

SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion

Travis Ingram^{1*} and D. Luke Mahler²

¹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA; and ²Center for Population Biology, University of California, Davis, CA, USA

Summary

1. We present a method, ‘SURFACE’, that uses the Ornstein-Uhlenbeck stabilizing selection model to identify cases of convergent evolution using only continuous phenotypic characters and a phylogenetic tree.
2. SURFACE uses stepwise Akaike Information Criterion first to locate regime shifts on a tree, then to identify whether shifts are towards convergent regimes. Simulations can be used to test the hypothesis that a clade contains more convergence than expected by chance.
3. We demonstrate the method with an application to Hawaiian *Tetragnatha* spiders, and present numerical simulations showing that the method has desirable statistical properties given data for multiple traits.
4. The R package `surface` is available as open source software from the Comprehensive R Archive Network.

Key-words: community convergence, contingency, determinism, ecomorph, Hansen model, OUCH, phylogenetic comparative methods, replicated adaptive radiation

Introduction

Convergent evolution is among the most powerful lines of evidence for the power of natural selection to shape organisms to their environment (Simpson 1953; Harvey & Pagel 1991; Losos 2011). The repeated evolution of similar phenotypes in similar environments implies a deterministic aspect of phenotypic evolution. In some evolutionary radiations, including African cichlids (Kocher *et al.* 1993), Caribbean *Anolis* lizards (Losos 2009) and Hawaiian *Tetragnatha* spiders (Gillespie 2004), communities of similar ecological specialists have evolved largely independently. Such clade-wide convergence can be interpreted as lineages independently responding to the same selective regimes, or equivalently, discovering the same adaptive peaks on a macroevolutionary adaptive landscape (Schluter 2000). In what follows, we do not distinguish between convergence and parallelism, as either represents evidence for nonrandom evolutionary change, and as we model convergence at the phenotypic level without addressing its underlying genetic and developmental basis (Simpson 1953; Arendt & Reznick 2008; Losos 2011).

While studies of convergence have resulted in many key insights into adaptation and adaptive radiation, several issues complicate the statistical detection of exceptional convergence in continuous traits. First, some lineages may evolve similar trait values by chance even in the absence of deterministic

convergence. Simulations demonstrate that even a random walk (Brownian motion) model of evolution can lead to considerable incidental convergence, especially if trait space is low-dimensional (Stayton 2008). If repeated convergent evolution is to be interpreted as adaptation to shared environments, the frequency of convergence should be distinguishable from what is expected by chance. More subtly, tests for convergence may be motivated by the observed similarity of sets of species, such as ‘ecomorphs’ that have evolved similar morphology in response to similar ecological conditions (Williams 1972; Gillespie 2004; Losos 2009). The *a priori* identification of ecomorphs creates the potential for bias in tests for convergence, for two reasons. First, nonecomorph species may be ignored in the analysis, exaggerating the extent of phenotypic clustering in a clade (Losos *et al.* 1998; Beuttell & Losos 1999). Second, testing whether ecomorphs are convergent in a set of traits has an element of circularity if those traits played any role in ecomorph designation. Tests for convergence should be able to rule out phenotypic similarity due to chance, and to avoid identifying candidate convergent species *a priori* when it is inappropriate to do so.

We present a new method for identifying convergent evolution without the *a priori* designation of ecomorphs or selective regimes. The method takes as input only a phylogenetic tree and continuous trait data, and fits a series of stabilizing selection models to identify cases where multiple lineages have discovered the same selective regimes. Our method is called ‘SURFACE’, a recursive acronym for ‘SURFACE Uses Regime Fitting with Akaike Information Criterion (AIC) to

*Correspondence author. E-mail: ingram@fas.harvard.edu

model Convergent Evolution'. It builds upon two recent developments in comparative phylogenetic analysis: methods allowing selective regimes to be 'painted' onto the branches of a phylogenetic tree (Hansen 1997; Butler & King 2004; Beaulieu *et al.* 2012), and data-driven stepwise algorithms that locate evolutionary shifts on a tree (Alfaro *et al.* 2009; Thomas & Freckleton 2012). SURFACE consists of a 'forward' stepwise phase in which selective regimes are added to the tree, followed by a 'backward' phase that identifies cases where the same regime is reached by multiple lineages (Fig. 1). This results in

an estimate of the macroevolutionary adaptive landscape that includes measures of the extent of phenotypic convergence.

IMPLEMENTATION

The SURFACE method is implemented as open source software in the R environment (R Core Team 2012), and is available as the extension package `surface` from the Comprehensive R Archive Network (CRAN). `surface` calls functions in the `ouch` package (Butler & King 2004) to fit

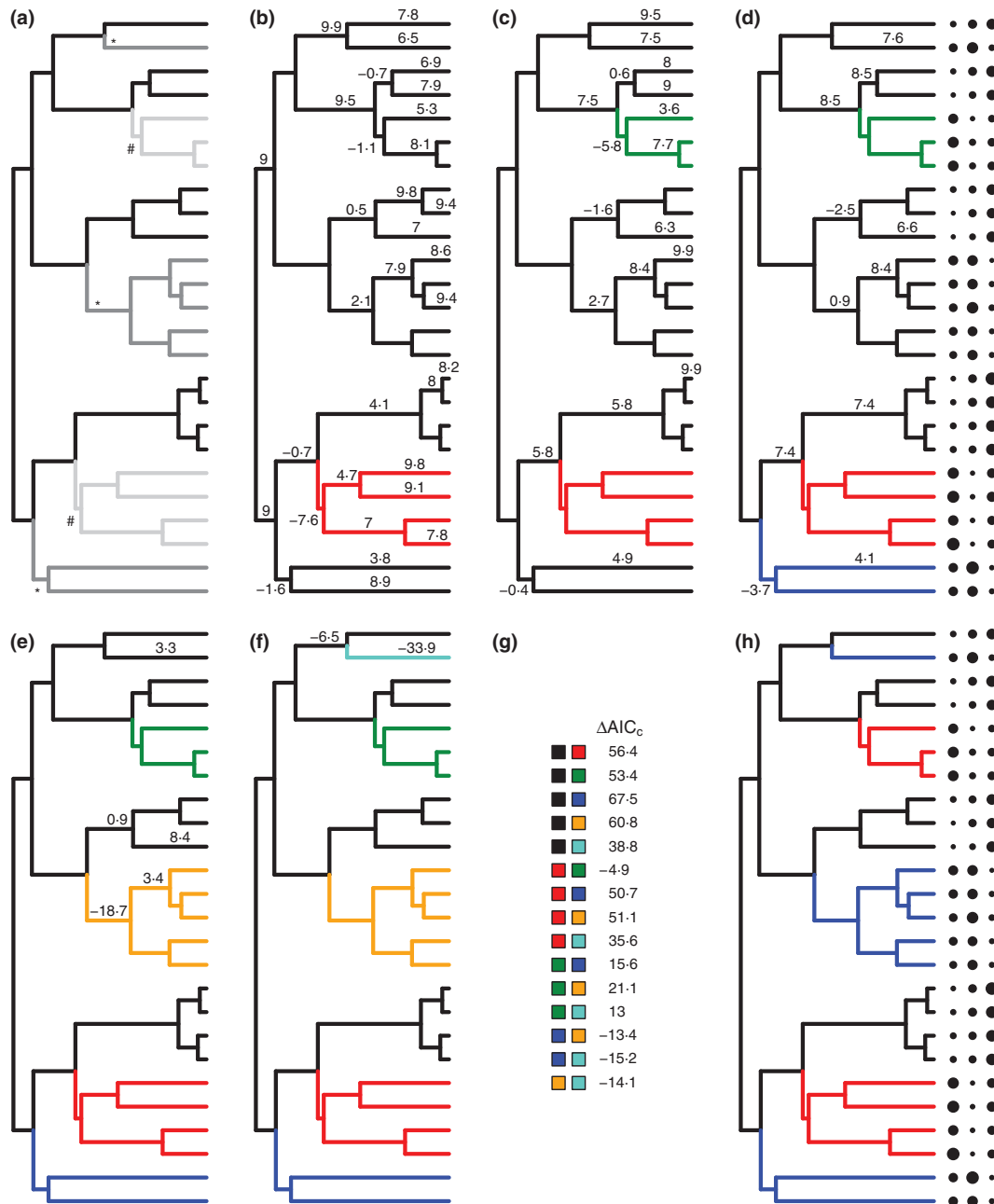


Fig. 1. The forward and backward phases of SURFACE. (a) Generating Hansen model used to simulate trait evolution on a pure-birth tree with a total depth of 10 My, painted with one ancestral and two convergent regimes (shifts denoted * and #). Three traits (values proportional to symbol size) were simulated with relatively rapid adaptation ($\alpha = 0.5$, $\sigma^2 = 0.25$). (b–f) Steps of the forward phase in which a regime shift is added to the branch with the lowest ΔAIC_c score (only values <10 shown). (g) ΔAIC_c values for each candidate pairwise regime collapse in the backward phase (all collapses were compatible and completed in one step). (h) Hansen model returned by SURFACE: in this case, all regime shifts were recovered.

models with selective regime shifts, and incorporates functions from the `ape` (Paradis, Claude & Strimmer 2004), `geiger` (Harmon *et al.* 2008), `pmc` (Boettiger, Coop & Ralph 2012), and `igraph` (Csardi & Nepusz 2006) packages. The two phases of the SURFACE algorithm are carried out by the functions `surfaceForward` and `surfaceBackward`, or by the wrapper function `runSurface`. These functions take as input a phylogenetic tree (this can contain polytomies, which should be left unresolved), and data for one or more continuous traits for each species in the tree. Other features of `surface` include the function `surfaceSimulate` for generating data sets, utilities for converting between data formats and accessing outputs, functions for visualizing the results of an analysis and a vignette that demonstrates the major features of the package.

Materials and methods

FORWARD PHASE: ADDING REGIME SHIFTS TO THE HANSEN MODEL

We model adaptive evolutionary scenarios using the Ornstein-Uhlenbeck (OU) process, a convenient representation of evolution towards adaptive peaks (Felsenstein 1988; Hansen 1997). Under the OU process, a continuous trait X evolves following:

$$dX(t) = \alpha[\theta - X(t)]dt + \sigma dB(t). \quad \text{eqn 1}$$

The magnitude of undirected, stochastic evolution is captured by σ , generally presented as the Brownian rate parameter σ^2 ($dB(t)$ is Gaussian white noise). The deterministic component of trait evolution is represented by α : the rate of adaptive evolution towards an optimum trait value θ (we favour this interpretation over the use of α as a measure of the strength of selection; for further discussion, see Hansen 2012). An intuitive transformation of α is the phylogenetic half-life $t_{1/2} = \log(2)/\alpha$: the expected time for a lineage to evolve halfway towards θ . Hansen (1997) showed that adaptive evolution can be modelled as an OU process in which lineages in different selective regimes are attracted to different optima, and Butler & King (2004) presented methods for specifying adaptive hypotheses by ‘painting’ multiple hypothesized regimes onto the branches of a phylogenetic tree. These ‘Hansen’ models can be used to test for convergence by evaluating support for a model in which multiple lineages shift to shared regimes corresponding to selective factors such as different habitats. SURFACE differs from most previous applications of the Hansen model in that the placement of regime shifts is guided by continuous trait data rather than *a priori* hypotheses about where convergence occurred.

The forward phase of SURFACE adds regimes to a Hansen model, starting with a model in which the entire clade is in a single regime. Maximum likelihood is used to estimate parameter values and model likelihoods L . Under any Hansen model, the trait data \mathbf{X} follow a multivariate normal distribution, with expectation $E[\mathbf{X}]$ and variance-covariance matrix \mathbf{V} (Butler & King 2004). Expected values for the n tip species depend on the regimes experienced during their evolutionary history. Covariances between species pairs depend on the duration of their shared ancestry and the total duration of time each has evolved independently, and higher values of α erode the signal of more ancient events. Nonlinear optimization is used to find the maximum likelihood estimate for α , from which calculation of the maximum likelihood estimates of σ^2 and θ is straightforward (detailed equations are in Butler & King 2004). We assume for computational tractability that α and σ^2 are constant across the tree (but see Beaulieu *et al.* 2012).

SURFACE takes as input one or more continuous trait measurements for each species, though it generally performs much better if the number of trait axes m is at least 2 (see below). If trait data are multidimensional, α , σ^2 and θ become vectors of length m . We assume that traits evolve independently, meaning there are no parameters representing correlated diffusion (σ^2) or adaptation (α) of different traits (but see Bartoszek *et al.* 2012). Because traits evolve independently, we can carry out separate likelihood estimations for each trait, then obtain an overall log-likelihood as the sum of the m log-likelihoods (equivalent to the logarithm of the product of the likelihoods) estimated for individual traits: $\log(L) = \sum_{i=1}^m \log(L_i)$.

After fitting the single-regime Hansen model, we begin the stepwise process of adding regime shifts. We generate candidate models by adding one regime shift at a time to the origin of each branch, causing it and its descendants to be attracted to a new optimum θ . For each candidate model, likelihood estimation is repeated and the log-likelihood is summed across traits. We measure model performance with the AIC, which balances improvements in the log-likelihood against increases in model complexity, and is often used to evaluate competing macroevolutionary models (Burnham & Anderson 2002; Butler & King 2004; Alfaro *et al.* 2009; Harmon *et al.* 2010; Boettiger, Coop & Ralph 2012). We use the finite sample size correction

$$\text{AIC}_c = -2 \log(L) + 2p + \left(\frac{2p(p+1)}{N-p-1} \right), \quad \text{eqn 2}$$

where the sample size N is the total number of trait values ($N = nm$) and p is the number of parameters in the model (Burnham & Anderson 2002). We define the number of parameters as $p = k + (k' + 2)m$. Of these parameters, k correspond to the placement of the regime shifts in the tree (counting the placement of the ancestral regime as a shift), while $k'm$ correspond to the k' optima ($\theta_{1...k'}$) estimated for each of the m traits ($k' = k$ during the forward phase). The additional $2m$ parameters are trait-specific estimates of α and σ^2 . Thus, the single-regime Hansen model has $1 + 3m$ parameters, and adding a regime adds $1 + m$ parameters: a ‘shift’ parameter and an estimate of θ for each trait under the new regime. This way, we account for the complexity both of the adaptive landscape (k' , the number of adaptive peaks) and of the clade’s evolutionary history (k), a distinction that becomes important during the second phase of the analysis.

The performance of each candidate model i is quantified as $\Delta\text{AIC}_c(i) = \text{AIC}_c(i) - \text{AIC}_c$, the difference between that model’s AIC_c and the AIC_c of the model from the previous iteration. Because AIC_c values are calculated after adding log-likelihoods across traits, a new model may improve the AIC_c by increasing the log-likelihood for all m traits simultaneously, or by increasing the log-likelihood for a subset of traits sufficiently to compensate for an unchanged log-likelihood for other traits. Whichever candidate model has the lowest (i.e. best) ΔAIC_c is selected, as long as the magnitude of the improvement exceeds a threshold ΔAIC_c^* ($\Delta\text{AIC}_c(i) < \Delta\text{AIC}_c^*$). In our experience it is generally effective to accept all AIC_c improvements ($\Delta\text{AIC}_c^* = 0$), but users can specify more conservative thresholds guided by conventional interpretations of AIC differences (e.g. $\Delta\text{AIC}_c^* = -2$ or -5 ; Burnham & Anderson 2002) or by simulation (Thomas & Freckleton 2012). Some users may wish to use Monte Carlo simulations to test whether each step constitutes a statistically significant improvement when the previous iteration is treated as the null model (the function `pmcSurface` provides an interface to the `pmc` package; Boettiger, Coop & Ralph 2012).

The regime shift corresponding to the best model is painted onto the tree and retained through subsequent iterations. This process of fitting candidate models and painting one new regime onto the tree per step is repeated until no candidate model meets the criterion $\Delta\text{AIC}_c < \Delta\text{AIC}_c^*$.

(Fig. 1b–f). The number of likelihood searches ($m[2n - k - 1]$) at each iteration can be reduced by using the option `exclude` to skip candidate models that performed poorly during the previous iteration and are thus highly unlikely to be among the best models. The outcome of the algorithm is fully determined by the ΔAIC_c values, although stochasticity can be introduced by using the option `sample_shifts` to randomly sample from among the models within a specified number of AIC_c units of the best model at each step. At the end of the forward phase, we have a fitted Hansen model with k regime shifts painted onto the tree, with m estimates of α and σ^2 , and a k -by- m matrix of optima θ .

BACKWARD PHASE: IDENTIFYING CONVERGENT REGIMES

During the backward phase of SURFACE, we carry out a second stepwise procedure to identify whether subsets of the k regimes can be ‘collapsed’ together to yield $k' < k$ distinct regimes. Recall that each new regime shift added to the Hansen model increased the number of parameters by $m + 1$, accounting for m new optima and one new regime shift. If we subsequently collapse two regimes into one convergent regime, the regime shift parameter remains in the model, but because the vector of optima for the regimes is constrained to be equal, the number of distinct regimes k' decreases by 1, and the number of parameters p decreases by m . While constraining regimes to have the same optima cannot increase the log-likelihood of the model, the AIC may nonetheless improve if the reduction in the number of parameters outweighs any decrease in the log-likelihood.

Starting with the model in which each shift is to a different regime, we move through all pairwise combinations of regimes i and j , and re-fit the model after collapsing them into a shared regime (again adding log-likelihoods across traits). We then calculate $\Delta AIC_{c(ij)}$ for each of these $k(k - 1)/2$ candidate models by comparison to the previous model, and determine which candidate models meet the criterion $\Delta AIC_{c(ij)} < \Delta AIC_c^*$ (Fig. 1g), indicating that the model improves when regimes i and j are treated as convergent. Importantly, not all proposed regime collapses that meet this criterion will necessarily be compatible with one other. For example, regimes A and B may each improve the AIC_c when collapsed with regime C, but not when collapsed with each other ($\Delta AIC_{c(AC)} < \Delta AIC_c^*$ and $\Delta AIC_{c(BC)} < \Delta AIC_c^*$, but $\Delta AIC_{c(AB)} \geq \Delta AIC_c^*$). One can decide either to accept only the pairwise collapse with the lowest ΔAIC_c at each iteration, or to identify any sets of compatible regime collapses.

We use techniques from graph theory to include as many collapses as possible at each step, subject to the condition that no pair of regimes that does not meet the $\Delta AIC_{c(ij)} < \Delta AIC_c^*$ criterion is collapsed. We build an undirected graph whose vertices are regimes, with edges drawn between pairs of regimes that meet this criterion. We then use the function `clusters` in the `igraph` package (Csardi & Nepusz 2006) to identify all connected components of this graph: clusters of directly or indirectly connected regimes that are not connected to regimes from any other cluster. If all pairwise ΔAIC_c values between regimes in a cluster are less than ΔAIC_c^* , all collapses are compatible and the cluster is collapsed into a single convergent regime. If some pairs of regimes in the cluster do not meet the $\Delta AIC_{c(ij)} < \Delta AIC_c^*$ criterion, only the pair of regimes in the cluster with the lowest ΔAIC_c is collapsed. As multiple regimes may be collapsed in a step, the updated model AIC_c is calculated using the updated numbers of distinct regimes k' and parameters p .

This stepwise procedure is repeated until no further collapses improve the model by ΔAIC_c^* (Fig. 1h). The final Hansen model produced by SURFACE has k regime shifts placed on the tree and $k' \leq k$

distinct regimes. The function `surfaceSummary` returns the details of where in the tree the inferred regime shifts occur, as well as estimates for the parameters of the OU model (α , σ^2 and θ) and parameters summarizing the features of the inferred macroevolutionary landscape (Table 1). The extent of convergence in the model can be quantified as $\Delta k = k - k'$, with Δk representing the simplification of the adaptive landscape (decrease in the number of regimes or peaks) when convergent regimes are identified during the backward phase. Alternative measures of convergence include c (the number of shifts that are towards convergent regimes occupied by multiple lineages), or either Δk or c scaled to the number of regime shifts ($\Delta k/k$ or c/k). One could also envision measures of convergence not currently implemented, such as the average magnitude of change in optimum position between convergent regimes and their ancestors.

HYPOTHESIS TESTING: IS CONVERGENCE EXCEPTIONAL?

The algorithm described above attempts to find the best painting of convergent and nonconvergent regime shifts on the tree, subject to the constraints of the stepwise approach. SURFACE may be used as an exploratory tool to visualize evolutionary patterns but can also be used to test biologically motivated hypotheses about the evolutionary history of the clade. We can ask if the clade is characterized by many shifts between adaptive peaks (Simpson 1953) by evaluating the number of inferred shifts k and the improvement of the AIC_c between the initial single regime OU model and the model returned by the forward phase. We can then evaluate the evidence for convergent evolution using the number of cases of convergence (Δk or c) and the further improvement of the AIC_c between the models returned by the forward and backward phases.

An issue that arises when we compare models in this way is that some regimes are often added when the true evolutionary model contained no regime shifts, and some cases of convergence are often identified when convergence was absent from the true model. If we are specifically interested in the number of shifts or cases of convergence, one solution is to simulate data sets under a ‘null’ model that lacks regime shifts and/or convergence, then run SURFACE on these simulated data sets to obtain a null distribution of the metric of interest (k , c or Δk). The proportion of values of this metric (including both null and observed values) at least as large as the observed value provides a P -value for a

Table 1. Parameters representing evolutionary processes and features of the adaptive landscape

Parameter	Interpretation
σ^2	Rate of stochastic evolution (one parameter per trait)
α	Rate of adaptation to optima (one parameter per trait)
$t_{1/2}$	Expected time to evolve halfway to an optimum; $\log(2)/\alpha$
θ	Optimum trait value (one parameter per regime per trait)
k	Number of regime shifts
k'	Number of distinct regimes (after collapsing convergent regimes)
k'_{conv}	Number of convergent regimes reached by multiple shifts
Δk	$k - k'$, the reduction in complexity of the adaptive landscape when accounting for convergence
$\Delta k/k$	Relative reduction in complexity of the adaptive landscape when accounting for convergence
c	Number of shifts that are towards convergent regimes occupied by multiple lineages
c/k	Proportion of shifts that are towards convergent regimes

significance test of whether the observed value exceeds the expectation under the null model. One could also compare the change in AIC_c during the forward and backward phases of the analyses of simulated and real data sets to evaluate whether the model improvement attributable to convergence, in addition to the number of cases of convergence, is exceptional.

We have implemented two types of null models that can be used for hypothesis testing: one captures the general temporal dynamics of diversification without regime shifts and the other includes regime shifts but not deterministic convergence. The first null model may be simple constant-rate Brownian motion, though one may also wish to consider models in which the rate of trait evolution declines over time (e.g. the 'Early Burst' or 'Time' models; Harmon *et al.* 2010; Mahler *et al.* 2010), or in which the volume of trait space is constrained (e.g. the single-regime OU model; Felsenstein 1988; Harmon *et al.* 2010). Null data sets can be simulated using the maximum likelihood parameters of the preferred model for each trait, and may be appropriate for testing either for more regime shifts or greater convergence than expected by chance.

The forward phase of SURFACE may fit relatively few regime shifts to data simulated under this type of null model, meaning the null distribution of c or Δk may be biased downwards because there are fewer regimes that can potentially be collapsed during the backward phase. To generate a null distribution of the extent of convergence that accounts for the presence of regime shifts, one can take as a starting point the Hansen model returned by the forward phase. Estimates of α and σ^2 and the placement of each of the k shifts are preserved, but new optimum values are sampled to break up any tendency for optima to cluster in trait space. To ensure that the volume of trait space will be comparable, new optima for each trait can be sampled from a normal distribution with the mean and variance of the estimated optima. Other options may be preferable if some optima are inferred to be well outside the range of trait data, which may truly reflect incomplete evolution towards distant optima, or may be a biologically unrealistic consequence of model assumptions such as α being constant across the tree or shifts only occurring at the origin of branches. Null data sets simulated under this Hansen null model may contain incidental convergence if optima are similar by chance, and can be used to ask whether the extent of convergence is exceptional even when the presence of regime shifts is accounted for.

APPLICATION TO HAWAIIAN *TETRAGNATHA* SPIDERS

To demonstrate these methods, we investigated patterns of morphological evolution in a well-studied adaptive radiation: Hawaiian spiders in the genus *Tetragnatha* (Gillespie, Croom & Palumbi 1994; Gillespie, Croom & Hasty 1997; Blackledge & Gillespie 2004; Gillespie 2004). Hawaiian *Tetragnatha* consists primarily of two diverse clades: the 'spiny-leg' clade and a clade of web builders (the more distantly related *T. hawaiiensis* has not radiated in Hawaii). Both clades show evidence for repeated convergent evolution, either in traits related to microhabitat such as size, colour and foraging behaviour (spiny-leg clade; Gillespie 2004) or in an extended phenotypic trait, web shape (web-building clade; Blackledge & Gillespie 2004). We used SURFACE to test for convergence in continuous morphological traits within and between the two clades. We note that, apart from body size, our analysis does not use traits that have previously been identified as convergent in *Tetragnatha* (Gillespie, Croom & Hasty 1997), which are either discrete characters or are unavailable for many species. The fact that convergence is known largely for traits and for subsets of the clade that differ from those included here means that while *Tetragnatha* is an interesting

group in which to demonstrate these methods, what follows should not be viewed as a test of existing hypotheses about how selection has shaped the evolution of this group.

We used an ultrametric phylogenetic tree from Harmon *et al.* (2010), containing 58 taxa – 25 spiny-leg and 33 web-building species – after the removal of *T. hawaiiensis*. The tree was constructed from mitochondrial sequence data using UPGMA, and was scaled to have a root age of 4.17 Ma. As trait data, we used log-transformed total length (TL) as a measure of body size, plus two size-independent morphological trait axes (Gillespie, Croom & Palumbi 1994; Harmon *et al.* 2010). To obtain these, we calculated phylogenetic residuals of four log-transformed traits (cephalothorax length, chelicera length, leg spine length and abdomen depth/length ratio) against log (TL), then retained the first two axes of a phylogenetic principal components analysis (pPCA) of these four size-adjusted traits (Revell 2009). The pPCA assumes a Brownian motion model of evolution, in contrast to the more complex Hansen models we are fitting. We confirmed that the axes were qualitatively similar even under a nonphylogenetic PCA, and were thus unlikely to be strongly affected by the method of calculation. Higher values of pPC1 were associated with longer cephalothoraxes and chelicerae, and higher pPC2 was associated with deeper abdomens and longer spines. We analysed this data set with SURFACE using a ΔAIC_c^* threshold of 0 and allowing multiple compatible regime collapses during each step of the backward phase. We also ran SURFACE on 99 data sets simulated under each of the Brownian motion and Hansen null models to evaluate whether convergence in this clade was greater than expected by chance.

The final Hansen model included 10 regime shifts and seven distinct regimes ($\Delta k = 3$) and $c = 6$ convergent shifts (Fig. 2). The AIC_c improved from 495.6 to 430.7 ($\Delta AIC_c = 64.9$) during the forward phase, then to a final AIC_c of 410.0 during the backward phase ($\Delta AIC_c = 20.7$; for comparison, the AIC_c of the Brownian motion model was 517.1). Three convergent regimes were present in the model, each of them reached by one regime shift in each of the two major clades. The majority of the taxa in the spiny-leg clade (21 of 25) were placed into one of the three convergent regimes, each of which was also discovered by a small subclade of one or two species in the web-building clade. Thus, while *Tetragnatha* research has focused on convergence within each of the two subclades (Blackledge & Gillespie 2004; Gillespie 2004), SURFACE found only cases of between-subclade convergence. The estimated phylogenetic half-lives of $t_{1/2} = 0.24, 0.06$ and 0.19 My for log (TL), pPC1 and pPC2, respectively, implied relatively fast adaptation towards optimum trait values (Hansen 2012). However, high rates of stochastic trait evolution ($\sigma^2 = 0.21, 19.6$ and 10.8 My^{-1}) reduced the distinctness of regimes (Fig. 2d and e).

Comparison to the null model simulations did not reveal significantly greater convergence in *Tetragnatha* than expected by chance. The simplification of the adaptive landscape ($\Delta k = 3$) was not exceptional compared to the distributions simulated under either Brownian motion ($P = 0.33$) or the Hansen null model ($P = 0.29$; Fig. 2c). Similarly, the number of convergent shifts ($c = 6$) did not significantly exceed either null distribution ($P = 0.22$) under either null model. Thus, we cannot reject a null hypothesis that the number of cases of convergence found by SURFACE is incidental, resulting from some optima being similar by chance. The failure to recover established cases of convergence is not especially surprising as, other than body size, the traits generally understood to be convergent in this group are discrete characters and thus are not suitable for fitting Hansen models. Thus, while this demonstration of SURFACE provides a useful illustration of the patterns of morphological evolution in Hawaiian *Tetragnatha*, it

should not be taken as refutation of more biologically motivated hypotheses about convergence.

SIMULATIONS TO ASSESS PERFORMANCE

We investigated the performance of SURFACE using two sets of simulations: one to quantify its ability to recover a true model, and the other to evaluate the statistical properties of the simulation-based hypothesis test. First, we ran SURFACE on data sets simulated under different Hansen models to examine the accuracy with which several aspects of the generating model were recovered. We quantified the proportion of branches containing regime shifts in the generating model that were correctly assigned shifts in the fitted model (ignoring the basal shift, which is always present). We also calculated the similarity between the true and fitted models based on the regime assignments of extant

species. The function `propRegMatch` determines for all pairs of taxa (or optionally all pairs of branches) whether they are in the same regime or in different regimes in the fitted Hansen model. It repeats this determination for the generating model, then calculates the proportion of pairs that have the same status in both models. As this proportion approaches 1, it indicates that the fitted model is accurately grouping taxa into the correct regimes. Finally, we compared the number of regime shifts (k) and distinct regimes (k') in the true and fitted Hansen models to assess the recovery of these general characteristics without regard to the specific placement of regimes.

We varied features of the generating model and tree to evaluate which have the greatest influence on the accuracy of SURFACE, with 10 replicates of each set of parameter values. One parameter was varied at a time, while the remaining parameters were set to default values (in boldface). We simulated data on $n = 32, 64$ or 128 taxon trees

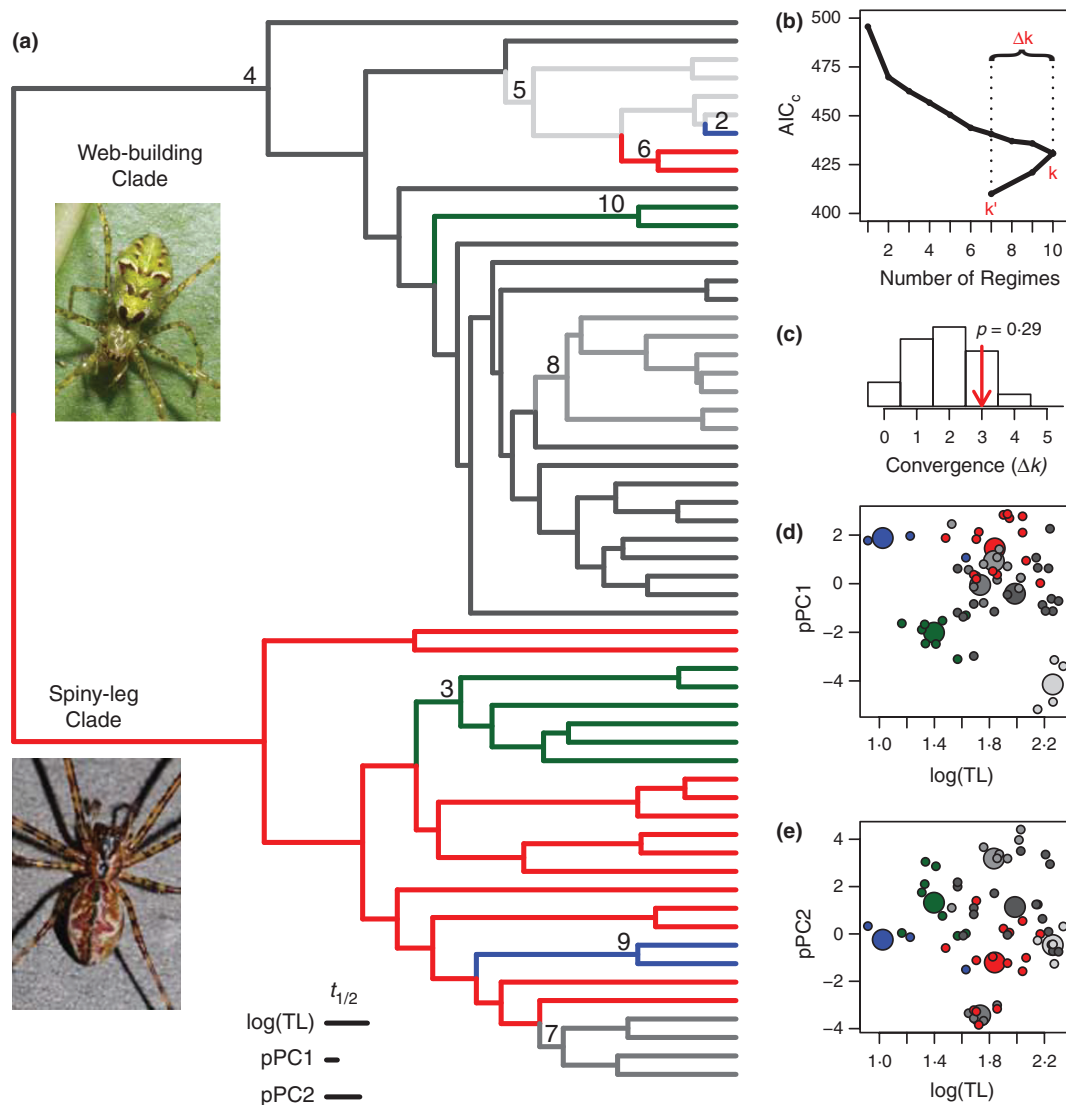


Fig. 2. Results of a SURFACE analysis of the two major clades of Hawaiian *Tetragnatha* spiders. (a) Phylogenetic tree, with `surfaceTreePlot` used to paint convergent (coloured) and nonconvergent (greyscale) regimes onto the branches. Numbers on branches indicate the order in which regime shifts were added during the forward phase, and estimated phylogenetic half-lives for each trait ($t_{1/2} = \log(2)/\alpha$) are shown below the tree. (b) Change in AIC_c during the forward and backward phases of the analysis. (c) Comparison of the observed extent of convergence Δk to the null expectation under a nonconvergent Hansen model. (d and e) Trait values for each species (small circles) and estimated optima (large circles), with regime colours matching those in the tree. Photographs were provided by R. Gillespie.

generated under a pure-birth model, and scaled to a total depth of 10 My. We randomly pruned 0%, 25% or 50% of the taxa from the tree to evaluate the effects of missing data. We simulated $m = 1, 2$ or 4 trait axes, with $\sigma^2 = 0.1$ and $\alpha = 0.25, 0.5$ or 1 (corresponding to an average of $t_{1/2} = 2.8, 1.4$ or 0.7 My to reach a new optimum). We sampled optima for each trait from a normal distribution with $SD = 1, 2$ or 4 , and added $k = \frac{1}{16}n, \frac{3}{16}n$ or $\frac{5}{16}n$ regime shifts towards $k' = \frac{1}{2}k, \frac{3}{4}k$ or k distinct regimes. Randomly selecting branches to receive shifts tends to result in shifts concentrated at the tips, so we also varied whether branches were sampled probabilistically in proportion to their time of origin, with a very recent branch being $1 \times, 0.1 \times$ or $0.01 \times$ as likely to be sampled as branches originating at the root.

Over the parameter space explored, SURFACE was more accurate for data simulated with more traits, faster adaptation and more regime shifts (Figs 3 and 4). Under most conditions, 70–80% of regime shifts were placed correctly, though the proportion was variable and was lower for data sets with few traits. The proportion of recovered shifts declined when shifts were sampled disproportionately towards the root of the tree, suggesting that recent shifts are more likely to be identified accurately. The proportion of pairs of tip taxa correctly assigned to either the same or different regimes was generally much higher, typically exceeding 90% (Fig. 3). Thus, the fitted model tended to resemble the generating model even when shift placement was imprecise, likely because a shift placed on a branch nearby in the tree to the true shift location can have a very similar effect on the expected distribution of traits among extant taxa. The broad characteristics of the true Hansen model k and k' were generally close to the true values as well (Fig. 4). SURFACE tended to slightly overestimate the number of shifts and regimes when there were few regime shifts, and to underestimate the true number when there were many shifts, few traits, or many missing taxa (though the latter is expected because some regimes will be unrepresented in the pruned tree). Processor times required for these

SURFACE analyses varied from a few minutes to over 4 h on Intel x86-64 eight-core processors with 24 Gb RAM, and increased primarily with tree size and to a lesser extent with the number of traits and regimes.

Our second set of simulations evaluated the statistical power and type I error of the simulation-based hypothesis test of whether a clade contains more convergence than expected by chance. We explored the effects of clade size n , number of traits m and the true extent of convergence Δk_{true} . We simulated 20 ‘real’ data sets with each of four levels of convergence under each of the following conditions: $n = 32, m = 4$; $n = 64, m = 1$; $n = 64, m = 2$; $n = 64, m = 4$. For 64-taxon trees, we randomly placed $k = 16$ shifts and set levels of convergence to $\Delta k_{\text{true}} = 0, 2, 6$ or 12 , while 32-taxon trees had $k = 8$ and $\Delta k_{\text{true}} = 0, 1, 3$ or 6 . All other parameters were set to the default values given above. We ran SURFACE on each data set to estimate the extent of convergence Δk_{obs} (results were similar when we measured convergence using c). Then, for each ‘real’ data set, we ran SURFACE on 50 data sets simulated under the Hansen null model to obtain a null distribution of the extent of convergence Δk_{null} . We computed P -values by comparison to the null distributions, and calculated approximate statistical power for each set of parameters as the proportion of significant results ($P < 0.05$) out of the 20 tests.

These analyses indicate that this simulation-based hypothesis test has fairly good power to detect a greater amount of convergence than expected by chance, given multidimensional trait data (Fig. 4). Power was lower for analyses of single traits, likely because more incidental convergence is expected in low-dimensional trait space (Stayton 2008). Type I error, estimated as the proportion of the 20 data sets that did not contain true convergence ($\Delta k_{\text{true}} = 0$) but that nonetheless resulted in significant tests, was 0 or 0.05 for each combination of n and m . These results suggest that the simulation-based test for exceptional convergence has appropriate statistical properties.

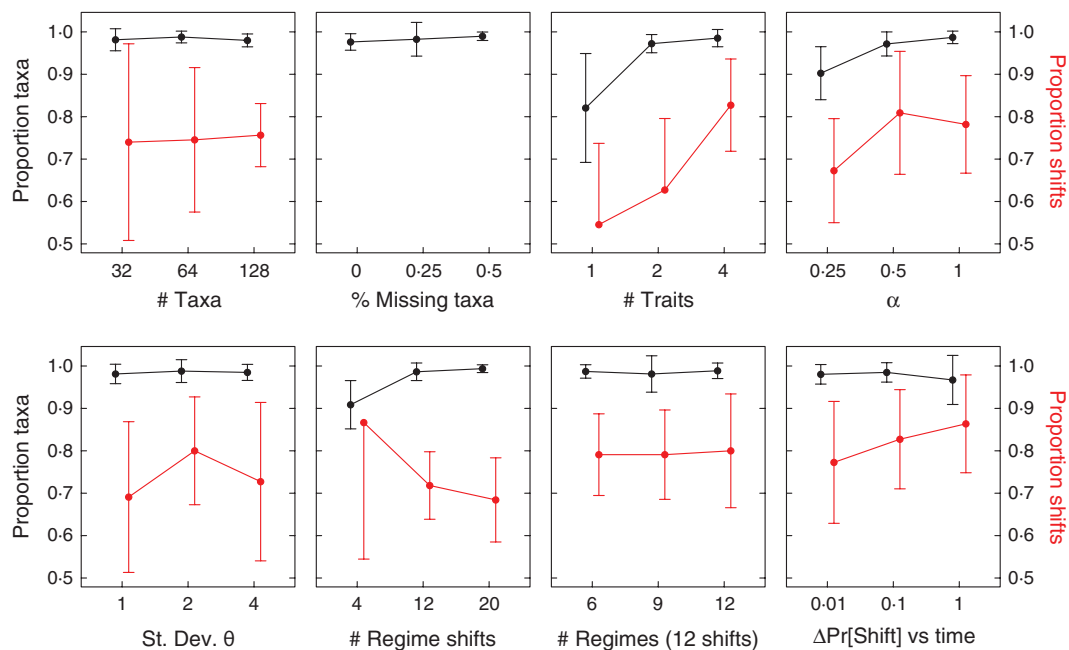


Fig. 3. Accuracy of SURFACE at recovering true Hansen models under a range of conditions. Black symbols correspond to the pairwise accuracy with which tip taxa are grouped into regimes (see text), and red symbols show the proportion of branches containing shifts in the generating model that are correctly inferred to have shifts. Error bars show standard deviation among 10 replicates (some are drawn one-sided to keep them within the panel). Note that when there are missing taxa pruned from the tree, the recovery of shift locations is not meaningful (as there is no 1 : 1 correspondence between many branches in the pruned and complete trees), and the pairwise recovery of regime assignments is only evaluated for the subset of taxa in the pruned tree.

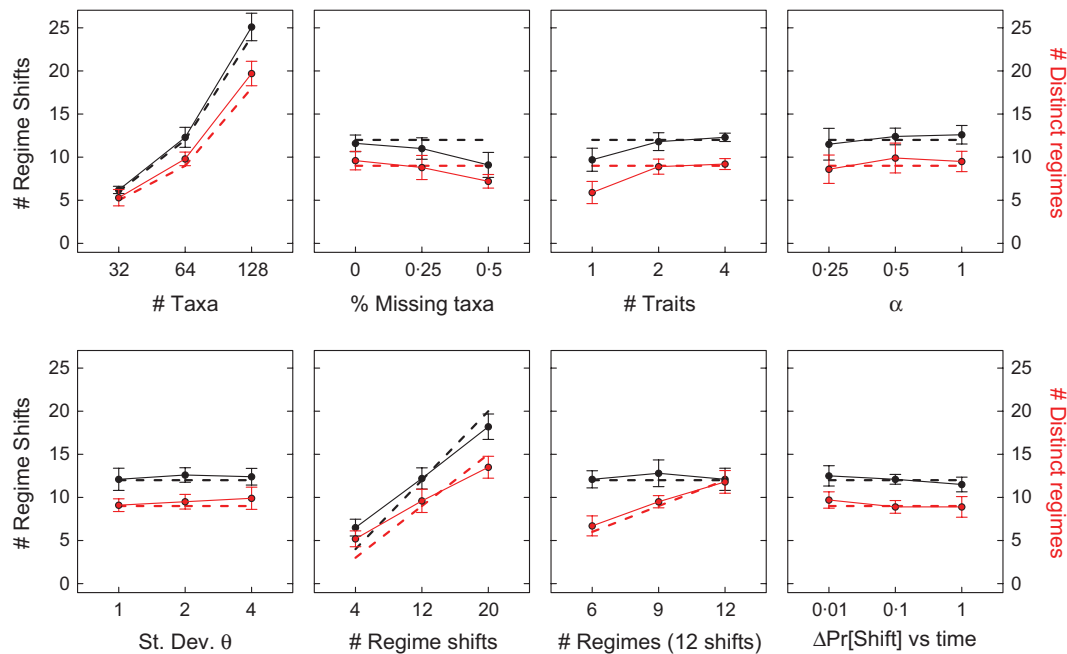


Fig. 4. Ability of SURFACE to recover the broad characteristics of a generating Hansen model: the number of regime shifts (black) and the number of distinct regimes (red). Error bars show means and standard deviations of 10 replicate simulations and dashed lines show true values. Parameter values and settings are as in Fig. 3.

Discussion

We have described a new method for inferring the macro-evolutionary adaptive landscape for a clade, allowing the assessment of phenotypic convergence given only a phylogenetic tree and continuous trait measurements. SURFACE fills a gap in the set of currently available phylogenetic comparative methods by combining features of two recent developments. First, recent applications of the OU model allow researchers to paint the branches of a tree with hypothesized selective regimes (Butler & King 2004; Beaulieu *et al.* 2012). This provides a powerful way to test whether taxa in similar environments have evolved similar phenotypes, but does not solve the issue of circularity if hypothetical regimes were identified in part based on the traits being modelled. The second development is methods for fitting shifts in evolutionary parameters to a tree without an *a priori* hypothesis about where the shifts should occur. The first of these methods, MEDUSA, uses stepwise AIC to locate shifts in speciation and extinction rates on a tree (Alfaro *et al.* 2009), and subsequent methods allow shifts in the Brownian rate of trait evolution σ^2 to be located using stepwise AIC (trait MEDUSA: Thomas & Freckleton 2012) or Bayesian Markov chain Monte Carlo methods (Eastman *et al.* 2011; Venditti, Meade & Pagel 2011; Revell *et al.* 2012). The method most similar to ours is MATICCE, which uses a model-averaging information theoretic approach to evaluate support for a number of candidate Hansen models (Hipp & Escudero 2010). The major differences are that SURFACE does not take candidate regime shift scenarios as inputs, and that it includes routines for evaluating whether regimes are convergent.

The main innovation of SURFACE is its backward phase, which assesses whether multiple regime shifts are towards the same regimes. Our simulations show that SURFACE performs fairly well at recovering the true convergent and nonconvergent regimes in simulated data sets under a range of conditions, particularly given multidimensional trait data and fast adaptation to new optima (Figs 3 and 4). In general, features that increase the degree to which taxa in the same regime are clustered in trait space should improve the performance of SURFACE. Greater trait dimensionality increases the likelihood of separation in at least one dimension, while more widely spaced optima, faster adaptation, or lower rates of stochastic evolution should lead to a greater signal of deterministic vs. stochastic evolutionary processes. We have also described how simulations can be used to test evolutionary hypotheses, such as whether the extent of convergence is greater than expected under a given null model. This test has good statistical power given multidimensional trait data and a moderate or high extent of convergence in the generating model (Fig. 5). We leave it to users to decide whether this null model approach is appropriate to test their hypotheses of interest, or if they prefer to make inferences strictly based on the model AIC_c and parameter values.

The ability to carry out data-driven tests for exceptional convergence presents an opportunity to re-evaluate clades that have previously been recognized as containing many cases of convergence. In this study, an analysis of morphological evolution in *Tetragnatha* spiders on Hawaii identified only a limited extent of convergence that occurred between subclades (Fig. 2), although it is important to note that we could not incorporate the discrete characters

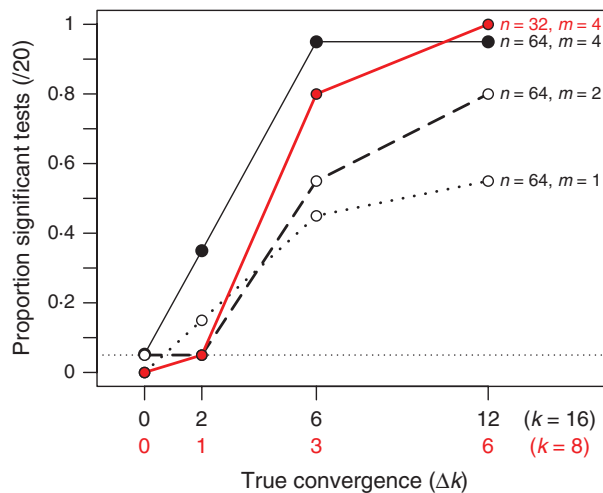


Fig. 5. Approximate statistical power and type I error of the simulation-based hypothesis test for unexpectedly high convergence, given different numbers of taxa (n) and traits (m) and different true levels of convergence (Δk). Each point shows the proportion of significant tests out of 20 'true' simulated data sets, using 50 resimulated data sets from a Hansen null model without convergence to generate a null distribution of Δk .

understood to be convergent within subclades (Gillespie, Croom & Hasty 1997; Gillespie 2004). SURFACE provides an opportunity to objectively evaluate the extent of convergence in other clades, including classic replicated radiations such as cichlids in Africa's Great Lakes (Kocher *et al.* 1993).

SURFACE may have several additional uses to researchers interested in a data-driven estimation of the adaptive landscape. For example, one may wish to test whether regime shifts are nonrandomly associated with biogeographical events, or are concentrated early or late in a clade's history, although it is important to remember that regime groupings of extant taxa and broad measures of convergence are recovered more reliably than precise positions of regime shifts (Figs 3 and 4). Other hypotheses may concern the inferred adaptive landscape, which could be compared to a landscape predicted on the basis of resource distributions and phenotype-resource use mapping (e.g. Schluter & Grant 1984). The methods described here can potentially be extended to allow simulation-based hypothesis tests tailored to a range of biologically motivated questions.

SURFACE infers a Hansen model without the need to paint selective regimes onto the tree in advance, but it may often be more appropriate to test specific adaptive hypotheses. This is particularly the case when the adaptive hypothesis was identified using traits other than those being fitted with the Hansen model. For example, one may test whether habitat use results in convergent evolution of morphology (Collar, Schulte & Losos 2011), or whether ecomorphs recognized by one set of traits (e.g. microhabitat and morphology) are also convergent in other traits (e.g. sexual dimorphism; Butler & King 2004). We do note that SURFACE should be robust to a 'trickle-down' effect that can mislead inference when hypothetical evolution-

ary shifts are placed on specific branches, whereby a shift on a phylogenetically nested branch may provide false support for a hypothesized shift on an earlier branch (Moore & Chan 2004; Revell *et al.* 2012).

The choice of traits will be an important component of any SURFACE analysis. First, as the method assumes that traits have independent rates of adaptation (α) and diffusion (σ^2), traits with strong evolutionary correlations should be avoided. SURFACE performs poorly when given only a single trait, but between two and four traits is often enough to ensure good performance (Figs 3 and 4), assuming these traits are in fact affected by the selective regime shifts. Including too many traits, especially axes lacking clear biological interpretation or unlikely to be involved in environmental adaptation, may limit the ability of SURFACE to recover convergence in ecomorphological traits. On the other hand, selection of only traits already believed to be convergent may predispose the analysis to finding a positive result. Researchers using SURFACE should ensure that the number and type of input traits are appropriate for addressing a given question in their clade of interest.

SURFACE uses stepwise AIC_c as a computationally tractable means of exploring the space of possible evolutionary scenarios (Alfaro *et al.* 2009; Thomas & Freckleton 2012). This stepwise approach has drawbacks: the constraint of adding one regime shift per step means the optimal configuration may not be found, and the answer can be sensitive to the choice of ΔAIC_c^* and the topology and branch lengths of the tree. While SURFACE can be run on multiple credible trees and can optionally allow stochasticity in the sequence of regimes added, there is still no guarantee of finding an optimal model or fully quantifying uncertainty. Future Bayesian methods may allow a more thorough exploration of model space and a better accounting for uncertainty in the placement and degree of convergence of regimes and in the phylogeny itself. In the meantime, SURFACE offers a valuable step forward in the application of comparative methods to test hypotheses about convergent evolution.

Many clades have long been understood to contain extensive convergence, but statistically appropriate methods for testing the extent of convergence have been lacking. SURFACE allows reassessment of such data sets, and can be used to test whether convergence is greater than expected by chance. As an objective tool for characterizing the macroevolutionary adaptive landscape of a clade, SURFACE provides many new opportunities to understand the dynamics of adaptive radiation.

Acknowledgements

This method was initially conceived during conversations with Luke Harmon and Chad Brock, and was improved by discussion with Jonathan Losos, Liam Revell, Brian Moore, and members of the Losos Lab. We thank Rosemary Gillespie for sharing data and discussing the *Tetragnatha* analysis, and Thomas Hansen and two anonymous reviewers for feedback that considerably improved this manuscript. Most simulations were run on the Odyssey cluster, supported by the FAS Science Division Research Computing Group at Harvard University. T.I. was funded by a National Science and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship, and D.L.M. was funded by a Center for Population Biology Postdoctoral Fellowship from the University of California at Davis.

References

- Alfaro, M.E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D.L., Carnevale, G. & Harmon, L.J. (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences of the USA*, **106**, 13410–13414.
- Arendt, J. & Reznick, D. (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology and Evolution*, **23**, 26–32.
- Bartoszek, K., Pienaar, J., Mostad, P., Andersson, S. & Hansen, T.F. (2012) A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology*, **314**, 204–215.
- Beaulieu, J.M., Jhwueng, D.C., Boettiger, C. & O'Meara, B.C. (2012) Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, **66**, 2369–2383.
- Beuttell, K. & Losos, J.B. (1999) Ecological morphology of Caribbean anoles. *Herpetological Monographs*, **13**, 1–28.
- Blackledge, T.A. & Gillespie, R.G. (2004) Convergent evolution of behavior in an adaptive radiation of Hawaiian web-building spiders. *Proceedings of the National Academy of Sciences of the USA*, **101**, 16228–16233.
- Boettiger, C., Coop, G. & Ralph, P. (2012) Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, **66**, 2240–2251.
- Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer Verlag, New York, NY.
- Butler, M.A. & King, A.A. (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist*, **164**, 683–695.
- Collar, D.C., Schulte, J.A. II & Losos, J.B. (2011) Evolution of extreme body size disparity in monitor lizards (*Varanus*). *Evolution*, **65**, 2664–2680.
- Csardi, G. & Nepusz, T. (2006) *The igraph software package for complex network research. InterJournal, Complex Systems*, 1695.
- Eastman, J.M., Alfaro, M.E., Joyce, P., Hipp, A.L. & Harmon, L.J. (2011) A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, **65**, 3578–3589.
- Felsenstein, J. (1988) Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, **19**, 445–471.
- Gillespie, R.G. (2004) Community assembly through adaptive radiation in Hawaiian spiders. *Science*, **303**, 356–359.
- Gillespie, R.G., Croom, H.B. & Hasty, G.L. (1997) Phylogenetic relationships and adaptive shifts among major clades of *Tetragnatha* spiders (Araneae: Tetragnathidae) in Hawai'i. *Pacific Science*, **51**, 380–394.
- Gillespie, R.G., Croom, H.B. & Palumbi, S.R. (1994) Multiple origins of a spider radiation in Hawaii. *Proceedings of the National Academy of Sciences of the USA*, **91**, 2290–2294.
- Hansen, T.F. (1997) Stabilizing selection and the comparative analysis of adaptation. *Evolution*, **51**, 1341–1351.
- Hansen, T.F. (2012) Adaptive landscapes and macroevolutionary dynamics. *The Adaptive Landscape in Evolutionary Biology* (eds E. Svensson & R. Calsbeek), pp. 205–226. Oxford University Press, Oxford, UK.
- Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E. & Challenger, W. (2008) GELGER: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.
- Harmon, L.J., Losos, J.B., Davies, T.J., Gillespie, R.G., Gittleman, J.L., Bryan Jennings, W., Kozak, K.H., McPeck, M.A., Moreno-Roark, F., Near, T.J., Purvis, A., Ricklefs, R.E., Schluter, D., Schulte, J.A. II, Seehausen, O., Sidlauskas, B.L., Torres-Carvajal, O., Weir, J.T. & Mooers, A.Ø. (2010) Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, **64**, 2385–2396.
- Harvey, P.H. & Pagel, M.D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK.
- Hipp, A.L. & Escudero, M. (2010) MATICCE: mapping transitions in continuous character evolution. *Bioinformatics*, **26**, 132–133.
- Kocher, T.D., Conroy, J.A., McKaye, K.R. & Stauffer, J.R. (1993) Similar morphologies of cichlid fish in Lakes Tanganyika and Malawi are due to convergence. *Molecular Phylogenetics and Evolution*, **2**, 158–165.
- Losos, J. (2009) *Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles*. University of California Press, Berkeley, CA.
- Losos, J.B. (2011) Convergence, adaptation, and constraint. *Evolution*, **65**, 1827–1840.
- Losos, J.B., Jackman, T.R., Larson, A., de Queiroz, K. & Rodríguez-Schettino, L. (1998) Contingency and determinism in replicated adaptive radiations of island lizards. *Science*, **279**, 2115–2118.
- Mahler, D.L., Revell, L.J., Glor, R.E. & Losos, J.B. (2010) Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. *Evolution*, **64**, 2731–2745.
- Moore, B.R. & Chan, K.M.A. (2004) Detecting diversification rate variation in supertrees. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (ed. O.R.P. Bininda-Emonds), pp. 487–533. Kluwer Academic, Dordrecht, The Netherlands.
- Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
- R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL: <http://www.R-project.org/>.
- Revell, L.J. (2009) Size-correction and principal components for interspecific comparative studies. *Evolution*, **63**, 3258–3268.
- Revell, L.J., Mahler, D.L., Peres-Neto, P.R. & Redelings, B.D. (2012) A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution*, **66**, 135–146.
- Schluter, D. (2000) *The Ecology of Adaptive Radiation*. Oxford University Press, Oxford, UK.
- Schluter, D. & Grant, P.R. (1984) Determinants of morphological patterns in communities of Darwin's finches. *American Naturalist*, **123**, 175–196.
- Simpson, G.G. (1953) *The Major Features of Evolution*. Columbia University Press, New York, NY.
- Stayton, C.T. (2008) Is convergence surprising? An examination of the frequency of convergence in simulated datasets. *Journal of Theoretical Biology*, **252**, 1–14.
- Thomas, G.H. & Freckleton, R.P. (2012) MOTMOT: models of trait macroevolution on trees. *Methods in Ecology and Evolution*, **3**, 145–151.
- Venditti, C., Meade, A. & Pagel, M. (2011) Multiple routes to mammalian diversity. *Nature*, **479**, 393–396.
- Williams, E. (1972) The origin of faunas. Evolution of lizard congeners in a complex island fauna: a trial analysis. *Evolutionary Biology*, **6**, 47–89.

Received 5 November 2012; accepted 18 January 2013

Handling Editor: Thomas Hansen