Chapter 8
An Introduction to Phylogenetic Path Analysis

Alejandro Gonzalez-Voyer and Achaz von Hardenberg

The questions addressed by macroevolutionary biologists are often impervious to experimental approaches, and alternative methods have to be adopted. The phylogenetic comparative approach is a very powerful one since it combines a large number of species and thus spans long periods of evolutionary change. However, there are limits to the inferences that can be drawn from the results, in part due to the limitations of the most commonly employed analytical methods. In this chapter, we show how confirmatory path analysis can be undertaken explicitly controlling for non-independence due to shared ancestry. The phylogenetic path analysis method we present allows researchers to move beyond the estimation of direct effects and analyze the relative importance of alternative causal models including direct and indirect paths of influence among variables. We begin the chapter with a general introduction to path analysis and then present a step-by-step guide to phylogenetic path analysis using the *d*-separation method. We also show how the known statistical problems associated with non-independence of data points due to shared ancestry become compounded in path analysis. We finish with a discussion about the potential effects of collinearity and measurement error, and a look toward possible future developments.

Both authors contributed equally to this work.

A. Gonzalez-Voyer (✉)
Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Av. Américo Vespucio SN, 41092 Sevilla, Spain
e-mail: alejandro.gonzalez@ebd.csic.es

A. von Hardenberg
Alpine Wildlife Research Centre, Gran Paradiso National Park, Degioz 11, 11010 Valsavarenche, Aosta, Italy
e-mail: achaz.hardenberg@pngp.it

L. Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, DOI: 10.1007/978-3-662-43550-2_8,

8.1 Phylogenetic Linear Models: Drawbacks and Limitations When Analyzing the Influence of Multiple Variables

Because comparative biologists address questions related to long-term processes, they are faced with an important practical obstacle: the time necessary to produce evolutionary change. Hence, as with many problems in ecology, evolution, and behavior, the questions addressed by comparative biologists are often impervious to experimental approaches and alternative methods have to be adopted. Phylogenetic comparative methods employ the results from replicated "natural experiments" across multiple extant taxa as the data with which to test evolutionary hypotheses. Repeated associations between putative functional traits and environmental variables (proxies for a selective regime) or among traits are taken as supporting the evolutionary hypothesis. However, although the approach is potentially a very powerful one, given that comparisons generally involve numerous species and hence span long periods of evolutionary change (Freckleton 2009), comparative biologists are constrained in the inferences or conclusions they can draw from their results. Correlations between traits or between traits and the environment in extant taxa do not address the question of evolutionary origin (Martins 2000). Indeed, an important limitation when dealing with processes having occurred in the distant past is that there is no information about the conditions during most of the evolutionary history of the process being analyzed. Therefore, although there is a relationship between traits in extant taxa and current environments, this does not necessarily mean that there was a relationship between traits and the environmental conditions when the adaptation arose (Martins 2000). Furthermore, correlations between traits and the environment or between traits do not necessarily imply that the environment or trait is the driving force for the observed phenotypic changes (Martins 2000). Indeed, all correlative data have the inherent limitation that there is no way to determine causality. Nonetheless, a comparative method does exist allowing researchers to determine the order of evolutionary transitions (contingency) in correlated discrete traits (Pagel and Meade 2006).

These caveats not withstanding, there are also limitations regarding inferences researchers can make about their results due to limitations of the most commonly employed statistical methods. Currently, when testing hypotheses about associations between traits or traits and the environment, the method most often employed by comparative biologists is based on linear models. Phylogenetic independent contrasts or phylogenetic generalized least squares (PGLS) methods allow to analyze covariation between traits or traits and the environment, controlling for non-independence of data points (correlated residuals) due to shared ancestry (Felsenstein 1985; Grafen 1989; Martins and Hansen 1997). In addition, PGLS allows to combine continuous and discrete traits in a single model without the need to code dummy variables as well as allowing for different models of trait evolution to be incorporated in the analyses (Martins and Hansen 1997; see Chaps. 5 and 6).

However, both methods present similar limitations, which are the same as those of traditional linear models. First, only a single-dependent variable can be analyzed at a time, although a more realistic reflection of the complexity of the multivariate relationships currently analyzed by comparative biologists would allow for simultaneous exploration of the effects of a number of predictor variables on a number of different outcomes. Second, in multivariate linear models, a particular variable can either be a predictor or a response; however, a particular phenotypic trait can be responding, for example, to the influence of the environment and in turn be itself the cause of changes in a second phenotypic trait, hence a single trait can be both a response and a predictor. In order to overcome these limitations of traditional multivariate linear models, path analysis was developed. Confirmatory path analysis (and structural equation modeling) is an extension of multiple regression, but it is superior to ordinary regression analysis in that it allows researchers to move beyond the estimation of direct effects and analyze the relative importance of alternative causal models including direct and indirect paths of influence among variables. In von Hardenberg and Gonzalez-Voyer (2013), we introduced phylogenetic path analysis (PPA), integrating PGLS with the *d-separation* method for path analysis developed by Shipley (2000a). The proposed method allows researchers to harness the power of path analysis to disentangle cause–effect relationships among variables with data leading to correlated residuals due to shared ancestry. In this chapter and in the online practical material (hereafter OPM) available at www.mpcm-evolution.org, we provide further information and a detailed tutorial about how to perform PPA using the open source statistical language R (R Development Core Team 2013).

8.2 The Philosophy of Path Analysis

Correlation does not imply causation. Back in our undergraduate statistics classes, we were all taught this scientific mantra (Fig. 8.1). This statement is so deeply embedded in our modern scientific culture that it even deserved its own Wikipedia page.[1] Indeed, it is undeniable that if *A* is related to *B*, this does not imply that *B* is caused by *A*, or that *A* is caused by *B*. Both variables may, for example, be caused by a third confounding variable C. Some simple examples will elucidate this point: A highly significant correlation exists between the number of breeding pairs of storks (*Ciconia ciconia*) and human birthrates in Europe (Matthews 2000). Does this imply that storks deliver babies? Another study suggests that scientific productivity (measured as the number of citations) of ecologists is inversely correlated with per capita beer consumption (Grim 2008). Does this mean that beer drinking is detrimental to your scientific career? If you are not willing to give up your passion for beer, you may nonetheless be able to compensate eating lots of

[1] http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation Retrieved June 4, 2014

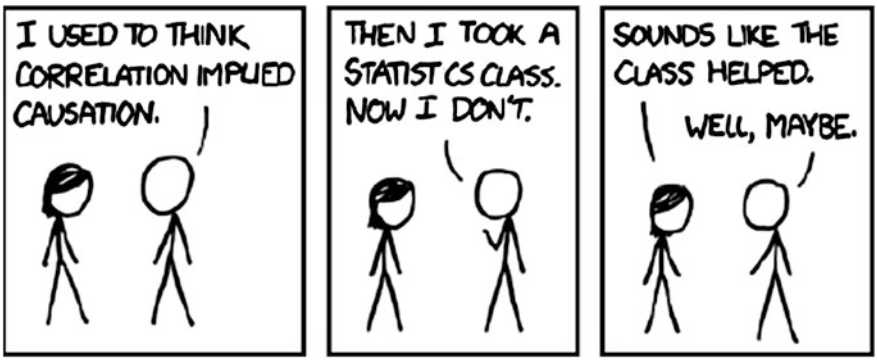

Fig. 8.1 Courtesy of XKCD (Distributed under a creative commons attribution-noncommercial 2.5 license)

chocolate. This at least is what a recent study in the New England Journal of Medicine suggests (Messerli 2012). The study shows a significant correlation between per capita chocolate consumption and the number of Nobel laureates per 10 million population in each country. All of these causal claims can easily be dismissed, taking into account other possible common causal variables, if not simply by logic. The main problem with the above-mentioned studies is that they are based on observational data rather than on controlled or randomized experiments (Fisher 1926), which are the commonly accepted scientific methods to infer causality. It would be great to be always able to use proper randomized or controlled experiments in all our studies, but this is obviously not possible, particularly in the case of comparative studies, where the unit of analysis is estimates of trait values for diverse species.

Correlation does not imply causation. This is what we have so dutifully learned. But is this completely true? Actually no. Indeed, without being afraid of saying a heresy, we can claim that *correlation always implies an underlying, unresolved causal structure* (Shipley 2000b). If we can rule out that the correlation between two variables is simply due to chance, there must be something that causes this relationship directly or indirectly through some other variables, even if we cannot necessarily identify the causes. The causal structure behind this correlation is indeed said to be *unresolved* because we cannot know, from the single correlation we can observe, how this correlation structure is built. Let us take a closer look at the "baby-delivering storks" data of Matthews (2000). The original data are available in the OPM available at http://www.mpcm-evolution.org as a "comma-separated values" (CSV) file. A tutorial showing how this data was analyzed and plotted using the open source statistical language R (R Development Core Team 2013), is also available in the aforementioned Web site. A quick glance at Fig. 8.2 strongly suggests that there is a relationship between number of storks and human births. Indeed, there is a significant relationship between the number of stork pairs and human birthrate with a p value of 0.008.

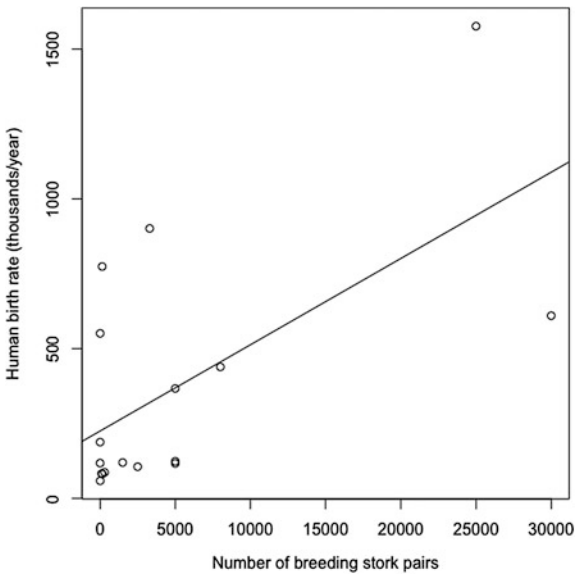

Fig. 8.2 Relationship between the number of breeding pairs of storks (*Ciconia ciconia*) and human birthrates in European countries (data from Matthews 2000)

Of course, none of us seriously believes that this means that storks really deliver babies[2]! Everybody will most likely agree that there is some other confounding variable which is the real, direct, or indirect cause of both the number of breeding stork pairs and human birthrates. As we said: *correlation always implies an underlying, unresolved causal structure*. Theoretically, we could test whether storks actually deliver babies doing a randomized experiment, for example, keeping constant the number of breeding stork pairs over a random selection of countries[3] in order to physically control for the variability in the number of stork pairs and at the same time excluding the effect of other possible confounding variables thanks to the randomization. However, this would undeniably be a very large scale and impractical experiment, not considering the moral implications it would have! What we can however do, if we have other factors which we suspect to be the true cause behind both human birthrates and the number of stork pairs, is to statistically control for the variability in these factors and thus see whether, controlling for the supposed common cause (in statistical jargon we would say: conditioning on it) the relationship between the number of storks and human birthrates still holds. We can try this using one of the other variables available in the dataset: area, which represents the surface size in squared kilometers of each country. It is indeed reasonable to think that larger countries host a higher number of stork pairs and at the same time have higher human birthrates, possibly indirectly through some other unmeasured variable. Indeed, there appears to be a very

[2] If you do, you can stop reading here!

[3] By hunting or, less drastically, translocating excess pairs from one country to an other.

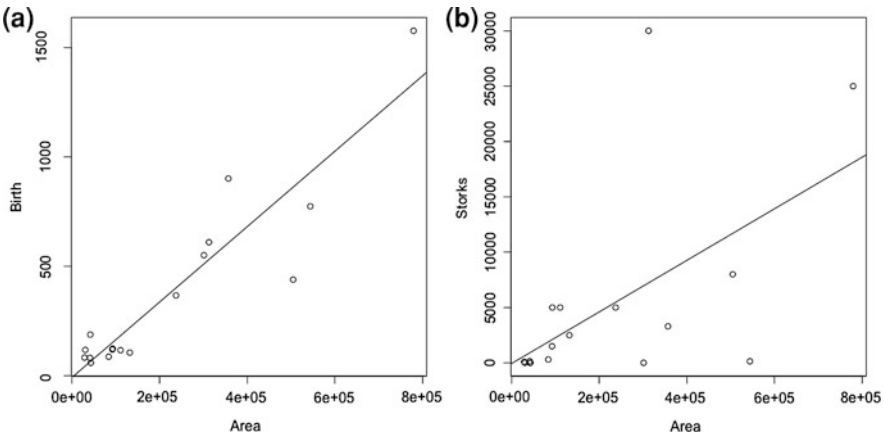

Fig. 8.3 **a** Relationship between human birthrate and country area; **b** relationship between the number of breeding pairs of storks and country area (data from Matthews 2000)

strong relationship between human birthrate and country area (Fig. 8.3a)! We can do the same for the relationship between the number of stork pairs and area. In this case also the relationship (Fig. 8.3b), even thought not as strong, is significant ($p = 0.0148$). We can now go back to our first linear model of the relationship between number of stork pairs and birthrates and statistically control for the effect of the confounding variable area, simply including this variable in the model transforming it into a multiple linear regression model of the kind[4]: Birth $\sim$ Area + Storks. Where Birth = birthrate, Area = country area, and Storks = number of stork pairs. With this model, while the effect of area on birthrate is highly significant ($p = 6.62e{-}06$), including this variable drastically changed the significance of the effect of the number of stork pairs to an unimpressive p value of 0.307 compared to the p value of 0.008 we obtained previously without conditioning on area! Technically, what we did is test the partial regression coefficient of the effect of the number of stork pairs on birthrate statistically controlling for the confounding effect of area and thus testing the statistical independence of the number of stork pairs from human birthrate. Is this enough to be able to claim that area is thus the common cause of both the number of breeding stork pairs and human birthrate? Sadly no. Indeed statistically, even if not logically, the result of this partial regression model may imply at least one alternative causal structure besides the above-mentioned hypothesis: The number of stork

[4] Note that here and in the rest of this chapter, we use the modified Wilkinson-Rogers notation for linear models (Wilkinson and Rogers 1973) widely used in statistical languages such as R. In this notation, the intercept is implicit and the tilde ($\sim$) separates the left-hand side from the right-hand side of the equation.

pairs may indirectly influence human birthrate through the area of countries.[5] While this alternative hypothesis does not necessarily make any logical sense in this case,[6] statistically, in the absence of further information, we cannot distinguish it from the hypothesis that area influences both the number of storks and birthrate, as the correlation pattern we observe among the variables (i.e., the partial regression model described above), can imply more than one underlying causal structure (as already mentioned that is why it is *unresolved*). However, while correlation implies an underlying, unresolved causal structure, *causation always implies a completely resolved correlational structure* (Shipley 2000a). This means that the hypothesized causal relationships among variables imply one, and only one, specific pattern of correlations and partial correlations (which in turn, however, can be cast by more than one causal model). Bill Shipley in his excellent book on cause and correlation in biology compares the pattern of correlations we can observe in nature to the shadows cast on a screen by a three-dimensional object, which in turn represents the causal structure behind the observable correlations (Shipley 2000b). A round shadow can be cast both by a ball as well as by a frisbee (i.e., the implied causal structure is unresolved), but the ball can cast only a round shadow (the implied correlational structure is completely resolved). This means that, at least in principle, we could test the "goodness of fit" of the correlational pattern we would expect to be cast by our hypothesized causal model, with the correlational structure we observe in the data. To be able to do this, we need a formal method to translate between the language of causality and the language of statistical probability. We also need an appropriate measure of the fit between the correlational pattern we observe in the data and the one that must exist given a specific causal structure. We will take a closer look at the recently developed methods permitting us to do exactly this, but first we need to define better the language of causality. To this end, let us complicate a bit our model of the causal relationships linking the various variables present in the Matthews (2000) dataset. For example, we can plausibly hypothesize that while area is the common cause of the number of stork pairs and human birthrate, this latter variable in turn is the causal parent of the number of inhabitants in each country[7] (Fig. 8.4). The causal model depicted in Fig. 8.4 is what in graph theory is called a directed acyclic graph (DAG). Squares represent variables, which in the language of graph theory are called "vertices." The directed arrows, called "edges," represent the hypothesized causal links. The graph is called "Acyclic" because in this kind of graph, a causal path (i.e., the path you can do following the edges passing from one vertex to the next along the causal model) never returns to the same starting vertex. A vertex in a DAG such as birthrate in Fig. 8.4 can be both a

[5] Implying that country size is somehow determined by the number of stork pairs inhabiting that country!

[6] Even though it is not necessarily more implausible than the hypothesis that storks deliver babies!

[7] In the data frame storks.dat this variable is called "Humans" and it is expressed as millions of inhabitants.

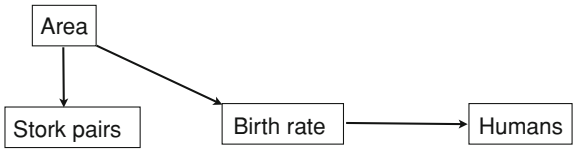

Fig. 8.4 Causal model of the relationship between the number of breeding pairs of storks, human birthrates in European countries, country area and population size depicted as a directed acyclic graph

dependent and an independent variable at the same time (in the language of graph theory you would say that it is both a causal parent and a causal child). We refer to Shipley (2000b) and Pearl (2009) for more details about the language of graph models. DAGs are the mathematical tool we use to formulate hypothesized models of causal relationships among variables. What we now need are formal methods to translate between the language of causality (which we represent with DAGs) and the language of statistical probability. These tools have been introduced to biologists only relatively recently and they go under the name of path analysis and structural equation models (of which classical path analysis is a special case). In the next sections, we will describe them in more detail, with a specific reference to past attempts in the literature to use these methods with data in which data points are represented by species with non-independent errors due to the underlying phylogeny. We will also introduce the *d*-separation-based technique for path analysis (Pearl 1988, 2009) and the *d*-sep test developed by Shipley (2000a), which are at the core of the method we recommend to use for phylogenetic path analysis (von Hardenberg and Gonzalez-Voyer 2013).

8.3 Structural Equations and *d*-Separation-Based Techniques

In structural equation models (SEM), the causal models are translated into a set of linear equations following the causal structure, and the parameters to be estimated from the data are specified. The expected pattern of covariance among the variables can thus be derived simply using the rules of covariance algebra. The free parameters are estimated by maximum likelihood minimizing the difference between the expected covariance matrix of the assumed model and the observed covariance in the data. Finally, we can calculate the probability that the minimum difference between the expected and observed covariance is different from zero (i.e., the observed covariance pattern deviates significantly from the covariance expected by the causal model). This method is appealing because it is based on maximum likelihood and it permits the inclusion of unmeasured latent variables. The latter is an important difference between SEM and path analysis based on *d*-separation, which cannot include latent variables. For a thorough review of SEM

methods, we point readers to Shipley (2000b) and Kline (2010). However, to make SEM methods amenable to work with an underlying phylogenetic signal not only must we compare the covariance matrix expected by the causal model with the covariance observed in the data, but also somehow include the expected covariance due to common ancestry. In Sect. 5, we review past attempts to develop phylogenetic SEMs (for examples, see Lesku et al. 2006; Santos 2009, 2012; Santos and Cannatella 2011). The method that we propose to use for phylogenetic path analysis (von Hardenberg and Gonzalez-Voyer 2013) follows a different approach and is based on the concept of *d*-separation developed by Judea Pearl and his collaborators (Geiger et al. 1990; Pearl 1988; Verma and Pearl 1988).

D-separation[8] is the 'missing link' between the language of causality, represented as directed acyclic graphs, and the language of statistical linear models. *D*-separation specifies the minimum set (called the basis set) of independence and conditional independence relationships (called *d*-separation statements) that hold true among all variables (the vertices) of the hypothesized causal model. In other words, it specifies the list of all, and only those, pairs of variables that are statistically independent conditioning on a set of other variables in the causal model. The minimum set of conditional independencies is determined in the following manner. First, list all pairs of non-adjacent vertices, i.e., the pairs of vertices that are *not* directly connected by an arrow (edge) in the directed acyclic graph. This gives a list of conditionally independent pairs of variables (these vertices are said to be *d*-separated). Second, list all the vertices with an arrow pointing directly to any of the conditionally independent variables in each pair, i.e., the causal parents of any of the two *d*-separated vertices. This gives the list of variables upon which the independent pairs of variables are conditioned, i.e., the variables that are statistically controlled to test the independence between the *d*-separated variables. Simply combining the two lists, we obtain the minimum set of conditional independence statements, which have to be true not to reject the hypothesized causal model. The conditional independence statements can be directly translated into statistical models using correlation, linear models, or other statistical tests that adequately fit the error structure of the data including nonparametric tests and permutation methods. The flexibility in the statistical methods that can be employed to test the conditional independencies is one of the important advantages of *d*-separation compared to SEM methods. To make the above clearer, we will go back to our "baby-delivering storks" example and the hypothesized causal model depicted in Fig. 8.4. In this simple example, the number of stork pairs is *d*-separated from birthrate (storks, birth), and from human population size (storks, humans). Furthermore, area is *d*-separated from human population size (area, humans). This gives us the following list: [(storks, birth), (storks, humans), and (area, humans)]. Let us now list the causal parents. For the first statement (storks, birth) we have only area, which is directly linked with both vertices. Following the notation proposed by Shipley (2004), we put the parent variables between curled

[8] D-separation is an acronym for "Directed" separation.

brackets, in this case: {Area}. For the second statement (storks, humans), we have two parent variables: area, directly causing storks, and birth, which is the causal parent of humans {Area, Birth}. For the last statement in this example (area, humans), we have only one causal parent which is birth directly causing humans {Birth}. The resulting list is [{Area}, {Area, Birth}, {Birth}]. As we mentioned above, combining these two lists we obtain the basis set of conditional independencies which must be true for the data to fit this model: [(Storks, Birth){Area}, (Storks, Humans){Area, Birth}, (Area, Humans){Birth}]. We can now translate these *d*-separation statements to statistical linear models in which we test the independence of the pairs of variables in round brackets conditioning on their parent vertices enclosed in curled brackets. The linear models we get are the following:

Birth $\sim$ Area + Storks

Humans $\sim$ Area + Birth + Storks

Humans $\sim$ Birth + Area

In the OPM (available at: http://www.mpcm-evolution.org), we show how you can test these models using *R*. You may have noticed that actually, we already tested the first of these linear models in Sect. 2 and found that, indeed, human birthrate is statistically independent from the number of stork pairs when conditioning on area with a *p value* of 0.307. The effect of storks on human population size is not significant when conditioning on area and birthrate ($p = 0.110$) as well as the effect of area on humans when conditioning on birthrate ($p = 0.6232$). The fact that none of the three hypotheses implied in the above independence statements is rejected, permits us to say that the hypothesized causal model depicted in Fig. 8.4 is a plausible explanation of the correlation patterns we observe among the variables.

Shipley (2000b) proposed to combine the *p* values using Fisher's C statistic which is calculated with the following formula:

$$C = -2\sum_{i=1}^{k}(\ln(p_i)) \tag{8.1}$$

where *k* is the number of conditional independencies in the minimum set and *p* their *p* value. The *C* statistic follows a χ^2 distribution with degrees of freedom (*df*)= 2 k. The *C* statistic therefore provides a convenient statistic for testing the goodness of fit of the whole path model. With this test (called the *d*-sep test), the path model is rejected, i.e., it does not provide a good fit to the data, if the p value of the *C* statistic is below the pre-specified alpha value (e.g., 0.05). We can now test the fit of our hypothesized causal model of the relationships among number of stork pairs, human birthrate and population size and country surface area. The C statistic has a value of 7.713, which, knowing that the number of conditional independencies *k* is 3, leads to a *p* value of the *d*-separation test of 0.26. This *p* value is larger than an alpha value of 0.05, and therefore, we can accept the model depicted in Fig. 8.4 as a plausible causal explanation of the

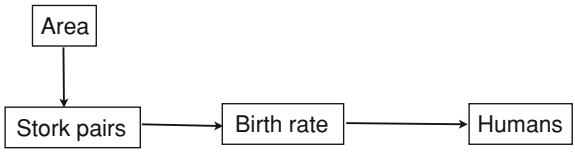

Fig. 8.5 Alternative causal model of the relationship between the number of breeding pairs of storks, human birthrates in European countries, country area, and population size depicted as a directed acyclic graph

relationships found among the variables in our dataset. If you are not convinced, and still believe that storks deliver babies, you can try an alternative model in which instead of having a direct causal link from area to birthrate, you have a direct causal link from the number of stork pairs to birthrate, while the other relationships stay the same as in the previous model. This alternative causal model is depicted as a DAG in Fig. 8.5. We leave it as a little exercise for the readers to obtain the minimum set and thus apply the *d*-sep test to the derived conditional independencies.[9] If you carefully followed all the steps, you should get a C value of 29.2 and a corresponding p value of 5.570647×10^{-5}, which is way below the alpha value of 0.05. This model is therefore rejected, and this should, we think, put the final word on the dispute of whether storks actually deliver babies! In the next section, we show how to apply this elegant and powerful method to data with an underlying phylogenetic signal, introducing in this way our proposed method for phylogenetic path analysis (von Hardenberg and Gonzalez-Voyer 2013).

8.4 A Step-by-Step Guide to Phylogenetic Path Analysis Using the *d*-Separation Method

The first step for any phylogenetic path analysis, as for any study in evolutionary biology, is to clearly define the hypothesis (or hypotheses) being tested. Although this may seem rather trivial to most readers, if not enough time is dedicated to clearly define the hypotheses to be tested, their predictions and underlying assumptions, the study can rapidly go astray and valuable time go to waste. A clear description of the hypotheses to be tested will be crucial for the next step: data collection. Although in the past, the limiting factor for comparative analyses was the lack of well-defined and robust phylogenies, at present the limitations are generally due to insufficient data. A well-defined hypothesis is important to guide researchers as to the data required to test it. We should stress the importance of careful data collection with particular attention to the importance of repeatability,

[9] All conditional independencies and full results for this model are provided in the online practical material (http://www.mpcm-evolution.org).

data that are representative of the species, and at the same time are also comparable across species (see Chap. 7).

The second step is to use graph theory to depict the hypotheses being tested as directed acyclic graphs. As mentioned previously, although path analysis is an extension of linear regression, it relies on path diagrams to depict the causal relationships between the variables. Because path analysis is a model-testing procedure, and not a model-developing one, all models to be tested should be based on theory and previous evidence. Once the hypotheses to be compared have been properly depicted by the directed acyclic graphs, the third step is to test the fit of each path model to the data.

As seen above, to test the fit of a path model to the data, we must first enumerate the minimum set of conditional independencies that must be true for the model to adequately fit the data. These conditional independencies can then be translated into linear models and tested with conventional statistical tests. Shipley (2009) showed that the *d*-separation method for path analysis can be extended to data with a hierarchical structure using generalized mixed models to test the conditional independencies in the minimum set. In von Hardenberg and Gonzalez-Voyer (2013), we extended the method further to include the particular case of interspecific comparisons, in which the lack of independence of data points and resulting correlation structure in the residuals violates assumptions of traditional statistical methods. We showed how the conditional independencies can be simply translated into phylogenetic generalized least squares models. Because the conditional independencies are being tested using linear regression models (PGLS) rather than correlations, the order in which we put the variables in the model is important and thus care must be taken when determining which vertex is the "*dependent variable*" and which vertex is the "*independent variable.*" Vertices that are causal children (at the end of the causal path separating the two vertices of interest) are *dependent variables*, while causal parents (at the beginning of the causal path) are the *independent variables*. To calculate the number of conditional independencies in the minimum set, the following handy formula can be employed (Shipley 2000b):

$$\frac{V!}{2 \times (V-2)!} - A \tag{8.2}$$

where V is the number of vertices in the directed acyclic graph and A is the number of edges (the arrows in our DAG). The test, for each conditional independency, involves determining whether indeed, vertices are uncorrelated when conditioning on the parents of each of them. A slightly special case, for defining conditional independencies, is when two vertices are separated by a collider vertex. A vertex is called a collider vertex when two edges from opposite directions in the causal path point toward it (e.g., $A \rightarrow B \leftarrow C$). A collider vertex is said to switch the causal path from *active* to *inactive*, that is, vertices in one side of it are unaffected by changes in vertices in the other side. Hence, when testing the conditional independency of vertices on either side of a collider, the collider is not included in the

conditioning set. For example, to test the conditional independency of (A, C) the conditioning set is $\{\phi\}$. The symbol ϕ is used to indicate that A, C are conditioned on no other variable. The last step is to combine the p values of the conditional independence tests using Fisher's C statistic and thus test the fit of the hypothesized causal model as shown in the previous section (Sect. 8.3).

To illustrate more clearly and in a greater extent the process of phylogenetic path analysis, we now present an empirical example, which we invite the readers to follow replicating it on their own computers using the R language. To make the example biologically meaningful and more intuitive, we propose the following evolutionary puzzle. The aim of the study is to identify the factors that determine geographical range size in the Rhinogrades. If you are unaware of this particular mammal order (Rhinogradentia, DE BURLAS Y TONTERIAS 1948), this comes as no surprise. The Rhinogrades (also called snouters) were endemic to the islands of the Hi-yi-yi Archipelago in the Pacific Ocean only discovered in 1941 but erroneously completely destroyed by secret nuclear experiments in the 1950s, causing the complete extinction of this highly diversified taxonomic group. The main characteristic of the Rhinogrades is that their noses evolved and diverged (in an analogous way to the beak in Darwin finches) into variegated forms with the most diverse functions. In particular, in most genera of the Rhinogradentia, the nose evolved into a complex locomotion organ (Fig. 8.6). For a full account of the natural history of the Rhinogrades, we refer the readers to Stümpke (1967). Previous studies suggest that, as in other mammalian species, there is an allometric relationship between range size and body mass. Some Rhinogradentia specialists suggest that species with larger range sizes also have larger litter sizes, because of higher resource availability. Nonetheless, given the allometric relationship between body size and litter size, it is still unclear whether the association between litter size and range size is causal or merely correlational. There is much discussion regarding the relationship between range size and nose length. On the one hand, range size has been proposed to directly affect nose length, given that Rhinogradentia use their nasal appendage for locomotion and hence larger range sizes select for longer-distance displacements. Alternatively, some experts suggest that the direction of causality should be turned upside down and that it is nose length that determines displacement distances, and therefore, species with longer noses are able to expand their range size. Finally, there is some consensus among Rhinograd experts that dispersal distance is determined by nose length. Based on this knowledge, we can construct a set of hypotheses of causal relationships among variables which we can depict using directed acyclic graphs with the five traits of interest. We will refer to the five traits of this example with acronyms for brevity: body mass = BM, litter size = LS, nose length = NL, dispersal distance = DD, and range size = RS. Figure 8.7 presents the models we proposed for the present example.

The nuclear disaster which destroyed the Hi-yi-yi islands, together with the Rhinogrades also brought down the Darwin Institute of Hi-yi-yi, where all the specimens and life history data of this group were conserved (Stümpke 1967). We therefore had no other choice but to resort to simulated data for our example.

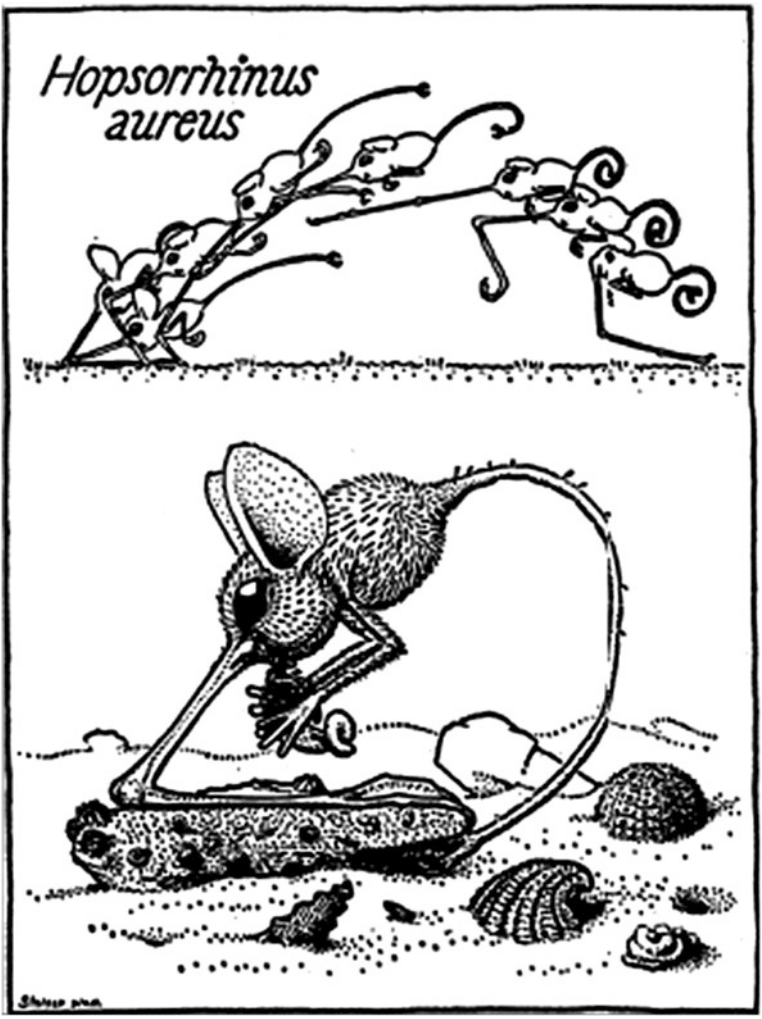

Fig. 8.6 A representative of the Rhinogradentia order: the *Hopsorrinhus aureus* belonging to the Hopsorrhinidae family (Snout Leapers sens. strict.), characterized by the peculiar nasal structure which permits them to move thanks to long backward leaps (Taken from Plate VI, in Stümpke 1967)

We simulated a phylogenetic tree of 100 species under a pure-birth model. Data for the five variables were simulated to evolve on the tree following a Brownian model with a lambda-transformed tree ($\lambda = 0.8$). The data were simulated to evolve with varying degrees of inter-correlation among the variables based on a pre-specified causal model. Variables directly linked in the path model presented correlations of 0.5, while variables with indirect links presented correlations that decreased proportionally with the number of variables separating them, with correlation decreasing by half for each variable in the indirect link (see OPM).

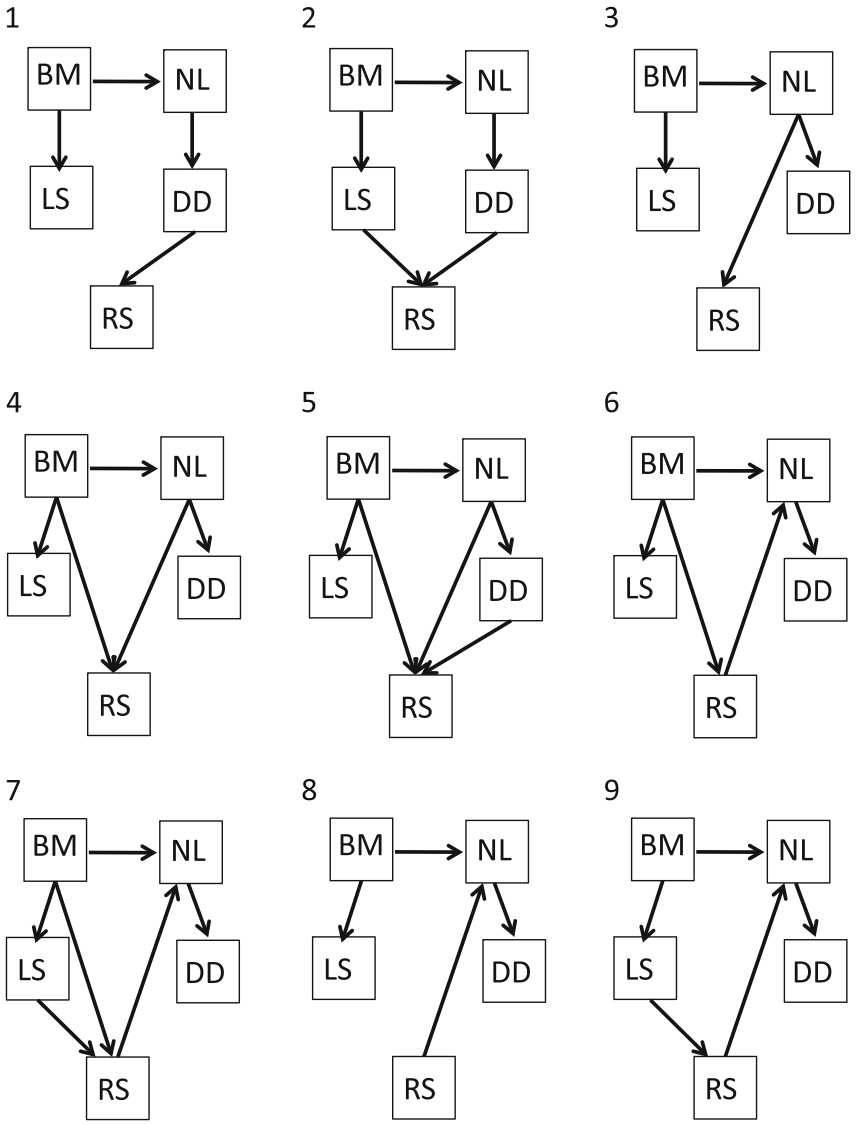

Fig. 8.7 Alternative path models depicting the relationship between body mass, litter size, nose length, dispersal distance, and range size in Rhinogradentia

Use of simulated data following a pre-specified path model is an excellent means to practice a novel approach. In the OPM (http://www.mpcm-evolution.org), we provide both the simulated data (in a file called rhino.csv) and phylogeny (in a file called rhino.tree) used in the Rhinogradentia example, as well as the R code used to simulate the phylogenetic tree and data to enable readers to simulate their own data under a different path model if they wish to do so. We also provide an online tutorial,

replicating all steps described in this section using R. In the first model in Fig. 8.7, there are 5 vertices and 4 edges, hence the minimum set contains 6 conditional independencies to be tested (as follows using the above-mentioned formula).

These are the conditional independencies in the basis set and their translation into linear models:

Conditional independencies	Linear models
(BM, DD) {NL}	DD ~ NL + BM
(BM, RS) {DD}	RS ~ DD + BM
(NL, LS) {BM}	LS ~ BM + NL
(DD, LS) {BM, NL}	LS ~ BM + NL + DD
(LS, RS) {BM, DD}	RS ~ BM + DD + LS
(NL, RS) {BM, DD}	RS ~ BM + DD + NL

It is important to note that since we are using linear models to test the conditional independencies (as opposed to correlations), care must be taken when determining which variable is the response and which the predictor. In such cases, to determine the order of variables in the conditional independency, always follow the direction of causality in the directed acyclic graph (as noted above). Note that in particular cases there is no a priori reason to define one variable as the "response" and the other as the "predictor" as each variable is at the end of a causal path. In such circumstances, the researcher must decide how to define the model to test the conditional independency and keep it constant in other models being compared. We can test the conditional independencies using one of the available statistical packages to perform PGLS and thus obtain the value of the C statistic as described above. Note that an important advantage of the approach we propose is that it allows for analyses to be done using the evolutionary model which best fits the data (Freckleton 2009; Freckleton et al. 2002; Grafen 1989; Hansen 1997; Pagel 1999). In the case of this particular example, given we simulated the data under a Brownian model we will use PGLS analyses with a maximum-likelihood estimate of the lambda parameter (Freckleton et al. 2002; Revell 2010). Following the same steps as for the first path model, we can also test the minimum set of conditional independencies for model 2. These are presented below with their translation into linear models:

Conditional independencies	Linear models
(BM, DD) {NL}	DD ~ NL + BM
(BM, RS) {DD, LS}	RS ~ DD + LS + BM
(NL, RS) {DD, LS, BM}	RS ~ DD + LS + BM + NL
(NL, LS) {BM}	LS ~ BM + NL
(DD, LS) {BM, NL}	LS ~ BM + NL + DD

Table 8.1 C statistic, number of conditional independencies tested (k), and p values of the C statistic for the 9 path models depicted in Fig. 8.7

Model	C statistic	k	p value
1	63.809	6	4.52×10^{-9}
2	62.769	5	1.08×10^{-9}
3	28.973	6	0.004
4	6.582	5	0.764
5	5.258	4	0.730
6	6.439	5	0.777
7	6.018	4	0.645
8	7.699	6	0.808
9	7.362	5	0.691

The reader can now test the conditional independencies of the remaining models depicted as DAGs in Fig. 8.7.[10] Table 8.1 presents the values of the C statistic for each model as well as the p value.

Based on the results in Table 8.1, we can already conclude that the first three models provide very poor fit to the data, because the p value of the C statistic is significant. Hence, we can reject the hypothesis that the correlation structure observed in the data is the result of these three proposed causal models. On the contrary, models 4–9 cannot be rejected, or in other words the correlation structure observed in the data could potentially result from any of these 6 models. As stated above, there is inevitably some uncertainty regarding the causal model that gives rise to the observed correlation structure in the data, and in this case we have identified 6 candidate causal models. This not very satisfactory! Shipley (2000b) proposed two competing models that can be compared based on the difference in the C statistics, which follows a χ^2 distribution with $\Delta\ df = df_{\text{model1}} - df_{\text{model2}}$. However, only nested models can be compared in this manner. Ideally, we would like to be able to compare among all models (including non-nested ones) and rank them based on some estimate of their goodness of fit (Burnham and Anderson 2002). In von Hardenberg and Gonzalez-Voyer (2013), we proposed to use an information theoretic approach alike to the classical Akaike Information Criterion (Akaike 1974) using a modified version of AIC, which we called the C statistic Information Criterion (CIC). This approach was first proposed, in the framework of non-phylogenetic path analysis, by Cardon et al. (2011). Use of an information theoretic approach requires that the measure of "goodness of fit" be based on maximum-likelihood estimates; hence to be able to apply such an approach to path analysis using d-separation, it is necessary to show that the C statistic, used to calculate this criterion, is equivalent to a maximum-likelihood estimate. Shipley (2013) recently provided such mathematical proof, validating the use of AIC (i.e.,

[10] All conditional independencies and full results for these models are provided in the online practical material.

CIC) to compare between non-nested models in the framework of *d*-separation path analysis. This should allay concerns of readers worried by the fact that the *C* statistic is calculated based on the *p* values of the conditional independency tests and we are now using it to estimate CICc, apparently combining frequentist and information theoretic approaches. To calculate CIC, we simply need to know the number of parameters estimated in the path model using the empirical data. In phylogenetic path analysis, we assume a multivariate normal distribution of errors and linear relationships between variables, because these are assumptions of the phylogenetic generalized least squares models used to test the conditional independencies (for use of CIC with models with different error distributions, see Shipley 2013). We employ here the formula to calculate CICc, the equivalent of CIC with a correction for small sample sizes. In any case, when the sample size is large relative to the number of parameters, CICc will converge on CIC. To calculate CICc:

$$\mathrm{CICc} = C + 2q \times \frac{n}{(n - 1 - q)} \tag{8.3}$$

where *C* is the *C* statistic for the particular model, *q* is the number of parameters estimated in the path model, and *n* is the sample size, in the case of phylogenetic path analysis the number of species. For a given path model, we are interested in the slopes of each of the causal links between the variables and the variances. For example, for model 1 in the Rhinogrades exercise, 9 parameters are estimated: the variance for body mass (BM), which is the only variable without any causal parent in the model, and the 4 slopes and the variances for the causal links. While in model 2, 10 parameters are estimated: the variance for body mass, 5 slopes for the causal links and 4 variances, because in this case range size is causally determined by both litter size and dispersal distance, and therefore, two slopes and one variance are estimated for these causal links (see Shipley 2013 for details). In cases in which the interest lies only in the slopes of the causal links between the variables, a quick way to obtain the number of parameters estimated in the model is simply to add the number of vertices and number of edges in the path model. Note that for models to be comparable using CICc, all models must have the same sample size (number of species), and therefore, the data set is reduced to the maximum number of species for which data for all variables is available. Furthermore, all compared models must also have the same number of vertices, although they can have different numbers of edges. Hence, to compare two models in which one variable has no causal link to any other (i.e., there is no edge between it and any other vertex in the model), the complete set of conditional independencies between this variable and all others in the model must be tested to calculate the *C* statistic. Indeed, such a model assumes that the "isolated" variable (unconnected to any other variable in the model) is conditionally independent from all the variables in the model, and this assumption must be tested (Cardon et al. 2011).

Now we can calculate CICc values for all the models, we are comparing in the Rhinogrades example. With the CICc values in hand, we can also rank the models

based on the difference in CICc (ΔCICc). ΔCICc is simply the CICc value of model *i* minus the value of the model with the lowest CICc ($CICc_{MIN}$). ΔCICc can be used in the very same way as it is normally done in standard model selection procedures using AIC (Cardon et al. 2011). Given that ΔCICc is measured in a continuous scale of *information,* the values are comparable among models. As a general rule of thumb, models with ΔCICc values < 2 are all considered to have substantial support (Burnham and Anderson 2002). The relative likelihood of a model *i* given the data *L* ($g_i|$data), provides information regarding the relative strength of evidence for a model compared to the others and can be computed, following Burnham et al. (2011):

$$\ell = L(g_i|\text{data}) = (\exp -(1/2)\Delta\text{CICc}_i) \tag{8.4}$$

Finally, CICc weights, the probability of each path model g_i, given the data and the set of models being compared, are also simple to compute as a measure of strength of evidence (Burnham et al. 2011):

$$w_i = \Pr\{\text{mod}(g_i)|\text{data}\} = \frac{l_i}{\sum_{j=1}^{R} l} \tag{8.5}$$

Use of CICc allows us to move from a hypothesis testing to a hypothesis comparison framework. Below we present the CICc values for all the tested models in the Rhinogrades example, including the number of estimated parameters in each model (q), ΔCICc, likelihoods and weights (Table 8.2).

Use of CICc allows for finer comparisons among models compared with what can be gained by simply looking at the C statistic and its associated p value. Table 8.2 presents a clear ranking of all models from the Rhinogrades example. Those with significant *C* statistics (models 1, 2, and 3) also show elevated CICc and ΔCICc values, indicating that they provide a very poor fit to the data. We can also however gain some insight about the six models with nonsignificant C statistics. Models 5, 7, and 9 provide a relatively poorer fit to the data than the other three models (4, 6, and 8) as the ΔCICc values are > 2. We cannot distinguish between models 4, 6, and 8, since they present very small differences in CICc, with all models showing ΔCICc values < 2. Note that care must be taken when comparing models with ΔCICc values < 2. As pointed out by Arnold (2010), also highlighted by Burnham and Anderson (2002: p. 131), for equivalent AIC[11] values, care must be taken when interpreting models based solely on ΔAIC (or ΔCICc) values. In some cases, models might not be truly "competitive" with top-ranking models, but appear to be based solely on low CICc values, because addition of an uninformative variable, or in the particular case of path analysis an uninformative causal link between variables, can

[11] CICc in the case of phylogenetic path analysis.

Table 8.2 Number of parameters estimated in each model (q) C statistic information criterion with correction for small sample sizes (CICc), ΔCICc, likelihoods (l_i), and CICc weights (ω_i) are shown for each model of the Rhinogradentia example

Model	q	CICc	ΔCICc	l_i	ω_i
8	9	27.700	0.000	1.000	0.349
6	10	28.911	1.211	0.546	0.190
4	10	29.054	1.354	0.508	0.177
9	10	29.834	2.134	0.344	0.120
5	11	30.258	2.558	0.278	0.097
7	11	31.018	3.318	0.190	0.066
3	9	48.973	21.273	2.402×10^{-05}	8.380×10^{-06}
1	9	83.809	56.109	6.548×10^{-13}	2.284×10^{-13}
2	10	85.240	57.540	3.201×10^{-13}	1.117×10^{-13}

lead to marginal changes in CICc values even though there is very little difference in the goodness of fit. Therefore, models with such uninformative causal links might present ΔCICc ≤ 2, generally interpreted as indicating "substantial level of empirical support" (Burnham and Anderson 2002: 170), although such an interpretation would be erroneous. Burnham and Anderson (2002) suggest that models having Δi [ΔCICc] within 0–2 values of the best model should be examined to check whether they differ from the best model by having 1 more parameter and also present essentially the same maximized log-likelihood value (in this particular case, similar C statistic). In such cases, the model with more parameters is not really supported, but presents marginal difference with the "best model" simply because one parameter is added to the model, although the fit of the model is not truly improved as measured by the log-likelihood value (C statistic). Returning to our example, we can see that models 4 and 6 differ by a single parameter from model 8, the best-fitting model. Model 4 also differs from model 8 in the direction of the causal link between range size and nose length, which as the reader might remember was the cause of much discussion among Rhinograd experts. These models also present small differences in C statistic with model 8 (model 4: difference = 1.35, model 6: difference = 1.21). Hence, following Burnham and Anderson (2002) models 4 and 6 might not be considered as supported and competitive to the same degree as the best-fitting model 8, even though they are within ΔCICc < 2. Note that we are by no means advocating selection of a single model over all others. Rather, following Burnham and Anderson (2002) and Arnold (2010), we highlight the need for caution when comparing models, above all that it should not be done mechanistically simply based on ΔAIC (ΔCICc) values. In applications of phylogenetic path analysis with empirical data, it is highly likely that more than one model will present small (< 2) ΔCICc values. Under such circumstances, conclusions should be drawn based on the set of most likely models.

In our example, Model 8 appears to be the best-fitting model. We can now calculate standardized path coefficients of the causal edges linking the variables

according to this model. Standardized path coefficients are particularly useful because, being standardized, they are comparable with each other, and therefore, we can compare the relative strength of each causal relationship in the model. To calculate them, we must first standardize the original data. To do this, we subtract the trait specific population mean from each value and divide by the standard deviation. In the specific case of the simulated data used in this example (given it is randomly drawn from a multivariate normal distribution with mean 0 and standard deviation of 1), the data are already standardized, therefore this step is not necessary.

We then use the standardized data to calculate the standardized path coefficients using PGLS analyses, following the causal paths in the directed acyclic graph. In the case of model 8, the path coefficients are as follows:

BM→ LS 0.4973 (±0.0893 s.e.)
BM→ NL 0.4614 (±0.0650 s.e.)
RS→ NL 0.5281 (±0.0572 s.e.)
NL→ DD 0.6285 (±0.0800 s.e.)

Had we truly competitive models, one way to account for this "model uncertainty" is model averaging (Burnham and Anderson 2002). In von Hardenberg and Gonzalez-Voyer (2013), we showed how standard model averaging procedures can be applied also in the context of phylogenetic path analysis, averaging the path coefficients of all models with CICc < 2 according to the CICc weights of each model, thus on the relative strength of the models in the averaged set of models.

What have we learned regarding the relationship between range size, nose length, and other traits in Rhinogradentia after employing phylogenetic path analysis to tackle the question? First, based on the best-supported model (ΔCICc ≤ 2), range size appears to be the causal parent of nose length, while litter size does not appear to be causally linked to range size. Moreover, the effect of range size on dispersal distance appears to be indirectly mediated through nose length. In other words, in Rhinograds, dispersal distance appears to be directly determined by nose length. Finally, given this entire example was based on data simulated following a pre-specified path model we can now ask how precise is phylogenetic path analysis in identifying the path model giving rise to the data? Well, quite accurate in fact! The model we used to simulate the data is actually model 8, which is the best-supported model based on CICc. Furthermore, model 9 is identical to model 8 except for the additional causal link between litter size and range size. Despite virtually identical C statistics, there is a difference in CICc of 2.13, which suggests CICc is adequately penalizing this model for the additional parameter. Finally, looking at the standardized path coefficients calculated above, we see that they are all roughly around 0.5, which is not surprising, but reassuring, as the data have been simulated with correlation coefficients of 0.5 for all the pre-specified direct links.

8.5 Phylogenetic Non-Independence of Data Points, Correlated Residuals, and the Problems with Inflated Type I Error

The interest in the present chapter is to apply path analysis to macroevolutionary questions, involving comparisons among numerous species. Attempting to convince readers of this book of the importance of accounting for non-independence of data points due to phylogenetic relatedness of species is like preaching to the choir.[12] Nonetheless, we present first the challenges associated with accounting for phylogenetic relatedness in path analysis and second demonstrate the extent of the problem if non-independence of data points is ignored when undertaking confirmatory path analysis using the *d*-separation method (von Hardenberg and Gonzalez-Voyer 2013). It is well known that interspecific comparative analyses violate the assumption of traditional statistical methods that data points are independent, indeed the varying degrees of shared ancestry of the species included in the analysis influences the expected similarity of trait values (Felsenstein 1985; Freckleton et al. 2002; Garland et al. 1992; Harvey and Pagel 1991). For linear models, the main problem is the correlation structure of the residuals that is determined by the degree of phylogenetic relatedness among species (Felsenstein 1985; Grafen 1989; Martins and Hansen 1997; Revell 2010; see Chap. 5). The consequences of not accounting for phylogenetic effects in statistical analyses of multispecies data are, among others, artificially inflated number of degrees of freedom, incorrectly estimated variances, and increased type I error rates of significance tests (Felsenstein 1985; Harvey and Pagel 1991; Martins et al. 2002; Martins and Garland 1991; Rohlf 2006). These problems, however, become compounded in path analysis because of the requirement of testing multiple structural equations (in the case of SEM) or all the conditional probabilistic independencies that must be true for the causal model to be correct (in the case of the *d*-sep test). Previous attempts at controlling for phylogenetic relatedness in path analysis exist. Among those having included an explicit description of how phylogenetic non-independence was controlled are Lesku et al. (2006) and Santos and Cannatella (2011) who used phylogenetic independent contrasts (PIC; Felsenstein 1985) as the data entered in a SEM. Use of independent contrasts allowed the authors to account for phylogenetic non-independence explicitly in their SEM. However, there are limitations associated with the use PIC. First, the method assumes the traits, and covariances between traits evolve following a strict Brownian motion model and performance can be compromised if the assumption is not met (Revell 2010), second, PIC assumes strictly linear relationships between traits (Quader et al. 2004). More recently, Santos (2012) combined two approaches to control for phylogenetic non-independence in SEM in a study aimed at analyzing the factors associated to rate of molecular evolution in poison frogs. First,

[12] All pun intended!

for a set of species trait values, he estimated the phylogenetic signal of each trat by estimating the maximum-likelihood value of λ, he then calculated PIC from a λ-transformed phylogeny using the ML estimate for each particular trait. For data on rate of molecular evolution he used an estimate of the variance-covariance matrix derived from a molecular phylogeny.

We proposed an alternative approach (von Hardenberg and Gonzalez-Voyer 2013) combining confirmatory path analysis using the *d*-separation method (Pearl 1988; Shipley 2000b) and phylogenetic generalized least squares (PGLS; Martins and Hansen 1997). The advantage of PGLS is that it can incorporate distinct models of trait evolution, can combine continuous and categorical variables in a single model without the need to code dummy variables, and provides the value of the *y*-intercept (Martins and Hansen 1997; see Chap. 5). Further, a key advantage of using PGLS is that it allows for path analyses to be undertaken using taxon-specific trait values rather than contrasts, facilitating interpretation of the results. Finally, in PGLS an evolutionary parameter is estimated simultaneously with model fit, which determines the amount of phylogenetic signal in the data (in the residuals of the model to be precise) and hence the necessary correction for the expected covariance in trait values resulting from phylogenetic relatedness, given the evolutionary model (Freckleton et al. 2002; Martins and Hansen 1997; Revell 2010). This is an important advantage because in some instances data may present a phylogenetic structure that is intermediate between that predicted by the evolutionary model and absence of phylogenetic correlation in the data (Freckleton et al. 2002; Revell 2010). Under such circumstances, PGLS models have been shown to outperform independent contrasts (Martins and Hansen 1997). These advantages of PGLS allow us to ensure that tests of conditional independencies are done with the adequate correction for phylogenetic signal in the residuals of each particular model. Note that the flexibility of the *d*-separation method also allows researchers to combine continuous, categorical, and discrete variables in their path models, because tests of conditional independencies can be done using phylogenetic ANOVA, or other appropriate statistical methods (see Chap. 12 for an introduction to phylogenetic-mixed models).

In von Hardenberg and Gonzalez-Voyer (2013), we used a simulation-based approach to explore the consequences of ignoring phylogenetic non-independence when undertaking confirmatory path analysis using the *d*-separation method. We simulated evolution of five hypothetical traits along a simulated phylogeny under the covariance matrix expected from the causal relationships among the traits derived from a specific pre-defined causal model. In order to analyze the effects of varying degrees of phylogenetic signal in the data, the simulations were run under six different scenarios with different degrees of lambda (λ), spanning from null to strong phylogenetic signal in the simulated data. When $\lambda = 0$ traits were simulated evolving along a star phylogeny, where trait evolution for each species is completely independent, while at the other extreme of $\lambda = 1$ traits were simulated to evolve following a pure Brownian motion model, where the degree of similarity between species traits is inversely proportional to the distance to the nearest common ancestor. For the four remaining scenarios, prior to simulating trait

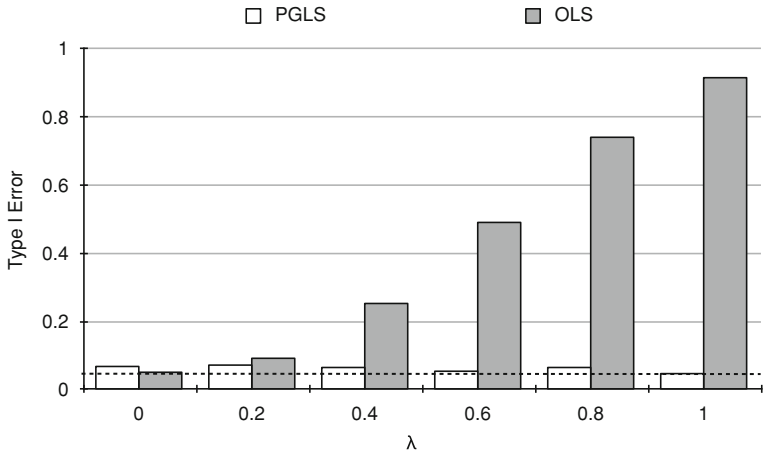

Fig. 8.8 Type I error of traditional (i.e., non-phylogenetic OLS) and phylogenetic (PGLS) path analysis under six simulated scenarios spanning low to high phylogenetic signal in the data

evolution, the phylogenetic tree was transformed based on values of λ ranging from 0.2 to 0.8 (i.e., 0.2, 0.4, 0.6, and 0.8). Tests of conditional independencies were done using the untransformed tree. One thousand datasets were simulated for each of the six scenarios, each with an underlying phylogenetic tree of a fixed, arbitrary size of 100 species. Each simulation of trait evolution was done using a different simulated phylogeny; hence simulations also incorporated the effects of varying phylogenetic topology. At each iteration, von Hardenberg and Gonzalez-Voyer (2013) calculated Fisher's C statistic and obtained a distribution of p values to determine the level of type I error (i.e., the probability of rejecting the null hypothesis, in this case the tested model, when it is true, testing the predicted set of conditional independencies consistent with the "true" underlying causal model) and the power (i.e., 1-the type II error, the probability of not rejecting the tested model when it is actually false, testing the predicted set of conditional independencies of a "wrong" causal model). These simulations were run both for d-sep tests ignoring phylogenetic effects and for the phylogenetically explicit d-sep test. The results of the first test, type I error, are shown in Fig. 8.8. It is clear that the type I error of "classical" path analysis, ignoring phylogenetic non-independence, increases rapidly with the degree of phylogenetic signal in the simulated data to reach values > 0.9 when traits are simulated to evolve via Brownian motion. On the contrary, although our phylogenetic path analysis method is slightly over-conservative, it nonetheless performs well under varying degrees of phylogenetic signal in the data. Figure 8.8 clearly demonstrates the importance (to say the least) of accounting for phylogenetic relatedness when undertaking path analysis using the d-separation method. However, power was in general comparable between "classical" path analysis, ignoring phylogeny, and phylogenetic path analysis (see Fig. 8.9). The high power of non-phylogenetic path analysis is not surprising. The

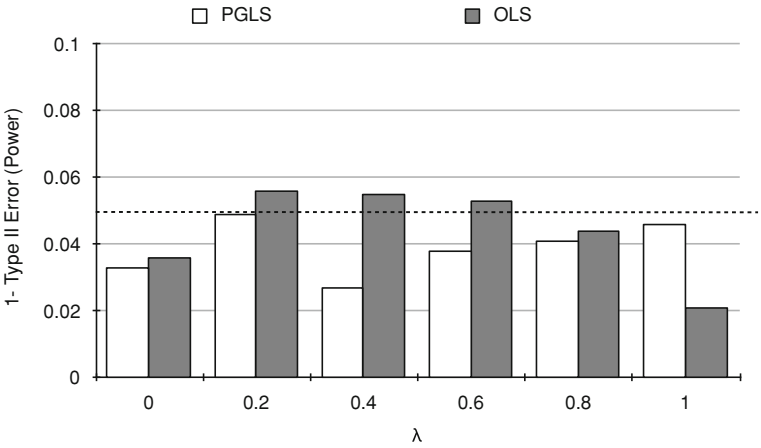

Fig. 8.9 Power of traditional (i.e., non-phylogenetic OLS) and phylogenetic (PGLS) path analysis under six simulated scenarios spanning low- to high-phylogenetic signal in the data

sagacious reader will have already guessed that the high power of non-phylogenetic path analysis is a consequence of the high type I error. Indeed when ignoring phylogenetic relationships, there is a higher probability of detecting significant correlations among traits, even if these are simply due to phylogenetic relatedness rather than true correlated evolution, with the result of a higher probability of rejecting the proposed model.

8.6 Does Collinearity Affect Path Analysis?

Literature on the effect of collinearity on Path Analysis is controversial. While some studies suggest that structural equation models (SEM) can effectively eliminate problems with collinearity (Pugesek and Grace 1998; Pugesek and Tomer 1995), others suggest it can be cause for concern (Petraitis et al. 1996; Grewal et al. 2004). As far as we know, no study has specifically dealt with the effect of multicollinearity on the *d*-separation method. Because the phylogenetic path analysis method we presented (von Hardenberg and Gonzalez-Voyer 2013) is based on the use of PGLS to test conditional independencies, violations of the assumptions of PGLS will inevitably undermine such tests. Least squares estimates of statistical model parameters are robust to moderate, even high, levels of collinearity (Freckleton 2011). However, estimates of parameter variance may be very sensitive affecting hypothesis tests, which would undermine confidence on tests of conditional independencies. Hence, strong collinearity can indeed be a problem, as long it is a problem for PGLS although it will be limited to the specific conditional independences we are testing. Our view is however, that the *d*-separation method can actually be an effective way to disentangle collinearity, at least

when it is not very strong. Indeed, the way you set up your path analysis model, and test for the independence among the variables to see if your model fits the data, you basically are testing for the presence of collinearity among your variables. Models with strong collinearity among the variables not directly causally linked will be rejected by the data and therefore will not be accepted as a possible explanation of the cause–effect relationships among the variables. On the other hand, collinearity between predictors could also affect the power of tests of conditional independencies because collinearity increases the standard error of partial regression coefficients. As collinearity increases, the ability to detect a significant effect (statistically non-zero partial regression slope) is reduced (Freckleton 2011). An often-unappreciated problem is the effect of measurement error, which is common for most (if not all) data employed in comparative analyses. Measurement error can result in underestimation of model parameters, even in the absence of collinearity, due to attenuation (Freckleton 2011). Bias increases when there is measurement error in combination with collinearity. Under such circumstances, attenuation leads to underestimation of the effect of the predictor with the weakest effect, while the predictor with stronger effect is over-estimated (Freckleton 2011). One possibility, which would need to be explored, is to include within-species variation in the models, for example, using mixed models (see Chaps. 7 and 10). By including several measurements per species for each trait, we could not only obtain a better estimate of the species mean but also obtain an estimate of the species-specific variation, which could potentially mitigate the effects of measurement error, although this has yet to be explored in the context of phylogenetic path analysis. We follow Freckleton (2009) and strongly suggest to always verify that the assumptions of the statistical methods employed to test the conditional independencies of the path model are met, this will ensure robust results of tests of conditional independencies.

8.7 Conclusions

The aim of this chapter was first to demonstrate in a didactic and easy to follow manner how to undertake a path analysis using the *d*-separation method (Shipley 2000b), while explicitly accounting for phylogenetic non-independence. As pointed out previously, the method we propose (von Hardenberg and Gonzalez-Voyer 2013) is not the only attempt (see for example Lesku et al. 2006; Santos 2009, 2012; Santos and Cannatella 2011). However, we think our method has some advantages, including, but not limited to, flexibility in the evolutionary model, ability to execute the analysis on the data as such rather than resorting to independent contrasts, and ability to include variables resulting in non-normal distribution of errors. Comparative methods are developing rapidly, for example, Chap. 9 in this book deals with phylogenetic logistic regression methods, which could in theory allow for phylogenetic path analysis including binary traits. Furthermore, the flexibility of PGLS would also allow for phylogenetic path

analysis to be undertaken accounting for variation in species traits (Martins and Hansen 1997), for example, using mixed models. The second aim of this chapter was to show how using phylogenetic path analysis novel questions in macroevolution can be addressed. Using our example with the simulated Rhinogradentia data, we showed how path analysis can help in disentangling evolutionary relationships between traits. For example, based on the results we can say, with some confidence, that litter size has no direct causal effect on range size in this fictitious mammalian order. We also show how phylogenetic path analysis can be employed to compare models with alternative causal relationships between variables. We must once again point out that the observed correlational pattern in the data can imply more than one underlying causal model, hence we might not always be able to distinguish between alternative causal models. Nonetheless, use of CICc, model comparison, and model averaging procedures can allow us to propose causal hypotheses among variables from the observed correlational patterns. Do we mean to say that employing this method we can do away with the limitations of comparative analyses for inferring causality pointed out at the beginning of the chapter? By no means! Such limitations are still there, and the statistical controls we use to disentangle cause–effect relationships are of course not comparable to the physical controls and randomizations we can apply in well-designed experiments. However, as stated at the beginning of the chapter, such an experimental approach is virtually impossible to carry out in the context of comparative analyses. Phylogenetic path analysis (using the *d*-separation method we propose or other approaches) may well be the only resort we have to infer causality in comparative studies. We must however keep in mind that path analysis is a hypothesis testing approach rather than a hypothesis-generating method. Carefully pondered and biologically meaningful, and supported, hypotheses of the causal relationships among studied traits must be presented before jumping into model testing. The end result of such a process is the confirmation of the plausibility of the proposed evolutionary causal model (although other alternative causal models can possibly explain the same observed correlation pattern), and probably more interestingly, the rejection of erroneous evolutionary causal models. We would therefore caution readers against overconfidence on the correctness of a causal model fitting the observed correlation structure; nonetheless we can be reasonably sure that rejected models are wrong. With all the uncertainties macroevolutionary studies must deal with, we think that the advantages provided by phylogenetic path analysis are not trivial. Furthermore, the causal model, or the set of models, we finally adopt as potential evolutionary explanations of the patterns we observe among the traits, can be formally challenged by alternative models in future studies involving new or better data. Such a process of presentation of a model (our causal hypothesis) and its provisory acceptance as plausible explanation of a causal phenomenon until it is confuted by an alternative model is at the very base of modern scientific methodology. We hope to have been successful in transmitting our enthusiasm for this method and to stimulate thought as to how it can allow you to tackle evolutionary questions in the context of comparative analyses.

Acknowledgments We thank László Zsolt Garamszegi for inviting us to write this chapter, as well as him and two anonymous referees for their useful comments and suggestions on a first draft of this chapter.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Arnold TW (2010) Uninformative parameters and model selection using Akaike's information criterion. J Wildl Manage 74(6):1175–1178

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York

Burnham KP, Anderson DR, Huyvaert KP (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol 65:23–35

Cardon M, Loot G, Grenouillet G, Blanchet S (2011) Host characteristics and environmental factors differentially drive the burden and pathogenicity of an ectoparasite: a multilevel causal analysis. J Anim Ecol 80:657–667

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125(1):1–15

Fisher RA (1926) The design of experiments, 1st edn. Oliver and Boyd, Edinburgh

Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22(7):1367–1375. doi:10.1111/j.1420-9101.2009.01757.x

Freckleton RP (2011) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol 65(1):91–101. doi:10.1007/s00265-010-1045-6

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat 160(6):712–726. doi:10.1086/343873

Garland TJ, Harvey PH, Ives AR (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst Biol 41:18–32

Geiger D, Verma T, Pearl J (1990) Identifying independence in Bayesian Networks. Networks 20:507–533

Grafen A (1989) The phylogenetic regression. Phil Trans Roy Soc B 326:119–157

Grewal R, Cote JA, Baumgartner H (2004) Multicollinearity and measurement error in structural equation models: Implications for theory testing. Mark Sci 23(4):519–529

Grim T (2008) A possible role of social activity to explain differences in publication output among ecologists. Oikos 117(4):484–487

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51(5):1341–1351

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Kline RB (2010) Principles and practice of structural equation modelling methodology in the social sciences, 3rd edn. Guilford Press, New York

Lesku JA, Amlaner CJ, Lima SL (2006) A phylogenetic analysis of sleep architecture in mammals: the integration of anatomy, physiology, and ecology. Am Nat 168(4):441–443

Martins EP (2000) Adaptation and the comparative method. Trends Ecol Evol 15(7):296–299

Martins EP, Diniz-Filho JA, Housworth EA (2002) Adaptation and the comparative method: a computer simulation study. Evolution 56:1–13

Martins EP, Garland T (1991) Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution 45(3):534–557

Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat 149(4):646–667

Matthews R (2000) Storks deliver babies ($p = 0.008$). Teach Stat 22(2):36–38

Messerli FH (2012) Chocolate consumption, cognitive function, and Nobel laureates. New Engl J Med 367(16):1562–1564

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877–884

Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov Chain Monte Carlo. Am Nat 167(6):808–825

Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan and Kaufmann, San Mateo

Pearl J (2009) Causality: models, reasoning and inference. Cambridge University Press, Cambridge

Petraitis PS, Dunham AE, Niewiarowski PH (1996) Inferring multiple causality: the limitations of path analysis. Funct Ecol 10:421–431

Pugesek BH, Grace JB (1998) On the utility of path modelling for ecological and evolutionary studies. Funct Ecol 12:853–856

Pugesek BH, Tomer A (1995) Determination of selection gradients using multiple regression versus structural equation models (SEM). Biometrical J 37:449–462

Quader S, Isvaran K, Hale RE, Miner BG, Seavy NE (2004) Nonlinear relationships and phylogenetically independent contrasts. J Evol Biol 17:709–715. doi:10.1111/j.1420-9101.2004.00697.x

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Meth Ecol Evol 1(4):319–329. doi:10.1111/j.2041-210X.2010.00044.x

Rohlf FJ (2006) A comment on phylogenetic correction. Evolution 60(7):1509–1515

Santos JC (2009) The implementation of phylogenetic structural equation modeling for biological data from variance-covariance matrices, phylogenies, and comparative analyses. The University of Texas at Austin, Austin

Santos JC (2012) Fast molecular evolution associated with high active metabolic rates in poison frogs. Mol Biol Evol 29(8):2001–2018

Santos JC, Cannatella DC (2011) Phenotypic integration emerges from aposematism and scale in poison frogs. Proc Natl Acad Sci USA

Shipley B (2000a) A new inferential test for path models based on directed acyclic graphs. Struct Equ Model 7(2):206–218

Shipley B (2000b) Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge

Shipley B (2004) Analysing the allometry of multiple interacting traits. Perspect Plant Ecol Evol Syst 6(235):241

Shipley B (2009) Confirmatory path analysis in a generalized multilevel context. Ecology 90:363–368

Shipley B (2013) The AIC model selection method applied to path analytic models compared using a d-separation test. Ecology 94(3):560–564

Stümpke H (1967) The Snouters: form and life of the Rhinogrades (trans: Doubleday & Company I). University of Chicago Press, Chicago

Team RDC (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Verma T, Pearl J (1988) Causal networks: semantics and expressiveness. In: Schachter R, Levitt TS, Kanal LN (eds) Uncertainty in artificial intelligence, vol 4. Elsevier, Amsterdam, pp 69–76

von Hardenberg A, Gonzalez-Voyer A (2013) Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. Evolution 67(2):378–387. doi:10.1111/j.1558-5646.2012.01790.x

Wilkinson GN, Rogers CE (1973) Symbolic description of factorial models for analysis of variance. Appl Stat 22(3):392–399