Correlated Evolution in the Small Parsimony Framework

Brendan Smith[1], Rebecca Buonopane[1],
Cristian Navarro-Martinez[1], S. Ashley Byun[1]
and Murray Patterson[2]

[1]Fairfield University, Fairfield CT 06824, USA
[2]Georgia State University, Atlanta GA 30303, USA

Abstract

When studying the evolutionary relationship between a set of species, the principle of parsimony states that a relationship involving the fewest number of evolutionary events is likely the correct one. Due to its simplicity, this principle was formalized in the context of computational evolutionary biology decades ago by, *e.g.*, Fitch and Sankoff. Because the parsimony framework does not require a model of evolution, unlike maximum likelihood or Bayesian approaches, it is often a good starting point when no reasonable estimate of such a model is available.

In this work, we devise a method for detecting correlated evolution among pairs of discrete characters, given a set of species on these characters, and an evolutionary tree. The first step of this method is to use Sankoff's algorithm to compute *all* most parsimonious assignments of ancestral states (of each character) to the internal nodes of the phylogeny. Correlation between a pair of evolutionary events (*e.g.*, absent to present) for a pair of characters is then determined by their (co-) occurrence patterns among their respective ancestral assignments. We implement this method: `parcours` (PARsimonious CO-occURrenceS) and use it to study the correlated evolution among vocalizations in the Felidae family, revealing some interesting results.

The `parcours` tool is freely available at `https://github.com/murraypatterson/parcours`

keywords: Parsimony · Evolution · Correlation

1 Introduction

The principle of parsimony first appeared in the context of computational evolutionary biology in [7, 3, 10, 14]. A few years later, Sankoff and Rousseau generalized this to allow the association of a (different) cost to each transition between a pair of states [35, 36]. A few years after that, Felsenstein noticed that parsimony could produce misleading results when the evolutionary rate of change on the branches of the phylogeny is high [12]. Because parsimony does not take branch length into account, it ignores the fact that many changes on a long branch — while being far from parsimonious — may not be so unlikely in this case, resulting in a bias coined as "long branch attraction". This paved the way for a proposed refinement to parsimony known as the maximum likelihood method [12, 13]. Because of an explosion in molecular sequencing data, and the sophisticated understanding of the evolutionary rate of change in this setting, maximum likelihood has become the prevalent framework for inferring phylogeny. Many popular software tools which implement maximum likelihood include RAxML [39], PHYLIP [11] and the Bio++ library [15]. Even more recently, some Bayesian approaches have appeared, which sample the space of likely trees using the Markov-Chain Monte Carlo (MCMC) method, resulting in tools such as MrBayes [17] and Beast [1].

However, in cases where there is no model of evolution, or no reasonable estimation of the rate of evolutionary change on the branches of the phylogeny, parsimony is a good starting point. In fact, even in the presence of such a model, there are still conditions under which a maximum likelihood phylogeny is always a maximum parsimony phylogeny [42, 40]. Finally, when the evolutionary rate of change of different characters is heterogeneous, *i.e.*, there are rates of change, maximum parsimony has even been shown to perform *substantially better* than maximum likelihood and Bayesian approaches [21]. A perfect example is cancer phylogenetics [16, 37, 9, 38], where little is known about the modes of evolution of cancer cells in a tumor. Due to its importance, this setting has seen a renewed interest in high-performance methods for computing maximum parsimonies in practice, such as SCITE [19] and SASC [33].

The setting of the present work is an evolutionary study of vocalizations among members of the family Felidae. Felids possess a range of intraspecific vocalizations for close, medium, and long range communication. These vocalizations can vary from discrete calls such as the spit or hiss to more graded calls such as the mew, main call, growl and snarl [41]. There is no evidence that Felid vocalizations are learned: it is more likely that these

calls are genetically determined [28, 8, 34]. While there are 14 major discrete and graded calls documented in Felidae, not all calls are produced by all species [30]. In [31], the authors map some of these calls to a molecular phylogeny of Felidae, to show that it was consistent with what was previously known, strengthening the argument that vocalizations are genetically determined. Here we consider a similar approach, but because (a) an obvious model of evolution is lacking in this case; (b) the possibility that vocalizations within a given group of species can evolve at considerably different rates [31]; and (c) that rates for specific characters can differ between different lineages within that group [31], it follows from [21] that parsimony is more appropriate than maximum likelihood or a Bayesian approaches.

In this work, we develop a general framework to determine correlated (or co-) evolution among pairs of characters in a phylogeny, when no model of evolution is present. We then use the framework to understand how these vocalizations may have evolved within Felidae, and how they might correlate (or have co-evolved) with each other. The first step of this approach is to infer *all* most parsimonious ancestral states (small parsimonies, of these Felid calls) in the phylogeny. Correlation between a pair of evolutionary events (for the respective pair of characters) is then determined by the ratio of the number of such events that *co-occur* on a branch of the phylogeny, and the number of independent occurrences of each event on the branches, over all pairs of most parsimonious ancestral reconstructions. We implement this approach in a tool called `parcours`, and use it to detect correlated evolution among 14 Felid vocalizations, obtaining both results that are consistent with the literature [27, 31] and some interesting new findings. While various methods for detecting correlated evolution exist, they tend to use only *phylogenetic profiles* [22], or are based on maximum likelihood [23, 5, 25], where a model of evolution is needed. Methods that determine co-evolution in the parsimony framework exist as well [26, 6], however they are aimed at reconstructing ancestral gene *adjacencies*, given extant co-localization information. Finally, while some of the maximum likelihood software tools have a "parsimony mode", *e.g.*, [11, 15], the character information must be encoded using restrictive alphabets, and there is no easy or automatic way to compute *all* most parsimonious ancestral states for a character. On the contrary, `parcours` takes character information as a *column-separated values* (CSV) file, inferring the alphabet from this input. In summary, our contribution is a methodology for inferring correlation among pairs of characters when no model of evolution is available, and implement it into a software that is so easy to use that it could also be used as a pedagogical tool.

This paper is structured as follows. In Sec. 2, we provide the background

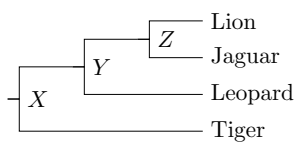
on computing *all* most parsimonious ancestral states in a phylogeny, given a set of extant states. In Sec. 3, we present our approach for computing correlation between pairs of characters from all such parsimonies. In Sec. 4, we describe an experimental analysis of the implementation, `parcours`, of our method on Felid vocalizations. Finally, in Sec. 5, we conclude the paper with a discussion of the results obtained, and future directions.

2 Small Parsimony

In computing a parsimony, the input is usually *character-based*: involving a set of species, each over a set of *characters* (*e.g.*, vocalizations), where each character can be in a number of *states* (*e.g.*, absent, present, etc.). The idea is that each species is in one of these states for each character — *e.g.*, in the Lion, the puff (one of the 14 vocalizations) is present — and we want to understand what states the ancestors of these species could be in. Given a *phylogeny* on these species: a tree where all of these species are leaves, a *small parsimony* is an assignment of states to the internal (ancestral) nodes of this tree which minimizes the number of changes of state among the characters along branches of the tree[1]. We illustrate this with the following example.

Suppose we have the four species: Lion, Jaguar, Leopard and Tiger, alongside the phylogeny depicted in Fig. 1a, implying the existence of the ancestors X, Y and Z. We are given some character puff which can be in one of a number of states taken from *alphabet* $\Sigma = \{\text{absent, present, unknown}\}$, which we code here as $\{0, 1, 2\}$ for compactness. If puff is present only in Lion and Leopard as in Fig. 1b, then the assignment of 0 (absent) to all ancestors would be a small parsimony. This parsimony has two changes of state: a change $0 \to 1$ on the branches $Z \to$ Lion and $Y \to$ Leopard — implying convergent evolution of puff in these species. Another small parsimony is the presence of puff in X and Y (absence in Z) — implying that puff was ancestral (in Y) to the Lion, Jaguar and Leopard, and was later lost in the Jaguar. A principled way to infer all small parsimonies is to use Sankoff's algorithm [35], which also makes use of a *cost matrix* δ, depicted in Fig. 1c, that encodes the evolutionary cost of each state change along a branch. Since the change $0 \to 1$ (absent to present) costs $\delta_{0,1} = 1$, and vice versa ($\delta_{1,0} = 1$), it follows that each of the small parsimonies mentioned above have an overall cost, or *score* of 2 in this framework. A

[1] A *large parsimony* is then a phylogeny which results in a small parsimony with the fewest changes (and this small parsimony)



(a)

sp. \ ch.	puff
Lion	1
Jaguar	0
Leopard	1
Tiger	0

(b)

δ	0	1	2
0	0	1	1
1	1	0	1
2	∞	∞	0

(c)

Figure 1: A (a) phylogeny and (b) the extant state of character puff in four species; and (c) the cost $\delta_{i,j}$ of the change $i \to j$ from state i in the parent to state j in the child along a branch of the phylogeny, *e.g.*, $\delta_{1,2} = 1$ ($\delta_{2,1} = \infty$).

simple inspection of possibilities shows that 2 is the minimum score of any assignment of ancestral states.

Sankoff's algorithm [35] computes all (small) parsimonies given a phylogeny and the extant states of a character in a set of species, and a cost matrix δ, *e.g.*, Fig. 1. The algorithm has a bottom-up phase (from the leaves to the root of the phylogeny), and then a top-down phase. The first (bottom-up) phase is to compute $s_i(u)$: the minimum score of an assignment of ancestral states in the subtree of the phylogeny rooted at node u when u has state $i \in \Sigma$, according to the recurrence:

$$s_i(u) = \sum_{v \in C} \min_{j \in \Sigma} \{s_j(v) + \delta_{i,j}\} \tag{1}$$

where C is the set of children of node u in the phylogeny. The idea is that the score is known for any extant species (a leaf node in the phylogeny), and is coded as $s_i(u) = 0$ if species u has state i, and ∞ otherwise. The score for the internal nodes is then computed according to Eq. 1 in a bottom-up dynamic programming fashion, starting from the leaves. The boxes in Fig. 2 depict the result of this first phase on the instance of Fig. 1, for example:

$$\begin{aligned} s_1(Y) = &\ \min\ \{s_0(\text{Leopard}) + \delta_{1,0},\ s_1(\text{Leopard}) + \delta_{1,1}, \\ &\qquad s_2(\text{Leopard}) + \delta_{1,2}\} \\ &+ \min\ \{s_0(Z) + \delta_{1,0},\ s_1(Z) + \delta_{1,1},\ s_2(Z) + \delta_{1,2}\} \\ = &\ \min\ \{\infty + 1,\ 0 + 0,\ \infty + 1\} + \min\ \{1 + 1,\ 1 + 0,\ \infty + 1\} \\ = &\ \min\ \{\infty,\ 0,\ \infty\} + \min\ \{2,\ 1,\ \infty\} = 0 + 1 = 1 \end{aligned} \tag{2}$$

$$\begin{aligned} s_0(X) = &\ \min\ \{s_0(\text{Tiger}) + \delta_{0,0},\ s_1(\text{Tiger}) + \delta_{0,1},\ s_2(\text{Tiger}) + \delta_{0,2}\} \\ &+ \min\ \{s_0(Y) + \delta_{0,0},\ s_1(Y) + \delta_{0,1},\ s_2(Y) + \delta_{0,2}\} \\ = &\ \min\ \{0 + 0,\ \infty + 1,\ \infty + 1\} + \min\ \{2 + 0,\ 1 + 1,\ \infty + 1\} \\ = &\ \min\ \{0,\ \infty,\ \infty\} + \min\ \{2,\ 2,\ \infty\} = 0 + 2 = 2 \end{aligned} \tag{3}$$

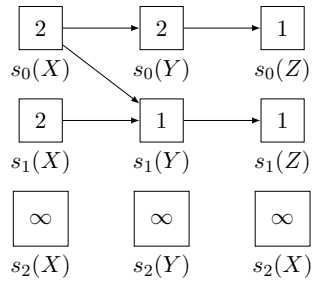


Figure 2: Structure resulting from Sankoff's algorithm on the instance of Fig. 1. The *boxes* contain the values of $s_i(u)$ computed in the bottom-up phase, for each state $i \in \Sigma$ and internal node $u \in \{X, Y, Z\}$ of the phylogeny. The *arrows* connecting these boxes were computed in the top-down phase.

In general, for a given character with $|\Sigma| = k$, this procedure would take time $O(nk)$, where n is the number of species, since we compute nk values, and computing each one takes (amortized) constant time.

After the bottom-up phase, we know the minimum score of any assignment of ancestral states (it is the minimum of those s_i, for $i \in \Sigma$, at the root), but we do not yet have an ancestral assignment of states. Here, since X is the root in Fig. 1a, we see from Fig. 2 that the minimum score for this example is 2, as we saw earlier. Note that this minimum may not be unique; indeed, here $s_0(X) = s_1(X) = 2$, meaning that in at least one parsimony, puff is absent (0) in X, and in at least one other parsimony, puff is present (1) in X (but never unknown). Now, to reconstruct one of these ancestral assignments of minimum score — the top-down phase of Sankoff's algorithm — we first assign to the root, one of the states $i \in \Sigma$ for which s_i is minimum. We then determine those states in each child of the root from which s_i can be derived (these may not be unique either), and assign those states accordingly. We continue, recursively, in a top-down fashion until we reach the leaves of the tree, having assigned all states at this point. For example, $s_0(X) = 2$, and can be derived in Eq. 3 from $s_1(Y)$ (and $s_0(\text{Tiger})$), which is in turn derived in Eq. 2 from $s_1(Z)$. This corresponds to the second parsimony mentioned earlier, where the puff was ancestral in Y, and later lost in the Jaguar. Notice that $s_0(X)$ can also be derived in Eq. 3 from $s_0(Y)$, which is in turn derived from $s_0(Z)$. This is the first parsimony where puff is absent (0) in all ancestors.

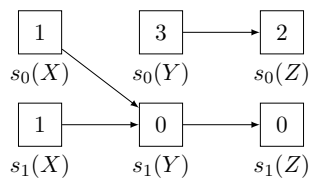
One can represent the structure of *all* parsimonies as a network (or graph) with a *box* for each s_i at each internal node of the phylogeny (*e.g.*, boxes of Fig. 2), and an *arrow* from box $s_i(u)$ to $s_j(v)$ for some node u and

its child v in the phylogeny, whenever $s_i(u)$ can be derived from $s_j(v)$ (*e.g.*, arrows of Fig. 2). A parsimony is then some choice of exactly one box in this network for each internal node of the phylogeny in such a way that they form an underlying directed spanning subtree in this network, in terms of the arrows which join the boxes. This spanning subtree will have the same structure as the phylogeny, in fact, and the choice of box will correspond to the choice of ancestral state of the corresponding internal nodes of this phylogeny. In Fig. 2, for example, the spanning subtree $s_0(X) \to s_0(Y) \to s_0(Z)$ corresponds to the parsimony where puff is absent in all ancestors, and $s_0(X) \to s_1(Y) \to s_1(Z)$ corresponds to the parsimony where puff was ancestral in Y, and lost in the Jaguar. Implicitly the leaves are also included in these spanning subtrees, but since there is no choice of state for extant species, they are left out for simplicity — see [4] for another example of this structure, with the leaves included. Notice, from Fig. 2 that there is a third solution $s_1(X) \to s_1(Y) \to s_1(Z)$ corresponding to a parsimony where puff was ancestral (in X) to all species here, and then lost in both the Tiger and the Jaguar. Note that for internal nodes u other than the root, that the choice of state $i \in \Sigma$ may not be one such that $s_i(u)$ is minimized. In general, for $|\Sigma| = k$, the procedure for building this structure (*e.g.*, in Fig. 2) would take time $O(nk^2)$, since at each node u, each of the k values $s_i(u)$ is derived from $O(k)$ values at the (amortized constant number of) children of u.

Note that there can be $O(n^k)$ parsimonies in general. Indeed, consider the extreme case where all costs in δ are 0: any choice of k states in each of the $O(n)$ internal nodes would be a parsimony. For possibly dense solution spaces, we first take note during the construction of this structure, of any states of each internal node which have the potential to be in an optimal solution. Such states will correspond to any box $s_i(u)$ for some non-root internal node u that is reachable from some minimum score root state via arrows in this structure. For example, in Fig. 2, such states are $s_0(u)$ and $s_1(u)$ for $u \in \{X, Y, Z\}$. Starting with, *e.g.*, $s_0(u)$, for each internal node u, we then enumerate all possible combinations of these potential states in *Gray code* order [24], allowing for time $O(k)$ updates. That is, suppose some combination (*e.g.*, $s_0(X) \to s_1(Y) \to s_1(Z)$) results in a spanning subtree of the network (it is a parsimony). This order ensures that the next combination differs by only one state (*e.g.*, $s_1(X)$), requiring only the update of its arrows with the (unchanged) states of the parent and children of this node (X) in the phylogeny. In this example, since there is an arrow between (new) $s_1(X)$ and (old) $s_1(Y)$, this new combination ($s_1(X) \to s_1(Y) \to s_1(Z)$) is also a spanning subtree. While this structure (Fig. 2) mentioned [4],

sp. \ ch.	roar
Lion	1
Jaguar	1
Leopard	1
Tiger	0

(a)



(b)

Figure 3: The (a) extant state of character roar in four species and (b) the structure resulting from Sankoff's algorithm on these extant states along with the phylogeny of Fig. 1a and cost of Fig. 1c. Values s_2 have been omitted for compactness, since they are all ∞, like in Fig. 2.

how to efficiently enumerate all its parsimonies is rarely discussed (that we know of).

3 Correlation

We want to determine correlated evolution among pairs of characters. Suppose we have a second character roar which is absent only in Tiger as in Fig. 3a, and we want to understand how this is correlated to the puff of Fig. 1b. An established approach to do this is to compare these phylogenetic profiles [22] computing, *e.g.*, their Jaccard (-Tanimoto) Index [18]

$$\frac{|A \cap B|}{|A \cup B|} \tag{4}$$

where A (resp., B) is the set of species containing the first (resp., second) character. In this example with puff and roar, respectively, $A = \{\text{Lion}, \text{Leopard}\}$ and $B = \{\text{Lion}, \text{Jaguar}, \text{Leopard}\}$, hence their Jaccard Index would be $2/3 \approx 0.66$.

This approach only uses information about extant states — however we also have a phylogenetic tree at our disposal. With both of these pieces of information together, we can detect correlated changes in state of pairs of characters along branches of the phylogenetic tree. Such a truly phylogenetic approach has been shown in [23] to improve detection by avoiding spurious correlations caused by large clades of species which may have similar characteristics for the historical reason of being closely related, rather than as a result of adaptive evolution. Since we have no model of evolution in our setting, we determine such changes in state of each character c in the phylogenetic tree by computing the set $P(c)$ of all small parsimonies. The

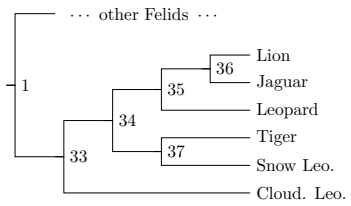
correlation between the change $a = \alpha \to \beta$ in state (for some pair α, β of states) of character c and change b in character d is then

$$\frac{\sum_{p \in P(c), q \in P(d)} |B_p(a) \cap B_q(b)|}{\sum_{p \in P(c)} |B_p(a)| \cdot \sum_{q \in P(d)} |B_q(b)|} \tag{5}$$

where $B_p(a)$ is the set of *branches* in the phylogenetic tree where change a occurred in the respective parsimony p. For example, given characters puff (Fig. 1b) and roar (Fig. 3a), we compute the correlation of change $0 \to 1$ (absent to present, or co-gain) in both of these characters in the phylogeny of Fig. 1a (according to cost of Fig. 1c) as follows. Character puff has the set $P(\text{puff}) = \{p_1, p_2, p_3\}$ of three parsimonies, with $B_{p_1}(0 \to 1) = \{Z \to \text{Lion}, Y \to \text{Leopard}\}$, $B_{p_2}(0 \to 1) = \{X \to Y\}$, and $B_{p_3} = \emptyset$. By inspecting Fig. 3b, character roar has the set $P(\text{roar}) = \{q_1, q_2\}$ of two parsimonies, with $B_{q_1}(0 \to 1) = \{X \to Y\}$ and $B_{q_2}(0 \to 1) = \emptyset$. It hence follows from Eq. 5 that the correlation of this co-gain event is $1/3 \approx 0.33$. The correlation of the co-loss $1 \to 0$ in both puff and roar is also $1/3$, while the correlation of $0 \to 1$ in puff and $1 \to 0$ in roar (and vice versa) are 0, because there is no overlap between sets $B_p(0 \to 1)$ and $B_q(1 \to 0)$ for any pair of parsimonies $p \in P(\text{puff})$ and $q \in P(\text{roar})$ (and vice versa). For completeness, the correlation of any combination of event $0 \to 2$ or $2 \to 0$ with any other event in either puff or roar is 0 because none of these events happen in any parsimony of puff or roar. This measure of correlation in Eq. 5 is similar in idea to the Jaccard Index 4 in that it measures the amount of event co-occurrence, normalized by the amount of independent occurrences of either event, except that this is over all pairs of parsimonies of the respective characters. In a similar approach [25], albeit in a maximum likelihood setting where there are branch lengths, the authors map the probabilities of gain and loss to each branch of the phylogeny in the most likely ancestral assignment. Correlation between such events in a pair of characters is then assessed by computing the mutual information between the associated sets of probabilities. Our approach could be viewed as a discrete analog of this.

4 Experimental Analysis

We used the following data to validate our approach, but also to find some interesting correlations which we describe below. The phylogeny is that of the family Felidae from Johnson et al. [20], and the extant states are for the 14 vocalizations as documented in Sunquist and Sunquist [41]. While the



(a)

sp. \ ch.	prus.	puff	main	m./gr.	roar	gr.
Lion	0	1	0	1	1	1
Jaguar	1	0	1	1	1	1
Leopard	0	1	1	1	1	1
Tiger	1	0	1	1	0	0
Snow Leo.	1	0	1	0	0	0
Cloud. Leo.	1	0	1	2	0	0

(b)

Figure 4: (a) Phylogenetic tree of [20] restricted to the subfamily Pantherinae, containing: Lion, Jaguar, Leopard, Tiger, Snow Leopard and Clouded Leopard. In this tree, ancestor 1 is common to all Felids. (b) Extant states in Pantherinae of six characters of [41], respectively: prusten, puff, main call, main call with grunt element, roaring sequence and grunt. Here, the states 0, 1 and 2 mean absent, present and unknown, respectively.

entire data are too large to depict here, Fig. 4 depicts a portion of the tree corresponding to the subfamily Pantherinae (Fig. 4a), along with the extant states of the six characters that exhibit variation in Pantherinae (Fig. 4b). Interestingly, it is among these six characters, along with the characters gurgle and wah wah (which are not shown here because they were uniformly absent in Pantherinae), that all of the interesting correlation results occur. Finally, we use the same cost matrix as in Fig. 1c, since some of the extant states of the vocalizations have been documented as *unknown*. Since this is an artifact of the collection process, by assigning a high evolutionary cost to the changes $0 \to 2$ and $1 \to 2$ (unknown to some known state), which makes little evolutionary sense, we mitigate the propagation of the unknown state to ancestral nodes in any parsimony, whenever some known state (0 or 1) can explain the data. For example, in the instance of Fig. 1, if the puff was instead unknown in the Jaguar and Tiger, then it would only have the unique parsimony where all ancestors X, Y and Z have state 1 (present). This is why we use the approach of Sankoff [35] (instead of, *e.g.*, Fitch [14]), because we need these transition-specific costs.

Our approach and resulting implementation into the software tool `parcours` begins with the efficient computation of all parsimonies described in Sec. 2, followed by the computation of correlation from these parsimonies, as described in Sec. 3. Since `parcours` hence starts with the same input as Sankoff's algorithm (*e.g.*, Fig. 1), we input the phylogenetic tree and extant states of the 14 vocalizations in Felidae described above, along with cost of Fig. 1c. From this, `parcours` returns the correlations of all possible events

between all pairs of characters according to Eq. 5. While each correlation is, by the definition in Eq. 5, a value between 0 and 1, it remains to show how we assess which such values are statistically significant. Since the Jaccard Index suffers from a similar drawback, studies [2, 32] aim to determine significance by first generating a distribution of Jaccard Index over all possible phylogenetic profiles for a given number (N) of species. Since this is an estimate of the likelihood of observing a certain Jaccard Index at random, the statistical significance of a given Jaccard Index can then be assessed by comparing to this (null) distribution. A similar approach is employed in [22] to assess significant functional overlap (*i.e.*, as computed by Jaccard Index) of groups of proteins. We take a similar approach here. For each character, we first generate (up to) 100 unique random permutations of its phylogenetic profile in the species. Then, for each pair of characters, we run `parcours` on the (up to) 10,000 pairs of respective permutations (along with the phylogeny and cost matrix), to generate a distribution of correlation. The significance of the correlation for the given pair of characters is then based on how it compares to this (null) distribution. Note that such distributions may not have 10,000 values because some profiles do not have 100 unique permutations, or some pairs of permutations result in 0 correlation. In this latter case, we do not count such values towards the distribution, in taking a very conservative approach on determining significance. Table 1 gives all events for all pairs of characters that we recorded as statistically significant. We discuss these results in the final Sec. 5.

5 Discussion

Indeed, `parcours` retrieved some results in this data that are consistent with the literature, validating this approach. For example, the three friendly close-range sounds: the prusten, the gurgle, and the puff, have been documented in Felidae. These calls are considered to be functionally equivalent, and each Felid species has only one of these call types in its acoustic repertoire [29]. The ancestral reconstructions obtained in the first step of `parcours` (Sec. 2) suggest that the gurgle is most likely to be the ancestral, close contact friendly call (possibly being found in the ancestor of all Felidae), which is consistent with past work [31]. Prusten is a close contact call found only in four species: the Tiger, the Clouded Leopard, the Snow Leopard and Jaguar. Ancestral reconstructions suggest that along the lineage leading to Pantherinae, the gurgle was lost and the prusten was gained, a significant correlation (Table 1, row 1) according to `parcours`.

ch. c	ch. d	ev. a	ev. b	corr.	mean	SDs	#vals
prusten	gurgle	$0 \to 1$	$1 \to 0$	0.25	0.049	8.23	5234
m./gr.	gurgle	$0 \to 1$	$1 \to 0$	0.25	0.051	8.35	5066
roar	puff	$0 \to 1$	$0 \to 1$	0.33	0.16	3.82	1775
roar	m./gr.	$0 \to 1$	$0 \to 1$	0.25	0.085	4.78	3100
grunt	puff	$0 \to 1$	$0 \to 1$	0.33	0.16	3.67	1726
grunt	m./gr.	$0 \to 1$	$0 \to 1$	0.25	0.086	4.76	2929
grunt	roar	$0 \to 1$	$0 \to 1$	1.0	0.11	23.09	2543
wah wah	gurgle	$1 \to 0$	$1 \to 0$	0.17	0.051	2.38	2584

Table 1: Pairs a, b of events (columns 3,4) in pairs c, d of characters (columns 1,2) of our study that were statistically significant. Column 5 gives the correlation of a and b (Eq. 5), and columns 6 and 7 give the mean correlation, and the number of SDs (standard deviations, or σ) from this mean in its distribution, respectively. Column 8 gives the number of values in this distribution. All values are reported with at least 2 decimal places (or 2 significant digits). A correlation was recorded as significant when it was at least 2σ from the mean, and the distribution had at least 1000 values.

However, `parcours` also retrieved some interesting new findings. For example, the origin of the puff is uncertain [31]. Ancestral reconstructions indicate that the puff could have a common origin in the ancestor of Lions, Leopards and Jaguars, or an independent origin in both Lions and Leopards (the running example of Fig. 1). Ancestral reconstruction of both the roaring sequence, a high intensity call characterized by regular species-specific structure, and the grunt, each resulted in a single most parsimonious solution, placing the origin of both calls in the common ancestor of Lions, Jaguars and Leopards. The significant correlation of both the gain of the puff and the gain of a roaring sequence and the grunt (Table 1, rows 3 and 5), further reinforced by the highly significant correlation of the roaring sequence and the grunt (Table 1, row 7), is consistent with the idea that the puff most likely evolved in the common ancestor of these cats, and thus, is not convergent. This illustrates an important strength of `parcours`: its leveraging of correlation between pairs (or clusters, in this case) of characters in order to disambiguate parsimonious scenarios involving a single character. While `parcours` currently computes only pairwise correlations, this opens up the idea of studying clusters of characters (of size larger than two), possibly adding this functionality to the methodology. Despite this, we also found the main call with grunt element to be significantly correlated to roar and grunt (Table 1, rows 4 and 6), while also being significantly correlated the gurgle (Table 1, row 2), which likely appeared before the an-

cestor of Lions, Leopards and Jaguars. This ambiguous case motivates the collection of more data, in order to have a more detailed understanding.

Finally, `parcours` indicated that the loss of a gurgle and a vocalization known as the wah wah were significantly correlated (Table 1, row 8). The wah wah is a widespread, close contact onomatopoeic call produced when two conspecific individuals approach one another [41]. Thus, the wah wah may be a type of displacement call. Documentation of the wah wah in Felidae is not complete, complicating efforts to elucidate the ancestral states of this call. As stated earlier, the gurgle is a friendly close contact call, likely ancestral based on our parsimonious reconstructions. The significant co-loss of the wah wah and the gurgle along the lineage leading to Pantherinae suggests that the wah wah is, in fact, an ancestral call. This example illustrates another important aspect of correlation analyses like `parcours`. Significant correlation can help to elucidate ancestral states when empirical data is lacking.

Immediate future work is to use `parcours` to discover correlations between these vocalizations and several ecological characteristics that we have found in the literature. In fact, some of our preliminary work on this indicates that changes in vocalization and increased body size, characteristic of most members of Pantherinae, are not correlated. This suggests that new vocalizations originated before the increase in body size. We hope to investigate more connections like these as we gather the data from the literature. Finally, while we used `parcours` on a vocalizations, it is a general-purpose framework. We could use it, *e.g.*, to find correlations among mutations in a tumor sample, given a tree (constructed by, *e.g.*, [19]), or even validate such a tree, based on correlations that are expected to be found. Given that this framework is general, another future direction is to gather more detailed information about these vocalizations (*e.g.*, mean frequency) from audio recordings, and reconstruct this information. This, along with ecological characteristics such as body size, and how they correlate, should give us an even more refined perspective.

References

[1] BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(214), 2007.

[2] Cesare Baroni-Urbani. A statistical table for the degree of coexistence between two species. *Oecologia*, 44(3):287–289, 1979.

[3] J.H. Camin and R.R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, pages 311–326, 1965.

[4] José Clemente, Kazuho Ikeo, Gabriel Valiente, and Takashi Gojobori. Optimized ancestral state reconstruction using sankoff parsimony. *BMC Bioinformatics*, 10(1), 2009.

[5] O. Cohen, H. Ashkenazy, E. Levi Karin, D. Burstein, and T. Pupko. CoPAP:. *Nucleic Acids Research*, 41:232–237, 2013.

[6] Wandrille Duchemin, Yoann Anselmetti, Murray Patterson, Yann Ponty, Sèverine Bérard, Cedric Chauve, Celine Scornavacca, Vincent Daubin, and Eric Tannier. DeCoSTAR: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biology and Evolution*, 9(5):1312–1319, 2017.

[7] A.W.F. Edwards and L.L. Cavalli-Sforza. The reconstruction of evolution. *Annals of Human Genetetics*, 27:105–106, 1963.

[8] G Ehret. Development of sound communication in mammals. In *Advances in the Study of Animal Behavior*, volume 11, pages 179–225, 1980.

[9] Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.

[10] James S. Farris. Methods for computing wagner trees. *Systematic Biology*, 19(1):83–92, 1970.

[11] Felsenstein. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989.

[12] Joseph Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1978.

[13] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[14] Walter M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[15] Laurent Guéguen, Sylvain Gaillard, Bastien Boussauand Manolo Gouy, Mathieu Groussin, Nicolas C. Rochette, Thomas Bigot, Davied Fournier, Fanny Pouyet, Vincent Cahais, Aurélien Bernard, Céline Scornavacca, Benoît Nabholz, Annabell Haudry, Loïc Dachary, Nicholas Galtier, Khalid Belkhir, and Julien Y. Dutheil. Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution*, 30(8):1745–1750, 2013.

[16] Iman Hajirasouliha and Benjamin J. Raphael. Reconstructing mutational history in multiply sampled tumors using perfect phylogeny mixtures. In *Algorithms in Bioinformatics*, pages 354–367, 2014.

[17] John P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*, 17:754–755, 2001.

[18] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societé Vaudoise des Sciences Naturelles*, 44(163):223–270, 1908.

[19] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biology*, 17(1):86, 2016.

[20] Warren E. Johnson, Eduardo Eizirik, Jill Pecon-Slattery, William J. Murphy, Agostinho Antunes, Emma Teeling, and Stephen J. O'Brien. The late miocene radiation of modern felidae: A genetic assessment. *Science*, 311(5757):73–77, 2006.

[21] B. Kolaczkowski and J.W. Thornton. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431(7011):980–984, 2004.

[22] Edward M. Marcotte, Ioannis Xenarios, Alexander M. van der Bliek, and David Eisenberg. Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 97(22):12115–12120, 2000.

[23] A. Meade and M. Pagel. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinf.*, 23(1):14–20, 2007.

[24] S. Mossige. An algorithm for Gray codes. *Computing*, pages 89–92, 1977.

[25] Murray Patterson, Thomas Bernard, and Daniel Kahn. Correlated evolution of metabolic functions over the tree of life. *bioRxiv*, 093591, 2016.

[26] Murray Patterson, Gergely Szöllősi, Vincent Daubin, and Eric Tannier. Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, 14(S4), 2013.

[27] Matteo Pellegrini, Edward M. Marcotte, Michael J. Thompson, David Eisenberg, and Todd O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[28] G. Peters. Vergleichende Untersuchung zur Lautgebung einiger Feliden (Mammalia, Felidae). *Spixiana*, 1:1–283, 1978.

[29] G. Peters. On the structure of friendly close range vocalizations in terrestrial carnivores (Mammalia: Carnivora: Fissipedia). *Zeitschrift für Säugetierkunde*, 49:157–182, 1984.

[30] G Peters. Vocal communication in cats. In *Great Cats–Majestic Creatures of the Wild*, pages 76–77, 1991.

[31] Gustav Peters and Barbara A. Tonkin-Leyhausen. Evolution of acoustic communication signals of mammals: Friendly close-range vocalizations in felidae (carnivora). *Journal of Mammalian Evolution*, 6(2):129–159, 1999.

[32] Raimundo Real and Juan M. Vargas. The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, 45(3):380–385, 1996.

[33] Camir Ricketts, Mauricio Soto Gomez, Murray Patterson, Dana Silverbush, Paola Bonizzoni, Iman Hajirasouliha, and Gianluca Della Vedova. Inferring cancer progression from Single-Cell Sequencing while allowing mutation losses. *Bioinformatics*, 2020.

[34] R. Romand and G Ehret. Development of sound production in normal, isolated and deafened kittens during the first postnatal months. *Developmental Psychobiology*, 17:629–649, 1984.

[35] David Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

[36] David Sankoff and Pascale Rousseau. Locating the vertices of a steiner tree in an arbitrary metric space. *Mathematical Programming*, 9(1):240–246, 1975.

[37] R. Schwartz and A.A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017.

[38] Mauricio Soto Gomez, Murray D. Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. gpps: an ILP-based approach for inferring cancer progression with mutation losses from single cell data. *BMC Bioinformatics*, 21(1):413, 2020.

[39] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[40] M. Steel and D. Penny. Two further links between MP and ML under the poisson model. *Applied Mathematics Letters*, 17(7):785–790, 2004.

[41] M. Sunquist and F. Sunquist. *Wild Cats of the World*. 2002.

[42] Chris Tuffley and Mike Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59(3):581–607, 1997.