ECOLOGICAL SOCIETY OF AMERICA

Ecology/Ecological Monographs/Ecological Applications

PREPRINT

The doi for this manuscript and the correct format for citing the paper are given at the top of the online (html) abstract.

Once the final published version of this paper is posted online, it will replace the preliminary version at the specified doi.

Submitted to *Ecological Applications*

Using phylogenetic information to predict species tolerances to toxic chemicals

Guillaume Guénard [a,d,*], *Peter Carsten von der Ohe* [b], *Dick de Zwart* [c], *Pierre Legendre* [d], *and Sovan Lek* [a]

[a] *Laboratoire Évolution et diversité biologique* (EDB) UMR 5174 *CNRS / Université Paul-Sabatier, 118 rte. de Narbonne 4R3-b1, 31062 Toulouse Cedex 9, France*

[b] *UFZ, Department of Effect-Directed Analysis, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, 04318 Leipzig, Germany*

[c] *Laboratory for Ecological Risk Assessment (LER), National Institute of Public Health and the Environment (RIVM), PO Box 1 NL-3720 BA Bilthoven, The Netherlands*

[d] *Département des sciences biologiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, QC, Canada H3C-3J7*

[*] *Corresponding author at: Département des sciences biologiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, QC, Canada H3C-3J7*

E-mail addresses: guillaume.guenard@gmail.com (G. Guénard), Dick.de.Zwart@rivm.nl (D. de Zwart), peter.vonderohe@ufz.de (P. C. von der Ohe), pierre.legendre@umontreal.ca (P. Legendre), and lek@cict.fr (S. Lek).

Abstract

Tolerance to toxic substances is a characteristic of an organism that determines whether it is able to withstand the concentrations occurring in its environment. The measurement of tolerance is therefore of fundamental importance when assessing the impact of anthropogenic chemicals on ecosystems and ecological communities. Although an appreciable amount of information on species tolerance to chemicals has been collected through the last 50 years, substantial gaps remain in our knowledge of tolerance relative to the diversity of organisms inhabiting aquatic ecosystems and the great and increasing number of chemicals released in these ecosystems. Within that context, methods allowing one to reliably and accurately estimate a species' tolerance using other known characteristics would be valuable. In the present study, we introduce an approach that uses phylogeny to estimate the tolerance of a species using that of a set of other species related to the focus species at different phylogenetic scales. We estimated phylogenies from molecular data (DNA sequences) or inferred them from taxonomy. Up to 83% of the among-species variation in tolerance (log-transformed median Lethal Concentration over 96 hours; LC_{50}) was found to be phylogenetically structured, and was therefore usable for making predictions. The ability of phylogenetic models to produce accurate estimates of species tolerances is seemingly related with the availability of information within species groups and the variation in pesticide tolerance within these groups. Toxicity models integrating phylogeny therefore appear suitable to assist in risk assessment.

Keywords: tolerance, phylogeny, molecular characters, phylogenetic eigenfunctions, phylogenetic model

Introduction

Tolerance to toxic substances is a trait that determines the ability of organisms to withstand the level of pollutants occurring in their environment and is thus central to assessing the effects of toxicity on biodiversity (e.g., the calculation of species sensitivity distributions; von der Ohe and Liess 2004, Postuma et al. 2002). Tolerance is commonly approximated using bioassays, which are controlled experiments where individuals are exposed, for a given amount of time, to different concentrations of a substance or a mixture of substances and an effect is observed on a portion of the population (e.g., death of 50% of the population over 48 hours of exposure, or inhibition of reproduction for 90% of the population after 96 hours of exposure). While being a useful trait for ecotoxicologists, estimating tolerance is costly (several thousands dollars needed per estimate), logistically challenging (lots of laboratory space and personnel must be be mobilized), and sometimes impossible for all important species since specimens need to be raised in captivity or collected alive in nature. There is a large number of substances known to be hazardous to organisms in the environment. The challenge faced by ecotoxicologists is to provide reliable estimates of tolerance for as many species-substance combinations as possible. This task is extremely difficult given the ever increasing number of potentially hazardous compounds that are introduced each year and the often broad variety of organisms inhabiting the ecosystems affected by anthropogenic releases. It is therefore of interest to find alternatives to the exhaustive testing of species-substance combinations. Methods allowing the estimation of tolerance using other features of organisms –for instance, trait values (e.g., morphological, physiological, biochemical and/or ecological traits) and/or their phylogeny with respect to other species having known tolerance– may represent such alternatives. Here we consider methods for predicting tolerance using a statistical modeling approach based on phylogeny.

Tolerance is the result of multiple subordinate traits related to the uptake of pollutants by the organisms, their metabolism (e.g., transport, accumulation, sequestration, activation / inactivation), and excretion. Dependence on a wide array of such subordinate traits may generate character correlation and (positive or negative) phylogenetic autocorrelation. Modeling approaches can take advantage of these correlations to estimate tolerance, lessening the complexity associated with the numerous toxic substances and species co-occurring in the environment. Character correlation occurs when the phenotype value of a given trait is correlated with that of another trait as a consequence of, for instance, their common reliance on similar subordinate traits influenced by genetic (e.g., pleiotropy, linkage disequilibrium) or environmental processes (e.g., correlational selection; Lande and Arnold 1983). The presence of character correlation implies that the value of a trait that is hard to measure can be, to some extent, estimated from that of a trait that is more easily obtained. That approach was used by Baird and Van den Brink (2007) to estimate tolerance (the median lethal concentration: the concentration that kills 50% of individuals of a population over a specified amount of time – LC_{50}) using species' traits related to morphology, life history, physiology and feeding ecology. The second trait property, phylogenetic autocorrelation, implies that trait values show dependence with respect to species' positions in a phylogeny and may occur over multiple scales. Positive phylogenetic autocorrelation implies that closely related species share more similar trait values in comparison to more distant ones, as a consequence of evolution proceeding slowly by means of a series of small steps, over a long time period (Blomberg et al. 2003, Bulchwalter et al. 2008, Diniz-Filho et al. 1998). Positive autocorrelation shows-up as large-scale structures in phylogenetic trait signals. These large-scale structures are characterized by large differences between species pairs from different high-order taxonomic groups and small differences between species pairs from the same high-order taxonomic groups. However, closely related species can vary markedly in individual traits as a result of differentiation among parent

species (e.g., inter-specific competition; Svanbäck and Bolnick 2007). By contrast, negative phylogenetic autocorrelation implies that closely related species have more dissimilar trait values than more distant species. Negative autocorrelation appears as small-scale structures in phylogenetic trait signals. These small-scale structures are characterized by large differences occurring between closely related species pairs (e.g., from the same low-order taxonomic groups) and small differences occurring between loosely related species pairs. As for character correlation, the phylogenetic autocorrelation of a trait such as tolerance may be attributed to subordinate traits, thereby reducing or enhancing its rate of change depending on the level of non-additivity of the effects of those subordinate traits on higher-order traits. Hence, the effect of a change in a given subordinate trait may be dampened by that of other, more conserved, subordinate traits (leading to small differences among closely related species) whereas a change of a similar magnitude, but on a different subordinate trait, may have exacerbating effects (leading to substantial differences among closely related species). Phylogenetic autocorrelation was found to reliably describe the extinction threat to amphibians (Corey and Waite 2008) and the bioaccumulation of cadmium in insects (Buchwalter et al. 2008) and of trace elements in fish (Jeffree et al. 2010). However, and in spite of their anticipated relevance, predictive modeling approaches based on phylogenetic autocorrelation remain sparse.

The purpose of the present study is to develop a statistical modeling approach for making predictions of species' tolerances to toxic substances based on information available from other species and their common phylogeny, which can be obtained using different methods. We achieved this by providing assessments of (1) the fraction of variation in the tolerance of a set of species to toxic substances that can be modeled by phylogeny and of (2) the predictive power of tolerance models based on phylogeny. Considering the wide range of information and techniques now available to reconstruct the evolutionary relatedness of species, phylogenetic modeling of species tolerance may

represent a critical step towards the improvement of toxicity assessment. The same approach could also be used to compute predictive models for any other species traits that exhibit phylogenetic autocorrelation.

Methods

Data sources and selection

We used a database of concentrations associated with different toxicological endpoints and effects for various substances, aquatic species, and exposure times (de Zwart 2002). That database has been compiled from three sources: 1) AQUIRE (USEPA 1984) from U.S. Environmental Protection Agency – Mid-Continent Ecology Division, 2) a compilation of pesticide toxicity made by the Centre for Substances and Risk Assessment (Netherlands National Institute of Public Health and the Environment; Crommentuijn et al. 1997, Tomlin 1997), and 3) another compilation of pesticide toxicity offered by the U.S. Environmental Protection Agency – Office of Pesticides Programs, Ecological Effects Branch. From that database we selected data of lethal concentration (LC_{50}) after 96 hours while excluding all entries with inequality indications (i.e., greater than or smaller than). We selected that particular endpoint-effect combination in order to obtain the greatest number of substance-species combinations (7 170 combinations over 8 848 entries, with 1 731 substances and 759 species involved). When multiple test values were found for one substance, quality checks such as water solubility were employed to eliminate odd data entries (e.g., unit transformation errors). If values differed by more than a factor of 30 from the closest one in a group of at least two other references, we discarded the aberrant value in order to remove outliers from the data set. Of all the remaining values for a given substance, we took the geometric mean as the valid experimental value. The remaining selection

procedure aimed at obtaining the largest set of species whose effect concentrations were available for as many substances as possible with no missing information. To obtain that species by compound table, we first classified species and chemicals by decreasing order of number of effect concentrations available and investigated the topmost elements of the resulting lists.

Obtaining phylogenies

Phylogenies can either be estimated using suitable characters, or obtained from the literature. A wide variety of phylogenetic inference methods now exists (e.g., maximum-parsimony, distance-based, maximum-likelihood, spectral, or Bayesian methods) whereas abundant, and rapidly increasing, information about molecular (DNA) characters is being made available on the Internet through organizations such as the U.S. National Center for Biotechnology Information (NCBI; URL: http://www.ncbi.nlm.nih.gov/). Phylogenies can also be found within the molecular taxonomy literature or from the Tree of Life project (ToL; Maddison et al. 2007). Our methodology can be used with any of these sources of phylogenetic information as long as they are considered reliable and accurate.

We used two different approaches to obtain phylogenies in the present study. The first involved the estimation of a tree from DNA sequences using a maximum-likelihood approach (Felsenstein 1981, Felsenstein and Churchill 1996; analysis performed using the software EMBOSS version 6.1.0-5, Rice et al. 2000). To do so, we obtained DNA sequences from NCBI's Nucleotide database which consisted, whenever available, of the entire mitochondrial genome as well as nuclear DNA sequences for 28S, 18S, and 5.8S ribosomal RNA transcripts and their internal transcribed spacers (ITS 1 and ITS 2). Then, we performed multiple sequence alignment on each gene separately using the computer program MUSCLE (v3.7; Edgar 2004). Finally, we concatenated these aligned sequences into a super alignment

of genes before estimating the tree. The resulting tree was used to assess the ability of phylogenetic autocorrelation at describing the tolerance of a set of species to multiple pesticides.

The second approach involved constructing a tree from information on taxonomic classification. For that purpose, we gathered information on a maximum of 19 taxonomic ranks from the ToL project for each species. Species with no available information for a given rank were assigned a generic taxon for that rank. We constructed the tree topology implied by the hierarchical structure of taxonomy and placed all taxa of a given rank at the same distance from the root.

Although the construction of a species tree from taxonomy may be the only solution available in the absence of suitable molecular information, readers must be warned that there are many situations in which these trees may not accurately represent the phylogeny. For instance, trees constructed from the taxonomy of species covering a wide range of high-order taxa may be congruent with molecular phylogenetics trees in term of their tree topology while their adequacy in representing branch lengths may remain questionable. The quality of a tree constructed from the taxonomy of species covering a narrower range of low-order taxa would be questionable both in terms of topology and branch lengths. As it is the case for modeling methods in general, the modeling approach described herein assumes that the explanatory factor that is provided (i.e. the phylogeny), and on which it depends, is correct. In most practical situations, trees estimated from molecular phylogenetic methods should therefore be preferred over trees constructed using taxonomic classification.

Constructing a phylogenetically-explicit model

We represented the structures of phylogenetic signals using eigenfunctions derived from a phylogenetic tree, a method also known as phylogenetic eigenvectors regression (PVR; Desdevises et al. 2003,

Diniz-Filho et al. 1998; Diniz-Filho et al. (1998) used only the first few eigenfunctions obtained by PCoA to represent the phylogeny, whereas Desdevises et al. (2003) used all eigenfunctions, as in the method described in the present paper). These eigenfunctions were computed from the phylogenetic covariance matrix **W** whose elements $w_{i,j}$ correspond to the length of path leading from the root of the tree to the first common ancestor of species *i* and *j*. The eigenvalues and eigenvectors associated with **W** after double centering were obtained by solving the equation:

(1) $$\boldsymbol{\Omega} = \mathbf{Q}\mathbf{W}\mathbf{Q} = \mathbf{U}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{U}^{\mathrm{T}} \quad , \quad \mathbf{Q} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} \quad ,$$

where matrix **U** has eigenvectors $\mathbf{u}_i$ as column vectors, diagonal matrix $\mathbf{D}_{\lambda}$ is a diagonal matrix of eigenvalues, **Q** is a centering matrix calculated from an $n \times n$ identity matrix $\mathbf{I}_n$ and a vector of *n* ones $\mathbf{1}_n$; *n* is the number of species and superscript t denotes matrix transposition. As consequences of the symmetry of **W** and its centering prior to eigenvalue decomposition, $n - 1$ non-zero and mutually orthogonal unit vectors are obtained, therefore defining an orthonormal basis against which trait variance can be decomposed with respect to phylogeny in a multiple scale fashion (Figure 1). That approach is similar to a principal coordinate analysis based on a similarity matrix (Gower 1966).

In the models developed in the present study, the response variable **y** is a vector whose elements are experimental LC_{50} values for a given species and compound. This variable was regressed against a design matrix **X** involving two factors and their interaction. The first factor describes the fraction of LC_{50} variability which is associated strictly with mean toxicity of each compound for all the species involved in the model and was represented using Helmert orthogonal contrast variables. Each instance

of a given compound was represented in the design matrix **X** by its scores on the contrast variables. The number of such contrast variables in the design matrix was $m - 1$, where m is the number of compounds considered. The second factor describes the LC_{50} variability which is associated strictly with the mean susceptibility of species for all the compounds involved in the model. It was represented in the design matrix using the species scores on each of the $n - 1$ eigenvectors obtained from Equation 1; the scores of any given species being repeated for each compound. The interaction term between the two factors describes the LC_{50} variability which is not accounted simply by adding the mean toxicity of compounds with the mean susceptibility of species, thereby allowing the model to represent cases where different compounds affect the species within the phylogeny in different ways. That interaction term was represented in the design matrix by set of all possible $(m - 1) * (n - 1)$ element-wise multiplication of any Helmert contrast variable describing mean compound toxicity with any variable describing species susceptibility at a particular phylogenetic scale from their position in the phylogeny. The design matrix included a column of ones to allow the estimation of the intercept of the model.

In order to avoid over-fitting, a column subset of the design matrix was selected when constructing the models. We obtained that subset by first including the factor representing the compounds (i.e., the intercept and Helmert contrasts) and then performing a forward-stepwise selection, using F-tests, of the variables representing phylogeny and the interactions between compounds and phylogeny. Family-wise (corrected) *p*-values of the inference tests performed for the stepwise addition of variables were obtained using the sequential Bonferroni procedure (Holm 1979). Finally, proportions of variation associated with the compounds, the phylogeny, and the compound-phylogeny interactions were estimated as their respective adjusted coefficients of determination. That approach is meant to provide a column subset $\mathbf{X_S}$ of the design matrix **X** that best fitted the response

variable while avoiding over-fitting. It does not, however, allow one to make predictions of LC_{50} values for additional species.

Making predictions

The approach to make predictions for additional species involves four steps. Firstly, the positions of the new species in the phylogenetic tree have to be taken from a previous analysis or estimated. In the case that the position has to be estimated, the new species must be added to the established phylogenetic tree (i.e. the one used to calculate **W**) without modifying the topology and branch lengths of the subset tree for the original species. Warning should be made here that redoing/repeating phylogenetic analysis with one or more additional species often results in the alteration of the original subset tree. Under these circumstances, the orthonormal basis must be recalculated and any model based on it rebuilt. We avoided this issue by including in the phylogenetic analysis, from the beginning, the species for which predictions were to be made; it was then possible to select the subset tree of the n species with known response variable to estimate the phylogenetic model, and then use the positions of the remaining q species to make predictions. Secondly, a $q \times n$ matrix $\mathbf{W}_{n+k}$ whose elements $w_{n+k,j}$ are the lengths of the paths leading from the root of the tree to the first common ancestor of a new species k and a species j within the model, is calculated. Thirdly, the projection scores $\mathbf{S}_{n+k}$ of the new species on the $n - 1$ eigenfunctions underlying the eigenvectors in **U** are obtained following Gower's approach for adding new observations in an existing principal coordinate analysis, by rearranging Equation 1 and performing a partial substitution of matrix **W** by $\mathbf{W}_{n+k}$ (Gower 1969; Figure 1: black markers):

(2) $$\mathbf{S}_{n+k} = \left\{ \mathbf{W}_{n+k} - \frac{1}{n}\left(\mathbf{1}_q \mathbf{1}_n^{\mathrm{T}} \mathbf{W} + \mathbf{W}_{n+k} \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}\right) + \frac{1}{n^2} \mathbf{1}_q \mathbf{1}_n^{\mathrm{T}} \mathbf{W} \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}} \right\} \mathbf{U} \boldsymbol{\Lambda}^{-1} \quad .$$

232

233 Finally, the last step involves using the scores of the new species as explanatory variables to calculate
234 predictions. Note that using scores obtained from species found to be outside the originally established
235 phylogeny (such as in see Figure 1: *species O*) to make predictions involves extrapolation beyond the
236 known range of phylogenetic variation of traits and should thus be avoided. Besides those involving
237 phylogenetic eigenfunctions, other approaches (based, for instance, on generalized least-squares
238 regression or autoregression) have been proposed to test for phylogenetic signals (e.g. Blomberg et al.
239 2003, Zheng et al. 2009) and to estimate trait values (e.g. Martins and Hansen 1997, Garland and Ives
240 2000, Rohlf 2001, Bokma 2008).

241 Constructing phylogenetic models through cross-validation

242 The above-described framework provides the possibility of using cross-validation as an alternative to
243 forward-stepwise multiple regression to obtain a phylogenetically-explicit predictive model. Cross-
244 validation allows a straightforward assessment of the ability of the approach to make predictions for
245 new species while avoiding the issue of over-fitting the model. Such an approach involves 1) removing
246 one species from an original dataset at a time, 2) calculating linear model coefficients (**b**) using the
247 remaining species, 3) predicting the value of the response from the removed species, and 4) reiterating
248 the first three steps for every species. In that case, linear coefficients (**b**) and standardized linear
249 coefficients (**β**) of the relationship between the response variable **y** (LC_{50} in the present study) and the
250 eigenvectors describing phylogeny (**U**) are calculated as:

251

(3) $$\mathbf{b} = \mathbf{U}^{\mathbf{T}}[y-\bar{y}] \quad , \quad \boldsymbol{\beta} = \frac{1}{\sqrt{[y-\bar{y}]^{\mathbf{T}}[y-\bar{y}]}}\mathbf{U}^{\mathbf{T}}[y-\bar{y}]$$

where $\bar{y}$ is the mean of the response variables, and the predicted values of the response variable ($\mathbf{y}_{predicted}$) are obtained from:

(4) $$\mathbf{y}_{predicted} = \bar{y}+\mathbf{S}_{n+i}\mathbf{b} \quad .$$

Since the observed values of the response variable are not involved in the calculation of their respective predictions, that approach has the advantage of conserving the independence of the observed and predicted values under the null hypothesis that the response is unrelated to phylogeny. Although that approach allows the use of every single eigenfunction in models, it does not, however, guarantee that all of them are relevant for making predictions. A simple method to obtain more generalizable models is to truncate the vector of linear coefficients **b** by assigning 0 to its elements that are associated with square standardized linear coefficients ($\boldsymbol{\beta}^2$) that are below a threshold chosen to minimize the mean squared error of the model (the mean of the squared differences between predicted and observed values), thereby filtering out irrelevant eigenfunctions. The cross-validation procedure was illustrated by selecting LC_{50} values (96 hours) for pesticide Carbaryl on all available species in the database and constructing a tree representing their phylogeny from information on their taxonomy.

Comparing observed with predicted tolerance

The comparison of observed and predicted tolerance values was performed at two levels. Firstly, a

global comparison of these values was made through the examination of the confidence intervals of the slope and intercept of a linear regression line with observed values on the ordinates and predicted values on the abscissa using $\log_{10}$-transformed LC_{50} values on a molecular basis. Secondly, a comparison was performed at the observation level by calculating the deviation factor *d* of a species *i* as:

(5) $$d_i = \begin{cases} 10^{(y_{pred\,i} - y_{obs\,i})} - 1 & \text{if} \quad y_{pred\,i} \geq y_{obs\,i} \\ 1 - 10^{(y_{obs\,i} - y_{pred\,i})} & \text{if} \quad y_{pred\,i} < y_{obs\,i} \end{cases}$$

where y_{obs} are the observed values and y_{pred} are those predicted by the model, both on a $\log_{10}$ scale. The deviation factor is the number of times tolerance is overestimated (positive values) or underestimated (negative values) by the model. For example, a value close to 0 means that the tolerance observed for a species is in close agreement with that predicted by the phylogenetic model. Similarly, a value of +10 means that the tolerance observed for a species is ten times lower than that predicted by the model while a value of –2 means that the tolerance observed for a species is twice as high as that predicted by the model.

All calculations and statistical analyses were performed using the R language and environment (version 2.10.1; R Development Core Team 2010). Database queries were done using package RMySQL (version 0.7-4; James and DebRoy 2009) and phylogenetic analyses with package ape (version 2.4-1; Paradis et al. 2004).

Results

Data mining

The best dataset that we found involved pesticides Malathion (CAS: 121-75-5), DDT (CAS: 50-29-3), Lindane (CAS: 58-89-9) and Carbaryl (CAS: 63-25-2), and 25 aquatic animal species, including 20 bony fish, 3 crustacean, and 2 insect species (see *Appendix A*, Table A1 for details). We built a first model (hereafter referred as model #1) using these data. In order to study the response of models to a varying number of substances with respect to species, we assembled three additional datasets by adding chemical substances, which resulted in a subsequent reduction in the number of species. We obtained the first additional dataset by adding Parathion (CAS: 56-38-2; 18 species: 13 fishes, 3 crustaceans, and 2 insects), the second by further adding Dieldrin (CAS: 60-57-1; 14 species: 10 fishes, 3 crustaceans, and 1 insect), and the third by further adding rotenone (CAS: 83-79-4) and Toxaphene (CAS: 8001-35-2; 11 species: 9 fishes, 1 crustacean, and 1 insect). The three additional models built from these datasets are hereafter referred as model #2, model #3, and model #4, respectively.

DNA sequences were available for 23 species out of 25 and the most widespread were those for the Cytochrome oxydase subunit 1 (COX1; 19 species) and the mitochondrial large (21 species) and small (18 species) ribosomal RNA subunits (see *Appendix A*, Table A2 for details). On average, 20.56 of the 44 sequences were available at a specific level (range: 5–44). We completed that set of sequence by borrowing sequences from other species within the same genus (2.08 sequences on average) or family (3.16 sequences on average), while an average of 18.2 sequences remained missing. The resulting super-alignment included from 3 246 to 24 433 base pairs (median: 16 503 base pairs).

Phylogenetic analysis

The tree obtained from DNA sequences placed most species within their known taxonomic group (Figure 2). The tree was rooted at the separation between arthropods and (bony) fish. The first separation on the arthropod subtree occurred between crustaceans and insects, and the second for crustaceans at the sub-ordinal level between eucarids (the speckled shrimp, *Metapenaeus monoceros*) and peracarids (represented by orders isospoda: *Asellus brevicaudus* and a amphipoda: *Gammarus lacustris*). On the fish subtree, the first separation occurred between ostariophysians and the remaining two teleost suborders, i.e., protacanthopterygii and acanthopterygii. Within the ostariophysians the separation first occurred at the ordinal level between cypriniforms and siluriforms, each represented by a single family (cyprinidae and ictaluridae, respectively). The second separation on the fish subtree occurred between protacanthopterygians, which is represented by family Salmonidae (ord. Salmoniformes), and acanthopterygians. On the salmonids subtree, the first separation occurred between genus *Oncorhynchus* (rainbow trout and coho salmon) and genera *Salmo* (brown trout) and *Salvelinus* (brook trout and lake trout), with the second separation occurring between the latter genera. Discrepancies of the constructed phylogeny with respect to taxonomy occurred on the acanthopterygians subtree. First, the stripped bass (*Morone saxatilis*, fam. moronidae) and spotted snakehead (*Channa punctatus*, fam. channidae) separated from other species of order perciformes rather than the species from order cyprinodontiformes (both fam. poeciliidae: the mosquitofish, *Gambusia affinis* and the guppy, *Poecilia reticulata*), as expected by taxonomy. Cyprinodontiforms species remained clustered within perciforms up to the sub-ordinal level where they separate from the Mozambique tilapia (*Oreochromis mossambicus*, fam. cichlidae). Following taxonomy, the stripped bass and spotted snakehead were expected separate from other perciforms at the sub-ordinal level.

These apparent discrepancies may outline the limit of the current DNA dataset for reconstructing the phylogeny of these species; we explored their possible impact on the modeling approach herein described by recalculating model #1 using a tree obtained from taxonomy (hereafter referred as model 1T).

Phylogenetic models of tolerance

The models describing LC_{50} variability among pesticides and as a function of species' phylogenetic structure explained from 61% (model #3) to 85% (model #1, Figure 2) of the observed variation in tolerance to pesticides (Table 1). By comparison, a model using the mean LC_{50} of all organisms for each of the pesticides (i.e., factor pesticide) only explained from 45% (model #2) to 49% (model #1) of that variability. The addition of phylogenetic information thus represents improvements ranging from 12% (model #3) to 26% (model #1) of the total LC_{50} variability, with phylogeny explaining from 24% (model #3) to 63% (model #1) of LC_{50} variability within pesticides. Model 1T slightly differed from model #1, but led to similar conclusions.

We found 67 species whose LC_{50} values (96 hours) for pesticide Carbaryl were available to illustrate the cross-validation procedure. These species included 35 fish, 18 crustaceans, eight insects, four mollusks, one amphibian, and one annelid. We estimated the β^2 threshold for the truncation of the vector of linear coefficients graphically as 0.000 52 from a plot of cross-validated mean standard error obtained by repeating the calculations for thresholds ranging from 0 (all eigenfunctions retained) to 0.015 (the expected β^2 if all 67 eigenfunctions were equally relevant) in steps of 0.000 01. The resulting cross-validated models explain 83% of the observed variation of the $\log_{10} LC_{50}$ for Carbaryl among these species (adjusted R^2; Figure 3). The regression slope (1.06; 95% confidence limits: 0.94 and 1.18)

and intercept (-0.01; 95% confidence limits: -0.15, 0.13) of the relationship between predicted and observed value was consistent with those of a 1:1 relationship and are not suggestive of a substantial prediction bias by the approach. Its ability to predict LC_{50} accurately within taxonomic group differed among high-order taxonomic groups, with the model representing from only 13% ($p > 0.05$) of $\log_{10}$ LC_{50} variability among mollusk species up to 81% ($p = 0.001$) of that among insects species (fish: 40% – $p < 0.000\ 1$, crustaceans: 68% – $p < 0.000\ 1$). The median deviation factor was 0.09 (range: -46 – 10) and ranged from -1.84 (mollusks) to 0.57 (insects; crustaceans: 0.11, fish: 0.05) among the four groups with more than one representative species (Figure 4). Overall, predictions for 64 species out of 67 (95.5%) had a deviation factor between -10 and +10 whereas a deviation factor between -1 and +1 was obtained for 41 (61.2%) species. The fish was the group whose tolerance was the most accurately represented by the models (median absolute deviation factor: 0.70), followed by crustaceans (0.94) insects (1.10), and mollusks (2.37).

Discussion

The approach herein described exemplifies how phylogeny could be used to predict tolerance to pesticides and other chemical substances. In spite of the relatively modest number of representative species available, the results of the present study suggest that the phylogenetic structuring of tolerance, quantified in terms of LC_{50}, accounted for almost one fourth to almost two thirds of the residual variation within sets of four to eight pesticides. When cross-validated against a single pesticide, Carbaryl, the phylogenetic prediction approach provided good estimates of observed LC_{50} values taken from published laboratory studies, using a reasonable amount of empirical information. Given the ever increasing availability of molecular information, more particularly in the form of DNA sequences, these results highlight an opportunity to stretch our current usage of the existing tolerance data through

phylogenetic-based estimation for species of unknown tolerances. The phylogenetic modeling framework developed in the present study seems, at least under certain circumstances, robust to discrepancies in its prediction basis (i.e., the phylogenetic tree), as illustrated by similarity of the results obtained by model #1 and model 1T, which was based on taxonomy. The robustness of phylogenetic models towards misspecified phylogenies has also been recently demonstrated for the Phylogenetic Generalized Least Squares regression, another method to construct phylogenetic models (Stone 2011).

The approach used was, in part, borrowed from that of the phylogenetic comparative method, whose purpose is to study the relationships between traits by mean of comparisons across species, while correcting for their respective phylogenetic autocorrelation. In the present study the fraction of trait variation which is organized with respect to phylogeny is exploited for making predictions. We ought to mention here that autocorrelation implies the violation of the assumption of independence of observations and may thus affect the outcome of statistical tests. It has been recognized that phylogenetic autocorrelation may render invalid the statistical tests of correlation between species traits (Feldsenstein 1985). This represents a serious shortcoming that sometimes fails to be addressed when using character correlation for predicting tolerance from other species' traits (e.g., Baird and Van den Brink 2007). As a potential solution, a model may use a phylogeny in conjunction with auxiliary traits related with tolerance to pesticides, possibly enhancing the capacity of the former. When constructing models involving auxiliary traits, one has, however, to keep in mind that any trait used as an explanatory variable may itself be phylogenetically autocorrelated. For example, body size has been shown to be related with a wide range of physiological and ecological attributes (Peters 1983) and may affect tolerance as well. However, the magnitude of body size is heavily structured by phylogeny at large scale and body size may also vary markedly, but within similar orders of magnitude, at smaller phylogenetic scales (e.g., within a family). Since both the tolerance and body size may be driven by the

same phylogenetic structures, the parameters (intercept and slope) of a regression involving these traits are likely to be biased and not representative of the general relationship between them. A general solution when integrating auxiliary traits in models is to use phylogenetic eigenfunctions, which correspond to the eigenfunctions selected for the phylogenetic model, as co-variables when estimating the relationship between the response (i.e., tolerance) and the auxiliary trait (e.g., body mass). Using phylogeny in that manner allows one to partial-out the phylogenetic components of the variation of these species traits before using them as explanatory variables. For instance, body size sometimes varies greatly during the ontogeny of organisms such as aquatic animals, and therefore is irrespective of their phylogeny. Hence, if tolerance to a given pollutant is related with body size, and one builds a model using many individuals of different sizes to represent each species, an important portion of the variation observed for tolerance cannot be represented using a phylogeny, but will be suitably accounted for by body size. In such a situation, a model involving body size as an auxiliary trait would explain a greater portion of the variation in tolerance than one involving phylogeny alone.

The ability of a phylogenetic model to make reliable predictions for a given taxonomic group may not only depend on the number of representative species involved in its construction, but also on the structure of the trait variation along the tree used to represent the phylogeny. For instance, 81% of the variability in the tolerance to Carbaryl among insect species was explained by the phylogeny. The relatively good accuracy of the model for predicting the tolerance of insect species to Carbaryl is driven by the small tolerance of the four plecopteran species (mean LC_{50} = 0.020 $\mu mol \cdot L^{-1}$) compared to that of hemipterans (mean LC_{50} = 1.75 $\mu mol \cdot L^{-1}$). Moreover, the large-scale component of the phylogenetic tolerance signal did accurately predict the tolerance of the only amphibian species (the Indian bullfrog; *Hoplobatrachus tigerinus*) to Carbaryl from that of the other vertebrate species (fish). On the other hand, the poor performance of the model at predicting the tolerance among mollusks is seemingly the

consequence of its inability to predict the greater tolerance of the Atlantic rangia (*Rangia cuneata*), the only bivalve species available, with respect to the other three gastropod species. These examples illustrate the two main requirements of the phylogenetic approach to accurately model the value of a trait such as tolerance: the accuracy of the method is dependent both on the degree of the phylogenetic autocorrelation of the trait (i.e., how much of the trait value is inherited from ancestral species) and the adequacy of the sampling (i.e., the number of members within taxonomic groups among which large difference in trait value are observed or expected). For instance, a phylogenetic model is expected to be inaccurate at evidencing very sensitive or tolerant species pertaining to a highly variable and poorly sampled genus. To this end, it is noteworthy that phylogenetic models cannot predict instances where outstandingly resistant populations arise by natural selection such, as resistance to pesticides (Ferro 1993, Nandula 2010) or cases of populations living in heavily polluted environments and showing high tolerance to local pollutants (e.g., Nacci et al. 2010). In these cases, a phylogenetic model can nevertheless be useful as a baseline to qualify organisms as resistant or sensitive when their observed tolerance are higher or lower than predicted by the model, respectively. Also noteworthy is the fact that the approach described in the present study carries the assumption, which is common among statistical modeling methods, that the set of species under study forms a representative sample of a larger group of species for which we want to estimate tolerance (i.e., the statistical population). In some groups, the tolerance of ubiquitous species occurring close to – and/or bearing economical value for – humans, may be better studied than that of rare species. Hence, a model involving a sample of exceptionally tolerant (or sensitive) species will consistently overestimate (or underestimate) the tolerance of species for whom tolerance data are not available.

Besides its direct application for predicting a single toxicological effect and endpoint, the approach described in the present study remains applicable in a multiple-effects or multiple-endpoints

model framework. Here we will suggest two approaches by which it can be achieved, although others may be applicable. The first possibility would be to use multivariate regression of a species × effect or species × endpoints response matrix instead of a single response vector as used in the present study. Such a relatively simple approach allows one to obtain several models describing the different effects and/or endpoints at once. The second, more elaborate possibility would be to combine the information on many different endpoints for a given species and calculate metrics describing their relationships to one another (e.g., the log ratio between concentration for observing effects x, y, z and LC_{50}, over the same amount of time), or with respect to a common tolerance baseline, and then applying multivariate regression to the resulting species × metrics response matrix. For both approaches, it would be possible to further the analysis of the results obtained by subjecting their resulting multivariate fitted and residual values to principal components analysis. The combination of these two methods, multivariate regression and principal component analysis, is known in community ecology as redundancy analysis (Rao 1964, Legendre and Legendre 1998).

Although the phylogenetic eigenfunctions approach considered in the present study relies on known chemicals for which toxicity was assessed empirically from bioassays, its flexible nature also allows it to be transposed to other frameworks based, for instance, on toxic modes of action (TMoA) or quantitative structure-activity relationships (QSAR; Ajmani et al. 2009, de Roode et al. 2003, Schultz et al. 2003, Russom et al. 1997, Von der Ohe et al. 2005). TMoA refers to the metabolic function which is the most adversely disturbed by a given chemical and most readily leads to the observed effect on the whole organism. Hence, different TMoA can be used as levels of a linear model factor, with individual chemicals acting through the same mode nested within its respective level (i.e., its respective mode of action). Models thus obtained could provide insight on how the sensitivity towards particular TMoA is structured into phylogeny and which groups are the most or the least susceptible, etc. QSAR models

seek to predict the biological activity of a compound from descriptors of its chemical structure. Since biological activity may vary among organisms as a consequence of their particular biochemical traits, it is conceivable that including phylogenetic eigenfunctions as a new set of parameters in QSAR models may improve their ability to predict the impact of new compounds on organisms from a given set of taxonomic groups. If such an approach proves successful, it would provide environmental protection agencies with more dependable tools to more readily screen across the growing list of emerging industrial compounds. Furthermore, organism-specific QSAR may benefit the chemical industry by providing insights on the theoretical innocuousness of compounds under development on the organisms that would specifically be exposed to it.

Acknowledgements

We are thankful to Cândida Shinn, Steven C. Walker, Michael Hellberg, and two anonymous reviewers for their helpful comments on earlier versions of the manuscript. This study was supported by the Marie Curie Research Training Network – Keybioeffects (MRTN-CT-2006-035695) and the Integrated Project MODELKEY (contract 511237-GOCE) of the 6th framework program of the European Commission. Guillaume Guénard also received support from the “Fond québécois de recherche sur la nature et la technologie (FQRNT)” and Peter von der Ohe received financial support through a “Deutsche Forschungsgesellschaft” (DFG – Bonn, Germany) fellowship (PAK 406/1).

References

Ajmani, S., Jadhav, K., and Kulkarni, S. 2009. Group-Based QSAR (G-QSAR): Mitigating Interpretation Challenges in QSAR. *QSAR Comb. Sci.* 28: 36-51

Baird, D. J. and Van den Brink, P. J. 2007. Using biological traits to predict species sensitivity to toxic substances. *Ecotox. Env. Safety* 67: 296-301

Blomberg, S., Garland, T., and Ives, A. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57: 717-745

Bokma, F. 2008. Detection of "punctuated equilibrium" by bayesian estimation of speciation and extinction rates, ancestral character states, and rates of anagenetic and cladogenetic evolution on a molecular phylogeny. *Evolution* 62: 2718-2726

Buchwalter, D. B., Cain, D. J., Martin, C. A., Xie, L., Luoma, S. N., and Garland, T. Jr. 2008. Aquatic insect ecophysiological traits reveal phylogenetically based differences in dissolved cadmium susceptibility. *Proc. Nat. Acad. Sci.* USA 105: 8321-8326

Corey, S. J. and Waite, T. A. 2008. Phylogenetic autocorrelation of extinction threat in globally imperilled amphibians. *Diversity & Distributions* 14: 614-629

Crommentuijn, T., Polder M. D., and van de Plassche, E. J. (1997). Maximum Permissible Concentrations and Negligible Concentrations for metals, taking background concentrations into account. Rep. #601501001, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands. URL: http://www.rivm.nl/bibliotheek/rapporten/601501001.pdf

de Zwart, D. 2002. Observed regularities in SSDs for aquatic species. pp. 133–152 In Posthuma, L., Suter, G.W., and Traas, T. (ed.) Species Sensitivity Distributions in Ecotoxicology. Lewis

Publishers, Boca Raton, FL, USA.

Desdevises, Y., P. Legendre, L. Azouzi and S. Morand. 2003. Quantifying phylogenetically-structured environmental variation. *Evolution* 57: 2647-2652.

Diniz-Filho, J. A. F. 2001. Phylogenetic Autocorrelation under Distinct Evolutionary Processes. *Evolution* 55: 1104-1109

Diniz-Filho, J. A. F., de Sant'Ana, C. E. R., and Bini, L. M. 1998. An eigenvector method for estimating phylogenetic inertia. *Evolution* 52: 1247-1262

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high throughput, *Nucleic Acids Res.* 32: 1792-1797

USEPA 1984. AQUIRE, Aquatic Information Retrieval Toxicity Database. Project Description, Guidelines and Procedures by R.C. Russom and A. Pilli. Duluth, MN, USA: United States Environmental Protection Agency, Environmental Research Laboratory-Duluth, Office of Research and Development. Report nr EPA 600/8-84-021.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Ecol.* 17: 368-376

Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.* 125: 1-15

Felsenstein, J. and Churchill, G. A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13: 93-104

Ferro, D. N. 1993. Potential for resistance to *Bacillus thuringiensis*: Colorado potato beetle (Coleoptera: Chrysomelidae) - a model system. *Am. Entomol.* 39: 38-44

Garland, T. Jr. 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

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate

analysis. *Biometrika* 53: 325-338

Gower, J. C. 1969. Adding a point to vector diagrams in multivariate analysis. *Biometrika* 55: 582-585

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6: 65-70

James, D. A. and DebRoy, S. 2009. RMySQL: R interface to the MySQL database (R package version 0.7-4). URL: http://CRAN.R-project.org/package=RMySQL

Lande, R. and Arnold, S. J. 1983. The measurement of selection on correlated characters. *Evolution* 37: 1210-1226

Legendre, P. and Legendre, L. 1998. Numerical Ecology – Second English edition, Elsevier Science B.V., Amsterdam, The Netherlands.

Maddison, D. R., Schulz, K. S., and Maddison, W. P. 2007. The tree of life web project. *Zootaxa* 1668: 19-40

Martin, E. P. and Hansen, T. F. 1997. Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. *Am. Nat.* 149: 646-667

Nacci, D. E.; Champlin, D., and Jayaraman, S. 2010. Adaptation of the estuarine fish *Fundulus heteroclitus* (Atlantic killifish) to polychlorinated biphenyls (PCBs). *Estuaries Coasts* 33: 853-864

Naudula, V. K. 2010. Glyphosate resistance in crops and weeds: history, development, and management. Wiley Publishing, NJ, USA.

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289-290

Peters, R. H. 1983. The ecological implications of body size. Cambridge University Press, New-York, NY, USA.

554 Posthuma, L., Suter, G. W., and Traas, T. 2002. Species Sensitivity Distributions in Ecotoxicology.
555 Lewis Publishers, Boca Raton, FL, USA.
556 R Development Core Team (2010). R: A language and environment for statistical computing. R
557 Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL
558 http://www.R-project.org
559 Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research.
560 Sankhyaá, Ser. A 26: 329-358
561 Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open
562 Software Suite. *Trends Genetics* 16: 276-277
563 Rolf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric
564 interpretations. *Evolution* 55: 2143-2160
565 de Roode, D., Hoekzema, C., de Vries-Buitenweg, S., van de Waart, B., and van der Hoeven, J. 2006.
566 QSARs in ecotoxicological risk assessment. *Regul. Toxicol. Pharm.* 45: 24-35
567 Russom, C. L., Bradbury, S. P., Broderius, S. J., Hammermeister, D. E., and Drummond, R. A. 1997.
568 Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow
569 (*pimephales promelas*). *Env. Toxicol. Chem.* 16: 948-967
570 Schultz, T. W., Cronin, M. T. D., and Netzeva, T. I. 2003. The present status of QSAR in toxicology. *J.*
571 *Mol. Struct. THEOCHEM* 622: 23-38
572 Stone, E. A. 2011. Why the phylogenetic regression appears robust to tree misspecification. *Syst. Biol.*
573 60: 245-260
574 Svanbäck, R. and Bolnick, D. I. 2007. Intraspecific competition drives increased resource use diversity
575 within a natural population. *Proc. Roy. Soc. Lond. B* 274: 839-844
576 Tomlin, C. D. S. (ed.) 1997. Pesticide Manual. Farnham, UK: British Crop Protection Council.

577 von der Ohe, P. C. and Liess, M. 2004. Relative sensitivity distribution of aquatic invertebrates to
578 organic and metal compounds. *Env. Toxicol. Chem.* 23: 150-156
579 von der Ohe, P. C., Kuhne, R., Ebert, R. U., Altenburger, R., Liess, M., and Schuurmann G. 2005.
580 Structural alerts – A new classification model to discriminate excess toxicity from narcotic
581 effect levels of organic compounds in the acute daphnid assay. *Chem. Res. Toxicol.* **18**: 536-
582 555.
583 Zheng, L., Ives, A. R., Garland, T., Larget, B. R., Yu, Y., and Cao, K. 2009. New multivariate tests for
584 phylogenetic signal and trait correlations applied to ecophysiological phenotypes of nine
585 Manglietia species. Funct. Ecol. 23: 1059-1069

Tables

Table 1. Statistical test results associated with the phylogenetic models describing among-pesticides and among-species variation of LC_{50}, and their associated coefficient of multiple determination ($\mathbf{R^2}$) and adjusted coefficients of multiple determination ($\mathbf{R^2_{adj.}}$). The additional model 1T corresponds to model 1 but was constructed on a phylogeny obtained from taxonomy rather than from molecular characters.

Model	Number of pesticides	Number of species	Factor	$F_{(df,df_{residual})}$	df	*p*-value		R^2	$R^2_{adj.}$	
						Test-wise	Family-wise		Ind. Factor	All Factors
#1	4	25	Pesticide	129.407	3	< 0.000 1		0.602	0.590	
			Phylogeny	23.815	5	< 0.000 1	< 0.000 1	0.185	0.141	0.847
			Interaction	16.558	3	< 0.000 1	< 0.000 1	0.077	0.048	
			Residual		88			0.136		
#2	5	18	Pesticide	33.489	4	< 0.000 1		0.478	0.453	
			Phylogeny	18.429	3	< 0.000 1	< 0.000 1	0.197	0.169	0.683
			Interaction	10.169	1	0.002	0.03	0.036	0.025	
			Residual		81					
#3	6	14	Pesticide	22.222	5	< 0.000 1		0.520	0.489	
			Phylogeny	12.070	1	0.000 9	0.006	0.056	0.045	0.612
			Interaction	14.470	1	0.000 3	0.003	0.068	0.056	
			Residual		76			0.356		
#4	8	11	Pesticide	26.296	7	< 0.000 1		0.562	0.524	0.734
			Phylogeny	18.460	1	< 0.000 1	0.002	0.056	0.045	
			Interaction	16.314	3	< 0.000 1	< 0.000 1	0.150	0.119	

Model	Number of	Number	Factor	$F_{(df,df_{residual})}$	df	*p*-value		R^2	$R^2_{adj.}$	
			Residual		76			0.232		
			Pesticide	102.061	3	< 0.000 1		0.602	0.590	
			Phylogeny	24.485	3	< 0.000 1	< 0.000 1	0.144	0.117	0.805
1T	4	25	Interaction	13.013	3	< 0.000 1	< 0.000 1	0.077	0.048	
			Residual		90			0.177		

592

Figures captions

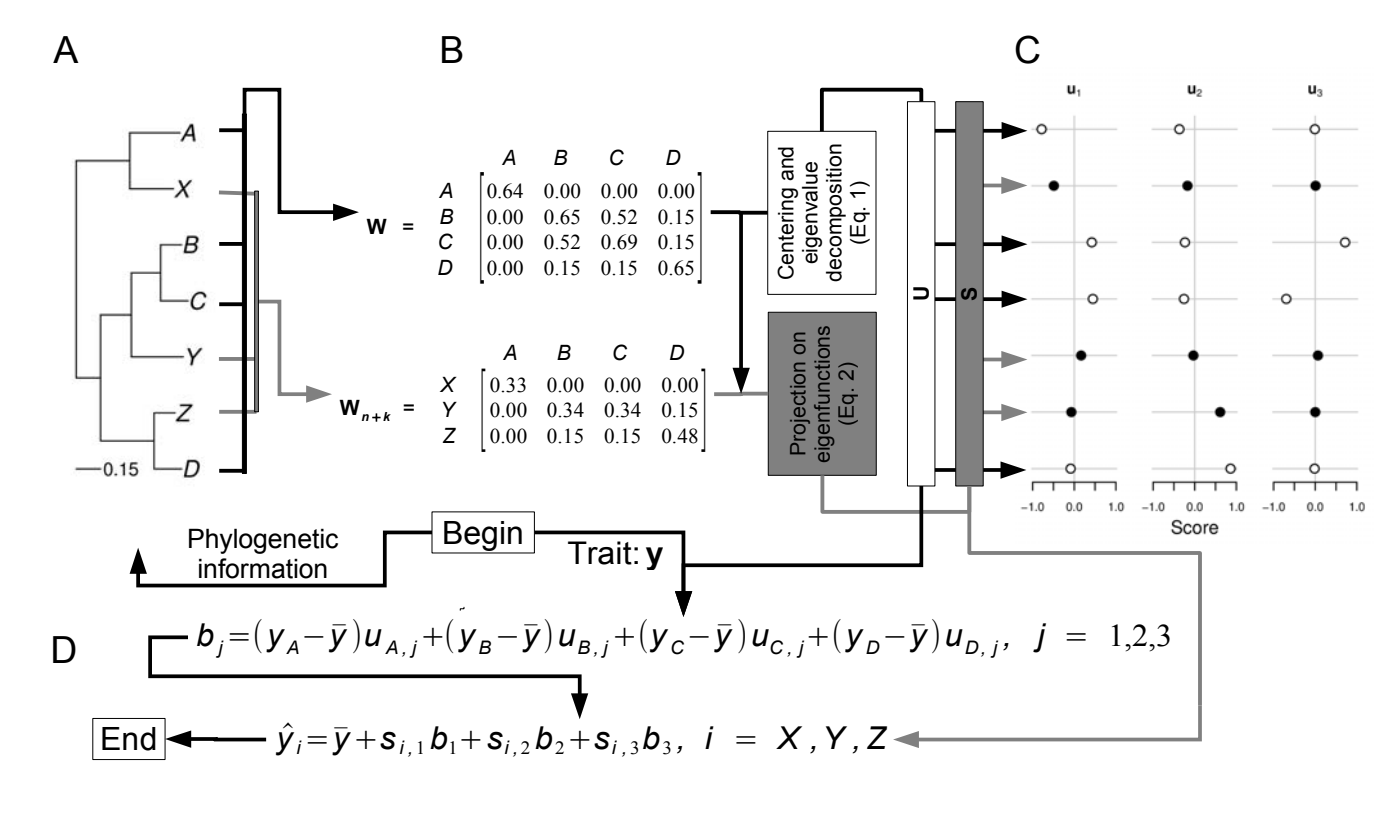
Figure 1. Example illustrating the approach for modeling trait values using phylogeny. We start with trait values (**y**, mean trait value: $\bar{y}$), which are known for species *A*-*D* and are estimated for species *X*-*Z* using phylogenetic information on all seven species. A) The phylogenetic information is used to estimate a tree. B) Phylogenetic covariance matrices (among species *A*-*D*: **W**, and between species *X*-*Z* and *A*-*D* $\mathbf{W}_{n+k}$) are obtained from the tree. C) These matrices are in turn used to obtain the species score matrices **U** (by eigenvalue decomposition after row and column centering on means) and **S** (by projection on the eigenfunction defined for species *A*-*D*). D) Score matrix **U** is used to estimate parameters **b** of a linear model that is finally used to estimate trait values $\hat{y}$.

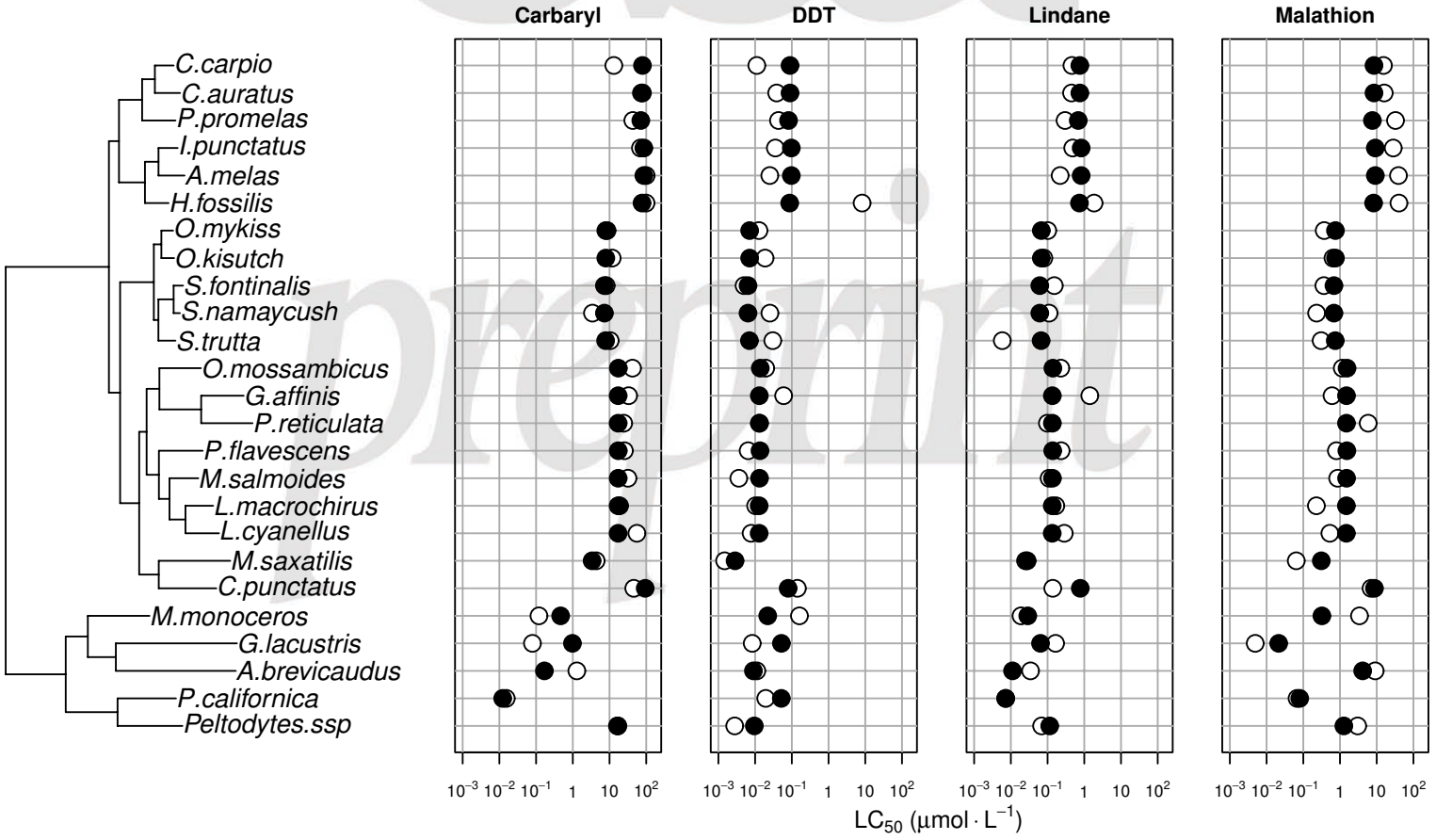
Figure 2. The 25-species model for Malathion, DDT, Lindane and Carbaryl. White and black markers represent the observed and fitted values, respectively.

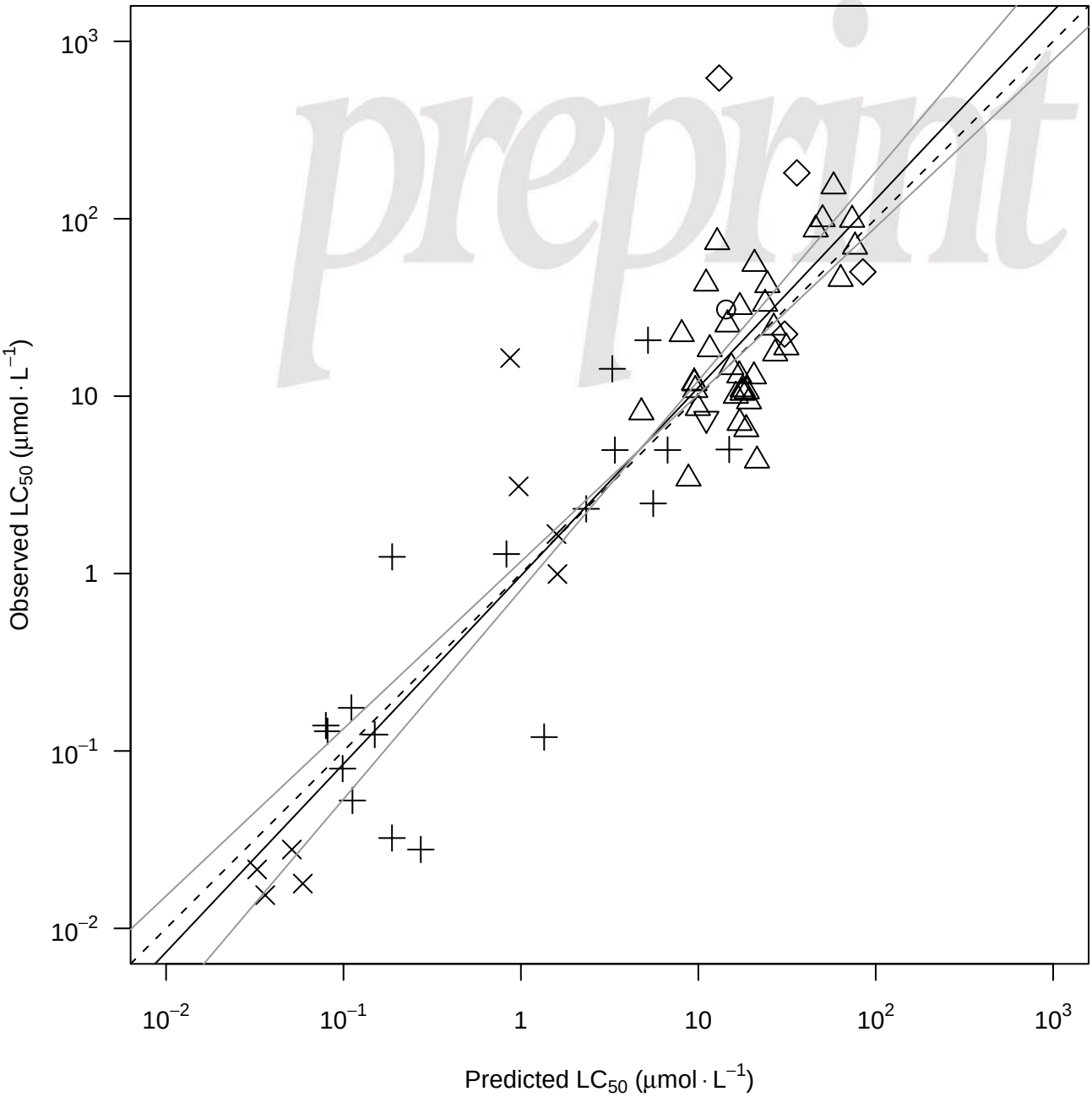
Figure 3. Relationship between predicted and observed $\log_{10}$ (LC_{50}) for Carbaryl (markers: ○ amphibian, △ fish, + crustaceans, × insects, ◇ mollusks, and ▽ annelid; regression line: solid black, confidence limits of the slope: solid grey, 1:1 line: dashed).

Figure 4. Deviation factor, i.e., the number of times tolerance is over- or underestimated by the phylogenetic model (overestimation: positive values underestimation: negative values; A; ○ : absolute value > 10, grey backgrounded marker: 1 < absolute value < 10, ● : absolute value < 1) and the LC_{50} values (B; ○ : observed values, ● : predicted values) with respect to their taxonomy (abbreviations: superphylum Lopho – Lophotrochozoa; phylum An – Annelida, Moll – Mollusca; Class Gas –

615 Gastropoda, Bi – Bivalvia, Am – Amphibia; subclass Or – Orthogastropoda, Ba – Basommatophora;
616 superorder En – Endopterygota, Exopteri – Exopterygota, Proacantho – Protacanthopterygii; order
617 Amphipo – Amphipoda, Is – Isopoda, Co – Coleoptera, Di – Diptera, He – Hemiptera, Pleco –
618 Plecoptera, Cy – Cyprinodontiformes, Salmonifo – Salmoniformes, Silu – Siluriformes; family Am –
619 Ampullariidae, Me – Melanopsidae, Gamma – Gammaridae, Po – Pontoporeiidae, Camb –
620 Cambaridae, Pal – Palaemonidae, Pe – Penaeidae, Ne – Nepidae, No – Notonectidae, Pt –
621 Pteronarcyidae, Pr – Perlidae, Pn – Perlodidae, Os – Osphronemidae, Ci – Cichlidae, Pc – Percidae, Te
622 – Terapontidae, Cent – Centrarchidae, Mo – Moronidae, Ch – Channidae, Cl – Clariidae, He –
623 Heteropneustidae, Ic – Ictaluridae).


A
B
C
D
A
X
B
C
Y
Z
D
0.15
W =
A B C D
A 0.64 0.00 0.00 0.00
B 0.00 0.65 0.52 0.15
C 0.00 0.52 0.69 0.15
D 0.00 0.15 0.15 0.65
W n+k =
A B C D
X 0.33 0.00 0.00 0.00
Y 0.00 0.34 0.34 0.15
Z 0.00 0.15 0.15 0.48
Centering and eigenvalue decomposition (Eq. 1)
Projection on eigenfunctions (Eq. 2)
U
S
u1
u2
u3
−1.0 0.0 1.0
Score
Phylogenetic information
Begin
Trait: y
D
$b_j=(y_A-\bar{y})u_{A,j}+(y_B-\bar{y})u_{B,j}+(y_C-\bar{y})u_{C,j}+(y_D-\bar{y})u_{D,j},\ j = 1,2,3$
End
$\hat{y}_i=\bar{y}+s_{i,1}b_1+s_{i,2}b_2+s_{i,3}b_3,\ i = X, Y, Z$





Deviation factor
50
25
10
3
1
0
−1
−3
−10
−25
−50
LC50 (µmol · L−1)
10^3
10^2
10
1
10^−1
10^−2
Species
Tubifex tubifex
Pila globosa
Melanopsis dufouri
Lymnaea acuminata
Rangia cuneata
Gammarus fasciatus
Gammarus italicus
Gammarus lacustris
Gammarus pseudolimnaeus
Echinogammarus tibaldii
Pontoporeia hoyi
Asellus brevicaudus
Parathelphusa jacquemontii
Parathelphusa masoniana
Scylla serrata
Procambarus acutus
Procambarus clarkii
Orconectes nais
Orconectes immunis
Macrobrachium dayanum
Macrobrachium lamarrei
Palaemonetes kadiakensis
Metapenaeus monoceros
Peltodytes ssp.
Aedes aegypti
Ranatra elongata
Notonecta undulata
Pteronarcella badia
Pteronarcys californica
Claassenia sabulosa
Skwala ssp.
Hoplobatrachus tigerinus
Pseudosphromenus cupanus
Oreochromis mossambicus
Perca flavescens
Therapon jarbua
Lepomis cyanellus
Lepomis macrochirus
Pomoxis nigromaculatus
Micropterus salmoides
Morone saxatilis
Channa punctatus
Channa striata
Gambusia affinis
Poecilia reticulata
Oncorhynchus mykiss
Oncorhynchus tshawytscha
Oncorhynchus kisutch
Oncorhynchus clarki
Salmo salar
Salmo trutta
Salvelinus fontinalis
Salvelinus namaycush
Clarias batrachus
Heteropneustes fossilis
Ameiurus melas
Ictalurus punctatus
Gila elegans
Pimephales promelas
Ptychocheilus lucius
Gibelion catla
Labeo rohita
Barbus conchonius
Barbus ticto
Carassius auratus
Cyprinus carpio
Cirrhinus mrigala
Family
Order
Superorder
Subclass
Class
Subphylum
Phylum
Superphylum
Infrakingdom
Am
Me
Gamma
Po
Pa
Pu
Camb
Pal
Pe
Amphipo
Is
Decapoda
Co
Di
Ne
No
He
Pt
Pr
Pn
Pleco
Peracarida
Eucarida
En
Exopteri
Or
Ba
Eumalacostraca
Pterygota
Gas
Bi
Malacostraca
Insecta
Crustacea
Mandibulata
An
Moll
Arthropoda
Lopho
Ecdysozoa
Protostomia
Os
Ci
Pc
Te
Cent
Mo
Ch
Perciformes
Cy
Acanthopterygii
Salmonidae
Cl
He
Ic
Cyprinidae
Salmonifo
Silu
Cypriniformes
Protacantho
Ostariophysi
Neopterygii
Am
Actinopterygii
Vertebrata
Chordata
Deuterostomia
Deuterostomia