

# Computational methods to study phenotype evolution and feature selection techniques for biological data under evolutionary constraints

Inaugural-Dissertation

zur

Erlangung des Doktorgrades  
der Mathematisch-Naturwissenschaftlichen Fakultät  
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

**Christina Kratsch**

geb. Tusche, aus Eisenach

Düsseldorf, Januar 2014

---

# RidgeRace for ancestral character state reconstruction and inference of phenotypic rates

---

## 8.1 Introduction

Many biological studies are interested in the evolution of ancestral states of one or several discrete and continuous characters on a phylogenetic tree (Chapter 6 reviews current methods). Typical examples are the absence, presence or state of genes or traits, environmental preferences of different species, measures of morphology or physiology, or of behavioral or metabolic properties (a comprehensive collection of examples can be found in Nunn, 2011). Comparative methods aim to determine genetic markers correlating with each other, or with a specific phenotype, and thus often require the

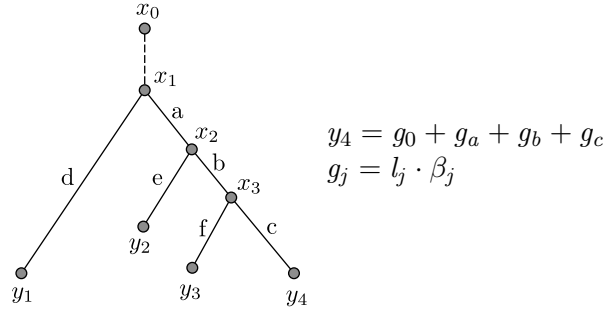


Figure 8.1: Model of phenotype evolution on a phylogenetic tree. The observed continuous character values at nodes  $y_i$  are the result of a sum of contributions on ancestral branches. A virtual branch “above” the root node  $x_1$  is contributing the global phylogenetic mean, i.e. the ancestral state of  $x_1$ .

reconstruction of ancestral values (Elliot, 2013). Such reconstructions are also of interest when fossil records cannot be retrieved, or when the phenotype of interest cannot be reconstructed from them, e.g. when studying environmental conditions for a particular species.

*RidgeRace* (Ridge Regression for Ancestral Character Estimation) is a method inspired by the least-squares optimization technique of Cavalli-Sforza and Edwards (1967). Similar to the concept of maximum parsimony, RidgeRace does not assume certain evolutionary rates at certain regions of the phylogeny, or a particular model of rate change over time. It treats phenotypic measurements at the terminal nodes of a phylogeny as sample observations and relies on a simple linear regression with  $L_2$ -Norm regularization, allowing phenotypic rates to vary at every branch. It estimates branch-wise rates and ancestral characters simultaneously, by inferring a regression model that describes the phenotypes observed at the terminal nodes best. As in the original BM model, we consider the leaf values to be the result of a weighted sum of intermediate contributions  $g_i$  created along the tree, beginning at the root (see Figure 8.1). The contributions represent the gain or loss in character value on each branch of the tree, with  $g_0$  holding a bias term representing the original contribution of the root node. The contribution  $g_j$  of a single branch  $j$  can be seen in analogy to the formulation of BM: the gain or loss in phenotype is dependent on the length  $l_j$

of branch  $j$  and of the speed  $\beta_j$  of the process. RidgeRace infers the  $\beta_j$  with ridge regression and is thus able to reconstruct the values of inner nodes. In a simulation study, we evaluate variations of Brownian Motion on randomly created trees and show that our method performs equally well or better than established implementations of two state-of-the-art reconstruction algorithms (Section 8.2). The branch weights  $\beta_j$  can be interpreted as phenotypic rates and provide insight into particularly interesting areas of the phylogeny (see Section 8.3). We believe that RidgeRace might be of use in studies aiming for reconstructions of ancestral character states of continuous characters when no definite assumptions can be made about the type of evolutionary process, or when the assumption of a model for phenotypic evolution is not appropriate at all. The latter might for example be the case in studies that rely only on a hierarchical clustering of samples instead of phylogenies. They can also be of use to judge the phenotypic impact of e.g. genetic changes or other types of events associated with branches of the phylogeny. When discrete data is provided in addition to the continuous phenotype and the underlying phylogeny, RidgeRace reconstructs genetic changes to the inner branches of the tree, and identifies those changes that occur on branches with a particularly high phenotypic change. We will demonstrate this with an example application for a cancer subtype stratification in Section 8.4. The last section of this chapter explains the technical details of RidgeRace and the preprocessing steps for the analyzed data.

## 8.2 Evaluation with simulated data

To evaluate the suitability of our method for ancestral character state reconstruction, we randomly created phylogenetic trees of increasing number of leaves. We evaluated two different settings, which we named the simple Brownian motion setting (BMS) and the extended setting using multiple regimes (ESMR). BMS refers to a standard simulation of Brownian Motion beginning at the root, creating ancestral values for each inner node. ESMR is an extension of BMS that divides the tree into  $\kappa$  regimes of Brownian Motion with different variation parameters. Three different methods were compared for

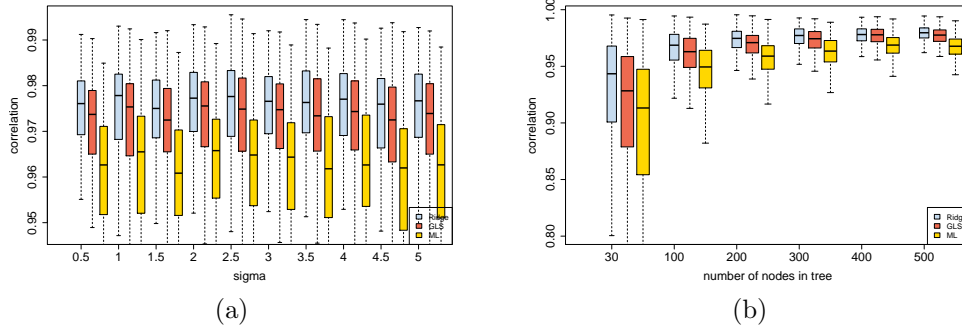


Figure 8.2: Pearson's correlation between inferred ancestral characters and true simulated values, when using maximum likelihood reconstruction (yellow), GLS (red), and Ridge-Race (blue). The plot shows (a) the dependence of performance on the standard deviation  $\sigma$  of the BM process or (b) when increasing the number of leaf nodes in the tree.

evaluation. We used the maximum likelihood method (“REML”) (Felsenstein, 1985) and the generalized least squares method (Martins and Hansen, 1997) provided by the function `ape::ace` as well as our ridge regression method. For each tree and each character assignment, we provided the tree structure and the leaf node assignment to the reconstruction method, which created a prediction for the assignment of inner nodes. To estimate the correctness of a reconstruction method, we computed Pearson's correlation between the predicted values and the true simulated values at those inner nodes. Our evaluation showed that the method performs comparative or up to 3 percent points better than other state-of-the-art techniques. For the BMS evaluation, Figure 8.2 shows that all three methods are able to reconstruct ancestral states very well, achieving correlation values between 85 and 98 percent, even for very small trees or very high variation values. However, RidgeRace shows consistently better correlation values than the two reference methods. Performance is independent of the variation parameter for all methods (Figure 8.2a), but does (not surprisingly) increase with the size of the tree (Figure 8.2b). A similar observation can be made for ESMR with variable rates. Similarly to the simple BMS, performance is independent of the size of the range from which standard deviations for the tree regimes are drawn, and again the correlation between predicted and true values increases with the number of nodes in

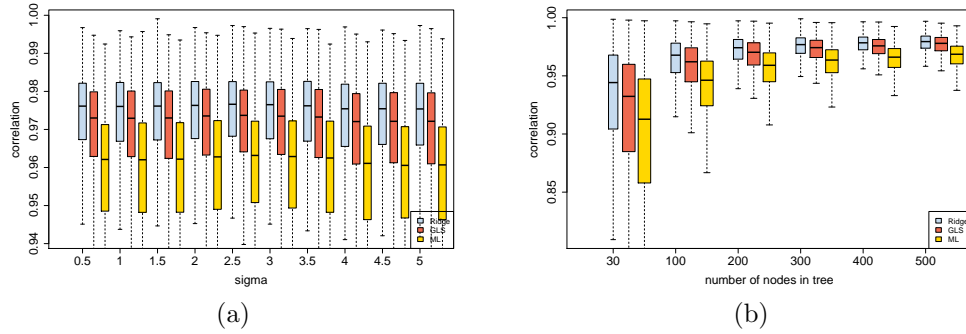


Figure 8.3: Pearson's correlation between inferred ancestral characters and true simulated values, with colors analogous to Figure 8.2. The plot shows (a) the dependence of performance on increasing the interval  $\mathcal{U}(0, s_G)$  from which the rates of the BM processes of each of the  $\kappa$  single regimes is drawn. Figure (b) shows performance when increasing the number of leaf nodes in the tree.

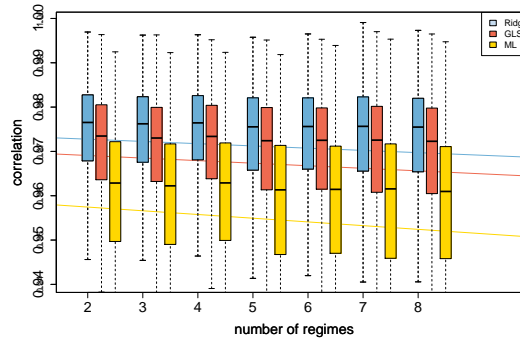


Figure 8.4: Pearson's correlation between inferred ancestral characters and true simulated values, with colors analogous to Figure 8.2. The plot shows the dependence of performance when increasing the number of regimes in the tree. Straight lines indicate a linear fit between the two variables.

the tree (see Figure 8.3). RidgeRace achieved correlation values consistently higher than the two other methods in all settings. When increasing the number of regimes, performance drops slightly for all three methods, with the slope of the linear fit being almost zero (see Figure 8.4, but still smallest (= slowest) for RidgeRace (RR:  $-5.06 \cdot 10^{-4}$ , GLS:  $-5.78 \cdot 10^{-4}$ , ML:  $-8.38 \cdot 10^{-4}$ , estimated using the R-function `stats::lm` (R Core Team, 2012)).

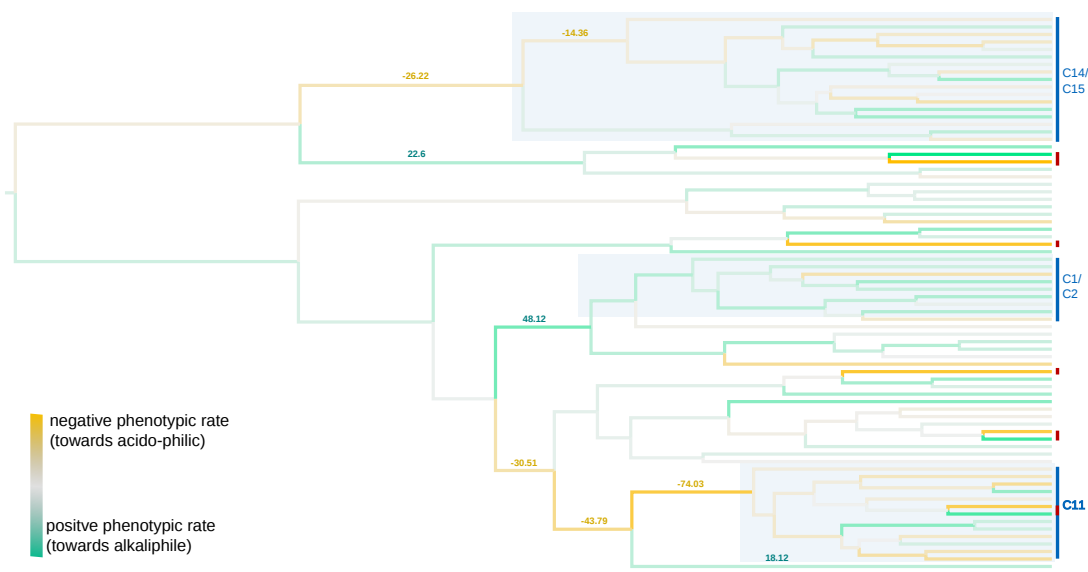


Figure 8.5: Visualization of inferred phenotypic rates (parameter  $\beta$ ) for a RidgeRace reconstruction of thaumarchaeota *amoA* sequences. Strong orange or green colors indicate high positive or negative rates. Rates are particularly high directly after speciation to one of the phenotypically most specialized clusters (indicated by blue bars, absolute rates indicated on branches), or at certain leaf nodes with strongly diverging phenotype (indicated by red bars).

### 8.3 Exemplary application on a thaumarchaeota data set

One advantage of RidgeRace over other methods is that it allows the phenotypic rate  $\beta_i$  to change on every branch of the tree. Typically, the phenotypic rate is assumed to be constant. In more complex models, phenotypic rates may vary, but have to be known in advance to reconstruct ancestral character states<sup>1</sup> (Harmon *et al.* (2010) offers a review of such models). Other techniques test for changes in the phenotypic rate in predefined regimes of the tree (McPeck, 1995; O’Meara *et al.*, 2006; Revell, 2008), but do not reconstruct ancestral character states beside the root state, and again require specific a priori assumptions on the locations of the regimes.

In a recent study on the ecology and evolutionary history of terrestrial thaumar-

<sup>1</sup>In fact, the GLSR approach is flexible enough to re-define the correlation structure provided by the phylogenetic tree in a completely arbitrary manner, allowing the inclusion of all kinds of branch transformations. However, it still requires the user to know the degree and position of the deviations from the basic model.

chaeta (Gubry-Rangin *et al.*, 2014, *in prep.*), we have analyzed the influence of the pH value of soil as an environmental factor that controls the adaptation of a specific lineage of archaea capable of ammonia oxidation (see Gubry-Rangin and Hai (2011) and Nicol *et al.* (2008) for details on the ecology and terrestrial distribution of thaumarchaeota). An additional RidgeRace analysis of the pH preferences of thaumarchaeota samples using a phylogeny inferred on *amoA* gene sequences (Figure 8.5) reconstructed the pH value of the root of the tree, i.e. the common ancestor of all thaumarchaeota, to 6.18, a value very similar to the reconstruction of 6.3 under a Brownian Motion model performed by the authors using the *ape* package in R (Paradis *et al.*, 2004; R Core Team, 2012)<sup>2</sup>. It also revealed that the phenotype (pH preference) has often evolved quicker on ancestral than on more recent branches of the tree, and in particular on branches directly after the separation of certain highly specialized pH clusters, such as the three main abundant clusters of terrestrial thaumarchaeota (marked by blue bars in Figure 8.5). This might indicate a particularly high speed of molecular adaptation. RidgeRace assigns similarly high rates to a few samples with a pH preference that strongly deviates from the mean of their clade (marked by red bars). This can be considered an artifact of the method, but it also used as an indicator of phenotypic outliers.

## 8.4 Example application to cancer data

### Cancer as a disease of evolution

According to the World Health Organization, cancer is a leading cause of disease-related deaths worldwide, responsible for 7.6 million deaths in the year 2008 (WHO, 2013a). The term “cancer” describes a variety of different diseases that may affect any part of the human body, with lung, stomach, liver, colon, and breast cancer being responsible for most of the cancer-associated deaths. The causes for cancer are not entirely described yet, but certain behavioral factors strongly contribute to the disease risk: about 30% of

<sup>2</sup>We performed Maximum Likelihood ratio tests to confirm that BM is indeed the most suitable model and that no signal of evolutionary trend is present in the data.



cancer deaths are related to high body mass index, low fruit and vegetable intake, lack of physical activity, and tobacco or alcohol consumption (WHO, 2013a).

Cancer is initiated by transformations in single cells that are caused by external factors such as physical, chemical, or biological carcinogens, or by a deficiency of cellular repair mechanisms (e.g. due to high age). The disease is the result of a complex interplay of genetic preconditions, external influences and interaction with the immune system. The main genetic factors (*hallmarks*) underlying the disease are summarized in two seminal papers by Hanahan and Weinberg (2000, 2011), which can be considered two of the most influential publications in the field (but see also Lazebnik 2010). At the time of creation of this thesis, the earlier paper has been cited more than 17,000 times<sup>3</sup>. For a wide variety of cancer types, recent studies identified genes that are significantly associated with cancer risk, onset, and progression (Kandoth *et al.*, 2013; Röhr *et al.*, 2013; TCGAN, 2008, 2012a,b, 2013).

Tumors are now considered as a heterogeneous population of cells that are the result of a shared process of evolution (Nowell, 1976). The author stated the concepts of clonal expansion and discussed the interplay between cancer therapy and the evolution of tumor cell subpopulations. By now, a large number of observations have provided evidence for this theory. Several studies confirmed that the degree of genetic diversity in a tumor cell population is a good predictor for malignancy (Maley *et al.*, 2006). In 2006, Merlo *et al.* suggested to include the genetic instability preceding this diversity as an additional hallmark of cancer, and the concept was included as an “enabling characteristic” in the 2011 update by Hanahan and Weinberg. There exists a multitude of known cancer diseases with a highly varying pathology and a similarly varying heterogeneity of involved pathways. However, there might also be a strong difference in tumor histologies between patients suffering from the same subtype (Yates and Campbell, 2012). Even within a single tumor tissue, sub-tissues show differing copy number profiles (Podlaha *et al.*, 2012).

---

<sup>3</sup>according to Google Scholar, accessed 09/09/2013.

### RidgeRace for integrating cancer study data

To demonstrate a possible application of RidgeRace integrating phenotypic and genotypic data, we studied an ovarian cancer data set, created by the TCGA research network, and recently analyzed with Network Based Stratification (Hofree *et al.*, 2013). Hofree *et al.* argue that somatic mutations are likely to contain the causal drivers of tumor progression, and that this type of data provides a promising source of information to identify clinically relevant sub-clusters (stratifications). The authors note that tumors are very heterogeneous, and genetic profiles are sparse and vary strongly between patients, making clustering and stratification a challenging task. Network based stratification is a new clustering method that smooths genetic profiles with the help of gene interaction networks, and Hofree *et al.* show that it produces clinically meaningful clusterings. We use a data set and the software provided by the authors (NBS, version 0.2, available from the authors website) to reconstruct a hierarchical clustering on somatic mutation data of ovarian cancer samples, creating a tree structure (Figure 8.6a).

Although it was not possible to check if our inferred clustering is completely identical to the one discussed by Hofree *et al.*, we similarly find that patients assigned to the smallest of the four subtypes show an increased survival time (Figure 8.6b, green cluster). A RidgeRace analysis of patient survival time as a phenotype consistently showed a strong positive increase in rate at the branch leading to that cluster (Figure 8.6a, marker  $m1$ ). Similarly, RidgeRace infers a decrease in survival time on the branch leading to the yellow cluster (marker  $m2$ ). Marker  $m3$  shows a rather small decrease in survival time, because the red cluster splits in distinct two subtypes with a successive second increase ( $m5$ ) or a decrease ( $m4$ ) in survival time, with branch ( $m4$ ) leading to the majority of the red cluster, which has the lowest survival time of all four clusters.

As suggested above, the RidgeRace reconstruction can be combined with the reconstruction of discrete genetic events. We mapped the binary data encoding the absence or presence of non-synonymous mutations in a selection of genes to the tree (see Section 8.5 for technical details). However, the mapping confirmed the diverse nature of the somatic mutations. Only *P53* was found to be mutated in almost all patients,

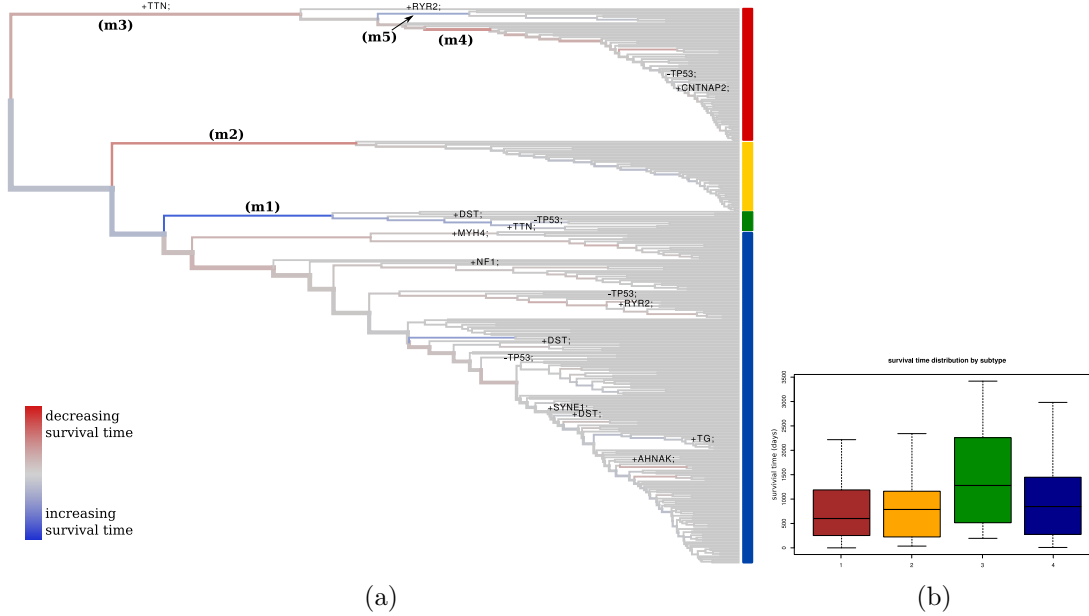


Figure 8.6: RidgeRace application to a clustering on somatic mutations inferred for an ovarian cancer data set. Colors on the side of the tree indicate subtypes inferred with Network Based Stratification (Hofree *et al.*, 2013). Branches are colored according to the phenotypic rate parameter  $\beta$ , thickness of branches is proportional to the number of nodes below them. Branches leading to leaf nodes were colored grey for improved visibility. Markers  $m1$  to  $m5$  indicate branches with strong changes in patient survival time. Changes in the absence or presence of mutations in selected genes are indicated on all branches with 4 or more children.

and was reconstructed to be mutated at the root of the tree. Beside *P53*, only *TTN* was reconstructed to appear on a higher level node, it is “gained” (mutated) at branch  $m3$ , and present in 83 of 85 patients of the red cluster. *RYR2* is gained on branch  $m5$  and present in 9 out of 85 patients of the red cluster. Besides these changes, no change appears on a branch higher than five levels below the root.

It is obvious that, although RidgeRace is able to correctly display the main structure of the phenotype, no significant association between genetic aberrations and change in survival rate is possible in this case. However, this demonstration provides many insights for future improvements:

- The survival rate as a phenotype might be a very biased measurement, since it is based on the time of diagnosis and the (potential) death of the patient. A

very late onset of therapy or simply a survival of the patient might influence this measurement.

- The phylogeny itself may be a source of bias: although Hofree *et al.* 2013 convincingly argue that their method produces clinically meaningful subclusters, the technique is based on a large number of parameters, among them the final number of main clusters, and the underlying gene interaction network (we used the parameters inferred by the authors). Hofree *et al.* suggest that future improvements of NBS may consider other alterations than non-synonymous mutations, or consider the length of genes.
- The subtypes identified by NBS may indeed represent the correct *genetic* stratification of the patients, nevertheless, their survival time may be dependent on many other factors, e.g. patient age and type of received therapy. Since RidgeRace is essentially a regression patient data, such information can easily be included as additional covariates, giving insight into their relevance relative to genetic factors.

## 8.5 Technical details of the method

### RidgeRace weight inference

RidgeRace is primarily intended as a method to estimate ancestral character states on a phylogenetic tree. As in the original BM model, we consider the leaf values to be the result of a weighted sum of intermediate contributions  $g_i$  created along the tree, beginning at the root (see Figure 8.1). The contributions represent the gain or loss in character value on each branch of the tree, so that, for example, the character value of sample  $y_4$  can be described as

$$y_4 = g_0 + g_a + g_b + g_c,$$

where  $a, b, c$  represent the branches in the tree, and  $g_0$  holds a bias term representing the original contribution of the root node. The contribution  $g_j$  of a single branch  $j$  can

be seen in analogy to the formulation of BM: the gain or loss in phenotype is dependent on the length  $l_j$  of branch  $j$  and of the speed  $\beta_j$  of the process, in analogy to the variance term  $\sigma^2$  in the BM model:

$$g_j = l_j \cdot \beta_j.$$

One can then write the solution for the vector of leaf phenotypes  $\mathbf{Y}$  in matrix form:

$$\hat{\mathbf{Y}} = \mathbf{L}\beta, \tag{8.1}$$

where

$$\mathbf{L}_{i,j} = \begin{cases} l_j & \text{if branch } j \text{ is on the way from the root to sample } i \\ 1 & \text{if } j = 0 \\ 0 & \text{else} \end{cases}$$

and  $\beta$  is a vector of length equal to the number of branches in the phylogeny, including a single virtual branch above the root to account for its original contribution  $g_0$ . Note that this scheme allows an easy inclusion of measurements at inner nodes, e.g. from fossil records. It is also suited to account for multiple measurements at single nodes simply by adding additional rows to  $\mathbf{Y}$  and  $\mathbf{L}$ .

Ridge regression then estimates a vector  $\hat{\beta}$  that explains the known observations  $\mathbf{Y}$  best:

$$\hat{\beta} = \arg \min_{\beta} \sum_i (y_i - (\mathbf{L}\beta)_i)^2 + \lambda \sum_j \beta_j^2, \tag{8.2}$$

and the text book solution (see, e.g. [Hastie \*et al.\*, 2009](#)) to this optimization problem is

$$\hat{\beta} = (\mathbf{L}^T \mathbf{L} + \lambda \mathbf{I})^{-1} \mathbf{L}^T \mathbf{Y}.$$

Equation 8.2 shows how the optimization tries to balance the leaf reconstruction error versus a regularization term. This second term forces the  $\beta_j$  and therefore the gains  $g_j$  to be distributed evenly across the tree, towards a globally constant rate (see [Hastie \*et al.\* \(2009\)](#) for a discussion on the influence of regularization terms). Without this

term, a trivial but undesirable solution to the optimization would set the gain at each terminal branch equal to the according terminal node value, leaving all other gains empty and making ancestral reconstruction impossible.

For a given estimate of  $\hat{\beta}$ , the vector  $\hat{\mathbf{A}}$  containing the phenotypic reconstruction of all inner nodes can then be computed analogous to equation 8.1:

$$\hat{\mathbf{A}} = \mathbf{L}'\hat{\beta}, \quad (8.3)$$

where

$$\mathbf{L}'_{i,j} = \begin{cases} l_j & \text{if branch } j \text{ is on the way from the root to ancestor } i \\ 1 & \text{if } j = 0 \\ 0 & \text{else} \end{cases}$$

Note that this formulation is very similar to the generalized least squares method proposed by [Martins and Hansen \(1997\)](#). They similarly suggest to infer ancestral character states as weighted average of leaf contributions, with weights according to the covariance between an ancestor and a leaf (see equ. (10) in [Martins and Hansen \(1997\)](#), and [Cunningham et al. \(1998\)](#) for a worked example):

$$\hat{\mathbf{A}} = \mathbf{W}\mathbf{Y} + \epsilon \quad (8.4)$$

$$\mathbf{W} = cov[\mathbf{A}, \mathbf{Y}] var[\mathbf{Y}]^{-1}, \quad (8.5)$$

where the covariance between inner node  $a$  and a leaf node  $y$  is defined as  $\sigma^2 T_{a,y}$ , with  $T_{a,y}$  being the distance between the root of the tree and the most recent common ancestor of  $a$  and  $y$  (see Figure 7.4 for an example). RidgeRace differs in the sense that it allows to estimate a weight  $\beta_j$  for every branch instead of assuming a constant rate  $\sigma^2$ , or, more general, predefined covariances between nodes. Extensions of the simple GLS approach under the Brownian motion model use more complex matrices  $\mathbf{W}$ . However, the design of  $\mathbf{W}$  has to be defined in advance based on specific model assumptions, whereas RidgeRace offers to estimate rates independently.

## Simulation study

We created random trees with an increasing number  $N$  of leafs using the function `rtree` in the R-package `textttape` (Paradis *et al.*, 2004; R Core Team, 2012). For a first evaluation (BMS), we simulated Brownian motion with variation  $\sigma^2$  along the branches of the tree, resulting in a character assignment for every inner or leaf node. The parameter  $\sigma^2$  was iteratively increased in each round of simulation. In the extended setting (ESMR), we simulated changing rates of evolution in the tree by dividing the tree into  $\kappa$  different regimes, and in every regime  $r_i$ , a new rate  $\sigma_i^2$  was drawn at random from a global uniform distribution in the interval  $\mathcal{U}(0, s_G)$ . To make the process increasingly more variable and difficult, the size of that interval  $s_G$  was iteratively increased with every simulation, thus allowing larger  $\sigma_i$  to be drawn. The simulation of Brownian motion in each regime was performed using the respective  $\sigma_i^2$ , resulting in a character assignment for all nodes. This process was repeated several times and for different parameters  $\sigma^2, s_G, \kappa$  and  $N$ . See supplementary text D and supplementary Figure D.1 and D.2 for details on the simulation algorithms and parameter settings. The random tree and the simulated values obtained at the leaf nodes were provided as input to RidgeRace, and an implementation of the ML and GLSR algorithms in the `ape` function for ancestral character state estimation (`ace`). Obtained reconstructed values were mapped back to the inner nodes of the tree and compared with the simulated ones using Pearson's correlation coefficient (leaf nodes were excluded).

## Analysis of thaumarchaeota data

The pH preferences of 425 thaumarchaeota samples of 16 subtypes and a phylogeny based on their *amoA* genes was obtained from a collaboration project (Gubry-Rangin *et al.*, 2014, *in prep.*). PH preferences were mapped to the leafs of the phylogeny and ancestral values were reconstructed with RidgeRace using  $\lambda = 10^{-5}$ . Inferred phenotypic rates  $\beta_j$  were visualized using FigTree (Rambaut, 2013). In the collaboration project, we used the given phylogeny for maximum likelihood ratio tests to infer the most suitable model of continuous phenotype evolution. Among others, we compared

Brownian Motion to models containing Ornstein-Uhlenbeck processes with different numbers of regimes, an early burst and a trend model. The ML ratio tests inferred BM to be the best fitting model.

### Preprocessing of cancer data

A binary matrix describing the absence or presence of non-synonymous mutations in 9850 genes for 325 patients was taken from the supplementary data provided by Hofree *et al.* (2013). As indicated by the authors in their article and supplementary material, we used their software with 4 clusters and the HM network, and default parameters, creating 1000 bootstrap samples. We then inferred a hierarchical clustering (average linkage) on the bootstrap similarity matrix with NBS methods and used the inferred topology as input tree for RidgeRace. Information on patients survival rate was downloaded from the TCGA database

(TCGAN, 2011). Phenotypic rates were inferred with RidgeRace as described above. The binary genetic profile of each patient was mapped to the leaf nodes and reconstructed to inner nodes with the Sankoff algorithm implemented in RidgeRace, using a simple 0/1 cost matrix and the ACCTRAN principle in case of ambiguities. “Mutations”, i.e. changes in genetic profile, were then reconstructed on the branches of the tree. Finally, the tree was visualized using FigTree (Rambaut, 2013).