PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Animal vocal sequences: not the Markov chains we thought they were

Journal:	*Proceedings B*
Manuscript ID:	RSPB-2014-1370.R1
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Kershenbaum, Arik; National Institute for Mathematical and Biological Syntehsis, Bowles, Ann; Hubbs SeaWorld Resarch Institute, BioAcoustics Freeberg, Todd; University of Tennessee, Psychology; Jin, Dezhe; Penn State University, 4Department of Physics and the Center for Neural Engineering Lameira, Adriano; University of Amsterdam, Institute for Biodiversity and Ecosystem Dynamics; Pongo Foundation, Bohn, Kirsten; Florida International University, Biological Sciences
Subject:	Behaviour < BIOLOGY
Keywords:	language evolution, renewal process, vocal complexity
Proceedings B category:	Behaviour

Animal vocal sequences: not the Markov chains we thought they were

Arik Kershenbaum[1], Ann E. Bowles[2], Todd M. Freeberg[3], Dezhe Z. Jin[4], Adriano R. Lameira[5,6], Kirsten Bohn[7]

[1]National Institute for Mathematical and Biological Synthesis, Knoxville, TN, USA

[2]Hubbs SeaWorld Research Institute, San Diego, CA 92109, USA

[3]Department of Psychology, University of Tennessee, Knoxville, TN, USA

[4]Department of Physics and the Center for Neural Engineering, Penn State University, University Park, Pennsylvania, USA

[5]Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Sciencepark 904, 1098 XH, The Netherlands

[6]Pongo Foundation, Papenhoeflaan 91, 3421 XN, The Netherlands

[7]Department of Biological Sciences, Florida International University, Miami, FL, USA

CORRESPONDING AUTHOR:

Arik Kershenbaum

Email: arik@nimbios.org

Tel: (865) 974-9334

SUMMARY

Many animals produce vocal sequences that appear complex. Most researchers assume that these sequences are well characterised as Markov chains, i.e. that the probability of a particular vocal element can be calculated from the history of only a finite number of preceding elements. However, this assumption has never been explicitly tested. Furthermore, it is unclear how language could evolve in a single step from a Markovian origin, as is frequently assumed, as no intermediate forms have been found between animal communication and human language. Here we assess whether animal taxa produce vocal sequences that are better described by Markov chains, or by non-Markovian dynamics such as the "renewal process", characterised by a strong tendency to repeat elements. We examined vocal sequences of seven taxa: Bengalese finches *Lonchura striata domestica*, Carolina chickadees *Poecile carolinensis*, free-tailed bats *Tadarida brasiliensis*, rock hyraxes *Procavia capensis*, pilot whales *Globicephala macrorhynchus*, killer whales *Orcinus orca*, and orangutans *Pongo spp.* The vocal systems of most of these species are more consistent with a non-Markovian renewal process than with the Markovian models traditionally assumed. Our data suggest that non-Markovian vocal sequences may be more common than Markov sequences, which must be taken into account when evaluating alternative hypotheses for the evolution of signalling complexity, and perhaps human language origins.

KEYWORDS: Language evolution, renewal process, vocal complexity

1. INTRODUCTION

Many species of animals produce vocalisations comprising multiple element types, combined into complex sequences. Some species have vocal repertoires of tens or even hundreds of discrete elements; others have only a handful, but use them to generate a wide variety of combinations. For example, an individual mockingbird *Mimus polyglottos* can mimic over 100 distinct song types of different species, and combine them into diverse sequences [1]. Even the rock hyrax, *Procavia capensis*, using no more than five discrete vocal elements, creates long vocal sequences that are rarely the same on repetition [2]. Thus, even species with few vocal elements can sometimes generate an apparently unbounded range of possible combinations. Such varied vocal behaviour raises the question of the role and origin of complexity in animal vocal communication, and the comparison of vocal complexity across taxa, including human speech.

Complexity seems easy to identify, but hard to define, and even harder to quantify [3]. Numerous metrics have been suggested to ascribe a value to the complexity of vocal repertoires. However, these metrics all rely, either explicitly or implicitly, on assumptions of the underlying process that generated sets of sequences. For instance, a frequently cited complexity measurement, Shannon entropy, is only appropriate when each element in a sequence is produced independently of all other elements (i.e. an independent production process) although the assumption of independence is rarely tested [4-7]. If vocal sequences are generated by a non-independent random process, however, Shannon entropy is probably not suitable for quantifying complexity [8]. Whether vocal sequences are random independent processes, or conform to some other non-independent stochastic model,

identifying the process operating is an essential task for quantifying and comparing sequence properties.

Beyond the application to complexity metrics, uncovering the processes underlying vocal sequence generation in animals may prove crucial to our understanding of language origins. Vocal complexity naturally brings to mind human language; however the comparison appears to be inappropriate. One of the main differences between language and non-human animal communication is the grammar used to produce sequences. Human language uses "context-free grammars" (CFGs) that are capable of generating recursive sequences and unbounded correlations [9,10]. In contrast, animal vocal sequences are usually described as “regular grammars”, the simplest class of formal grammars [11], and many researchers have analysed animal vocalisations as such (e.g. [12-15]). Regular grammars correspond to Finite State Automata, because they comprise a set of rules that could instruct a simple machine (automaton) to move between a (finite) number of well-defined states. In the case of vocal sequences, each state is an acoustic element. Finite State Automata can be deterministic, for example *syllable ‘A’ is always followed by syllable ‘B’*. They can also be probabilistic (pFSA), in which multiple possible transitions between states are governed by fixed probabilities; for example, *syllable ‘A’ is followed by syllable ‘B’ 90% of the time, and by syllable ‘C’ 10% of the time*. In contrast to deterministic Finite State Automata, different sequences can be generated each time a pFSA is used. pFSAs are an example of a Markov chain [16], the most common model used to examine animal vocal sequences [14]. The pFSA (or Markovian) paradigm assumes that future occurrences (or the probability of each future occurrence) are entirely determined by a finite number of past occurrences. This property of a stochastic sequence is known as the Markov property. For example, the probability of the next syllable in a sequence being of type “*A*” is determined by the types of the immediately preceding syllables – or at most some finite number of preceding syllables.

pFSAs remain popular for characterising animal vocal sequences [11,14], as the mechanism for producing Markov chains is easily understood, and simple neural mechanisms for implementing them have been postulated, based on neuroanatomical observations [17,18]. However, Markov chains are insufficient for producing the complexity of any human language [9] and there exist grammatical structures that no pFSA can generate, in particular tree-like syntax such as "the hyrax ate the grass that grew near the rock under the tree" [11]. Furthermore, no intermediate grammatical form exists between pFSA models, and the CFG of human language [9]. It is not clear what adaptive force could drive the gradual evolution of CFGs, in a species that uses only pFSA vocal communication. In computer science, the addition of register memory, which provides the ability to count the number of repetitions of a syllable, appears to be a simple transition from regular to context-free automata [19]. However, such models have not been described in animal communication.

Despite the widespread use and simplicity of pFSA, there are other, non-Markovian stochastic processes, in particular models where future occurrences are determined by the (infinite) entirety of preceding events [20]. Non-Markovian processes have been used to describe (non-vocal) animal behaviour, for instance the renewal process (RP) model in the reproductive behaviour in sticklebacks, canaries, and *Drosophila* [21], and the psychohydraulic model of motivation (PHM) proposed by Konrad Lorenz [22] for basic drives such as hunger. Although we are not aware of any prior work using non-Markovian processes to describe vocal behaviour, they seem likely candidates for vocal production. For example, non-Markovian mechanisms are able to describe both rapid shifts among vocal elements and long strings of repeated elements. Here we test whether vocal sequences in several species are more consistent with a Markovian pFSA model, or a non-Markovian process, such as the RP or PHM. Non-Markovian stochastic processes like the RP have properties somewhat between the pFSA and the CFG, and the investigation of language

evolution would not be complete without consideration of other, biologically realistic, sequence-generating mechanisms.

Both RP and PHM models are considered non-Markovian because they do not rely on finite memory. In RP models, a particular behaviour (for instance, production of a particular vocal syllable) is repeated for some probabilistically determined time. Transitions between syllables of different types are still defined by a transition table as with a pFSA, but the number of repeats of each syllable in between transitions may be drawn from a distribution, e.g. Poisson. Although at first surprising, it can be shown that the sequence generated by such a process is non-Markovian [23], and cannot be well described by a pFSA. The RP does not fit the Markovian paradigm of finite memory, since the Poisson tail is unbounded. The PHM also relies on a nominally unbounded memory; in this case the probability of a particular syllable occurring increases with the time since its last occurrence, and then falls to a minimum as soon as the syllable is used.

We gathered vocal sequences from seven taxa: the Bengalese finch *Lonchura striata domestica* [24], Carolina chickadee *Poecile carolinensis* [25,26], free-tailed bat *Tadarida brasiliensis* [13], rock hyrax *Procavia capensis* [2], short-finned pilot whale *Globicephala macrorhynchus* [27], killer whale *Orcinus orca* [28], and orangutan *Pongo abelii* and *P. pygmaeus wurmbii* [29]. For comparison with a human sequence corpus, we also analysed letter-order in a sample of English (the text of the play *Hamlet* [30]), although the intention was not to imply that letter-order in human language has any relevance to the evolution of vocal sequences in animals. These sequences were coded for distinct vocal elements (syllables) as described in the above-cited previous works. We aimed to match these sequences to the most likely generation model, from a range of possible models of varying complexity by testing each species' sequences against stochastic production from six different prospective processes: (a) a zero-order Markov process, (b) a first-order Markov

process, (c) a second-order Markov process, (d) a hidden Markov model, (e) a renewal process, and (f) a psychohydraulic process.

2. METHODS

Description of the stochastic processes

First, we describe each of these processes in detail. Consider a sequence S of n elements [1…n], taken from a set of C different element types. The zero-order Markov process (ZOMP) defines a production process where each element is generated according to a fixed prior probability $\boldsymbol{\pi}$, independent of the preceding elements, so that the probability of the n^{th} element S_n being of type i=[1…C] is given by $\Pr(S_n=i)=\pi_i$. In the first-order Markov process (FOMP), the probability that the n^{th} element will be of type i is determined only by the preceding element j, and the C x C transition matrix $\mathbf{T}$, which defines the probability that element i will occur after element j, so that $\Pr(S_n=i \mid S_{n-1}=j)=T_{j,i}$. Similarly, the second-order Markov process (SOMP) defines the probability of the n^{th} element in terms of the two preceding elements: $\Pr(S_n=i \mid S_{n-1}=j,\ S_{n-2}=k)=U_{(j,k),i}$. Note that the size of the second-order transition matrix $\mathbf{U}$ is of size C^2 x C, which indicates the rapid increase in sample size required for accurate estimates of the transition probabilities, as the order of a Markov process increases [31].

The hidden Markov model (HMM) [32] provides a more parsimonious and memory-efficient representation of higher-order Markov processes, and has been used successfully to capture the characteristics of vocal sequences from different species, e.g. [24,33]. As with the traditional Markov models, in generating an HMM sequence, successive elements are chosen

probabilistically given the current state. Unlike traditional Markov models, though, in the HMM the states themselves are not explicitly defined in terms of preceding sequences of a fixed number of elements, but are constructed from the data by an Expectation-Maximisation optimisation known as the Baum–Welch algorithm [32]. This allows the HMM to represent a combination of low- and high-order Markov relationships within the same model.

The renewal process (RP) is defined by a first-order transition matrix, which determines the pFSA transitions between different elements. This matrix **R** is defined in a similar way to the FOMP transition matrix **T**, but with zeroes along the main diagonal. Instead, those self-transitions are generated by a separate stochastic process. In this case, we define the number of repeated elements as being drawn from a Poisson distribution with mean μ, with a separate Poisson distribution for each element type *i.* A graphical description of the differences between the RP and Markovian processes can be found in [8], and is reproduced in the ESM (ESM Figure 1).

Although the psychohydraulic model (PHM) has not previously been used to describe animal communication, it forms a useful counterpoint to the RP. Whereas in a RP model, repeated elements occur more often than would be expected in a Markov model, in a PHM repeated elements are less common than expected. We implement a simplified PHM by defining for each element type i a function of the form $A_i(t_i) = 1 - e^{-\kappa_i t_i}$ where t_i is the time elapsed since element i last appeared, k_i is an element-specific rate constant, and A_i is the equivalent of what Lorenz coined the "action-specific energy", i.e. the driving motivational force that builds up within an animal until a particular behaviour is precipitated. The probability of the next element being of type i is then given by $\Pr(i) = A_i \Big/ \sum_i A_i$.

187 **Generation of synthetic sequences**

188 We determined the maximum likelihood estimator parameters for each of the processes,
189 given the empirical data. For the ZOMP, the parameter $\boldsymbol{\pi}$ is simply the observed prior
190 probabilities of each of the element types *i*. For the FOMP and SOMP, the matrices **T** and **U**
191 are estimated from the number of occurrences of the specific transitions between element
192 types within the observed sequences.

193 For the HMM, the parameters of the model are calculated from the empirical data using the
194 standard Viterbi algorithm [34]. In any HMM implementation, the number of states is a
195 crucial factor in the model performance, therefore we optimised the number of hidden states
196 by minimising the Akaike Information Criterion [35]. To do this, we calculated the log
197 likelihood of generating the training sequence from the trained HMM, and used the number
198 of hidden states as the number of parameters in the information criterion calculation.

199 For the RP, the matrix of transitions between different elements **R** is calculated as for the
200 FOMP, and the means $\boldsymbol{\mu}$ of the repeated element Poisson distributions are estimated from the
201 empirical distributions of number of repeats, separately for each element type *i*. For the PHM,
202 the rate constants $\boldsymbol{\kappa}$ are also estimated from the empirical distributions of the intervals
203 between elements of the same type.

204 Having extracted the maximum likelihood estimator parameters for each model, we used
205 these to generate artificial sequences based on each model, where the sequence lengths
206 matched those of the original vocal data sets (Figure 1). An overview of the data set sizes and
207 sequence lengths is given in Table 1, and the data themselves are available in the ESM
208 (data.xls). For each species, we generated 200 artificial data sets using each of the model
209 types.

210

211 **Comparison of artificial and recorded sequences**

212 Determining whether a particular sequence, vocal or otherwise, is Markovian or not is a non-
213 trivial problem. Rigorous tests for finite-state sequences exist [36], but are not easily applied
214 to data sets of limited size. When limited data are available, transition probabilities are poorly
215 estimated by the small number of transitions observed. In addition, rare states may be
216 completely absent. Previous authors have used multiple measures of the statistical properties
217 of the sequences, such as n-gram distribution [24]. However, these techniques measure
218 aggregate similarity, and do not directly address the similarity of the individual sequences.
219 Aggregate comparisons may be an effective way of comparing very long sequences, where
220 the probability distribution of n-grams would be expected to be limiting. However, they
221 would be less accurate when comparing short strings such as those found in real recordings,
222 and when the processes generating these strings may not be stationary (for instance due to
223 shifting motivational state and responses to external events). We used a more direct method
224 by comparing each simulated sequence with the corresponding original data, and calculating
225 the edit (Levenshtein) distance [37] between the pair of sequences. Levenshtein distance
226 measures the minimum number of insertions, deletions, and replacements necessary to
227 convert one sequence into another, and has been used for assessing vocal syntax in previous
228 studies [2,38,39]. This distance gives a measure of dissimilarity between the simulated and
229 original sequences, which we then averaged over the entire data set. We calculated the
230 Levenshtein distance between corresponding sequences, both in the simulated and original
231 data, to generate a pairwise distance matrix. We then repeated this for 200 randomly
232 generated sequence data sets for each model and each species.

Having measured the mean Levenshtein distance between a data set and the maximum likelihood estimator prospective models, we used Multi-Dimensional Scaling [40] to convert the Levenshtein distance matrix to a series of Cartesian vectors, one for each sequence, which preserved to the greatest extent possible the pairwise Levenshtein distances between all of the sequences. The transformation of the distance matrix to a feature-space matrix, allowed us to use classification algorithms for assigning the simulated data to the most likely model. For each data set of N sequences, we used the Matlab function *cmdscale* to convert the N x N distance matrix to a matrix Y consisting of a series of N vectors of length p, where $p<N$ is the minimum dimensionality in which the N points can be embedded, i.e. where the pairwise distances between the points are conserved. We then reduced the dimensionality of each vector to length $q \leq p$, where q is the number of eigenvalues E of $Y \cdot Y'$ for which E is positive, and the change in successive eigenvalues $\Delta E=[E(r)-E(r+1)]/E(r)$, $r=[1 \ldots p-1]$ is greater than 1%.

We used both a naïve Bayesian classifier and a Z-test to determine from which of the six generation models the original sequences were most likely to have been drawn. The naïve Bayesian classifier [41] calculated the posterior probabilities of belonging to each of the six model clusters, in q-dimensional space, given the distribution of the 200 sequences for each of the six models. The model with the highest posterior probability was chosen as the candidate model. We then performed an additional Z-test, using the Matlab *normfit* function, to compare the mean distance of the original data to 200 simulated samples of the candidate model. We used a Monte Carlo method to take into account the variation within each model, and give an estimate of the probability that the observed data were drawn from a distribution characterised by the 200 simulated samples of the candidate model. Simulated sequences that are very similar to each other (low variance) are clustered together in distance space, whereas simulated sequences with a high variance are spread out in distance space (see Figure 2).

Therefore, any particular empirical data set is more likely to fall within 95% confidence limits of a high variance model than a low variance one.

Higher order Markov models are, by definition, generalisations of lower order models, and in particular, the HMM is a generalisation of any arbitrary order Markov model. Therefore, it might appear that an HMM model must necessarily provide a maximum likelihood estimator of the model parameters that is at least as accurate as lower order models, if less parsimonious (having a greater number of model parameters). However, we compared the original sequences directly to the corpus of generated sequences, so our similarity metric more broadly measured the appropriateness of each model, and often showed a better fit from the lower order models (Figure 3). We also performed an Analysis of Variance, and post-hoc Tukey test, to assess whether the Levenshtein distances between the original sequences and their corresponding simulated sequences are significantly different among the different models.

Given that transition probability estimates are likely to be inaccurate for small sample sizes, we tested the robustness of our conclusions by repeating the analyses using smaller subsets of the empirical data. We sub-sampled each data set and determined the best-fit model for each sample size. If the data set in its entirety is of sufficient size to estimate the best-fit model, we expect that the best-fit model would be consistent between the larger and full sample sizes.

3. RESULTS

For the purpose of visualisation, the naïve Bayesian classifier is illustrated in Figure 2 with a 2-dimensional embedding, rather than a full q-dimensional embedding (although the 2-

dimensional embedding is in general insufficient to capture the distribution of the sequences in Levenshtein distance-space). Figure 2 illustrates the location of the simulated sequences in distance-space, the location of the original data, and the domains of the classifier. Note that the spread of the simulated sequences varies substantially between models. For those models where the simulated sequences are tightly grouped (small Levenshtein distance between them), the Z-test is more likely to reject the hypothesis that the original data belong to this model, as the variance of the simulated sequences is small, and the original data are likely to fall several standard deviations from the mean of the simulated cluster. Figure 3 shows the results of the Z-test; comparing the distribution of distances within the simulated data set of the candidate model, and the distance of the original data from the simulated set. Where the original data distance is far from the intra-model distances, the data are unlikely to have been drawn from the model.

Table 2 shows the results of the Bayesian classification for each of the eight species (including English), along with the result of the Z-test for the most likely candidate model. The Shapiro-Wilk test for normality did not reject a normal distribution for any of the best fit models, supporting the use of a Z-test. ESM Table 1 shows the results of the Z-tests for all models. Of the seven non-human species, none show clear Markovian behaviour. The Bengalese finch, Carolina chickadee, free-tailed bat, pilot whale, and killer whale appear most similar to the non-Markovian RP, and the Z-test does not reject the RP model (P=0.824, P=0.989, P=0.764, P=0.586, P=0.646 respectively). The orangutan and the hyrax are most similar to the Markovian FOMP, but for both of these species the FOMP is a poor fit to the data, and the vocal sequences are sufficiently different that we reject the null hypothesis of belonging to that model (orangutan P=0.026, hyrax P<0.001). Letter order in a sample of English writing appears to follow a Markovian ZOMP model (P=0.914). The PHM did not appear to be a good model for any of the data sets tested.

306 The test of robustness by varying sample size showed that for all data sets used, except the
307 pilot whale (which passed the Z-test) and the hyrax (which failed the Z-test), the conclusion
308 of best-fit models was consistent at larger sub-sample sizes (see ESM Figure 2).

309

310 4. DISCUSSION

311

312 Our results show that the vocal sequences of over half of the species studied - the Bengalese
313 finch, the Carolina chickadee, the free-tailed bat, the pilot whale, and the killer whale - can be
314 better described as non-Markovian renewal processes, rather than traditional first-order,
315 second-order, or arbitrary-order hidden Markov models. We cannot reliably identify a
316 stochastic process generating the sequences of the hyrax or the orangutan, and it would be
317 interesting to investigate why these vocalisations are qualitatively different from the others
318 studied, whether because of phylogeny, functionality, or other constraints.

319 This diversity of production models is quite unexpected, as previous works have
320 overwhelmingly used the Markovian paradigm as a starting-point for the analysis of animal
321 vocal sequences [11,14]. Although putative Markovian generation processes are popular,
322 partly because of their simplicity, and partly because of the clear role that they fill in the
323 Chomsky hierarchy [9,10], it is inappropriate to assume that they adequately describe the true
324 generation process, simply because of their utility. Indeed, it seems simplistic to assume that
325 animals would primarily generate their vocal sequences based solely on a small number of
326 preceding elements. Renewal processes, in which a certain element is repeated until the
327 animal is “tired of it” (whether physically, cognitively, or only figuratively), are alternative

328 models, and indeed we have shown the RP to be a better approximation for the vocalisations
329 of most of the species we examined.

330 Repetition has long been recognised as a feature of animal behaviour, e.g. eventual variety in
331 birdsong [42], and non-vocal behavioural repetition [21], although the mechanisms
332 responsible may be diverse [43,44]. Mating and aggressive displays make use of repetition to
333 augment the magnitude of the display signal [43], and repeated displays appear more
334 effective in attracting a mate or deterring a rival in species including songbirds [45] and
335 fallow deer *Dama dama* [46]. However, a tradeoff must exist between the benefit of signal
336 repetition, and energetic costs or physiological constraints [45,47]. Such a tradeoff may be a
337 proximal cause of a non-monotonic distribution of the number of repeats, such as the Poisson
338 distribution of the proposed renewal process, which appears to be consistent with our
339 empirical data.

340 Characterising vocal sequences as Markov chains places animal vocal sequences in the
341 category of regular grammars, and distinguishes them from the more complex context-free
342 structure of human language. However, we have shown that the oft-cited conclusion that all
343 animal communication conforms to regular grammars [11,18] is misleading. Indeed, little
344 mention has been made in the literature of non-Markovian alternatives to the pFSA grammar.
345 No attempt has been made until now to test whether animal vocal sequences are indeed most
346 likely generated by pFSAs, or instead by some other, non-Markovian, stochastic process. It
347 has been pointed out [48,49] that insufficient attention has been given to the different levels
348 of complexity in pFSAs of different types (i.e. different orders, vs. HMMs), and we extend
349 this observation to non-Markovian processes. Claims that certain species such as European
350 starlings *Sturnus vulgaris* perceive vocal sequences with a grammar more complex than
351 regular-grammar, have been met with scepticism [50-52]. However, our findings do not point

to *greater* grammatical complexity, but to *different* grammatical processes, something so far barely examined in the literature.

For this comparison, we have used a small but diverse set of data. Some of the data sets, such as the sequences obtained from orangutans, were necessarily rather small because of the difficulty of working in the wild with an inaccessible, endangered, and semi-solitary species. The sequences from the pilot whales potentially contained biasing information, since the audio recorders attached to the animals could also detect the calls of other individuals. However, such a bias would tend to produce a more independent (ZOMP-like) sequence, whereas our findings for the pilot whale indicated a low probability of independent generation. The pilot whale data were also unusual in that they consisted primarily of stereotyped calls, and few non-stereotyped calls; the occurrence of such sequences is likely highly context dependent [27]. This could indicate either atypical behaviour or, possibly, unusually communicative behaviour. We believe that, despite these limitations, inclusion of these species helps to broaden the scope of our comparison, since primates and cetaceans are mammalian orders recognised as having particularly sophisticated acoustic communication.

Estimating the parameters of probabilistic models from small sample sizes is necessarily problematic [53,54]. For example, a FOMP transition table for the English alphabet is a matrix of size 26 x 26 = 676 cells, and assuming that at least 10 observations are required to provide a reasonable estimate of each transition probability, then at least 6760 transitions must be made. In practice, the required number of observations is much more, as some transitions may be rarely observed. For a SOMP, more than 10^5 observations are required. In most cases, our data fall far short of the desirable number of observations. However, reasonable estimates of model parameters can often be made with surprisingly small sample sizes [8,54]. The results of our test for robustness show that, with the exception of the pilot whale data, wherever a clear best-fit model is indicated, the chosen model is consistent

between larger sub-sample sizes, and we believe that this is an indication of reliability of our conclusions.

We found that the letter order in the English language is best modelled by a ZOMP, whereas previous work has indicated that a SOMP is a more appropriate model [55]. However, English is clearly neither a zero- nor a first-order Markov process, and so the resemblance of a corpus of English letters to one or the other, may be more dependent on the metric used to assess similarity, rather than the underlying stochastic processes. Information theoretic approaches (e.g. [36]) naturally lean toward the second-order Markovian paradigm; for example, because the letter 't' is so often followed by the letter 'h'. However, we believe that our approach of comparing the string similarity of sequences generated by putative models, provides a more useful comparison in the field of animal communication research, although possibly less useful for analysing human texts.

To the best of our knowledge, no extant species other than humans have a true language, with an unlimited ability to communicate abstract concepts [56]. Although many non-human animal species have essential precursor abilities, such as vocal production learning [57], contextual reference [58-61], and non-semantic syntax [2,27,62], only humans have a grammatical structure that is sufficiently complex for true linguistic potential [56]. Since no non-human species demonstrate proto-linguistic grammars, proposed mechanisms for the evolution of language in humans remain speculative, e.g. [63]. Among theories of language origin that posit language evolving from systems like extant non-human animal communication, it is debated whether language arose as a gradual adaptation of simpler vocal communication systems [64] or gestural systems [65], or whether essential linguistic abilities arose suddenly or at least very rapidly [11]. Although a conceptual path between regular and supra-regular grammars is well accepted in the computer science literature [19], an important question is whether an incremental *evolutionary* path exists between the pFSA regular

grammars that heretofore were considered common in animals, and the CFG linguistic structures that exist in humans. The incremental hypothesis must explain the lack of "proto-languages" in the animal kingdom, representing a link between animal and human linguistic capabilities [66]. Conversely, saltationary or rapid-evolution hypotheses must provide a convincing and evolutionarily plausible mechanism that could explain the qualitative gap between the regular grammar of animal communication and the context-free grammar of human language. Examples would include metric (timing) features [11], or a synthesis of multiple regular grammars [63] as a "bridge" between the two capabilities. Recent work has indicated that complex syntax can develop as the result of simple neurological changes; for example, in Bengalese finches, which have syntax qualitatively more complex than their wild ancestors [67].

Our findings appear to indicate that pFSA is *not* the ubiquitous nature of animal vocal sequences, and this requires re-evaluation of both gradual and saltational hypotheses. Application of our analysis to more species, and the use of more putative non-Markovian stochastic models, may reveal intermediate steps between known Markovian animal grammars and human context-free grammars, narrowing the gap between human and non-human animal communicative abilities.

ACKNOWLEDGEMENTS

AK funding: National Institute for Mathematical and Biological Synthesis, sponsored by the National Science Foundation, U.S. Department of Homeland Security, U.S. Department of Agriculture (#EF-0832858), The University of Tennessee, Knoxville.

425 Pilot whale permits: US NMFS 1121-1900, 981-1578, Bahamas 01/09, 02/07, 02/08;
426 funding: SERDP, ONR, NOAA, US Navy Environmental Readiness Division; call
427 classification: Laela Sayigh, Nicola Quick, Gordon Hastie, Peter Tyack.

428 ARL permissions: Indonesian RISTEK, Indonesian PHKA, Leuser Ecosystem Management
429 Authority (BPKEL); support: Universitas Nasional Jakarta, Sumatran Orangutan
430 Conservation Programme, Borneo Orangutan Survival, Carel van Schaik, Maria van
431 Noordwijk, Serge Wich; funding: The Menken Funds (University of Amsterdam).

432 Killer whale call classification: Jessica Crance, Juliette Nash; support: SeaWorld Parks.

433 DZJ funding: NSF IOS-0827731.

434

435 REFERENCES

436 [1] Gammon, D. E. & Altizer, C. E. 2011 Northern Mockingbirds produce syntactical patterns of
437 vocal mimicry that reflect taxonomy of imitated species. *J. Field Ornithol.* **82**, 158-164.

438 [2] Kershenbaum, A., Ilany, A., Blaustein, L. & Geffen, E. 2012 Syntactic structure and geographical
439 dialects in the songs of male rock hyraxes. *Proc R Soc Lond B Biol Sci.* **279**, 2974-2981.

440 [3] Edmonds, B. 1999 What is Complexity?-The philosophy of complexity per se with application to
441 some examples in evolution. In *The Evolution of Complexity* (eds. F. Heylighen & D. Aerts), pp. 1-
442 16. Dordrecht: Kluwer.

443 [4] Da Silva, M. L., Piqueira, J. R. C. & Vielliard, J. M. E. 2000 Using Shannon entropy on measuring
444 the individual variability in the rufous-bellied thrush *Turdus rufiventris* vocal communication. *J.*
445 *Theor. Biol.* **207**, 57-64.

446 [5] McCowan, B., Doyle, L. R. & Hanser, S. F. 2002 Using information theory to assess the diversity,
447 complexity, and development of communicative repertoires. *J Comp Psychol.* **116**, 166-172.

448 [6] Suzuki, R., Buck, J. R. & Tyack, P. L. 2006 Information entropy of humpback whale songs. *J.*
449 *Acoust. Soc. Am.* **119**, 1849-1866.

450 [7] Doyle, L. R., McCowan, B., Hanser, S. F., Chyba, C., Bucci, T. & Blue, J. E. 2008 Applicability
451 of information theory to the quantification of responses to anthropogenic noise by southeast Alaskan
452 humpback whales. *Entropy.* **10**, 33-46.

453 [8] Kershenbaum, A. 2013 Entropy rate as a measure of animal vocal complexity. *Bioacoustics.*

454 [9] Chomsky, N. 2002 *Syntactic structures, 9th edition*. The Hague: de Gruyter Mouton.

455 [10] Jäger, G. & Rogers, J. 2012 Formal language theory: refining the Chomsky hierarchy.
456 *Philosophical Transactions of the Royal Society B: Biological Sciences.* **367**, 1956-1970.

457 [11] Berwick, R. C., Beckers, G. J., Okanoya, K. & Bolhuis, J. J. 2012 A bird's eye view of human
458 language evolution. *Front Evol Neurosci.* **4**.

459 [12] Robinson, J. G. 1979 An analysis of the organization of vocal communication in the titi monkey
460 *Callicebus moloch. Z. Tierpsychol.* **49**, 381-405.

461 [13] Bohn, K. M., Schmidt-French, B., Schwartz, C., Smotherman, M. & Pollak, G. D. 2009
462 Versatility and stereotypy of free-tailed bat songs. *PLoS ONE.* **4**, e6746.

463 [14] ten Cate, C. & Okanoya, K. 2012 Revisiting the syntactic abilities of non-human animals: natural
464 vocalizations and artificial grammar learning. *Philos Trans R Soc Lond B Biol Sci.* **367**, 1984-1994.

465 [15] Briefer, E., Osiejuk, T. S., Rybak, F. & Aubin, T. 2010 Are bird song complexity and song
466 sharing shaped by habitat structure? An information theory and statistical approach. *J. Theor. Biol.*
467 **262**, 151-164.

468 [16] Grinstead, C. M. & Snell, J. L. 1997 Chapter 11: Markov chains. In *Introduction to probability*
469 (eds. C. M. Grinstead & J. L. Snell), pp. 405-470. Providence, RI: Amer Mathematical Society.

[17] Jin, D. Z. 2009 Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC. *Phys Rev E.* **80**, 051902.

[18] Katahira, K., Suzuki, K., Okanoya, K. & Okada, M. 2011 Complex sequencing rules of birdsong can be explained by simple hidden Markov processes. *PloS One.* **6**, e24516.

[19] Minsky, M. L. 1967 *Computation: finite and infinite machines*. NJ, USA: Prentice-Hall, Inc.

[20] Van Kampen, N. 1998 Remarks on non-Markov processes. *Brazilian Journal of Physics.* **28**, 90-96.

[21] Cane, V. R. 1959 Behaviour sequences as semi-Markov chains. *J R Stat Soc Series B Stat Methodol.* **21**, 36-58.

[22] Lorenz, K. Z. 1950 The comparative method in studying innate behavior patterns. In *Physiological Mechanisms in Animal Behavior*, pp. 221-268. Oxford: Academic Press.

[23] Nelson, R. 1995 *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modeling*. New York: Springer Verlag.

[24] Jin, D. Z. & Kozhevnikov, A. A. 2011 A compact statistical model of the song syntax in Bengalese finch. *PLoS Comput Biol.* **7**, e1001108.

[25] Freeberg, T. M. 2012 Geographic variation in note composition and use of chick-a-dee calls of Carolina chickadees (*Poecile carolinensis*). *Ethology.* **118**, 555-565.

[26] Freeberg, T. M. 2008 Complexity in the chick-a-dee call of Carolina chickadees (*Poecile carolinensis*): associations of context and signaler behavior to call structure. *Auk.* **125**, 896-907.

[27] Sayigh, L., Quick, N., Hastie, G. & Tyack, P. 2012 Repeated call types in short-finned pilot whales, *Globicephala macrorhynchus. Mar. Mamm. Sci.* **29**, 312-324.

[28] Crance, J. L., Bowles, A. E. & Garver, A. 2014 Evidence for vocal learning in juvenile male killer whales, *Orcinus orca,* from an adventitious cross-socializing experiment. *J. Exp. Biol.* **217**, 1229-1237.

[29] Lameira, A. R., de Vries, H., Hardus, M. E., Hall, C. P., Mitra-Setia, T., Spruijt, B. M., Kershenbaum, A., Sterck, E. H., van Noordwijk, M. & van Schaik, C. 2013 Predator guild does not influence orangutan alarm call rates and combinations. *Behav. Ecol. Sociobiol.* **67**, 519-528.

[30] Shakespeare, W. The Tragedy of Hamlet, Prince of Denmark **2014**. http://shakespeare.mit.edu/hamlet/full.html, Accessed 2014.

[31] Doyle, L. R., McCowan, B., Johnston, S. & Hanser, S. F. 2011 Information theory, animal communication, and the search for extraterrestrial intelligence. *Acta Astronaut.* **68**, 406-417.

[32] Cappé, O., Moulines, E. & Rydén, T. 2005 *Inference in Hidden Markov Models*. New York: Springer Science Business Media.

[33] Reby, D., André-Obrecht, R., Galinier, A., Farinas, J. & Cargnelutti, B. 2006 Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (*Cervus elaphus*) stags. *J. Acoust. Soc. Am.* **120**, 4080-4089.

[34] Garivier, A. 2012 The Baum-Welch algorithm for hidden Markov Models: speed comparison between octave / python / R / scilab / matlab / C / C++. http://www.math.univ-toulouse.fr/~agarivie/Telecom/code/index.php, Accessed 2013.

[35] Burnham, K. P. & Anderson, D. R. 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer Verlag.

[36] DeDeo, S. 2012 Evidence for non-finite-state computation in a human social system. *arXiv preprint arXiv:1212.0018*.

[37] Ristad, E. S. & Yianilos, P. N. 1998 Learning string-edit distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* **20**, 522-532.

[38] Gil, D. & Slater, P. J. 2000 Song organisation and singing patterns of the willow warbler, *Phylloscopus trochilus. Behaviour.* **137**, 759-782.

[39] Garland, E. C., Lilley, M. S., Goldizen, A. W., Rekdahl, M. L., Garrigue, C. & Noad, M. J. 2012 Improved versions of the Levenshtein distance method for comparing sequence information in animals' vocalisations: tests using humpback whale song. *Behaviour.* **149**, 1413-1441.

[40] Cox, T. F. & Cox, M. A. A. 2000 *Multidimensional scaling*. Berlin: Springer.

[41] Duda, R. O., Hart, P. E. & Stork, D. G. 2012 *Pattern Classification*. New York: John Wiley & Sons.

[42] Catchpole, C. K. & Slater, P. J. B. 2003 *Bird song: biological themes and variations*. Cambridge: Cambridge Univ Press.

[43] Payne, R. J. H. & Pagel, M. 1997 Why do animals repeat displays?. *Anim. Behav.* **54**, 109-119.

[44] Mowles, S. L. & Ord, T. J. 2012 Repetitive signals and mate choice: insights from contest theory. *Anim. Behav.* **84**, 295-304.

[45] Podos, J. 1997 A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution.*, 537-551.

[46] Jennings, D. J., Gammell, M. P., Carlin, C. M. & Hayden, T. J. 2005 Win, lose or draw: a comparison of fight structure based on fight conclusion in the fallow deer. *Behaviour.* **142**, 423-439.

[47] Draganoiu, T. I., Nagle, L. & Kreutzer, M. 2002 Directional female preference for an exaggerated male trait in canary (*Serinus canaria*) song. *Proc. Biol. Sci.* **269**, 2525-2531.

[48] Hurford, J. R. 2011 *The Origins of Grammar: Language in the Light of Evolution II*: Oxford University Press.

[49] Petkov, C. I. & Wilson, B. 2012 On the pursuit of the brain network for proto-syntactic learning in non-human primates: conceptual issues and neurobiological hypotheses. *Philos Trans R Soc Lond B Biol Sci.* **367**, 2077-2088.

[50] Gentner, T. Q., Fenn, K. M., Margoliash, D. & Nusbaum, H. C. 2006 Recursive syntactic pattern learning by songbirds. *Nature.* **440**, 1204-1207.

[51] Beckers, G. J., Bolhuis, J. J., Okanoya, K. & Berwick, R. C. 2012 Birdsong neurolinguistics: songbird context-free grammar claim is premature. *Neuroreport.* **23**, 139-145.

[52] Corballis, M. C. 2007 Recursion, language, and starlings. *Cognitive Science.* **31**, 697-704.

[53] Cover, T. M. & Thomas, J. A. 1991 *Elements of information theory*, pp. 18-21. New York, NY: John Wiley & Sons, Inc.

[54] Hausser, J. & Strimmer, K. 2009 Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J Mach Learn Res.* **10**, 1469-1484.

[55] Kundu, A. & He, Y. 1991 On optimal order in modeling sequence of letters in words of common language as a Markov chain. *Pattern Recognit.* **24**, 603-608.

[56] Hauser, M. D., Chomsky, N. & Fitch, W. 2002 The faculty of language: What is it, who has it, and how did it evolve?. *Science.* **298**, 1569-1579.

[57] Janik, V. M. & Slater, P. J. 1997 Vocal learning in mammals. *Adv. Study Behav.* **26**, 59-99.

[58] Slobodchikoff, C., Kiriazis, J., Fischer, C. & Creef, E. 1991 Semantic information distinguishing individual predators in the alarm calls of Gunnison's prairie dogs. *Anim. Behav.* **42**, 713-719.

[59] Cheney, D. L. & Seyfarth, R. M. 2005 Constraints and preadaptations in the earliest stages of language evolution. *The Linguistic Review.* **22**, 135-159.

[60] Clay, Z. & Zuberbühler, K. 2011 Bonobos extract meaning from call sequences. *PloS One.* **6**, e18786.

[61] Arnold, K. & Zuberbühler, K. 2012 Call combinations in monkeys: Compositional or idiomatic expressions. *Brain Lang.* **120**, 303-309.

[62] Shapiro, A. D., Tyack, P. L. & Seneff, S. 2010 Comparing call-based versus subunit-based methods for categorizing Norwegian killer whale, *Orcinus orca*, vocalizations. *Anim. Behav.* **81**, 377-386.

[63] Miyagawa, S., Berwick, R. C. & Okanoya, K. 2013 The emergence of hierarchical structure in human language. *Front Psychol.* **4**.

[64] Jackendoff, R. 2011 What is the human language faculty?: Two views. *Language.* **87**, 586-624.

[65] Tomasello, M. 2008 *Origins of Human Communication*: MIT press Cambridge.

[66] Clark, B. 2013 Syntactic theory and the evolution of syntax. *Biolinguistics.* **7**, 169-197.

[67] Katahira, K., Suzuki, K., Kagawa, H. & Okanoya, K. 2013 A simple explanation for the evolution of complex song syntax in Bengalese finches. *Biol. Lett.* **9**, 20130842.

572 FIGURE LEGENDS

573

574 Figure 1. Flow diagram illustrating the calculation of the distance metric between model and
575 empirical data. Empirical sequences (1) are used to derive maximum likelihood estimator
576 parameters for each of the models (2). Using these parameters, simulated sequences are
577 generated (3) and compared to the corresponding original sequences (4). The average edit
578 distance between these pairs of sequences is a measure of similarity between sequence and
579 model (5).

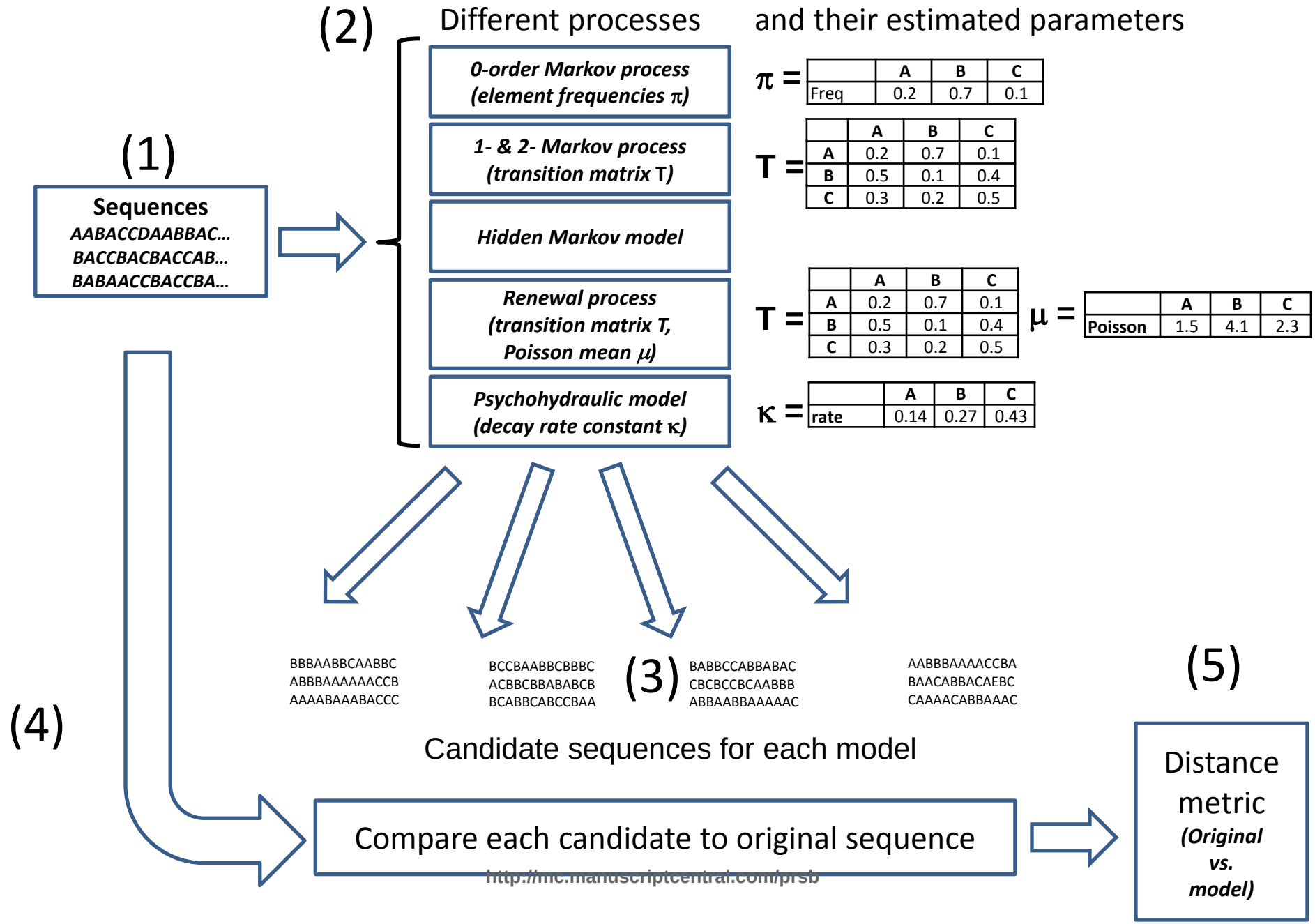
580

581 Figure 2. Location of the simulated sequences and original data (black circle) in 2-
582 dimensional Levenshtein distance space. Coloured points indicate only the first 30 randomly
583 generated sequences from each model, for clarity: ZOMP (red), FOMP (green), SOMP
584 (blue), HMM (cyan), RP (magenta), and PHM (yellow). Solid colours indicate the domains
585 of the naïve Bayesian classifier for each model type.

586

587 Figure 3. Histograms of the Levenshtein distances of simulated sequences from each other
588 (blue bars) for the best fit model (indicated in the title of each panel), and the fitted normal
589 distribution (green line) using the Matlab *normfit* function. The red line shows the mean
590 Levenshtein distance of the original data from the simulated sequences, and the *P* value
591 indicates the probability of this mean distance (or greater) having been drawn from the
592 distribution of simulated sequences (Z-test).

593


(1)
Sequences
AABACCDAABBAC…
BACCBACBACCAB…
BABAACCBACCBA…
(2)
Different processes
and their estimated parameters
0-order Markov process (element frequencies π)
1- & 2- Markov process (transition matrix T)
Hidden Markov model
Renewal process (transition matrix T, Poisson mean μ)
Psychohydraulic model (decay rate constant κ)
π =
A B C
Freq 0.2 0.7 0.1
T =
A B C
A 0.2 0.7 0.1
B 0.5 0.1 0.4
C 0.3 0.2 0.5
T =
A B C
A 0.2 0.7 0.1
B 0.5 0.1 0.4
C 0.3 0.2 0.5
μ =
A B C
Poisson 1.5 4.1 2.3
κ =
A B C
rate 0.14 0.27 0.43
BBBAABBCAABBC
ABBBAAAAAACCB
AAAABAAABACCC
BCCBAABBCBBBC
ACBBCBBABABCB
BCABBCABCCBAA
(3)
BABBCCABBABAC
CBCBCCBCAABBB
ABBAABBAAAAAC
AABBBAAAACCBA
BAACABBACAEBC
CAAAACABBAAAC
(5)
(4)
Candidate sequences for each model
Compare each candidate to original sequence
Distance metric (Original vs. model)

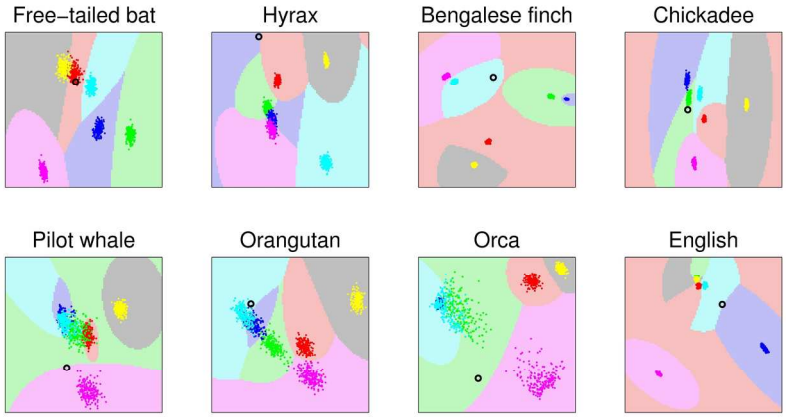


Figure 2. Location of the simulated sequences and original data (black circle) in 2-dimensional Levenshtein distance space. Coloured points indicate only the first 30 randomly generated sequences from each model, for clarity: ZOMP (red), FOMP (green), SOMP (blue), HMM (cyan), RP (magenta), and PHM (yellow). Solid colours indicate the domains of the naïve Bayesian classifier for each model type.
188x89mm (300 x 300 DPI)

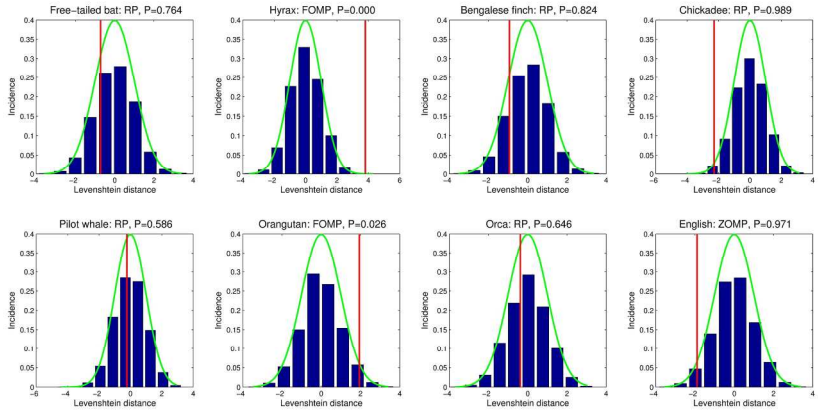


Figure 3. Histograms of the Levenshtein distances of simulated sequences from each other (blue bars) for the best fit model (indicated in the title of each panel), and the fitted normal distribution (green line) using the Matlab normfit function. The red line shows the mean Levenshtein distance of the original data from the simulated sequences, and the P value indicates the probability of this mean distance (or greater) having been drawn from the distribution of simulated sequences (Z-test).
218x98mm (300 x 300 DPI)

Table 1. Summary of the data sets used and their characteristics.

Species	Number of element types	Number of sequences	Total sequence length	Source
Free-tailed bat *Tadarida brasiliensis*	3	71	514	(21)
Hyrax *Procavia capensis*	5	263	3296	(2)
Bengalese finch *Lonchura striata domestica*	7	2130	27858	(32)
Chickadee *Poecile carolinensis*	7	4246	37094	(33, 34)
Pilot whale *Globicephala macrorhynchus*	20	18	246	(15)
Orangutan *Pongo spp.*	7	32	373	(31)
Killer whale *Orcinus orca*	5	8	224	(52)
English language	25	455	3816	(36)

Table 2. Results of the Bayesian classifier to find the best fit model to the observed data. Embedding dimension shows the number of multidimensional scaling dimensions used for the classification, and $P<0.05$ in the final column indicated by (*) shows that the Z-test rejects the hypothesis that the data belong to the best fit model.

Species	Best fit model	Embedding dimension	*P*
Free-tailed bat	RP	5	0.764
Hyrax	FOMP	7	<0.001*
Bengalese finch	RP	7	0.824
Chickadee	RP	8	0.989
Pilot whale	RP	14	0.586
Orangutan	FOMP	9	0.026*
Killer whale	RP	14	0.646
English	ZOMP	6	0.914