SHAPE AND PHYLOGENY

A thesis submitted to The University of Manchester for the degree of

DOCTOR OF PHILOSOPHY in the Faculty of Life Sciences

2014

CEFERINO VARÓN GONZÁLEZ

LIST OF CONTENTS

Final word count: 38143

LIST OF FIGURES

Chapter 2:

Chapter 3:

LIST OF TABLES

Chapter 3:

Chapter 4:

ABSTRACT

Institution: The University of Manchester
Name: Ceferino Varón González
Degree Title: PhD Evolutionary Biology
Thesis Title: Shape and phylogeny
Date: 2014

Geometric morphometrics, the science about the study of shape, has developed much in the last twenty years. In this thesis I first study the reliability of the phylogenies built using geometric morphometrics. The effect of different evolutionary models, branch-length combinations, dimensionality and degrees of integration is explored using computer simulations. Unfortunately in the most common situations (presence of stabilizing selection, short distance between internal nodes and presence of integration) the reliability of the phylogenies is very low. Different empirical studies are analysed to estimate the degree of evolutionary integration usually found in nature. This gives an idea about how powerful the effect of integration is over the reliability of the phylogenies in empirical studies. Evolutionary integration is studied looking at the decrease of variance in the principal components of the tangent shape space using the independent contrasts of shape. The results suggest that empirical data usually show strong degrees of integration in most of the organisms and structures analysed. These are bad news, since strong degree of integration has devastating effects over the phylogenetic reliability, as suggested by our simulations. However, we also propose the existence of other theoretical situations in which strong integration may not translate into convergence between species, like perpendicular orientation of the integration patterns or big total variance relative to the distance between species in the shape space. Finally, geometric morphometrics is applied to the study of the evolution of shape in proteins. There are reasons to think that, because of their modular nature and huge dimensionality, proteins may show different patterns of evolutionary integration. Unfortunately, proteins also show strong functional demands, which influence their evolution and that cause strong integration patterns. Integration is then confirmed as a widespread property in the evolution of shape, which causes poor phylogenetic estimates.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Alternative format justification and contribution of the authors

This thesis is written under the alternative format. The submission in the alternative format avoided the rewriting of my work and therefore it saved time that could be destined to extend and to improve the quality of the research. It also helped to practice the scientific writing and my career perspectives, since we were able to submit the first chapter to a journal before the ending of the PhD.

The first chapter, 'Reliability of Estimating Phylogenies from Shape and Similar Multidimensional Data', is currently under review in Systematic Biology. Chris Klingenberg has provided the original idea. I have implemented the methods and contributed in the interpretation of the results. Simon Whelan has contributed with some essential tips.

In the second chapter, 'Patterns of Integration in Geometric Morphometrics Studies and Their Role in Phylogenetics', Chris Klingenberg has provided the original idea. I have implemented the methods and contributed in the interpretation of the results. Simon Whelan has contributed with some of the methods. It is going to be submitted to Journal of Evolutionary Biology.

The third chapter, 'Geometric Morphometrics: A Useful Tool to Study Protein Evolution', is going to be submitted to Plos One. In it, Chris Klingenberg has provided the original idea and helped with the interpretation of the results. Simon Lovell has contributed with essential advices about the nature of the data used and its implications on the results. I have implemented the methods and contributed with the interpretation of the results. Simon Whelan has contributed with some essential tips.

ACKNOWLEDGEMENTS AND DEDICATION

Now that I am finishing my PhD I would like to start thanking the people who recommended me for this position more than three years ago. José María Gómez Reyes and Juan Pedro Martínez Camacho are not only the lecturers that introduced me to evolution and made me consider a career in science. They are also the people who trusted in me for this challenge and the effort devoted to this thesis is my way of thanking them for that.

I would like to especially thank Chris Klingenberg for his teaching and his helpful advice beyond the topics mentioned in this thesis. Thanks also to my advisors Simon Whelan and Simon Lovell, as well as my thesis examiners Rob Samson and Leandro Monteiro for their valuable comments and suggestions for different chapters of this thesis.

The work explained from here on also reflects fruitful discussions with people with whom I have shared daily work in the lab. Some of them are Eli Mora, Raffaella Bravi, Marta Linde and Hugo Benítez. I am in special debt to Elis Damasceno for the daily discussions, jokes and comments that have enriched my work and my stay in England.

Finally, I would like to thank many other people who have given me their support and that have made my life better during this time. I am also especially grateful to the members of the University of Manchester Handball Club.

Esta tesis se la dedico a toda la gente que desde España me ha apoyado durante estos tres años. En especial a mis padres y mi hermano, por ser mi red salvavidas particular en este ejercicio de equilibrismo. A Hannah RF y Juanjo PA, por ser mi tiento casi diariamente. A Julia, por hacerme mirar siempre al frente.

The author

Ceferino Varón González (Almería, Spain, 1987) is a graduate student at the University of Manchester. He obtained his bachelor in Biology at the Universidad de Granada (Spain) in July 2011, finishing with a GPA of 1.797 out of 4 (7.0/10). Although the elaboration of scientific reviews was a frequent task during the degree, no formal research was required to obtain the degree. He had a remarkable performance in the subjects directly related to evolution: first class honours in Biological Anthropology (Prof. M. C. Botella) and second-class honours, upper division, in both Population Genetics and Evolution (Prof. J. P. M. Camacho) and Theoretical and Evolutionary Ecology (Dr. J. M. Gómez Reyes).

In September 2011, he registered into an MPhil program in Evolutionary Biology at the University of Manchester, under the supervision of Prof. Chris Klingenberg. During the first months of his MPhil he learnt the basics of computer programming and geometric morphometrics. At the end of that academic year Ceferino, Dr. Simon Whelan and Prof. Klingenberg agreed to switch his MPhil candidacy status into a PhD candidacy with the approval of the transfer report examiners: Dr. Casey Bergman and Dr. William Sellers.

This thesis is the result of that first year as a master student plus two more years as a PhD student, which constitutes the whole scientific production of Ceferino.

1. INTRODUCTION

Geometric morphometrics is the science that studies shape. It has been used in a wide variety of contexts, of which evolution is the main one. It is a science under continuous development (Adams, et al. 2013) and there are many areas within evolutionary biology in which great improvements have been achieved in the last few years, e.g. the application of geometric morphometrics to the study of plant evolution (Gómez, et al. 2006) or evo-devo (Lawing and Polly 2010, Salazar-Ciudad and Marín-Riera 2013). Macroevolution has been one of the areas in which more interest and development has been lately (Klingenberg and Gidaszewski 2010, Catalano and Goloboff 2012, Klingenberg and Marugán-Lobón 2013, Monteiro 2013, Adams 2014c, Adams 2014b, Adams 2014a). In this thesis, I study geometric morphometrics in this context, first studying the relationship between shape and phylogenetics and then developing a morphometric method to study macroevolution in proteins, in which the problems found to estimate phylogenies in anatomical structures might not apply or be less severe.

1.1. Geometric morphometrics and Phylogenetics

The utility of geometric morphometrics to build reliable phylogenies has been a controversial topic since twenty years ago, when the study of shape restarted to develop and there were the first attempts to study shape using geometric morphometrics in macroevolution (Bookstein 1994, Zelditch, et al. 1995, Naylor 1996, Rohlf 1998, Swiderski, et al. 1998, Zelditch, et al. 1998, Monteiro 2000). Some authors proposed that geometric morphometrics can be useful to build phylogenies due to its multidimensional nature (González-José, et al. 2008). The fact that there is so much data available to describe the shape of each specimen supposes that there are 'many sources of variation' (MacLeod and Forey 2002) and therefore convergence should be less likely (chapter 1, no integration case). The great ability to discriminate

between different specimens using geometric morphometrics in empirical studies has reinforced that idea. In contrast to the one-dimensional data, which makes the reliability of the phylogenies to be very low (Lynch 1989), shape data is supposed to perform better.

Under that logic, different studies have been published trying to find the appropriate methods to build reliable phylogenies. The first attempts to infer phylogenies using shape characters used the partial warp scores (Fink and Zelditch 1995, Zelditch, et al. 1995, Swiderski, et al. 1998, Zelditch, et al. 1998, Clouse, et al. 2011) or principal component scores (MacLeod 2002, González-José, et al. 2008, Aguilar-Medrano, et al. 2011, González-José, et al. 2011) as cladistic characters. However, these procedures were criticized for the lack of biological meaning in the case of the former and their decomposition of the different dimensions in different characters in both cases (Bookstein 1994, Naylor 1996, Adams and Rosenberg 1998, Rohlf 1998, Monteiro 2000, Adams, et al. 2011, Zelditch, et al. 2012).

Shape data is characterized by some covariation (integration) between shape variables (Klingenberg 2013). Therefore, shape must not be decomposed into separate characters but treated as a multidimensional and multivariate character (Klingenberg and Gidaszewski 2010). This fact was taken into account by some other studies, in which clustering techniques or statistical approaches are used (Marcus, et al. 2000, Cannon and Manos 2001, Lockwood, et al. 2004, Caumul and Polly 2005, Couette, et al. 2005, Macholán 2006, Cardini and Elton 2008). However, the success of these methods in reconstructing phylogenies has been difficult to assess. In the first place, the results in the empirical studies are a consequence of specific evolutionary pathways that are unknown and therefore it is difficult to know the evolutionary diversification necessary in the organisms for these phylogenies to be reliable. In second place, most of the phylogenies built using shape data show partial agreement with other reliable sources (e.g.

molecular data), so the reliability depends at some extent on the expectations of the researcher. That explains the publication of studies in which the phylogenies are not completely correct but the authors are optimistic about the possibilities of shape data in this regard (Catalano and Goloboff 2012, Smith and Hendricks 2013).

Integration produces a concentration of the variation in specific directions of the shape space (Wagner 1984) and therefore it can substantially decrease the sources of independent variation and promote convergence (Goswami, et al. 2014). In the extreme, integration would transform shape into univariate data, where the phylogenetic estimates have been shown to be unreliable (Lynch 1989). That fact, together with some bad results obtained in empirical data (Klingenberg and Gidaszewski 2010), it has added in a good part of the morphometrician community a cautious attitude towards the use of shape to build phylogenies. As Joe Felsenstein pointed out in the morphomet emailing list (4th Dec 2013), 'For GM [Geometric morphometrics] coordinates or in fact for any quantitative traits, it is not appropiate to use them for inferring phylogenies unless you have some way of dealing with the covariances'. However, it has not been quantified how much integration affects the phylogenetic reconstruction in practice or whether the methods prepared to deal with covariances (Felsenstein 2002), beyond the specific requirements in terms of dimensionality and species sample size, significantly improve the reliability.

1.2. Integration and convergence

In order to have a complete knowledge about the reliability of the phylogenies built using geometric morphometrics it is not enough to know how different degrees of integration affect phylogenetic accuracy. It is also important to know whether these conditions are met in empirical studies, so

the patterns of integration that are usually found in morphometric studies can be compared with the results found for these patterns in our simulations.

Integration is a widespread feature in morphometric data (Olson and Miller 1958, Cheverud 1996, Wagner and Altenberg 1996, Marroig and Cheverud 2001, Bookstein, et al. 2003, Monteiro, et al. 2005, Young and Badyaev 2006, Lockwood 2007, Klingenberg 2008, Hallgrímsson, et al. 2009, Gonzalez, et al. 2011, Klingenberg 2013, Armbruster, et al. 2014, Goswami, et al. 2014), which has been found in different organisms from mammals to plants. Indeed, different mechanisms promoting integration are present in the vast majority of the organisms (Klingenberg 2014). They all produce coordinated responses in different parts of the structures, which translates into patterns of phenotypic integration. These patterns seem to be different depending on the organisms (Jamniczky and Hallgrímsson 2009, Gómez, et al. 2014, Goswami, et al. 2014) but the variance of the eigenvalues of the covariance matrix of shape data (Wagner 1984, Young 2006) is a common consequence of all of them. The variance of the eigenvalues is a way of looking at the concentration of variation in specific directions of the shape space during the evolution of the species. Comparing this measure obtained from empirical data with the strength of integration set in our simulations we can assess whether the conditions in nature are favourable for the phylogenetic reconstruction using geometric morphometrics or not.

Homoplasy is behind the mistakes in the phylogenetic reconstruction, since it reflects similarity in shape between unrelated species. However, the importance of homoplasy goes beyond that and it is important to understand the evolution of the organisms (Losos 2011). Although integration is not essential for the production of convergence, it can promote it significantly (Losos 2011, Goswami, et al. 2014). The combination of different patterns of integration with different orientations of the major directions of variation and

different total variances can result in interesting conclusions to understand the evolution of species and the production of convergence.

1.3. Proteins and integration

Integration has been found in almost all the morphological structures in which it has been studied using geometric morphometrics. However, in molecules like proteins, it has been only been studied from the perspective of the sequence (Trifonov and Frenkel 2009). Just as anatomical structures, proteins can have strong functional demands. Functional integration, the 'association between parts interacting in some functional context' (p. 4) (Klingenberg 2014) is a feature that should be present in all the proteins, since the functionality is determined by the correct interaction in first place with the chaperons (Hartl and Hayer-Hartl 2002), which are responsible for the correct protein folding. In all the enzymes, in addition, their function is conditioned by the correct interaction between different parts of the proteins and different parts from different proteins.

Unlike the anatomical structures, where integration has been proposed to be the ancestral state, evolving towards modularity with time (Wagner and Altenberg 1996, Goswami, et al. 2014), the evolution of the proteins suggests modularity would be in principle the ancestral state. New proteins can appear after duplication of preexisting proteins (Lynch 2000). At early stages after that, some parts in the new proteins duplicate forming new modules (Halabi, et al. 2009). At that point, the modularity in the protein would be very strong (Trifonov and Frenkel 2009). Then, recombination between the new and the pre-existing modules happens so the protein would gain new functions (Hopfner, et al. 1998, Lynch 2000). Integration is then a feature that would appear later in the evolution.

The study of integration in proteins can give a new perspective over the differences in the evolution of anatomical and protein structures. The use

of geometric morphometrics in this regard would be stimulating for both fields, structural biology and geometric morphometrics, since it will give new perspectives about the study of shape to both sciences.

1.4. Summary and aims

The first chapter of this thesis is a theoretical approach to the relationship between shape and phylogenetic inference, taking into consideration different features of the morphometric data that can affect the reliability of the phylogenies built using geometric morphometrics. Using computer simulations is a good start, because they allow us exploring a wide range of values in the factors that may affect the ability of geometric morphometrics to reconstruct correct phylogenies.

The next chapters study one of these factors specifically, integration, over the evolution of shape in a phylogenetic context. Since the first chapter is based on computer simulations, in the next two we explore empirical data, so we can check how widespread integration is in nature and how strong it is. Two different kinds of empirical data are studied, so our study is fairly extensive and strong evidence about the nature of integration is obtained. Empirical data from anatomical studies, that is the kind of data managed in geometric morphometrics, is studied in first place (second chapter). Along with it, we also assess the relationship between different features of the integration patterns and convergence, which contributes to errors during the process of phylogenetic inference studied before.

In the last chapter of this thesis, geometric morphometrics is applied to the study of protein shape. Integration and convergence are then studied under the same perspective that has been applied to anatomical structures to a new set of structures, proteins. Using the same technique will make a fair comparison between integration and convergence found in morphology and integration and convergence found in proteins. Therefore, we extend the

study of these features to the molecular world, where different effects and principles may apply. Molecular data may reveal new and exciting features regarding the relationship between integration, shape and phylogenetics.

1.5. References

Adams DC. 2014a. A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. Systematic Biology.

Adams DC. 2014b. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. Evolution.

Adams DC. 2014c. Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. Systematic Biology 63:166-177.

Adams DC, Cardini A, Monteiro LR, O'Higgins P, Rohlf FJ. 2011. Morphometrics and phylogenetics: Principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. Journal of Human Evolution 60:240-243.

Adams DC, Rohlf FJ, Slice DE. 2013. A field comes of age: geometric morphometrics in the 21st century. Hystrix 24:7-14.

Adams DC, Rosenberg MS. 1998. Partial warps, phylogeny, and ontogeny: A comment on Fink and Zelditch (1995). Systematic Biology 47:168-173.

Aguilar-Medrano R, Frédérich B, de Luna E, Balart EF. 2011. Patterns of morphological evolution of the cephalic region in damselfishes (Perciformes: Pomacentridae) of the Eastern Pacific. Biological Journal of the Linnean Society 102:593–613.

Armbruster WS, Pélabon C, Bolstad GH, Hansen TF. 2014. Integrated phenotypes: understanding trait covariation in plants and animals.

Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Bookstein F. 1994. Can biometrical shape be a homologous character? In: Hall BK editor. Homology: The hierarchial basis of comparative biology. New York, Academic Press, p. 197-227.

Bookstein FL, Gunz P, Mitteroecker P, Prossinger H, Schaefer K, Seidler H. 2003. Cranial integration in Homo: singular warps analysis of the midsagittal plane in ontogeny and evolution. Journal of Human Evolution 44:167-187.

Cannon CH, Manos PS. 2001. Combining and comparing morphometric shape descriptors with a molecular phylogeny: the case of fruit type evolution in Bornean *Lithocarpus* (Fagaceae). Systematic Biology 50:860–880.

Cardini A, Elton S. 2008. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. Biological Journal of the Linnean Society 93:813-834.

Catalano SA, Goloboff PA. 2012. Simultaneously mapping and superimposing landmark configurations with parsimony as optimality criterion. Systematic Biology 61:392-400.

Caumul R, Polly PD. 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (*Marmota*, Rodentia). Evolution 59:2460–2472.

Cheverud JM. 1996. Developmental integration and the evolution of pleiotropy. American Zoologist 36:44-50.

Clouse RM, de Bivort BL, Giribet G. 2011. Phylogenetic signal in morphometric data. Cladistics 27:337–340.

Couette S, Escarguel G, Montuire S. 2005. Constructing, bootstrapping, and comparing morphometric and phylogenetic trees: A case study of New World monkeys (Platyrrhini, Primates). Journal of Mammalogy 86:773–781.

Felsenstein J. 2002. Quantitative characters, phylogenies, and morphometrics. In: MacLeod N, Forey PL editors. Morphology, Shape & Phylogeny. London, Taylor & Francis, p. 27-44.

Fink WL, Zelditch ML. 1995. Phylogenetic analysis of ontogenetic shape transformations: A reassessment of the piranha Genus Pygocentrus (Teleostei). Systematic Biology 44:343-360.

Gómez JM, Perfectti F, Camacho JPM. 2006. Natural selection on *Erysimum mediohispanicum* flower shape: Insights into the evolution of zygomorphy. American Naturalist 168:531-545.

Gómez JM, Perfectti F, Klingenberg CP. 2014. The role of pollinator diversity in the evolution of corolla-shape integration in a pollination-generalist plant clade. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Gonzalez PN, Oyhenart EE, Hallgrímsson B. 2011. Effects of environmental perturbations during postnatal development on the phenotypic integration of the skull. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 316B:547-561.

González-José R, Escapa I, Neves WA, Cúneo R, Pucciarelli HM. 2008. Cladistic analysis of continuous modularized traits provides phylogenetic signals in Homo evolution. Nature 453:775-778.

González-José R, Escapa I, Neves WA, Cúneo R, Pucciarelli HM. 2011. Morphometric variables can be analyzed using cladistic methods: A Reply to Adams et al. Journal of Human Evolution 60:244-245.

Goswami A, Smaers JB, Soligo C, Polly PD. 2014. The macroevolutionary consequences of phenotypic integration: from development to deep time. Philosophical Transactions of the Royal Society of London B Biological Sciences 369:20130254.

Halabi N, Rivoire O, Leibler S, Ranganathan R. 2009. Protein sectors: Evolutionary units of three-dimensional structure. Cell 138:774-786.

Hallgrímsson B, Jamniczky HA, Young NM, Rolian C, Parsons TE, Boughner JC, Marcucio RS. 2009. Deciphering the palimpsest: Studuying the relationship between morphological integration and phenotypic covariation. Evolutionary Biology 36:355-376.

Hartl FU, Hayer-Hartl M. 2002. Molecular chaperons in the cytosol: from nascent chain to folded protein. Science 295:1852-1858.

Hopfner KP, Kopetzki E, Kresse GB, Bode W, Huber R, Engh RA. 1998. New enzyme lineages by subdomain shuffling. Proceedings of the National Academy of Sciences of the United States of America 95:9813-9818.

Jamniczky HA, Hallgrímsson B. 2009. A comparison of covariance structure in wild and laboratory muroid crania. Evolution 63:1540–1556.

Klingenberg CP. 2008. Morphological integration and developmental modularity. Annual Review of Ecology, Evolution and Systematics 39:115-132.

Klingenberg CP. 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. Hystrix 24:43-58.

Klingenberg CP. 2014. Studying morphological integration and modularity at multiple levels: concepts and analysis. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Klingenberg CP, Gidaszewski NA. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Systematic Biology 59:245-261.

Klingenberg CP, Marugán-Lobón J. 2013. Evolutionary covariation in geometric morphometric data: analyzing integration, modularity, and allometry in a phylogenetic context. Systematic Biology 62:591–610.

Lawing AM, Polly PD. 2010. Geometric morphometrics: recent applications to the study of evolution and development. Journal of Zoology 280:1-7.

Lockwood CA. 2007. Adaptation and functional integration in primate phylogenetics. Journal of Human Evolution 52:490-503.

Lockwood CA, Kimbel WH, Lynch JM. 2004. Morphometrics and hominoid phylogeny: Support for a chimpanzee-human clade and differentiation among great ape subspecies. Proceedings of the National Academy of Sciences of the United States of America 101:4356-4360.

Losos JB. 2011. Convergence, adaptation, and constraint. Evolution 65:1827-1840.

Lynch M. 1989. Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation. Evolution 43:1-17.

Lynch M. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

Macholán M. 2006. A geometric morphometric analysis of the shape of the first upper molar in mice of the genus *Mus* (Muridae, Rodentia). Journal of Zoology (London) 270:672–681.

MacLeod N. 2002. Phylogenetic signals in morphometric data. In: MacLeod N, Forey PL editors. Morphology, shape and phylogeny. London, Taylor & Francis, p. 100–138.

MacLeod N, Forey PL. 2002. Introduction: morphology, shape and phylogenetics. In: MacLeod N, Forey PL editors. Morphology, Shape and Phylogeny. London, CRC Press.

Marcus LF, Hingst-Zaher E, Zaher H. 2000. Application of landmark morphometrics to skulls representing the orders of living mammals. Hystrix 11:27–47.

Marroig G, Cheverud JM. 2001. A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of New World monkeys. Evolution 55:2576-2600.

Monteiro LR. 2000. Why morphometrics is special: The problem with using partial warps as characters for phylogenetic inference. Systematic Biology 49:796-800.

Monteiro LR. 2013. Morphometrics and the comparative method: studying the evolution of biological shape. Hystrix.

Monteiro LR, Bonato V, dos Reis SF. 2005. Evolutionary integration and morphological diversification in complex morphological structures: Mandible shape divergence in spiny rats (Rodentia, Echimyidae). Evolution & Development 7:429–439.

Naylor GJP. 1996. Can partial warp scores be used as cladistic characters? In: Marcus LF, Corti M, Loy A, Naylor GJP, Slice DE editors. Advances in morphometrics. New York, Plenum Press, p. 519–530.

Olson EC, Miller RL. 1958. Morphological integration. Chicago, University of Chicago Press.

Rohlf FJ. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny. Systematic Biology 47:147-158.

Salazar-Ciudad I, Marín-Riera M. 2013. Adaptive dynamics under development-based genotype-phenotype maps. Nature 497:361-364.

Smith UE, Hendricks JR. 2013. Geometric morphometrics character suites as phylogenetic data: Extracting phylogenetic signal from gastropod shells. Systematic Biology 62:366-385.

Swiderski DL, Zelditch ML, Fink WL. 1998. Why morphometrics is not special: Coding quantitative data for phylogenetic analysis. Systematic Biology 47:508-519.

Trifonov EN, Frenkel ZM. 2009. Evolution of protein modularity. Current Opinion in Structural Biology 19:335-340.

Wagner GP. 1984. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: evidence for a nonrandom organization of

quantitative character variation. Journal of Mathematical Biology 21:77-95.

Wagner GP, Altenberg L. 1996. Complex adaptations and the evolution of evolvability. Evolution 50:967-976.

Young NM. 2006. Function, ontogeny, and canalization of shape variance in the primate scapula. Journal of Anatomy 209:623-636.

Young RL, Badyaev AV. 2006. Evolutionary persistence of phenotypic integration: Influence of developmental and functional relationships on complex trait evolution. Evolution 60:1291-1299.

Zelditch ML, Fink WL, Swiderski DL. 1995. Morphometrics, homology, and phylognetics: Quantified characters as synapomorphies. Systematic Biology 44:179-189.

Zelditch ML, Fink WL, Swiderski DL, Lundrigan BL. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny: A reply to Rohlf. Systematic Biology 47:159-167.

Zelditch ML, Swiderski DL, Sheets HD. 2012. Geometric morphometrics for biologists: a primer. San Diego, Academic Press.

Chapter 2

Reliability of Estimating Phylogenies from Shape and Similar Multidimensional Data

CEFERINO VARÓN GONZÁLEZ [1], SIMON WHELAN[1,2]

AND CHRISTIAN PETER KLINGENBERG[1]

[1]*Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, United Kingdom;*

[2]*Dept. of Evolutionary Biology, EBC, Uppsala University, Norbyägen 18D, 75236 Uppsala, Sweden*

Abstract

In recent years, there has been some controversy whether multidimensional data such as geometric morphometric data on shape or information on gene expression can be used for estimating phylogenies. This study addresses this question with simulations of evolution in multidimensional phenotype spaces. The simulations use the four-taxon case, where there are just three possible tree topologies, and a comprehensive scheme to cover different combinations of branch lengths. Under an evolutionary model of isotropic Brownian motion, phylogeny can be estimated reliably if dimensionality is high. If phenotypic variation is integrated so that most variation is concentrated in one or a few dimensions, the reliability of phylogenetic estimates is seriously reduced, corresponding to the reduced dimensionality of the data. Taking into account integration by using Mahalanobis distance in estimating phylogenies can restore phylogenetic reliability only in part, depending on the sample size used to estimate patterns of integration. Evolutionary models with stabilizing selection produce unreliable estimates, which are little better than picking a phylogenetic tree at random. Finally, a best-case scenario with an evolutionary model of mutual repulsion among taxa produces reliable estimates of phylogeny, but cannot be considered a realistic model of evolution in natural clades. Overall, the simulations suggest that multidimensional data, under plausible evolutionary models, do not produce reliable estimates of phylogeny.

2.1. INTRODUCTION

Quantitative multidimensional data are often good descriptors of biological structures or functions and it is therefore sensible to examine whether they can be used as sources of information for phylogenetics. Some early studies that pioneered phylogenetics were based on considerations of multidimensional spaces of allele frequencies for multiple loci (Cavalli-Sforza and Edwards 1967) and several more recent studies have estimated phylogenetic trees from data on gene expression (Enard et al. 2002; Rifkin et al. 2003; Uddin et al. 2004; Brawand et al. 2011), but most such analyses have used morphometric data on the shapes of organisms or their parts (Fink and Zelditch 1995; Zelditch et al. 1995; Naylor 1996; Swiderski et al. 1998; Zelditch et al. 1998; MacLeod 2002; Lockwood et al. 2004; Polly 2004; González-José et al. 2008; Catalano et al. 2010; Aguilar-Medrano et al. 2011; Goloboff and Catalano 2011; Catalano and Goloboff 2012; Smith and Hendricks 2013; Watanabe and Slice 2014). It remains contentious, however, whether the phylogenies estimated from quantitative multidimensional variables are reliable.

During the last two decades, a number of different aspects concerning this issue have been discussed. Some studies have suggested that partial warp scores (Fink and Zelditch 1995; Zelditch et al. 1995; Swiderski et al. 1998; Zelditch et al. 1998; Clouse et al. 2011) or principal component scores (MacLeod 2002; González-José et al. 2008; Aguilar-Medrano et al. 2011; González-José et al. 2011) can be used as cladistics characters. These proposals, however, have been criticized for various reasons, especially concerning the decomposition of phenotypic spaces into distinct characters (Bookstein 1994; Naylor 1996; Adams and Rosenberg 1998; Rohlf 1998; Monteiro 2000; Zelditch et al. 2004; Adams et al. 2011). Some authors have advocated methods, different in many respects from the Procrustes approach

used almost ubiquitously in geometric morphometrics, that use landmarks as characters in cladistic analysis (Catalano et al. 2010; Goloboff and Catalano 2011; Catalano and Goloboff 2012). An alternative is to use methods that do not divide the phenotypic variation into characters, but infer phylogenies from distances among taxa using clustering techniques (e.g., neighbor joining or UPGMA) or using statistical approaches such as maximum likelihood (e.g., Marcus et al. 2000; Cannon and Manos 2001; Lockwood et al. 2004; Caumul and Polly 2005; Couette et al. 2005; Macholán 2006; Cardini and Elton 2008). Such estimates of phylogenies may be unreliable, however, because theoretical studies and computer simulations have demonstrated that random evolutionary processes such as Brownian motion frequently produce convergence, so that phenotypic distance may not be a good indicator of time since divergence (Lynch 1989; Stayton 2008).

These debates raise the question of how the quality of estimated trees can be assessed. So far, the majority of such assessments have compared trees obtained from morphometric data to trees hypothesized to reflect true phylogenetic history (Cole et al. 2002; Lockwood et al. 2004; Cardini and Elton 2008; González-José et al. 2008; Klingenberg and Gidaszewski 2010; Catalano and Goloboff 2012). Many of these studies produced partial agreement in the trees, which is somewhat ambiguous: adherents of a particular method can emphasize that the trees are partly correct, critics can point out that other aspects are wrong. Similarly, Smith and Hendricks (2013, p. 377) "consider it impressive" that morphometric characters were able to allocate 33–45% of taxa successfully to their positions in a phylogenetic tree, whereas skeptics might argue that this indicates a clear majority of failures. A possible solution for this problem is to use computer simulations of simple trees, as it has been used for testing methods to infer phylogenies from molecular data (Huelsenbeck and Hillis 1993; Hillis et al. 1994; Huelsenbeck 1995). Simulations have been used in the context of geometric

morphometrics to examine the effect of different evolutionary scenarios on ancestor–descendant distances and to explore the consequences on phylogenetic inference (Polly 2004). However, the simulations were conducted only under restricted sets of parameters (e.g., dimensionality, patterns of integration, branch lengths) and simulation results were evaluated qualitatively.

This study uses an extensive set of simulations to analyze how accurately phylogenies can be estimated using quantitative multidimensional data. Adopting an approach similar to that used by Huelsenbeck and Hillis (1993) in molecular phylogenetics, we use the four-taxon case as the simplest situation where different unrooted trees are possible, we systematically examine the effects of different combinations of branch lengths in the phylogeny, and we implement several evolutionary models. In addition, we vary the characteristics of the phenotypic variation, namely its dimensionality and the patterns of integration among the variables (Klingenberg 2008; Goswami et al. 2014). Together, these simulations furnish estimates of the accuracy of phylogenies inferred from multidimensional data under a wide range of conditions. The results provide new and decisive information to the debate about the role of multidimensional data in phylogenetics.

2.2. Materials and Methods

2.2.1. Simulation strategy

Complex phenotypes can be represented in a multidimensional space, in which evolving populations can be represented as points according to their average phenotypes. Examples of such multidimensional spaces are the space of gene expression (e.g., Brawand et al. 2011) and shape tangent spaces (Dryden and Mardia 1998; Kendall et al. 1999). Evolution of the mean phenotype in a population corresponds to movement of the respective point

through the space.

Our strategy follows the one used by Huelsenbeck and Hillis (1993) for molecular sequence data, but uses models of evolution in multidimensional phenotypic spaces. The simulation uses four taxa because this is the minimal number for which there are several different unrooted trees (three different trees). We repeatedly run an evolutionary simulation for four taxa in a phenotypic space (Fig. 2.1). We then estimate the unrooted tree from the resulting multidimensional phenotypes by computing the tree length for all three possible tree topologies. The proportion of simulations in which the shortest tree matches the tree topology used in the simulation, the proportion of correct estimates, is a natural measure of reliability of the phylogeny reconstruction. Because there are only three possible trees (Fig. 2.2a), it is feasible to evaluate all three possible trees for each simulation and the analyses are therefore guaranteed to find the optimal tree in each simulation. Also, it is clear which tree is correct and which ones are incorrect and there is none of the ambiguity about whether a reconstructed tree is "mostly correct" or "incorrect in some fundamental features", as it occurs in discussions of empirical examples involving more taxa.

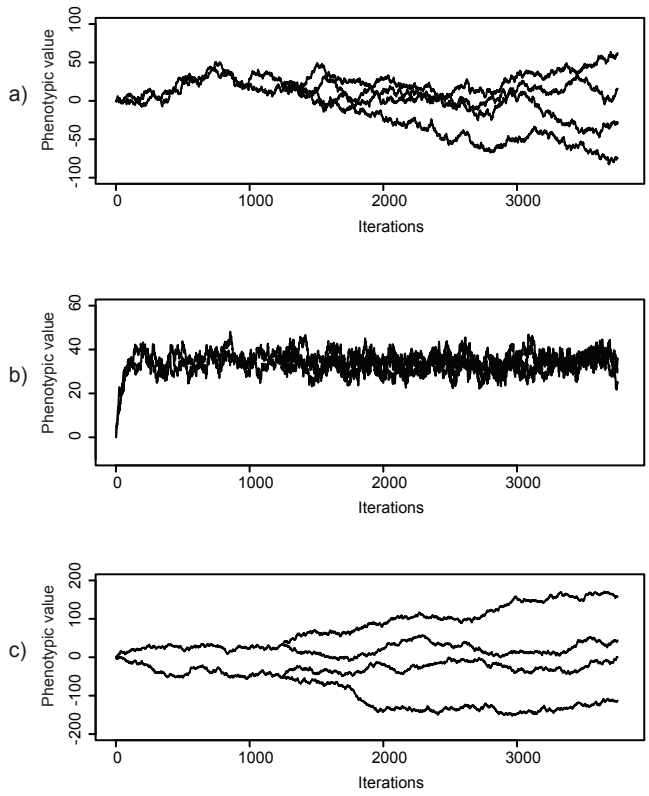

***Figure 2.1**. Examples of the three different evolutionary models used in the study. a) Brownian motion model. At each iteration the phenotypic values change randomly (note that there is a computational shortcut with identical outcome, as the phenotypic change on each branch can be simulated with a single random variable drawn from a normal distribution with appropriate variance). b) Stabilizing selection. At each iteration the phenotypic values change a certain amount towards the phenotypic optimum (a phenotypic value of 35 in this case) and also have a low amount of random movement. c) Repulsion model. At each iteration the phenotypic values move a certain amount away from those of other lineages and also have a low amount of random movement. Note that crossings are much more unlikely than using Brownian motion model.*

To examine the effects of variation in branch lengths, we adopt the systematic exploration of different combinations of branch lengths used in the

computer simulations of Huelsenbeck and Hillis (1993). Branch lengths reflect the opportunity for evolutionary change along the branches of a phylogeny, and result jointly from the rate of evolutionary change and the time interval corresponding to the respective branch of the phylogeny. We conduct two different sets of simulations, one to analyze the effects of the relative lengths of internal versus terminal branches (Figure 2.2b) and another set to study the effect of long-branch attraction and related difficulties for phylogeny reconstruction (Fig. 2.2c). In both cases, we divide the five branches in two groups, within which all the branches have the same length. In the first case, one group contains the four peripheral branches and second group consists of just the central branch (Fig. 2.2b). The phylogenetic reconstruction should be easier when the central branch is long relative to the peripheral branches, because this situation provides ample opportunity for the two internal nodes to diverge, while each of them is likely to remain close to its two adjoining terminal nodes. Conversely, if the central branch is very short relative to the terminal branches, so that the tree approaches a polytomy, all four taxa are expected to be roughly equidistant to one another and which tree fits the data best is mainly due to random variation. Note that if the central branch actually has length zero, the three possible unrooted trees represent the true tree equally well, so the evaluation of the phylogenetic reconstruction does not make sense. Whereas these expectations are fairly straightforward, it is not clear to what extent intermediate combinations of branch lengths provide reliable estimates of phylogeny. Our simulations aim to establish this under several evolutionary models.

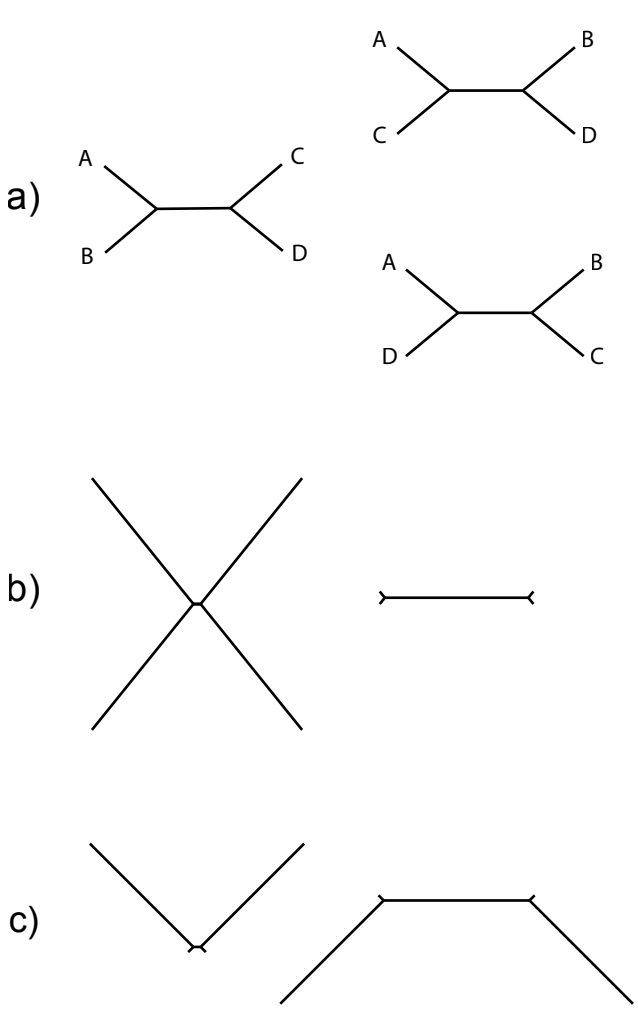

***Figure 2.2**. The three possible unrooted trees and two ways to vary branch lengths in the simulations. a) True phylogenetic tree simulated (left) and the two other possible tree topologies (right). b) Variation in branch lengths contrasting terminal versus central branches. All the terminal branches share a length and the central branch has a different length. The relative lengths of the two sets of branches are varied from 1;20 to 20:1. When the central branch is very long relative to the terminal branches (right), it is expected that estimating the phylogeny should be easy. c) Variation in branch lengths contrasting two terminal branches with the three remaining branches (2-versus-3 scenario). The situation at the left, where two terminal branches at either end of the central branch are much longer than the remaining three branches, is well known to be particularly challenging.*

In the second set of simulations, the central branch and one peripheral branch at either end of it have one branch length and the other two peripheral branches have another branch length (Fig 2c). This arrangement of relative branch lengths has been shown to pose potential challenges to phylogenetic

methods (Felsenstein 1978; Huelsenbeck and Hillis 1993; Huelsenbeck 1995). Remotely related terminal nodes may appear close because they are linked to the rest of the tree by long branches and therefore it may have a negative effect over the phylogenetic reconstruction. This situation has long been known as long-branch attraction or heterotachy, where the rate of evolutionary changes differs among lineages in the phylogeny, and has been widely studied in molecular phylogenetics (Wiens and Hollingsworth 2000; Bergsten 2005; Philippe et al. 2005; Wägele and Mayer 2007; Degtjareva et al. 2012). It is less clear, however, whether this problem has similarly serious effects on phylogeny estimation from multidimensional phenotypes.

In both sets of simulations, each set of relative branch lengths varies from 1 to 20. Therefore, we have situations in which one set of branches is much longer than the other (up to 20 times) and all the intermediate situations including that in which all the branches have the same length. For each set of simulations and evolutionary model, we start with the univariate case and then increase the dimensionality to 2, 3, 5, 10, 20, 50 and 100 dimensions.

For each number of dimensions and combination of branch lengths, we repeat the simulations and the phylogenetic reconstruction 5000 times to obtain a percentage of how many times the phylogenetic method chooses the correct tree. This percentage constitutes our estimate of phylogenetic reliability or phylogenetic accuracy. By comparing the phylogenetic reliability under different conditions we can assess the effect of the dimensionality, branch lengths and integration.

All the simulations are implemented using the R 2.10 statistical package (R Core Team 2013). The package 'mnormt' (Genz and Azzalini 2013) is used to implement the multivariate approach.

2.2.2. Brownian Motion

The first set of simulations assumes an evolutionary process according to a Brownian motion model, because this type of models has been of fundamental importance in discussions about phylogenies and quantitative traits (Cavalli-Sforza and Edwards 1967; Felsenstein 1973; Lynch 1989; Felsenstein 2002; Stayton 2008). This model assumes that the phenotype of each lineage evolves by a small amount for each short time span, and that this change is equally likely in every direction of the phenotypic space. In other words, the traits corresponding to the dimensions of the phenotypic space are evolving randomly, with equal evolutionary rates and independently of one another. The resulting evolutionary trajectory is a random walk through the phenotypic space (Fig. 2.1a). Under Brownian motion, there is an association between the time since the splitting of two lineages and the expected distance between the corresponding phenotypes, providing a possible basis for estimating phylogeny. Nevertheless, this association is not deterministic, but has a substantial stochastic component of variation, so that estimating the phylogeny from the distances between the phenotypes of the terminal nodes is inevitably fraught with a degree of uncertainty (Lynch 1989).

To conduct simulations under a Brownian motion model, random walks of lineages through the phenotypic space can be implemented explicitly (Figure 1a). It is more efficient, however, to obtain changes along the branches in the phylogeny directly as random vectors drawn from multivariate normal distributions with variances proportional to the respective branch lengths and zero covariances among variables (this follows from the multivariate version of the central limit theorem; e.g., Mardia et al. 1979). The phenotypes for the four terminal nodes can then be obtained by combining these changes in accordance with the true phylogenetic tree (Fig. 2.2a).

The reconstruction of the phylogeny from the phenotypes of the terminal nodes is done by computing the tree length for each of the three possible trees using squared-change parsimony (Maddison 1991; McArdle and Rodrigo 1994) and accepting the shortest tree. This estimate is also the maximum-likelihood estimate of the phylogeny under the Brownian motion model (Maddison 1991; Felsenstein 2004). This procedure was repeated with different numbers of dimensions and combinations of branch lengths, as described above.

For Brownian motion, the absolute scale of the branch lengths has no effect on the distribution of relative arrangements of taxa in phenotype space, other than the scale of distances between them. As branch lengths increase, it is expected that taxa will be farther from each other, but still will form the same relative patterns. This is different from molecular evolution, where there are saturation effects if the product of time and substitution rate becomes very large, because there are only four possible nucleotides (or 20 amino acids). Therefore, in each set of simulations, only the ratio of branch lengths in the two groups of branches needs to be varied, but not the absolute branch length. Accordingly, reliability is presented as the percentage of correctly estimated phylogenies, plotted as a function of the ratio of branch lengths in the two groups of branches.

2.2.3. Brownian Motion with Phenotypic Integration

The model of equal and independent evolution of all phenotypic traits is not a realistic representation of biological data, where integration among traits is virtually ubiquitous (Olson and Miller 1958; Cheverud 1996; Wagner et al. 2007; Klingenberg 2008, 2013). Integration means that traits are correlated with each other and that variation is concentrated in certain directions in phenotypic space (Wagner 1984; Klingenberg 2008; Pavlicev et al. 2009).

We include two more sets of simulations to investigate the effects of integration on estimation of phylogeny from multidimensional traits (Figure 2.3). One model simulates very strong integration, in which one dimension accounts for 80% of the total variation and all the other ones take up the remaining 20% of variation in equal amounts. In another simulation, the relative amount of variance decreases in an exponential manner from one dimension to the next, so that each eigenvector of the covariance matrix is 60% of the preceding eigenvector. At the end, all the eigenvalues are standardized so they sum up 1.

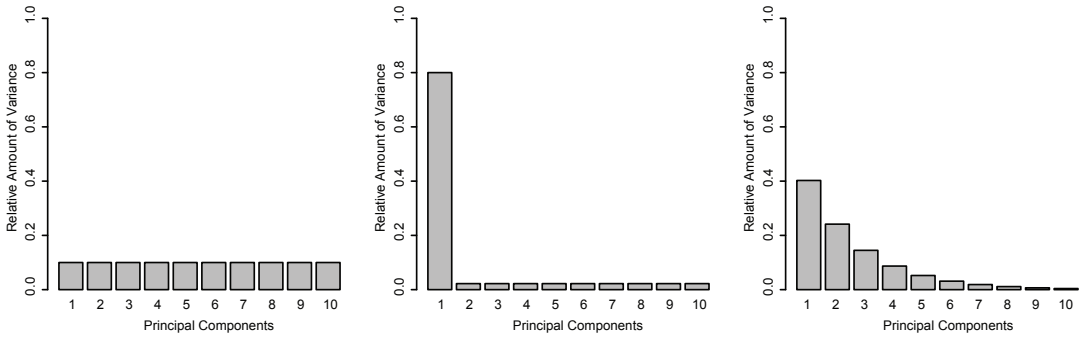

Figure 2.3*. Examples of the models of integration used in the study (shown here for 10 dimensions). In the scenario on the left, there is no integration: all the PCs have the same amount of variation. In the centre, the scenario of strong integration is shown, where the first PC has the 80% of the variation and the rest of the PCs share the 20%. On the right, the distribution of values follows an exponential model, which is a more realistic model of morphological integration.*

Because the method for estimating phylogeny is based on the distances between phenotypes of the different taxa, the rotation of the coordinate system used does not influence the results. Because of this invariance under rotation, we can choose any coordinate system without loss of generality. Accordingly, we use the principal components (PCs) of the

evolutionary covariance matrix as the coordinate system for our simulations, so that evolutionary changes in the resulting coordinates are uncorrelated with one another. We can therefore simulate the evolutionary change on each branch by independently drawing random deviations from normal distributions with variances obtained as the eigenvectors of the evolutionary covariance matrix multiplied with the respective branch length.

For this first set of simulations, tree length was computed using squared-change parsimony, where changes in every direction of phenotypic space are treated in the same way. This corresponds to the practice of many empirical studies that have estimated phylogenetic trees from untransformed morphometric variables (e.g., Lockwood et al. 2004; Caumul and Polly 2005; Couette et al. 2005; Cardini and Elton 2008; Smith and Hendricks 2013).

2.2.4. Estimates of Phylogeny Compensating for Integration

A way to take into account integration is the use of Mahalanobis distances in estimating phylogenies (Felsenstein 1973; Felsenstein 2002). Mahalanobis distances are based on a transformation of the phenotypic space that produces a modified space where the within-taxon variation is equal in every direction, so that the distance between groups becomes a measure of the degree of separation between them (Klingenberg and Monteiro 2005). In this modified space, therefore, the effect of integration within groups has been removed. This change, however, comes with other fundamental changes in the relative arrangement of taxon averages and in the scaling of different dimensions.

If the evolutionary covariance matrix were known, therefore, the phenotypic space could be scaled by the inverse of this matrix, transforming the space to a new space in which the distances are Mahalanobis distances and in which the Brownian motion model for evolutionary change would apply. In practice, however, the evolutionary covariance matrix is not known, but

must be estimated from the available data, which is exceedingly difficult if the phylogeny itself is also unknown (Felsenstein 1973; Felsenstein 2002). In principle, the phylogeny and evolutionary covariance matrix could be estimated simultaneously, but severe limits on the relative number of taxa and dimensions of the phenotypic space apply (Felsenstein 2002).

For the purpose of this study, and probably somewhat overoptimistically, we assume that the evolutionary covariance matrix is proportional to the within-taxon covariance matrix and the relevant information can therefore be obtained by characterizing patterns of individual variation within taxa. In practice, the pooled within-taxon covariance matrix can be computed from the deviations of individuals from the respective taxon means. In this study, we will simulate this situation by using a combined sample. This sample is created as a set of random numbers taken from a normal distribution with mean = (0,…,0) and the same covariance matrix used in the simulation of the averages. We conduct two sets of simulations, one with smaller and the other with larger sample sizes. In the first set of simulations with Mahalanobis distances we are taking a sample of 40 individuals when the dimensionality of the phenotype space is d = 1, 2, 3, 5, 10 and 20. When d = 50 we create a sample of 60 and when d = 100, the sample is of 120 individuals. These sample sizes are fairly typical of the sample sizes available in taxonomic studies, or may even be generous compared to many actual studies with similar dimensionality and only four taxa. In a second set of simulations, with larger sample sizes, we are taking a sample of 80 individuals when d = 1, 2, 3, 5, 10 and 20, a sample of 110 when d = 50 and a sample of 220 when d = 100. These sample sizes are untypically large for taxonomic studies.

For each simulation, three types of evolutionary covariance matrices were used: no integration (classical Brownian motion), strong integration with 80% of the total variance in the first eigenvalue, and the exponential model,

in which each successive eigenvalue is 60% of the preceding one (Fig. 2.3). The same covariance matrix was used to simulate both the evolutionary divergence along the branches of the phylogeny and variation in the sample of individuals. That sample was then used to compute a within-taxon covariance matrix for each simulation round, from which the transformation of the phenotype space into a Mahalanobis space was computed. In the transformed space, tree lengths were computed by squared-change parsimony (amounting to a criterion of minimal squared Mahalanobis distance).

2.2.5. Stabilizing Selection Model

Stabilizing selection appears to be widespread (e.g., Estes and Arnold 2007) and it can potentially have serious effects on estimates of phylogeny from the traits it affects (Polly 2004). The simulations of evolution under stabilizing selection were conducted as explicit random walks, starting from a root of the phylogeny located at the midpoint of the central branch (Fig. 2.1b). When stabilizing selection is simulated, at each interval from time t to $t + 1$, each population changes its position from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ following the equation $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha(\boldsymbol{\theta} - \mathbf{x}_t) + \boldsymbol{\sigma}$, where $\boldsymbol{\alpha}$ is a coefficient indicating the strength of stabilizing selection, $\boldsymbol{\theta}$ is the position of the adaptive peak, and $\boldsymbol{\sigma}$ is an isotropic random deviation, drawn from a multivariate normal distribution with zero mean and an identity matrix as the covariance matrix. The coefficient $\boldsymbol{\alpha}$ can take values from zero (in this case, the model will be the same as Brownian motion) to unity (in that case, the phenotype will be returned exactly to the optimum at each iteration, and will only deviate by the random effect newly added in that round).

Each simulation consists of a number of iterations that is determined by the branch lengths of the phylogeny in each simulation run. The branch lengths are varied in steps of 5 iterations from 10 to 105 iterations, according

to the branch lengths required for the simulation (Fig. 2.2b, c).

We conduct separate simulations with weak and strong stabilizing selection, which use values of α = 0.05 and α = 0.3 respectively. The simulations start with two populations at the root of the phylogeny, both with initial phenotypes $\mathbf{x}_0$ = (0,…,0). To test for the effect of the initial conditions, we conducted separate simulations where the starting point coincides with the optimal phenotype, **θ**= (0,…,0). A separate set of simulations is conducted for the situation where the starting point is at a distance to the optimum, which was set to **θ**= (35, 0, …, 0) (Fig. 2.1b). This is equivalent to a model that initially contains a component of directional selection, which then diminishes as each lineage approaches the optimum phenotype.

2.2.6. Repulsion Model

The final evolutionary model in this study was designed to produce the best possible conditions for the estimation of phylogeny from multidimensional phenotypes. To minimize the opportunity for convergence and parallelism, this model assumes that there is mutual repulsion among evolving lineages. A possible biological scenario for such repulsion is ecological character displacement, where morphologically and ecologically similar forms are under selection for divergence (Brown and Wilson 1956; Adams and Rohlf 2000; Adams 2010). We emphasize, however, that our aim is not to provide a realistic model of character displacement, but that the repulsion model is intended as a most favorable scenario for phylogeny estimation from multidimensional phenotypes. Some of the choices we made in implementing the model reflect this goal.

In the repulsion model, the phenotype of the *i*-th lineage changes in the interval from time *t* to *t* + 1 according to the following equation:

$$\mathbf{x}_{i,t+1} = \mathbf{x}_{i,t} + \alpha \left(\sum_{j \neq i} \frac{1}{\|\mathbf{x}_{i,t} - \mathbf{x}_{j,t}\|} \left(\mathbf{x}_{i,t} - \mathbf{x}_{j,t} \right) \right) + \boldsymbol{\sigma}.$$

In this equation, the parameter **α** quantifies the strength of repulsion effect and was set to a value of 1.5, whereas the expression $\|\mathbf{x}_{i,t} - \mathbf{x}_{j,t}\|$ denotes the distance between the phenotypic values of populations *i* and *j* at time *t*. Therefore, all the lineages in the simulations move away from each other at each unit of time by an amount that is inversely proportional to the distance between them. The parameter **σ** stands for the random component of variation, which is implemented as a random vector drawn from a multivariate isotropic distribution with zero mean and an identity matrix as the covariance matrix. This random component is therefore the same as isotropic Brownian motion. The number of model iterations, according to branch length, goes from 10 to 105 in increments of 5.

2.3. RESULTS

The simulations using Brownian motion show that the relative branch lengths and dimensionality play an important role in phylogenetic reconstruction (Fig. 2.4a, d). The phylogenetic reliability consistently improves as the central branch increases in length relative to the terminal branches (from left to right in the diagrams in Fig. 2.4a). In addition, as dimensionality increases, the region of high phylogenetic reliability expands toward shorter relative lengths of the central branch.

At low dimensionality, the situation where the central and two terminal branches vary in length relative to the two remaining terminal branches (Fig. 2.4d) appears more challenging than the situation where the central branch is contrasted to the four terminal branches (Fig. 2.4a). In the 2-versus-3 simulations, there are always two terminal branches that are at least as long as the central branch, providing opportunities for homoplasy by parallel or convergent evolution (Fig. 2.4d). For high dimensionality, however, the probability of homoplasy is reduced because there are many more directions in which taxa can diverge. As a result, the estimates are quite reliable even

when the central branch is short relative to two terminal branches connected to it at either end.

Phenotypic integration has a strong negative effect on the accuracy of phylogenetic estimates (Fig 4b, c, e, f). In the simulations with integration, reliability improves much less with increasing dimensionality than it does for isotropic Brownian motion and therefore the results are always worse than the corresponding simulations with isotropic variation. The effect of integration is similar to that of a reduction of the dimensionality. When one dimension contains 80% of the total variation, regardless of overall dimensionality, the relation between branch length ratios and reliability resembles that for the simulation with two dimensions and isotropic variation (compare Fig. 2.4b to Fig 2.4a, Fig. 2.4e to Fig. 2.4d). With the exponential model of integration, the relations of branch length ratios to reliability all resemble those for 3 or 5 dimensions under the model of isotropic variation (compare Fig. 2.4c to Fig 2.4a, Fig. 2.4f to Fig. 2.4d). In both these cases, most of the variation is contained in a single dimension or a few dimensions regardless of the overall dimensionality of the phenotypic space. Phenotypic integration effectively reduces the dimensionality of variation and therefore reduces the reliability of phylogeny estimation.

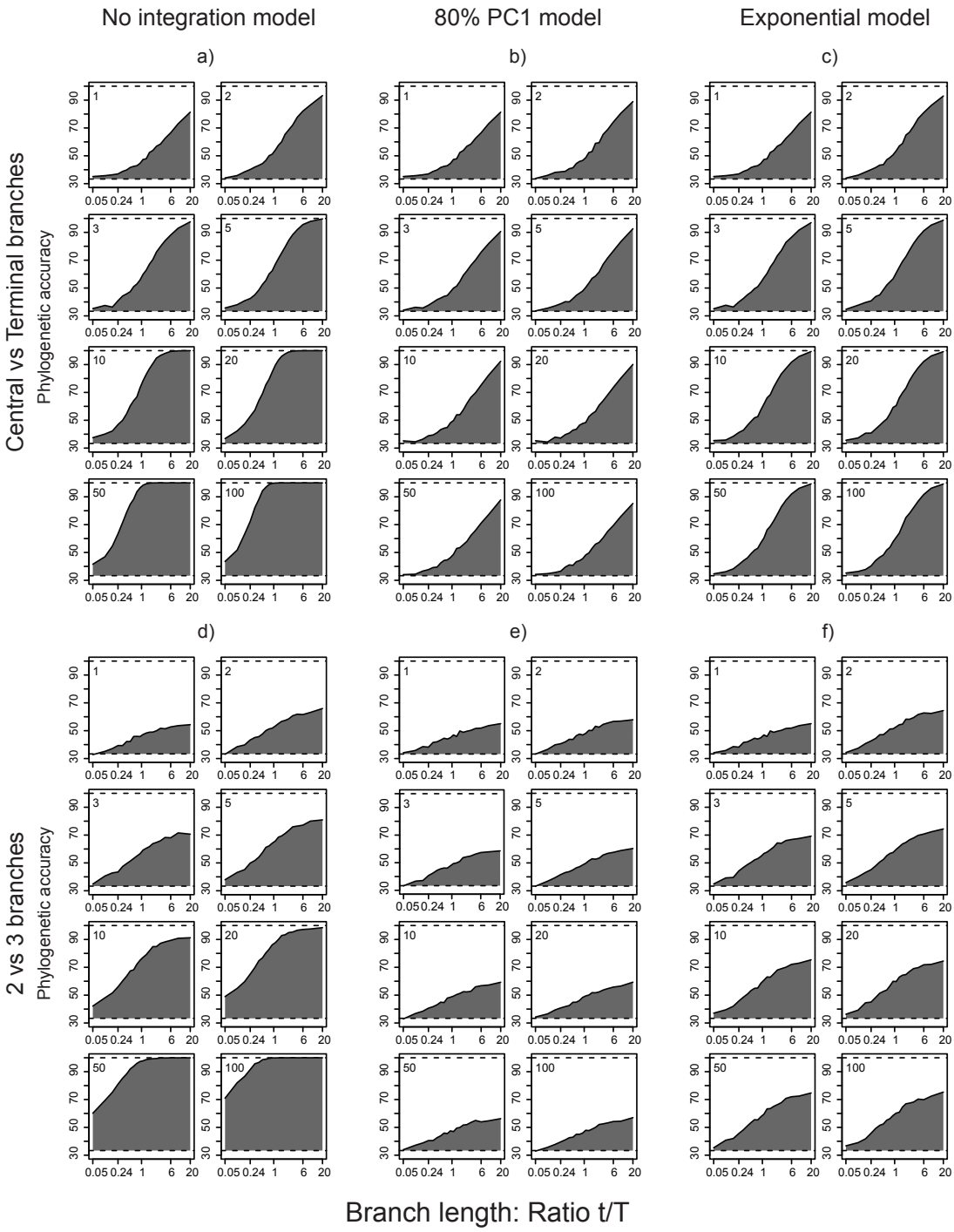

***Figure 2.4**. Estimation of the phylogenetic reliability under the Brownian motion model. The y-axis values are the percentages of right choices of the phylogenetic method and the x-axis values the logarithm of the ratio between the two sets of branches (central : terminal branches in a) to c) and 3-branch set : 2-branch set in d) to f)). In each diagram the number of dimensions used is indicated in the top left corner.*

Using Mahalanobis distance to estimate phylogeny tends to improve reliability if there is phenotypic integration (compare Fig. 2.5b, e and Fig. 2.6b, e versus Fig. 2.4b, e; Fig. 2.5c, f and Fig. 2.6c, f versus Fig. 2.4c, f). How much the effect of integration can be ameliorated by using Mahalanobis distances depends on the sample size and, to a lesser degree, on the number of dimensions with a significant amount of variation (Fig 2.5 and 2.6). Unsurprisingly, simulations using larger sample size show a greater improvement of phylogenetic accuracy than those with a smaller sample size (Fig 2.5 versus Fig 2.6). Nevertheless, neither set of simulations reaches the level of reliability for Brownian motion without integration (Fig. 2.4a, d), which would be expected if the true evolutionary covariance matrix were used for computing the Mahalanobis distances. The consequences of inaccurate estimates of the evolutionary covariance matrix are particularly evident if Mahalanobis distance is used for reconstructing phylogenies when there is no phenotypic integration (Fig. 2.5a, d, Fig. 2.6a, d). In this case, using Euclidean distance in the phenotypic space would be the optimal procedure, and the drop in reliability from the simulations using Euclidean distance (Fig. 2.4a, d) is a direct consequence of the error in estimating the evolutionary covariance matrix. As expected, this drop is less accentuated for large sample size (Fig. 2.5a, d) than for small sample size (Fig. 2.6a, d).

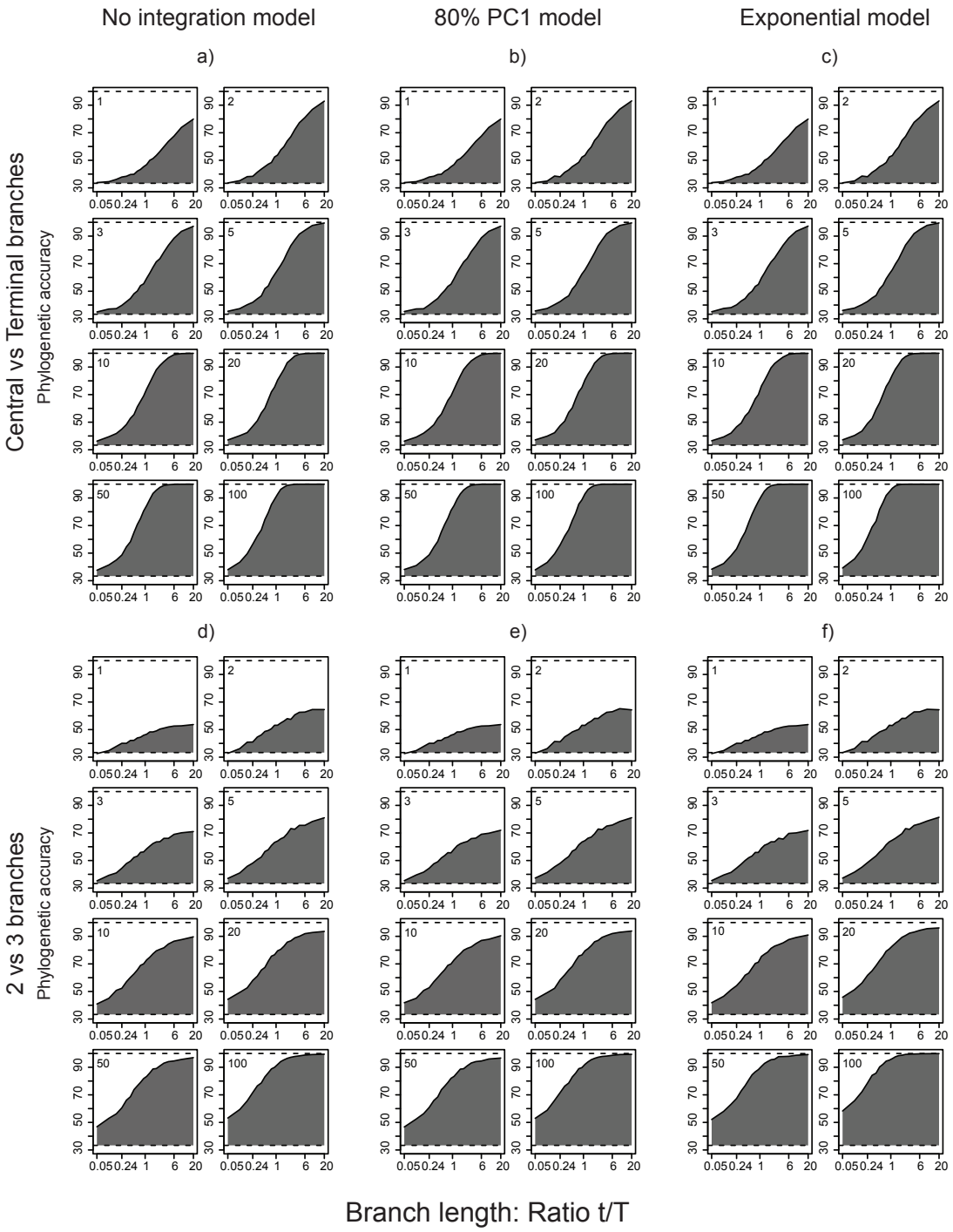

***Figure 2.5**. Estimation of the phylogenetic reliability, using Mahalanobis distances to take integration into account. Large sample sizes were used to estimate covariance matrices. For further explanations, see Fig. 2.4.*

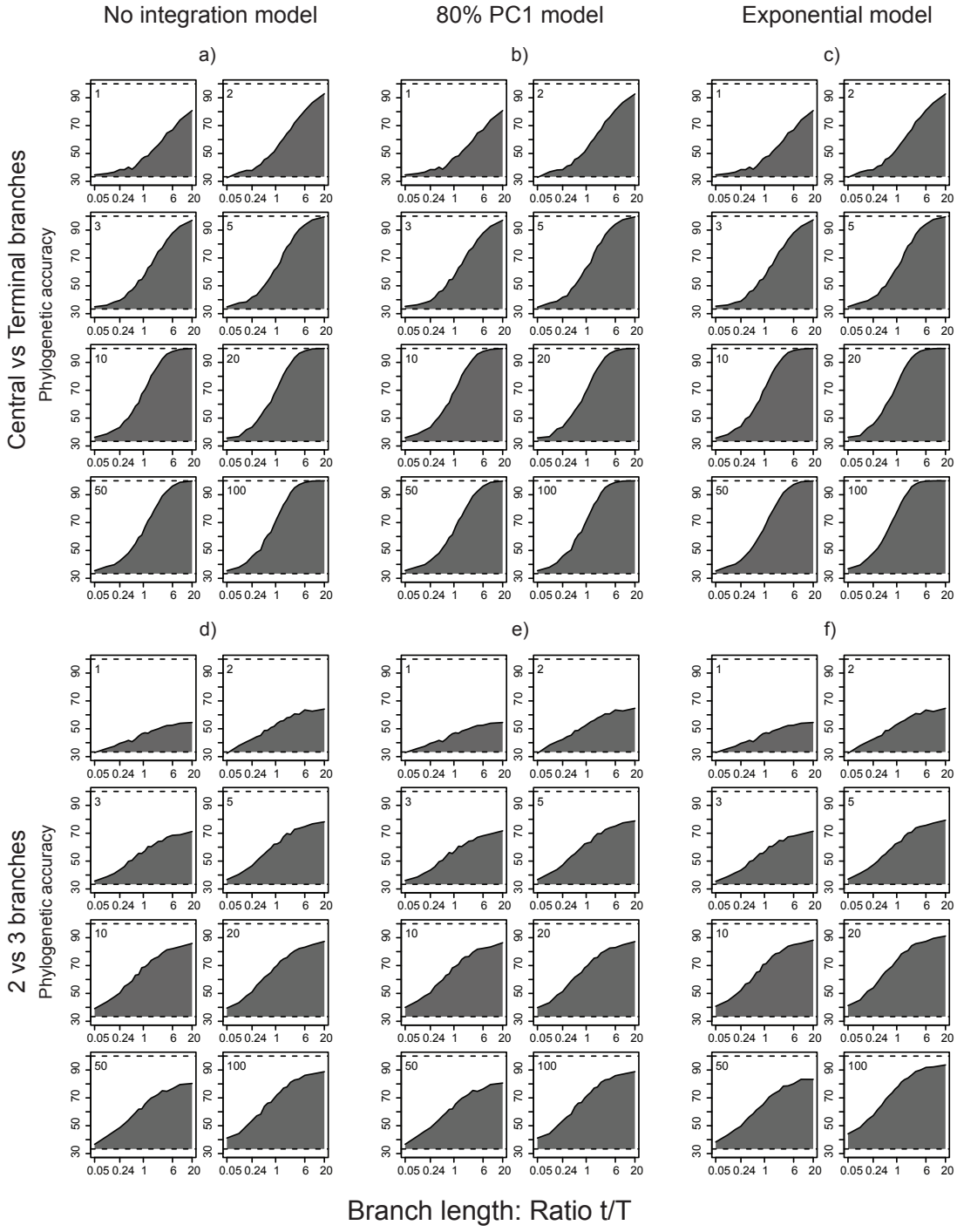

***Figure 2.6**. Estimation of the phylogenetic reliability, using Mahalanobis distances to take into account integration. Covariance matrices were estimated based on small sample sizes. For further explanations, see Fig. 2.4.*

Because the simulations with stabilizing selection were run over time and the outcome may differ according to the time scale when the starting phenotype is displaced from the optimum, we use a presentation where the absolute branch lengths are plotted in graphs (Fig. 2.7). Under stabilizing selection, phylogenetic reliability is not much better for most combinations of branch lengths than drawing trees randomly (Figure 2.7). If stabilizing selection is weak, the accuracy of the estimates is better where the terminal branches are very short (at the bottom of the diagrams in Fig. 2.7a, b), especially when the dimensionality is high. If the terminal branches are longer, phylogenetic accuracy is low because all lineages have time to approach the optimum regardless of ancestry. For the 2-versus-3 branch simulations, reliability is best if all branches are short and more or less equal (lower-left corners of the diagrams in Fig. 2.7d, e; note that this situation, with all branches short, is similar to the lower-left corners of the diagrams in Fig. 2.7a,b). A particular situation occurs for the simulation with weak stabilizing selection with an initial phenotype at some distance from the optimum, because there are some situations with very unequal branch lengths where phylogenetic reliability is worse than randomly drawing trees (left edges of the diagrams in Fig. 2.7e). In this situation, only the lineages of the two long terminal branches have time to approach the optimum and incorrect tree ((A,D),(B,C)) tends to be shorter than the correct tree ((A,B),(C,D)). With strong stabilizing selection, there is no branch length combination in which the phylogenetic reliability is perceptibly better than for drawing a phylogeny at random. This is true regardless of dimensionality, and it makes no noticeable difference whether the simulations start with the optimal phenotype or at a distance from it (Fig. 2.7c, d; simulations starting at the optimal phenotypes not shown because the graphs look the same).

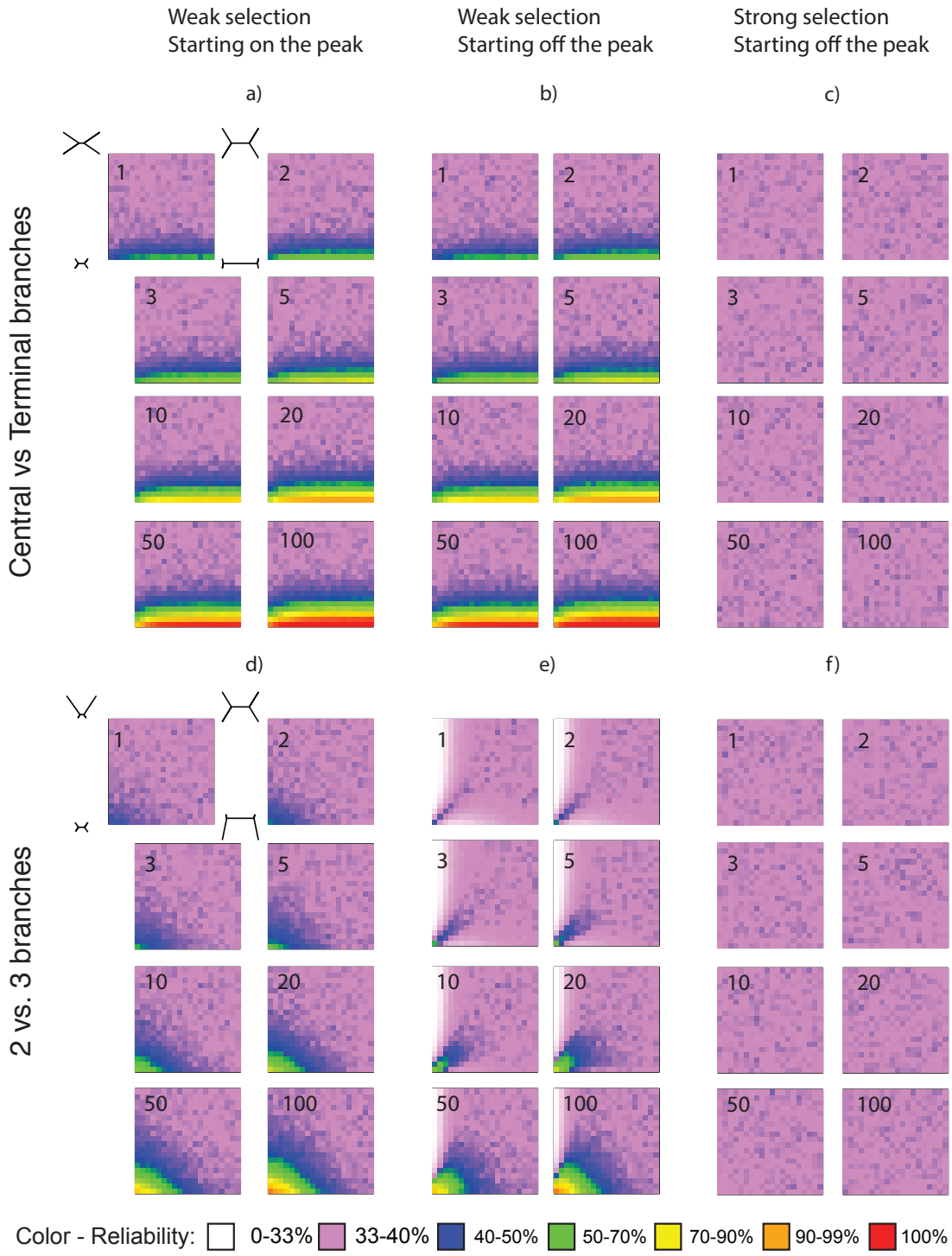

Figure 2.7. *Phylogenetic reliability in the simulations using evolutionary models with stabilizing selection. In each diagram the y-axis represents an increasing in the length of the central branch and the x-axis an increasing of the terminal branches (central : terminal branches in a) to c) and 3-branch set : 2-branch set in d) to f)). In each diagram the number of dimensions*

used is indicated in the top left corner. The color legend at the bottom indicates the reliability of each branch-length combination.

As expected, the repulsion model gives much better results (Figure 2.8). Even in the univariate case, there are no branch-length combinations that yield an accuracy of less than 70%. Intriguingly, if there is more than one dimension and the terminal branches are long relative to the central branch, phylogenetic reliability is slightly worse than in the univariate case. This drop in reliability relates to the fact that the mutual repulsion among taxa can cause them to form a near-symmetric square (in 2 dimensions) or tetrahedron (in spaces with 3 or more dimensions) from which the phylogeny is difficult to infer. For phenotypic spaces with many dimensions, repulsion combines with the tendency for divergence for the random component of evolution, so that almost all the branch-length combinations yield a good or excellent reliability (Fig. 2.8).

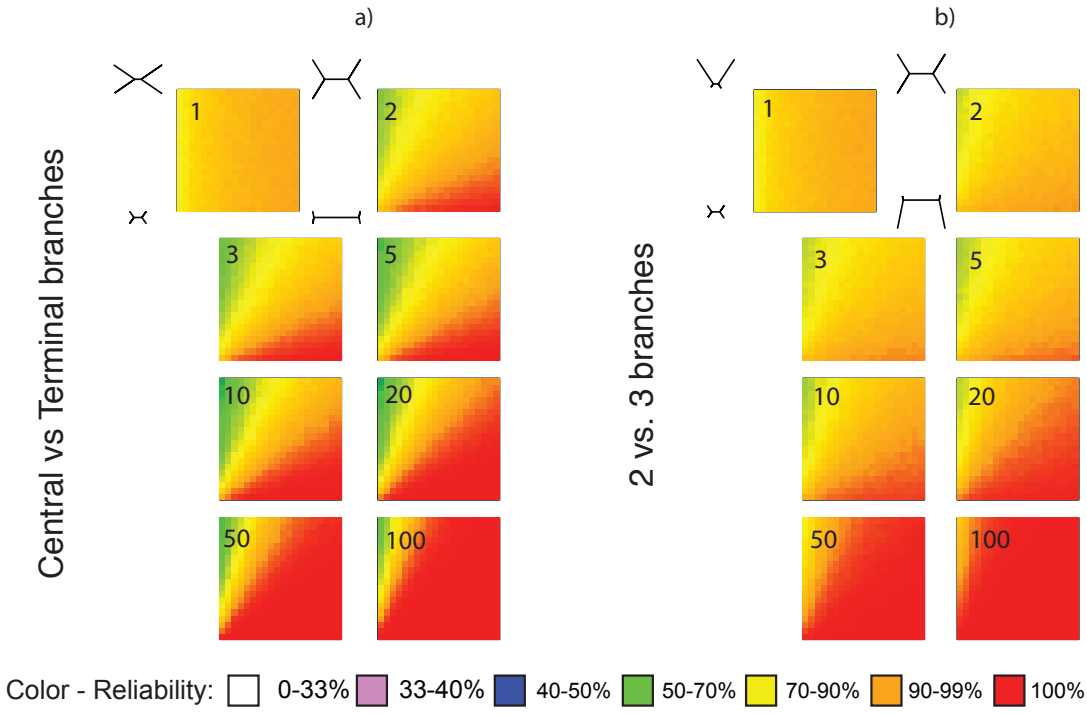

***Figure 2.8**. Phylogenetic reliability in simulations using the repulsion model of evolution. (a) Simulations using the contrasts of the relative lengths of*

central versus peripheral branches. (b) Simulations using the two-versus three branch contrasts. In each diagram the number of dimensions used is indicated in the top left corner. The color legend at the bottom indicates the reliability of each branch-length combination.

2.4. DISCUSSION

The simulations in this study have shown that the accuracy of phylogenetic estimates from multidimensional phenotypes depends greatly on the evolutionary model and on the dimensionality of the phenotype under study. Isotropic Brownian motion with high dimensionality yields good estimates of phylogenetic trees, but it is an unrealistic scenario for real biological data. Phenotypic integration, which is ubiquitous in real data, severely reduces the reliability of phylogenetic estimates because it decreases the effective dimensionality (Fig. 2.4). Using Mahalanobis distances to address the problem of integration is a viable solution in principle, but our simulations show that this solution does not work well even with large sample sizes and under a variety of favorable assumptions (Fig. 2.5, 2.6). Furthermore, stabilizing selection has a devastating effect on phylogenetic estimates because it systematically erodes the phylogenetic signal in the data (Fig. 2.7). The most consistently successful simulations are those using the repulsion model (Fig 2.8), which was constructed specifically as a best-case scenario for estimating phylogeny but is probably an unrealistic model for real evolutionary phenomena. Here we explore these results further and evaluate them in light of published evidence to assess their possible implications.

Most simulations show that the relative lengths of branches in the phylogeny are a key factor for the reliability of phylogenetic estimates. If the internal branch is very long relative to all four terminal branches, the phylogeny can be obtained reliably under all evolutionary scenarios except strong stabilizing selection (Fig. 2.4–2.8), as is to be expected for all

reasonable methods of phylogeny estimation. Conversely, if the internal branch is very short relative to the terminal branches, the phylogeny approaches an unresolved polytomy and no method can perform much better than randomly drawing a tree from the three possible topologies. The key question therefore is about the results of the simulations between these extreme situations and, particularly, for the situations where two terminal branches at opposite ends of the terminal branch differ in their lengths from the three remaining branches (Fig. 2.1c), a situation that has long been recognized to pose challenges for phylogeny estimation due to rate heterogeneity and long-branch effects (Felsenstein 1978; Huelsenbeck and Hillis 1993; Huelsenbeck 1995). This situation indeed proved difficult in many of the simulations, so that high reliability is only achieved if dimensionality is very high (Fig. 2.4–2.8). Nevertheless, although many simulations produced results that seem close to picking a phylogeny at random, even the 2-versus-3 branch scenarios did not systematically yield an incorrect estimate of the phylogeny (except Fig. 2.7e). This differs from studies relating to molecular data, where the 2-versus-3 branch scenario, with certain combinations of evolutionary models and estimation methods, can systematically yield incorrect estimates of the phylogeny (Felsenstein 1978; Huelsenbeck and Hillis 1993; Huelsenbeck 1995). With multidimensional continuous traits and the evolutionary models studied here, it appears that random errors in phylogeny estimation are far more important than systematic errors.

The evolutionary models and the dimensionality of the phenotype are very important to explain the results of any simulation study concerning phylogenetic accuracy (Huelsenbeck 1995; Swofford et al. 1996). Brownian motion has been widely used as a model for the evolution of phenotypic traits in one- or multidimensional settings (Cavalli-Sforza and Edwards 1967; Felsenstein 1973; Lynch 1989; Polly 2004; Stayton 2008). Brownian motion is an evolutionary model that is favorable for estimating phylogeny because the

expected distance between taxa increases with the time of separation (Lynch 1989). Yet, a difficulty has been pointed out because this distance also has a high variance (a coefficient of variation of 1.4 for one-dimensional Brownian motion) and therefore may often lead to convergence, reversals and parallel evolution that may produce erroneous phylogenetic estimates (Lynch 1989; Stayton 2008; Klingenberg and Gidaszewski 2010). The squared distance between the phenotypes at either end of a branch of the phylogeny, up to a scaling factor the expected change along the branch, follows a chi-squared distribution with as many degrees of freedom as there are dimensions in the phenotypic space (this follows from the Pythagorean theorem and the definition of the chi-squared distribution with *n* degrees of freedom as the sum of squared values of *n* mutually independent random variates drawn from the standard normal distribution). The coefficient of variation for the chi-squared distribution is the square root of two divided by the square root of the degrees of freedom. The relative variability of the phenotypic distances thus diminishes with increasing degrees of freedom, so that the phenotypic distances are a better reflection of the branch lengths with increasing dimensionality of the phenotype. Therefore, studying phenotypes with high dimensionality has been proposed as one way of increasing phylogenetic reliability (Felsenstein 1973; Polly 2004; González-José et al. 2008; Stayton 2008). Note, however, that a substantial improvement is only achieved with dimensionalities that are quite high: the coefficient of variation is 0.44 for 10 dimensions, 0.2 for 50 dimensions, 0.14 for 100 dimensions, and 200 dimensions are necessary for a coefficient of variation of 0.1. The benefits of high dimensionality also can be understood intuitively because there is always just one direction in which two points can converge toward each other, but with increasing dimensionality, there are more and more directions in which two points can move away from each other. Convergence is very likely in the univariate case, as shown in previous studies (Lynch 1989), but it becomes

less probable as more dimensions are added (Fig. 2.3a,d; Stayton 2008). High dimensionality also can alleviate the problems of long-branch attraction and differences in evolutionary rates among branches in the phylogeny (Fig 3d). This high dimensionality is typical for studies using genetic data (Cavalli-Sforza and Edwards 1967; Brazil et al. 2008) or shape (Cardini and Elton 2008; González-José et al. 2008; Piras et al. 2010; von Cramon-Taubadel and Smith 2012; Smith and Hendricks 2013). In principle, therefore, it seems that phylogenetic accuracy might be improved by adding more data, even though these beneficial effects will diminish with the number of dimensions already included.

In practice, however, the variation in the data often does not "fill" the entire dimensionality of the phenotypic space, but is concentrated mostly in a few of the available dimensions because of integration (Olson and Miller 1958; Cheverud 1996; Klingenberg 2008, 2013; Goswami et al. 2014). The simulations in this study show that integration is a significant force that reduces the effective dimensionality of variation and therefore also limits the reliability of phylogenetic estimates (Fig 2.4b, c, e, f). In the simulations where one dimension accounts for 80% of the total variation, even in spaces of high dimensionality, phylogenetic reliability resembles that of a model with isotropic variation in one or two dimensions (compare Fig. 2.4a to Fig. 2.4b, Fig. 2.4d to Fig. 2.4e). This reflects the fact that most variation indeed is in a single dimension. Likewise, in the simulations using the exponential model of integration, even with high dimensionality, phylogenetic reliability does not exceed that for an isotropic model with 3–5 dimensions (Fig. 2.4a, d versus 2.4c, f). Because most variation is concentrated within just a few dimensions and this distribution remains essentially the same no matter how many additional dimensions are included, the overall dimensionality of the phenotypic space seems immaterial for phylogenetic reliability. It appears from these simulations that integration may be an important problem for

phylogenetic reconstruction. This raises the question whether the simulations of integration are realistic at all. The scenario in which 80% of the variation is contained in the first principal component was designed to be extreme and probably exceeds the level of integration in real data, although some examples come relatively close (e.g., analyses with more than 60% of variation among species in the PC1; Klingenberg et al. 2012). The exponential model of integration is more realistic, as numerous examples show comparable or greater strengths of interspecific integration in geometric morphometric data (Monteiro et al. 2005; Sidlauskas 2008; Friedman 2010; De Esteban-Trivigno 2011b, a; Monteiro and Nogueira 2011; Brusatte et al. 2012; Santana and Lofgren 2013; Baab et al. 2014; Martín-Serra et al. 2014; Watanabe and Slice 2014), although some other studies found somewhat weaker integration, albeit still with most variation concentrated in just a few dimensions (Figueirido et al. 2010; Chamero et al. 2013; Klingenberg and Marugán-Lobón 2013; Sherratt et al. 2014). Altogether, the exponential model of integration used in the simulation seems to be fairly realistic by comparison with empirical data, so that the results of those simulations need to be taken seriously. Even though including additional data in a phylogenetic analysis seems a tempting way of improving phylogenetic reliability, it is therefore unlikely to make much of a difference unless the new data are more or less independent of those already included. Because of the near-ubiquity of phenotypic integration (Klingenberg 2013), this is unlikely to be successful. Morphological integration is a serious problem for phylogenetic reliability.

In principle, the adverse effects of integration can be mitigated by using Mahalanobis distance in the process of estimating phylogeny (Felsenstein 1973; Felsenstein 2002). If the correct evolutionary covariance matrix is used to compute Mahalanobis distances, this eliminates the effects of integration and phylogenetic reliability therefore should be the same as for Brownian motion with no integration. Our simulations with estimated

covariance matrices show some improvements of phylogenetic reliability (Fig. 2.5, 2.6), but phylogenetic reliability is not restored completely to the levels for Brownian motion without integration (Fig. 2.4a, d). Reliability is better if large sample sizes are used (Fig. 2.5), but the simulations also indicate that the mitigating effects are reduced if smaller sample sizes are used (Fig. 2.6). The drop in phylogenetic reliability when estimated Mahalanobis distances are applied in a situation without integration directly shows the effects of error in estimating the within-taxon covariance matrix (compare Fig. 2.5a,d and Fig. 2.6a,d to Fig. 2.4a,d). When assessing what implications these simulations have for empirical studies, we need to take into account that we made some assumptions that are favorable for phylogenetic reliability. The sample sizes varied from 40 to 120 for the simulations with smaller sample sizes, from 80 to 220 for the simulations with larger sample sizes. These sample sizes, from 10 to 30 or 20 to 55 specimens per taxon, are comparable to some studies with relatively large sample sizes (Lockwood et al. 2004; Cardini and Elton 2008), but often studies using geometric morphometrics are based on sample sizes at the lower end of this spectrum (Aguilar-Medrano et al. 2011; Smith and Hendricks 2013), and some studies use individual specimens as taxa (MacLeod 2002; González-José et al. 2008). In addition, the simulations assume that the pattern of evolutionary covariation can be estimated from the pattern of within-taxon covariation. This assumption is motivated by results from quantitative genetic theory, stating that the pattern of evolutionary divergence is a scaled version of the pattern of genetic variation within populations if evolution is due to random drift (Lande 1979). This reasoning contains further auxiliary assumptions. First, the phenotypic patterns of integration must be sufficiently similar to the genetic ones to serve as a proxy for them, a point that remains contentious (Cheverud 1988; Willis et al. 1991; Kruuk et al. 2008). Second, the patterns of within-taxon variation need to be constant across the whole phylogeny, so that it is

possible to obtain a single estimate of within-taxon variation; this assumption also is under continuing debate (Steppan et al. 2002; Arnold et al. 2008). Some morphometric studies have found similar patterns of integration within and between populations or species (Monteiro et al. 2005; Drake and Klingenberg 2010; Smith 2011; Klingenberg et al. 2012; Goswami et al. 2014), but this kind of similarity cannot be generally expected and there are case studies demonstrating that the patterns and strength of phenotypic integration can evolve (Jamniczky and Hallgrímsson 2009; Sanger et al. 2012; Gómez et al. 2014). All these assumptions built into our models are favorable for using of Mahalanobis distances in estimating phylogeny, but it is doubtful how realistic they are. It is therefore possible that these simulations give an overoptimistic picture and that using Mahalanobis distance to account for integration is less promising than it may appear from our simulation results.

When the evolutionary model used in the simulations is stabilizing selection, the results are rather discouraging because the phylogenetic reliability, with most of the combinations of branch lengths, is little better than for picking a tree at random (Fig. 2.7). When stabilizing selection is weak, some reliable results can be obtained when dimensionality is high and the terminal branches are short (Fig. 2.7a, b). When stabilizing selection is strong, phylogenetic reliability is low in all simulations regardless of the branch lengths. Under stabilizing selection, all the different lineages evolve toward the adaptive peak regardless of their ancestry and, once arrived at the optimum phenotype, each of the lineages tend to return immediately from any random movement away from the oprtimum. Because stabilizing selection affects each lineage regardless of ancestry, it will erode any phylogenetic signal that results from phenotypic deviations shared by sister lineages just after splitting from each other. Once the lineages have split, each evolves separately in a balance between the addition of new random

variation and the constant corrections toward the optimum phenotype, which can remove any relation to ancestral phenotypes. If stabilizing selection is sufficiently strong or if the terminal branches of the phylogeny are sufficiently long, the relative positions of the lineages in phenotype space have no association with their relatedness. Therefore, under these conditions, reconstructing the phylogeny from phenotypic values may indeed be no better than picking a tree at random. Studies of quantitative phenotypes such morphological traits and gene expression have found extensive evolutionary conservation, which is consistent with the view that stabilizing selection is widespread (Rifkin et al. 2003; Estes and Arnold 2007; Hunt 2007; Kalinka et al. 2010; Gallego Romero et al. 2012). It is therefore likely that problems similar to those in our simulations will occur in many studies that attempt to estimate phylogenies from multidimensional phenotypes.

In addition to stabilizing selection, the simulations that start at a distance from the optimal phenotype contain a component of directional selection that is present until the evolving lineages reach the region of the fitness peak. For many simulations, this additional component of selection has little effect (compare Fig. 2.7a to Fig. 2.7b; no visible difference with strong stabilizing selection). Under weak stabilizing selection, however, there are circumstances when that directional component affects phylogeny reconstruction from the phenotypes so that it systematically yields the wrong tree, that is, it performs worse than picking a tree at random. This happens in the two-versus-three branch simulations when two branches are relatively long and the remaining three branches are very short (left edges of the diagrams in Fig. 2.7e). Because of the short branches, two lineages at opposite ends of the (short) central branch remain near the starting point, whereas the two lineages corresponding to the long branches undergo parallel evolution toward the optimal phenotype. As a result, the arrangement of taxa in phenotype does not reflect their relatedness. This scenario is probably not a

realistic one for natural evolution, but it highlights the pitfalls of evolution when there are directional trends, so that reconstruction of ancestral values from measurements of descendants faces serious difficulties (Oakley and Cunningham 2000; Webster and Purvis 2002), in combination with the long-established problems concerning long branches (Felsenstein 1978; Huelsenbeck and Hillis 1993).

Whereas evolution under a model of Brownian motion, in principle at least, can continue without bounds, models of stabilizing selection ensure that phenotypes sooner or later converge toward the optimal phenotype and remain in that region. If stabilizing selection is sufficiently strong or the branches are sufficiently long, there is therefore no longer an association between the time of separation and the phenotypic distance between taxa. In other words, the phenotype loses any phylogenetic signal it may have had (see the upper-right regions of the diagrams on Fig. 2.7). This phenomenon is analogous to the problem of substitution saturation in molecular data, when the product of substitution rate and branch lengths is so large that each position is expected to have undergone multiple substitutions and therefore loses phylogenetic information. This is different from the other models used in this study, where no such phenomenon exists and phenotypic differences are expected to increase with time. In real organisms, however, there cannot be an indefinite amount of change. Simulations of Brownian motion can easily produce phenotypes that are clearly non-functional (Polly 2004), so that it might be preferable to view the models as restricted to a domain of phenotype space within which phenotypes are viable. In practice, however, this problem is may not be a serious problem because the amount of phenotypic variation tends to be within tractable limits even at large phylogenetic scale (Marcus et al. 2000; Friedman 2010; Sallan and Friedman 2012; Klingenberg and Marugán-Lobón 2013; Sherratt et al. 2014). Any effects of such boundaries, if they exist, would manifest themselves as

stabilizing selection and therefore would probably be detrimental to phylogenetic reliability.

The repulsion model was specifically designed as a best-case scenario for phylogenetic analysis, and it is therefore not too surprising that phylogenetic reliability was high under a wide range of conditions (Fig. 2.8). The results are unusual by comparison to the other models because the simulation in a one-dimensional phenotypic space has a better phylogenetic reliability than those two- or higher-dimensional spaces when the length of the central branch is short relative to the terminal branches. This is because, if there are multiple dimensions, taxa can evolve "past one another" by moving in different directions to form nearly symmetric configurations such as squares (in 2 dimensions) or tetrahedral for which phylogenies are difficult to estimate. By contrast, when there is only a single dimension, taxa consistently push one another in opposite directions. The repulsion model, although favorable for phylogeny reconstruction, is not a realistic evolutionary scenario. The idea of repulsion is loosely based on the phenomenon of character displacement, where competition due to the presence of morphologically and ecologically similar taxa leads to selection for divergence (Brown and Wilson 1956; Stuart and Losos 2013). Geometric morphometric studies have suggested character displacement in a range of taxa including salamanders, shrews and bats (Rácz and Demeter 1998; Adams and Rohlf 2000; Gannon and Rácz 2006). Nevertheless, it is unclear how widespread character displacement may be because there are only few case studies where a sufficient range of evidence is available to rule out possible alternative explanations (Stuart and Losos 2013). Regardless of this question, the repulsion model, as used in this study, is not intended as a realistic model of character displacement. In the model, repulsion is active over larger distances in phenotypic space than it would be plausible for character displacement (e.g., scaling the repulsion effect by the inverse of the square or

a higher power of distance might provide a more realistic model), but in this way the model contains a more consistent tendency for lineages to diverge from each other and thus provides more favorable conditions for phylogeny reconstruction. Yet, even under these very optimistic conditions, phylogenetic reliability is not perfect unless dimensionality is high or the central branch of the phylogeny is relatively long. In the presence of morphological integration, which concentrates variation in just a few dimensions of the phenotypic space, there is therefore still an appreciable probability that phylogenetic estimates are erroneous.

In conclusion, the simulations in this study demonstrate that estimates of phylogenetic trees derived from multidimensional data tend to be reliable only under specific conditions that are unlikely to apply in naturally evolving clades. Widespread phenomena such as morphological integration and stabilizing selection can severely limit phylogenetic reliability, under some circumstances to the extent that estimating phylogenies from multidimensional data is little better than picking trees at random (Fig. 2.7). This conclusion is more pessimistic than several earlier assessments of the suitability of morphometric data for estimating phylogenies (e.g., MacLeod 2002; Lockwood et al. 2004; Polly 2004; Caumul and Polly 2005; Smith and Hendricks 2013), but is in line with earlier studies that found discrepancies between morphometric estimates and independently obtained phylogenetic trees (e.g., Marcus et al. 2000; Caumul and Polly 2005; Cardini and Elton 2008; Klingenberg and Gidaszewski 2010). The simplicity of the four-taxon simulations, where there is no ambiguity whether an estimate is correct or incorrect, and the broad spectrum of different combinations of branch lengths and evolutionary models makes our conclusions robust and decisive. We understand that the results are frustrating to some investigators, particularly to paleontologists, because morphometric data may be the only or at least most easily available data for many fossil and even some extant taxa

(MacLeod 2002; Smith and Hendricks 2013). Where possible, other data such as DNA sequence information can be used instead, which suffers from these difficulties to a lesser extent and where vast amounts of information are available (Rannala and Yang 2008). Even where such alternatives are not available, however, we think it is preferable to recognize the limitations of phylogenetic inference from such data, rather than to use approaches that may provide unreliable results.

2.5. REFERENCES

Adams D.C. 2010. Parallel evolution of character displacement by competitive selection in terrestrial salamanders. Bmc Evolutionary Biology 10:72.

Adams D.C., Cardini A., Monteiro L.R., O'Higgins P., Rohlf F.J. 2011. Morphometrics and phylogenetics: Principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. Journal of Human Evolution 60:240-243.

Adams D.C., Rohlf F.J. 2000. Ecological character displacement in *Plethodon*: biomechanical differences found from a geometric morphometric study. Proceedings of the National Academy of Sciences of the USA 97:4106–4111.

Adams D.C., Rosenberg M.S. 1998. Partial warps, phylogeny, and ontogeny: A comment on Fink and Zelditch (1995). Systematic Biology 47:168-173.

Aguilar-Medrano R., Frédérich B., de Luna E., Balart E.F. 2011. Patterns of morphological evolution of the cephalic region in damselfishes (Perciformes: Pomacentridae) of the Eastern Pacific. Biol J Linn Soc 102:593–613.

Arnold S.J., Bürger R., Hohenlohe P.A., Ajie B.C., Jones A.G. 2008. Understanding the evolution and stability of the **G**-matrix. Evolution 62:2451–2461.

Baab K.L., Perry J.M.G., Rohlf F.J., Jungers W.L. 2014. Phylogenetic, ecological, and allometric correlates of cranial shape in Malagasy lemuriforms. Evolution 68:1450–1468.

Bergsten J. 2005. A review of long-branch attraction. Cladistics 21:163–193.

Bookstein F. 1994. Can biometrical shape be a homologous character? In: Hall B. K. editor. Homology: The hierarchial basis of comparative biology. New York, Academic Press, p. 197-227.

Brawand D., Soumillon M., Necsulea A., Julien P., Csárdi G., Harrigan P., Weier M., Liechti A., Aximu-Petri A., Kircher M., Albert F.W., Zeller U., Khaitovich P., Grützner F., Bergmann S., Nielsen R., Pääbo S., Kaessmann H. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343-350.

Brazil M., Thomas D.A., Nielsen B.K., Winter P., Wulff-Nilsen C., Zachariasen M. 2008. A novel approach to phylogenetic trees: *d*-dimensional geometric Steiner trees. Networks 53:104–111.

Brown W.L., Wilson E.O. 1956. Character displacement. Systematic Zoology 5:49-64.

Brusatte S.L., Sakamoto M., Montanari S., Harcourt Smith W.E.H. 2012. The evolution of cranial form and function in theropod dinosaurs: insights from geometric morphometrics. Journal of Evolutionary Biology 25:365–377.

Cannon C.H., Manos P.S. 2001. Combining and comparing morphometric shape descriptors with a molecular phylogeny: the case of fruit type evolution in Bornean *Lithocarpus* (Fagaceae). Systematic Biology 50:860–880.

Cardini A., Elton S. 2008. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. Biol J Linn Soc 93:813-834.

Catalano S.A., Coloboff P.A., Giannini N.P. 2010. Phylogenetic morphometrics (I): the use of landmark data in a phylogenetic framework. Cladistics 26:539-549.

Catalano S.A., Goloboff P.A. 2012. Simultaneously mapping and superimposing landmark configurations with parsimony as optimality criterion. Systematic Biology 61:392-400.

Caumul R., Polly P.D. 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (*Marmota*, Rodentia). Evolution 59:2460–2472.

Cavalli-Sforza L.L., Edwards A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570.

Chamero B., Buscalioni Á.D., Marugán-Lobón J. 2013. Pectoral girdle and forelimb variation in extant Crocodylia: the coracoid–humerus pair as an evolutionary module. Biol J Linn Soc 108:600–618.

Cheverud J.M. 1988. A comparison of genetic and phenotypic correlations. Evolution 42:958–968.

Cheverud J.M. 1996. Developmental integration and the evolution of pleiotropy. Amer. Zool. 36:44–50.

Clouse R.M., de Bivort B.L., Giribet G. 2011. Phylogenetic signal in morphometric data. Cladistics 27:337–340.

Cole T.M., III., Lele S.R., Richtsmeier J.T. 2002. A parametric bootstrap approach to the detection of phylogenetic signals in landmark data. In: MacLeod N., Forey P. L. editors. Morphology, shape and phylogeny. London, Taylor & Francis, p. 194–219.

Couette S., Escarguel G., Montuire S. 2005. Constructing, bootstrapping, and comparing morphometric and phylogenetic trees: A case study of New

World monkeys (Platyrrhini, Primates). Journal of Mammalogy 86:773–781.

De Esteban-Trivigno S. 2011a. Buscando patrones ecomorfológicos comunes entre ungulados actuales y xenartros extintos. Ameghiniana 48:189–209.

De Esteban-Trivigno S. 2011b. Ecomorfología de xenartros extintos: análisis de la mandíbula con métodos de morfometría geométrica. Ameghiniana 48:381–398.

Degtjareva G.V., Valiejo-Roman C.M., Samigullin T.H., Guara-Requena M., Sokoloff D.D. 2012. Phylogenetics of *Anthyllis* (Leguminosae: Papilionoideae: Loteae): Partial incongruence between nuclear and plastid markers, a long branch problem and implications for morphological evolution. Molecular Phylogenetics and Evolution 62:693-707.

Drake A.G., Klingenberg C.P. 2010. Large-scale diversification of skull shape in domestic dogs: disparity and modularity. American Naturalist 175:289–301.

Dryden I.L., Mardia K.V. 1998. Statistical shape analysis. New York, John Wiley & Sons.

Enard W., Khaitovich P., Klose J., Zöllner S., Heissig F., Giavalisco P., Nieselt-Struwe K., Muchmore E., Varki A., Ravid R., Doxiadis G.M., Bontrop R.E., Pääbo S. 2002. Intra- and interspecific variation in primate gene expression patterns. Science 296:340-343.

Estes S., Arnold S.J. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. American Naturalist 169:227-244.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics 25:471–492.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Biology 27:401–410.

Felsenstein J. 2002. Quantitative characters, phylogenies, and morphometrics. In: MacLeod N., Forey P. L. editors. Morphology, Shape & Phylogeny. London, Taylor & Francis, p. 27-44.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA, Sinauer Associates.

Figueirido B., Serrano-Alarcón F.J., Slater G.J., Palmqvist P. 2010. Shape at the cross-roads: homoplasy and history in the evolution of the carnivoran skull towards herbivory. Journal of Evolutionary Biology 23:2579–2594.

Fink W.L., Zelditch M.L. 1995. Phylogenetic analysis of ontogenetic shape transformations: A reassessment of the piranha Genus Pygocentrus (Teleostei). Systematic Biology 44:343-360.

Friedman M. 2010. Explosive morphological diversification of spiny-finned teleost fishes in the aftermath of the end-Cretaceous extinction. Proceedings of the Royal Society of London, B Biological Sciences 277:1675–1683.

Gallego Romero I., Ruvinsky I., Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics 13:505–516.

Gannon W.L., Rácz G.R. 2006. Character displacement and ecomorphological analysis of two long-eared *Myotis* (*M. auriculus* and *M. evotis*). Journal of Mammalogy 87:171–179.

Genz A., Azzalini A. 2013. mnormt: The multivariate normal and t distributions.

Goloboff P.A., Catalano S.A. 2011. Phylogenetic morphometrics (II): Algorithms for landmark optimization. Cladistics 27:42-51.

Gómez J.M., Perfectti F., Klingenberg C.P. 2014. The role of pollinators in the evolution of corolla shape integration in a pollination-generalist plant clade. Philosophical Transactions of the Royal Society of London B Biological Sciences 369:20130257.

González-José R., Escapa I., Neves W.A., Cúneo R., Pucciarelli H.M. 2008. Cladistic analysis of continuous modularized traits provides phylogenetic signals in Homo evolution. Nature 453:775-778.

González-José R., Escapa I., Neves W.A., Cúneo R., Pucciarelli H.M. 2011. Morphometric variables can be analyzed using cladistic methods: A Reply to Adams et al. Journal of Human Evolution 60:244-245.

Goswami A., Smaers J.B., Soligo C., Polly P.D. 2014. The macroevolutionary consequences of phenotypic integration: from development to deep time. Philosophical Transactions of the Royal Society of London B Biological Sciences 369:20130254.

Hillis D.M., Huelsenbeck J.P., Cunningham C.W. 1994. Application and accuracy of molecular phylogenies. Science 264:671–677.

Huelsenbeck J.P. 1995. Performance of phylogenetic methods in simulation. Systematic Biology 44:17–48.

Huelsenbeck J.P., Hillis D.M. 1993. Success of phylogenetic methods in the four-taxon case. Systematic Biology 42:247-264.

Hunt G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proceedings of the National Academy of Sciences of the USA 104:18404–18408.

Jamniczky H.A., Hallgrímsson B. 2009. A comparison of covariance structure in wild and laboratory muroid crania. Evolution 63:1540–1556.

Kalinka A.T., Varga K.M., Gerrard D.T., Preibisch S., Corcoran D.L., Jarrells J., Ohler U., Bergman C.M., Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. Nature 468:811–814.

Kendall D.G., Barden D., Carne T.K., Le H. 1999. Shape and shape theory. Chichester, Wiley.

Klingenberg C.P. 2008. Morphological integration and developmental modularity. Annual Review of Ecology, Evolution and Systematics 39:115-132.

Klingenberg C.P. 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. Hystrix 24:43–58.

Klingenberg C.P., Duttke S., Whelan S., Kim M. 2012. Developmental plasticity, morphological variation and evolvability: a multilevel analysis of morphometric integration in the shape of compound leaves. Journal of Evolutionary Biology 25:115–129.

Klingenberg C.P., Gidaszewski N.A. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Systematic Biology 59:245-261.

Klingenberg C.P., Marugán-Lobón J. 2013. Evolutionary covariation in geometric morphometric data: analyzing integration, modularity and allometry in a phylogenetic context. Systematic Biology 62:591–610.

Klingenberg C.P., Monteiro L.R. 2005. Distances and directions in multidimensional shape spaces: Implications for morphometric applications. Systematic Biology 54:678-688.

Kruuk L.E.B., Slate J., Wilson A.J. 2008. New answers for old questions: the evolutionary quantitative genetics of wild animal populations. Annual Review of Ecology, Evolution and Systematics 39:525–548.

Lande R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

Lockwood C.A., Kimbel W.H., Lynch J.M. 2004. Morphometrics and hominoid phylogeny: Support for a chimpanzee-human clade and differentiation among great ape subspecies. P Natl Acad Sci USA 101:4356-4360.

Lynch M. 1989. Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation. Evolution 43:1-17.

Macholán M. 2006. A geometric morphometric analysis of the shape of the first upper molar in mice of the genus *Mus* (Muridae, Rodentia). Journal of Zoology (London) 270:672–681.

MacLeod N. 2002. Phylogenetic signals in morphometric data. In: MacLeod N., Forey P. L. editors. Morphology, shape and phylogeny. London, Taylor & Francis, p. 100–138.

Maddison W.P. 1991. Squared-change parsimony reconstructions of ancenstral states for continuous-valued characters on a phylogenetic tree. Systematic Zoology 40:304-314.

Marcus L.F., Hingst-Zaher E., Zaher H. 2000. Application of landmark morphometrics to skulls representing the orders of living mammals. Hystrix 11:27–47.

Mardia K.V., Kent J.T., Bibby J.M. 1979. Multivariate analysis. London, Academic Press.

Martín-Serra A., Figueirido B., Palmqvist P. 2014. A three-dimensional analysis of morphological evolution and locomotor performance of the carnivoran forelimb. PLoS ONE 9:e85574.

McArdle B., Rodrigo A.G. 1994. Estimating the ancestral states of a continuous-valued character using squared-change parsimony: An analytical solution. Systematic Biology 43:573-578.

Monteiro L.R. 2000. Why morphometrics is special: The problem with using partial warps as characters for phylogenetic inference. Systematic Biology 49:796-800.

Monteiro L.R., Bonato V., dos Reis S.F. 2005. Evolutionary integration and morphological diversification in complex morphological structures: Mandible shape divergence in spiny rats (Rodentia, Echimyidae). Evolution & Development 7:429–439.

Monteiro L.R., Nogueira M.R. 2011. Evolutionary patterns and processes in the radiation of phyllostomid bats. Bmc Evolutionary Biology 11:137.

Naylor G.J.P. 1996. Can partial warp scores be used as cladistic characters? In: Marcus L. F., Corti M., Loy A., Naylor G. J. P., Slice D. E. editors. Advances in morphometrics. New York, Plenum Press, p. 519–530.

Oakley T.H., Cunningham C.W. 2000. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. Evolution 54:397–405.

Olson E.C., Miller R.L. 1958. Morphological integration. Chicago, University of Chicago Press.

Pavlicev M., Cheverud J.M., Wagner G.P. 2009. Measuring morphological integration using eigenvalue variance. Evolutionary Biology 36:157–170.

Philippe H., Zhou Y., Brinkmann H., Rodrigue N., Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. Bmc Evolutionary Biology 5:50.

Piras P., Colangelo P., Adams D.C., Buscalioni A., Cubo J., Kotsakis T., Meloro C., Raia P. 2010. The *Gavialis-Tomistoma* debate: the contribution of skull ontogenetic allometry and growth trajectories to the study of crocodylian relationships. Evolution & Development 12:568-579.

Polly P.D. 2004. On the simulation of the evolution of morphological shape: multivariate shape under selection and drift. Palaeontologia Electronica 7:7A.

R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Rácz G.R., Demeter A. 1998. Character displacement in mandible shape and size in two species of water shrews (*Neomys*, Mammalia: Insectivora). Acta Zoologica Academiae Scientiarum Hungaricae 44:165–175.

Rannala B., Yang Z. 2008. Phylogenetic inference using whole genomes. Annual Review of Genomics and Human Genetics 9:217–231.

Rifkin S.A., Kim J., White K.P. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nature Genetics 33:138-144.

Rohlf F.J. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny. Systematic Biology 47:147-158.

Sallan L.C., Friedman M. 2012. Heads or tails: staged diversification in vertebrate evolutionary radiations. Proceedings of the Royal Society of London, B Biological Sciences 279:2025–2032.

Sanger T.J., Mahler D.L., Abzhanov A., Losos J.B. 2012. Roles for modularity and constraint in the evolution of cranial diversity among *Anolis* lizards. Evolution 66:1525–1542.

Santana S.E., Lofgren S.E. 2013. Does nasal echolocation influence the modularity of the mammal skull? Journal of Evolutionary Biology 26:2520–2526.

Sherratt E., Gower D.J., Klingenberg C.P., Wilkinson M. 2014. Evolution of cranial shape in caecilians (Amphibia: Gymnophiona). Evolutionary Biology:in press.

Sidlauskas B. 2008. Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. Evolution 62:3135–3156.

Smith H.F. 2011. The role of genetic drift in shaping modern human cranial evolution: a test using microevolutionary modeling. International Journal of Evolutionary Biology 2011:145262.

Smith U.E., Hendricks J.R. 2013. Geometric morphometrics character suites as phylogenetic data: Extracting phylogenetic signal from gastropod shells. Systematic Biology 62:366-385.

Stayton C.T. 2008. Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J Theor Biol 252:1-14.

Steppan S.J., Phillips P.C., Houle D. 2002. Comparative quantitative genetics: evolution of the **G** matrix. Trends in Ecology & Evolution 17:320–327.

Stuart Y.E., Losos J.B. 2013. Ecological character displacement: glass half full or half empty? Trends in Ecology & Evolution 28:402-408.

Swiderski D.L., Zelditch M.L., Fink W.L. 1998. Why morphometrics is not special: Coding quantitative data for phylogenetic analysis. Systematic Biology 47:508-519.

Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M. 1996. Phylogenetic inference. In: Hillis D. M., Moritz C., Mable B. K. editors. Molecular systematics. Sunderland, MA, Sinauer, p. 407–514.

Uddin M., Wildman D.E., Liu G., Xu W., Johnson R.M., Hof P.R., Kapatos G., Grossman L.I., Goodman M. 2004. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. Proceedings of the National Academy of Sciences of the USA 101:2957–2962.

von Cramon-Taubadel N., Smith H.F. 2012. The relative congruence of cranial and genetic estimates of hominoid taxon relationships: Implications for the reconstruction of hominin phylogeny. Journal of Human Evolution 62:640-653.

Wägele J.W., Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. Bmc Evolutionary Biology 7:147.

Wagner G.P. 1984. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: evidence for a nonrandom organization of quantitative character variation. J. Math. Biol. 21:77–95.

Wagner G.P., Pavlicev M., Cheverud J.M. 2007. The road to modularity. Nature Reviews Genetics 8:921-931.

Watanabe A., Slice D.E. 2014. The unitility of cranial ontogeny for phylogenetic inference: a case study in crocodylians using geometric morphometrics. Journal of Evolutionary Biology 27:doi: 10.1111/jeb.12382.

Webster A.J., Purvis A. 2002. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. Proceedings of the Royal Society of London, B Biological Sciences 269:143–149.

Wiens J.J., Hollingsworth B.D. 2000. War of the iguanas: Conflicting molecular and moprhological phylogenies and long-branch attraction in iguanid lizards. Systematic Biology 49:143-159.

Willis J.H., Coyne J.A., Kirkpatrick M. 1991. Can one predict the evolution of quantitative characters without genetics? Evolution 45:441–444.

Zelditch M.L., Fink W.L., Swiderski D.L. 1995. Morphometrics, homology, and phylognetics: Quantified characters as synapomorphies. Systematic Biology 44:179-189.

Zelditch M.L., Fink W.L., Swiderski D.L., Lundrigan B.L. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny: A reply to Rohlf. Systematic Biology 47:159-167.

Zelditch M.L., Swiderski D.L., Sheets H.D., Fink W.L. 2004. Geometric morphometrics for biologists: a primer. San Diego, Elsevier.

Chapter 3

Patterns of Integration in Geometric Morphometrics Studies and Their Role in Phylogenetics

CEFERINO VARÓN GONZÁLEZ [1], SIMON WHELAN[1,2]

AND CHRISTIAN PETER KLINGENBERG[1]

[1]*Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, United Kingdom;*

[2]*Dept. of Evolutionary Biology, EBC, Uppsala University, Norbyägen 18D, 75236 Uppsala, Sweden*

Abstract

Integration is a common feature in shape data, which produces concentration of the variation in some directions of the shape space. It plays an important role in the diversification of species and has been proposed as one enhancer of convergence between species. In this study, we analyse how strong patterns of evolutionary integration are in empirical data. We fit some simple models (linear, exponential, logistic and Gompertz) to the decrease of variation in the principal components of the independent contrasts of shape. Then, we look at the coefficients of the best fitting function so the strength of decrease can be assessed, used as an estimate of integration. We take into account the effect of allometry and the use of weighted squared-change parsimony in our estimates. Finally, we study the effect of different patterns of evolutionary integration on convergence using computer simulations. The empirical datasets we analyse reveal strong patterns of integration in a wide sample of structures and organisms, the exponential model being the function that fits best. Allometry, as expected, is one present integrating factor in many of the studies. These results confirm the important role of strong patterns of evolutionary integration in promoting convergence between species with similar directions of evolution in the phenotypic space. However, our simulations show theoretical situations in which strong patterns of integration can also prevent convergence. That is the case when the directions of the major axes of variation are perpendicular or when the amount of evolution is big relative to the distance between the species.

3.1. INTRODUCTION

Integration is a fairly well studied feature in geometric morphometrics. It is defined as the covariation between different parts of one structure (Olson and Miller 1958, Wagner and Altenberg 1996, Wagner, et al. 2007, Klingenberg 2013, Goswami, et al. 2014). This covariation between different parts of the morphological structures is a product of many different biological processes acting at an earlier level of expression, such as genetic (pleiotropy, linkage desequilibrium), developmental and environmental mechanisms (Klingenberg 2014). These processes shape variation in organisms producing concentration in some specific directions of shape space. This differential variation in different directions of shape space in species has been defined many times as a constraint to evolution, based on the fact that there are directions of shape space in which evolution is more likely (Armbruster, et al. 2014, Goswami, et al. 2014). However, the directions of the shape space in which most variation is concentrated may facilitate the evolution along these directions (Schluter 1996, Renaud, et al. 2006, Armbruster, et al. 2014, Goswami, et al. 2014).

One effect of evolutionary integration, i. e. the concentration of variation along specific directions during the evolution of the species, is a reduction of dimensionality and therefore a high probability of homoplasy (Goswami, et al. 2014) (Chapter 1). In the extreme, integration would approach the unidimensional problem, where homoplasy has been theoretically shown (Lynch 1989). Integration would then cause problems in phylogenetic reconstruction, where convergence causes distantly-related species to be closer in shape space than anyone of them to their sister nodes (Chapter 1). However, we can theoretically expect some changes in the patterns of integration during the diversification of the species. Heterochrony, for example, can cause the adults of a species to evolve under a pattern of

integration followed by the infants of other species (Goswami, et al. 2014). Unfortunately, these changes have been shown to be small between close-related species (Goswami, et al. 2014).

Different approaches have been used in the literature to assess the degree of integration in a structure, or from a different perspective, the presence or absence of modularity (Klingenberg 2013). Partial least squares, matrix correlation and the ordination of covariance matrices are some of them (Klingenberg 2013). In this study we will focus just on the analysis of the variance in the eigenvalues obtained from the principal component analysis (Wagner 1984, Young 2006, Pavlicev, et al. 2009). Many studies analysing integration have used correlation matrices between the different variables, normally length measurements (Haber 2011). However, while the magnitude of the relationship between different parts of the shape may be discarded in traditional morphometrics, this is less appropriate when dealing with geometric morphometrics, since computing the correlation matrix distorts the scale in the shape space. Therefore, the rotation of the configurations during the Procrustes superimposition influences the eigenvalues of the correlation matrix. Exploring the covariance matrix is a better option, which has been implemented in a number of studies (Hallgrímsson, et al. 2009, Jamniczky and Hallgrímsson 2009, Ivanović and Kalezić 2010, Gonzalez, et al. 2011, Jojić, et al. 2011, Gómez-Robles and Polly 2012).

Here we analyse the patterns of evolutionary integration in different datasets and run some simulations to explain the effect of integration on the probability of convergence between two clades. We fit simple mathematical models to empirical measures of the degree of evolutionary integration, so a general picture about this feature in morphometric data is obtained. This will help to obtain a general view about the strength of integration in nature and to understand the variation in the evolution of shape. The effect of evolutionary allometry and the information about branch lengths in the

phylogenies over integration are also analysed. Then, some simulations will explain the effect of different patterns of integration over the probability of convergence between two clades in a two-dimensional space. In those situations in which the probability of convergence is very low phylogenies could be built reliably.

3.2. Materials and Methods

3.2.1. Patterns of evolutionary integration in empirical data

The patterns of evolutionary integration are inferred using morphometric data of a group of species and a phylogeny that represents the evolutionary relationship between them. Independent contrasts of shape can then be obtained (Felsenstein 1985, Rohlf 2001, Klingenberg and Marugán-Lobón 2013). The independent contrasts recover the pattern of evolutionary divergence across the clade under study. This is a scaled version of the pattern of integration of each species under the assumption of evolution by drift, as predicted by the quantitative genetics theory (Lande 1979). Here we infer the pattern of evolutionary phenotypic integration by obtaining the variance in the eigenvalues obtained from a principal component analysis (PCA) of the independent contrasts of shape.

We study a set of empirical datasets obtained from 16 different studies using geometric morphometrics (Klingenberg and Gidaszewski 2010, Álvarez, et al. 2011, Abe and Lieberman 2012, Brusatte, et al. 2012, Foth, et al. 2012, Klingenberg, et al. 2012, Álvarez, et al. 2013, Foth and Rauhut 2013, Klingenberg and Marugán-Lobón 2013, Baab, et al. 2014, Watanabe and Slice 2014), collected from Dryad using the keywords 'geometric morphometrics phylogen*' and unpublished studies (Table 3.1). These are all the studies uploaded to the repository in which all the morphometric raw data and a phylogeny were available. Unpublished data from the authors of this study is

also included. In those studies in which there are different structures analysed (Klingenberg, et al. 2012) the morphometric data of each one is treated separately (Table 3.1). The same applies to the studies (Foth, et al. 2012, Watanabe and Slice 2014) in which two possible phylogenies are considered, so calculations are repeated using both phylogenies. We also do separate calculations for the weighting process in the phylogenies that contain branch lengths information. By estimating the integration patterns using branch-length information and not using it we can check the importance of this decision of the researchers in the estimation of evolutionary integration. From each of these datasets two datasets are treated separately: one where the branch lengths of the phylogeny are the same and one where the branches are different. Note that in some of these studies the dimensionality is higher than the species sample size, which results in an effective dimensionality of n-1 (Table 3.1). In symmetric structures just the symmetric component is used for the analyses, so the dimensionality is reduced.

First author	Year	Structure	S	d
Klingenberg	2010	Drosophila wings	9	8
Benítez*	2014	Drosophila wings	59	26
Klingenberg	2012	Whole leaf	20	18
		Distal leaflet	20	12
		Lateral leaflet	20	12
Foth	2012	Pterosauria skulls	31	29
Foth	2013	Theropods skulls	41	36
		Paraves skulls	14	13
Klingenberg*		Beetles heads	271	13
		Beetles pronotums	290	6
Klingenberg*		Waterfowls	21	19

Klingenberg	2013	Birds skulls	160	18
Álvarez	2013	Rodents	24	23
Álvarez	2011	Rodents	17	16
Varón-González*		Serine proteases	27	26
		Active site	27	5
		Substrate-specificity area	27	25
Varón-González*		Serine proteases	29	28
		Active site	29	5
		Substrate-specificity	29	19
Brusatte	2012	Dinosaurs skulls	35	34
Abe	2012	Trilobites	61	31
Baab	2014	Lemurs	33	32
Watanabe	2014	Crocodiles	10	9

Table 3.1. *List of the studies that have been analyzed. In the first column it is written the first author of the study and in the second column the year. The studies marked with an * are unpublished. The third column reflects the structure analyzed. The four and fifth columns give the species sample size and the dimensionality of the dataset.*

Evolutionary allometry has been suggested to be an integrating factor (Klingenberg 2013). The effect of size over shape usually follows a linear function, so it causes concentration in one specific direction of the shape space, i. e. integration. The strength of this effect depends on how much allometric variation there is in comparison to the rest of variation (Klingenberg 2013). In order to study its integrative effect, in this study we

also decompose the datasets in which we find allometry in: a) the original dataset, b) a dataset with the residuals of the regression of the independent contrasts of shape over the independent contrasts of centroid size and c) a dataset with the residuals of the regression of the independent contrasts of shape over the independent contrasts of the log-transformed centroid size. Note that in those datasets in which the dimensionality is higher than the species sample size these residuals looses one extra dimension. In total, 71 sets of PC relative variances in each PC are obtained using MorphoJ (Klingenberg 2011).

We use four models that can recover the decreasing in variation in each PC (Figure 3.1), using only those dimensions that contain some variation. We use a least-squares procedure to fit the functions of these models to the empirical values and to obtain the error of each fitting. The function with the least error is chosen as the best-fitting model. The model choice is also done using the corrected version of the Akaike's information criterion (AICc) (Akaike 1973, Sugiura 1978). This corrected version is less prone to be biased due to small sample sizes (Hurvich 1991).

The first function we use is the linear model ($y=bx+a$), i.e. the simplest model, being y the relative amount of variation in each PC and x each PC. We also test the exponential model ($y=e^{a}e^{bx}$), in which the decreasing of variation in each PC is much faster in the first PCs for the same b than the linear model and therefore integration would be stronger. Then, two more models are tested based on the exponential model: a logistic model ($y=100e^{a}e^{bx}/(1+e^{a}e^{bx})$) and a Gompertz model ($y=100e^{-e(a+bx)}$) in which the variation in the first PCs approach an asymptote of 100% (Figure 1). The Gompertz model also allows the decreasing of variation to be smoother in the last PCs for the same b than in the logistic model, so it represents a model in which there is less integration. The logistic and the Gompertz curves add one flatter part (that approaching the asymptote of 100%) to the exponential

model. They are more complex and represent weaker integration models than the exponential for the same *b*, since the decay in variation is more gradual in the first PCs in these parts and more abrupt in the case of the last PCs in the case of the Gompertz model (Figure 3.1). In all the models the parameter *a* is related to the amount of the variation (that is very similar, since the relative amount of variance always varies from 0 and adds up to 100) and the parameter *b* the strength of the decay in variation. We look at the latter to study the strength of integration. As we measure the decay of variation and therefore this parameter is negative, the bigger its absolute value is the more integration there is. We use this measure to check the effect of using branch-length information in the estimation of the independent contrasts and the removal of the allometric effect over integration.

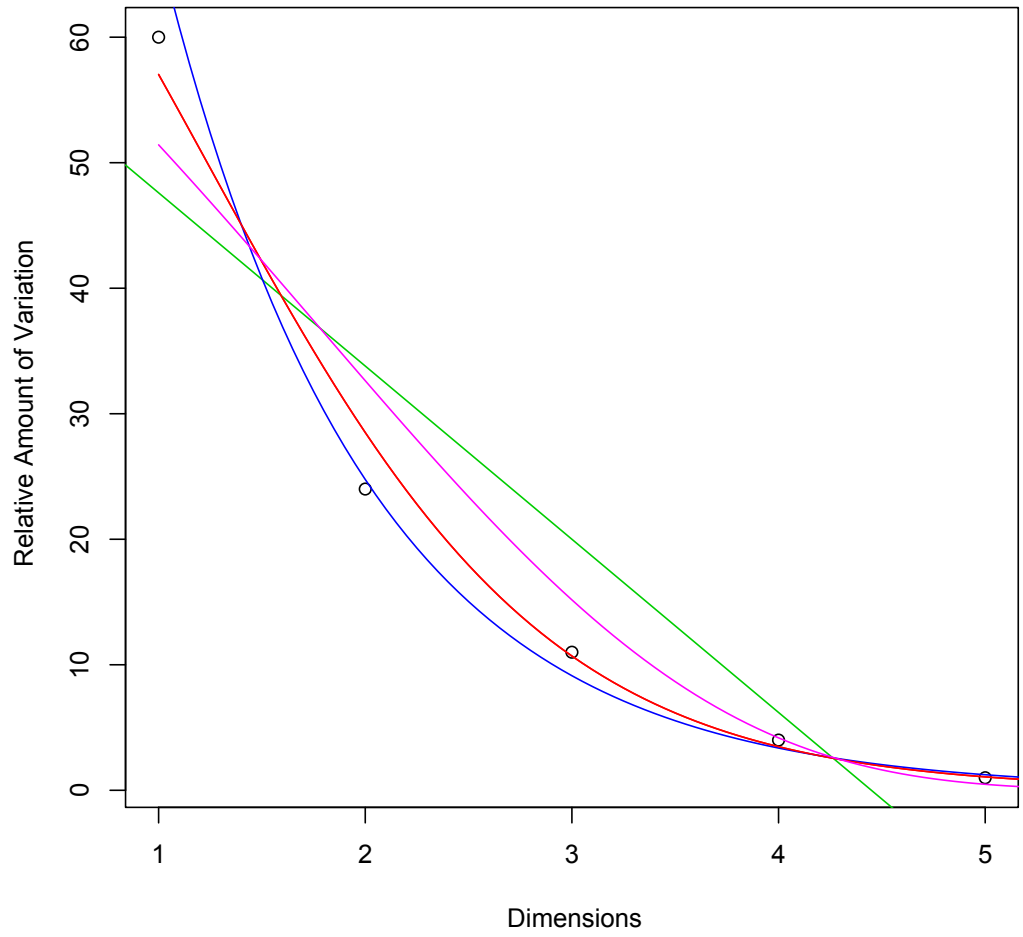

***Figure 3.1**. Example of fitting with five different dimensions of variation. In the x axis there is the numbering of five principal components. In the y axis the relative amount of variation is represented. The white dots represent the relative amount of variation in each principal component obtained from the independent contrasts of shape in an empirical study. Each line represents the best fitting function for each kind of model (green - lineal model, blue - exponential model, red - logistic model, pink - Gompertz model).*

In order to control for the number of clades and dimensionality we use simulations. We create a sample of species distributed in a n-dimensional

shape space supposing no integration at all. Then, we run a PCA and we check which function fits better the decreasing of variation in the PCs. Replicating the process 1000 times we obtain a percentage of preferred models for a specific combination of dimensionality and sample size. We test the combinations of dimensionality and species sample size observed in the empirical datasets (Table 3.1). We check the probability of obtaining those functions in the absence of integration for the same combinations of dimensionality and sample size as in the empirical studies. If there is a high probability of obtaining a specific function for a combination of dimensionality and sample size and we obtained that function in the dataset with those conditions, they may be affecting our estimations about the best-fitting function.

3.2.2. Integration and probability of convergence

Once the pattern of evolutionary integration is studied in empirical studies, we study the theoretical relationship between the strength of integration and convergence using simulations. This will give an estimate of the impact of the patterns of integration over the reliability of the phylogenetic inference in empirical studies. This can also explain patterns of diversification in shape. The patterns of integration obtained in the empirical datasets reflect the pattern of integration during the evolution of each species when evolution happens by drift, as predicted by the quantitative genetic theory (Lande 1979). We can simulate the evolution of just two species in a phenotypic space in presence of different patterns of integration and look at the probability of them to converge.

We set two species one unit of distance far away in a bivariate phenotypic space. We fix one of them and let the other one evolve according to a Brownian motion model of evolution. We simulate three different patterns of integration, in which one dimension accounts for 95%, 80% and 60% of

the variation and the other dimension take up the rest. If after the evolution the Euclidean distance between the two clades has decreased then convergence has happened (Figure 3.2). Replicating this process 1000 times we obtain a percentage, our estimation of the probability of convergence.

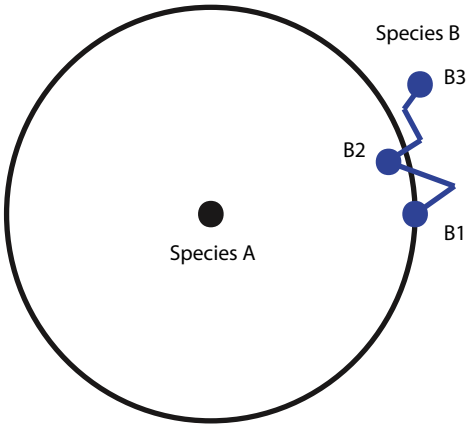

***Figure 3.2**. Simulations in which the effect of the different variances, patterns of integration and orientations over the probability of homoplasy between two clades (species A and B) is studied. Species A is fixed and Species B, starting in B1, evolves according to a Brownian motion model of evolution. If the distance between A and B has decreased after the evolution (e.g. if the evolution of B ends up in B2), then convergence has happened. If the distance increases (e. g. if the evolution of B ends up in B3), then convergence has not happened. In this example the orientation between the major axis of evolution of B and the axis connecting both species at the beginning of the simulation is 90 degrees.*

In addition to the strength of the integration pattern, there are two more features that can influence the probability of convergence between two clades and that we take into account in our simulations. The first of them is the orientation of the pattern of evolutionary integration in relation to the fixed clade. This is defined as the angle between the axis connecting both clades at the beginning of the simulation and the major axis of variation in the evolving clade. An angle of 0 degrees represents the situation in which the evolving clade moves mainly in the axis that connects both clades. A similar situation (but less extreme) is expectable in closely-related species, so

the area of the shape space in which they have evolved is mostly the same. In our simulations, we use orientations of 0, 20, 45, 60 and 90 degrees.

The third important feature to estimate the probability of convergence is the ratio between the total variance in the evolution of the clade and the distance between the two clades at the beginning of the simulations. Since this distance is set to one, the ratio is equal to the total variance in the evolution. Under the Brownian motion model of evolution, the total variance is determined by the multiplication of the rate of evolutionary change and the time this clade is evolving (Felsenstein 2004). We run the simulations using total variances of 0.1, 0.5, 1, 1.5 and 2.

We simulate the evolution of the clade by obtaining a random vector from a bivariate normal distribution with mean (0,0), i.e. the starting point, and a covariance matrix defined as the multiplication of those total variances times a diagonal matrix with the relative amount of variance in each dimension. This random vector is then multiplied by a rotation matrix, which gives the orientation. The result is a vector, which represents the position of the evolved clade. R 3.0.2 (R Core Team 2013) has been used for both the fitting process and the simulations. The package MuMIn has been used for the AICc estimation of each model.

3.3. RESULTS

3.3.1. Patterns of evolutionary integration in empirical data

The exponential function is the model that best fits the majority of patterns of integration, in 60 out of 71 datasets under the least-squares criterion and in 62 out of 71 datasets using the AICc (Table 3.2). Using the AICc, the logistic function is chosen as the best model for 8 of the remaining datasets and the Gompertz function for one. Using the least-squares criterion, the logistic function is found as the best model in 9 of the remaining cases.

The Gompertz curve is chosen as the best function in the other 2 datasets. The coefficient of decay in the exponential functions varies, for base e, from -1.004 when the distal leaflet of Potentilla is analysed to -0.183 in a group of serine proteases, very close to the coefficient of -0.186 of the South American rodents' cranium (Fig. 3.3). For the 60 datasets in which the exponential model is the preferred, the average coefficient of decay of variation is -0.428. The median is -0.352, a value between the coefficients obtained for the Pterosauria skulls using two different phylogenies.

First author	Structure	Least-squares Best model	% Exp fun.	Coeff. of decay	AICc best model
Klingenberg	Drosophila wings	Exponential	1.40	-0.56	Exponential
Benítez	Drosophila wings	Exponential	1.2	-0.27	Exponential
Klingenberg	Whole leaf	Exponential*	0	-0.56	Exponential*
	Distal leaflet	Exponential*	7.5	-1.00	Exponential
	Lateral leaflet	Exponential	7.5	-0.93	Exponential
Foth	Pterosauria skulls	Exponential	0	-0.35	Exponential*
Foth	Theropods skulls	Exponential	0	-0.21	Exponential
	Paraves skulls	Exponential	0.1	-0.39	Exponential*
Klingenberg	Beetles	Exponential	38.9	-0.34	Exponential

	heads				
	Beetles pronotums	Exponential	48.7	-0.61	Logistic
Klingenberg	Waterfowls	Exponential	3.2	-0.28	Exponential
Klingenberg	Birds skulls	Exponential	31.5	-0.23	Exponential
Álvarez	Rodents	Exponential	0	-0.19	Exponential
Álvarez	Rodents	Exponential	0	-0.38	Exponential
Varón-González	Serine proteases	Exponential	0	-0.18	Logistic*
	Active site	Exponential	37.8	-0.63	Exponential
	Substrate-specificity area	Exponential	0	-0.38	Exponential
Varón-González	Serine proteases	Logistic*	0	-	Exponential
	Active site	Logistic*	40.3	-	Exponential
	Substrate-specificity	Exponential	0.6	-0.38	Exponential
Brusatte	Dinosaurs skulls	Exponential	0	-0.23	Logistic*
Abe	Trilobites	Exponential	0	-0.28	Exponential
Baab	Lemurs	Exponential	0	-0.19	Exponential
Watanabe	Crocodiles	Logistic	1.20%	-	Exponential

***Table 3.2**. Results of the model-fitting process. The first two columns represent again the first author and the structure analyzed. The best fitting function using the least-squares criterion is mentioned in the third column. The next column gives the probability of obtaining a logarithmic model as the*

*best fitting function in our simulations using the same ratio sample/dimensionality in absence of integration. The fifth column gives the coefficient of decreasing of variation for the raw data in each structure. The last column gives the selected method by AICc. Those datasets in which the preferred model changes using weighted squared-change parsimony or allometric residuals are labeled with an *.*

There is no systematic effect of the weighted version of the squared-change parsimony on this coefficient of integration. In three structures the weighted version increases the integration while in seven of them it decreases it. When we check for the effect of allometry over integration, we find that it is hidden by the fact that in some of the datasets the number of dimensions is in principle bigger than the number of species (Fig. 3.3). This reduces the dimensionality. The coefficients in those datasets reveal a higher integration in the datasets where the effect of allometry has been removed. In most of the rest of the datasets, where the number of species is bigger than the dimensionality and therefore removing the effect of allometry does not remove any dimension, the coefficient of decay of variation is smaller when the effect of allometry is not removed. The exceptions are the theropods skulls, where the coefficient varies from -0.210 when no allometry is taken into account to -0.214 when the effect of allometry is removed using the logarithmic centroid size, and the substrate specificity area of one group of serine proteases, where the coefficient varies from -0.534 to -0.552 (Table 3.2, Fig. 3.3).

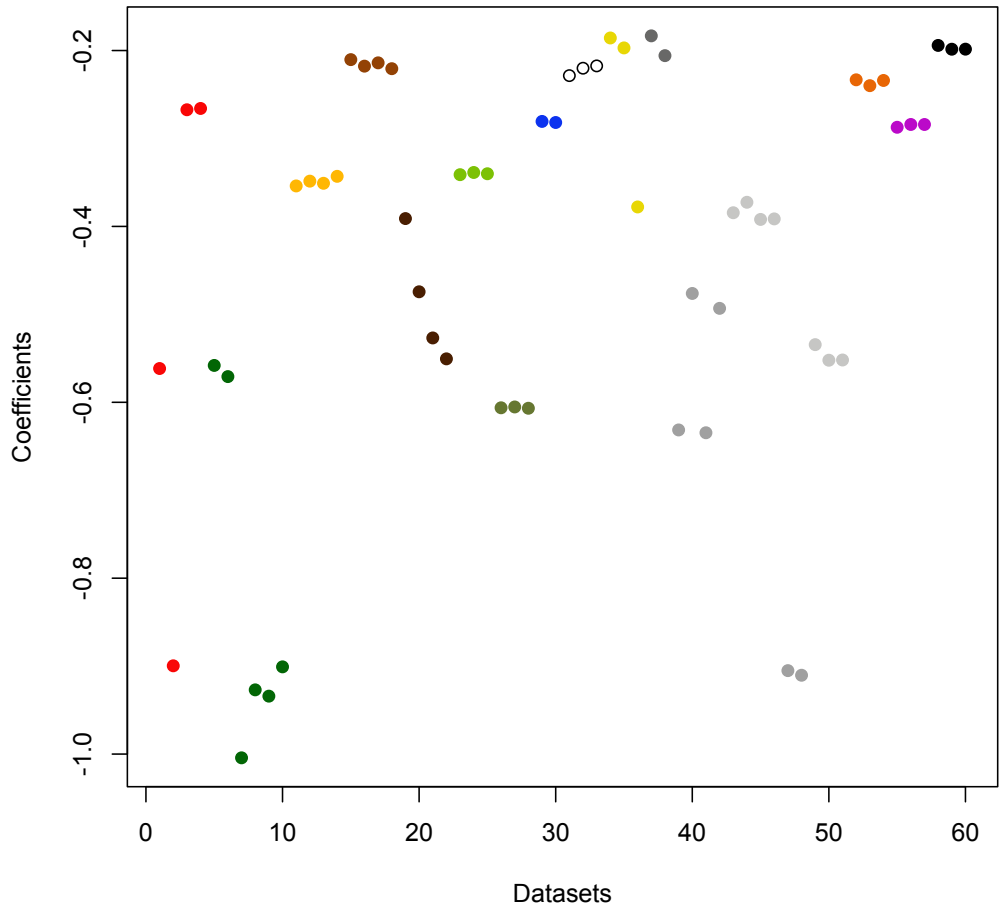

Figure 3.3. *Representation of the different coefficients of decay of variation in those studies in which the exponential model has been chosen. In the x axis the 60 different datasets are represented. The y axis represents the coefficients. Different colours represent different structures. Clusters are obtained in many cases: these are different datasets where the same structure is analysed and the variation in the coefficient is very small. For example, when the same structure is analysed using the raw data, the allometric residuals for the centroid size and the allometric residuals for the log-transformed centroid size. Note that in some of these cases one extra dimension is lost when the allometric residuals are used, so the coefficients are not entirely comparable.*

In our first set of simulations, in which we estimate the best-fitting curves for the different combinations of dimensionality and species sample size in absence of integration, we find that the probability of finding the logarithmic model as the best model is the highest when the taxa sample is much bigger than the dimensionality (Table 3.2). There are five datasets that have a sample size/dimensionality ratio with which the probability of obtaining the exponential model in absence of integration is significant. In one of these datasets it was found during the fitting process that the logistic model fits the best, so no artefact due to the sample size/dimensionality ratio seem to be acting. In the beetle pronotum we find that in absence of integration the linear model is almost as likely to be found due to the sample size/dimensionality ratio (41%). However, during the fitting of the empirical values of variation in each PC we found that the difference in the fitting errors between these two models (linear and exponential) is substantial. Therefore, the results found should not be due to the ratio sample size/dimensionality. In the case of the bird skulls the Gompertz function is much more favoured (42%). Just in the beetle heads and one of the active site datasets there are reasons to think that the fitting process may have favoured the function obtained independently of the integration pattern.

3.3.2. Probability of convergence in the two-dimensional space

The probability of convergence between two species in a bivariate space approximates to 50% when the ratio variance/distance between the taxa decreases. The evolving species changes so little that the small amount of evolution towards the fixed species or away from it determines whether there is convergence or not. The probability of convergence in this case depends only on the little component of the variation that is towards the other clade or away from it. Big variance always avoids convergence: because the taxa are very close relative to expected evolution at the start, it is unlikely

that they end up even closer. The evolving taxon always ends up at a bigger distance no matter the direction of the change (Figure 3.4). When integration is taken into account (Tables 3.3-3.5) the probability of convergence increases with the degree of integration if the angle between the evolving species and the vector between both species at the beginning is low. The evolution of the clade towards the other or away from it determines the appearance or not of convergence. When the angle is 0 degrees and the pattern of integration is extreme, the probability of convergence tends to 50%. In this case, that there would be variation just in one dimension, the evolving clade could move just towards the fixed clade or away from it. If the angle between the evolutionary trajectories is close to ninety degrees then the probability of convergence decreases with the degree of integration. In this case, there is no evolution towards the fixed species but perpendicular to it and therefore the evolving species can just move away from the other species. When the degree of integration is low, the angle does not matter (Figure 3.4).

var/ori	0	20	45	60	90
0.01	48.7	47.8	51	49	43.3
0.5	49.5	44.7	43.2	36.3	24.4
1	47.7	45.8	34.6	29.4	19.4
1.5	45.9	41.5	34.7	27.2	17.5
2	41.1	36.8	31.5	22.1	18.6

Table 3.3*. Simulations in which a very strong degree of integration is considered (95% of variation in the first dimension and 5% in the second). The rows are the ratios of variance/distance between taxa. The columns are the different orientations of the integration pattern (expressed as the degrees of the angle between the pattern of integration and the vector between both*

taxa at the beginning of the simulations. The results are the proportion of simulations ending up at a lower distance than at the beginning for the combination of integration pattern, orientation of the evolution and total variance. Colour legend: Probability of convergence > 45% (red), >40% (orange), <25% (blue).

var/ori	0	20	45	60	90
0.01	50.2	47.8	51.1	49.4	46.9
0.5	45.5	47.4	42.7	35.5	30
1	39.4	41	33.1	30.8	25.7
1.5	36.3	34.8	28	26.9	21.9
2	33.4	33.1	25.8	23.6	21.7

***Table 3.4**. Simulations in which a strong degree of integration is considered (80% of variation in the first dimension and 20% in the second). The rows are the ratios of variance/distance between taxa. The columns are the different orientations of the integration pattern (expressed as the degrees of the angle between the pattern of integration and the vector between both taxa at the beginning of the simulations. Colour legend: Probability of convergence > 45% (red), >40% (orange), <25% (blue).*

var/ori	0	20	45	60	90
0.01	50.7	47.6	48.9	48.7	50
0.5	43.5	42	41.3	36.8	38.9
1	36.1	34.6	33.9	35.1	33.2
1.5	33.7	32.3	31.6	31.6	27.3
2	27.4	30.3	27.5	27.1	25.5

***Table 3.5**. Simulations in which a weak degree of integration is considered (60% of variation in the first dimension and 40% in the second). The rows are the ratios of variance/distance between taxa. The columns are the different orientations of the integration pattern (expressed as the degrees of the angle between the pattern of integration and the vector between both taxa at the beginning of the simulations. Colour legend: Probability of convergence > 45% (red), >40% (orange), <25% (blue).*

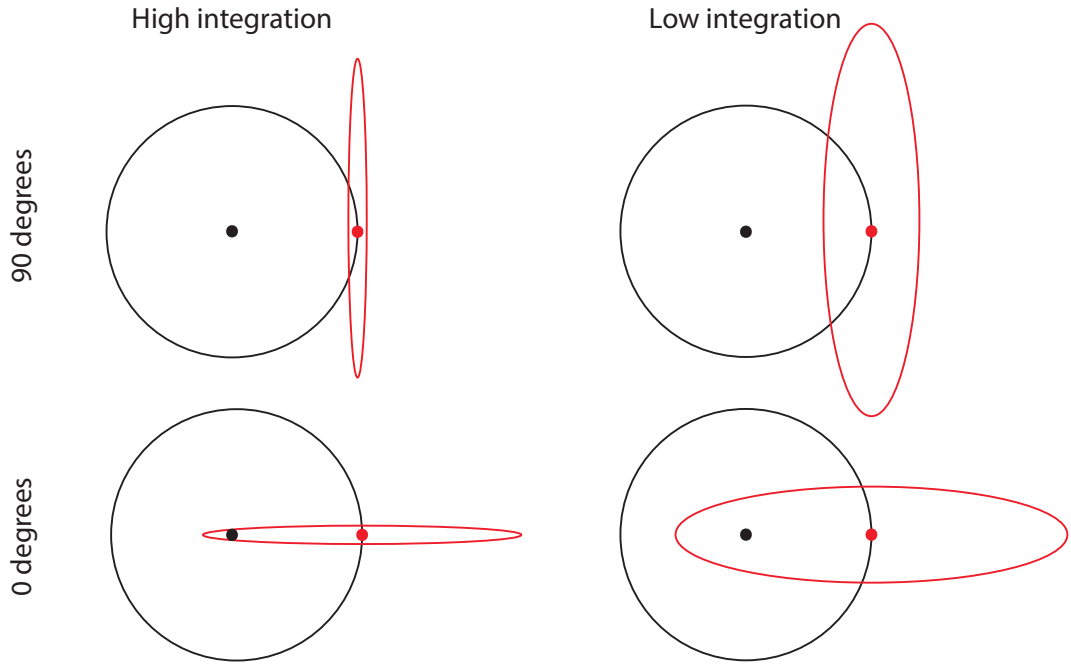

***Figure 3.4.** Illustration of the results of the simulations studying the probability of convergence under different patterns of integration and orientations of the major axis of variation. Each diagram represents one bivariate space in which there are two species separated by one unit of distance, one in the centre of the black centre (with a radius of one unit) and one in the centre of the red ellipse (over the radius of the circle). We let the red species to evolve under a Brownian motion model of evolution and a specific degree of integration and orientation. On the left we see the situations where there is high integration and on the right the situation where*

there is a lower degree of integration. On the top we see the situations where the evolution of the red species follows an orientation of 90 degrees with respect to the axis connecting both species at the beginning. On the bottom the evolution happens under an orientation of 0 degrees. After the evolution the red species will finish in one point within the red ellipse. The area of the red ellipse falling into the black centre determines how likely is convergence to happen between the two species. We can see that in the situation in the bottom left corner (high integration and 0 degrees) 50% of the red ellipse falls into the black centre, so the probability of convergence is bigger than in any other case. In the top left situation (high integration and 90 degrees) the area of the red ellipse falling into the black centre is minimum.

3.4. DISCUSSION

Evolutionary integration is present in all the morphometric data. The decreasing in variation in each direction of the shape space follows an exponential curve in the vast majority of the datasets analysed (Table 3.3). This suggests that shape is a character that evolves mainly in few of the possible directions of the shape space in a vast majority of the structures, since no systematic effect of the ratio sample size/dimensionality has been found (Table 3.3). The degree of integration is smaller in all the datasets in which the effect of evolutionary allometry has been removed and sample size is relatively large, hence the dimensionality does not change, confirming its integrative role. The fact that the variation is concentrated in few dimensions during the evolution of the species is one important feature in the enhancing of the probability of convergence between them when the directions of variation in the evolution of both species is similar and the amount of variation is too. However, patterns of high integration in two species evolving in perpendicular directions would avoid convergence. When the total variance

is big relative to the distance between the species, convergence is very unlikely too.

We have found an overwhelming support for the exponential function as the best function describing the pattern of evolutionary integration in most of the organisms and structures. This function represents the model with the most abrupt decrease in variation in the PCs. Although a small number of studies is analysed, it is the preferred curved in the vast majority of the datasets even when a heterogeneous set of structures are analysed. In addition, the results in the simulations suggest that there are just two structures in which a preference towards the exponential model due to the ratio between dimensionality and number of species may be expected. In absence of integration, the eigenvalues are known to converge when the sample is much bigger than the dimensionality and to be more spread out when they are similar (Johnstone 2001, Bickel and Levina 2008). The logarithmic decrease in variation supposes a strong integration in most of the structures used in geometric morphometrics. Evolution in shape is favoured in just few directions of the shape space, while it is constrained in all the others.

On average, the integration coefficient in the exponential function is -0.4282. That means that each PC accounts for the variation of about the 65% of the variation of the previous PC. There are some extreme cases, like the distal leaflet of the genus Potentilla, in which this coefficient decreases to -1.0042, so the variation in one PC is of about the 35% of the previous one. This is a dramatic decrease and represents the most strong integration pattern of the datasets analysed. On the other extreme, we have the shape of a group of serine proteases and the cranium of a group of South American caviomorph rodents, where each PC is about the 83% of the previous one, so consecutive PCs are fairly similar.

The fact that few studies are used may be of importance in the analysis of the effect of allometry and the weighting process over the degree

of integration, even when the results are coherent with the current knowledge about these features. This is especially remarkable due to the small ratio dimensionality/species sample size used in many morphometric datasets, which can 'hide' their effect in the integration coefficient via the removal of one dimension. Allometry has been proposed to be an integrating factor in morphological structures (Klingenberg 2013). The increase in the gene expression behind the increase in size, for example, may reinforce pleiotropic effects or developmental patterns and therefore enhance integration. In biological structures the effect of size over shape is often close to be linear and therefore 'a considerable proportion of the total shape variation may be concentrated in the direction of the allometric effects' if these are big enough (Klingenberg 2013).

There is a less straightforward relationship between integration and the weighting process during the mapping of the data on the phylogeny. An unweighted mapping of the data on the phylogeny minimizes the distance of the whole phylogeny in the shape space. It places the internal nodes where this distance is minimized and therefore they tend to be around the average position of the terminal nodes. The weighted version of the mapping adds some 'flexibility' to some branches and removes it from others, therefore the position between the internal nodes is less regular. The relationship between the weighting process and the relative amount of variation obtained in each PC depends on the specific case and there is no general effect that can be extracted from our results.

Strong patterns of integration, as the ones we have found in the empirical data, have an effect over the probability of two species to convergence in the shape space. They are therefore implicated in the production of similar phenotypes independently of natural selection (Losos 2011). The main consequence of this is different scenarios for the impact of integration over the reliability of the phylogenetic reconstruction. Integration

constraints the evolution of one clade in a specific direction of the shape space and therefore the probability of convergence depends on the angle between this major axis of variation and the direction of the axis connecting both clades. When the angle between these two axes is low then a high degree of integration increases the probability of convergence and makes the phylogenetic reconstruction fairly unreliable (chapter 3.1), since they would share a unidimensional shape space (Lynch 1989). This situation is expected under our assumption of stability in the pattern of integration in the whole phylogeny, which may be frequent when close related species are studied (Goswami, et al. 2014). Therefore, the phylogenies built using geometric morphometrics and closely related species are in principle less reliable than those using distantly related species (chapter 3.1). It would be in those cases in which the species evolve under different patterns of integration, with different orientations and variations, when a high degree of integration in both of them prevents convergence. In terms of phylogenetic reconstruction, this situation can produce partial success, e.g. when the families of species were successfully clustered but not the species within families (Cardini and Elton 2008). While the integration pattern may be conserved within a family of species, it may be different on a larger scale (Goswami, et al. 2014). Therefore, the probability of convergence would be much higher within clades than among them. There are, however, exceptions to these general patterns, since the pattern of integration can evolve at a smaller phylogenetic scale (Gómez, et al. 2014).

For the simulations we have made some theoretical assumptions that may be controversial, starting with the premise that shape evolves according to a Brownian motion model of evolution, for which there is evidence against (Estes and Arnold 2007). Although it has been shown empirically that the pattern of genetic or phenotypic variation is similar to the pattern of evolutionary divergence (Monteiro, et al. 2005, Drake and Klingenberg 2010,

Smith 2011, Klingenberg, et al. 2012, Goswami, et al. 2014), this expectation may not be fulfilled for all the cases. We have also assumed that the patterns of integration within-species are stable along the whole phylogeny (Steppan, et al. 2002, Arnold, et al. 2008), assumption that is not entirely clear.

In this study we surveyed the degree of integration found in empirical studies and its implications over the evolution of the species. The role of integration in macroevolution has been assessed from many different perspectives lately (Armbruster, et al. 2014, Goswami, et al. 2014, Klingenberg 2014), explaining the effect of this feature over the evolution of the species. However, although our results suggest that a stable pattern of integration at the macroevolutionary level is a source of convergence, other features like the change in the amount of variation or the orientation between evolutionary trajectories still need further study to accurately assess the effect of the concentration of variation in some dimensions over the evolution of the species.

3.5. REFERENCES

Abe FR, Lieberman BS. 2012. Quantifying morphological change during an evolutionary radiation of Devonian trilobites. Paleobiology 38:292-307.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F editors. 2nd International Symposium on Information Theory. Budapest, Akademia Kiado, p. 267-281.

Álvarez A, Perez SI, Verzi DH. 2011. Early evolutionary differentiation of morphological variation in the mandible of South American caviomorph rodents (Rodentia, Caviomorpha). Journal of Evolutionary Biology 24:2687-2695.

Álvarez A, Perez SI, Verzi DH. 2013. Ecological and phylogenetic dimensions of cranial shape diversification in South American caviomorph rodents (Rodentia: Hystricomorpha). Biological Journal of the Linnean Society.

Armbruster WS, Pélabon C, Bolstad GH, Hansen TF. 2014. Integrated phenotypes: understanding trait covariation in plants and animals. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Arnold SJ, Bürger R, Hohenlohe PA, Ajie BC, Jones AG. 2008. Understanding the evolution and stability of the **G**-matrix. Evolution 62:2451–2461.

Baab KL, Perry JMG, Rohlf FJ, Jungers WL. 2014. Phylogenetic, ecological, and allometric correlates of cranial shape in Malagasy lemuriforms. Evolution 68:1450–1468.

Bickel PJ, Levina E. 2008. Regularized estimation of large covariance matrices. The Annals of Statistics 36:199-227.

Brusatte SL, Sakamoto M, Montanari S, Harcourt Smith WEH. 2012. The evolution of cranial form and function in theropod dinosaurs: insights from geometric morphometrics. Journal of Evolutionary Biology 25:365–377.

Cardini A, Elton S. 2008. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. Biological Journal of the Linnean Society 93:813-834.

Drake AG, Klingenberg CP. 2010. Large-scale diversification of skull shape in domestic dogs: disparity and modularity. American Naturalist 175:289–301.

Estes S, Arnold SJ. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. American Naturalist 169:227-244.

Felsenstein J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1-15.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA, Sinauer Associates.

Foth C, Brusatte SL, Butler RJ. 2012. Do different disparity proxies converge on a common signal? Insights from the cranial morphometrics and evolutionary history of Pterosauria (Diapsida: Archosauria). Journal of Evolutionary Biology 25:904-915.

Foth C, Rauhut OWM. 2013. Macroevolutionary and morphofuntional patterns in theropod skulls: A morphometric approach. Acta Palaeontologica Polonica 58:1-16.

Gómez JM, Perfectti F, Klingenberg CP. 2014. The role of pollinator diversity in the evolution of corolla-shape integration in a pollination-generalist plant clade. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Gómez-Robles A, Polly PD. 2012. Morphological integration in the hominin dentition: Evolutionary, developmental, and functional factors. Evolution 66:1024-1043.

Gonzalez PN, Oyhenart EE, Hallgrímsson B. 2011. Effects of environmental perturbations during postnatal development on the phenotypic integration of the skull. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 316B:547-561.

Goswami A, Smaers JB, Soligo C, Polly PD. 2014. The macroevolutionary consequences of phenotypic integration: from development to deep time. Philosophical Transactions of the Royal Society of London B Biological Sciences 369:20130254.

Haber A. 2011. A comparative analysis of integration indices. Evolutionary Biology 38:476-488.

Hallgrímsson B, Jamniczky HA, Young NM, Rolian C, Parsons TE, Boughner JC, Marcucio RS. 2009. Deciphering the palimpsest: Studuying the

relationship between morphological integration and phenotypic covariation. Evolutionary Biology 36:355-376.

Hurvich CM. 1991. Bias of the corrected AIC criterion for underfitted regression and time series models. Biometrika 78:499-509.

Ivanović A, Kalezić ML. 2010. Testing the hypothesis of morphological integration on a skull of a vertebrate with a biphasic life cycle: a case study of the alpine newt. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 314B:527-538.

Jamniczky HA, Hallgrímsson B. 2009. A comparison of covariance structure in wild and laboratory muroid crania. Evolution 63:1540–1556.

Johnstone IM. 2001. On the distribution of the largest eigenvalue in principal components analysis. The Annals of Statistics 29:295-327.

Jojić V, Blagojević J, Vujošević M. 2011. B chromosomes and cranial variability in yellow-necked field mice (*Apodemus flavicollis*). Journal of Mammalogy 92:396-406.

Klingenberg CP. 2011. MorphoJ: an integrated software package for geometric morphometrics. Molecular Ecology Resources 11:353-357.

Klingenberg CP. 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. Hystrix 24:43-58.

Klingenberg CP. 2014. Studying morphological integration and modularity at multiple levels: concepts and analysis. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Klingenberg CP, Duttke S, Whelan S, Kim M. 2012. Developmental plasticity, morphological variation and evolvability: a multilevel analysis of morphometric integration in the shape of compound leaves. Journal of Evolutionary Biology 25:115–129.

Klingenberg CP, Gidaszewski NA. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Systematic Biology 59:245-261.

Klingenberg CP, Marugán-Lobón J. 2013. Evolutionary covariation in geometric morphometric data: analyzing integration, modularity, and allometry in a phylogenetic context. Systematic Biology 62:591–610.

Lande R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

Losos JB. 2011. Convergence, adaptation, and constraint. Evolution 65:1827-1840.

Lynch M. 1989. Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation. Evolution 43:1-17.

Monteiro LR, Bonato V, dos Reis SF. 2005. Evolutionary integration and morphological diversification in complex morphological structures: Mandible shape divergence in spiny rats (Rodentia, Echimyidae). Evolution & Development 7:429–439.

Olson EC, Miller RL. 1958. Morphological integration. Chicago, University of Chicago Press.

Pavlicev M, Cheverud JM, Wagner GP. 2009. Measuring morphological integration using eigenvalue variance. Evolutionary Biology 36:157–170.

R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Renaud S, Auffray JC, Michaux J. 2006. Conserved phenotypic variation patterns, evolution along lines of least resistance, and departure due to selection in fossil rodents. Evolution 60:1701-1717.

Rohlf FJ. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations. Evolution 55:2143-2160.

Schluter D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766-1774.

Smith HF. 2011. The role of genetic drift in shaping modern human cranial evolution: a test using microevolutionary modeling. International Journal of Evolutionary Biology 2011:145262.

Steppan SJ, Phillips PC, Houle D. 2002. Comparative quantitative genetics: evolution of the **G** matrix. Trends in Ecology & Evolution 17:320–327.

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics-Theory and Methods 7:13-26.

Wagner GP. 1984. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: evidence for a nonrandom organization of quantitative character variation. Journal of Mathematical Biology 21:77-95.

Wagner GP, Altenberg L. 1996. Complex adaptations and the evolution of evolvability. Evolution 50:967-976.

Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nature Reviews Genetics 8:921-931.

Watanabe A, Slice DE. 2014. The utility of cranial ontogeny for phylogenetic inference: a case study in crocodylians using geometric morphometrics. Journal of Evolutionary Biology 27:1078-1092.

Young NM. 2006. Function, ontogeny, and canalization of shape variance in the primate scapula. Journal of Anatomy 209:623-636.

Chapter 4

Geometric Morphometrics: A Useful Tool to Study Protein Evolution

CEFERINO VARÓN GONZÁLEZ [1], SIMON WHELAN[1,2], SIMON LOVELL[1]

AND CHRISTIAN PETER KLINGENBERG[1]

[1]*Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, United Kingdom;*

[2]*Dept. of Evolutionary Biology, EBC, Uppsala University, Norbyägen 18D, 75236 Uppsala, Sweden*

Abstract

Shape is a fundamental feature for the correct performance of the proteins, which guarantees the proper interaction of them with their environment. Different methods have been used to study protein structure in studies of evolution. However, the most common technique used to study evolution of shape in anatomical structures, geometric morphometrics, has been used very few times for molecular data. In this study we explain how to implement geometric morphometric methods in protein data to study different aspects of the evolution of shape demonstrated in serine proteases. We test whether there is phylogenetic signal for shape in this family of proteins, as well as the role of the changes in size during the diversification of the family in the patterns of diversification in shape. Then, the relationship between variation in shape and specific functions of the proteins is established. Finally, the relationship between similarity in shape and similarity in sequence is explored. In addition to the complete protein shapes, specific parts of them are studied including the active site and the area of the proteins involved in substrate specificity. The results obtained by applying these methods to a set of serine proteases are congruent with the literature existent: there is phylogenetic signal although some convergence can be identified, evolutionary allometry plays an important role in the diversification of shape and the shape in the area involved in substrate specificity can discriminate between functional groups in these proteins. Geometric morphometrics can be useful in the description and explanation of the evolution of shape in protein studies.

4.1. INTRODUCTION

In the last twenty years there has been a great development of the set of techniques used to analyse shape and size, all of them forming the science known as geometric morphometrics (Adams, et al. 2013). This field has enhanced the study of the evolution of shape in a very wide range of organisms and contexts, from biomechanical performance (O'Higgins and Milne 2013) to sexual selection (Sanger, et al. 2013).

Geometric morphometrics is formed by a set of ideas and statistical techniques that are now well-established (Zelditch, et al. 2012). Shape is defined as all the geometrical features of a configuration but its position, orientation and size. It can be composed of a set of quantitative continuous variables, which define the position of a group of landmarks within a configuration in either 2D or 3D. It is studied using multivariate statistics, so the variation in the position of the landmarks can be assessed altogether. The analyses are both descriptive and explanatory, so regressions using size, analyses of symmetry and covariation tests are also included in the common geometric morphometrics 'toolkit' (Adams, et al. 2013). Shape studies are capable of not only describing the variation but also explaining the presence of that variability to some extent.

The methods used in geometric morphometrics are not restricted to morphological structures at the organismal scale. There are no special conditions that constrain the application of geometric morphometrics to anatomical structures. Indeed, these tools have been applied successfully to a broad range of organisms: e.g. mammals (Klingenberg 2013), plants (Gómez, et al. 2014), rotifers (Fontaneto, et al. 2007), algae (Neustupa and Šťastný 2006, Neustupa and Nēmcová 2007, Neustupa, et al. 2010) or protists (Poulíčková, et al. 2010). However, very few molecular studies have been published using them (Adams and Naylor 2000, Theobald and Wuttke 2006a,

Theobald and Wuttke 2008). Despite the fact that the importance of the three-dimensional conformation in the evolution of proteins has been widely assessed in the literature and many different approaches has been applied to study it (Russell, et al. 1997, Martin, et al. 1998, Orengo, et al. 1999, Todd, et al. 1999, Thornton, et al. 2000, Chelliah, et al. 2004, Socolich, et al. 2005, Goldstein 2008, Siltberg-Liberles, et al. 2011, Meyer and Wilke 2013), no recent attempt of applying the set of techniques developed lately in geometric morphometrics has been done. This is surprising, since the study of protein evolution could benefit from the techniques geometric morphometrics offers in obtaining a new perspective about the evolution of molecular structures.

However, the application of the methods used in geometric morphometrics to proteins supposes a conceptual and methodological challenge, since changes in the evolution of the shape of the proteins may be different to the changes in the evolution of the anatomical structures (e. g. more abrupt or relatively larger). These difficulties can show up in the application of the common techniques successfully used in morphological studies and in the explanation of the results. In this study we analyse the potentials and the pitfalls of the current techniques developed in geometric morphometrics in their application to study the evolution of the shape in proteins. We have two different aims: in first place to confirm different features about the evolution of shape in proteins using these new methods (e.g. allometric patterns or the importance of functional demands). That will give an estimate of the reliability of these new tools and will show the efficiency of these methods to study protein evolution. In second place, the results will also allow us to suggest new hypotheses about the evolution of protein shape. In studies about anatomical structures, patterns that can be studied in proteins (e.g. evolutionary allometry) have been related to other ones that have not been explored in these molecules (e.g. integration) (Klingenberg 2013). If we find common patterns in proteins and anatomical

structures we may be able to propose new hypotheses about the evolution of protein shape. We apply geometric morphometrics tools to a group of chymotrypsin-like serine proteases (Table 4.1) (Rawlings and Barrett 1993, Barrett and Rawlings 1995, Krem and Di Cera 2001), a very well known family of proteins.

4.2. Materials and Methods

4.2.1. Protein Shape Data

The raw data used in geometric morphometrics is a set of x, y and z coordinates of the position of specific points, landmarks, taken in an arbitrary reference system. In anatomical structures, these landmarks can be located in any position of a given structure. In the case of the proteins, the landmarks are the different atoms of the structure. As it also happens in morphological studies, the data collection in crystallography is not free from different kinds of measurement errors (Borek, et al. 2003). Since geometric morphometrics can detect very subtle differences in shape, the application of these techniques requires the protein structures to be well resolved. That increases the accuracy of the position of the landmarks. The parameters used to assess the quality of the data have been the resolution, R and Rfree factors (Table 4.1). The crystallographic methods detect areas in a given space that resist the crossing of electrons or X-ray. It is assumed that the structure of interest is embedded in that area of electron-density. In crystallography, the resolution refers to the magnitude of that area. Resolution can be viewed as an empty contour surface, e.g. a cylinder, in which the structure under study is. The resolution would be the diameter of that cylinder. The smaller that diameter is, the better the signal is. The shape of the area of electron-density tends to resemble the shape of the molecule under study. Therefore the better the determination should be. The resolution of a resolved structure

represents the limits in which the data collection has been made or, in other words, 'the amount of data (...) used in the structure determination' (Chapter 8, p. 185) (Rhodes 2010). However, the situation is better than just having a 'cloud' in which our structure should be embedded: we have some a priori models about the shape of the structure. The R-factor reveals the coherence between the observed data in the structure determination and the data calculated from a model for that specific protein (Morris, et al. 1992). The lower it is, the more agreement there is between what our model predicts and the data obtained. In other words, the coherence between the shape of the cloud of electron density we have obtained and the shape of the prediction we have for that structure. A low R-factor is a measure of good data reliability. The R-factor is correlated with the resolution, so the smaller the area of electron density (resolution) the better the determination of the structure is (R-factor) (Kleywegt and Jones 1997, Rhodes 2010). The Rfree factor is used as a cross-validation parameter for the R factor (Brünger 1992). The fit of the predicted model for our structure in the electron density cloud may not be optimal. The fitting is a statistical process (e.g. maximum likelihood) that tries to optimize the matching between electron-density cloud and predicted model. However, it can be trapped in local optima. The Rfree factor checks the quality of the fitting using some of the data not used in the previous fitting. In the case that the fitting process reaches an optimum, the predicted model does not fit in the obtained data well enough even when it could. We would be under-interpreting the data. Alternatively, it may happen that during the fitting the model has been changed in an unrealistic way to fit the data, so it has over-interpreted the data. In those cases in which there have not been any over or under-interpretation of the data it converges to the same value as the R factor.

In order to study evolution it is necessary to limit studies to homologous regions of the structures. If we identify those areas of the

structure that are homologous and we leave out the rest of the data, the interpretations of the results can be made in an evolutionary context. Hypotheses of homology between areas of the different structures can be obtained using sequence alignments (Wong, et al. 2008, Blackburne and Whelan 2013), so corresponding amino acids can be identified. There are different options to establish homology hypotheses depending on the proteins that are going to be compared. In the situation in which the proteins studied are evolutionary distant, the use of genetic data can help. The protein sequences may be completely different but in the genetic data there may be homologous regions where relatively few nucleotides have change. As a consequence of these nucleotides changes, the protein sequence signal may be lost. However, homologous regions may be identified in the DNA and then translate that to hypothesis of homology between amino acids. In the opposite situation, in which all the proteins belong to the same family and are relatively conserved, the DNA sequence alignment should not give much more information than the protein sequence alignment. In any case, the method used to obtain the sequence alignment also needs to be considered depending on the proteins of interest (Blackburne and Whelan 2012). Different methods can give the same results with proteins that share the same distance in the sequence space between them. The study of proteins with different distances may require sequence alignment methods especially careful with the identification of homologous amino acids (Blackburne and Whelan 2012, Blackburne and Whelan 2013). Once the homologous amino acids are identified, their side chains can be discarded to study the overall shape of the proteins. At the end, just the xyz coordinates of the alpha carbons should be included in the analyses.

In our study, we used a family of 27 serine proteases taken from the Protein Data Bank (Berman, et al. 2000). Three of these proteins, the neuropsin and the first two salmon trypsins, have two different chains that

have been treated as different proteins, so the dataset is finally composed of 30 structures (Table 4.1). Although all the proteins belong to the same evolutionary family and therefore are relatively conserved, the sequence distances between them are very different. These conditions are not ideal then for the part of the analyses that involve the protein sequence. Some databases like HOMSTRAD (Mizuguchi, et al. 1998) offer groups of homologous proteins with a high sequence similarity and trustworthy sequence alignments but with some limitations in the quality of the structural data, which we have prioritized, given that the structural analysis is where the novelty of this study lies.

Protein (PDB ID)	**Species**	**Molecular Replacement (PDB ID)**	**Res. (Å)**	**Physiological Process**
Azurocidin (1A7S)	Homo sapiens	Preliminar Azurocidin	1.12	Immunological
Trypsin (1FNI)	Sus scrofa	Bovine β-Trypsin	1.63	Digestive
Trypsin (1H4W)	Homo sapiens	Human Trypsin I (1TRN)	1.7	Other
Trypsin (1HJ8)	Salmo salar	Anionic salmon trypsin form 2 (1BIT)	1	Digestive
Pro-chymase (1NN6)	Homo sapiens	Activated human chymase (1klt)	1.75	Immunological
Kallikrein (1SPJ)	Homo sapiens	Human K6 structure (1LO6)	1.7	Other
Chymotrypsin (1YPH)	Bos taurus	Yes (but not published)	1.34	Immunological

Urokinase (4FU9)	Homo sapiens	Yes (but not published)	1.6	Coagulation
Trypsin (4I8H)	Bos taurus	Bovine Trypsin (3MFJ)	0.75	Digestive
Trypsin (4M7G)	Streptomyces erythraeus	S. griseus Trypsin (1SGT)	0.81	Digestive
Matriptase (1EAX)	Homo sapiens	No	1.3	Immunological
Factor VIIa (1KLI)	Homo sapiens	Factor VIIa (1CVW)	1.69	Coagulation
Chymase (1KLT)	Homo sapiens	Homology construct of human chymase	1.9	Immunological
Elastase (1GVK)	Sus scrofa	PPE BCM7 structure (1QIX)	0.95	Immunological
Thrombin-like venom (3SG9)	Gloydius saxatilis	AaV-SP-I (1OP0)	1.43	Coagulation
Prostasin (3DFJ)	Homo sapiens	Human plasma kallikrein (2ANW)	1.45	Coagulation
Complement Factor D (1BIO)	Homo sapiens	Factor D molecule B in triclinic cell	1.5	Immunological
Factor Xa (3FFG)	Homo sapiens	No	1.54	Coagulation
Trypsin III (2ZPS)	Salmo salar	Salmon trypsin (1MBQ)	1.55	Digestive
Thrombin	Homo sapiens	Human thrombin	1.55	Coagulation

(3U69)		(1VZQ)		
Trypsin II (2ZPR)	Salmo salar	Salmo trypsin (1MBQ)	1.75	Digestive
Trypsin I (2ZPQ)	Salmo salar	Salmo trypsin (1MBQ)	1.9	Digestive
Neuropsin (1NPM)	Mus musculus	Bovine pancreatic β-Trypsin	2.1	Other
Trypsin (2F91)	Pontastacus leptodactylus	N	1.2	Digestive
AHV (4E7N)	Agkistrodon halys	Yes (Not published)	1.75	Coagulation
Viper venom (3S9B)	Daboia rusellii siamensis	ACC-C (2AIQ)	1.9	Coagulation

***Table 4.1**. Table of the proteins used in the morphometric study. From left to the right, the columns express the protein used, the organisms where they have been obtained, the molecules used in the molecular replacement (where molecular replacement has been used), the resolution and the physiological function in which they are involved. The data about the molecular replacement has been obtained from the original publications (Kang, et al. , Razeto, et al. , Zeng, et al. , McGrath, et al. 1997, Jing, et al. 1998, Karlsen, et al. 1998, Kishi, et al. 1999, Deepthi, et al. 2001, Leiros, et al. 2001, Friedrich, et al. 2002, Katona, et al. 2002a, Katona, et al. 2002b, Sichler, et al. 2002, Reiling, et al. 2003, Laxmikanthan, et al. 2005, Fodor, et al. 2006, Koizumi, et al. 2008, Rickert, et al. 2008, Toyota, et al. 2009, Quan, et al. 2010, Huang, et al. 2011, Nakayama, et al. 2011, Figueiredo, et al. 2012, Liebschner, et al. 2013, Blankenship, et al. 2014)*

We use two different programs for the alignment in order to test the importance of the hypotheses of amino acid homology and the ability of different methods to deal with it. We execute from the European Bioinformatics Institute website (http://www.ebi.ac.uk/): webPRANK (Löytynoja and Goldman 2010) and MUSCLE software (Edgar 2004). These methods belong to two different kinds of multiple sequence alignment programs: while the former is 'intended to produce evolutionarily realistic alignments', the latter is a 'similarity based' one (Blackburne and Whelan 2013) (Page 645). The main technical difference between these two methods is the treatment of the indels during the alignment process. While MUSCLE tries to minimize the number of them webPRANK treats insertions and deletions differently using phylogenetic information (Löytynoja and Goldman 2008).

The outcome of each sequence alignment is a hypothesis of homologous amino acids in the 30 structures and the phylogeny used for the alignment. In the case of MUSCLE this phylogeny contains information about branch lengths, while the phylogeny obtained from webPRANK assumes equal branch lengths. Based on the alignment, the structural data of the homologous amino acids is collected. These data can be different, although the expectation is that many amino acids are proposed as homologous in both alignments due to the conservation in the family. We leave out the side chains of the morphometric analyses and focus on the backbone. In those amino acids in which we find A and B atoms in the PDB file we always choose the A one (Word, et al. 1999). The subsequent morphometric analyses are done independently for the dataset collected using MUSCLE for the alignment and the dataset collected using webPRANK for the alignment.

One extra outcome in the alignment using MUSCLE is a similarity matrix for all the structures built by other 'similarity based' sequence alignment program, CLUSTALW 2.1 (Larkin, et al. 2007). This similarity

matrix is built comparing proteins in pairs and giving a score at each residue position, so it penalizes gap openings, gap extensions and uses a protein weight matrix that evaluates the similarity of each possible amino acid to each other.

4.2.2. Procrustes Superimposition and Size

The morphometric analyses start with the superimposition of the different structures. A wide range of superimposition techniques have been used in protein structural biology (Adams and Naylor 2000, Theobald and Wuttke 2005, Theobald and Wuttke 2006b, Theobald and Wuttke 2006a, Hirsch and Habeck 2008, Theobald and Wuttke 2008, Fang, et al. 2009, Hasegawa and Holm 2009, Liu, et al. 2009, Mechelke and Habeck 2010, Sun, et al. 2012, Theobald and Steindel 2012, Gapsys and de Groot 2013). Geometric morphometrics uses a technique called Procrustes superimposition, which consists in three steps. First, all the structures are shifted to one arbitrary position, so the centroid (average point of all the landmarks) of all the structures falls in the same position. Then, all the structures are rescaled to the same size, usually the same centroid size, i.e. squared root of the the sum of the squared distances between each landmark and the centroid (Dryden and Mardia 1998). Finally, the structures are rotated so the variance in the position of the landmarks is minimized via a least-squares method. After this process, position, orientation and size are standardized (Dryden and Mardia 1998). This technique has received some criticism, since least-squares does not take into account the different variance of the atoms (the landmarks) can have (Theobald and Wuttke 2006a, Theobald and Wuttke 2008). However, although we agree that heterocedasticity is a property certainly common both in proteins and in anatomical structures, other techniques proposed to solve this problem require further assumptions. Procrustes superimposition has performed reasonably well in anatomical

studies and gives support to the rest of the methods developed in geometric morphometrics, including the theoretical construction of the shape space (Dryden and Mardia 1998). The shape space is a theoretical curved space where all the possible shapes are represented. It has certain properties, such as its dimensionality, determined by the Procrustes superimposition. The dimensionality of the shape space is 3 times the number of landmarks (for 3D data) minus 7, the degrees of freedom lost during the superimposition. These properties remain unexplored for different superimposition methods. An important feature is the fact that for most of biological data the tangent shape space is a good local approximation to the shape space (curved) (Marcus, et al. 2000). This projects a portion of a curved space into a flat tangent space, so it extensively facilitates the mathematics and statistics behind the methods. However, when dealing with evolutionary distant proteins, it may be reasonable to check whether this equivalency still holds, since such features have not been shown with molecular data.

We use four different datasets for the morphometric analyses, in which the 30 structures are included. The first one is composed by the alpha-carbons of the whole proteins and the other three are just specific parts of them: 2) the catalytic tryad plus one amino acid involved in the formation of the oxyanion hole (active site), the Ser^{195}, His^{57} and Asp^{102} plus the Gly^{193}, 3) the part of the protein involved in the substrate specificity, formed by the aminoacids 190-192, 214-215 and 225-228 in the webPRANK dataset and by the same aminoacids plus the aminoacid 189 in the MUSCLE dataset (Perona and Craik 1995, Hedstrom 2002, Polgár 2005) and 4) the aminoacids proposed to be evolutionary markers in this family, the catalytic tryad plus Ser^{214} and Pro^{225} (Krem and Di Cera 2001).

4.2.3. Comparative methods

The data in macroevolutionary studies is influenced by its phylogenetic structure, so comparative methods are required to track the evolutionary changes in shape in the proteins (Felsenstein 1985). In this sense, comparative methods have been used to study morphological structures in geometric morphometrics for some time now (Klingenberg and Marugán-Lobón 2013, Monteiro 2013) and there should not be any difference for protein studies.

The use of independent contrasts allows the study the evolution of shape by removing the effect of phylogenetic relatedness (Felsenstein 1985, Rohlf 2001, Klingenberg and Marugán-Lobón 2013). In other words, the comparison between sister nodes avoids finding common shape features between structures that are closer in the phylogeny. For that, apart from the shape data a reliable phylogeny with the different structures under study is needed.

The extraction of reliable phylogenetic information from molecular data can be done during the sequence alignment, using again DNA or the residues sequence depending on the evolutionary proximity of the proteins used. Once a reliable phylogeny is obtained, we can assess the distribution of our data in the shape space. For that, we can run first by a principal component analysis (PCA), so the variance in shape in our sample and the distances between proteins is observed. Then, it is possible to map the data on a reliable phylogeny using squared-change parsimony (Maddison 1991, McArdle and Rodrigo 1994), so the position of the internal nodes of the phylogeny in the shape space and the evolution in shape between nodes are estimated. This allows the observation of the distribution of the different structures in shape space and possible clusters of them. The overall phylogenetic signal can be assessed via permutation test, i.e. whether or not the pattern of shape

similarity in the family of proteins follows the pattern of phylogenetic relatedness (Klingenberg and Gidaszewski 2010).

In our example, the phylogenies obtained in the alignments are mapped using the weighted squared-change parsimony in the case of the dataset obtained using MUSCLE and the unweighted version in the case of the dataset obtained using webPRANK.

4.2.4. Allometry

Size is a factor commonly found to influence the evolution of shape. The change in shape due to the change in size is called allometry and it has been extensively studied in geometric morphometrics (Monteiro 1999, Piras, et al. 2010, Sidlauskas, et al. 2011, Klingenberg and Marugán-Lobón 2013, Voje, et al. 2014, Watanabe and Slice 2014). Although size is standardized during the superimposition, it is stored as a variable for further analyses, so the evolutionary relationship size-shape can be studied using a multivariate linear regression of the independent contrasts of shape on the independent contrasts of centroid size (Klingenberg and Marugán-Lobón 2013).

However, the effect of size on the evolution of shape may be very different in proteins to morphological structures. Whereas the centroid size usually varies smoothly in the evolution of anatomical structures, relatively few changes in a protein sequence can bend the protein and therefore change its centroid size drastically. This is the case, for example, when globular and fibrous proteins are studied. Fortunately, protein volumes has been shown to be under conservation in proteins (Gerstein, et al. 1994), so big changes may be uncommon. In any case, they may reveal important biological features of the proteins. Outliers can also be left out to test their influence on the results concerning the whole family.

In those analyses in which allometry is found, analyses using the residuals of the regression are encouraged. This is the variance that is left

after removing the predicted effect of size, so analyses removing the effect of size are interested to know at what extend some results are an effect of size.

4.2.5. Structure and function

Size is one factor that can explain the evolution of shape in a structure but it is not necessarily the most important one. Functional demands also influence the protein structure (Goldstein 2008). Geometric morphometrics can test hypothesis of functional diversification (De Esteban-Trivigno 2011). Some clusters of proteins with the same function can appear when using a PCA, suggesting a relationship between shape and function. However, this technique is not ideal to discriminate groups in the dataset since it does not take into account any group structure. For that purpose, we use a canonical variate analysis (CVA) (Albrecht 1980). We first need to group the specimens according to their function and the variation within these groups is estimated. Then, the space is transformed so the variance is increased in those directions in which the within-groups variance is smaller and decreased in those ones in which the within-groups variance is bigger. That maximizes the separation among groups (Klingenberg and Monteiro 2005). However, because CVA is based on the variance of the groups, it is sensitive to low sample sizes. Whereas that may not be a problem when specific parts of the proteins are used (as long as the number of landmarks is not very high), this is the case when we analyse the whole proteins. To avoid that problem, a between-group PCA can be run (Klingenberg and Spence 1993, Boulesteix 2002, Mitteroecker, et al. 2005, Mitteroecker and Bookstein 2011). In it, first a PCA of the proteins averaged by the functional groups is run. Then, a new PCA is run for all the proteins using the scores of the previous PCA.

For the serine proteases, we classify the specimens in three groups according to the physiological function in which they are involved: immunological, coagulation, digestive and other (Blasi, et al. 1987, Perona

and Craik 1997, Roach, et al. 1997, Belaaouaj, et al. 1998, Jing, et al. 1998, Friedrich, et al. 2002, Katona, et al. 2002a, Krem and Di Cera 2002, Sichler, et al. 2002, Reiling, et al. 2003, Wątorek 2003, Terayama, et al. 2005, Yousef, et al. 2005, Fodor, et al. 2006, Gallwitz and Hellman 2006, Rickert, et al. 2008, Toyota, et al. 2009, Huang, et al. 2011, Nakayama, et al. 2011, Porter, et al. 2012, Blankenship, et al. 2014) (Table 1).

4.2.6. Sequence and structure

The sequence is the last factor influencing shape that we studied. It is possible to know by how much the similarity in the sequences is related to the shape and therefore at what extend the sequences are responsible of all the previous results. For that, we use the partial least squares (PLS) (Rohlf and Corti 2000). Two blocks of data are included in this analysis: the shape data, represented by the coordinates obtained after the Procrustes superimposition (Procrustes coordinates), and a set of variables reflecting the diversification of the sequences. In order to obtain the latter, we first get a similarity matrix from the sequence alignment, which is transformed into a dissimilarity matrix subtracting each measure from 100. This similarity matrix is usually obtained with the sequence alignment and its calculation depends on the specific software. Then a Principal Coordinate Analysis (PCoA) (Gower 1966) is run using that. PCoA produces a set of variables reflecting the diversification of the sequences. The PCo Scores obtained are used to run the PLS along with a permutation test, where the null hypothesis is the independence between them and the shape variables. The RV coefficient gives us a percentage of correlation between these variables (Klingenberg 2009). However, here sample size also matters, so it may be advisable to use a reduced number of PCo Scores.

The morphometric analyses were all run in MorphoJ (Klingenberg 2011). The removal of side chains of the data taken from the PDB and the

analyses regarding the PCoA were run using R version 3.0.2 (R Core Team 2013). The removal of some non-homologous aminoacids was done using Microsoft Excel for Mac 2011 version 14.4.3.

4.3. RESULTS

Using both the dataset obtained from the alignment with the evolutionary-based alignment method and the dataset obtained from the alignment with the similarity-based method there is little variation in size in this family of proteins. The test for evolutionary allometry is significant (webPRANK dataset: p-value=0.0014; % predicted: 12.27%; MUSCLE dataset: p-value=0.0342; % predicted: 6.60%) (Figure 4.1). We also find a stronger effect of the size over shape in both datasets when we test the evolutionary allometry using the active site (webPRANK dataset: p-value=0.0015; % predicted= 49.0665%; MUSCLE dataset: p-value=<0.0001; % predicted= 94.003%), the part involved in substrate specificity (webPRANK dataset: p-value= 0.0155; % predicted= 26.0165%; MUSCLE dataset: p-value= <0.0001; % predicted= 98.2451%) or the evolutionary markers subdivision (webPRANK dataset: p-value= 0.0015; % predicted= 24.99%; MUSCLE dataset: p-value= <0.0001; % predicted= 99.4915%). However, in all these datasets there is (at least) one outlier regarding size (Figure 4.1). When we repeat these analyses removing the protein causing the outlier, the Prochymase, the results are different. In the active site there is not evolutionary allometry (webPRANK dataset: p-value=0.5843; % predicted: 2.3974%; MUSCLE dataset: p-value=0.0631; % predicted: 7.4291%). We still find evolutionary allometry for the substrate specificity area (webPRANK dataset: p-value=0.0042; % predicted: 23.9865%; MUSCLE dataset: p-value<0.0001; % predicted: 18.2093%) and the evolutionary markers (webPRANK dataset: p-value=0.0013; % predicted: 25.8719%; MUSCLE dataset: p-value=0.0005; % predicted: 23.8659%) but much softer.

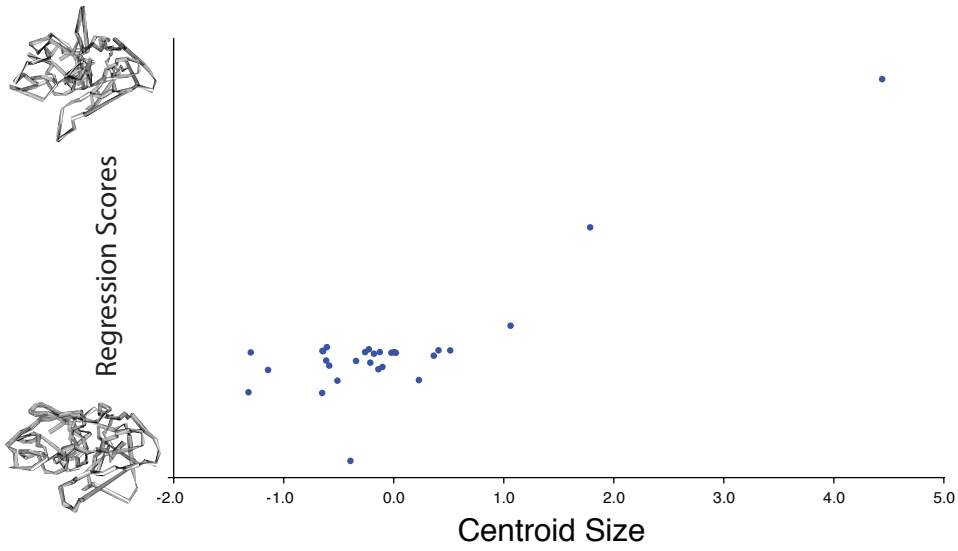

***Figure 4.1**. Scores of the Evolutionary Allometry regression using the independent contrasts of the whole proteins and the data obtain from the alignment with webPRANK. Each dot in the graph represents one contrast. In the y-axis the change associated to the difference in centroid size can be observed. On the top, the proteins appear partly unfolded while on the bottom the proteins are more compact.*

This allometric effect is especially strong in the case of the analyses in which the datasets aligned with the similarity-based method are used and just some part of the protein is included. In these cases the contrast between chymase and prochymase is always situated apart from the rest of the contrasts. Since the transition from the structure of the prochymase to the chymase is mediated by a conformational change, so the change between them in important parts of the protein as the active site and parts involved in the substrate specificity is not only evolutionary but also functional.

Following the theory of gene duplication followed by functional differentiation (Lynch 2000), the phylogenies present clades in which the proteins share the physiological function (Figures 4.2 & 4.3). Therefore, it is

not surprising to find in both datasets a strong phylogenetic structure in the shape data when we use the whole protein (p-values<0.0001) (Figures 4.4 & 4.5). However, there are some differences that can be observed in the PCAs. In the PCA run over the evolutionary-based dataset, the proteins that are far from the rest are the snake venoms, the factor VIIa and the matriptase, revealing uncommon shape features that may relate to their uncommon function (Figure 4.4). The pro-chymase is very close to the chymase, as expected, given that they are indeed two different forms of the same protein. All the trypsins but those from a bacteria and a crustacean, evolutionarily remotely related organisms from the other organisms, are clustered together. In the PCA run over the similarity-based dataset, however, the proteins have a distribution more difficult to explain from the evolutionary point of view (Figure 4.5). The pro-chymase and the chymase are far away from the rest of proteins and the structures of the trypsins are more different relatively to the other proteins.

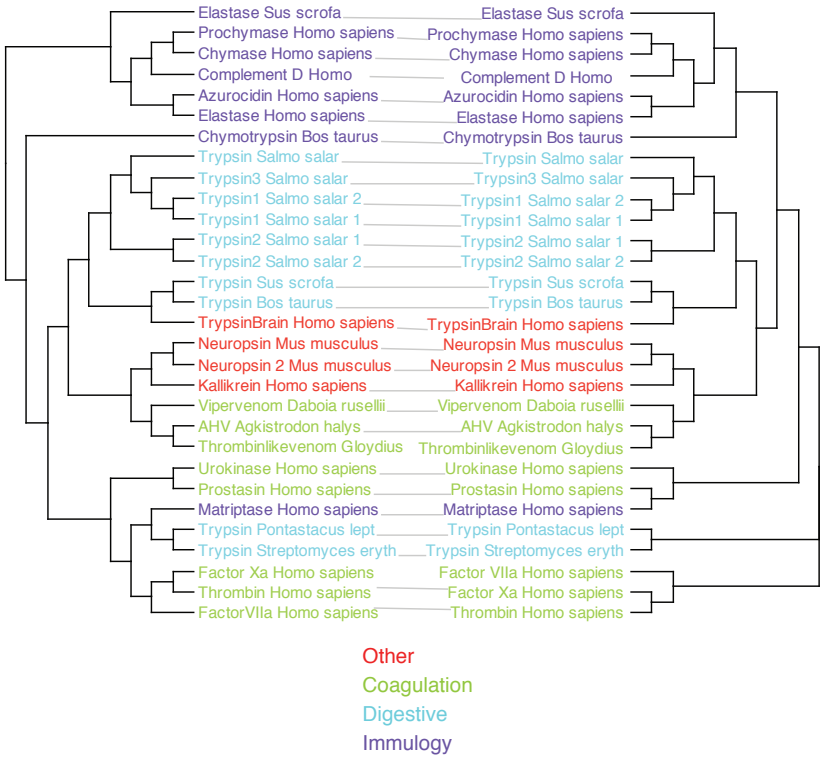

Figure 4.2. *Phylogenies obtained from the sequences in webPRANK (left) and MUSCLE (right). The first letters refer to the protein and the last 2 letters are the initials of the species.*

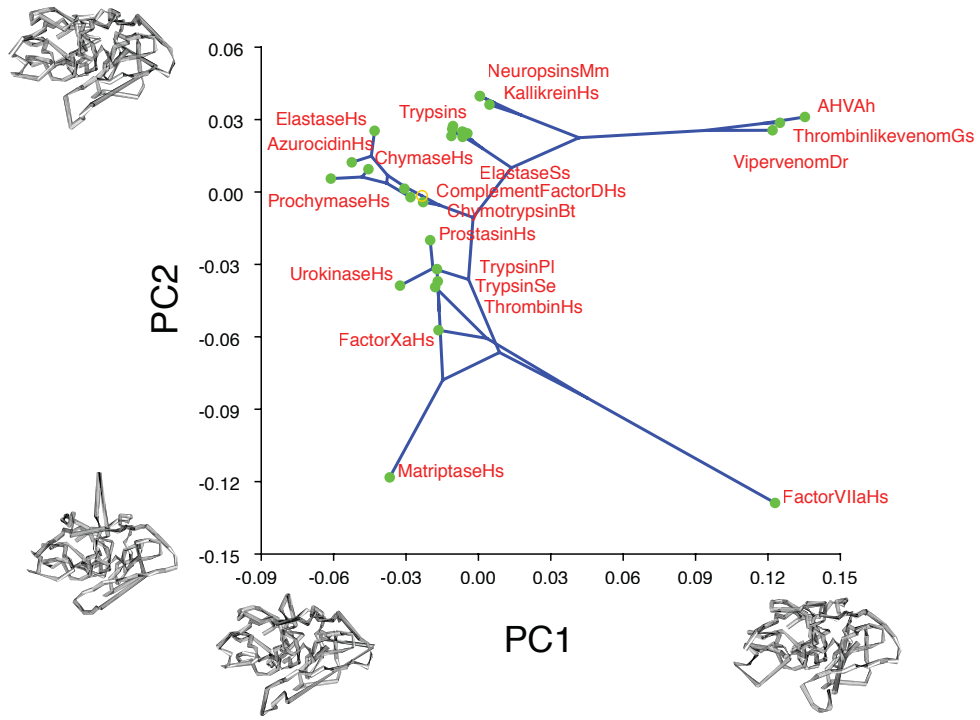

Figure 4.3. *PC Scores of shape mapped on the phylogeny using webPRANK and the whole proteins.*

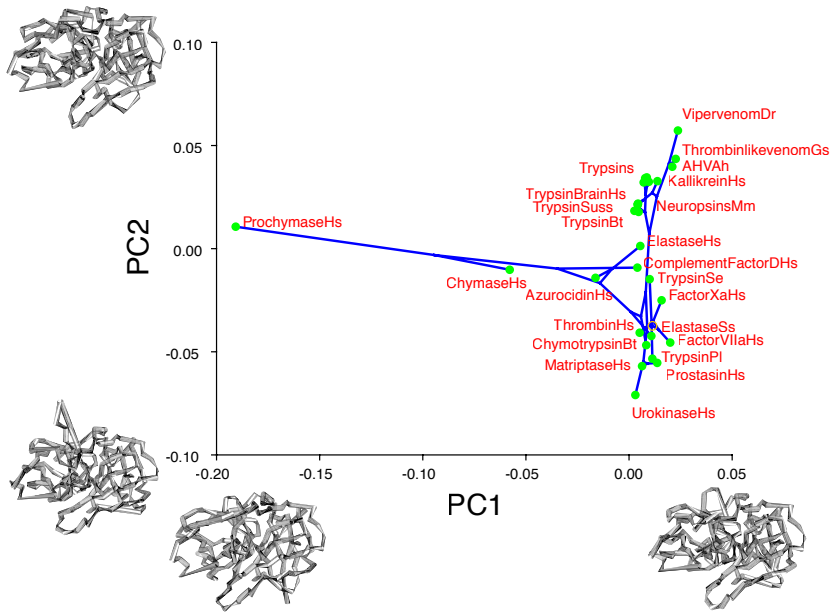

Figure 4.4*. PC Scores of shape mapped on the phylogeny using MUSCLE and the whole proteins.*

When only the active site is used, there is no phylogenetic signal (webPRANK dataset: p-value=0.8150; MUSCLE dataset: p-value=0.9908) (Figure 4.6). A marginally nonsignificant phylogenetic signal is found when the residuals of that regression are taken (webPRANK dataset: p-value=0.0591; MUSCLE dataset: p-value=0.0529). When we use the amino acids involved in substrate specificity (Hedstrom 2002), we again do not find phylogenetic signal (webPRANK dataset: p-value=0.2238; MUSCLE dataset: p-value=0.9741). If we remove the effect of size we find marginal nonsignificance for the dataset obtained from the webPRANK alignment (p-value=0.0545) but complete significance for the dataset from the MUSCLE alignment (p-value=0.0029). The only test that gives different results for the datasets obtained using the evolutionary-based method and the similarity-based one is that one regarding the phylogenetic signal using the evolutionary markers identified by Krem and Di Cera (2000, 2001) (webPRANK dataset: p-value=0.0007; MUSCLE dataset: p-value=0.9826).

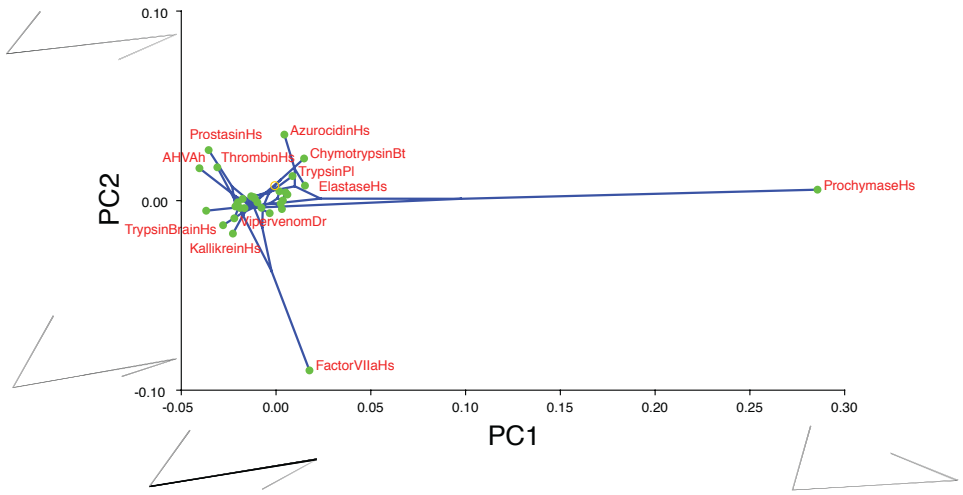

Figure 4.5*. PC Scores of shape mapped on the phylogeny using webPRANK and the active site.*

The CVA analysis, a discriminant analysis, reveals that there is not a good discrimination between functional groups using either of the alignments in the active site. However, when we use the amino acids involved in the general substrate-specificity the discrimination between all the groups is highly significant (in both datasets all the p-values<0.005) (Figure 4.7). When we use the between-group PCA, so it is used a PCA of the amino acids averaged by function to run a PCA on these scores, we find a good differentiation between functional groups, especially when the evolutionary-based method has been used (Figure 4.8). There are many results that are very similar using either of the alignments and the choice of homologous aminoacids.

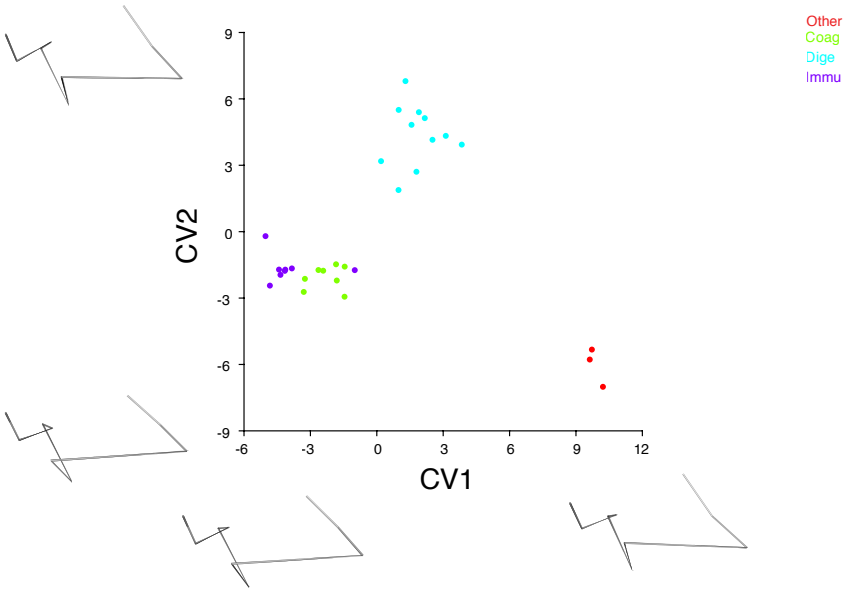

***Figure 4.6**. CVA of shape using the part of the protein involved in substrate specificity and webPRANK.*

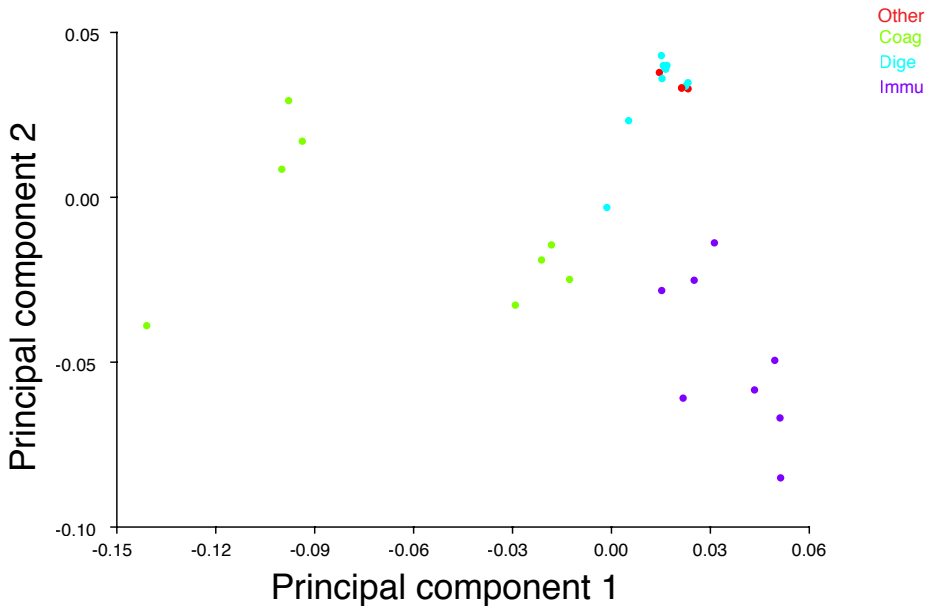

***Figure 4.7**. Results of the between-group PCA. PCA of shape of the whole proteins using the dataset obtained from webPRANK and the PC Scores from the PCA of the dataset average by function.*

Finally, the permutation test of association between the principal coordinate scores of the sequence space and the shape variables gives significant results (p-value=<0.0001, RV coefficient= 0.7746). When just the first five Principal Coordinate Scores are used, so the sample size should not be a problem, these results hold (p-value<0.0001; RV coefficient= 0.6823) (Figure 4.9). In the PLS, the proteins have a strong phylogenetic signal in the shape space (p-value<0.0001). However, functional groups do not appear completely clustered (Figures 4.10 & 4.11).

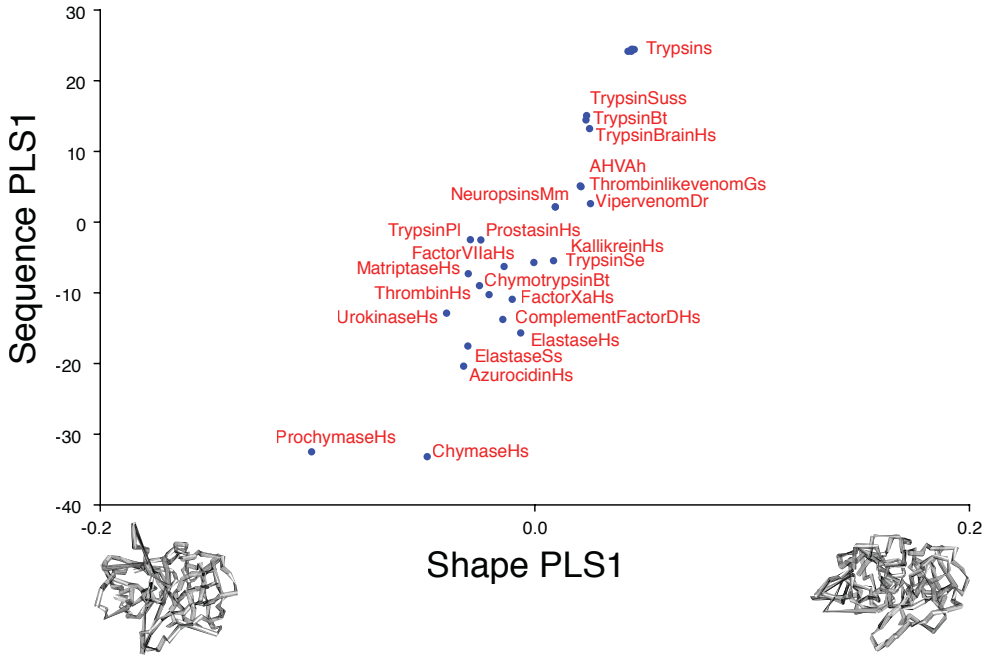

***Figure 4.8**. Partial Least Squares run using the Procrustes Coordinates of the whole proteins (Block 1) and the PCo Scores of the dissimilarity matrix obtained in MUSCLE (Block 2).*

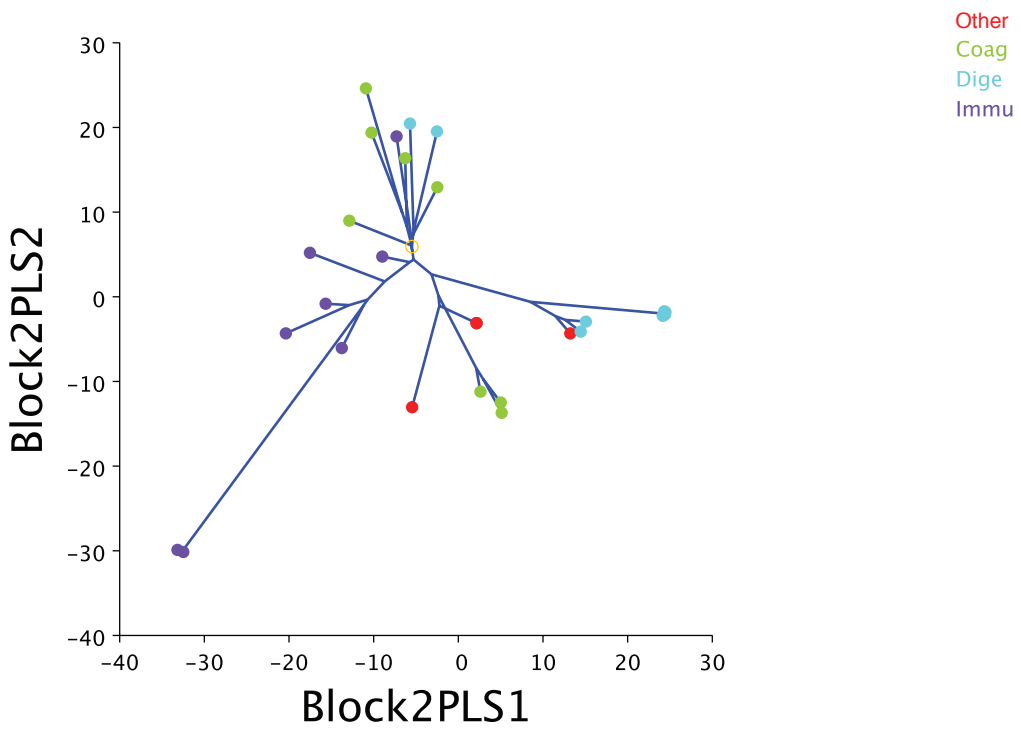

***Figure 4.9**. Sequence space in the Partial Least Squares: two dimensions of top covariation.*

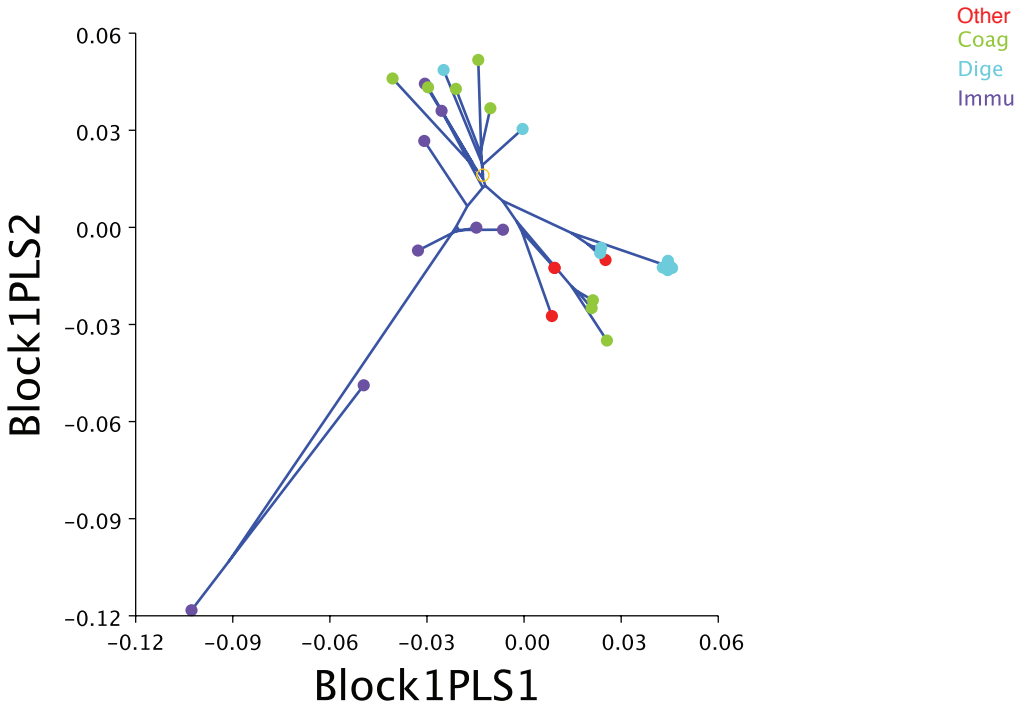

***Figure 4.10**. Shape space in the Partial Least Squares: two dimensions of top covariation.*

4.4. DISCUSSION

Geometric morphometrics is a very powerful tool to find subtle differences in shape between proteins and propose hypotheses about their evolution. In first place, we have been able to confirm previous features studied in the literature about serine proteases: the constraint in size, the phylogenetic signal in the shape of these proteins, specific convergent cases associated to special functional demands and a strong relationship between sequence and structure. Therefore, geometric morphometrics can be used to study these features in other families of proteins that are not known enough. In second place, our results are analogous to those found in many anatomical studies, so there are forces (sequence variation, size, functional demands) that are common in the generation of the shape diversification. That gives an opportunity to propose new hypotheses about the evolution of the structure in proteins. For example, these forces behind shape variation in the evolution of serine proteases (functional demands, allometry)have been suggested to be integrating factors in anatomical structures (Klingenberg 2008, Armbruster, et al. 2014, Klingenberg 2014). If we establish an analogy with protein evolution, we can suggest an important role of evolutionary integration in the evolution of the serine proteases, which could be tested in the future.

4.4.1. Importance of the Raw Data

The importance of choosing evolutionarily corresponding landmarks has been emphasized in geometric morphometrics (Bookstein 1991, Zelditch, et al. 2012). In proteins, this process can be done using a sequence alignment. Our results using webPRANK and MUSCLE suggest that the former, as expected due to previous studies (Blackburne and Whelan 2012, Blackburne and Whelan 2013), is more coherent with previous evolutionary studies in this family. The results obtained with the similarity-based approach looks more sensitive to the functionality of the proteins, as illustrated by the

positions of the chymase and the prochymase in the shape space, where they are distant from the rest of proteins. This big difference may be due to the conformational change that is needed to convert prochymase into chymase (Reiling, et al. 2003). However, both methods gave relatively similar results in the tests we run, given that there are many amino acids in common using both methods and that, for those areas within the proteins, the resulting differences where due just to differences in the phylogeny. In studies in which a family of proteins is studied and therefore all the proteins are evolutionarily conserved, we do not expect the results to be very different. In this case, the results obtained using either webPRANK or MUSCLE are coherent with the literature published about serine proteases.

Once the homologous amino acids are chosen, the structural data needs to be collected. However, as landmark data in morphological studies, crystallographic data is not free from methodological problems and measurement errors. Even if the measurement errors (Borek, et al. 2003) associated to our data are reasonable, the information obtained from the structure may not be enough to explain biological features (Wlodawer, et al. 2008). Proteins are not stable structures with an immutable shape, so they vary depending on whether they are active or inactive and whether they are bound to other molecules. This is especially pertinent when the proteins are obtained from databases, as in this study, so the structures were obtained for different purposes: e.g. at different conformational stages or bound to other different molecules. In addition, most of these structures were collected using molecular replacement (Table 4.1), meaning that the structure of the proteins has been obtained using the structure of other known similar proteins and therefore the coordinates of some atoms in different proteins are not independent, so some similarities obtained in the results can be a product of this dependence.

4.4.2. Morphometric explanations

Our findings support some general knowledge about the structure of the proteins, such as the observed degree of size conservatism. Small variations in size are associated with changes in shape (Fig. 4.1), so when the distance between specific atoms increases the chemical interactions can have the same effect over shape. However, size conservatism is not complete in our sample, it is altered by some abrupt changes in the bending of the proteins that caused big changes in size (Fig. 4.1). The identification of abrupt changes can be very interesting biologically. In this case, there has been a big change in size in the branch connecting the matriptase from humans with the ancestor originating two trypsins: one from a bacteria and one from a crayfish. The big evolutionary distance between these organisms could explain such difference in size. Abrupt changes in the folding of the proteins can be one generator of allometry in proteins. The importance of allometry is that size can be one driver of integration (Klingenberg and Marugán-Lobón 2013). When in the active site and the area involved in substrate specificity we remove the effect of size over shape, we can obtain phylogenetic signal in our shape data. Therefore, in serine proteases, size variation in these areas can be one driver for the convergence in shape.

When the whole protein is studied, the appearance of phylogenetic signal is expected, since in general the proteins within phylogenetic clades share the same function (Figure 4.2) (Krem, et al. 2000, Krem and Di Cera 2001), with some exceptions (Yousef, et al. 2003). That makes extreme forms and specific events of convergence more noticeable and interesting. This is the case of the venoms, far away in the scatter plot, which obviously have a different mechanism of action than the rest of coagulation factors. Factor VIIa, another protein far from the rest of the sample, has a regulation mechanism via one cofactor, which is different to the rest of serine proteases (Sichler, et al. 2002). The presence or lack of phylogenetic signal found in

different parts of the protein can give some information about the evolution of shape in these areas (suggesting an important role in the biology of the proteins). The part of the protein involved in substrate specificity, a binding pocket for the side chains of the substrates, illustrates that the shape in this part is a possible adaptation to specific physiological processes so the phylogenetic signal has been lost. CVA has allowed us to differentiate functional groups when the part involved in substrate specificity has been used. However, a caveat is necessary here. Although the dimensionality here is much lower than when the whole protein is used and the results are better, the ratio between sample size and dimensionality is still too low. That should not be a problem when this hypothesis is tested using the between-group PCA for the whole protein, where the same results hold. This is compatible with the fact that the substrate specificity has driven the evolution of these proteins (Perona and Craik 1997, Esmon and Mather 1998, Krem, et al. 2000, Rose and Di Cera 2002). In other circumstances, in which the ratio sample size/dimensionality is a bigger problem, other techniques such as the cross-validation procedure can be used (Lachenbruch 1967). We also found that the variation in shape in the active site does not recover the pattern of phylogenetic relatedness and the CVA using this area cannot discriminate between groups either. In this case, allometry is playing a major role in the formation of convergence.

Finally, we have studied the relationship between the sequence and the shape, not only important to understand the evolution of this family but also to be extrapolated to others (Dessailly, et al. 2009, Williams and Lovell 2009, Marks, et al. 2011). As expected, the relative amount of shape variation that is associated to the sequence is high, especially bearing in mind that just the evolutionarily corresponding amino acids are included in the analysis. The collection of more data would allow a more accurate assessment of this relationship as well as the origin of the shape in variation in areas

where there is hardly any sequence variation, such as the active site. In addition, more data would also allow a more accurate application of the CVA and the opening of a broad new kind of studies involving the integration within the shape of the proteins (Trifonov and Frenkel 2009). In proteins like the serine proteases, where the different domains have probably originated by duplications (Halabi, et al. 2009), this kind of studies would be very promising.

4.5. Conclusion and Future Work

Overall, geometric morphometrics has been able to track the evolution of the protein structure in the serine proteases, identifying important features such as size changes or functional demands. The study of shape using geometric morphometrics also allows the association between shape and sequence in proteins, which in turn could be associated to DNA sequences. That would establish a pathway to study the evolution of the structure in proteins from their origin. The results obtained are coherent with the literature published in these proteins and the methods can be applied in many other fields where protein shape is important, as studies about coevolution, crystallography or immunology. In future studies we strongly recommend doing the collection of the data and avoid the collection via databases, so the previous pitfalls can be avoided. The replication of the data would also help for that purpose and would add many possibilities for the implementation of new methods.

4.6. References

Adams DC, Naylor GJP. 2000. A comparison of methods for assessing the structural similarity of proteins. In: Guerra C, Istrail S editors.

Mathematical methods for protein structure analysis and design. Heidelberg, Springer, p. 109-115.

Adams DC, Rohlf FJ, Slice DE. 2013. A field comes of age: geometric morphometrics in the 21st century. Hystrix 24:7-14.

Albrecht GH. 1980. Multivariate analysis and the study of form, with special reference to canonical variate analysis. American Zoologist 20:679-693.

Armbruster WS, Pélabon C, Bolstad GH, Hansen TF. 2014. Integrated phenotypes: understanding trait covariation in plants and animals. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Barrett AJ, Rawlings ND. 1995. Families and clans of serine proteases. Archives of Biochemistry and Biophysics 318:247-250.

Belaaouaj A, McCarthy R, Baumann M, Gao Z, Ley TJ, Abraham SN, Shapiro SD. 1998. Mice lacking neutrophil elastase reveal impaired host defense against gram negative bacterial sepsis. Nature 4:615-618.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Research 28:235-242.

Blackburne BP, Whelan S. 2012. Measuring the distance between multiple sequence alignments. Bioinformatics 28:495-502.

Blackburne BP, Whelan S. 2013. Class of multiple sequence alignment algorithm affects genomic analysis. Molecular Biology and Evolution 30:642-653.

Blankenship E, Vukoti K, Miyagi M, Lodowski DT. 2014. Conformational flexibility in the catalytic triad revealed by the high-resolution crystal structure of *Streptomyces erythraeus* trypsin in an unliganded state. Acta Crystallographica Section D 70:833-840.

Blasi F, Vassalli JD, Danø K. 1987. Urokinase-type plasminogen activator: Proenzyme, receptor and inhibitors. The Journal of Cell Biology 104:801-804.

Bookstein F. 1991. Morphometric tools for landmark data: geometry and biology. Cambridge, Cambridge University Press.

Borek D, Minor W, Otwinowski Z. 2003. Measurement errors and their consequences in protein crystallography. Acta Crystallographica Section D Biological Crystallography 59:2031-2038.

Boulesteix FL. 2002. A note on between-group PCA. International Journal of Pure and Applied Mathematics 19:359-366.

Brünger AT. 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355:472-475.

Chelliah V, Chen L, Blundell TL, Lovell SC. 2004. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. Journal of Molecular Biology 342:1487-1504.

De Esteban-Trivigno S. 2011. Ecomorfología de xenartros extintos: análisis de la mandíbula con métodos de morfometría geométrica. Ameghiniana 48:381–398.

Deepthi S, Johnson A, Pattabhi V. 2001. Structures of porcine beta-trypsin-detergent complexes: the stabilization of proteins through hydrophilic binding of polydocanol. Acta Crystallographica Section D 57:1506-1512.

Dessailly BH, Redfern OC, Cuff A, Orengo CA. 2009. Exploiting structural classifications for function prediction: towards a domain grammar for protein function. Current Opinion in Structural Biology 19:349-356.

Dryden IL, Mardia KV. 1998. Statistical shape analysis. New York, John Wiley & Sons.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Esmon CT, Mather T. 1998. Switching serine protease specificity. Nature Structural Biology 5:933-937.

Fang Y, Liu YS, Ramani K. 2009. Three dimensional shape comparison of flexible proteins using the local-diameter descriptor. BMC Structural Biology 9.

Felsenstein J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1-15.

Figueiredo AC, Clement CC, Zakia S, Gingold J, Philipp M, Pereira PJB. 2012. Rational design and characterization of D-Phe-Pro-D-Arg-derived direct thrombin inhibitors. Plos One 7:e34354.

Fodor K, Harmat V, Neutze R, Szilágyi L, Gráf L, Katona G. 2006. Enzyme:Substrate hydrogen bond shortening during the acylation phase of serine protease catalysis. Biochemistry 45:2114-2121.

Fontaneto D, Herniou EA, Boschetti C, Caprioli M, Melone G, Ricci C, Barraclough TG. 2007. Independently evolving species in asexual bdelloid rotifers. Plos Biology 5:e87.

Friedrich R, Fuentes-Prior P, Ong E, Coombs G, Oehler R, Pierson D, Gonzalez R, Huber R, Bode W, Madison EL. 2002. Catalytic domain structures of MT-SP1/Matriptase, a matrix-degrading transmembrane serine proteinase. Journal of Biological Chemistry 277:2160-2168.

Gallwitz M, Hellman L. 2006. Rapid lineage-specific diversification of the mast cell chymase locus during mammalian evolution. Immunogenetics 58:641-654.

Gapsys V, de Groot BL. 2013. Optimal superpositioning of flexible molecule ensembles. Biophysical Journal 104:196-207.

Gerstein M, Sonnhammer ELL, Chothia C. 1994. Volume changes in protein evolution. Journal of Molecular Biology 236:1067-1078.

Goldstein RA. 2008. The structure of protein evolution and the evolution of protein structure. Current Opinion in Structural Biology 18:170-177.

Gómez JM, Perfectti F, Klingenberg CP. 2014. The role of pollinator diversity in the evolution of corolla-shape integration in a pollination-generalist plant clade. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Gower JC. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325.

Halabi N, Rivoire O, Leibler S, Ranganathan R. 2009. Protein sectors: Evolutionary units of three-dimensional structure. Cell 138:774-786.

Hasegawa H, Holm L. 2009. Advances and pitfalls of protein structural alignment. Current Opinion in Structural Biology 19:341-348.

Hedstrom L. 2002. Serine protease mechanism and specificity. Chemical Reviews 102:4501-4523.

Hirsch M, Habeck M. 2008. Mixture models for protein structure ensembles. Bioinformatics 24:2184-2192.

Huang K, Zhao W, Gao Y, Wei W, Teng M, Niu L. 2011. Structure of saxthrombin, a thrombin-like enzyme from *Gloydius saxatilis*. Acta Crystallographica Section F 67:862-865.

Jing H, Babu YS, Moore D, Kilpatrick JM, Liu XY, Volanakis JE, Narayana SVL. 1998. Structures of native and complexed complement factor D: Implications of the atypical His57 conformation and self-inhibitory loop in the regulation of specific serine protease activity. Journal of Molecular Biology 282.

Kang YN, Stuckey JA, Nienaber V, Giranda V. Crystal structure of the urokinase. [To be Published].

Karlsen S, Iversen LF, Larsen IK, Flodgaard HJ, Kastrup JS. 1998. Atomic resolution structure of human HBP/CAP37/azurocidin. Acta Crystallographica Section D Biological Crystallography 54:598-609.

Katona G, Berglund GI, Hajdu J, Gráf L, Szilágyi L. 2002a. Crystal structure reveals basis for the inhibitor resistance of human brain trypsin. Journal of Molecular Biology 315:1209-1218.

Katona G, Wilmouth RC, Wright PA, Berglund GI, Hajdu J, Neutze R, Schofield CJ. 2002b. X-ray structure of a serine protease acyl-enzyme complex at 0.95-Å resolution. Journal of Biological Chemistry 277:21962-21970.

Kishi T, Kato M, Shimizu T, Kato K, Matsumoto K, Yoshida S, Shiosaka S, Hakoshima T. 1999. Crystal structure of neuropsin, a hippocampal protease involved in kindling epileptogenesis. Journal of Biological Chemistry 274:4220-4224.

Kleywegt GJ, Jones TA. 1997. Model building and refinement practice. Methods in Enzymology 277:208-230.

Klingenberg CP. 2008. Morphological integration and developmental modularity. Annual Review of Ecology, Evolution and Systematics 39:115-132.

Klingenberg CP. 2009. Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. Evolution & Development 11:405-421.

Klingenberg CP. 2011. MorphoJ: an integrated software package for geometric morphometrics. Molecular Ecology Resources 11:353-357.

Klingenberg CP. 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. Hystrix 24:43-58.

Klingenberg CP. 2014. Studying morphological integration and modularity at multiple levels: concepts and analysis. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Klingenberg CP, Gidaszewski NA. 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. Systematic Biology 59:245-261.

Klingenberg CP, Marugán-Lobón J. 2013. Evolutionary covariation in geometric morphometric data: analyzing integration, modularity, and allometry in a phylogenetic context. Systematic Biology 62:591–610.

Klingenberg CP, Monteiro LR. 2005. Distances and directions in multidimensional shape spaces: Implications for morphometric applications. Systematic Biology 54:678-688.

Klingenberg CP, Spence JR. 1993. Heterochrony and allometry: Lessons from the water strider genus *Limnoporus*. Evolution 47:1834-1853.

Koizumi M, Fujino A, Fukushima K, Kamimura T, Takimoto-Kamimura M. 2008. Complex of human neutrophil elastase with 1/2SLPI. Journal of Synchroton Radiation 15:308-311.

Krem MM, Di Cera E. 2001. Molecular markers of serine protease evolution. The EMBO Journal 20:3036-3045.

Krem MM, Di Cera E. 2002. Evolution of enzyme cascades from embryonic development to blood coagulation. Trends in Biochemical Sciences 27:67-74.

Krem MM, Rose T, Di Cera E. 2000. Sequence determinants of function and evolution in serine proteases. Trends in Cardiovascular Medicine 10:171-176.

Lachenbruch PA. 1967. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics 23:639-645.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R*, et al.* 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948.

Laxmikanthan G, Blaber SI, Bernett MJ, Scarisbrick IA, Juliano MA, Blaber M. 2005. 1.70 Å X-ray structure of human *apo* kallikrein 1: Structural changes upon peptide inhibitor/substrate binding. Proteins: Structure, Function and Bioinformatics 58:802-814.

Leiros HK, McSweeney SM, Smalås AO. 2001. Atomic resolution structures of trypsin provide insight into structural radiation damage. Acta Crystallographica Section D 57:488-497.

Liebschner D, Dauter M, Brzuszkiewicz A, Dauter Z. 2013. On the reproducibility of protein crystal structures: five atomic resolution structures of trypsin. Acta Crystallographica Section D 69:1447-1462.

Liu YS, Fang Y, Ramani K. 2009. Using least median of squares for structural superposition of flexible proteins. BMC Bioinformatics 10.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632-1635.

Löytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics:579.

Lynch M. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

Maddison WP. 1991. Squared-change parsimony reconstructions of ancenstral states for continuous-valued characters on a phylogenetic tree. Systematic Zoology 40:304-314.

Marcus LF, Hingst-Zaher E, Zaher H. 2000. Application of landmark morphometrics to skulls representing the orders of living mammals. Hystrix 11:27–47.

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D structure computed from evolutionary sequence variation. Plos One 6:e28766.

Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JBO, Taroni C, Thornton JM. 1998. Protein folds and functions. Structure 6:875-884.

McArdle B, Rodrigo AG. 1994. Estimating the ancestral states of a continuous-valued character using squared-change parsimony: An analytical solution. Systematic Biology 43:573-578.

McGrath ME, Mirzadegan T, Schmidt BF. 1997. Crystal structure of phenylmethanesulfonyl fluoride-treated human chymase at 1.9 Å. Biochemistry 36:14318-14324.

Mechelke M, Habeck M. 2010. Robust probabilistic superposition and comparison of protein structures. BMC Bioinformatics 11.

Meyer AG, Wilke CO. 2013. Integrating sequence variation and protein structure to identify sites under selection. Molecular Biology and Evolution 30:36-44.

Mitteroecker P, Bookstein F. 2011. Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. Evolutionary Biology 38:100-114.

Mitteroecker P, Gunz P, Bookstein FL. 2005. Heterochrony and geometric morphometrics: a comparison of cranial growth in *Pan paniscus* versus *Pan troglodytes*. Evolution & Development 7:244-258.

Mizuguchi K, Deane CM, Blundell TL, Overington JP. 1998. HOMSTRAD: A database of protein structure alignment for homologous families. Protein Science 7:2469-2471.

Monteiro LR. 1999. Multivariate regression models and geometric morphometrics: the search for causal factors in the analysis of shape. Systematic Biology 48:192-199.

Monteiro LR. 2013. Morphometrics and the comparative method: studying the evolution of biological shape. Hystrix.

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein structure coordinates. Proteins: Structure, Function and Genetics 12:345-364.

Nakayama D, Ben Ammar Y, Miyata T, Takeda S. 2011. Structural basis of coagulation factor V recognition for cleavage by RVV-V. FEBS Letters 585:3020-3025.

Neustupa J, Nēmcová Y. 2007. A geometric morphometric study of the variation in scales of Mallomonas striata (Synurophyceae, Heterokontophyta). Phycologia 46:123-130.

Neustupa J, Skaloud P, Stastny J. 2010. The molecular phylogenetic and geometric morphometric evaluation of *Micrasterias Crux-Melitensis/M-Radians* species complex. Journal of Phycology 46:703-714.

Neustupa J, Šťastný J. 2006. The geometric morphometric study of Central European species of the genus *Micrasterias* (*Zygnematophyceae, Viridiplantae*). Preslia 78:253-263.

O'Higgins P, Milne N. 2013. Applying geometric morphometrics to compare changes in size and shape arising from finite elements analyses. Hystrix.

Orengo CA, Todd AE, Thornton JM. 1999. From protein structure to function. Current Opinion in Structural Biology 9:374-382.

Perona JJ, Craik CS. 1995. Structural basis of substrate specificity in the serine proteases. Protein Science 4:337-360.

Perona JJ, Craik CS. 1997. Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. Journal of Biological Chemistry 272:29987-29990.

Piras P, Colangelo P, Adams DC, Buscalioni A, Cubo J, Kotsakis T, Meloro C, Raia P. 2010. The *Gavialis-Tomistoma* debate: the contribution of skull ontogenetic allometry and growth trajectories to the study of crocodylian relationships. Evolution & Development 12:568-579.

Polgár L. 2005. The catalytic triad of serine peptidases. Cellular and Molecular Life Sciences CMLS 62:2161-2172.

Porter ML, Blasic JR, Bok MJ, Cameron EG, Pringle T, Cronin TW, Robinson PR. 2012. Shedding new light on opsin evolution. Proceedings of the Royal Society B-Biological Sciences 279:3-14.

Poulíčková A, Veselá J, Neustupa J, Škaloud P. 2010. Pseudocryptic diversity versus cosmopolitanism in diatoms: a case study on *Navicula cryptocephala* Kütz. (Bacillariophyceae) and morphologically similar taxa. Protist 161:353-369.

Quan ML, Pinto DJP, Rossi KA, Sheriff S, Alexander RS, Amparo E, Kish K, Knabb RM, Luettgen JM, Morin P*, et al.* 2010. Phenyltriazolinones as potent factor Xa inhibitors. Bioorganic & Medical Chemistry Letters 20:1373-1377.

R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Rawlings ND, Barrett AJ. 1993. Evolutionary families of peptidases. Biochemical Journal 290:205-218.

Razeto A, Galunsky B, Kasche V, Wilson KS, Lamzin VS. High resolution structure of native bovine alpha-chymotrypsin. [To be Published].

Reiling KK, Krucinski J, Miercke LJ, Raymond WW, Caughey GH, Stroud RM. 2003. Structure of human pro-chymase: a model for the activating transition of granule-associated proteases. Biochemistry 42:2616-2624.

Rhodes G. 2010. Crystallography made crystal clear: A guide for users of macromolecular models. Burlington: Elsevier Science.

Rickert KW, Kelley P, Byrne NJ, Diehl RE, Hall DL, Montalvo AM, Reid JC, Shipman JM, Thomas BW, Munshi SK*, et al.* 2008. Structure of human prostasin, a target for the regulation of hypertension. Journal of Biological Chemistry 283:34864-34872.

Roach JC, Wang K, Gan L, Hood L. 1997. The molecular evolution of the vertebrate trypsinogens. Journal of Molecular Evolution 45:640-652.

Rohlf FJ. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations. Evolution 55:2143-2160.

Rohlf FJ, Corti M. 2000. Use of two-block partial least-squares to study covariation in shape. Systematic Biology 49:740-753.

Rose T, Di Cera E. 2002. Substrate recognition drives the evolution of serine proteases. The Journal of Biological Chemistry 277:19243-19246.

Russell RB, Saqi MAS, Sayle RA, Bates PA, Sternberg JE. 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. Journal of Molecular Biology 269:423-439.

Sanger TJ, Sherratt E, McGlothlin JW, Brodie ED, Losos JB, Abzhanov A. 2013. Convergent evolution of sexual dimorphism in skull shape using distinct developmental strategies. Evolution 67:2180-2193.

Sichler K, Banner DW, D'Arcy A, Hopfner KP, Huber R, Bode W, Kresse GB, Kopetzki E, Brandstetter H. 2002. Crystal structures of uninhibited factor VIIa link its cofactor and substrate-assisted activation to specific interactions. Journal of Molecular Biology 322:591-603.

Sidlauskas BL, Mol JH, Vari RP. 2011. Dealing with allometry in linear and geometric morphometrics: a taxonomic case study in the *Leporinus cylindriformis* group (Characiformes: Anostomidae) with description of a new species from Suriname. Zoological Journal of the Linnean Society 162:103-130.

Siltberg-Liberles J, Grahnen JA, Liberles DA. 2011. The evolution of protein structures and structural ensembles under functional constraint. Genes 2:748-762.

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. 2005. Evolutionary information for specifying a protein fold. Nature 437:512-518.

Sun J, Tang S, Xiong W, Cong P, Li T. 2012. DSP: a protein shape string and its profile prediction server. Nucleic Acids Research 40:W298-W302.

Terayama R, Bando Y, Yamada M, Yoshida S. 2005. Involvement of neuropsin in the pathogenesis of experimental autoimmune encephalomyelitis. Glia 52:108-118.

Theobald DL, Steindel PA. 2012. Optimal simultaneous superpositioning of multiple structures with missing data. Bioinformatics 28:1972-1979.

Theobald DL, Wuttke DS. 2005. Divergent evolution within protein superfolds infered from profile-based phylogenetics. Journal of Molecular Biology 354:722-737.

Theobald DL, Wuttke DS. 2006a. Empirical Bayes hiearchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. Proceedings of the National Academy of Sciences of the United States of America 103:18521-18527.

Theobald DL, Wuttke DS. 2006b. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics 22:2171-2712.

Theobald DL, Wuttke DS. 2008. Accurate structural correlations from maximum likelihood superpositions. Plos Computational Biology 4.

Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. 2000. From structure to function: approaches and limitations. Nature Structural Biology 7:991-994.

Todd AE, Orengo CA, Thornton JM. 1999. Evolution of protein function, from a structural perspective. Current Opinion in Chemical Biology 3:548-556.

Toyota E, Iyaguchi D, Sekizaki H, Tateyama M, Ng KKS. 2009. A structural comparison of three isoforms of anionic trypsin from chum salmon (*Oncorhynchus keta*). Acta Crystallographica Section D 65:717-723.

Trifonov EN, Frenkel ZM. 2009. Evolution of protein modularity. Current Opinion in Structural Biology 19:335-340.

Voje KL, Hansen TF, Egset CK, Bolstad GH, Pélabon C. 2014. Allometric constraints and the evolution of allometry. Evolution 68:866-885.

Watanabe A, Slice DE. 2014. The utility of cranial ontogeny for phylogenetic inference: a case study in crocodylians using geometric morphometrics. Journal of Evolutionary Biology 27:1078-1092.

Wątorek W. 2003. Azurocidin - inactive serine proteinase homolog acting as a multifunctional inflammatory mediator. Acta Biochimica Polonica 50:743-752.

Williams SG, Lovell SC. 2009. The effect of sequence evolution on protein structural divergence. Molecular Biology and Evolution 26.

Wlodawer A, Minor W, Dauter Z, Jaskolski M. 2008. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. The FEBS journal 275:1-21.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. Science 319:473-476.

Word JM, Lovell SC, RIchardson JS, Richardson DC. 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of Molecular Biology 285:1735-1747.

Yousef GM, Kopolovic AD, Elliott MB, Diamandis EP. 2003. Genomic overview of serine proteases. Biochemical and Biophysical Research Communications 305:28-36.

Yousef GM, Obiezu CV, Luo LY, Magklara A, Borgoño CA, Kishi T, Memari N, Michael IP, Sidiropoulos M, Kurlender L*, et al.* 2005. Human tissue kallikreins: From gene structure to function and clinical applications. Advances in clinical chemistry 39:11-79.

Zelditch ML, Swiderski DL, Sheets HD. 2012. Geometric morphometrics for biologists: a primer. San Diego, Academic Press.

Zeng F, Li X, Teng M, Niu L. Crystal structure of AhV_TL-I, a glycosylated snake-venom thrombin-like enzyme from *Agkistrodon halys*. [To be Published].

5. General Discussion

5.1. Shape and phylogenetics

Shape data does not provide reliable phylogenetic estimates (chapter 1). There are different factors that cause problems for the reliable estimations of phylogenies, starting with the evolutionary model. Shape is one part of the phenotype that is important in the evolution of the structures and has been frequently a target for natural selection at a microevolutionary level (Gómez, et al. 2006, Martínez-Abadías, et al. 2012). At a macroevolutionary level, stabilizing selection seems to be the prominent evolutionary model in nature (Estes and Arnold 2007, Hunt 2007, Haller and Hendry 2014). Stabilizing selection is one expected generator of convergence, given that it 'pulls' the different species in the shape space towards one specific point, the optimum (Hansen 1997, Butler and King 2004). This locates the species closer in the shape space and makes it easier for the random component of the evolution, inherent to any finite population, to shuffle the relative position of the species. Brownian motion is another model which has been found in paleontological data (Hunt 2007), although less frequently. This is the model expected to apply when populations evolve by genetic drift, as shown by quantative genetics theory (Lande 1979). It does not have any 'pulling' effect and the expected distance among lineages increases with time (Lynch 1989). Therefore, it is considered a favourable model for estimating phylogenies. However, this expected distance between different lineages has much variance and that fact can result in poor estimates (Stayton 2008).

Our simulations suggest that the estimates of phylogenetic reliability assuming an isotropic Brownian motion model of evolution are related to the shape dimensionality. Homoplasy is more likely when the dimensionality of the shape data is low (chapter 1). This is the case of some studies in geometric morphometrics where landmarks are sometimes difficult to find and

therefore low dimensionality (around 15 dimensions) is used (Cole, et al. 2002, Cardini 2003, Guill, et al. 2003, Moraes, et al. 2004, Nagorsen and Cardini 2009). In principle, however, there is a good group of studies in which the reliability should be fairly good assuming isotropic variation, since it is not very uncommon to find more than one hundred dimensions in morphometric datasets (Marcus, et al. 2000, Cardini and Elton 2008, González-José, et al. 2008, Piras, et al. 2010, von Cramon-Taubadel and Smith 2012). The problem is that even in the cases in which the dimensionality is very high, morphological integration can significantly decrease the reliability of the phylogenies (chapter 1).

Morphological integration plays an important role in the explanation of the shape diversification of the species (Olson and Miller 1958, Cheverud 1996, Wagner and Altenberg 1996, Monteiro, et al. 2005, Young and Badyaev 2006, Hallgrímsson, et al. 2009, Klingenberg 2013, Armbruster, et al. 2014, Goswami, et al. 2014). The covariation between different parts of a structure is known as integration and it can be seen as a concentration of the variation in the shape space during the evolution of the species (Wagner 1984, Young 2006), or in other words, the fact that evolution happens more frequently in specific directions of the shape space than in others. If integration is extremely high the shape space approaches to a univariate space, where the phylogenetic reliability has been theoretically shown to be very poor (chapter 1) (Lynch 1989). Although extreme, some morphometric empirical datasets approach to that situation (Klingenberg, et al. 2012). Indeed, the degree of integration is usually strong in empirical studies (chapter 2), so the intuition about integration as a source of convergence is justified (Losos 2011). Integration is a widespread force in nature (Olson and Miller 1958, Cheverud 1996, Wagner and Altenberg 1996, Marroig and Cheverud 2001, Bookstein, et al. 2003, Monteiro, et al. 2005, Young and Badyaev 2006, Lockwood 2007, Klingenberg 2008, Hallgrímsson, et al. 2009, Gonzalez, et al. 2011,

Klingenberg 2013, Armbruster, et al. 2014, Goswami, et al. 2014), frequent in both anatomical and molecular structures (chapter 2).

This thesis contributes then not only with a theoretical suggestion about the reliability of estimating phylogenies using shape data. It also shows with solid evidence that the conditions of evolutionary models (Hunt 2007) and strength of integration (chapter 2) assumed in the theoretical part of the work are met in empirical data. That gives decisive information about why geometric morphometrics data is not suitable for phylogenetic reconstruction.

5.2. Integration and evolution

The biological origins of integration are very diverse: functional and environmental pressures affecting the genetic architecture (pleiotropy, linkage disequilibrium) or developmental processes are the most common of them (Klingenberg 2014). The biological role of integration has been suggested to be the maintenance of specific features that are essential in the shape during evolution (Klingenberg 2008). Under this perspective the limitation of the occupancy in some directions of the shape specific would reveal the presence of essential features for survival and reproduction in these areas. Integration would also cause the reorganization of the whole shape of one structure when selection affects just a limited part of it (Albertson, et al. 2003, Martínez-Abadías, et al. 2012), although whether this reorganization follows the direction of selection or not is controversial (Marroig, et al. 2009, Goswami, et al. 2014).

Anisotropic variation has important implications for the evolution of the species in the shape space. The directions of the shape space in which more evolutionary variation is concentrated are seen as directions of the shape space in which evolution is 'easier'. These directions are known as lines of least resistance (Schluter 1996, Renaud, et al. 2006). The implications of these patterns of variation over evolution are heavily conditioned by their

orientation and total variance (chapter 2). The total variance under Brownian motion is related to the potential for evolution (Felsenstein 2004). On average, the distance from the starting point of the evolution increases with the variance (chapter 1). The species with more total variance have more opportunity for evolution. In some species under artificial selection, an abnormally huge total variance can be found (Drake and Klingenberg 2010). These species would have especially big opportunity for adaptation. However, the direction in which the species evolve in the shape space does not depend on integration. In the extreme, two species (with completely integrated phenotypes) would drift in a multidimensional phenotypic space along one line each, but these lines (the trajectories of each species) can have a similar orientation or a completely different one (perpendicular) (chapter 2). In both cases, the species phenotypes are completely integrated. Therefore, two species drifting can approach or move away from each other in the phenotypic space in presence of strong integration depending on the orientation of their lines of least resistance. In presence of complete integration and perpendicular orientations the species would move away from each other (chapter 2). In presence of complete integration and similar orientations the species would approach. The similarity among integration orientations in different species in nature is something controversial and that ultimately refers to what we know about the evolution of integration. This thesis has studied the patterns of integration as a stable feature during the evolutionary diversification. This is based on the fact that close related species would share similar genetic and developmental features and therefore the patterns of integration would be similar. However, there is little evidence about how general this supposition is.

5.3. Shape and proteins

The evolution of the proteins has some differences with the evolution of anatomical structures. The standard model of evolution in these molecules implies the duplication of the proteins and the subfunctionalization of the new molecules (Lynch 2000). This loss of function and expression in the new proteins could relax their shape integration. There were reasons to think that proteins could have shown a low degree of integration, also because the dimensionality is huge (chapter 3) and therefore there are more sources of variation to 'compensate' the functional demands of certain parts of the structure in charge of specific functions. Indeed, when we compare the decreasing in variation of each PC of the independent contrasts of shape in this structure with this decreasing in some anatomical structures, the integration found in these structures is relatively low (chapter 2). However, they still present an exponential decrease of variation (chapter 2), which has been shown to promote convergence relatively highly (chapter 1).

Convergence is found in protein shape, something that is caused by functional demands (chapter 3) (Todd, et al. 1999, Thornton, et al. 2000, Goldstein 2008, Dessailly, et al. 2009, Goldstein and Pollock 2012). Although natural selection is probably behind some this convergence in the structure of the proteins (Siltberg-Liberles, et al. 2011), integration can also enhance it (chapter 3). Differences in size, mediated by differences in the folding, are behind some of this integration (chapter 3).

The last chapter of this thesis provides a new perspective on the study of the evolution of the proteins as well as on the application of geometric morphometrics. The results suggest a similar situation to that obtained in some anatomical features, e.g. in some anatomical structures as skulls in which there are some parts of them especially linked to function, like mandible (Figueirido, et al. 2013). Integration is then certainly present in these structures. However, the difference between the evolution of these

molecules and the evolution of anatomical structures (e.g. their evolution mediated by duplications or abrupt changes in size) suggest that there may be different patterns in the evolution of integration and the evolution of allometry (chapters 3). More work specifically looking at these features with better sample sizes is needed to disentangle their role in the evolution of protein shape.

5.4. Future work

There are conflicting hypotheses about how the patterns of integration evolve, resulting in assumptions in our work that are of importance for the results, e.g. orientation of the main axes of variation (chapter 1 & chapter 2). As in the results of Goswami et al. (2014), our results also remark 'the importance of considering the exact pattern of trait covariances in predicting long-term trait evolution' (p. 6). More work is needed in this area, so more reliable assumptions can be made. In the case that the evolution of integration varies much among different families of species, the probability of convergence for different clades could be estimated. That would bring interesting cases from the biological point of view, so they can explain different pressures that can shape covariance matrices, e.g. through functional demands (chapter 3), may be acting.

The extensive research in the integration patterns in proteins would bring much information about the differences between the patterns of integration in anatomical and molecular structures. Protein structures are divided in discrete parts (the amino acids) and their effect over the integration patterns can be analysed, so experimentation in the relationship between integration and modularity can be set up. Along with tests for the different modularity and integration methods (Klingenberg 2013) and the study of the link with different biological processes like folding or interaction between proteins, the research in this area could bring many possibilities.

5.5. References

Albertson RC, Streelman JT, Kocher TD. 2003. Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. Proceedings of the National Academy of Sciences of the United States of America 100:5252-5257.

Armbruster WS, Pélabon C, Bolstad GH, Hansen TF. 2014. Integrated phenotypes: understanding trait covariation in plants and animals. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Bookstein FL, Gunz P, Mitteroecker P, Prossinger H, Schaefer K, Seidler H. 2003. Cranial integration in Homo: singular warps analysis of the midsagittal plane in ontogeny and evolution. Journal of Human Evolution 44:167-187.

Butler MA, King AA. 2004. Phylogenetic comparative analysis: A modelling approach for adaptive evolution. The American Naturalist 164:683-695.

Cardini A. 2003. The geometry of the marmot (Rodentia : Sciuridae) mandible: Phylogeny and patterns of morphological evolution. Systematic Biology 52:186-205.

Cardini A, Elton S. 2008. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. Biological Journal of the Linnean Society 93:813-834.

Cheverud JM. 1996. Developmental integration and the evolution of pleiotropy. American Zoologist 36:44-50.

Cole TM, III., Lele SR, Richtsmeier JT. 2002. A parametric bootstrap approach to the detection of phylogenetic signals in landmark data. In: MacLeod N, Forey PL editors. Morphology, shape and phylogeny. London, Taylor & Francis, p. 194–219.

Dessailly BH, Redfern OC, Cuff A, Orengo CA. 2009. Exploiting structural classifications for function prediction: towards a domain grammar for protein function. Current Opinion in Structural Biology 19:349-356.

Drake AG, Klingenberg CP. 2010. Large-scale diversification of skull shape in domestic dogs: disparity and modularity. American Naturalist 175:289–301.

Estes S, Arnold SJ. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. American Naturalist 169:227-244.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA, Sinauer Associates.

Figueirido B, Tseng ZJJ, Martín-Serra A. 2013. Skull shape evolution in durophagous carnivorans. Evolution.

Goldstein RA. 2008. The structure of protein evolution and the evolution of protein structure. Current Opinion in Structural Biology 18:170-177.

Goldstein RA, Pollock DD. 2012. Modeling protein evolution. Computational Modeling of Biological Systems:311-325.

Gómez JM, Perfectti F, Camacho JPM. 2006. Natural selection on *Erysimum mediohispanicum* flower shape: Insights into the evolution of zygomorphy. American Naturalist 168:531-545.

Gonzalez PN, Oyhenart EE, Hallgrímsson B. 2011. Effects of environmental perturbations during postnatal development on the phenotypic integration of the skull. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 316B:547-561.

González-José R, Escapa I, Neves WA, Cúneo R, Pucciarelli HM. 2008. Cladistic analysis of continuous modularized traits provides phylogenetic signals in Homo evolution. Nature 453:775-778.

Goswami A, Smaers JB, Soligo C, Polly PD. 2014. The macroevolutionary consequences of phenotypic integration: from development to deep

time. Philosophical Transactions of the Royal Society of London B Biological Sciences 369:20130254.

Guill JM, Heins DC, Hood CS. 2003. The effect of phylogeny on interspecific body shape variation in darters (Pisces : Percidae). Systematic Biology 52:488-500.

Haller BC, Hendry AP. 2014. Solving the paradox of stasis: squashed stabilizing selection and the limits of detection. Evolution 62:483-500.

Hallgrímsson B, Jamniczky HA, Young NM, Rolian C, Parsons TE, Boughner JC, Marcucio RS. 2009. Deciphering the palimpsest: Studuying the relationship between morphological integration and phenotypic covariation. Evolutionary Biology 36:355-376.

Hansen TF. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341-1351.

Hunt G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proceedings of the National Academy of Sciences of the USA 104:18404–18408.

Klingenberg CP. 2008. Morphological integration and developmental modularity. Annual Review of Ecology, Evolution and Systematics 39:115-132.

Klingenberg CP. 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. Hystrix 24:43-58.

Klingenberg CP. 2014. Studying morphological integration and modularity at multiple levels: concepts and analysis. Philosophical Transactions of the Royal Society of London B Biological Sciences 369.

Klingenberg CP, Duttke S, Whelan S, Kim M. 2012. Developmental plasticity, morphological variation and evolvability: a multilevel analysis of morphometric integration in the shape of compound leaves. Journal of Evolutionary Biology 25:115–129.

Lande R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

Lockwood CA. 2007. Adaptation and functional integration in primate phylogenetics. Journal of Human Evolution 52:490-503.

Losos JB. 2011. Convergence, adaptation, and constraint. Evolution 65:1827-1840.

Lynch M. 1989. Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation. Evolution 43:1-17.

Lynch M. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

Marcus LF, Hingst-Zaher E, Zaher H. 2000. Application of landmark morphometrics to skulls representing the orders of living mammals. Hystrix 11:27–47.

Marroig G, Cheverud JM. 2001. A comparison of phenotypic variation and covariation patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of New World monkeys. Evolution 55:2576-2600.

Marroig G, Shirai LT, Porto A, de Oliveira FB, De Conto V. 2009. The evolution of modularity in the mammalian skull II: evolutionary consequences. Evolutionary Biology 36:136-148.

Martínez-Abadías N, Esparza M, Sjøvold T, Gonzalez-Jose R, Santos M, Hernández M, Klingenberg CP. 2012. Pervasive genetic integration directs the evolution of human skull shape. Evolution 66:1010-1023.

Monteiro LR, Bonato V, dos Reis SF. 2005. Evolutionary integration and morphological diversification in complex morphological structures: Mandible shape divergence in spiny rats (Rodentia, Echimyidae). Evolution & Development 7:429–439.

Moraes EM, Spressola VL, Prado PRR, Costa LF, Sene FM. 2004. Divergence in wing morphology among sibling species of the *Drosophila buzzatti*

cluster. Journal of Zoological Systematics and Evolutionary Research 42:154-158.

Nagorsen DW, Cardini A. 2009. Tempo and mode of evolutionary divergence in modern and Holocene Vancouver Island marmots (*Marmota vancouverensis*) (Mammalia, Rodentia). Journal of Zoological Systematics and Evolutionary Research 47:258-267.

Olson EC, Miller RL. 1958. Morphological integration. Chicago, University of Chicago Press.

Piras P, Colangelo P, Adams DC, Buscalioni A, Cubo J, Kotsakis T, Meloro C, Raia P. 2010. The *Gavialis-Tomistoma* debate: the contribution of skull ontogenetic allometry and growth trajectories to the study of crocodylian relationships. Evolution & Development 12:568-579.

Renaud S, Auffray JC, Michaux J. 2006. Conserved phenotypic variation patterns, evolution along lines of least resistance, and departure due to selection in fossil rodents. Evolution 60:1701-1717.

Schluter D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766-1774.

Siltberg-Liberles J, Grahnen JA, Liberles DA. 2011. The evolution of protein structures and structural ensembles under functional constraint. Genes 2:748-762.

Stayton CT. 2008. Is convergence surprising? An examination of the frequency of convergence in simulated datasets. Journal of Theoretical Biology 252:1-14.

Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. 2000. From structure to function: approaches and limitations. Nature Structural Biology 7:991-994.

Todd AE, Orengo CA, Thornton JM. 1999. Evolution of protein function, from a structural perspective. Current Opinion in Chemical Biology 3:548-556.

von Cramon-Taubadel N, Smith HF. 2012. The relative congruence of cranial and genetic estimates of hominoid taxon relationships: Implications for the reconstruction of hominin phylogeny. Journal of Human Evolution 62:640-653.

Wagner GP. 1984. On the eigenvalue distribution of genetic and phenotypic dispersion matrices: evidence for a nonrandom organization of quantitative character variation. Journal of Mathematical Biology 21:77-95.

Wagner GP, Altenberg L. 1996. Complex adaptations and the evolution of evolvability. Evolution 50:967-976.

Young NM. 2006. Function, ontogeny, and canalization of shape variance in the primate scapula. Journal of Anatomy 209:623-636.

Young RL, Badyaev AV. 2006. Evolutionary persistence of phenotypic integration: Influence of developmental and functional relationships on complex trait evolution. Evolution 60:1291-1299.