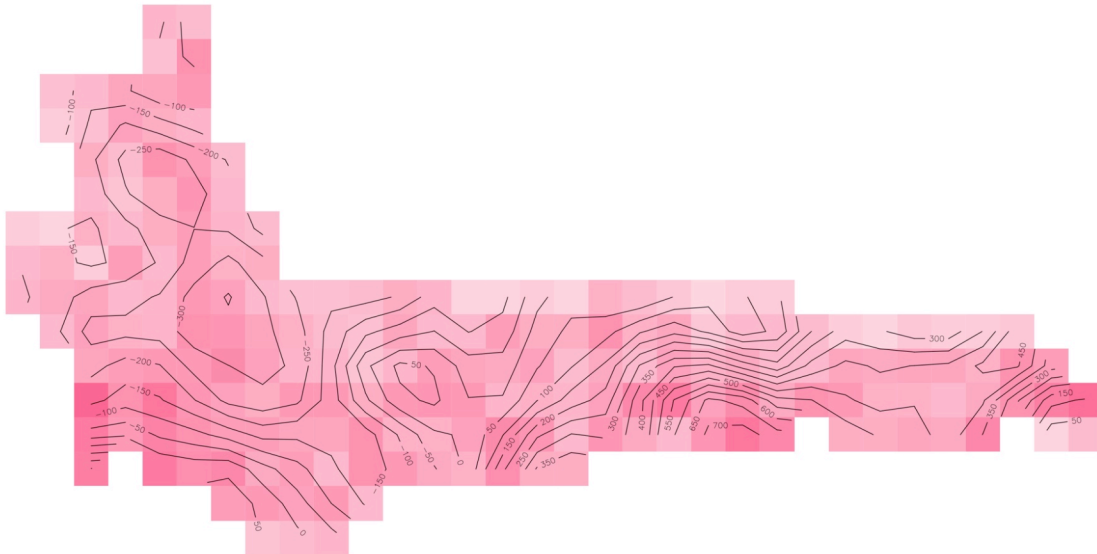




European Distributed Institute of Taxonomy (EDIT)

WP3.3: The bioinformatics toolbox



The EDIT toolbox will help guide conservation strategies. A map of phylogenetic diversity across the Cape Floristic Region, overlaid with a contour surface showing geographically-structured deviation from the expected phylogenetic diversity for the observed number of taxa. Data from Forest et al. 2007; this paper in *Nature* uses techniques designed by the author of this report and which are recommended for implementation in the bioinformatics toolbox.

REPORT 2 – February 2007
Dr Richard Grenyer & Dr Vincent Savolainen
Royal Botanic Garden, Kew

Executive Summary, incl. recommendations & timeframe for realization

Taxonomic information is central to understanding phylogenies, ecology, global change scenarios and in forming conservation strategy.

We can make a fundamental difference to the way taxonomy and taxonomists are integrated into modern biological science by making taxonomic information easier to use.

Taxonomic information is increasingly available in digital form via the internet. Geographic location data, phylogenies and molecular sequences are also increasingly available in this form.

This report describes a piece of software – the EDIT bioinformatics toolbox – which will be able to

- a) Combine these new digital sources of data with researchers' own data so that biodiversity science can be done *faster*,
- b) Allow easy access to cutting-edge numerical techniques so that biodiversity science can be done *better*,
- c) Deliver these benefits in an easy-to-use package that is free to use even on low-end hardware with any existing operating system.

We recommend the toolbox is completed, and that:

The toolbox design is for a set of custom database and statistical tools that run on industry standard server software on top of a self-contained, self-configuring operating system.

The toolbox can be deployed on the user's desktop PC, or as a server on a local network for workgroup or institutional use. The toolbox is accessed via a standard web browser, in which instructions are issued via a friendly graphical interface, and in which publication-quality graphical output is presented.

The design brief has taken four months of EDIT funding, and has already generated a publication in *Nature* using proof-of-concept code.

The toolbox now needs realistic development resources. A single highly-skilled postdoctoral biologist/programmer will be able to create and deploy the software in 4 years.

Introduction

Increasingly, the outputs of the taxonomic process are making the transition from paper publications to the digital realm (Godfray 2002). As they do so, it becomes imperative that the users of taxonomic output are equipped with tools to facilitate the use of taxonomic effort. A broad analogy can be drawn with the evolution of the GenBank facility at the NCBI (<http://www.nlm.ncbi.nih.gov>) over the last ten years. From being a static facility for the deposition of curated molecular sequences, GenBank has evolved into an experimental resource in its own right, a mass of multiple data types, queried and analysed remotely by automated agents, and capable of being the sole raw material for entire scientific projects (Michalickova *et al.* 2002). This goal, of easy access to the raw material of biodiversity studies by humans and software, is what e-taxonomists and software designers need to be aiming for.

This report details some design goals and features for a suite of software tools to enable part of this process. The EDIT phyloinformatic toolbox is intended to synthesise three sorts of data:

- a) taxonomic entities,
- b) phylogenetic trees,
- c) geographic distribution data,

with cutting-edge evolutionary and ecological techniques, and to therefore enhance and promulgate the spread of biodiversity analysis amongst taxonomists, systematists, ecologists and conservationists. There are therefore three main components to the toolbox:

- a) data acquisition agents,
- b) biodiversity analysis tools,
- c) human interface methods.

To see how these components may need to be designed, it will be helpful to visualise two potential users of the toolbox, who lie at opposite ends of the knowledge spectrum.

First is the taxonomist themselves. The only diagnostic feature of a taxonomist is a deep knowledge of specimens of the organisms with which they work, and consequently a data-driven opinion of the taxa into which those specimens group. In summary, taxonomists are collators of information concerning the pattern of evolution and seek to infer the process that gave rise to them. They may have extensive molecular expertise and data, they may have profound ecological knowledge of their chosen taxon, they may have extensive field experience and knowledge of where their taxon of choice are to be found, and

they may have contributed heavily to conservation projects to protect them and their habitat – but not all taxonomists have all or any of these. Consequently, there is considerable benefit to be gained from a software tool that can easily allow taxonomists to extend their work into the areas of phylogenetics, macroevolution, macroecology and conservation.

Second is the ecologist or conservation biologist who works at or above the landscape scale. Here, the only diagnostic feature is a wish to incorporate as many sources of biodiversity information as possible to better understand the functioning of a given piece of the world's surface. In essence, they study a pattern that results from exactly the same evolutionary process as the taxonomist, but which is hidden by a different spatial and temporal scale and extent. Consequently, they may not even be aware that taxonomy and systematics, as methods of inferring the evolutionary process that generates biodiversity, can contribute directly to the study of community assembly, or directly inform conservation prioritisation.

Breaking down these knowledge barriers between fields of enquiry that are really studying two sets of patterns generated by a single process is the core function of the phyloinformatic toolbox and is most effectively done by the creation of simple software tools to allow users to see the benefits quickly, and in their own fields.

Integrating taxonomy into ecology, evolution & conservation

In what ways, therefore, would it be best to integrate the taxonomic process into the study of ecology, evolution and conservation? In answering this question, we also define the nature of the analytical functions that the toolbox should possess, and also gain an understanding of the type of data that it must handle (and by extension the formats in which that data is usually stored); these are key design questions for the software.

The core linkage between the taxonomic process and biodiversity science is the use of taxa as numerical units for analysis. Establishing vital rates in macroevolution (per-lineage speciation (λ) and extinction rates (μ); Nee 2006), calculating the endemic species or feature quotient in an area (Faith *et al.* 2004) or evaluating the fit under neutrality (Condit *et al.* 2002) of an observed community structure to the possible local pool: all these biodiversity tasks can be radically affected by the choice of species considered 'real' for the purposes of analysis (Agapow *et al.* 2004). This choice is most commonly amongst species definitions, but higher taxa (Davies *et al.* 2004) and within-species taxa (Beerli and Felsenstein 1999) are often also counted.

The importance of this choice is often overlooked, but the following analogy makes it very clear. Demography is a core component of many biodiversity studies: estimates of changes in population over time underpin (for example) population viability analyses (Beissinger and Westphal 1998), our understanding of natural selection in wild populations (Coulson *et al.* 2006), epidemiology (Haydon *et al.* 2006) and population genetics (Beerli and Felsenstein 1999). At its core, demography requires a count of individuals of a given class at a given time, and the conclusions follow precisely from these counts. When these counts are of individuals, the matter is clear-cut, but if the definition of individuals were to be based upon the same decision process as taxonomists use to define taxa (and therefore their number), the sensitivity of the conclusions to that decision process would clearly be a major factor to be evaluated before the demographic conclusions could be justified. In biodiversity studies, however, this is rarely done: taxonomic statements, in the form of checklists, are chosen usually by familiarity, availability or expediency, and only occasionally after objective evaluation. Of course, the analogy between defining individuals and defining taxa is flawed: we have a universal “individual concept” (at least in most taxa (Queller *et al.* 2003) but no universal species concept (Isaac *et al.* 2004) – and until we do the best we can hope for is an easy mechanism for assessing the sensitivity of our biodiversity analyses to a given set of taxonomic entities. By forcing the choice of taxonomic units to be explicit, and by making the subsequent analyses simple and repeatable, the phyloinformatic toolbox will provide the best tool yet for establishing this sensitivity analysis as a standard component of best scientific (Isaac *et al.* 2004; Agapow and Sluys 2005).

The benefits of synthesised taxonomic, phylogenetic and geographic data

Several fields of research effort will benefit from the proposed linkage, via the toolbox, of taxonomic, phylogenetic and distributional data. Broadly speaking these fall into the subject areas of ecology, macroecology, macroevolution and conservation strategy. It is important to bear in mind that these benefits are bi-directional: not only does the specialist in any of these fields benefit from the easy access to taxonomic data that the toolbox offers, but the taxonomist (who may have some or all of this data in the first place) will benefit by being able to extend his or her work into fields which (for reasons detailed below) may be currently inaccessible without significant effort.

“...many ecologists are either unaware of the potential benefits of knowing about the phylogenetic relationships in their communities or are deterred by the unfamiliarity of molecular techniques and phylogenetic methods and the accompanying terminology. Similarly, many systematists are unaware of the fascinating ecological questions that can be addressed using the phylogenies they produce or the ways in which knowledge of community composition might bear on studies of character evolution, diversification rate, and historical biogeography.”

(Webb *et al.* 2002)

Here, then, we detail some of the potential uses for an integrated toolbox and in each section note the features of the toolbox that would be required to perform them.

Phylogenetics and macroevolution

Perhaps the simplest use of synthesised taxonomic and phylogenetic information is in macroevolutionary studies; there is no spatial data necessary and the raw material for understanding the processes that have shaped the evolutionary past and present of the group in question is simply the phylogeny that joins them.

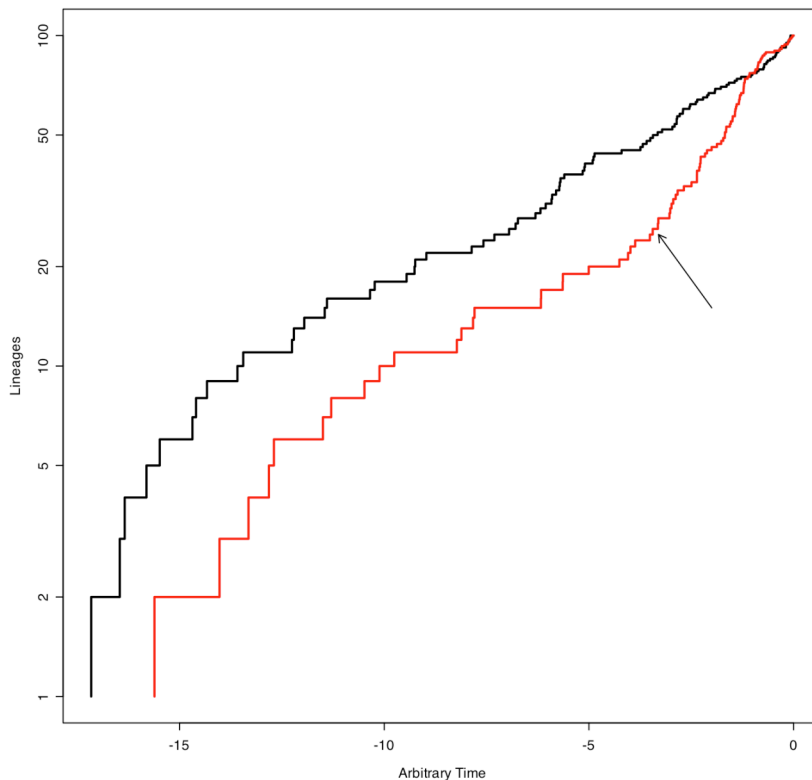


Figure 1: Semilogarithmic lineage-through-time plots of simulated phylogenies as grown and visualised in APE/R and as implemented in the toolbox prototype. The gradient of the trace in semi-log space is the rate of exponential growth, with constant gradient (black) indicating rate constancy, and changes in gradient (red) indicating rate shifts (arrowed).

There is a considerable body of theory detailing how a phylogeny can reveal different properties of the evolutionary regime that shaped them. In essence, this is a study of differential vital rates: the propensity with which a lineage gives rise to new lineages (the per-lineage speciation rate, λ), and extinction

rate, μ (μ), which combine to give a net rate of diversification usually given as r (Nee 2006). Analyses restricted solely to λ (or where μ is not significantly different from zero such that $r = \lambda$: pure-birth processes (Yule 1924) can reveal interesting temporal patterns of origination rate variation across and within phylogenies (e.g. tiger beetles (Barraclough and Vogler 2002); damselflies (Turgeon *et al.* 2005)), and can infer reasons for differences within and amongst subclades (e.g. (Ribera *et al.* 2001). Such differences are best visualised by plotting the accumulation of phylogenetic lineages through time in semi-logarithmic plots (see Figure 1) since pure-birth models are exponential in normal space and linear in semi-logarithmic space; deviations in the gradient show departures from rate constancy. When μ is differentiable from zero (birth-death models, see (Nee *et al.* 1995), there will be a tendency for the gradient of semi-logarithmic lineage-through-time curve to increase towards the present day: the gradient through much of the curve is a constant dictated by r , but towards the present lineages have not had the requisite amount of time for μ to act, leading to an increase in apparent speciation rate. This property is can be diagnosed (Nee 2001) and allows calculation of temporal trends in both λ and μ (i.e. in some cases the signal of a mass extinction can be recovered from a molecular phylogeny of extant species (Nee *et al.* 1994).

Macroevolutionary inference is not limited to cases where phylogenies have a relative or actual time axis (i.e. they can also be applied to cladograms). The topology of the phylogeny itself can yield important information. Under a pure-birth model of cladogenesis (usually referred to as the Equal-Rates Markov (ERM) model), the topology of a phylogeny retains a characteristic and diagnosable shape, because each lineage does not differ in its chances of bifurcating within a given time period. Several metrics exist (Kirkpatrick and Slatkin 1993) for quantifying the shape (or more accurately the imbalance) of phylogenies and significance tests exist to quantify the level of deviation shown by an empirical phylogeny from an ERM process (Mooers and Heard 1997). Further, novel likelihood methods (Chan and Moore 2002; McConway and Sims 2004) now exist which can pinpoint the location of diversification rate shifts within a large phylogeny based purely upon topology without reference to time. It is also possible, should data on phenotypic or genotypic characters be available to the researcher, to locate diversification rate shifts on a phylogeny, and test for their correlation with evolutionary changes in phenotype and genotype.

In summary, macroevolutionary analyses such as the above can determine if, when, where and why major events in the evolutionary history of whole and multiple clades occurred. Further, they can do so using only the same information that many systematists use to 'simply' work out relationships. Consequently, a toolbox which makes these analyses easy and simple to the systematist will have an immediate effect (see Figure 2). The integration of

taxonomic choice into the process is also vital, since there is evidence that macroevolutionary patterns can be influenced by choice of species concept and consequent taxonomic identities (Smith and Patterson 1988; Isaac and Purvis 2004)

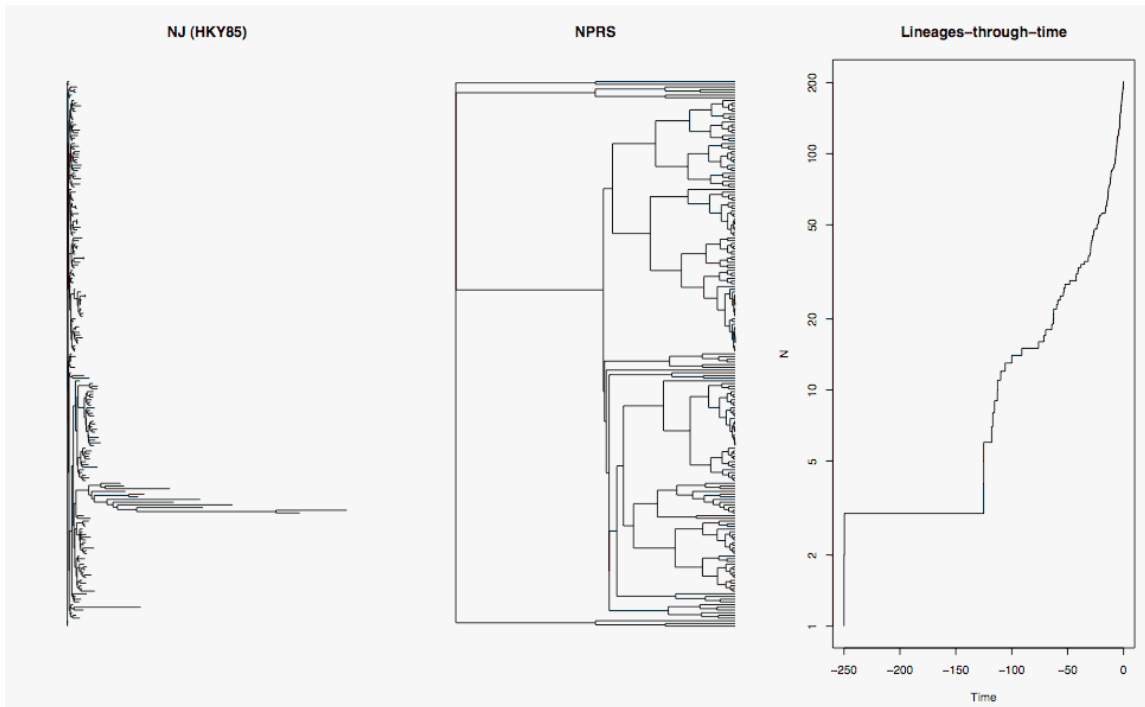


Figure 2: Proof-of-concept macroevolutionary output from the toolbox. The tree on the left was made by the toolbox from sequences provided by the WP6 exemplar taxon group (palms), although they could as easily have been downloaded directly from GenBank. Functions in R applied a simple model of molecular evolution (HKY85) to a neighbour-joining algorithm. Subsequently, further R functions applied the NPRS (non-parametric rate smoothing) algorithm (Sanderson 1997) to transform the metric neighbour-joining tree to an ultrametric tree with branch-lengths in unit time (centre). Finally, a LTT (lineages-through-time) plot was produced suggesting several interesting changes in diversification rate throughout the history of palms.

Ecology and macroecology

By providing an easy way for the non-specialist to access sequence data, existing and novel phylogenetic hypotheses, and distributional data, the toolbox will allow field and theoretical ecologists to integrate phylogenetic data into studies of local communities. The utility of phylogenetic data in community ecology has been known for some time (Brooks and McLennan 1991), but are best divided (Webb *et al.* 2002) into two areas: the study of pure community assembly rules (i.e. whether communities are assembled from species drawn from a local species pool but at random from their phylogeny); and the study of niche construction rules (i.e. whether the taxa in a community are drawn from a

local pool because of evolved differences in phenotype and are dominated by competitive exclusion rules, or evolved similarities in phenotype and dominated by habitat filtering rules (Webb 2000). Outside the community paradigm, there is also the simple question of whether an area has x species because they moved there or evolved there (Ricklefs and Schluter 1993). These questions can all be addressed by the specialist and non-specialist alike with a tool that can integrate phylogenies and spatial data, linked via a user-controlled list of taxon identities.

Macroecology is a loosely defined but vibrant new field at the interface between ecology, biogeography and macroevolution (Brown 1995). As most frequently practiced, macroecology involves documenting and analysing the distribution of taxon richness patterns (Figure 3) across large areas of space (Gaston 2000; Cardillo *et al.* 2005; Grenyer *et al.* 2006). Consequently, the main analytical tools are correlation statistics that can control for spatial non-independence (Diniz Filho and Bini 2005), and general linear modelling techniques which can deal with multivariate correlation structures with unusual distributions of error variance (Storch *et al.* 2006).

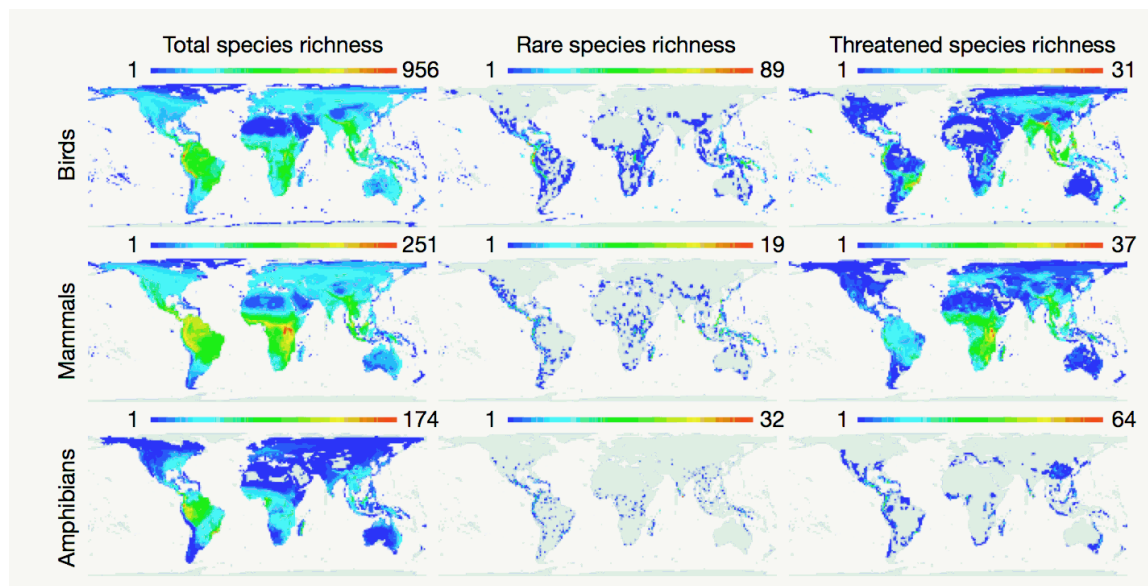


Figure 3: Figure from Grenyer *et al.* (2006) showing the distribution of richness amongst vertebrate species across the world’s surface. As designed here, the toolbox will be able to populate such grids by remotely querying distributional data sources such as GBIF.

From the perspective of the toolbox user, macroecological analysis is a logical progression: once taxonomic and phylogenetic information are assembled in the toolbox, the final ingredient for conservation area selection (see below) is spatial information concerning the distribution of the taxonomic entities being considered. Consequently, if conservation area selection possible, then so is

macroevolutionary analysis for very little extra coding effort. Primarily, the functions required of the toolbox are to be able to take location information (either in the form of point localities, vector range maps or raster presence-absence data at arbitrary pixel location and dimension) and convert it into a richness surface at a resolution and location specified by the user. Handler functions to convert 2D richness surfaces to arbitrary 1D richness gradients would also be beneficial as such transformations are non-trivial at present. Note that it will also be necessary to warn the user when potential data error can be introduced in some of these processes, for example in the transposition of raster (image) data from one set of cell locations to another.

One further subgroup of analyses that is often considered a subset of the macroecological paradigm are the various phylogenetic comparative methods (PCMs). At the risk of generalising, PCMs are statistical techniques which convert the pattern of dissimilarities amongst taxa derived from a phylogeny into a prior expectation of dissimilarity amongst observed data for the species (Pagel 1999). This allows the partitioning of observed relationships amongst species traits into components that could be considered “phylogenetic” and “adaptive”; although more a more accurate depiction would be “well explained” and “not well explained” by a posited model of trait evolution (Freckleton and Harvey 2006). In essence, once the user has selected a model of evolution that they can justify for their particular research question, PCMs allow the users to see patterns and infer causation amongst their observations free from the confounding effects of inherited similarity of phenotype.

Conservation

Taxonomic information has a core role to play in informing conservation strategy. Not only have the defects of taxon choices been demonstrated to play a very important part in determining the effects of a given action (Isaac *et al.* 2004), but when linked to phylogenetic or systematic information, taxonomic information allows a whole suite of novel and informative biodiversity metrics to be calculated, and when spatial data are linked to those taxon concepts a further range of powerful conservation area selection techniques can be employed. Both these goals are some of the most important potential uses of the toolbox.

Phylogenetic Diversity metrics (PD).

Conventional conservation strategy, when it operates above the species level, tends to assume that all species are equal and that preserving the largest number of species is a goal in and of itself (Mace *et al.* 2003). Whilst this logic is fine when there is no need to prioritise amongst species, when resource pressures force prioritization then the assumption of equal worth may not be justified (McIntyre *et al.* 1992). Phylogenetic information provides one amongst

many ways of performing this prioritization (May 1990). The pair-wise distances amongst species in a phylogeny are a measure of the biological differences amongst them that have accumulated over the course of evolution (Faith 1994); there is clear analogy here with the underlying rationale for the use of PCMs (Owens and Bennett 2000). Put another way, the worth of a set of species is expressed not simply as their number, but as the sum of the accumulated genotypic and phenotypic divergence between them; the currency units of diversity: this is best expressed simply as the **sum of the branch lengths connecting the species on a phylogeny**. Consequently a set of three species of mouse would rate a much lower diversity score than did a set comprising a mouse, a sloth and a panda. This intuitively appealing and powerful metric is known as PD (phylogenetic diversity) (Faith 1992), although other related metrics exist (Vane-Wright *et al.* 1991).

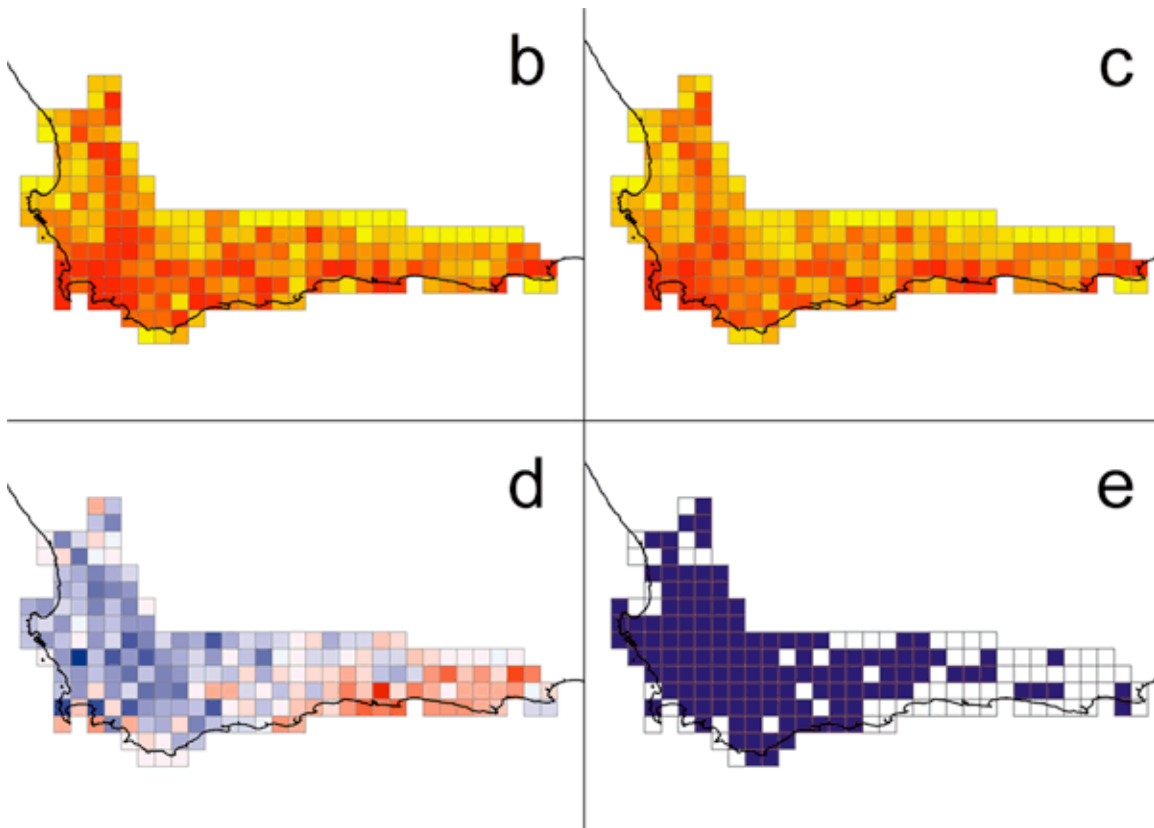


Figure 4: Compound figure from Forest *et al.* (2007) showing the highly congruent distribution of taxon richness (b) and PD (c) can obscure geographical pattern in the expected amount of PD for a given level of richness, as shown by residuals from a non-parametric regression (d; +ve = red, -ve = blue) and by randomisation tests (e; significantly low PD = blue)

Recently, the use of PD metrics has been called into question (Rodrigues *et al.* 2005) because it was found through simulation that PD tends to scale in a predictable way with taxon richness. However, in a recent paper that utilised

some of the prototype functions from the toolbox (Forest *et al.* 2007), it was shown that even a close correspondence between taxon richness values and PD values can hide an important decoupling. That is: some areas have more or less PD than you would expect given their number of species (see Figure 4), and that this variation means that choosing areas to maximise the number of taxa protected does not protect the maximum possible PD. This decoupling of diversity metrics has some quite profound implications for conservation strategy should the tools to examine it be made more widely available.

Conservation Area Selection.

Many problems in conservation biology fall into one of two main classes: a) how much resource is needed to achieve a goal? and b) given a level of resources what is the best goal to achieve? When applied to conservation area selection (Williams *et al.* 2005), these questions become:

- 1) **What is the smallest number of areas with which all species in a set can be protected (the *Species Set Covering Problem - SSCP*), and where are they?**
- 2) **What is the largest number of species that can be protected with *n* areas (the *Maximal Covering Species Problem - MCSP*), and which species and areas are they?**

Both these classes of problems are solved efficiently by a branch of mathematics known as “integer programming” (IP). In many cases, the solutions to these combinatorial problems are guaranteed to be optimal: although when there are many possible solutions it may not be practical to elucidate all (or even many) of them, for many conservation questions it is sufficient to know that one course of action is not sub-optimal.

Expressing the two questions as IP problems following Williams *et al.* (Williams *et al.* 2005), with a binary integer objective function to be minimized or maximised, and a set of linear constraints makes it apparent how solutions are found:

The SSCP:

$$\begin{aligned} & \text{Minimize } \sum_{j \in J} x_j \\ & \text{subject to } \sum_{j \in M_i} x_j \geq 1 \quad \forall i \in I \end{aligned}$$

where I is the set of all species i , J is the set of all areas j inhabited by any i and M_i is the set of all j inhabited by i . The decision variable x_j is constrained to be binary, and the linear constraint forces at least one area from the range of each species to be chosen.

The MCSP:

$$\begin{aligned} &\text{Maximise} && \sum_{i \in I} y_i \\ &\text{subject to} && y_i \leq \sum_{j \in M_i} x_j \quad \forall i \in I \\ &\text{and} && \sum_{j \in J} x_j = P \end{aligned}$$

where the binary decision variable y_i takes 1 only if species i is chosen from amongst I , but where y_i can take 1 only if 1 or more of areas j from the range of i is chosen. This choice of area j is limited only by the single constraint that the sum of (binary) variable x_j is equal to the fixed number of areas P that represents available resources.

Alone, these powerful techniques would be a valuable addition to the toolbox. However, it is possible to integrate the MCSP with the PD metric in order to ask the question “how much PD can I protect with a given level of resources?”. By comparing this answer with that to the species equivalent question to find out if simply protecting as many species as possible is a suboptimal choice for a given situation. The translation to the PD case (Rodrigues and Gaston 2002) is accomplished by letting I represent the set of branches i in the phylogeny, and M_i to be the set of areas inhabited by any species descended from that branch, then solving by maximising y_i as before. Note that the solution to the PD equivalent of the SSCP is identical to the original SSCP, because to protect the entire tree one must protect every species.

The design of the toolbox

We are now in a position to make a number of design requirements of the proposed toolbox. Firstly it must be network aware: incoming taxonomic, phylogenetic and geographic data sources are all accessed via the internet. Secondly, it must have a strong relational database core: not only does the taxonomic data require relational lookups in order to be able to implement handle multiple synonyms, but in addition several of the analytic functions (richness mapping, PD area selection) would benefit from easy access to a relational database. Third, users must be able to integrate these data into a number of complex simulation, mathematical and statistical analyses easily, and

produce graphical output that is immediately suitable for publication. Fourth, the user must be able to do all this without significant training in use, and via a standardised user interface. And finally, this should all be accomplished with minimum maintenance throughout the lifetime of the tool.

It would be possible to achieve all these goals with a single executable written in a low-level programming language such as C or its derivatives. There are several reasons why such an outcome may not be the optimal way to proceed:

- 1) **Redundant programming tasks.** Whilst there exist numerous code libraries for various biological data processing tasks in low-level languages (e.g. C or one of its descendants), there would necessarily be considerable programmer effort invested in combining and linking those routines together to perform a given toolbox function. For example, in order to demonstrate (using Pybus' γ statistic) a significant speciation rate change in a user's tree, the programmer would have to a) transfer information from a phylogenetic data format handler such as the NEXUS class library (Lewis 2003), to a binary tree data structure – see discussion in (Berry *et al.* 2005); b) write an algorithm to traverse that data structure correctly and calculate nodal depths; c) write an algorithmic implementation of the γ -statistic with reference to the original description (since the license allowing reuse of code from existing implementations is unclear (Pybus and Rambaut 2002) d) correctly calculate the null distribution of γ with reference to standard numerical techniques (Press *et al.* 1992) and f) return the statistic and associated p value with a portion of the user interface.

Whilst none of these steps is impossible, they are certainly time-consuming. In addition, they require that the initial programmer is both highly competent at the language of choice *and* a highly competent statistician and biologist, and that both statistical and biological expertise is available for support and error correction throughout the product's life.

- 2) **Time/feature-limited programming tasks.** There are simply too many components, each of them fully-featured and significant programming tasks in themselves, for the toolbox to be coded from scratch, or with reliance upon free class libraries in a compiled high-level language. Further, assuming the time and resources were available to code up a relational database for example, it is likely that the performance and feature set of the resulting component would be less than are currently available in pre-existing packages. This is particularly true of the GIS and scientific graphics components.
- 3) **OS-limited programming tasks.** Probably more than any other factor, the choice of target operating system has limited the take-up of taxonomic and phylogenetic software by the wider community. As a prime example, PAUP*,

the tool of choice for the majority of systematists for the last decade, has been responsible for the population of many systematists offices with Apple hardware that is usually not present in the laboratories of their ecologist neighbours. Consequently the ecological audience has not been well served with phylogenetic programs suitable for the novice.

It might, therefore, be imagined that developing a multi-platform application, compiled against a number of major operating systems (Windows, OSX, Linux), would be the sensible solution. The existing software landscape indicates exactly how difficult this is to achieve in practice. In essence, only command-line tools such as MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) have reliably colonised all three environments. With the notable exception of TreeView (Page 1996), very few GUI (graphical user interface) programs have become ubiquitous because the developer effort of maintaining three separate code bases (due to the lack of a reliable cross-platform GUI kit) is prohibitive. And yet a user-friendly, consistent GUI is a prerequisite for the toolbox.

As a result of these considerations, we can state several features of the toolbox “design philosophy”:

- 1) Minimise novel coding by using existing code and software as much as possible.
- 2) The toolbox is not to be a single standalone program, but a suite of existing software tools, modified for biodiversity analysis.
- 3) Maintenance of the toolbox should, as much as possible, benefit from the IT expertise of the host institution, and not be contingent upon the biological or mathematical abilities of a single person.

Further, we can suggest some desiderata:

- 4) The toolbox should not be limited to a single operating system,
- 5) The toolbox should be as freely distributed as possible,

The sum of these requirements leads us to an interesting design decision. Since the best databases, GIS tools and analytical functions already exist as separate software packages, with their own user base, maintenance communities and distribution mechanisms, the EDIT toolbox will not be a piece of new software, but an integration of existing tools: this integration takes places at the operating system level, not at the executable level, and once some coding is done to allow the different software packages to communicate (Figure 5), maintenance of the toolkit becomes not a programming task but a systems administration task in which most organisations have considerable experience.

The EDIT toolbox is therefore best actualised as a Live LINUX distribution. To the layman, this means simply that it will be distributed as either a CD or preferably a USB stick. It will run on any current ix86 hardware setup (note this includes all recent Apple hardware) and self-configure to the local hardware. The toolbox itself will contain a lean operating system, which simultaneously maximises performance (even on older hardware) and ensures that local variations in available libraries, permissions or versions of operating systems do not prevent the toolbox from functioning. It is becoming quite commonplace to assemble entire linux-based operating systems for a particular purpose – for example the UK Natural Environment Research Council produce just such a live linux distribution, tailored for sequence bioinformatics use (BioLinux - <http://envgen.nox.ac.uk/biolinux.html>). Tools to tailor, make and configure USB-based linux distributions are available at <http://www.live-linux.org>.

Before moving onto precise design features, it is worth bearing mind what the structure of the toolbox will be. On top of the Linux core operating system, the toolbox will comprise the following GPL (i.e. freely available, modifiable and redistributable with attribution) licensed software:

a) ***The statistical and graphical environment R*** (<http://www.r-project.org>), which will handle all analysis, scientific functions, and graphical output. In addition, R provides interface functionality to GIS, database and system calls, so the “glue” code that gets the different components passing information between each other can be written in R. If convenient, however, it could be implemented as shell scripts. The final important R feature is obtained through an add-on package, R-PHP [<http://dssm.unipa.it/R-php/>]. R-PHP simply shifts the R input and output streams from the command line to a suitably configured webserver. The user issues R commands and receives R output (text and graphics) using a standard web browser. Further, R can be linked in to any web application written in PHP (an industry standard language), so that a full easy-to-use GUI for the toolbox can be written in PHP, and accessed via a web browser. Finally, R can issue SQL, SOAP, XML and Javascript queries through network sockets, so that queries of remote data sources can all be handled from within R itself.

b) ***The industry-leading relational database postgresql*** (<http://www.postgresql.org>), which will have a dual purpose. First it will underly the PHP web application that allows the toolbox components to be coordinated. Second it will be an analytic database in which users can place data downloaded from network resources, and that can be passed to or result from calculations in R. In addition, postgresql has a GIS-component, PostGIS (www.postgis.org), that adds support for geographic referencing and manipulation to database objects.

c) **The industry-standard webserver Apache (www.apache.org) and application language PHP (www.php.org)** which will run the PHP web application that interfaces R and postgresql. The user will interact with the toolbox by pointing their web browser at their own machine. This has the added advantage that the whole toolbox can be run on a central server for institutional use with multiple users. This would simply require an addition to the PHP code for user authentication and data storage; code libraries to do this are available under the GPL at a number of repositories.

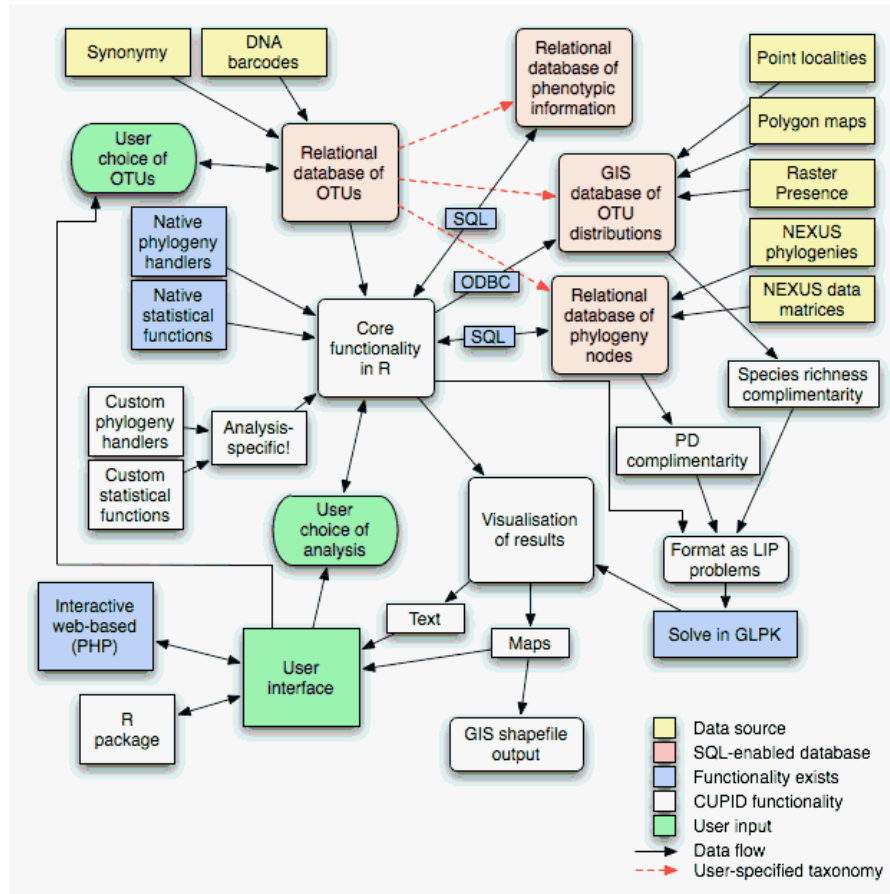


Figure 5: Design schematic for the toolbox. The working title of the proof-of-concept software is CUPID (“Kew PD”), hence white boxes represent “glue” code in R that needs to be written to achieve the design goals specified in this document, whereas blue code exists already as part of the various toolbox components and add-on packages. OTUs (operational taxonomic units) represent the taxa that the user has decided are ‘real’ for the purposes of a given analysis; as can be seen, the design of the toolbox reinforces to the user that taxonomic choices underlie the quantitative results of their analyses.

Network Data Sources

Note: an API is a generic term for the instructions by which a program (e.g. an online database or a software library) can be interacted with by a user or their program. Well-written databases and software have a published API distributed with them, so that third-parties can write their own software to interact correctly with the original program.

Taxonomy

The toolbox has been designed to be non-prescriptive. It must be able, at the users instructions, to query remote data sources that are increasingly available to obtain data that would have been previously very time-consuming for the user to access. However, each of the network data inputs must have a user-specified analogue: if a user wants simply to input their own set of taxonomic choices, phylogenetic trees and/or distribution data, and simply use the toolbox as an analysis platform, that should be encouraged, whilst still allowing the user with none of the above data to access each of the data types through the internet.

In many ways, taxonomic data is the most difficult type still to access through the internet. Indeed, much of the EDIT programme is designed to facilitate this process. Consequently, it is difficult to design precise handlers for remote taxonomic data sources until the design of the EDIT outputs themselves is finalised. Several other data sources exist, however.

Firstly, and perhaps most importantly, is the CATE design (www.cate-project.org) for taxon-specific taxonomic clearing-houses. A CATE node, once the design is finalised, will be able on interrogation via the internet to provide a consensus taxonomy for its group, together with the assenting and dissenting opinions that underlie the consensus. Clearly, this information is of the highest importance for the non-taxonomist who wishes simply to know which taxa the taxonomic community currently regards as real. Therefore, building an R handler in the toolbox to communicate with the CATE API considered a very high priority should the CATE model be adopted upon its final launch.

Secondly, various commercial taxonomic resources are being developed and might reach stability in the lifetime of the development process. Chief amongst these are the ZooBank project (www.zoobank.org), ITIS (<http://www.itis.org>), IPNI (<http://www.ipni.org>), the Species2000 list (<http://www.sp2000.org>), and uBio (<http://www.ubio.org>). Whilst interrogating these lists via HTTP queries is simple to do, it should be noted that many of these resource yet has the synonymy information that really makes taxonomic data warehousing useful. It should therefore be noted that some development effort might need to be put into developing taxon-specific query agents for resources where this information

is available. One good example would be the Mammalian Species Of The World (<http://nmnhgoph.si.edu/msw/>) (Wilson and Reeder 2005) website and database, for which synonym information can be queried, and which will have a stable API and web address on its launch in mid-2007.

Phylogeny

Again, many of the major phylogenetic resources on the internet are in a state of transition. Currently, the pre-eminent resource is TreeBASE (<http://www.treebase.org>), which is manually searchable but which does not have an API for agent use. A new revision of the existing database is planned that will have a direct SQL API, but it is not at present available and is behind schedule. Furthermore, a complete revision (TreeBASE II) is planned as part of the NSF CIPRES initiative, and will have a SOAP API which will be query-able from within R. At present, however, access to TreeBASE I will have to be through some of the “screen-scrape” HTML access functions that are present within the apTreeshape package in R (Bortolussi *et al.* 2006)

More immediate success can be had with the automated retrieval of sequence data from networked resources. R contains several add-on packages which facilitate the querying and downloading of tagged sequence data. Of particular note are the packages APE (Paradis *et al.* 2004), SEQINR (Charif and Lobry 2006) and the BioConductor suite (<http://www.bioconductor.org>). The R-package APE (Paradis *et al.* 2004) contains functions to estimate parameters in multiple models of molecular evolution, and to fit topologies via neighbour-joining algorithms. Topology-evaluation under ML and tree-searching should be available within APE within a year.

As a result of the generally poor availability at present of phylogenetic information over the internet, it is imperative that the toolbox be able to read and write standard (Newick, NEXUS) format phylogenetic trees that have been created in other more standard phylogenetic packages. Such functionality can already be found in the APE package. It is worth considering whether several of the GPL phylogenetic tools (for example PHYLIP (Felsenstein 2007) MrBayes (Huelsenbeck and Ronquist 2001) and r8s (Sanderson 1997)) should be included in the toolbox for ease of use.

Geographic data

Increasingly large amounts of geographic data are available via internet resources, and with an underlying GIS database such as PostGIS, that can be operated through handler functions in R. Making use of this distributional information is a key feature of the toolbox. However, to begin, it is worth noting that R already has within it the ability to read and write ESRI format shapefiles

and ArcInfo coverages. Consequently, a facility for user to upload their own distributional data for analysis is an imperative. Likewise, once geographic analysis has been completed, the ability for the user to output results in one of these formats is also important.

The preminent resource for distributional data via the internet is the GBIF facility and the local nodes that distribute such information. However, there is still considerable fluidity (see, for example, <http://newportal.gbif.org>) in the APIs for accessing GBIF distributional data, and considerable effort will need to be expended to integrate the toolbox with this rapidly changing target.

One other notable geographic resource that deals with the biodiversity data for north and south America is NatureServe (<http://www.natureserve.org>). NatureServe is beginning to implement APIs for much of its data. The spatial data, however, is often hedged around with limitations on usage and which seriously undermine its utility as a biological resource. However, it is possible that at some time in the future realistic usage of the dataset will become possible and a close watch should be kept upon the development of the API (http://services.natureserve.org/about/species_location_data.jsp) as a target.

Various local and continental atlas projects (the VLIZ datacentre – http://www.vliz.be/EN/Data_Centre/Data_Centre_intro; the Atlas Flora Europaea project - <http://www.fmnh.helsinki.fi/english/botany/afe/index.htm>) have been digitized over the last few years; many of these data providers are involved in the Edit project in various forms, and considerable effort needs to be put into allowing the toolbox to interface with these data providers. Many of the datasets are already integrated into GBIF in various forms, but a direct API would ease and simplify the access process.

References

- Agapow, P. M., O. R. P. Bininda-Emonds, et al. (2004). "The impact of species concept on biodiversity studies." Quarterly Review of Biology **79**(2): 161-179.
- Agapow, P. M. and R. Sluys (2005). "The reality of taxonomic change." Trends in Ecology & Evolution **20**(6): 278-280.
- Barracough, T. G. and A. P. Vogler (2002). "Recent diversification rates in North American tiger beetles estimated from a dated mtDNA phylogenetic tree." Molecular Biology and Evolution **19**(10): 1706-1716.
- Beerli, P. and J. Felsenstein (1999). "Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach." Genetics **152**(2): 763-773.
- Beissinger, S. R. and M. I. Westphal (1998). "On the use of demographic models of population viability in endangered species management." Journal of Wildlife Management **62**(3): 821-841.
- Berry, A., A. Sigayret, et al. (2005). "Maximal sub-triangulation in preprocessing phylogenetic data." Soft Computing **10**(5): 461-468.
- Bortolussi, N., E. Durand, et al. (2006). "apTreeshape: statistical analysis of phylogenetic tree shape." Bioinformatics **22**(3): 363-364.
- Brooks, D. R. and D. A. McLennan (1991). Phylogeny, Ecology and Behaviour: A Research Program In Comparative Biology. Chicago, University of Chicago Press.
- Brown, J. H. (1995). Macroecology. Chicago, University Of Chicago Press.
- Cardillo, M., C. D. L. Orme, et al. (2005). "Testing for latitudinal bias in diversification rates: an example using New World birds." Ecology **86**: 2278-2287.
- Chan, K. M. A. and B. R. Moore (2002). "Whole-tree methods for detecting differential diversification rates." Systematic Biology **51**(6): 855-865.
- Charif, D. and J. R. Lobry (2006). SequinR: a contributed package to the R project for statistical computing. Structural approaches to sequence evolution: molecules, networks and populations. U. Bastolla, M. Porto, H. E. Roman and M. Vendruscolo. New York, Springer Verlag.
- Condit, R., N. Pitman, et al. (2002). "Beta-diversity in tropical forest trees." Science **295**(5555): 666-669.
- Coulson, T., T. G. Benton, et al. (2006). "Estimating individual contributions to population growth: evolutionary fitness in ecological time." Proceedings of the Royal Society B-Biological Sciences **273**(1586): 547-555.
- Davies, T. J., T. G. Barracough, et al. (2004). "Darwin's abominable mystery: Insights from a supertree of the angiosperms." Proceedings of the National Academy of Sciences of the United States of America **101**(7): 1904-1909.
- Diniz Filho, J. A. F. and L. M. Bini (2005). "Modelling geographic patterns in species richness using Eigenvector-based spatial filters." Global Ecology and Biogeography **14**(2): 177-185.
- Faith, D. P. (1992). "Conservation evaluation and phylogenetic diversity." Biological Conservation **61**(1): 1-10.
- Faith, D. P. (1994). "Genetic Diversity and Taxonomic Priorities for Conservation." Biological Conservation **68**(1): 69-74.

- Faith, D. P., C. A. M. Reid, et al. (2004). "Integrating phylogenetic diversity, complementarity, and endemism for conservation assessment." Conservation Biology **18**(1): 255-261.
- Felsenstein, J. (2007). PHYLIP (Phylogeny Inference Package).
- Forest, F., R. Grenyer, et al. (2007). "Preserving the evolutionary potential of floras in biodiversity hotspots." Nature **445**: 757-760.
- Freckleton, R. P. and P. H. Harvey (2006). "Detecting Non-Brownian Trait Evolution in Adaptive Radiations." PLoS Biology **4**(11): e373.
- Gaston, K. J. (2000). "Global patterns in biodiversity." Nature **405**: 220-227.
- Godfray, H. C. J. (2002). "Towards taxonomy's 'glorious revolution'." Nature **420**(6915): 461-461.
- Grenyer, R., C. D. L. Orme, et al. (2006). "Global distribution and conservation of rare and threatened vertebrates." Nature **444**: 93-96.
- Haydon, D. T., D. A. Randall, et al. (2006). "Low-coverage vaccination strategies for the conservation of endangered species." Nature **443**(7112): 692-695.
- Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogeny." Bioinformatics **17**: 754-755.
- Isaac, N. J. B., J. Mallet, et al. (2004). "Taxonomic inflation: its influence on macroecology and conservation." Trends in Ecology & Evolution **19**(9): 464-469.
- Isaac, N. J. B. and A. Purvis (2004). "The 'species problem' and testing macroevolutionary hypotheses." Diversity and Distributions **10**(4): 275-281.
- Kirkpatrick, M. and M. Slatkin (1993). "Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree." Evolution **47**(4): 1171-1181.
- Lewis, P. O. (2003). The NEXUS class library, version 2.0, University of Connecticut.
- Mace, G. M., J. L. Gittleman, et al. (2003). "Preserving the Tree of Life." Science **300**(5626): 1707-1709.
- May, R. M. (1990). "Taxonomy as Destiny." Nature **347**(6289): 129-130.
- McConway, K. J. and H. J. Sims (2004). "A likelihood-based method for testing for nonstochastic variation of diversification rates in phylogenies." Evolution **58**(1): 12-23.
- McIntyre, S., G. W. Barrett, et al. (1992). "Species Triage - Seeing Beyond Wounded Rhinos." Conservation Biology **6**(4): 604-&.
- Michalickova, K., G. D. Bader, et al. (2002). "SeqHound: biological sequence and structure database as a platform for bioinformatics research." Bmc Bioinformatics **3**.
- Mooers, A. O. and S. B. Heard (1997). "Evolutionary process from phylogenetic tree shape." Quarterly Review of Biology **72**(1): 31-54.
- Nee, S. (2001). "Inferring speciation rates from phylogenies." Evolution **55**(4): 661-668.
- Nee, S. (2006). "Birth-Death Models in Macroevolution." Annual Review of Ecology and Systematics **37**: 1-17.
- Nee, S., E. C. Holmes, et al. (1994). "Extinction Rates Can Be Estimated from Molecular Phylogenies." Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **344**(1307): 77-82.
- Nee, S., E. C. Holmes, et al. (1995). Estimating extinction from molecular phylogenies. Extinction Rates. J. L. Lawton and R. M. May. Oxford, Oxford University Press.
- Owens, I. P. F. and P. M. Bennett (2000). "Quantifying biodiversity: a phenotypic perspective." Conservation Biology **14**(4): 1014-1022.

- Page, R. D. M. (1996). "TREEVIEW: An application to display phylogenetic trees on personal computers." Computer Applications in the Biosciences **12**: 357-358.
- Pagel, M. (1999). "Inferring the historical patterns of biological evolution." Nature **401**(6756): 877-884.
- Paradis, E., J. Claude, et al. (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics **20**(2): 289-290.
- Press, W. H., B. P. Flannery, et al. (1992). Numerical Recipes in C: The Art of Scientific Computing. Cambridge, UK, Cambridge University Press.
- Pybus, O. and A. Rambaut (2002). "GENIE: estimating demographic history from molecular phylogenies." Bioinformatics **18**: 1404-1405.
- Queller, D. C., E. Ponte, et al. (2003). "Single-gene greenbeard effects in the social amoeba *Dictyostelium discoideum*." Science **299**(5603): 105-106.
- Ribera, I., T. G. Barraclough, et al. (2001). "The effect of habitat type on speciation rates and range movements in aquatic beetles: inferences from species-level phylogenies." Molecular Ecology **10**: 721-735.
- Ricklefs, R. E. and D. Schluter (1993). Species diversity: regional and historical influences. Species Diversity in Ecological Communities: Historical and Geographical Perspectives. R. Ricklefs and D. Schluter. Chicago, Chicago University Press: 350-363.
- Rodrigues, A. S. L., T. M. Brooks, et al. (2005). Integrating phylogenetic diversity in the selection of priority areas for conservation: does it make a difference? Phylogeny and Conservation. A. Purvis, J. L. Gittleman and T. M. Brooks. Cambridge, UK, Cambridge University Press.
- Rodrigues, A. S. L. and K. J. Gaston (2002). "Maximising phylogenetic diversity in the selection of networks of conservation areas." Biological Conservation **105**(1): 103-111.
- Ronquist, F. and J. P. Huelsenbeck (2003). "MRBAYES 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**: 1572-1574.
- Sanderson, M. J. (1997). "A nonparametric approach to estimating divergence times in the absence of rate constancy." Molecular Biology and Evolution **14**: 1218-1231.
- Smith, A. B. and C. Patterson (1988). "The Influence of Taxonomic Method on the Perception of Patterns of Evolution." Evolutionary Biology **23**: 127-216.
- Storch, D., R. G. Davies, et al. (2006). "Energy, range dynamics and global species richness patterns: reconciling mid-domain effects and environmental determinants of avian diversity." Ecology Letters **9**(12): 1308-1320.
- Turgeon, J., R. Stoks, et al. (2005). "Simultaneous Quaternary radiations of three damselfly clades across the Holarctic." American Naturalist **165**(4): E78-E107.
- Vane-Wright, R. I., C. J. Humphries, et al. (1991). "What to protect? Systematics and the agony of choice." Biological Conservation **55**: 235.
- Webb, C. O. (2000). "Exploring the phylogenetic structure of ecological communities: An example for rain forest trees." American Naturalist **156**(2): 145-155.
- Webb, C. O., D. D. Ackerly, et al. (2002). "Phylogenies and community ecology." Annual Review of Ecology and Systematics **33**: 475-505.
- Williams, J. C., C. S. ReVelle, et al. (2005). "Spatial attributes and reserve design models: A review." Environmental Modeling & Assessment **10**(3): 163-181.
- Wilson, D. E. and D. M. Reeder (2005). Mammal Species of the World. Baltimore, MD., Johns Hopkins University Press.

Yule, G. U. (1924). "A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, FRS." Philosophical Transactions of the Royal Society, London. Series B **213**(21-87).