MorphoCode: coding quantitative data for phylogenetic analysis

P. Schols*, C. D'hondt, K. Geuten, V. Merckx, S. Janssens, E. Smets

Laboratory of Plant Systematics. Institute of Botany and Microbiology. K. Arenberg 31, 3001 Leuven

* peter.schols@bio.kuleuven.ac.be

Received: 1 December 2003 - Accepted: 15 July 2004

Abstract

Coding quantitative data has always been one of the most difficult steps in morphological phylogenetics. A robust and theoretically well-supported method for coding quantitative data was proposed by Thiele (1993). This method relies on differential weighting of gaps between coded states within one character. The downside of gap weighting is that it requires many calculations. In order to provide an easy way of integrating quantitative data into phylogenetic analyses, we have created MorphoCode, an open-source software project. MorphoCode implements a variation on Thiele's method and offers the user the possibility to import data as tab-separated text, to choose the number of character states and to export the newly coded data to a NEXUS file. MorphoCode is available for free on the internet: http://www.kuleuven.ac.be/bio/sys/mc.

Introduction

Today, most phylogenies published are exclusively based on gene sequences. While we do not doubt that molecular data far eclipses morphological data when it comes to the construction of a robust phylogeny, a phylogenetic hypothesis that is exclusively based on molecular data is narrow. As shown by the recent discussion on DNA taxonomy (Tautz et al. 2003), a classification system not supported by morphological synapomorphies will have difficulties to become widely accepted (Lipscomb et al. 2003). Revealing monophyletic groups is one of the major aims of systematics, but finding morphological characters supporting those groups is equally important.

Another advantage of integrating morphology into phylogenetic analyses is the possibility to trace character evolution. For the first time in history, the evolution of morphological features of organisms can be reconstructed confidently because of the power of molecular data. This will dramatically alter our understanding of character evolution, and will complement recent advancements in Evo-Devo (Hawkins 2002). When working with fossil organisms, taxonomists can not rely on DNA sequences to reconstruct phylogeny, making morphological and morphometric characters the only source of information (Gabunia et al. 2000).

However, coding morphological features of organisms to reconstruct their phylogeny is a controversial issue. In contrast to analyses of molecular data, in which the definition of characters and character states is virtually automatic, morphological character analysis requires interpretation and taking methodological decisions (Wiens 2001). Most morphological characters are actually quantitative data. In order to circumvent coding problems, taxonomists tend to describe these quantitative data with a vocabulary that hides the quantitative nature of the data (e.g. pollen surface reticulate). When no terminology exists, characters are often divided in arbitrary states (e.g. pollen larger than 50 µm vs. pollen smaller than 50 µm). These coding methods are rather difficult to reproduce and they ignore a lot of information contained within the data.

Gap coding

Most alternative solutions for this problem are gap-coding techniques. They are based on the distributions of mean values for a certain character. In general, these methods rank the taxa along a scaled attribute axis into states. They differ mainly in the degree to which they will divide the attribute axis at points where the gaps between means exceed a predefined value (Chapill 1989). While these methods are more reproducible, they are sampling dependent: removing one taxon from the sampling could alter the coding of the entire character. Also, some gaps are retained while others are ignored, resulting in loss of information.

Gap weighting

A theoretically well supported method for coding morphometric data was proposed by Thiele (1993). This method, called gap weighting, considers the order and distribution of the means for a certain character and converts them to ordered, multistate characters where the distance between the means is represented by the distance between the ordered character states in the matrix. In this method, the following formula is applied to every character state of every character in the matrix:

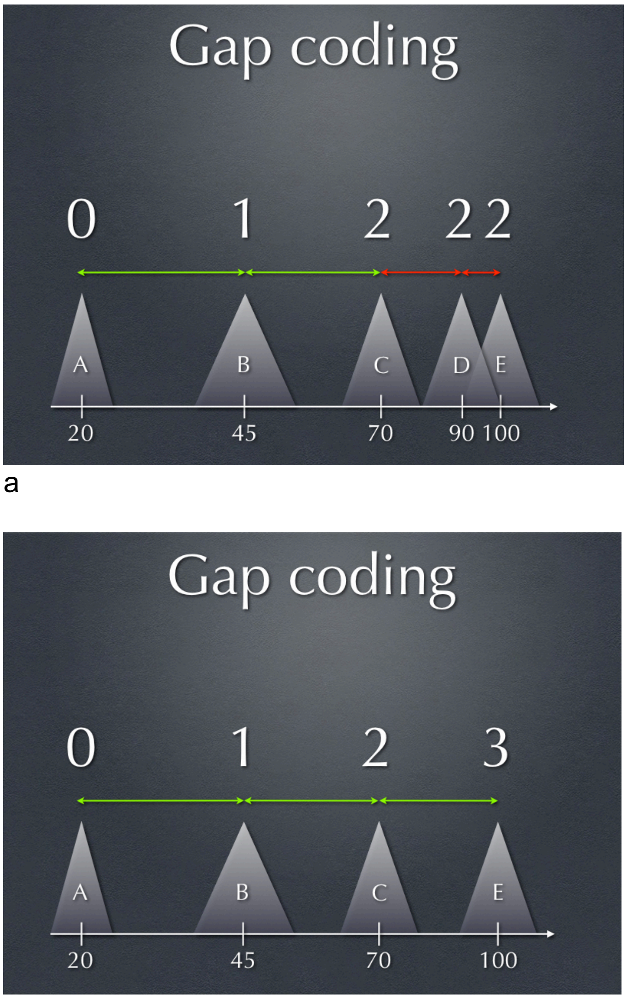

Fig. 1a. A continuous character (e.g. pollen size) with an average value ranging from 20 to 100 µm for species A-E. When we plot the average values for those five species on an axis and apply a gap coding technique, we create a new character state if the gap between two subsequent values is larger than a certain value (e.g 24 µm). The difference between taxon A and taxon B is 25 µm, so taxon A gets character state 0 and taxon B is coded as 1. The difference between taxa B and C is 25 µm as well so the average value for C is coded as 2. The difference between C and D is only 20 µm and therefore D is coded as 2. D and E are only 10 µm apart so taxon E gets character state 2 as well.

Fig. 1b shows that this kind of coding is very sampling dependent: if we omit taxon D from our sampling, taxon E is coded as 3, in stead of 2.

$x_{new} = n * [(x - min) / (max - min)]$

Max and min are the maximum and minimum mean value of the character across all species and x is the mean value of the current taxon and n is the number of allowed character states (n is limited to 10 in Nona and 32 in PAUP*, see Thiele 1993).

The downside of this method is that it requires many calculations: for a modest matrix of 35 taxa and 40 characters, Thiele's formula needs to be applied to all 1.400 character states. This might be the main reason why this method, although published in 1993, has only been used in a few papers throughout the taxonomic literature (Wiens 2001, Chu 2002).

To provide an easier way of applying the gap-weighting method to quantitative data, we have created MorphoCode, an open-source software project. MorphoCode implements Thiele's method and offers the user the possibility to import data as tab-separated text, choose the number of character states, perform all necessary calculations and export the newly coded data to a NEXUS file (Maddison et al. 1997). This NEXUS file can be combined with molecular data to run a combined molecular / morphological analysis. MorphoCode and its source are available for free on the MorphoCode website: http://www.kuleuven.ac.be/bio/sys/mc under the GNU GPL.

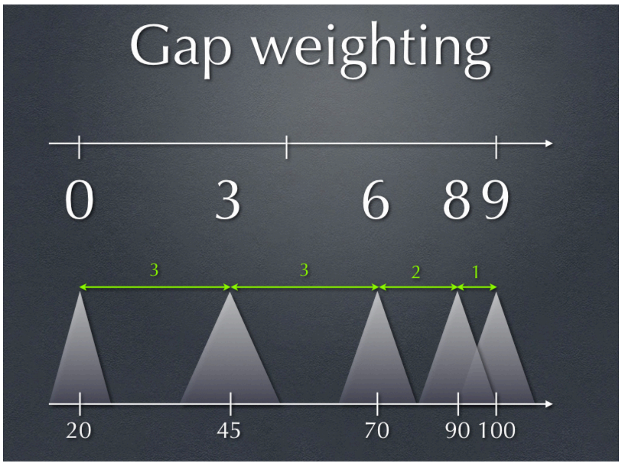

Fig. 2. With gap weighting, the range of values, (e.g. 20-100 µm) is standardized by converting the average values into values from 0-9. This means that the smallest mean value in the sampling (20 µm) is coded as 0 and the largest mean value (100 µm) is coded as 9. All other states are coded proportionally as well and will have values ranging from 0 to 9. Because the character is treated as ordered, the number of steps between two taxa is proportional to the distance between the average values of those species for any given character. Adding or removing taxa does not affect the coding.

ACKNOWLEDGEMENTS

PS holds a research grant from the IWT (n° 991293). This research is supported by a grant from the Research Fund of the K.U.Leuven (OT/01/25).

REFERENCES

Chappill J.A. 1989. Quantitative characters in phylogenetic analysis. *Cladistics* **5**: 217-234.

Chu P. 2002. A Morphological Test of the Monophyly of the Cardueline Finches (Aves: Fringillidae, Carduelinae). *Cladistics* **18**: 279-312.

Gabunia L., Vekua A., Lordkipanidze D., Swisher C.C., Ferring R., Justus A., Nioradze M., Tvalchrelidze M., Antón S.C., Bosinski G., Jöris O., Lumley M.A., Majsuradze G. & Mouskhelishvili A. 2000. Earliest Pleistocene Hominid Cranial Remains from Dmanisi, Republic of Georgia: Taxonomy, Geological Setting, and Age. *Science* **288**: 1019-1025

Hawkins J. 2002. Evolutionary developmental biology: impact on systematic theory and practice and the contribution of systematics. In: Cronk, Q.C.B., Bateman, R.M. and Hawkins, J.A. eds. *Developmental Genetics and Plant Evolution*, 32-51. Taylor and Francis, London.

Lipscomb D., Platnick N. & Wheeler, P. 2003. The intellectual content of Taxonomy: a comment on DNA taxonomy. *Trends Ecol. & Evol.* **18**: 65-66.

Maddison D.R., Swofford D.L. & Maddison W.P. 1997. NEXUS: an extendible file format for systematic information. *Syst. Biol.* **46**: 590-621.

Tautz D., Arctander P., Minelli A., Thomas R. & Vogler A. 2003. A plea for DNA taxonomy. *Trends Ecol. & Evol.* **18**: 70-74.

Thiele K. 1993. The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics* **9**: 275-304.

Wiens J. 2001. Character analysis in morphological phylogenetics: problems and solutions. *Syst. Biol.* **50**: 689-699.