PhyloPars: estimation of missing parameter values using phylogeny

Jorn Bruggeman1*, Jaap Heringa2 and Bernd W. Brandt2

1Department of Theoretical Biology, Faculty of Earth & Life Sciences and 2Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, The Netherlands

ABSTRACT

A wealth of information on metabolic parameters of a species can be inferred from observations on species that are phylogenetically related. Phylogeny-based information can complement direct empirical evidence, and is particularly valuable if experiments on the species of interest are not feasible. The PhyloPars web server provides a statistically consistent method that combines an incomplete set of empirical observations with the species phylogeny to produce a complete set of parameter estimates for all species. It builds upon a state-of-the-art evolutionary model, extended with the ability to handle missing data. The resulting approach makes optimal use of all available information to produce estimates that can be an order of magnitude more accurate than ad-hoc alternatives. Uploading a phylogeny and incomplete feature matrix suffices to obtain estimates of all missing values, along with a measure of certainty. Real-time cross-validation provides further insight in the accuracy and bias expected for estimated values. The server allows for easy, efficient estimation of metabolic parameters, which can benefit a wide range of fields including systems biology and ecology. PhyloPars is available at: http://www.ibi.vu.nl/programs/phylopars/.

INTRODUCTION

Quantitative predictions on the behavior of organisms, populations and ecosystems require accurate values for the metabolic parameters associated with cellular and physiological processes. These parameters range from the substrate affinities, processing rates and yield coefficients of single enzyme-mediated chemical reactions, to the growth rate and maintenance requirement of individual organisms or populations. Ideally, such parameters are measured directly by observing the species (or strain) of interest in experiments. However, this can be a costly, time-consuming and difficult process. It may even be impossible: many species that play key roles in nature have never been successfully kept in culture or captivity. In such contexts, methods that exploit alternative information sources to estimate parameter values can be very valuable. One such source of information is found in direct observations on phylogenetically related species: given that evolution changes most features gradually over time, two species that separated recently in evolutionary history would be expected to behave similarly. The best estimate for a metabolic parameter could thus be derived from observations on related species.

Intuitively, one might expect that the best estimate for a missing parameter is given by the observed value of the most closely related species available. Certainly, evolutionary models generally agree that the best estimate for the parameter of a species is given by the closest node possible: the true parameter value of its parent in the phylogeny (1). However, this value is usually unknown. First, the parent may be extinct, in which case we can only obtain indirect information on its original parameter value by sampling descendants. Second, most observations are samples subject to measurement error, rather than perfect measurements of the true value. Therefore, any empirically available estimate of the parent value is subject to error. We can reduce this error by considering more distantly related species in the phylogeny: every sampled species in the phylogeny tells a little about the possible value of the missing parameter. An optimal reconstruction of the parameter value thus combines observations throughout the phylogeny, weighed according to phylogenetic proximity. This is a dominant idea in comparative feature analysis (2,3) and ancestral state reconstruction (4–6), and occasionally surfaces in other applications such as sequence alignment (7,8) and comparative sequence analysis (9).

Another source of information on the value of an missing parameter may be found in other observations on the species of interest: its other features may tell a lot about the missing value. A prime example is the maximum body size of species, which is commonly associated with a
the value 1 observed for species D ('nearest neighbor model'). ('mean model'), or insert the phylogenetically nearest observed value:
of the other observed feature values (red numbers only) as estimate
make such complete use of available information: they use the average
based on all observations. Simple ad-hoc approaches typically do
netic correlation between the two features. This estimate is in turn
of the other feature (blue value 5) through its estimate of the phyloge-
phylogenetic proximity, and additionally involves the observed value
observed values of that feature (red numbers) weighed according to
mark). PhyloPars reconstructs the missing value from the other
features (red and blue numbers) and one missing value (question
known in terms of both topology and branch lengths; the
plethora of features through theory for metabolic organi-
ization (10) and empirical allometric scaling laws (11). A
good indication of the value of certain metabolic
parameters could therefore be given by the size of the
species, or in fact any observed parameter that is known
correlate with the feature of interest. It is tempting to
identify and use such correlations directly through regres-
sion analysis applied to observations across species. However,
such observations are not independent due to
interspecific phylogenetic relationships, leading to overes-
timation of the correlations (3). Information on parameter
correlations can provide a valuable aid for reconstructing
missing parameter values, but strikingly, their identifica-
tion and use again requires a phylogeny-aware analysis.
The PhyloPars web server offers an efficient, statistically
consistent method that exploits both phylogenetic rela-
tionships and parameter correlations to estimate missing
parameters of the species within a phylogeny (Figure 1).
It builds upon a state-of-the-art model (12,13) that addi-
tionally accounts for intraspecific variability and measure-
ment error. This model is extended to handle missing data.
The resulting approach makes optimal use of all available
observations to produce estimates that can be an order of
magnitude more accurate than ad-hoc alternatives.

Underlying model
PhyloPars operates on a phylogeny and an incomplete
feature matrix that describes the available observations
on one or more continuous features (e.g. metabolic par-
eters), for subsets of nodes (species or strains) that may
fully or partially overlap. The phylogeny is assumed to be
known in terms of both topology and branch lengths; the
latter are assumed to be proportional to evolutionary
time.

Phylogenetic variability
The role of the phylogeny is represented by a ‘Brownian
motion’ phylogenetic model that assumes feature values
change through genetic drift (1,2,9,12,14). A consequence
of the model is that all feature values for all nodes com-
bined can be described by a single multivariate normal
distribution. The covariances of the distribution depend
on the topology and branch lengths of the phylogeny,
and on the rates and correlations of the evolution of
the different features. These rates and correlations are
described by N(N + 1)/2 ‘phylogenetic’ covariances (12),
with N denoting the number of features. They are easily
transformed into more readily interpretable phylogenetic
regression slopes (15,16) and phylogenetic correlations
(12). It is worth noting that the term ‘phylogenetic corre-
lation’ has been used to describe different concepts (12,17);
we follow Felsenstein in using it exclusively to refer to
joint evolution of a pair of features.

Phenotypic variability
The phylogenetic model is extended to account for
phenotypic variability: the fact that a single sample is
not necessarily the mean of the species under study
(12,13). Such variability may be due to measurement
error or to intraspecific variation of the feature. It can
be incorporated in the phylogenetic model by introducing
a layer of variability between the species level and the
observation, effectively behaving as extension of the
evolutionary path (12). This is comparable to the role of
non-heritable residual variability in phylogenetic mixed
models (2,18).

We assume that observed correlations between features
are exclusively due to evolutionary processes (and not,
for instance, to the measurement process itself). This facil-
itates extending the model with the ability to use incom-
plete observations, i.e. observations on a species that
include only a subset of all features. As observations
may originate from different sources, with sets of observed
species per feature not necessarily overlapping, this func-
tionality is indispensable for many purposes. Phenotypic
variability is assumed equal for all species (but can differ
between features), and is described by N unknown ‘phe-
notypic’ variances (12). It may be noted that this differs
from the approach taken by Ives et al. (13), who permit
phenotypic variability to differ between species (as well as
between features); values for the phenotypic variability are
there prescribed rather than estimated.

Procedure
Both the phylogenetic covariances and the phenotypic
variances are initially unknown and need to be estimated
from the input data in order to reconstruct missing fea-
ture values. This is typically done through maximum
likelihood estimation (MLE) (2,12–14).

With the phylogenetic and phenotypic covariances
known, estimation of missing parameters is straight-
forward: the optimal phylogenetic and phenotypic
Phytoplankton represents the lowest tropic level in aquatic food webs, and as such govern processes on all scales: from small lakes to the global climate. Their influence is hard to quantify as many plankton species cannot be cultured. This therefore presents an ideal test case for the PhyloPars approach.

The dataset was compiled by J.B. from 38 literature sources and contains over one thousand measurements on 12 different features of 114 species. For the present study we have selected a subset of features: cell length, diameter, surface area, volume, maximum growth rate and phosphate affinity. It may be obvious that the first four features all describe aspects of cell size, and are likely to be positively correlated. Their joint inclusion is intentional: it demonstrates the ability of PhyloPars to uncover and exploit correlations between features.

To our knowledge, there does not exist a complete phytoplankton phylogeny based on molecular evidence. Therefore, we resort to using the Linnaean taxonomy with branch lengths of 1 between ranks. While this is undoubtedly a very crude approximation to the topology and branch lengths of the true phylogeny, other work indicates that even such qualitative phylogenies can contribute information on feature evolution (22).

The sample feature matrix contains a total of 289 observed values for a total of 242 phylogenetic groups (species and ancestors), leaving 1163 missing values. Both the feature matrix and phylogeny are available at the PhyloPars home page. Results obtained with default settings are shown in Figure 2.

Phylogenetic and phenotypic variability

In the first section of its results (Figure 2A), PhyloPars presents maximum likelihood estimates for the phylogenetic and phenotypic variability. Phylogenetic standard deviations provide a measure of the rate of feature evolution. Phenotypic standard deviations quantify the variability due to measurement error and/or intraspecific variation. The proportion of variance accounted for by the phylogeny is also presented, and may be used to compare phylogenetic and phenotypic contributions to the total observed variability (18). This proportion usually includes a contribution by natural selection (23); further analyses (24,25) might be used to disentangle phylogenetic and selection components. For reference, the table also lists a summary of cross-validation results (see below). Finally, phylogenetic correlations are shown; these indicate if pairs of features are likely to co-evolve.

For the phytoplankton example phylogenetic variability clearly plays an important role: the phylogeny explains more than 50% of the total variability of all features. It is also apparent that PhyloPars correctly recognizes likely correlations between features: correlations between length, diameter, surface area and volume all exceed 0.5.

Estimates for missing values

In the second result section (Figure 2B), estimates for all features of all nodes are presented in one single ‘supplemented’ feature matrix. Feature values that are also present in the input data as observation are indicated by covariances are first combined with the tree topology to calculate the covariances between the observations and the missing values. These are subsequently used to express the estimate of each missing value as the product of all original observations and an estimate-specific set of associated weights (formally: regression coefficients).

In order to assess the validity of the model result, one can additionally perform cross-validation: each observed parameter is excluded from the input data, and then re-estimated with the MLE-derived phylogenetic and phenotypic covariances to determine prediction error and bias.

WEB SERVER IMPLEMENTATION

Input

The web server accepts an uploaded phylogeny in Newick format (http://evolution.genetics.washington.edu/phylip/newicktree.html) and a feature matrix with observations as tab-separated text file. The latter can contain missing values. If there is good evidence that either phylogenetic correlations or phenotypic variability are absent in the input dataset, the user can additionally disable correlated evolution of features (phylogenetic correlations will be set to zero) or phenotypic variability (phenotypic variances will be set to zero), respectively. This restricts the freedom of the model and will then correctly decrease the uncertainty associated with estimated missing values.

Processing

The web server first constructs the full multivariate normal model that specifies the likelihood of observing the provided feature matrix, given the phylogeny, phylogenetic covariances and phenotypic variances. The covariance matrix of the model combines phylogenetic and phenotypic components in such a way that analytic calculation of the optimal phylogenetic and phenotypic covariances is not possible (12,13). Therefore, we resort to iterative numerical maximization of the likelihood. Phylogenetic and phenotypic covariances are first transformed into a set of unbounded parameters through log Cholesky parameterization (19); this permits unconstrained optimization. The likelihood is then maximized with the Broyden-Fletcher-Goldfarb-Shanno algorithm (20).

All processing logic is implemented in Python (http://www.python.org). For optimization and advanced linear algebra we use SciPy (http://www.scipy.org), which in turn encapsulates LAPACK (21). All output plots are generated with MatPlotLib (http://matplotlib.sourceforge.net). A mathematical description of the methodology is provided as online Supplementary Data. Total processing time depends on the number of features and nodes under study; a test case with 242 nodes, 6 features and 289 observations is processed in under 3 min.

WORKED EXAMPLE

We have applied the PhyloPars method to an extensive database of freshwater phytoplankton parameters. Phytoplankton represents the lowest tropic level in aquatic
Figure 2. Output of the PhyloPars web server consists of three sections: (A) estimates for phylogenetic and phenotypic variability, (B) estimates for missing parameter values, and (D) cross-validation details. Individual parameter estimates can be clicked to obtain a detailed report in a pop-up window; an example for the maximum growth rate of *Chroococcus* is shown in (C).
a trailing asterisk. It is worth noting that the estimated value of a parameter may differ from its original observation when phenotypic variability is allowed: in that case the observation is a single sample, whereas the supplemented feature matrix lists the expected (mean) value for the representative species. The supplemented feature matrix is also available as a downloadable tab-separated text file.

Clicking on a value in the supplemented feature matrix opens a detailed report (Figure 2C) that visualizes the contributions of all observations to the point estimate (black vertical bars and blue area). Additionally, it presents the standard deviation of the estimate. Together the point estimate and standard deviation specify the marginal likelihood of the estimated value (a normal distribution), plotted as a red curve.

Results for the example clearly demonstrate that estimates generally differ even for closely related species (e.g. Anabaena sp.)—this is a direct result of the PhyloPars capability to include observations on other features in its estimates through phylogenetic correlations.

Cross-validation
In the last result section (Figure 2D), detailed cross-validation reports are presented. These visualize the distribution of estimation errors, i.e. the differences between observations and their estimates in cross-validation. Error distributions are also plotted for two simple ad-hoc models: a mean model that assumes the best estimate for a missing value is given by the mean of all observations (valid if phylogenetic variability is absent), and a nearest neighbor model that assumes the best estimate for a missing value is given by the phylogenetically closest observation (a common method of estimating unknown parameters). These allow the user to judge the improvement of the PhyloPars model over the ad-hoc models. Values are also presented for the expected bias (the mean of all differences between estimates and observations in cross-validation), and the expected error (the mean of all absolute differences). These provide an indication of the accuracy of the estimates for missing values.

For the example, neither of the ad-hoc models has a definite advantage over the other, and the PhyloPars evolutionary model always improves upon both. This improvement can be very large: the mean error in the estimate for cell surface area is 25% compared to over 300% for the ad-hoc models. Additionally, the bias of the PhyloPars model is near zero for all features: a definite advantage over the other, and the PhyloPars model has no bias by definition, but its errors are relatively large.

DISCUSSION AND CONCLUSION
PhyloPars delivers estimates of missing feature values that can be an order of magnitude more accurate than those of ad-hoc alternatives. The maximum precision achievable depends in part on the accuracy of the topology and branch lengths of the user-supplied phylogeny. For a limited number of species, accurate phylogenies based on molecular evidence are available. For instance, TreeFam (26) and Pfam (27) offer phylogenetic trees based on gene and protein similarity, respectively. Branch lengths can then to some extent be expected to represent evolutionary time. If a detailed phylogeny is not available for the species of interest, a qualitative tree derived from resources such as the NCBI taxonomy (28) could be used instead. The worked example demonstrates convincingly that even a taxonomy-based phylogeny with arbitrary branch lengths allows PhyloPars to improve considerably upon alternative models.

A valid concern is to what extent the accuracy of PhyloPars predictions depends on the underlying ‘Brownian motion’ evolutionary model. It has been argued that this model overemphasizes the randomness of evolutionary change, at the expense of directional change due to natural selection (23). Not surprisingly, several alternative models of evolution have been proposed (17,29). However, the mathematical framework that underpins the Brownian motion model can be motivated independently on first-principle statistical grounds (15,30). Accordingly, the model has been shown to deliver accurate predictions even for data generated with alternative evolutionary models (31). If no detailed information is available on the processes that governed evolutionary change, the Brownian motion model provides a robust base model for phylogenetic analyses.

Within the context of the Brownian motion evolutionary model, the best estimate for a feature is given by that of its parent in the phylogeny. Estimating missing feature values thus equates to reconstructing ancestral states. Several stand-alone applications are capable of this, e.g. Pagel’s BayesTraits (http://www.evolution.reading.ac.uk), Swofford’s PAUP (http://paup.csit.fsu.edu), and Mesquite with the PDAP:PDTREE package (http://mesquiteproject.org). However, none of these applications incorporates phenotypic variability of features, which plays an important role in practice (12,13), and several do not accept datasets with missing values or cannot process these reliably. The ‘contrast’ program in Felsenstein’s PHYLIPI (http://evolution.genetics.washington.edu/phylip.html) deserves special mention, as it performs all preprocessing needed for subsequent ancestral state reconstruction and can handle phenotypic variability. However, it again does not permit missing values, which for our example would imply that only 18 out of 289 observations (3 species out of 114) could be used.

To our knowledge there does not exist a stand-alone or web application that offers a straightforward means of performing the complex task of phylogeny-based reconstruction of missing parameter values, comparable in ease of use and visual support to that offered by the PhyloPars web server. PhyloPars for the first time discloses valuable, theoretically advanced methods from evolutionary biology to experimentalists and modelers alike. Its results can serve several purposes: (i) estimates may be used directly to predict the behavior of species, populations or ecosystems, (ii) quantitative models can use PhyloPars predictions to supplement existing knowledge, either by directly incorporating a subset of its estimates or by using the predicted marginal likelihood as prior distribution in Bayesian context, (iii) estimates can aid...
experimental design by providing a prior indication of the feasible range for metabolic parameters. We feel that such functionality may benefit a wide range of fields, including ecology and systems biology.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We thank Jorg Brunner and two anonymous referees for their comments.

FUNDING
Computational Life Sciences program of the Netherlands Organisation for Scientific Research (NWO) (grant number 635.100.009 to J.B.); ENFIN, a Network of Excellence funded by the European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health’ (contract number 635.100.009 to J.B.).

REFERENCES
1. Felsenstein, J. (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. Am. J. Human Genet., 25, 471–492.
2. Lynch, M. (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution, 45, 1065–1080.
3. Felsenstein, J. (1985) Phylogenies and the comparative method. Am. Nat., 125, 1–15.
4. Cunningham, C.W., Omland, K.E. and Oakley, T.H. (1998) Reconstructing ancestral character states: A critical reappraisal. Trends Ecol. Evol., 13, 361–366.
5. Schluter, D., Price, T., Mooers, A.O. and Ludwig, D. (1997) Likelihood of ancestor states in adaptive radiation. Evolution, 51, 1699–1711.
6. Garland, T., Midford, P.E. and Ives, A.R. (1999) An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. Am. Zool., 39, 374–388.
7. Altschul, S.F., Grisham, A.J. and Lipman, D.J. (1989) Weights for data related by a tree. J. Mol. Biol., 207, 647–653.
8. Vingron, M. and Sibbald, P.R. (1993) Weighting in sequence space – a comparison of methods in terms of generalized sequences. Proc. Natl Acad. Sci. USA, 90, 8777–8781.
9. Stone, E.A. and Sidow, A. (2007) Weighting in sequence space – a comparison of methods in terms of generalized sequences. Proc. Natl Acad. Sci. USA, 90, 8777–8781.
10. Stone, E.A. and Sidow, A. (2007) Constructing a meaningful evolutionary average at the phylogenetic center of mass. BMC Bioinformatics, 8.
11. West, G.B., Brown, J.H. and Enquist, B.J. (1997) A general model for the origin of allometric scaling laws in biology. Science, 276, 122–126.
12. Felsenstein, J. (2008) Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. Am. Nat., 171, 713–725.
13. Ives, A.R., Midford, P.E. and Garland, T. (2007) Within-species variation and measurement error in phylogenetic comparative methods. Syst. Biol., 56, 252–270.
14. Revell, L.J. and Harmon, L.J. (2008) Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. Evol. Ecol. Res., 10, 311–331.
15. Garland, T. and Ives, A.R. (2000) Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat., 155, 346–364.
16. Lavin, S.R., Karasov, W.H., Ives, A.R., Middleton, K.M. and Garland, T. (2008) Morphometrics of the avian small intestine compared with that of nonflying mammals: a phylogenetic approach. Physiol. Biochem. Zool., 81, 526–550.
17. Blomberg, S.P., Garland, T. and Ives, A.R. (2003) Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution, 57, 717–745.
18. Housworth, E.A., Martins, E.P. and Lynch, M. (2004) The phylogenetic mixed model. Am. Nat., 163, 84–96.
19. Pinheiro, J.C. and Bates, D.M. (1996) Unconstrained parametrizations for variance-covariance matrices. Stat. Comput., 6, 289–296.
20. Nocedal, J. and Wright, S.J. (2006) Numerical Optimization. 2nd edn, Springer, New York.
21. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. et al. (1999) LAPACK Users’ Guide. Third edn, Society for Industrial and Applied Mathematics, Philadelphia, PA.
22. Martins, E.P. and Garland, T. (1991) Phylogenetic analyses of the correlated evolution of continuous characters – a simulation study. Evolution, 45, 534–557.
23. Westoby, M., Leishman, M.R. and Lord, J.M. (1995) On misinterpreting the phylogenetic correction. J. Ecol., 83, 531–534.
24. Desviestes, Y., Legendre, P., Aouazi, L. and Morand, S. (2003) Quantifying phylogenetically structured environmental variation. Evolution, 57, 2647–2652.
25. Cubo, J., Ponton, F., Laurin, M., de Margerie, E. and Castanet, J. (2009) Phylogenetic signal in bone microstructure of sauropsids. Syst. Biol., 58, 562–574.
26. Ruan, T., Li, H., Chen, Z.Z., Coghlan, A., Coin, L.J.M., Guo, Y., Heriche, J.K., Hu, Y.F., Kristiansen, K., Li, R.Q. et al. (2008) Treefam: 2008 update. Nucleic Acids Res., 36, D735–D740.
27. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. et al. (2008) The Pfam protein families database. Nucleic Acids Res., 36, D281–D288.
28. Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 37, D5–D15.
29. Mooers, A.O., Vamosi, S.M. and Schluter, D. (1999) Using phylogenies to test macroevolutionary hypotheses of trait evolution in cranes (Gruinae). Am. Nat., 154, 249–259.
30. Grafen, A. (1989) The phylogenetic regression. Philos. Trans. Roy. Soc. Lond. Ser. B–Biol. Sci., 326, 119–157.
31. Martins, E.P., Dnim, J.A.F. and Housworth, E.A. (2002) Adaptive constraints and the phylogenetic comparative method: A computer simulation test. Evolution, 56, 1–13.