Developmental Characters in Phylogenetic Inference and Their Absolute Timing Information

MICHEL LAURIN* AND DAMIEN GERMAIN

UMR CNRS 7207 “Centre de Recherches sur la Paléobiodiversité et les Paléoenvironnements,” Muséum National d’Histoire Naturelle, Département Histoire de la Terre, Bâtiment de Géologie, Case Postale 48, 43 rue Buffon, F-75231 Paris Cedex 05, France;

*Correspondence to be sent to: UMR CNRS 7207 “Centre de Recherches sur la Paléobiodiversité et les Paléoenvironnements,” Muséum National d’Histoire Naturelle, Département Histoire de la Terre, Bâtiment de Géologie, Case Postale 48, 43 rue Buffon, F-75231 Paris Cedex 05, France; E-mail: michel.laurin@upmc.fr.

Received 19 April 2010; reviews returned 18 October 2010; accepted 16 February 2011

Associate Editor: Todd Oakley

Abstract.—Despite the recent surge of interest in studying the evolution of development, surprisingly little work has been done to investigate the phylogenetic signal in developmental characters. Yet, both the potential usefulness of developmental characters in phylogenetic reconstruction and the validity of inferences on the evolution of developmental characters depend on the presence of such a phylogenetic signal and on the ability of our coding scheme to capture it. In a recent study, we showed, using a new method (called the continuous analysis) using standardized time or ontogenetic sequence data and squared-change parsimony outperformed event pairing and event cracking in analyzing developmental data on a reference phylogeny. Using the same simulated data, we demonstrate that all these coding methods (event pairing and standardized time or ontogenetic sequence data) can be used to produce phylogenetically informative data. Despite some dependence between characters (the position of an event in an ontogenetic sequence is not independent of the position of other events in the same sequence), parsimony analysis of such characters converges on the correct phylogeny as the amount of data increases. In this context, the new coding method (developed for the continuous analysis) outperforms event pairing; it recovers a lower proportion of incorrect clades. This study thus validates the use of ontogenetic data in phylogenetic inference and presents a simple coding scheme that can extract a reliable phylogenetic signal from these data. [Developmental biology; evo-devo; heterochrony; parsimony; phylogeny; resolution; simulations.]

The relationship between evolution and development has inspired generations of biologists and is currently experiencing a rapid growth under the label “evo-devo,” as the publication of recent books (Hall 2007), reviews (Minelli 2007), and research papers in leading journals (e.g., Minelli et al. 2006; Jaekel and Wake 2007; Penet et al. 2007; Harrison and Larsson 2008) attests. One of the most exciting recent advances in this field is the possibility to study the evolution of developmental characters on reference phylogenies. This was made possible, to a large extent, by the formalization of the method called “event pairing” (Smith 1996; Velhagen 1997; Jeffery, Bininda-Emonds, et al. 2002, Jeffery, Richardson, et al. 2002), that enables the compilation of data matrices of discrete states that can be exploited using standard phylogenetic analysis software, such as PAUP* (Swofford 2003), MacClade (Maddison D.R. and Maddison W.P. 2003), Nona (Goloboff 2005), Mesquite (Maddison W.P. and Maddison D.R. 2010), TNT (Goloboff et al. 2008), and Winclada (Nixon 1999).

Below, we present briefly event pairing, discuss some of the reasons that have led to its formalization, and present an alternative, called continuous analysis, that we recently proposed in a study on the evolution of cranial ossification sequence in 23 species of urodeles based on an empirical data set (Germain and Laurin 2009). We then test, using simulations, the hypothesis that developmental data coded as event pairs or as recommended for continuous analysis, includes phylogenetic data that could be used to produce reliable phylogenies. We also test the relative performance of both coding schemes.

We further explore the properties of the standardization method that we proposed for continuous analysis and its potential advantages over other coding schemes. Given the recent surge of interest in evo-devo (Bininda-Emonds et al. 2002, 2003; Buckley et al. 2007) and the widespread idea that several developmental characters are highly conserved (Poe 2006; Fröbisch et al. 2007), the problem of developmental character nonindependence in the context of phylogenetic analyses is becoming a pressing issue in systematics. This is demonstrated by four related points. First, developmental biologists are often interested in determining which phylogeny their data support (e.g., Werneburg and Sánchez-Villagra 2009). Second, to study the evolution of such characters, a phylogenetic signal must be present; otherwise, character optimization would yield unreliable results (Laurin 2004, 2005). Thus, the very foundation of the blossoming field of evo-devo would be undermined if it were shown that developmental characters do not usually include a phylogenetic signal. Third, the small quantity of accessible developmental data does not imply that these data are trivial; when combining data of different kinds, several authors opt for partitioned analyses or more sophisticated statistical techniques to see if various kinds of data support the same topology (e.g., Lee 2009; Lee and Camens 2009). Fourth, several developmental characters concerning mineralized or silicified structures can fossilize; such data have figured prominently in the debates about the origin of various taxa, such as lissamphibians (e.g., Schoch 2006; Fröbisch et al. 2007; Germain and Laurin 2009), and have been used to assess various evolutionary questions in a wide range of tracheophytes (Niklas 1997; Boyce and Knoll 2002) and of metazoans, such as trilobites (Hughes et al. 2006) and ostracods (Watabe and Kaesler 2004).
Event pairing

Below, we assess the performance of the coding scheme used for event pairing in phylogenetic analyses. Event pairing is based on comparisons between pairs of events whose relative chronology is scored. Event pairing requires comparing the relative timing of every possible pair of events of a matrix, typically coded as 0 if Event A happens before Event B, 1 if both events are simultaneous, and 2 if Event A happens after Event B. Because this yields a symmetrical matrix, only half of it (minus the diagonal) needs to be coded. Therefore, if \( n \) events (characters) are analyzed, the event-pairing matrix will include \( n^2(n−1)/2 \) event-pair characters.

Event pairing was developed to overcome the difficulty of interspecific comparisons of ontogeny because it is sometimes difficult to find strictly equivalent stages, and even the number of stages can differ. Optimizing the resulting characters on a reference phylogeny yields changes that represent heterochrony. However, a single event moving in a sequence will generate several linked changes; for instance, if events occur in the sequence ABCDEFG in an ancestor and in the sequence BCDAEFG in one of its descendants, event pairing would be expected to yield changes in the order between A and B, A and C, and A and D. Note that taken individually, these data do not indicate whether A or B has moved (in the first pairwise comparison), whether A or C has moved (in the second comparison), or whether A or D has moved. However, because events B, C, and D did not move compared with other events, it would be easy to conclude that only A has moved. But many cases are more complex, with several events moving simultaneously in a sequence, so complementary methods called “event-pair cracking” (Jeffery, Bininda-Emonds, et al. 2002; Jeffery, Richardson, et al. 2002), search-based optimization (Schulmeister and Wheeler 2004), and “Parsimov” were developed to isolate the “significant” changes (Jeffery et al. 2005). These methods rely on identifying the elements moving in the greatest number of pairwise comparisons, using parsimony and various algorithms to select the lowest number of changes required to explain the observed pattern.

A drawback of event pairing, event-pair cracking, search-based optimization, and Parsimov is that they are complex and time consuming to implement (especially the last two), and Schulmeister and Wheeler (2004, p. 55) admitted that their method of search-based optimization could be applied only to “a very small number of events.” The inconsistent results yielded by event pairing in some situations have led other authors to seek alternatives as a basis to evaluate potential heterochronies (Harrison and Larsson 2008).

Little has been done to assess the statistical properties of event-pair coding for phylogenetic analysis, or of cracking for heterochrony detection, using simulations. Koenemann and Schram (2002) simulated four data sets to compare the performance of event pairing and of a more direct coding and of various phylogenetic inference methods. Bininda-Emonds et al. (2003) simulated 10,000 matrices representing developmental sequences, but only to determine if their data set on mammalian development included more (or fewer) simultaneous events and if the data had a more clumped distribution than randomly generated sequences; they did not use their simulation to assess performance of event pairing. Extensive simulation-based tests of the performance of event pairing for phylogenetic inference are performed below for the first time.

Time and Sequences in Heterochrony Analysis

Below, we examine the potential usefulness of ontogenetic time data through simulations. Using unstandardized size to assess the timing of developmental events may obscure patterns, as pointed out by Smith (2001) and Koenemann and Schram (2002), so time was often discarded from more recent analyses. Perhaps this decision was not entirely justified, as a reanalysis of the data used by Smith (2001) suggests. These data, borrowed from Alberch et al. (1979; figs. 1 and 3), represent the relative timing of ossification of various cranial bones in the ontogeny of three urodele species, Triturus vulgaris, Ambystoma texanum, and Ambystoma mexicanum. Smith (2001) showed that when the timing of events was assessed by the body size (in mm) at which they occur, A. mexicanum looks very different from the two other species (Fig. 1a). On the other hand, when the events are simply ranked according to the sequence in which they occur in T. vulgaris (Fig. 1b), similarities between both species of Ambystoma, which were expected because of their relatedness, become obvious. From this, it could be concluded that body size is a poor proxy for developmental time and that using sequences is better (although Smith expressed a more moderate opinion), which would validate the use of methods that do not rely on time or a proxy, such as event pairing.

Examination of the data suggests that the strong deviation between the pattern of A. mexicanum and the other species simply reflects its much larger size; at the end of the examined developmental series, it measures 188 mm, as opposed to 25 mm in A. texanum and 33 mm in T. vulgaris. Using standardized size as a proxy of time (Fig. 1c and Table 1) resolves these problems and highlights similarities between both species of Ambystoma. Furthermore, this metric shows that the last events in the sequence of T. vulgaris occur at very similar body sizes (which does not necessarily indicate close timing), a pattern that is hidden by the use of sequences. Standardized size also reveals that in A. mexicanum, Event I (prootic ossification) occurs at a much smaller size (compared with its maximal size) than in A. texanum (Fig. 1c), a pattern not so obvious when looking at sequence data (Fig. 1b). Thus, these data, far from supporting the claim that body size is an invalid criterion to place events in ontogeny, indicate that if size is standardized, this criterion can reveal more information than sequences.
FIGURE 1. Development of three urodele species assessed through three scales. The data, originally presented by Alberch et al. (1979, figs. 1 and 5) and reused by Smith (2001, table 2 and fig. 3) are presented in Table 1. The occurrence of Events A–M is assessed either by the size (mm) of the specimen at which they occur (a), by their position in the sequence (b) of *Triturus vulgarus*, or by the standardized size of specimens (c). The events correspond with the first ossification of the following bones: A = splenial, vomer, and palatine; B = dentary; C = pterygoid, squamosal, and premaxillary; D = prearticular; E = frontal; F = parietal; G = opisthotic; H = first basibranchial and orbitosphenoid; I = prootic; J = quadrate; K = nasal; L = maxilla and parasphenoid; M = prefrontal.

Continuous Analysis

Using simulations, we analyze the statistical properties of a new coding method for heterochrony analysis called “continuous analysis,” that was proposed by Germain and Laurin (2009). This method relies on standardizing event times (Fig. 2) on the same scale in all species (typically, using an interval ranging from 0, the time of the first event, to 1, the time of the last event). Standardized time (Fig. 3a) or sequence (Fig. 3b) data can then be analyzed to infer ancestral conditions (ancestral developmental times of various events) using squared-change parsimony (Maddison 1991), as implemented in Mesquite (Maddison W.P. and Maddison D.R. 2010). Confidence intervals (CIs) on all nodal values can be computed using phylogenetically independent contrasts (Felsenstein 1985); these are the same as maximum likelihood (ML) estimates of the CIs and can be computed using the PDAP:DPTREE module for Mesquite (Midford et al. 2003). Heterochronies are considered statistically significant if the measured (for terminal taxa) or estimated (for nodes) value of a taxon is outside the 95% CI of its hypothetical ancestor (the node directly below). Thus, this method requires a phylogeny with estimated branch lengths.

Relative time data can be used directly, if available. If not, standardized sequence data can be used instead. If there are \( n \) events occurring at various times in the ontogeny, the interval (typically between 0 and 1) is broken into \( n - 1 \) intervals of equal value. For instance, with six events, as in the data simulated by Germain and Laurin (2009), the values would be multiples of 0.2, and a sequence in which events ABCDEF occur in this order in ontogeny would result in the values 0, 0.2, 0.4, 0.6, 0.8, and 1 being attributed to these events (such even spacing between events is automatically produced when sequence data are used, but not if time data are used). The general formula to standardize between 0 and 1, whether time, a proxy of time, or position is used is

\[ X_s = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \]

where \( X_s \) is the standardized value of \( X \) (time, a proxy, or sequence), \( X_i \) is an individual value of \( X \), \( X_{\text{min}} \) is the lowest observed value of \( X \), and \( X_{\text{max}} \) is the largest observed value of \( X \) (in the ontogeny of a taxon). If only sequence data are available and if there are simultaneous events, the average rank of simultaneous events should be used as recommended by Smith (2001; table 2). Germain and Laurin (2009) reported that using standardized sequences decreased power to detect heterochronies (compared with using standardized time); below, we test both coding schemes to determine their relative efficiency for phylogenetic analyses.

An example (Fig. 2) is subjected to continuous analysis (Germain and Laurin 2009) on standardized time to detect heterochronies, to show that time data can reveal heterochronies that would go unnoticed in event pairing because no change occurs in the sequence of events. The data were analyzed using the PDAP module for Mesquite (Midford et al. 2003) using the tree shown in Figure 2c.
FIGURE 2. Example showing how using standardized developmental time can reveal changes in developmental rates that would go unnoticed if only sequence information were used. (a) Absolute timing of Events A–F. T = taxon. Note that there is no change in sequence (hence, event pairing would detect no heterochrony). (b) Standardized time for the same data. (c) Phylogeny used to analyze the data through the continuous analysis. The branch on which significant heterochronies in Events B–E occurred is identified by its gray shade. Branch lengths are relative; the units could represent Ma, among other possibilities.

The continuous analysis (Germain and Laurin 2009) on standardized time data presented in Figure 2 detects significant heterochronies in Events B and C (moving toward earlier standardized time) and D and E (moving toward later standardized time) on the branch leading to the node that includes only taxa 4 and 5, which is the correct answer because we introduced these heterochronies when generating the data. No other heterochronies are detected for these data. Of course, event pairing (even with methods such as event-pair cracking and Parsimov) or the continuous analysis on sequences yield no changes because there is no variation in the sequences.

Continuous analysis is reasonably robust to missing data. Analysis of the data from Figure 3h using the tree shown in Figure 2C does not reveal any additional spurious heterochronies (Table 2) resulting from the absence of data for Events B and C (for taxa 2 and 6). All heterochronies detected in the presence of missing data (Fig. 3d–i) are also detected with the complete data set (Fig. 3a–c).

Testing the Phylogenetic Usefulness of Developmental Data

This study represents an attempt at assessing the usefulness of developmental data in phylogenetic analyses, a topic on which the scientific community is sharply divided. Developmental sequences are not ideal characters with which to perform phylogenetic analysis because the characters are not independent (the position of Event D in a sequence is not independent from the position of Events C and E, for instance). Event pairing compounds this problem by generating multiple characters for each event (each of the \( n \) events is represented by \( n - 1 \) characters in an event-pair matrix), as recognized by Jeffery, Richardson, et al. (2002, p. 301), although matrices for continuous analysis are exempt from this problem. Despite this caveat, phylogenetic analyses have often been performed on event-pair matrices to see which topology was supported (e.g., Jeffery, Richardson, et al. 2002; Sánchez-Villagra 2002; Schoch 2006; Werneburg and Sánchez-Villagra 2009), and Bininda-Emonds et al. (2002, p. 297) suggested that event pairing “may also yield data that can be used in phylogeny reconstruction.” In fact, because developmental characters evolve like any other characters (developmental times may evolve according to a Brownian motion model and developmental sequences, according to a Markov model), there is no a priori reason to exclude them from consideration in phylogenetic inference. However, this raises the question of how to code them to recover the most resolved and most accurate topology. This study attempts to answer this question using the simulated data previously analyzed by Germain and Laurin (2009). Because simulated data may not match empirical biological data, we also analyze two empirical developmental data sets. We also discuss some advantages of the proposed standardization method for heterochrony analysis and for reconstruction of ancestral ontogenies, and we assess the impact of missing data on such analyses. We also discuss theoretical and empirical justifications of the standardization procedure that is required by the continuous analysis.
**FIGURE 3.** Simple example showing the relationship between absolute developmental times (often unknown), here presented on a standardized scale, and developmental sequences. This example is matrix replicate 2 of the third data set used by Germain and Laurin (2009). Heterochronies compared with the extra group are marked in bold in b–c, e–f, and h–i. Standardized times of each event are shown in the upper tables, either without (a) or with (d, g) missing data. Resulting standardized sequences are shown in the middle row (b, e, h). Either can be subjected to comparative analysis, although times (upper tables) are better, when available. The lower tables (c, f, i) show data transformed into states for phylogenetic analysis (c, f), or data that still needs to be rounded off to the nearest integer to yield such data (i). In the middle column (d–f), the position of events in taxa in which an event is missing is aligned (on either side of the missing event) based on the position of the same events in taxa without missing data. In the right column, the standardized sequences are produced by breaking the interval (normally, 0 to 1) into \( n \) intervals, where \( n \) represents the number of events of different time or sequence position.

**METHODS**

**Simulation and Transformation of Data for Phylogenetic Analyses**

The data generated by Germain and Laurin (2009) include three sets of 100 matrices including time and sequence information about six characters in eight terminal taxa and the hypothetical ancestors. These three sets include characters of increasingly complex evolutionary simulated on the tree shown by Germain and Laurin (2009). Heterochronies compared with the extra group are marked in bold in b–c, e–f, and h–i. Standardized times of each event are shown in the upper tables, either without (a) or with (d, g) missing data. Resulting standardized sequences are shown in the middle row (b, e, h). Either can be subjected to comparative analysis, although times (upper tables) are better, when available. The lower tables (c, f, i) show data transformed into states for phylogenetic analysis (c, f), or data that still needs to be rounded off to the nearest integer to yield such data (i). In the middle column (d–f), the position of events in taxa in which an event is missing is aligned (on either side of the missing event) based on the position of the same events in taxa without missing data. In the right column, the standardized sequences are produced by breaking the interval (normally, 0 to 1) into \( n \) intervals, where \( n \) represents the number of events of different time or sequence position.

The second data set included two additional directional changes of low magnitude that represented significant heterochronies (two characters had their value increased or decreased by a given number on one branch each). The third data set included changes of larger magnitude and more overlap between the simulation boundaries of successive characters. See Germain and Laurin (2009, appendix 5) for more information on the simulation settings.

Because the third set of matrices is expected to include the largest number of phylogenetically informative characters, it was selected to assess the suitability of both event-pair and continuous analysis matrices for parsimony-based inference of phylogenies. For each character in these matrices, absolute and relative (sequence) timing data are available. We expect that using absolute timing would give better results than using only sequence data, but most phylogenetic analysis...
software cannot use continuous characters directly, TNT being the exception (Goloboff et al. 2006). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001). An alternative would be to build complex stepmatrices with gap weighting for each character (Wiens 2001).

Two empirical developmental data sets were analyzed: the first (online Appendix 1, available from http://www.sysbio.oxfordjournals.org/) was extracted by Koenemann and Schram (2002) from the data on amniotes presented by Jeffery, Bininda-Emonds, et al. (2002) and includes 11 species and 29 events (Fig. 4); the second (online Appendix 2) is an expanded version of the urodele data analyzed by Germain and Laurin (2009), to which we have added two gymnophionans, four anurans, and the temnospondyl Apateon, representing an outgroup. This lissamphibian data set includes 30 taxa and 12 events, but 2 taxa were excluded (a hybrid and Aneides, for which there are mostly missing data because the ontogenetic series is very coarse and insufficient to determine the sequence of ossification), resulting in an analyzed matrix of 28 taxa (although the 30 taxa are presented in the appendix). The additional data on Apateon originally came from Schoch (2006), the data on gymnophionans come from Müller et al. (2005; fig. 4), and those on anurans were extracted from Yeh (2002; table 1). The impact of “ranked events” scoring of Koenemann and Schram (2002) (the use of raw sequence position) was investigated because these authors concluded that the use of such unstandardized sequence position data was inadequate and that event pairing was a more accurate method. Comparisons with our standardized continuous coding enable us to determine which of the three coding schemes (the third being standardized sequence) is the most appropriate.

Table 1. Size, rank, and standardized size data for urodele species

<table>
<thead>
<tr>
<th>Events</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triturus vulgaris</td>
<td>5</td>
<td>5.5</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>17</td>
<td>18</td>
<td>25</td>
<td>30</td>
<td>31</td>
<td>31</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>Ambystoma texanum</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>22</td>
<td>18</td>
<td>25</td>
<td>8</td>
<td>12</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Ambystoma mexicanum</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>20</td>
<td>26</td>
<td>188</td>
<td>91</td>
<td>91</td>
<td>11</td>
<td>23</td>
<td>22</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>T. vulgaris</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>A. texanum</td>
<td>1.5</td>
<td>1.5</td>
<td>5.5</td>
<td>3.5</td>
<td>5.5</td>
<td>7.5</td>
<td>12</td>
<td>10.5</td>
<td>13</td>
<td>3.5</td>
<td>7.5</td>
<td>9</td>
<td>10.5</td>
</tr>
<tr>
<td>A. mexicanum</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>13</td>
<td>11.5</td>
<td>11.5</td>
<td>2.5</td>
<td>8</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>T. vulgaris</td>
<td>0</td>
<td>0</td>
<td>167</td>
<td>56</td>
<td>167</td>
<td>278</td>
<td>833</td>
<td>611</td>
<td>1000</td>
<td>56</td>
<td>278</td>
<td>444</td>
<td>611</td>
</tr>
<tr>
<td>A. texanum</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>51</td>
<td>85</td>
<td>1000</td>
<td>452</td>
<td>452</td>
<td>0</td>
<td>68</td>
<td>62</td>
<td>249</td>
</tr>
<tr>
<td>A. mexicanum</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>20</td>
<td>26</td>
<td>188</td>
<td>91</td>
<td>91</td>
<td>11</td>
<td>23</td>
<td>22</td>
<td>55</td>
</tr>
</tbody>
</table>
| Notes: Data originally come from Alberch et al. (1979), but were reused by Smith (2001, p. 176). Xs = (Xi-Xmin)/(Xmax-Xmin), where Xs is the standardized value of X (time, a proxy, or sequence), Xs is an individual value of X, Xmin is the lowest observed value of X, and Xmax is the largest observed value of X. Standardized time was multiplied by 1000, to save space in the table. Note that there was a small error in Smith’s (2001) table: for Ambystoma mexicanum, the cells scored “2” for rank should have been scored “2.5.” Ra = rank in the sequence of events; SS = specimen size (mm); SSS = specimen standardized size.

Table 2. Impact of missing data on heterochrony detection in continuous analysis

<table>
<thead>
<tr>
<th>Changes on branch leading to taxa</th>
<th>Char</th>
<th>T 1 (2–7)</th>
<th>T 2 (3–7)</th>
<th>T 3 (4–5)</th>
<th>T 4</th>
<th>T 5 (6–7)</th>
<th>T 6</th>
<th>T 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: The data from Figure 3h were analyzed on the tree shown in Figure 2c. The range of taxa above each branch is given by two numbers separated by a dash; for instance “(2–7)” designates the smallest clade that includes taxa 2–7. For some taxa, changes cannot be assessed because of the missing data; these are designated by a dash (–). 0 = absence of change; 1 = presence of change on the branch leading to the indicated taxon. Char = character; T = taxon.
Phylogenetic Analyses

Given that the simulated matrices include eight taxa and only six characters, it is impossible to get fully resolved trees; the two empirical data sets suffer from similar limitations. Thus, to assess the suitability of both methods, we compared the number of correct and of incorrect clades recovered by the 100 matrices. A correct clade is one that is found in the tree used to generate the simulated data (Germain and Laurin 2009; fig. 2), or that reflects current understanding of tetrapod phylogeny (for empirical data sets); an incorrect clade is one that does not exist in the simulation tree, or that departs from the established consensus (for empirical data sets). The method that recovers the highest number of correct clades and the lowest number of incorrect clades should be preferred. An analogous method was used by Bininda-Emonds and Sanderson (2001) to assess the properties of “total evidence” and MRP supertree methods, and the usefulness of simulations in statistics in general is well established (e.g., Legendre 2000).

To test the hypothesis that an increase in the amount of developmental data would result in better phylogenetic estimates, we combined the 100 matrices into 25 or 5 supermatrices (including 4 or 20 times more characters than the original matrices). We then compared the performance of the analyses based on data sets of different sizes.

All phylogenetic analyses were performed using PAUP* 4.0 (Swofford 2003) using simple parsimony with ordered states. Note that we did not perform continuous analysis here; thus, squared-change parsimony was not used because the purpose of our analyses is to recover phylogenies, not to detect heterochronies. However, our analyses are meant to show whether coding standardized sequences or event pairs perform best in phylogenetic inference. The simulated matrices did not include an outgroup and the basal dichotomy of the simulated tree is between large clades. Using the matrices as such would have been difficult because the position of the root would have been ambiguous. Therefore, an ideal outgroup was added. It was produced using the initial character values (positions in the sequences) used in the simulations. Because the PDSIMUL module (Garland et al. 1993) of PDAP (Garland et al. 2002) used these values at the root, this outgroup is analogous to a fossil linked to the root by a very short branch or a conservative extant species preserving the ancestral sequence. Note that some ancestors have been putatively identified in some taxa with a good fossil record (Webster and Purvis 2002). As the same ideal outgroup is used for all analyses (of event pairing and of continuous matrices), this should not bias the comparisons of relative performance. Only 59% of the original characters are parsimony informative, and given the redundancy of event-pairing characters, only 16% of those are parsimony informative. All characters were considered ordered.

The two empirical data sets were similarly analyzed by scoring the number of correct and incorrect clades. The reference (“correct”) phylogeny follows Hackett et al. (2008) for avian phylogeny, Asher (2007) for placental phylogeny, Wiens et al. (2005: fig. 8) for urodele phylogeny, Marjanović and Laurin (2007) for anuran phylogeny, and Laurin and Reisz (1995) for the position of turtles. This choice of source phylogenies is to an extent personal and subjective (we simply used phylogenies that appeared plausible, or that appeared to be well-supported by evidence), but these were selected before the developmental data sets were analyzed. Therefore, the choice of reference trees should not bias the results. Because both data sets include more than 10 taxa, more than 10 states per character can occur (and often do occur) because once the times are standardized, there can be as many states as taxa. This creates a dilemma because not all phylogenetic analysis software can accommodate more than 10 states per character. To assess the impact of the number of character states, both coding methods were used (one limited to 10 states and another allowing more states per character). For the data of Jeffery, Bininda-Emonds, et al. (2002), this results in a maximum of 11 character states; for our lissamphibian data, this results in up to 21 states per character (for 30 taxa).

The low number of empirical data sets (two, as opposed to 100 simulated matrices) makes drawing statistically valid conclusions challenging. To minimize this problem, we have performed an additional analysis on these matrices to determine whether they included a phylogenetic signal. The length of whole matrices over the reference tree (Figs. 4a and 5a) was compared with the length of a population of 10,000 equiprobable trees generated by Mesquite. The proportion of random equiprobable trees that imply fewer or as many steps as the reference tree is the probability that the association between the character matrix and the phylogeny is random. To better distinguish between coding methods when both yielded equal probabilities, we also looked at the difference between the number of steps implied by the shortest equiprobable tree and the reference tree; the largest difference should reflect the strongest phylogenetic signal.

RESULTS

Phylogenetic Analyses of Simulated Data

Both methods gave little phylogenetic resolution on the original matrices from Germain and Laurin (2009), with 0.73–1.05 correct clades recovered by the original matrices (of six characters), yielding by analysis of the event-pair and continuous (standardized) matrices, respectively (Table 3). As expected, resolution increased in both cases with the use of supermatrices, yielding on average 1.96 (event pairing) and 2.76 (continuous analysis) correct clades/supermatrix with the supermatrices of 24 characters and more than four correct clades/supermatrix of 120 characters. In all cases, continuous analysis yielded better resolution.
Continuous analysis also outperformed event pairing in all analyses, according to the proportion of correct clades among those recovered (Table 3). Using the initial 100 simulated matrices (6 characters each; 15 event-pairing characters), the matrices for continuous analyses yielded 105 correct clades and 83 incorrect (artifactual) clades. Event-pairing matrices yielded only 73 correct clades and 84 incorrect clades (Table 3). Using the 25 supermatrices of 24 characters each (60 event-pairing characters each), results were more similar (Table 3), with proportions of correct clades of 0.719 for continuous analysis and 0.681 for event pairing. With the largest supermatrices (120 characters; 300 event-pairing characters), continuous coding led to a perfect score (no incorrect clades), and event pairing had a good score (proportion of correct clades of 0.913). Thus, both methods seem to converge toward the correct phylogeny as the amount of data increases (Table 3).

**Phylogenetic Analyses of Empirical Developmental Data**

Analysis of empirical developmental data shows proportions of correct clades slightly lower than for simulated data (Table 4). The results, merged over both
FIGURE 5. continued
data sets, show that continuous coding with multiple states (more than 10 per character) performs best, followed by continuous coding with 10 states per character, and by event pairing. The ranked events scoring used by Koenemann and Schram (2002) performed worst.

The tests of phylogenetic signal using equiprobable trees show that both matrices under all coding methods contain some phylogenetic signal (Table 5). The probabilities ($P$ values) are lowest for continuous coding with more than 10 states per character and event pairing. The difference in the number of steps between the reference tree and the shortest equiprobable tree is highest for continuous coding with more than 10 states per character.

**TABLE 3.** Performance of continuous analysis and event-pairing in phylogeny reconstruction assessed through simulation of matrices on a known phylogeny.

<table>
<thead>
<tr>
<th>Number of matrices and of characters per matrix</th>
<th>Method</th>
<th>Number of correct clades recovered</th>
<th>Number of artificial clades recovered</th>
<th>Proportion of correct clades</th>
<th>Correct clades/matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>100, 6</td>
<td>Continuous</td>
<td>105</td>
<td>83</td>
<td>0.559</td>
<td>1.05</td>
</tr>
<tr>
<td>100, 6 (15)</td>
<td>Event pairing</td>
<td>73</td>
<td>84</td>
<td>0.465</td>
<td>0.73</td>
</tr>
<tr>
<td>25, 24</td>
<td>Continuous</td>
<td>69</td>
<td>27</td>
<td>0.719</td>
<td>2.76</td>
</tr>
<tr>
<td>25, 24 (60)</td>
<td>Event pairing</td>
<td>49</td>
<td>23</td>
<td>0.681</td>
<td>1.96</td>
</tr>
<tr>
<td>5, 120</td>
<td>Continuous</td>
<td>23</td>
<td>0</td>
<td>1.000</td>
<td>4.60</td>
</tr>
<tr>
<td>5, 24 (300)</td>
<td>Event pairing</td>
<td>21</td>
<td>2</td>
<td>0.913</td>
<td>4.20</td>
</tr>
</tbody>
</table>

Notes: Three related data sets were used. First, 100 matrices were composed of six characters simulated on the phylogeny shown in Figure 2. The resulting event-pairing matrices included 15 characters. Second, 25 supermatrices were compiled using four matrices of six characters each (for the continuous analysis), or 15 characters each (event pairing), which resulted in supermatrices of 24 (continuous analysis) or 60 (event pairing) characters each. Third, five supermatrices were compiled using 20 of the initial matrices each; these included 120 characters (300 event-pair characters). In all cases, an extra group with a state corresponding with the root was added prior to the parsimony analysis. In the first column, the number of resulting event-pair characters is given in parentheses.

**TABLE 4.** Performance of coding methods for phylogenetic inference on empirical data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Coding method</th>
<th>Correct clades</th>
<th>Incorrect clades</th>
<th>Proportion correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amniote</td>
<td>Continuous, &gt; 10 states</td>
<td>2</td>
<td>1</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Continuous, ≤ 10 states</td>
<td>2</td>
<td>4</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Ranked events</td>
<td>2</td>
<td>5</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Event pairing</td>
<td>2</td>
<td>3</td>
<td>0.57</td>
</tr>
<tr>
<td>Lissamphibia</td>
<td>Continuous, &gt; 10 states</td>
<td>3</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Continuous, ≤ 10 states</td>
<td>5</td>
<td>12</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Ranked events</td>
<td>1</td>
<td>6</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Event pairing</td>
<td>3</td>
<td>21</td>
<td>0.13</td>
</tr>
<tr>
<td>Sum of both</td>
<td>Continuous, &gt; 10 states</td>
<td>5</td>
<td>4</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Continuous, ≤ 10 states</td>
<td>7</td>
<td>16</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Ranked events</td>
<td>3</td>
<td>11</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Event pairing</td>
<td>7</td>
<td>24</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Notes: The amniote data set refers to part of the data presented by Jeffery, Bininda-Emonds, et al. (2002), as reused by Koenemann and Schram (2002). Note that the taxon that identified as “monkey” by Koenemann and Schram (2002) appears to be the spectral tarsier (not a monkey, but a haplorhine primate), and that not all taxa studied by Jeffery, Bininda-Emonds, et al. (2002) were used by Koenemann and Schram (2002). The lissamphibian data set is from Germain and Laurin (2009). States were ordered, contrary to what Koenemann and Schram (2002) did. Performance is based on a comparison between the recovered tree and the established consensus. See the text for the source of reference phylogenies.

**Figure 5.** Phylogenetic analyses of the lissamphibian data used by Germain and Laurin (2009), with additions from Yeh (2002) and Müller et al. (2005). (a) Established consensus; (b–e) strict consensus of the shortest trees yielded by a phylogenetic analysis of the data when coded according to (b), continuous coding, with potentially as many states as taxa; (c) continuous coding, with no more than 10 states/character; (d) ranked events; (e) event pairing. Named clades are all considered correct. Lissamphibia was not counted among the correct taxa yielded by the analyses because it was produced by the choice of outgroup and rooting option. C = correct clade (unnamed); U = clade of unknown status (could be correct or wrong); W = wrong (artifactual) clade.

**DISCUSSION**

**Performance of Event Pairing and of the Continuous Analysis**

The low number of correct clades recovered per matrix and the relatively high proportion of artificial clades are probably attributable mostly to the low proportion of parsimony-informative characters in the original matrices and the low character-to-taxon ratio in the matrices (six characters for eight terminal taxa, in the simulated matrices). This conclusion is supported by the fact that both the resolution and proportion of correct clades improved as the amount of data increased (Table 3). Thus, the suggestions that “…the nonindependence of the [event-pair] characters would introduce considerable problems in the [phylogenetic] analysis” (Smith 1997, p. 1671) and “…developmental sequences treated as event-pairs can at best be mapped in robust phylogenies to examine heterochrony but not used as data to generate phylogenetic hypotheses” (Sánchez-Villagra 2002, p. 272) appear too pessimistic. Our results show that, with enough data, using developmental data leads to the correct phylogeny. However, Sánchez-Villagra (2002, p. 271) was correct in stating that “[n]onindependence probably characterizes many phylogenetic data sets, both molecular and morphological, but this problem is exacerbated when dealing with
event-pairs characters…” Our results clearly show that if sequences are to be analyzed to produce phylogenies, this would best be done by treating the position of each character in time or in the sequence, rather than by generating event pairs.

The ability to recover real clades was worse for empirical developmental data than for simulated data, but the amount of empirical developmental data is too small to determine if this is attributable to mere stochastic variations, or if this reflects a real feature of developmental data. The low character-to-taxon ratio (2.64 and 0.43, for the amniote and lissamphibian data sets, respectively) provides a possible explanation. The phylogenetic usefulness of developmental data is better demonstrated (with high statistical confidence) by the equiprobable tree test (Table 5). Both continuous coding with more than 10 states per character and event pairing yield the smallest possible probability (< 0.0001, using 10,000 equiprobable trees) for both data sets. However, the difference in the number of steps implied by the shortest equiprobable tree and the reference tree (larger numbers are better); LRT = length of reference tree; LSET = length of shortest equiprobable tree; P = probability that the data do not show association with the reference tree.

Role of Developmental Data in Phylogenetic Inference

Developmental data could be used in phylogenetic inference as any other type of data, and indeed, this has occasionally been done recently (e.g., Werneburg and Sánchez-Villagra 2009). Another potential use of such data is to resolve the polytomes that are often obtained using various types of data in phylogenetic analyses. In that case, the relatively small amount of available data (when compared with nucleotide sequences) would matter less than in a combined approach. Most phylogenetic data matrices incorporate only discrete characters, which suggests that sequences rather than standardized time should be used. However, some phylogenetic analysis software can now use quantitative characters, which would allow use of standardized time directly into phylogenetic analyses (Goloboff et al. 2006). This would be preferable because our results show that part of the phylogenetic signal is lost when time data are converted into sequences.

The reliability of developmental data in phylogenetic inference presumably depends heavily on the evolutionary rates of these characters and on the taxon sampling. Fortunately, various tests can assess if data sets contain significant phylogenetic signal (Hillis 1991; Diniz-Filho et al. 1998; Blomberg et al. 2003; Laurin 2004).

Developmental Time and Sequence Data in Evo-Devo

Going back to using standardized time (allowed, but not required for continuous analysis) to compare the development of taxa might seem like a step backward because event pairing was developed precisely to avoid the difficulty of finding comparable developmental stages (Jeffery, Richardson, et al. 2002). These problems are created by the lack of universal stable characters in development. Jeffery, Richardson, et al. (2002) mentioned that “…absolute chronological age is an unsatisfactory metric in comparative studies, because of interspecific variation in developmental rates…” Jeffery, Bininda-Emonds, et al. (2002) argued that “factors such as temperature, nutrition, and inraspecific (genetic) variation all affect the rate of development.” Similar considerations led Bininda-Emonds et al. (2002) to suggest that no absolute temporal scale can be used.

These problems do not appear so limiting as to preclude the possibility of devising a simple efficient method to analyze evolution of absolute developmental time, as shown above. Although we recognize that there are no universally applicable developmental stages, several events occur in very old taxa and could be used as independent markers of developmental time. These include (the list is not exhaustive) the following: fertilization (for all sexually reproducing organisms), beginning of cell cleavage (for all metazoans and many colonial eukaryotes), formation of the blastula, beginning of gastrulation (for metazoans), formation of the first somite (vertebrates), hatching (for most oviparous taxa), birth (for viviparous taxa), and sexual maturity or age at which the first mating usually occurs (sexually...
reproducing organisms). These events can be subject to heterochrony, but they are nevertheless useful markers of ontogenetic development. Note that even a substantial distance between the events selected to standardize time (or sequences) and the bulk of the events whose evolution is of interest would not preclude using continuous analysis.

If no time marker is selected, it is possible to select the time at which the first and last events occur in all the species to standardize time, even if the position of these events varies in the developmental sequence. For instance, if a series of 10 events (A–J) is studied in eight taxa, and if, in eight of these taxa, Event A occurs first and Event J occurs last, whereas in one of these taxa, Event B occurs first and, in another taxon, Event H occurs last, the first and last events (A and J in most cases, but B and J in one taxon, and A and H in another) can be used to set the time interval used for standardization. In fact, several proponents of event pairing have standardized developmental time to compare relative timing of various developmental events (e.g., Smith 1997, fig. 4).

Variations in intraspecific developmental rate attributable to factors such as temperature, nutrition, and genetic variation can be minimized by the simple standardization procedure proposed above. If such factors affect the timing of all events in similar ways and if their effect does not vary greatly within the part of the ontogeny under study, such variations are eliminated by our standardization procedure. However, if they affect the events under study differently, or if they accelerate part of the ontogeny and slow down another part, problems will obviously arise. Hopefully, this catastrophic scenario is not a common situation (and most known factors, such as temperature or other environmental factors, tend to have systemic effects, rather than local ones), but only application of our method to several data sets can reveal how common such problems are in developmental data.

Germain and Laurin (2009, tables 4–6; appendices 9–14) showed that all analytical methods (Parsimov on eventpairing, continuous analysis on standardized time, or sequences) had low Type I error rate, but all displayed a greater error rate on terminal than on internal branches. This property of Parsimov was previously unknown and may explain why Bininda-Emonds et al. (2003, p. 341), using the related method of event-pair cracking, found that “far fewer shared changes were inferred to have occurred on internal branches than did unshared changes on terminal branches (on average, 3.3 vs. 22.8 events, respectively).” Fortunately, Bininda-Emonds et al. (2003, p. 344) focused the discussion of their results on shared changes because the terminal changes were “difficult to interpret objectively.” Nevertheless, they suggested (p. 345) that “[m]any of these lineages [main clades of placental mammals] are characterized by few or no timing shifts (e.g., primates, rodents, or artiodactyls). Instead, heterochrony among mammals appears to have occurred much more recently in evolutionary time, possibly implying an increased role in speciation.” The higher Type I error rate of all heterochrony analyses on terminal branches and the lower power to detect real heterochromies in internal branches (Germain and Laurin 2009, table 7) provide a simpler explanation: the difference in heterochrony rate is an artefact of the analytical methods. Performing a continuous analysis on standardized time (or a proxy) should minimize such artifactual patterns (Germain and Laurin 2009, table 4).

Smith (2001 p, 173) argued that “[m]any developmental events are by nature dependent on a sequence of prior events as found with induction, signalling cascades, or expression of regulatory genes.” Although she presented this as a justification of concentrating analyses on developmental sequences rather than times, there are two problems with this argument. First, the fact that many studies (cited above) have used event pairing and have uncovered fairly labile evolutionary patterns shows that the type of data commonly analyzed by this technique is not strongly constrained by phenomena such as signalling cascades, to the extent that these cascades are thought to evolve very slowly. To take the example that first drew our attention to event pairing (cranial ossification sequences in urodeles), appearance of the premaxilla may not induce appearance of the vomer, although it precedes it chronologically in many taxa; the presence of reversals shows that any causal relationship that may exist in that order is not strongly constrained in evolution, although it may be within individual species. Of course, some events are constrained; for instance, somites cannot form before gastrulation has occurred, but such constrained events may not form the majority of data analyzed for heterochromies. Second, if the events of interest were constrained to stay in the same order, analyzing sequences would be uninteresting (no change could be detected), but analyzing developmental timing of events could reveal heterochromies.

Quantitative Methods in Evo-Devo

The suggestion to use quantitative methods such as phylogenetic independent contrasts and squared-change parsimony in evo-devo can be compared in some ways with the introduction of statistical methods in phylogenetic inference, such as ML and Bayesian methods. ML has generally not been used in previous developmental studies, with very few exceptions (Oakley et al. 2005; Guo et al. 2007). As Felsenstein (2001) nicely summarized, “Statistical inference of phylogenies almost didn’t happen” because of bitterly conflicting views (between evolutionary systematists, cladists, and numerical taxonomists) on the principles under which phylogeny should be inferred. We hope that our method, which uses well-known mathematical techniques, will be received more favorably, given the increasing prevalence of quantitative methods in evolutionary biology. Developmental biology may be lagging behind other biological fields in this respect, because, as Bininda-Emonds et al. (2003, p. 341) stated,
“comparative developmental biology is still predominantly a non-quantitative discipline”.

Continuous analysis, through its use of squared-change parsimony, relies on the assumption that small changes in timing of developmental events are more likely to occur than large changes of timing (or position in sequence). As Harrison and Larsson (2008, p. 380) noted, there is a debate about the relevance of the magnitude of changes in timing. Fortunately, recent tests of this hypothesis using 1 (Poe and Wake 2004) and 13 (Poe 2006) empirical data sets support the evolutionary model in which changes between adjacent positions in sequences are more frequent than changes between distant positions; this is what Poe and Wake (2004) called the AJ (adjacency) model. By contrast, other evolutionary models, called the “early conservation” (EC), “hourglass” (HG), and “adaptive penetration” (AP) models, that have been assumed (but generally not tested) in several developmental studies, were not supported by the analyses of Poe and Wake (2004) and Poe (2006). Because performance of evolutionary analytical techniques is often dependent on adequacy of the assumed evolutionary model (Martins et al. 2002), these studies provide some support for use of continuous analysis. On the contrary, the method of Harrison and Larsson (2008) and Parsimov (Jeffery et al. 2005) minimize only the number of changes, not their magnitude, and are thus less consistent with Poe’s (2006) findings about the evolution of developmental sequences.

Continuous analysis could be developed further to examine other evolutionary problems. A potential application concerns the identification of developmental or evolutionary modules of characters that can be detected through correlated evolutionary patterns (Schlosser 2001; Smith 2001). Correlation between the timing or position of characters could be investigated using phylogenetically independent contrasts (Felsenstein 1985), using standardized time or sequences. Guo et al. (2007) developed a ML-based method to do this with unstandardized time, and Oakley et al. (2005), with gene expression data. If the composition of the modules is not clearly defined prior to analysis, their composition could be investigated through multivariate statistics such as principal component analyses performed on independent contrasts. Continuous analysis has one important advantage over Smith’s (2001) heterochrony plots: it can use data from all taxa simultaneously to assess heterochronies, whereas heterochrony plots rely on comparison with a single outgroup to assess heterochronies between two other taxa. Thus, Smith (2001, fig. 6) had to make multiple two-by-two comparisons of taxa to demonstrate the likely coevolution of sets of developmental events in amphibians, without being able to assess the statistical significance of the pattern. This limitation of heterochrony plots and of a few other methods, was previously discussed by Bininda-Emonds et al. (2002). Poe’s (2004) test of modularity of sequences is more similar to our suggestion in relying on sister-group comparisons, but it uses sequences, rather than changes in the sequences and thus departs more strongly from the initial formulation of phylogenetic independent contrasts (Felsenstein 1985). Other methods have been proposed to detect correlated characters, although they do not necessarily use developmental data (e.g., Geeta 2003). Thus, continuous analysis could provide a simple efficient method to test for correlated evolution and modularity using developmental data.

Smith (2001) commented that “[d]evelopmental sequences have lost their place as a central component of the concept of heterochrony only over the past 20 years with the current, almost exclusive, focus on size and shape.” We agree that many types of data cannot be studied using allometry, but it is possible that part of the popularity of such studies (e.g., Kim et al. 2002; Berge and Penin 2004; Montes et al. 2007) in the last decades is that simple efficient statistical techniques were available (Jolicoeur 1963; Cubo et al. 2008), in addition to an explicit theoretical framework (e.g., Gould 1977). Continuous analysis brings some of this operational simplicity to the study of timing and sequences of events. It also offers a simple solution to what constitutes a significant difference (or change), a problem first encountered when event pairing was developed (Smith 1997, p. 1671) and never truly solved for event pairs. Thus, we hope that this method will prove useful in the field of evo-devo, at least as an alternative to event pairing.

**Supplementary Material**

Supplementary material, including online-only appendices, can be found at [http://www.sysbio.oxfordjournals.org/](http://www.sysbio.oxfordjournals.org/).

**Funding**

This work was supported by the CNRS (operating grant of UMR 7207) and the Ministry of Research of France.

**Acknowledgments**

Several comments by anonymous referees and by the associate editors T. Oakley and R. DeBry improved the draft.

**References**


Diniz-Filho J.A.F., de Sant’Ana C.E.R., Bini L.M. 1998. An eigenvec-


