I go through their argument and some responses to it, I will clear up the relationship between Popper’s degree of corroboration and probability interpretations, because this has led to some confusion in the exchange between advocates of ML and CP on these matters.

7.1 Corroboration and Probability Interpretations

Siddall & Kluge (1997: 313) argue two contentious claims: that Popper’s degree of corroboration exemplifies logical probability, and that likelihood denies corroboration (Siddall & Kluge 1997: 329). I will argue that these two claims are both misunderstandings.[17] Both claims are buried in part of a somewhat confused argument for the exclusivity of cladistic parsimony (CP) in phylogenetics. That is, the argument I discuss in this section, as Siddall & Kluge (1997) present it, is tied

[17] It is unfortunate that de Quieroz & Poe (2001) spend so much time responding to these issues; meanwhile, avoiding the more important question that Siddall & Kluge (1997) have raised: Which probability interpretation is appropriate for phylogenetics?

up with another argument against ML that I will consider in 7.2. I agree with Haber (2005) that it is important to separate the two arguments that Siddall & Kluge (1997) make, but I have formulated them somewhat differently than he has. The first argument, which Haber (2005) calls *the argument from falsification*, also requires a premise (P1) exalting Popper's epistemology. This premise plays no role in the Popper debate, for both sides assume that his philosophy of science is "correct." I formulate the argument slightly differently than Haber (2005: 832) has, re-baptising it *the argument from corroboration*:

The Argument from Corroboration

P1. Corroboration is the only method for evaluating scientific hypotheses

P2. CP conforms to the tenants of corroboration

P3. ML does not conform to the tenants of corroboration

P4. CP is the only phylogenetic method that conforms to the tenants of corroboration

C. Thus, CP is the only phylogenetic method that conforms to the one and only method for evaluating scientific hypotheses

The two contentious claims are buried in P2 and P3. These two premises are at stake in the Popper debate; unpacking them involves understanding the tenants of corroboration, and the methods of CP, and ML. I will argue that corroboration pays no special credence to a particular probability interpretation, while pointing out that likelihood and corroboration, although different in important respects, do share some important similarities.

From the bare expression of its formula, given below, it is hard to see how corroboration in anyway "exemplifies" logical probability.

Degree of Corroboration

$$\mathrm{C}(h,e,b) = \frac{\mathrm{P}(e,hb) - \mathrm{P}(e,b)}{\mathrm{P}(e,hb) - \mathrm{P}(eh,b) + \mathrm{P}(e,b)}$$

Corroboration, like any formula, would require a particular interpretation to exemplify logical probability, namely a logical interpretation, and this Siddall & Kluge (1997) do not provide. This misunderstanding becomes especially clear when Siddall & Kluge (1997: 313) juxtapose corroboration with Bayes' theorem.

Bayes' Theorem

$$\mathrm{P}(h,e) = \frac{\mathrm{P}(e,h)\mathrm{P}(h)}{\mathrm{P}(e)}$$

Bayes' theorem is said by Siddall & Kluge (1997: 313) to typify the calculus of frequency probability. Yet here too we require a particular interpretation of the formula. Bayes' theorem may be interpreted either along Bayesian lines, which is in terms of beliefs (the subjective interpretation), or along frequentist lines, which is in terms of the frequency of two or more events. With this in mind, it becomes terribly unclear how they are expressing Bayes' theorem as "typifying the calculus of frequency probability."[18]

Furthermore, notice that both corroboration and Bayes' theorem make some use of the likelihood function.

[18] Corroboration, as Siddall & Kluge (1997: 313) present it, has a separate term *b*, for background knowledge. This is not a big difference. Bayes' theorem can of course also be expanded to include a term for background knowledge (or background beliefs), and when it is not included in the formalism background beliefs are taken as implicit. Popper too, in many of the formulations of his degree of corroboration, from time to time formulates corroboration without the extra term for background knowledge.

Likelihood

L(h,e) = P(e,h)

But notice that Bayes' theorem, unlike C(h,e,b) or L(h,e), is *prima facie* meant to assign a probability to a hypothesis in light of some given evidence, and this, as we have discussed, is what Popper argued against. At any rate, Bayes' theorem has little to do with ML, though likelihood is of course integral. Furthermore, Popper himself understood that both corroboration and likelihood, as de Quieroz and Poe (2001: 309) readily point out, are not meant to assign probabilities to hypotheses, but "are intended to measure the acceptability of the hypothesis . . ." (Popper 1959: 388). Popper's general complaint, to which corroboration is meant to be a remedy, is with epistemologies that assign probabilities to hypotheses; it is clear from likelihood's formal representation, in that it assigns a probability to the evidence given a hypothesis, that likelihood does not violate Popper's complaint (de Quieroz & Poe 2001: 309).

Corroboration, as Popper developed it, is based on only the mathematical calculus of probability. So corroboration is not inconsistent with any particular interpretation of said calculus. That is, corroboration is not just consistent with logical probability, but also with frequency probability (de Quieroz & Poe 2001: 310). A more general understanding of probability is an understanding under what Popper calls a 'formal' system, which "should be susceptible of many different interpretations . . . for example, (1) the classical interpretation, (2) the frequency interpretation, and (3) the logical interpretation . . ." (Popper 1959: 318). So, for this reason, the frequency interpretation of probabilities is not more or less consistent than the logical interpretation, *contra* Siddall & Kluge (1997), with Popper's degree of corroboration.

Not only this, but Popper (1959) himself at one time preferred, for reasons independent of his notion of corroboration, the frequency interpretation of probability before later (Popper 1983: 347-401) moving to the propensity interpretation. The propensity interpretation is like the frequency interpretation, but it attributes the physical experimental set-up with a tendency to produce long-run frequencies. Popper's philosophy is certainly systematic, so it would seem odd, to say the least, that he worked hard at developing corroboration along a logical interpretation of probability, while developing, at the same time, different versions of the frequency and propensity interpretation.[19]

Also corroboration is usually discussed through paradigmatic examples within the context of the predictive sciences, namely physics, and these examples avoid any special problem for phylogenetics that might be presented to it as a historical science. The predictive sciences involve repeatable experiments, which in turn, can be used to form a reference class wherein we can apply a frequency (or propensity) interpretation of probability. That is, if corroboration was designed to fit physics as a paradigmatic case of science, then it should be designed to handle frequency or propensity probabilities.

[19] Again, the only interpretations of probability that Popper consistently attacked are subjective interpretations.

Thus, it is clear that corroboration is not intended to pay special credence to any particular interpretation of probability. Even if we could make the case that Popper intended corroboration to pay heed to a particular interpretation, it would be easier to make the case that corroboration fits with the frequency or propensity interpretation than it would be to make the case that it fits with the logical interpretation. So CP loses some points here. Popper developed corroboration so that it was applicable to all sciences independently of which probability interpretation is appropriate for a particular science. Frequency probabilities are just as admissible whether we apply corroboration or likelihood to the science. Thus, the decision for or against the appropriateness of a probability interpretation has to come from a source other than corroboration.

7.2 Reference Class Problems

Siddall & Kluge (1997) argue that statistical methods that employ a frequency interpretation of probability are inappropriate for phylogenetics because the science deals with unique particulars. A frequency interpretation cannot properly assess historically unique particulars, so maximum likelihood (ML) is unsuitable for phylogenetics. [20] A logical interpretation of probability can deal with this situation by bearing the data in the simplest form on the competing

[20] One may want to draw an analogy here with other historical sciences that deal with unique particulars like Cosmology. But although Cosmology deals with a unique particular, the creation of the universe, hypotheses about the origin of the universe still generate predictions, which we can test, e.g., the current level of cosmic background radiation. Despite this, we can to some extent make predictions about mutations in biology. If we know, for example, that one molecule is more stable at a certain temperature, and that a certain organism has an environment at this temperature, and then we can predict that this molecule is less likely to mutate. At any rate, there are parallels and differences between historical sciences, but the purpose of this thesis is to examine the debate in biology independently of other related sciences.

hypotheses, so cladistic parsimony (CP) is suitable for phylogenetics. This part of Siddall & Kluge's (1997) argument privileges the status of CP, but does so on the basis of probability, rather than on the basis of corroboration as we saw with the last part of their argument. We may deem their argument against ML, as *the argument from probability* in accordance with Haber (2005: 832) formulating it somewhat differently as follows:

The Argument from Probability

P1. The evolutionary history of life on earth concerns unique historical entities.

P2. Unique historical entities cannot be evaluated with frequency probabilities.

C1. Thus, the evolutionary history of life on earth cannot be assessed with frequency probabilities.

P3. ML assesses the evolutionary history of life on earth by way of frequency probabilities.

C2. Thus, ML is inappropriate for assessing the evolutionary history of life on earth.

Again, the above scheme does not exactly capture all the moves in Siddall & Kluge's (1997) argument, but highlights its main points. This re-formulation, unlike Haber's (2005: 832), does not restrict the argument to concern only phylogenetic trees, and makes it somewhat less ambiguous in respect to the metaphysical sense of possibility at issue. This last point will come clearer as we provide motivation for the first two premises. I will argue that this debate about the admissibility of frequency probabilities clearly rests on a metaphysical debate, which further involves an adequate empirical interpretation, and that corroboration has no obvious stake in these matters.

Starting with P1, we find that Siddall & Kluge (1997: 314-15) first explicate the historical nature of phylogenetics by way of a neat little dialogue that in some ways parallels standard "twin earth" and natural kind arguments from metaphysics (*cf*. Putnam 1973; 1975; Kripke 1980). However, they use these similar strategies to argue for a version of species as individuals (SAI) or "the individuality thesis", rather than for a thesis about biology involving natural kinds. There is no need to rehash in too much detail or evaluate their particular argument within the larger context of the metaphysical aspects of the species debate here, because de Quieroz & Poe (2003) seem to accept their position, at least to a certain extent.[21] Accordingly, we will be content to explicate their understanding of SAI, and briefly elaborate their argument for it. Their understanding of SAI is broader than a thesis merely concerning the nature of species. They contend that: "Biological life is earth-bound through a historically singular continuum of common ancestry" (Siddall & Kluge 1997: 315). On this view Life is one individual thing composed of other individual things. For example, you, or any other human being, are an individual who is a part of other individuals: *homo sapiens*, Homonidae, Primates, Mammalia, Life etc., by virtue of your common ancestry (Siddall & Kluge 1997: 317).

Consider the example of what we may call the functional character "wings" (Siddall & Kluge 1997: 315; Wiley's (1981) three different types of character above).

[21] However, not everyone in the arena of phylogenetics accepts a thesis like this. Natural kinds are still to some degree on the table (*cf*. Franz 2005).

"Wings", when considered functionally, permit flying: birds, bats, and flies fly with their "wings". But the function of flying that "wings" permits does not confer identity on what we are here calling "wings", because birds, bats and flies have their "wings" from different origins. That is, "wings" are homoplasies. But when we consider a robin's wings and a stork's wings, they do have this identity because they have been derived from a common ancestor: these wings are called homologues. Thus, Siddall & Kluge argue that phylogenetics is concerned with the explanation of historical particulars and the events that form new particulars, particulars and events to which origin and location in space and time are relevant. These historical particulars in phylogenetics are things like individual clades (monophyletic groups), lineages, organisms, synapomorphies, and evolutionary transformations: "Each evolutionary transformation is unique in a spatio-temporally restricted, historical sense" (Siddall & Kluge 1997: 316)[22].

Moving on to P2, we cannot apply frequency probability to these historical particulars, because they are unique and interdependent. Frequency probability,

[22] However, other authors have pointed out that there may be some restrictions on this view in respect to certain parts of Life:

> Less inclusive individuals, such as demes, are historical insofar as they are extended in time, but their unity is due also to cohesive and integrative processes, making them *contemporary* individuals. In contrast, *historical* individuals are united only by common history, not current interactions (Grant 2002: 101).

The difference between what Taran Grant (2002) has called *historical* and *contemporary* individuals needs to be spelled out further—especially in terms of whether *the argument from probability* is relevant to contemporary individuals—but at least Grant's distinction shows reasons to not consider every part of Life to fall under the same ontological category. Despite historical uniqueness not applying to demes, it seems that it will also apply to identical nucleotide structures in the DNA. Although these structures may appear to be identical, and for all intents and purposes they may be indistinguishable to us, they still may be distinguished in principle due to their spatio-temporal origins; even if we cannot, due to our epistemic situation, discover these origins we must still assume them.

Siddall & Kluge (1997: 317) argue, requires independent particulars that belong to classes of concurrently possible instances. That is, frequency probability requires some type similarity to form a reference class. In the case of phylogenetic trees, for example, frequency probabilities would require a set of simultaneously possible trees. However, from their metaphysical argument it seems to follow that only one tree can be "true", and thus, that all others are necessarily false (Siddall & Kluge 1997: 317). This is a somewhat confusing claim, but it seems to entail that if it were the case that Life evolved differently than it did, then it would not be Life, but Life*, if we read them as saying only one tree is actual, and all others are impossible.

Another reason to be suspicious of the frequency approach is that it is supposed to converge on the correct hypothesis, the correct phylogenetic tree or clade, as the data becomes infinite. However, "neither a correct tree nor infinite data [will] *ever*. . . be available" (Siddall & Kluge 1997: 319). Also they reject the frequency probability approach to phylogenetics because "we are not faced with large numbers or repetitive cases. So long as time is taken to be linear, history has occurred but once. Large numbers and their generalities cannot be relevant to finite singular cases" (Siddall & Kluge 1997: 331). So for these reasons, in phylogenetics we should not care about "the long run", but only about particular cases. Furthermore, CP, being based on logical probability, avoids these problems, and is to be preferred to ML, which does not.

7.3 Reference Class Solutions

In regards to the *argument from probability* de Quieroz & Poe (2003) do not directly respond to Siddall & Kluge (1997), but respond by way of Kluge's (2001a) attack on de Quieroz & Poe (2001). This in itself presents a problem for an adequate reconstruction of de Quieroz & Poe's (2003) response, because it seems clear from a careful reading of all three papers that the original worry of Siddall & Kluge (1997) is never addressed.[23] In order to make up for this deficiency, we will consider, in part, how Haber (2005) has addressed the worry.

Haber's (2005) response to Siddall & Kluge's (1997) argument can be reconceptualised in terms of our above reformulation of *the argument from probability*. Haber's (2005) response involves arguing against P2, by noting a difference between our epistemic and ontological situation in phylogenetics. Instead of talking just about the actual evolutionary history of life, we must also talk about our hypotheses and models about this history. Only hypotheses can be said to be more or less likely in terms of the available data. The actual history of life just is. This distinction between our ontological and epistemological situation is similar to a

[23] That the worry is never addressed is clear when we read that de Quieroz & Poe (2003: 354) claim to be confused about the historical events that Kluge (2001b) is referring to. As we discussed above, Siddall & Kluge (1997) are making a more general claim about historical particulars in phylogenetics (with perhaps the limitations noted by Grant (2002)). These particulars include anything that is itself a part of Life: individual clades (monophyletic groups), lineages, organisms, synapomorphies, and evolutionary transformations. These historical particulars are unique and provide no relevant reference class for a frequency interpretation of probability. The response of de Quieroz & Poe (2003: 354) is a comparision of evolutionary transformations (one such historical event from phylogenetics) to coin tosses, which they also consider to be unique historical events. Coin tosses much more obviously form a reference class: they are independent of each other (unless the coin keeps changing its bias with each toss), and there are more salient features to distinguish them as a type of event than there are evolutionary transformations. That is, again, that there does not seem to be a relevant reference class for the unique historical individuals that phylogenetics concerns itself. Thus, de Quieroz & Poe's (2003) response seems inadequate.

more detailed distinction made by Vogt (2007). Vogt (2007: 402-403) argues that our epistemic situation is such that we have particular structures that are indistinguishable from each other. Consider the following oversimplified example. In molecular data an *A* that changed to a *C* and then back to an *A* over a period of time does not look different than an *A* that has remained constant over that period of time. All we have now is an *A* at a particular position in a gene sequence. Because of this, we need to classify the types of transformation processes using evolutionary models that can give rise to these indistinguishable structures. So, it seems reasonable to evaluate how frequently specific types of processes may have taken place that could have resulted in one of some set of indistinguishable structures. Let us now consider a brief and oversimplified example to help spell this idea out further.

So, for example, human beings lost their tails when they evolved from some more remote ancestor. We conjecture this for many reasons: all of our closest ancestors have tails, it seems our vertebrae are consistent with once having had tails, and human foetuses seem to begin to but then fail to develop a tail. The loss of our tails is a transformation event that we have inferred. On a most parsimonious reconstruction of this history, we would assume that this event happened once. However, there are many stories that are consistent with what we may observe today: one possible alternative being that we lost our tail, regained our tail, and then lost it again. This type of event is called a reversal. We may form a reference class of similar events from evolutionary reversals (e.g., whales went from the sea, to the land, to the sea again; frogs had teeth, lost them, and now some have them again), using this reference class, and the length of time between the distant ancestor of humans having a tail and today (no-tail), we can evaluate how likely it is that an evolutionary reversal has occured. If we assume that evolutionary reversal is always a possibility, since we cannot distinguish it from the transformation event occurring only once in history, the longer the period of time is, the more likely the evolutionary reversal is. This example is somewhat cooked-up, but it gets the general idea across. The type of example becomes more plausible and prevalent when we deal with molecular data, (As, Cs, Gs, and Ts), which would be even harder to distinguish by mere observation.

This example unfortunately does not make clear how exactly one would form a reference class. Because the events we wish to consider in phylogenetics can be classified in many different ways, it seems problematic to choose one particular way to classify them over others, because the probability of the event occurring will change dependent on the way we classify them. One way of dealing with this sort of problem may be to abandon a frequency interpretation of probability and use a propensity interpretation.

Popper's propensity interpretation of probability was developed, as Haber (2005: 836) rightly notes, to deal with problems with the frequency interpretation. Rather than take probability as a property of a given sequence, the propensity interpretation takes probability as a property of a physical state of affairs that has the propensity to produce a sequence. Under this interpretation a singular event does not need to be part of a given sequence to be assigned a probability. In the case of a single coin toss, the probability of heads would be determined by the physical properties of the coin. A singular event can be assigned a probability based on the physical state of affairs that produced it, even though it only occurred once. So now we can understand the actual evolutionary history of life as a singular event and "the probability of the actual character state distributions as a propensity of the conditions in which this distribution was produced. These conditions are generally described by biologists as models of evolution" (Haber 2005: 837). So this, it seems, qualifies as a viable alternative, although it has yet to be pursued in phylogenetics. Advocates of CP might respond here by asking whether propensities are playing the appropriate explanatory role in phylogenetics (Haber 2005: 837), but we will save our discussion of explanatory power for later on, and will approach it from a different angle than that pursued here.

In concluding Chapter Seven we should note what we have been able to show, and what we have been able to clear up. First, there is little motivation for claiming that corroboration pays heed to only the logical interpretation of probability. We cannot say that advocates of ML are bad Popperians, that they are inconsistent with corroboration on that grounds that ML requires a frequency interpretation. Logical, frequency, and propensity interpretations may all accord with corroboration. Second, likelihood and corroboration are not entirely at odds with each other, there are some relevant similarities between the two formulas. Third, there are problems with forming reference classes in phylogenetics, because the science concerns unique historical particulars. These difficulties are not insurmountable for the ML view, and I presented some strategies for responding to them. One strategy involved distinguishing our epistemic situation from our metaphysical situation. That is, there are epistemically indistinguishable structures that are metaphysically distinguishable, and certain types of processes give rise to indistinguishable structures. A second strategy, that Popper himself may have attempted, is to adopt a propensity interpretation of probability, but other than Haber (2005), no one has suggested this in phylogenetics. I will now move to discuss what else corroboration can tell us about the debate between ML ad CP in terms of severity of test.

Chapter Eight: Severity of Test and Phylogenetic Hypotheses

I have shown that corroboration is often simply identified by Popper as a report of the severity of tests that a hypothesis has passed relative to its competitors. The approach to interpreting severity of test in terms of CP is relatively straightforward, whereas the ML approach is somewhat more complicated. I will first unpack the somewhat baffling ML approach of de Quieroz & Poe (2001; 2003). Briefly, they wish to decouple test severity from corroboration, while retaining corroboration as a method to rank hypotheses, but only when we consider what they call "standard phylogenetic analysis." In tandem with decoupling test severity from corroboration, de Quieroz & Poe try to show that in the specific case of standard phylogenetic analysis we can ignore the term P(e,b).

Following the ML approach I will address the CP approach as a response to de Quieroz & Poe's position.

8.1 Ignoring Test Severity

Essentially de Quieroz & Poe want to argue that P(e,*b*) can be ignored, and that comparing degrees of corroboration is just comparing likelihoods under what they call "standard phylogenetic analysis." This strategy involves separating two notions pertaining to Popper's degree of corroboration: the evaluation of competing hypotheses; the evaluation of different tests in terms of severity (de Quieroz & Poe 2001: 318). They argue that we can ignore P(*e*,*b*) under standard phylogenetic analysis, which amounts to ignoring the evaluation of the severity of different tests.[24] However, they do want to retain the aspect of corroboration that allows one to evaluate competing hypotheses, and this they argue, is done by comparing their likelihoods (P(*e*,*h*). There are two stages to de Quieroz & Poe's (2001) argument, because there are two sorts of procedures that a ML analysis consists of. The first stage is to show that standard phylogenetic analysis tests a given set of trees with a particular set of evidence, *e*, under a particular phylogenetic method (e.g., CP or ML) (de Quieroz & Poe 2003: 360). The next is to show that a given set of models ($M_{1,}$ $M_{2,}$ $M_{3, \ldots}$) or methods (like CP) can be tested with respect to a single tree. In each stage de Quieroz & Poe provide reasons for ignoring P(*e*,*b*). After I go through the two stages of their argument, I will make it clear how they are misunderstanding severity of test.

[24] In their appendix they do note several situations, outside of standard phylogenetic analysis, where severity of test may come in to play in other relevant areas of biology.

I will now discuss the first stage of de Quieroz & Poe's argument in further detail. In this stage they wish to show that while we test competing tree hypotheses, our background knowledge, and our data set are both constant. Our background knowledge in phylogenetics, de Quieroz & Poe (2001) maintain, consists mainly of three groups of assumptions: descent with modification, the assumption that the relationships under examination conform to a tree-like pattern, and the assumptions required by whichever method we are using. The last group contains an optimality criterion, either CP or ML, and "various propositions concerning character transformation (e.g., character state order, character and state weights, transformation probabilities, among-site rate variation) . . ." (de Quieroz & Poe 2001: 312). So, we evaluate a given set of hypotheses (h_1, h_2, h_3, . . .), a set of trees or topologies, with a particular set of background knowledge, b, the latter conforming to the criteria just mentioned. That is, whatever b may be, it is held constant whenever we evaluate our given set of hypotheses; and also, they maintain that a given set of hypotheses will be evaluated by the same evidence e, which consists of observed character states. This now makes P(e,b) a constant for each hypothesis, h_1, h_2, h_3, . . ., thus, the corroboration value for each hypothesis is determined only by the value of P(e,hb). And this, they argue is equivalent to P(e,h) of the likelihood equation (de Quieroz & Poe 2001: 312).

The argument that P*(e,hb)* of the corroboration expression is equivalent to P(e,h) is based on de Quieroz & Poe's understanding of the model in statistical inference as equivalent to Popper's understanding of background knowledge. In

Popperian terms, their argument is that both the model and *b* are tentatively taken as "unproblematic." That is, models are assumed true for the purposes of testing competing tree topologies.

The second stage of de Quieroz & Poe's argument is to show that ML can assign a degree of corroboration to models (or methods like CP), as well as genealogical hypotheses. To evaluate models (or methods), we treat them as part of the hypothesis, *h*, and not as background knowledge, *b*. That is, we problematize our models. A hypothesis, *h*, under a likelihood analysis, is a model, *M*, plus a particular topology, *T*. We evaluate *h* in terms of some set of observed character data, *e*, by determining how likely *h* makes *e*:

$$\mathrm{P}(e \mid M,T)$$

Again, a set of hypotheses could be one model plus a set of particular topologies, or one topology plus a set of models. We have already discussed the former situation, so we need now only discuss the latter. In the latter case the models, each with its own particular components or parameters, would be problematized and our background knowledge, *b*, would consist in whatever assumptions are common to the models under consideration, e.g., descent with modification, and a particular assumed tree. In this case, as in the first stage of their argument, *b* would be the same for each competing model. Again, we would be considering the same set of data, *e*, as well. So, if both *e* and *b* are constant, then P(*e*,*b*) is constant, and we need only determine P(*e*,*hb*). Whichever hypothesis or hypotheses has/have the highest value for P(*e*,*hb*) is/are optimal. But, as they argued above, and as we will discuss further below, P(e,hb) of the corroboration expression is equal to P(e,h) of the likelihood expression. So whichever hypothesis or hypotheses has/have the maximum likelihood is/are optimal. So when evaluating different models determining the degree of corroboration is the same as determining their likelihoods.

So now that de Quieroz & Poe have showed that P(e,b) is constant under the two different stages of "standard phylogenetic analysis," they wish to show that P(e,b) can be ignored in these situations. Again, they want to claim that: "different kinds of analyses emphasize different aspects or components of corroboration: (1) the evaluation of rival hypotheses in terms of the results of a test—which is based primarily on $p(e \mid hb)$, and (2) the evaluation of different tests in terms of their severity—which is based primarily on $p(e \mid b)$" (de Quieroz & Poe 2001: 351). This disambiguation, I wish to suggest, serves no purpose.

To try and understand this further, let us first consider in more detail what Popper says about the relationship between likelihood and corroboration. Popper does not say that when P(e,b) is *constant* that likelihood will be a good approximation of corroboration. Popper says that likelihood will be a good approximation of corroboration only when P(e,b) is really small (Popper 1959: 413). Considering the formalism, the latter seems intuitively clear because if P(e,b) is very small then P(e,hb) – P(e,b) ≈ P(e,hb). So perhaps ignoring P(e,b) on the grounds that it is constant a misunderstanding.

Even if standard phylogenetic analysis is as described and P(*e*,*b*) is constant, then surely we can still evaluate for test severity. It does not matter that it is one test, one group of data, *e*, which evaluates different hypotheses, given that the background knowledge is also constant. One and the very same test can still have relative degrees of testability for competing hypotheses. That is, if each hypothesis can be assigned a different likelihood, then certainly we could subtract some constant value for P(*e*,*b*) from that likelihood. So, for example, if we have three hypotheses: h_1, h_2, and h_3, where the value for P(*e*,*hb*) with respect to h_1 = 0.8, h_2 = 0.7 and h_3 = 0.6, then h_1 would be the ML preferred hypothesis.. If we further have a constant value of P(*e*,*b*) = 0.2 for the analysis, then the situation would be as they described. Thus, with respect to a value for with respect to a value for P(*e*,*hb*) – P(*e*,*b*), we would get h_1 = 0.6, h_2 = 0.5 and h_3 = 0.4. With these latter numbers we can assess the severity of test.

I think it well help to understand my point if we consider what Sober (2004: 650) calls "ordinal equivalency":

> This idea is easy to understand by considering the Fahrenheit and Centigrade scales of temperature. These are ordinally equivalent, meaning that for any two objects, the first has a higher temperature-in-Fahrenheit than the second precisely when the first has a higher temperature-in-Centigrade than the second. The two scales induce the same ordering of objects.

My dummy numbers above gives to different rankings of hypotheses, but these two rankings are ordinally equivalent. Sober (2004) introduced ordinal equivalency to compare CP and ML, but I have only applied his notion to ML as conceived of by corroboration (by putting in some number for P(*e*,*b*)) or as conceived of by likelihood.

But, if we were to have an ordinally equivalent ranking of hypotheses, and a lower value for P(*e*,*b*) under CP than ML, say for example P(*e*,*b*) = 0.1, and this value is constant, then $h_1 = 0.7$, $h_2 = 0.6$ and $h_3 = 0.5$. That is, the measure of test severity would be higher under CP than ML, so its degree of corroboration would be higher. However, I must maintain that it is not clear how to get any real or precise values for P(*e*,*b*) under either CP or ML. I have just used dummy numbers to get the point across that severity of test cannot always be assessed with precision.

At any rate, de Quieroz & Poe are correct that when P(*e,b*) is constant, all we do is compare the likelihoods (P(*e*,*h*)) of competing hypotheses. ML can give us a precise value for P(*e*,*h*), whereas CP unfortunately cannot. CP merely focuses on the hypothesis with the fewest independent evolutionary changes. So it seems ML has a stronger claim to following Popper's formalism. But, though corroboration is a method for choosing among competing theories, it is a method to choose between them according to the severity of tests they have passed or failed. So to say that severity of test can be ignored seems to completely miss the point of what corroboration is supposed to formalize.

8.2 Increasing Test Severity

ML and CP have different background assumptions. P(*e*,*b*) for CP is *ceteris paribus* lower because it makes fewer background assumptions than ML. [25] That is, all things being equal, CP will conduct more severe tests than ML, which is a win by Popper's account. If we were to use corroboration to choose between all the trees

[25] This last claim, that CP has a lower value for P(*e*,*b*), will be explored further in the next section on explanatory power (Chapter Nine).

generated by both CP and ML, then CP seems to win. In other words, if we use Popper's idea of severity of test to test methods, not just trees generated within a single method, we no longer have constancy of P(*e*,*b*), and so we cannot just pay attention to likelihoods. Just paying attention to likelihoods is what ML wants to do.

So although ML may be able to test and problematize its background assumptions, this seems irrelevant to assessing severity of test, when compared with CP, if CP has fewer background assumptions. If, for any situation, two methods are ordinally equivalent, whichever method has a lower value for P(*e*,*b*) will be the method that tested the competing hypotheses more severely. If either CP or ML has a lower value for P(*e*,*b*) while being ordinally equivalent, then that method tests its hypotheses more severely. Because ML requires a model plus descent with modification, and CP, as Kluge understands it, only involves assuming descent with modification, it seems CP would, *prima facie*, have a lower value for P(*e*,*b*).

Nonetheless, this result seems to be a trade-off because ML has the advantage of testing different evolutionary models when it keeps tree topologies constant. Advocates of CP must downplay this advantage, which they do from the context of Popper's comments on background knowledge: "By our background knowledge *b* we mean any knowledge (relevant to the situation) which we accept—perhaps only tentatively—while we are testing *h*" (Popper 1983: 236). (Here *h* represents a particular hypothesis.) In CP we only, or so its advocates maintain, assume descent with modification.[26] At issue here is the claim, from advocates of CP, that ML makes inadmissible additions to background knowledge. The strategy that the CP camp uses against ML is to highlight the "tentative" part of the acceptance of background knowledge (Farris 2000: 385-386). Popper claims that hypotheses are also "tentatively" accepted when they have high degrees of corroboration. So, both corroborated hypotheses and background knowledge are "tentatively" accepted. Thus, background knowledge should consist of highly corroborated hypotheses, or at least should *not* consist of hypotheses that have a significant amount of evidence against them.

Models of the evolutionary process are of course often unrealistic idealizations, but this alone does not mean that they have a low degree of corroboration. Thus, the claim of Farris (2000: 391-392) that these models have a low degree of corroboration seems to require further motivation. Especially when Farris admits that the "No Common Mechanism" model of Tuffley & Steel (1997), which gives an equivalency between ML and CP, is realistic, and acceptable.[27] An ML advocate may maintain that they can test models, by moving them from acting as an auxiliary hypothesis (or background knowledge) to a main hypothesis, and that these models can acquire a degree of corroboration, even if they often are, in some sense, unrealistic.

[26] I will leave the claim that we only need assume descent with modification aside for now and will return to it when I speak more about explanatory power.

[27] He does of course give reasons other than the equivalency for the acceptability of their model: "Tuffley and Steel (1997) introduced a model called No Common Mechanism (NCM), in which characters may—but are not required to—vary their relative rates independently, both within and between branches. Because the independent variation is taken only as a possibility, not as a requirement, NCM would apply to almost any situation, and so may be accepted as realistic" (Farris 2008: 827).

In concluding this section, we may note that it seems hard for corroboration alone to decide between CP and ML in terms of severity of test. On the one hand, CP, as I have argued, seems to test its hypotheses more severely all things being equal. That is, if Kluge's version of CP is correct, then CP assumes less about evolution than ML, which makes the evidence (character data) less probable given CP's assumptions alone, which makes its severity of test values higher than ML for genealogical hypotheses, even if we assume that ML and CP are ordinally equivalent. This point about severity of test became even clearer after we deconstructed de Quieroz & Poe's (2001) misguided argument that P(*e*,*b*) can be ignored under "standard phylogenetic analysis." However, this claim by advocates of CP about minimum background knowledge is continually ignored by advocates of ML, who believe that within the CP method lies an implicit model of evolution which it can test by ML's own methodology, provided that this model is made explicit. It is true that ML can test various models against each other by keeping a specific topology constant, but an advocate of CP will retort that these models are all designed to be unrealistic, and will try to show that evolutionary models have a low degree of corroboration. In the next section we will further examine, as we compare ML and CP under an analysis of explanatory power, to what extent the assumptions of CP are more minimal than ML.

Chapter Nine: Phylogenetic Hypotheses and Explanatory Power

In Chapter Three I showed how Popper defined explanatory power in terms of probabilities that pertain to the evidence, *e*, the hypothesis, *h,* and background knowledge, *b*, and with the following formal representation:

Explanatory Power

$$\mathrm{E}(h,e,b) = \frac{\mathrm{P}(e,hb) - \mathrm{P}(e,b)}{\mathrm{P}(e,hb) + \mathrm{P}(e,b)}$$

The important thing to note about Popper's definition of explanatory power, as his definition of severity of test, and corroboration, is the probability values in the numerator. That is, the higher the value of $\mathrm{P}(e,hb) - \mathrm{P}(e,b)$ is, the higher the explanatory power of the hypothesis is.

Conjoined with Popper's definition of explanatory power is his convention to reject *ad hoc* auxiliary hypotheses. A hypothesis is considered *ad hoc* when it is brought in only to explain the evidence that the main hypothesis cannot. Thus, an *ad hoc* hypothesis diminishes the explanatory power of the main hypothesis that it supports.

The strategy for CP is to show that their methodology maximizes explanatory power over ML along these two lines of thought. In section 9.1 I will discuss their argument that CP requires the minimum amount of background knowledge (has a low value for $\mathrm{P}(e,b)$), which increases the explanatory power of the hypotheses it tests (has a high value for $\mathrm{P}(e,hb) - \mathrm{P}(e,b)$). In section 9.2 I will discuss their argument, which claims that CP prefers main hypotheses that require the minimum amount of *ad hoc* auxiliary hypotheses.

Following this discussion, in section 9.3 I will explore an argument employed by the ML camp to cause problems for these two CP strategies. This is the argument that CP has an implicit model.

9.1 Descent with Modification

The general strategy for CP, in terms of adhering to Popper's understanding of explanatory power, is in many ways similar to the strategy for adhering to his understanding of severity of test that I considered in Chapter Eight. In this section I will show how advocates of CP argue that evolution, understood as descent with modification, is the only required part of our background knowledge.

There are two possible responses to this strategy. The first is from ML advocates, who argue that more assumptions are required (e.g., de Quieroz & Poe 2001; 2003). The second is from pattern-cladists or transformed cladists, who argue that fewer assumptions are required.[28] Of the latter view Kluge (2001a) already provides some serious philosophical criticism, and since de Quieroz & Poe (2001; 2003) accept at least his minimum assumptions we need not discuss the second response to the general CP strategy. I will now briefly elaborate Kluge's (2001a) view that evolution, understood as descent with modification, is necessary and sufficient for explanation in phylogenetics.

[28] Pattern-cladists (e.g., Platnick 1979; Brower 2000) reject the requirement of evolution from our assumptions. Typically pattern-cladists are skeptical about reconstructing phylogeney; this is because they're worried about circularity: phylogenetic trees test evolutionary hypotheses, but are constructed by using evolutionary theory. One of Andrew Brower's (2000) main points in favour of his view is that pattern-cladism increases explanatory power, because it requires fewer assumptions. Kluge (2001a) counters this point by arguing, along the lines of the deductive-nomological model of explanation, that Brower's pattern-cladism fails to be explanatory because there is no general law with which one can explain shared derived traits.

Kluge (2001a), following his earlier work (Kluge 1984: 26), distinguishes between two types of parsimony in cladistics: what he now calls phylogenetic and plausibility parsimony. This distinction is made because a most parsimonious cladogram may either minimize *ad hocisms* of homoplasy (phylogenetic parsimony) or postulate fewer natural processes (plausibility parsimony) (Kluge 2001a: 201). Kluge is not exactly clear about this distinction, but the point seems to be that plausibility parsimony minimizes natural processes, if only implicitly, by assuming *a priori* that some type of evolutionary change is improbable. That is, Kluge maintains that plausibility parsimony is the types of parsimony that de Quieroz & Poe (2001) believe contain implicit models, whereas phylogenetic parsimony has no such implicit model. Kluge wishes to advocate phylogenetic parsimony as understood by Farris (1983), which, it can be argued, operates only by minimizing *ad hoc* hypotheses of homoplasy. I will discuss Farris' (1983) argument for this view of parsimony in the following section.

Phylogenetic parsimony requires the assumption of evolution. Kluge (2001a: 201) understands evolution to be Darwin's principles of "descent with modification,"[29] which he interprets as follows:

Descent	*Modification*
Species evolve from other species, as opposed to species being created independently of one another.	The traits of species are transformations of prior states, as opposed to abstract timeless, relations of character states

[29] Although Kluge now states that this is all that is required as background knowledge, he does explicate in an earlier coauthored paper (Farris *et al.*1970: 172-174) a more rigorous understanding of the required assumptions (there called axioms) of phylogenetics

Kluge is never entirely clear on how these principles operate in parsimony analysis, but when I consider Farris' (1983) argument in section 9.2, we can motivate their denial of the charge from the ML camp that I will consider in more detail in section 9.3.

At any rate, Kluge and others from the CP camp (e.g., Farris 2000) do not want models (or at least unrealistic models) to be part of our background knowledge. However, there are other things that they will admit: "[B]ackground knowledge might be anything else one claims to know . . . concerning species history, but not the phylogeny in question" (Kluge 1997a: 88). Essentially this means that for a given hypothesis about phylogeny, *h*, we may test *h* in view of other well-corroborated hypotheses that pertain to other species that are related, but are outside the group of taxa we wish to consider.

9.2 Homoplasy and *Ad hoc* Manoeuvres

I will now consider in more detail why the most parsimonious hypothesis is the most explanatory hypothesis in terms of its dismissal of *ad hoc* hypotheses. This is a difficult point, so I will try stating it in a couple of different ways. It is helpful to lay this idea out by framing it as a response to a very common objection to CP. The objection is that CP assumes homoplasy is rare in evolution (e.g., Felsenstein 1978; Saether 1986: 4). Kluge (2001a), again, believes that this objection is formed on a failure to delineate between phylogenetic parsimony and plausibility parsimony. Before I get to a detailed discussion of this objection and response, I will briefly frame four justifications for the CP.

CP chooses the hypothesis that requires the smallest number of independent evolutionary changes in order to explain the observed shared-character data that the taxa under consideration has. A homoplasy gives us two different evolutionary steps for one observed shared-character state, so CP chooses the tree with the least amount of homoplasy. Since CP reduces the amount of homoplasy, people who prefer other methods have thought that CP has an implicit model that assumes that homoplasy is rare.

However, there are several justifications that CP advocates have used for their procedure that must be considered. Here are four:

(1.) Hennig (1966): homoplasy should not be postulated beyond necessity.

(2.) Wiley (1975): the most parsimonious genealogical hypotheses are least falsified by the available evidence.

(3.) Farris (1983): the most parsimonious genealogical hypotheses are those that minimize requirements for *ad hoc* hypotheses of homoplasy.

(4.) Kluge (1997a): the most parsimonioius genealogical hypotheses are the most corroborated hypotheses.

None of these justifications are strictly speaking incompatible, but can be seen as a progressive elaboration on a theme. Each of (1-4), explicitly or implicitly, gives reasons for saying that the genealogical hypothesis to be preferred is the one that requires the fewest homoplasies (the fewest similar character traits that have evolved from separate origins). In what follows, I will focus largely upon elaborating

(4), in terms of explanatory power, but this elaboration requires some help from the other justifications.

The most common objection to CP is that it assumes unrealistically that homoplasy is rare in evolution. This assumption would be unrealistic because we know that a myriad of homoplasies have occurred. For example, in unrelated lineages many different animals have developed "wings" for flight. The objection stems from two sources: partly because CP does indeed minimize homoplasy, and partly on the basis of elaborate statistical arguments.[30] But what is relevant for my purposes here, is that the objection implies that CP assumes something more about the evolutionary process than just descent with modification: that is, CP assumes that homoplasy is rare.[31]

The objection that CP assumes that homoplasy is rare is met with responses like the following. The more homoplasies required by a genealogical hypothesis, the more evidence is unexplained: "The explanatory power of a genealogy is . . . diminished only when the hypothesis of kinship requires *ad hoc* hypotheses of homoplasy" (Farris 1983: 23). Only homoplasies that are merely implied by a tree, that have no supporting evidence of their own, are considered *ad hoc* (Farris 2008: 826). A genealogy, given descent with modification, does not explain shared derived characters that are incongruent with it: "A genealogy is able to explain observed

[30] Unfortunately I do not have space to analyze the strengths and weaknesses of such statistical arguments in detail, but these should not play a proper role in our analysis of the Popper debate.

[31] This objection may have been initially levelled more at Hennig's (1966) maxim, so I have chosen to deal with the response in Kluge's (1997a) terminology, who relates the justification of parsimony to corroboration, which is what we, after all, are interested in. Be that as it may, Kluge's (1997a) work is not meant to be an outright rejection of the work of Farris (1983), Wiley (1975), or even Hennig (1966), but an elaboration.

points of similarity among organisms just when it can account for them as identical by virtue of inheritance from a common ancestor" (Farris 1983: 18). The explanatory power of a genealogy is reduced each time it has to give a different origin for a shared feature (Farris 1986: 15). If wings are shared by storks and robins by way of a common ancestor, then a genealogy that reflects this is explanatory. On the other hand, "wings" are shared by birds, butterflies and bees, but not by way of a common ancestor, so we need to appeal to other considerations, perhaps environmental, to provide an explanation of why they all share "wings."[32]

To explain this response further let us consider a hypothetical example (modified from Farris 1983: 13-14; Sober 1983: 341-342). Consider three taxa, A, B, and C, and 10 independent characters that are weighted equally.[33] From the assumption of bifurcating evolution, there are three hypotheses for our three taxa, three logical possibilities for how they may be related to each other:

$h_1 = \text{(AB)C}$

$h_2 = \text{(AC)B}$

$h_3 = \text{(BC)A}$

[32] We also want homoplasies to be independent hypotheses: "If two characters were logically or functionally related so that homoplasy in one would imply homoplasy in the other, then homoplasy in both would be implied by a single *ad hoc* hypothesis (Farris 1983: 19-20). That is to say, only independent lines of evidence should be used in evaluating genealogies. Homoplasies conflict with genealogical hypotheses that try to account for why characters are shared among organisms through descent. We look for extrinsic evidence to see if we can resolve the conflict, re-evaluating the characters under consideration and the like, but: "If a conflicting character survives all attempts to remove it by searching for such evidence, then the conclusion of homoplasy in that character . . . satisfies the usual definition of homoplasy" (Farris 1983: 10).

[33] See Kluge (1997b) for why an explanation of why his version of CP requires equally weighted characters along the lines of corroboration. I have unfortunately had to bracket this issue in in the interests of space.

The notation "(AB)C", for example, indicates the hypothesis that the taxa A and B are more closely related to each other than either are to taxa C. We use the discovered 10 characters to test these three hypotheses. Now, suppose that we discover that 9 of our characters are *potential* synapomorphies (shared-derived states) consistent with h_1, and that these characters are all in the ancestral state for taxa C. And suppose that we discover that character 10, on the other hand, is a *potential* synapomorphy consistent with h_3, and that this state is in the ancestral position for taxa A.

CP interprets this data as supporting h_1 over h_3. The reason is that h_1 requires only 1 homoplasy to account for character 10, because h_1 alone cannot explain that character as a result of a shared lineage. Similarly h_3 requires 9 homoplasies to account for characters 1-9. That is, h_1 requires fewer evolutionary steps than h_3. Farris' (1983) idea is that h_1, for example, only "speaks" about the 1 incongruent character: h_1 says that that character is a homoplasy. (Whereas, h_3, on the other hand, "says" more, it says that characters 1-9 are homoplasy.) But, h_1 is "silent" about the remaining 9 characters (they may be either homoplasy or homology). Thus, CP does not involve an *a priori* reduction of the amount of homoplasy in the world, since the hypotheses it prefers are silent about the characters that are congruent with them.

But we need to say more about how evidence, putative synapomorphies that are incongruent with a genealogical hypothesis, counts against that hypothesis. If we recast the above example from Farris (1983) in Wiley's (1975) terms, h_1 would be refuted if character 10 were in *reality* (not just *potentially*) a homology, and h_3 would be refuted if any of characters 1-9 were in reality a homology[34], so we choose the one that has more potential falsifiers. There is a problem here with our epistemic situation. We can never really know whether a character is a homology, we can only infer it from the available present-day data. I will now consider Sober's reconstruction of Wiley's (1975) position, in order that I may explain further how evidence counts against hypotheses, but we must keep in mind, from Farris' (1983) argument, what a hypothesis "speaks" about, and what it is "silent" about.

How do homoplasies required by a hypothesis falsify it? Since h_1 requires character 10 to be a homoplasy, this hypothesis then, strictly speaking, would be false. One might accept that h_1 is indeed false, but that given the evidence available it is the one best corroborated (or the one that is the least false). But this seems like a poor way of conceptualizing what is going on.

As Sober (1983: 339) explains, this type of situation cannot be an example of what he calls "strong falsification," where a statement, here the one required homoplasy (or incongruent derived state), *deductively* implies the falsehood of h_1 (or at least that either h_1 or the character distribution is false). So under CP, if h_1 requires the dismissal of the character distribution as false, then this is an *ad hoc* requirement, because an *ad hoc* hypothesis dismisses evidence as false. But, since homoplasy is *always* a possible interpretation of a character, as we saw Farris (1983) pointing out, character distributions themselves cannot deductively imply the falsehood of a cladogram.

The situation must therefore be an example of “weak falsification,” where the one homoplasy required by h_1 counts against it, but does not deductively entail that h_1 is false. Thus, CP prefers the hypothesis with the least evidence against it.[35] Kluge (1997a: 87) recasts the result of this hypothetical example along Sober’s (1983) lines as follows:

1. A cladogram alone does *not* imply the derived states of a congruent synapomorphy are homologous (are of a common origin);
2. A cladogram *by itself* does imply the derived states of an incongruent synapomorphy are homoplasious (are of independent origin).

So, again, CP is forced to minimize hypotheses of homoplasy by choosing the cladistic hypothesis that requires the least instances of homoplasy, because a hypothesis of homoplasy reinterprets evidence that would otherwise count against a hypothesis. CP dismisses hypotheses of homoplasy because they are *ad hoc*; and if we were allowed to invoke hypotheses of homoplasy whenever we like, in order to save a particular cladistic hypothesis, then we would be obfuscating the relationship between theory and evidence, and letting our subjective judgments guide our choice. So, the hypothesis with the least *ad hoc* manouevers, at least in the case of the most parsimonious cladogram, is the most corroborated hypothesis, because of the relationship between *ad hoc* hypotheses and explanatory power. Homoplasies cannot be explained by particular cladograms, but homologies can. Given descent with modification, some shared-derived states, and a cladogram, we

[35] Vogt (2007; 2008) tries to reopen this point by claiming that cladograms are unfalsifiable in principle, and that they represent metaphysical hypotheses. Although he does reference Sober (1983), he does not address the adequacy of Sober’s distinction between “strong” and “weak” falsification. As well he quotes Kluge (1997a) as addressing this issue in similar terms, but provides no argument against it.

can explain the congruent derived states, the derived states that do fit with the cladogram, as homologies. But, the more homoplasies a cladogram needs, the less it explains.

In the next section I will briefly recapitulate why advocates of CP consider their method to be superior to ML in terms of explanatory power. Following this I will reconstruct de Quieroz & Poe's answer to this strategy, which picks up where the most common objection to CP, that it assumes homoplasy is rare, leaves off.

8.3 Implicit Models

Advocates of CP argue that they can select hypotheses with higher explanatory power than those selected by ML. This argument is relatively straightforward, and is very similar to the strategy with severity of test, but requires that CP has a set of necessary and sufficient assumptions for inferring phylogeny. If CP's assumptions are necessary and sufficient, then any additional assumptions, like the evolutionary models of ML, are superfluous. This is because ML already assumes descent with modification. If additional assumptions do not increase the degree of explanatory power, they are to be rejected.

ML is a very complicated method, and the hypotheses ML considers optimal, like many scientific hypotheses, do not of themselves yield probabilistic predictions, but require probabilistic assumptions (collectively termed "the model") to be evaluated with respect to likelihood:

> Cladistic groupings do not, by themselves, confer probabilities on character distributions . . . However, when transition probabilities are assigned to the branches of a tree and the root is assumed to be plesiomorphic . . . , likelihoods become well defined (Sober 1983: 345).

In Popperian terminology, models would be auxiliary assumptions. Models in phylogenetics are probabilistic descriptions of the evolutionary process that generated the observed character states (our evidence). ML needs some (probabilistic) model, but no specific set of such models, and because CP, as its advocates argue, requires no model, it is superior in terms of explanatory power. This of course is an oversimplification, but I will fill it out further when I consider de Quieroz & Poe's (2001) response to this strategy.

The response of de Quieroz & Poe's (2001: 312-313) is, in a sense, a form of *tu quoque*, that is, they argue that CP also involves a model of the evolutionary process[36]. Unfortunately, CP has an implicit model, which is therefore untestable. ML is superior to CP because it makes its own assumptions explicit thereby making them testable. de Quieroz & Poe's strategy I have already discussed somewhat in the section on severity of test, but I will consider it in more detail here and approach from a somewhat different angle. The "implicit model" strategy is often applied against CP by its various detractors; one form of this argument, which contends that CP assumes homoplasy is rare, I already addressed. However, this type of argument is put in a different light here because de Quieroz & Poe make it

[36] The recent result of Tuffley & Steel (1997) gives us a likelihood model called 'No Common Mechanism' (NCM) with explicit probabilistic assumptions. NCM gives equivalent results to CP. But an evaluation of this kind, whether it is with NCM, or with some other model, although important, is strictly speaking not relevant when working within the context of corroboration: "The fact that parsimony can be derived under very different types of models also casts doubt on the notion that one can evaluate a method justified on logical grounds by simply evaluating statistical models that happen to produce similar results" (Goloboff 2003: 92). In fact, the whole argument that de Quieroz & Poe (2001) provide to show that CP has probabilistic assumptions comes indirectly from the basis of statistical, and performance based theoretical virtues. These virtues may need to be evaluated against the Popperian virtues that we are exploring, but within the context of corroboration, and this is what de Quieroz & Poe (2001; 2003) claim to be entering, these virtues need to be put aside.

in a Popperian guise, rather than on the basis of statistical, performance based analyses of competing methodologies.

The argument that CP involves implicit and untestable probabilistic assumptions is made by de Quieroz & Poe (2001: 312-313) on the basis of their interpretation of Popper's understanding of background knowledge. The specific pieces of background knowledge in dispute are those made with a particular optimality criterion, whether it is ML or CP. These, they contend, are "various propositions concerning character transformation (e.g., character state order, character and state weights, transformation probabilities, among-site rate variation) . . ." (de Quieroz & Poe 2001: 312). Corroboration is defined in terms of probabilities, and for them corroboration in phylogenetics is simply P(e,h). Thus, Popper's degree of corroboration requires probabilistic assumptions, which are met by the use of explicit probabilistic models, the bread and butter of ML. Unfortunately CP lacks these explicit models; CP cannot provide a "basis for translating the minimum number of character transformations required by a tree into the probability of the observed character states among taxa given that tree" (de Quieroz & Poe 2001: 312). Essentially, without saying something more about the evolutionary process, at least implicitly, CP cannot calculate the degree of corroboration of any particular tree.[37] de Quieroz & Poe's argument is somewhat confusing, because it involves the premise that P(e,h) = C(h,e,b), which I showed was based on misunderstanding that severity of test can be ignored in standard

[37] What these implicit assumptions may be de Quieroz & Poe do not discuss, but they make a referral to other papers regarding the precise nature of the probabilistic assumptions that are required by CP (de Quieroz & Poe 2001:313). This issue is not relevant to my thesis, so I will forego it.

phylogenetic analysis. However, I will try and make a case for their position independently of this confusion.

One can re-construe de Quieroz & Poe's (2001) argument in even more Popperian terms, without having to make $P(e,h) = C(h,e,b)$. With this in mind, the objection could be construed as follows. The use of a model for ML gives an added explanatory bonus: a model can explain the probability of observed character states with precision among taxa given a particular tree. Advocates of CP do admit that "[p]arsimony seems to evaluate no precise probabilities such as likelihood . . ." (Farris 2000: 391). But the CP camp believes this is because they follow Popper's corroboration, which uses background knowledge, and that models are inadmissible as background knowledge. The tension between the two methods is here in some ways difficult to resolve. From what I have considered in the discussion of severity of test there does not seem to be any principled reason why we cannot allow models, at least tentatively, into our background knowledge. With this in mind, perhaps the tension is coming from another source. That is, there is an incompatibility in the aims of explanation between ML and CP. ML requires some information to be inferred about branch length for its models to operate, (under some simpler models multiple characters are assumed to have the same branch length), CP pays no attention to branch length, but chooses the most parsimonious hypothesis that requires the least amount of *ad hoc* hypotheses of homoplasy.

If this analysis is correct, then it is not clear that Popper's understanding of explanatory power alone can decide between ML and CP. What it may come down to is a debate about the goals of phylogenetic systematics. Either we are interested in reconstructing phylogeny alone, or we wish to say how our reconstruction can confer probabilities on observational data, given some more information about branch length. This debate, in turn, may be supported by a metaphysical debate about which sort of interpretation of probability is applicable to an explanatory and historical science. However, de Quieroz & Poe (2001) make a separate argument based on an analysis of one of Kluge's (1997a: 88) arguments, which we will now consider.

A further criticism of Kluge's (1997a) argument that descent with modification is a sufficient assumption to justify CP is offered by de Quieroz & Poe (2001: 313-315). Kluge's (1997a: 88) argument asks us to consider three taxa, A, B, and C. He argues that with only descent with modification as our background knowledge the synapomorphies that characterize (AB)C, (BC)A, and (CA)B should be all equally likely. These synapomorphies can be represented with the following notation, where 110, 011, or 101 respectively characterize the above hypotheses. The notation '1' means a derived state and '0' an ancestral state. So, for example, if we tag a character as '110', then both the taxa A and B share it in its derived state, and C has it in its derived state. (Kluge seems to mean that they are as equally likely as well as 000, 111, etc., but perhaps excludes them, because these latter character patterns, in not being synapomorphies, are uninformative about ancestry). But, Kluge argues, if our evidence is a large majority of synapomorphies that characterizes one of the hypotheses, say (AB)C, then this would be unlikely given *only* the background knowledge of descent with modification (making P(*e*,*b*) small) but it would not be unlikely given background knowledge plus a postulated rooted cladogram (AB)C (making P(*e*,*hb*) large). Thus, the cladogram (AB)C is corroborated to the degree to which synapomorphies (the observed derived character states) that characterize the taxa A and B are observed, because the value for P(*e,hb*) – P(*e*,*b*) increases the more synapomorphies that characterize (AB) are observed.

Contrary to this, de Quieroz & Poe (2001: 314) argue that descent with modification by itself cannot imply anything about the probabilities of the different possible character patterns for the three taxa considered. That is, descent with modification cannot infer that the distributions 000, 100, 010, etc., are all equally likely. To do this one requires a postulation of a probability distribution or a process for generating such a distribution, and they quote Popper to back them up on this: "statistical conclusions can only be derived from statistical premises" (Popper 1959: 208). Unfortunately, de Quieroz & Poe argue, descent with modification meets neither requirement, and: "If no probability distribution or generating process is specified, then no distribution of states is any more or less likely than any other, and this is true with the additional assumption of a tree-like model of descent" (de Quieroz & Poe 2001: 314). Achieving a degree of corroboration *C* for a particular tree requires P(*e*,*hb*), but the *b* (descent with modification) is not enough to calculate P(*e*,*hb*): "Kluge's statement that certain character parameters should be equally likely, given only the assumption of descent with modification has no basis.

Instead, the statement itself is an additional unjustified probabilistic assumption" (de Quieroz & Poe 2001: 314). Thus, descent with modification is insufficient background knowledge.

But perhaps, when Kluge said that certain character parameters should be equally likely, given only descent with modification, he made a slip of the pen. Alternatively, he might have meant "equally likely" in some non-technical sense. At any rate, this is certainly a mistake on his part, and it has obfuscated the clarity of his position. Thus, de Quieroz & Poe are right for catching him on this. Next I will consider de Quieroz & Poe's next move, and the way they bring branch length to the fore.

de Quieroz & Poe (2001) recast Kluge's (1997a) argument (that we considered above) to consider differences in branch length as relevant. So, again, let us consider '0' to be an ancestral state and '1' a derived state, and the hypothetical ancestor to have all characters with state 0. Suppose again that the evidence, *e*, exhibits more characters with the pattern 110 in taxa ABC relative to 101 and 011. Given this information they argue that:

> it would be incorrect to conclude that characters with the pattern 110 have the highest probability for the topology (AB)C. The reason is that the probability of the pattern . . . is greater on one (or both) of the alternative topologies than on . . . (AB)C . . . for certain patterns of inequality in the probabilities of change among branches (de Quieroz & Poe 2001: 314).

Under standard ML, sharing more derived traits (synapomorphies) does not necessarily imply a closer relationship between taxa (de Quieroz & Poe 2003: 354). ML was originally developed to deal with the simulated long-branch attraction in the “Felsenstein Zone” (hereafter “LBAP” for “long-branch attraction problem”). ML uses models to deal with LBAP by modeling the probability of character transformation, i.e., homoplasy, as a function of branch length (de Quieroz & Poe 2003: 354). Some authors (Felsenstein 1978) state that ML should be used as a last resort (during long-branch attraction cases), rather than as a general principle.

To understand LBAP further let us examine de Quieroz & Poe’s (2001: 315) theoretical example, which follows the work of Felsenstein (1978). Consider three taxa, A, B, and C, where A(BC) *ex hypothesi* is the true grouping. If the probabilities of character change on the branches that end in A and B is high, but the one that connects the common ancestor of B and C to C is low (see Fig. 4 below), then the probability of observing a shared derived state (synapomorphy) in A and B, which does not occur in C, as the result of convergence (homoplasy) is high, and the probability of observing a derived state (synapomorphy) in B and C, which does not occur in A, as the result of inheritance (homology) is low. And even if the latter, homology, does occur, the probability of a reversal on the long B branch is also high. Because of this, the character pattern 110 is higher, so CP would choose (AB)C, even though, *ex hypothesi* it is false and A(BC) is true.

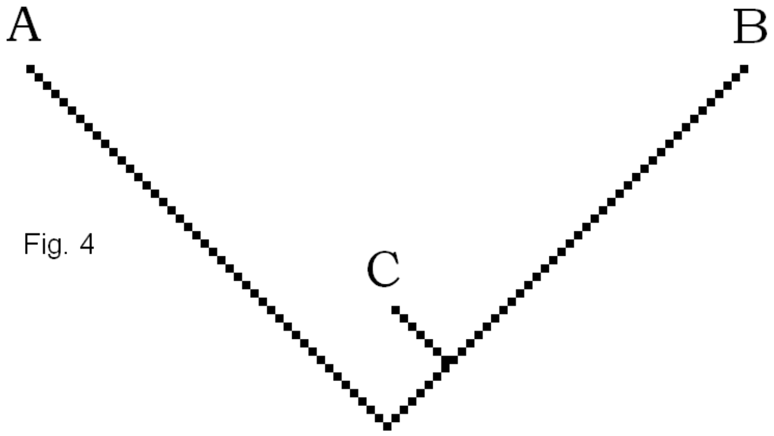


Fig. 4

Because of LBAP, the numbers of shared derived states that are the result of homoplasy are related to the lengths of the various branches of the tree by way of a model. So, for AB to be more corroborated the taxa A and B must share more derived states with each other than they do with C than would be expected given the lengths of the various branches. For AB to be contradicted (falsified, less corroborated) either A or B must share more derived states with C than with each other than would be expected given the lengths of the various branches.

In response to de Quieroz & Poe's point about branch length, we may in some respects say that a decision here goes beyond the bounds of merely applying corroboration to phylogenetic systematics. This debate involves determining whether branch length is relevant to inferring phylogeny. Branch length was shown to be relevant in a simulation where ML outperforms CP in the "Felsenstein Zone" (where two non-sister branches are long and others are short), but this has to an extent been rebutted, by showing that CP outperforms ML in a simulation that concocts the "Farris Zone" (where two sister branches are long and others are short). Yet, even given this rebuttal, most advocates of CP do not put much stock in simulations: "These simulations serve only to illustrate how competing discovery operations[38] may potentially mislead in different ways but they do not provide a rational basis for choice of discovery operation" (Grant 2002: 103).

Again, what this issue about branch length may come down to is a debate about the goals of phylogeneticists. That is, whether or not branch length is relevant to determining phylogeny. Those who advocate CP and also respond to the "Felsenstein Zone" with the "Farris Zone," seem to believe that it is relevant (e.g., Siddall 1998). Others put forth arguments dismissing the ferocity of the outperformance of CP by ML in mere simulations (e.g., Farris 1983; Sober 1983). One may also respond that none of these theoretical arguments seem to tell us anything about what evolution is actually like, but only how either ML or CP may potentially mislead us if we are actually in one of these "Zones" (Grant 2002: 103). So, for de Quieroz & Poe's argument to work they would perhaps need to make it independently of any performance simulation, otherwise it seems irrelevant to an application of corroboration to phylogenetic systematics, because it will be dismissed for the same old reasons. Alternatively, they might attempt a similar argument, in the context of corroboration, for the inclusion of branch length as relevant to inferring phylogeny, as perhaps an explanatory theoretical term, but it is unclear how to formulate such an argument.

In concluding this section, it seems even harder for corroboration to decide between ML and CP in terms of explanatory power, than it was in terms of severity

[38] By "discovery operations" Grant is referring to methods like ML and CP.

of test. On the one hand, CP makes fewer assumptions, so according to Popper's understanding of explanatory power, the hypotheses CP generates have greater explanatory power. Since ML requires models, the hypotheses ML considers have lower explanatory power. Advocates of ML counter this move by stating that their method can explain more things, and that the assumptions CP makes are insufficient to give the explanations that we are interested in. Advocates of CP can then counter this move by saying that we are not interested in making these explanations; our method reconstructs phylogeny, which is what we should really be interested in, without models. I have argued that this shows a tension of interests that corroboration itself cannot resolve.

Part Four: Concluding Remarks

Chapter Ten: Analyzing the Results

10.1 Corroboration and the Popper Debate

If one scores the points I have made here for or against either CP or ML on the basis of corroboration, then it seems we are not left with a clear winner. Advocates of CP have a longer history of basing their view on Popper's epistemology, but CP does not have as strong of a claim as they make out. I have shown reasons to believe, assuming Popper's corroboration to be the correct scientific methodology, that an application of corroboration to phylogenetics does not exclude one method or another. That is, it is clear that arguments for the exclusivity of CP do not follow from simply applying corroboration to phylogenetics. This becomes especially clear in the CP-defensive strategy against statistical models being used to understand their methodology:

> Advocates of parsimony generally deny the applicability of statistical methods because of the detailed information about evolutionary processes that is required by the statistical models (Farris 1983), the necessary uniqueness of the historical events phylogenetic methods aim to discover (Siddal & Kluge 1997), and the superfluity of the assumptions of statistical models (Kluge & Grant 2006) (Grant & Kluge 2008: 1052).

These three different strategies are employed defensively, but also on the basis of retreating to corroboration as a justification, independent of any statistical justification for CP. The order of this dialectic can be conceived as follows: ML tries to understand CP statistically, in response CP employs defensive strategies against justifying their method statistically, and then attempts to justify their method in terms of corroboration. Of the three types of strategies mentioned by Grant & Kluge (2008) in the above quotation I have paid the most attention to the latter two. In the remainder of this section I will clarify to what degree these defensive strategies relate to a grounding of CP in terms of corroboration with respect my evaluation of the Popper debate in phylogenetics in terms of probability, severity of test and explanatory power. I will also review the extent to which ML can base itself in corroboration.

In terms of probability interpretations, there is little guidance to be offered from corroboration, because logical, frequency, and propensity interpretations all accord with it. Thus, we cannot simply apply corroboration, and determine which method is correct based on which interpretation of probability it favours. There also do not seem to be any relevant differences between likelihood and corroboration with respect to the general attitude they both display toward the probability calculus. For example, both are designed to assess the probability of our evidence in terms of our conjectured hypotheses, and not the probability of hypothesis in terms of our evidence. There are problems for ML, in that it favours a frequency approach to probability, because there are reference class problems in phylogenetics, provided we accept some version, similar to Siddall & Kluge's (1997) version of species as individuals. These problems are not entirely insurmountable, but advocates of ML have to play defence. They either have to reject the frequency interpretation and adopt the propensity interpretation, or keep the frequency interpretation, but make a sharp distinction between our epistemic and metaphysical situation with respect to the content of phylogenetics. On the other hand, the logical interpretation of probability, that Siddall & Kluge (1997) claim CP favours, is not entirely spelled out, and is further confused by their claim that corroboration “exemplifies” said interpretation. In accordance with the logical interpretation of probability, there is a clear rubric for determining how many possible bifurcating cladograms there are given a set of taxa. This rubric does of course simplify the issue at hand, and ignores the possibility of hybridization, and reticulation (although this problem may also apply to ML) amongst other things. At any rate, if the CP camp is basing their allegiance to the logical interpretation of probability by way of their allegiance to corroboration, then this strategy is at the very least misguided. It is misguided because corroboration pays no special allegiance to one probability interpretation.

In terms of severity of test, we see a clear tension between CP and ML that cannot entirely be resolved via corroboration. CP, I have argued, seems to test its hypotheses more severely, whereas ML tests more hypotheses. That is, CP assumes less about evolution than ML, which makes the evidence (character data) less probable, which makes CP’s severity of test values higher than ML for genealogical hypotheses, even if we assume that ML and CP are ordinally equivalent. This point requires rejecting de Quieroz & Poe’s (2001) misguided argument that P(e,b) can be ignored under “standard phylogenetic analysis.” However, this claim by advocates of CP about minimum background knowledge is continually ignored by advocates of ML, who believe that within the CP method lies an implicit model of evolution which can be tested by ML’s own methodology, provided that this model is made explicit. It is true that ML can test various models against each other by keeping a specific topology constant, but an advocate of CP will retort that these models are all unrealistic (although some advocates do accept Tuffley & Steel's (1997) "No Common Mechanism" model).

In terms of explanatory power, we also see a clear tension between CP and ML that cannot entirely be resolved via corroboration. The situation here can be set in terms of the problem of homoplasy—the problem of how to explain homoplasy in phylogenetics. CP minimizes homoplasy in its reconstructions of evolutionary history. From this methodological minimization, ML advocates infer, based partly upon elaborate statistical arguments, that CP makes an extra, implicit assumption that homoplasy is rare. This assumption is unjustified, especially as phylogeneticists uncover more and more cases of homoplasy as they explore gene sequencing and molecular data. Advocates of CP that follow Farris (1983) defend themselves from this charge, by contending that homoplasy is minimized only because it is an *ad hoc* auxiliary hypothesis. Although this minimization does not directly tie into the theory of corroboration, it does tie into Popper's methodological convention to reject *ad hoc* hypotheses.

Also, we have a similar claim advanced by the CP camp to the one we have considered with respect to severity of test. That is, since CP makes less assumptions, and P(e,b) is lower than it would be given any model of the evolutionary process that ML could consider, the hypotheses chosen as optimal by CP have a higher explanatory power. The ML camp may lose this explanatory power, but they seem to be interested in explaining homoplasy as a function of branch length. Perhaps this could be reconceptualised as ML having an independent goal: estimating branch length. Roughly speaking, and similarly to severity of test, CP explains its hypotheses with more power, whereas ML explains more things, with less power. If each of two methods has some independent goals, and some similar goals, then we need to assess how these goals conflict, and whether or not one is relevant or detrimental to the other. It seems that advocates of CP are saying that estimating branch length is detrimental to reconstructing phylogeny, while advocates of ML are saying that it is relevant.

10.2 The Popper Debate and Corroboration

Given the tensions between the goals of CP and ML that I have argued for in Part Three, I will now determine what these aspects of the Popper debate can tell us about the general theory of corroboration. It should be said, although it is somewhat obvious, that it is not as simple as it would seem to apply an epistemological programme like corroboration to a science like phylogenetics. This is due, in part, to the fact that corroboration is designed to deal with competing hypotheses, whereas we have been considering a debate between different methodologies. Although CP and ML have some similar goals, they also have different goals and some of these stem from their different solutions to problem of homoplasy. In my opinion it is the goals, and not corroboration that need to be evalutated, for: “The capacity to assess Popperian corroboration neither justifies nor excludes any phylogenetic method” (Faith & Trueman 2001: 331). That is, corroboration, at least in terms of the two methods considered here, is compatible with either, and exclusive to neither.

In terms of simulations, like the “Farris Zone” or the “Felsenstein Zone”, it does not seem that either of these can count against a method from the context of corroboration. Corroboration cannot assess the value of a simulation. It can only assess empirical tests we have performed, and tests we may perform, but it is not yet clear how these simulations can be realized empirically. It seems that if the simulations were to pile up more on one side than the other, then we would have some (non-empirical) evidence that favours one method over another. In a recent review of long-branch attraction (the “Felsenstein Zone”), we see that there are more and more of what seem to be empirical cases that exhibit the features described in the simulation. Unfortunately this does not seem to be a clear win for ML, because as Bergsten (2005) shows, ML can also succumb to the long-branch attraction problem.

If there is a clear case to be made for a specific interpretation of probability being exclusive to a science, then, for the reasons discussed above, corroboration cannot aid in this decision. In phylogenetics this decision seems to have much more to do with the proper interpretation of metaphysical arguments. Corroboration cannot seem to clearly adjudicate between methods by way of explanatory power when different methods recognize different phenomena. At least those on the CP side seem to deny branch length and models of evolutionary rates as relevant to inferring phylogeny. Corroboration also cannot seem to adjudicate by way of severity of test between methods when different methods are interested in testing different things. At least those on the CP side seem to deny the relevancy of the testing of models that ML performs.

Despite all this, there are no strong reasons to reject the application of corroboration to phylogenetics. Corroboration may still be applied within a methodology. And an application of corroboration to a debate between methodologies will help explicate where the tension lies, where other considerations come into play, and where it is that we may concentrate our efforts to decide the debate. From my analysis, I have been able to show most clearly where the tension lies between CP and ML, and in some cases where other considerations come into play, specifically in their respective stances on probability interpretations, and has shown that if we make the "Felsenstein Zone" and "Farris Zone" simulations have empirical content, then we can potentially use this as leverage against either method. However a full decision between CP and ML would require a comparison against all the other competing methods within phylogenetics.

Bibliography

Bergsten, Johannes. (2005). A review of long-branch attraction. *Cladistics*, *21*: 163-193.

Braithwaite, R. B. (1953). *Scientific Explanation*. New York: Harper & Brothers.

Brower, Andrew V. Z. (2000). Evolution is not a necessary assumption of cladistics. *Cladistics*. *16*: 143-154.

Duhem, Pierre. (1954). *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Faith, Daniel P., and John W. H. Trueman. (2001). Towards an inclusive philosophy for phylogenetic inference. *Systematic Biology*, *50*, 331–350.

Farris, James S., Arnold G. Kluge, and Michael J. Eckardt (1970). A numerical approach to phylogenetic systematics. *Systematic Zoology*, *19*(2): 172-189.

Farris, James S. (1983). The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), *Advances in Cladistics*, Vol. 2. Columbia University Press, New York. pp. 7–36.

Farris, James S. (1986). On the boundaries of phylogenetic systematics. *Cladistics, 2*(1): 14-27

Farris, James S. (2000). Corroboration versus "strongest evidence". *Cladistics*, *16*: 385–393.

Farris, James S. (2008). Parsimony and explanatory Power. *Cladistics*, *24*: 825–847.

Felsenstein, Joseph. (1978). Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, *27*: 401-410.

Franz, Nico M. (2005). Outline of an explanatory account of cladistic practice. *Biology and Philosophy*, *20*: 489–515.

Goloboff, Pablo A. (2003). Parsimony, likelihood, and simplicity. *Cladistics*, *19*: 91–103.

Grant, Taran. (2002). Testing methods: the evaluation of discovery operations in evolutionary biology. *Cladistics*, *18*: 94-111.

Grant, Taran & Arnold G. Kluge. (2008). Clade support measures and their adequacy. *Cladistics*. *24*: 1051-1064.

Haber, Matthew. (2005). On probability and systematics: possibility, probability, and phylogenetic inference. *Systematic Biology*, *54*(5): 831-841.

Hacking, Ian. 1965. *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hennig, Willi. (1966). *Phylogenetic Systematics*. (translated by D. Dwight Davis and Rainer Zangerl). Chicago: University of Illinois Press, Urbana.

Howson, Colin & Peter Urbach. (1993). *Scientific Reasoning*: *The Bayesian Approach*, 2nd ed. Peru, Illinois: Open Court.

Kearney, Maureen. (2008). Chapter II: philosophy and phylogenetics: historical and current connections. In the *Cambridge Companion to Biology*. (eds. Hull and Ruse). Cambridge: Cambridge University Press: 211-232.

Kluge, Arnold. G. (1983). Cladistics and the classification of the great apes. In *New Interpretations of Ape and Human Ancestry*. (ed. R. L. Ciochon, and R. S. Corruccini, Eds). New York: Plenum Publishing Corporation. pp. 151–177.

Kluge, Arnold. G. (1988). Parsimony in vicariance biogeography: A quantitative method and a Greater Antillean example. *Systematic Zoology, 37*: 315–328.

Kluge, Arnold G. (1997a). Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics, 13*: 81–96.

Kluge, Arnold. G. (1997b). Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic systematics. *Zoological Scripta*, *26*: 349-360.

Kluge, Arnold. G. (2001a). Parsimony with and without scientific justification. *Cladistics*, 17: 199-210.

Kluge, Arnold G. (2001b). Philosophical conjectures and their refutation. *Systematic Biology*, *50*: 322–330.

Kripke, Saul. (1980). *Naming and Necessity*. Oxford: Basil Blackwell.

Laporte, Joseph. (2004). *Natural Kinds and Conceptual Change.* Cambridge: Cambridge University Press.

Lienau, Kurt E. and Rob DeSalle. (2009). Evidence, content and corroboration and the tree of life. *Acta Biotheoretica*, 57: 187–199.

Platnick, N. I. (1979). Philosophy and the transformation of cladistics. *Systematic Zoology*. 28. 4: 537–546.

Popper, Karl R. (1958). A third note on degree of corroboration or confirmation. *The British Journal for the Philosophy of Science*, *8*, 32: 294-302.

Popper, Karl R. (1959). *Logic of Scientific Discovery.* New York: Basic Books.

Popper, Karl R. (1965). *Conjectures and Refutations*. London: Routledge.

Popper, Karl R. (1983). *Realism and the Aim of Science*. (ed. W. W. Bartley, III). New York: Routledge.

Popper, Karl R. (2009). *The Two Fundamental Problems of the Theory of Knowledge*. New York: Routledge.

Putnam, H. (1973). Meaning and reference. *Journal of Philosophy*, 70: 699–711.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7: 215–271.

de Queiroz, Kevin, and Steven Poe. (2001). Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Systematic Biology*, *50*: 305–321.

de Queiroz, Kevin, and Steven Poe. (2003). Failed refutations: further comments on parsimony and likelihood methods and their relationship to Popper's degree of corroboration. *Systematic Biology, 52*: 352–367.

Quine, W. V. O. (1953). Two dogmas of empiricism. In *From a Logical Point of View*. Cambridge, MA: Harvard University Press. pp. 20-46

Rowbottom, Darrell P. (2010). Corroboration and auxiliary hypotheses: Duhem's thesis revisited. *Synthese*, 177: 139–149.

Saether, Ole A. (1986). The myth of objectivity—post-Hennigian derivations. *Cladistics*, *2*: 1-13..

Siddall, Mark and Arnold G. Kluge. (1997). Probabilism and phylogenetic inference. *Cladistics*, *13*: 313–336.

Sober, Elliott. (1983). Parsimony in systematics: philosophical issues. *Annu. Rev. Ecol. Systematics*, *14*: 335–357.

Sober, Elliott. (2004). The contest between parsimony and likelihood. *Systematic Biol*ogy, *53*(4): 644–653.

Tuffley, C., Steel, M., (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol., 59*: 581–607.

Vogt, Lars. (2007). A falsificationist perspective on the usage of process frequencies in phylogenetics. *Zoologica Scripta*, 36: 395–407.

Vogt, Lars. (2008). The unfalsifiability of cladograms and its consequences. *Cladistics*, *24*: 62–73.

Whewell, William. (1847). *The Philosophy of the Inductive Sciences Founded Upon Their History*. London: John W. Parker, West Strand.

Wiley, E. O. (1975). Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. *Systematic Zoology*, *24*. 2: 233-243.

Wiley, E. O. (1981). *Phylogenetics: the theory and practice of phylogenetic systematics.* Toronto: John Wiley & Sons.