Quantifying similarity in animal vocal sequences: which metric performs best?

Arik Kershenbaum[1*], Ellen C. Garland[2]

[1]Department of Zoology, University of Cambridge, Cambridge, England

[2]School of Biology, University of St. Andrews, St. Andrews, Fife, KY16 9TH, Scotland

*Author for correspondence: arik.kershenbaum@gmail.com

Short title: Measuring sequence similarity

Word count: 7431

SUMMARY

1. Many animals communicate using sequences of discrete acoustic elements which can be complex, vary in their degree of stereotypy, and are potentially open-ended. Variation in sequences can provide important ecological, behavioural, or evolutionary information about the structure and connectivity of populations, mechanisms for vocal cultural evolution, and the underlying drivers responsible for these processes. Various mathematical techniques have been used to form a realistic approximation of sequence similarity for such tasks.
2. Here, we use both simulated and empirical datasets from animal vocal sequences (rock hyrax, *Procavia capensis*; humpback whale, *Megaptera novaeangliae*; bottlenose dolphin, *Tursiops truncatus*; and Carolina chickadee, *Poecile carolinensis*) to test which of eight sequence analysis metrics are more likely to reconstruct the information encoded in the sequences, and to test the fidelity of estimation of model parameters, when the sequences are assumed to conform to particular statistical models.
3. Results from the simulated data indicated that multiple metrics were equally successful in reconstructing the information encoded in the sequences of simulated individuals (Markov chains, n-gram models, repeat distribution, and edit distance), and data generated by different stochastic processes (entropy rate and n-grams). However, the string edit (Levenshtein) distance performed consistently and significantly better than all other tested metrics (including entropy, Markov chains, n-grams, mutual information) for all empirical datasets, despite being less commonly used in the field of animal acoustic communication.
4. The Levenshtein distance metric provides a robust analytical approach that should be considered in the comparison of animal acoustic sequences in preference to other commonly employed techniques (such as Markov chains, hidden Markov models, or Shannon entropy). The recent discovery that non-Markovian vocal sequences may be more common in animal communication than previously thought, provides a rich area for future research that requires non-Markovian based analysis techniques to investigate animal grammars and potentially the origin of human language.

42 Keywords: Sequence, animal communication, vocal, edit distance, Markov, stochastic processes

INTRODUCTION

Many animals communicate using sequences of discrete acoustic elements, the best known example being bird song, which is composed of multiple notes combined in a distinctive order. These sequences are often complex, non-stereotyped, and potentially open-ended; that is, individuals may use an almost unlimited repertoire of sequences by making subtle or large variations to the order of notes (reviewed in Catchpole & Slater 2003). The role of such sequences varies among species. In some cases, sequences appear to advertise male quality through sequence complexity, e.g., in marsh warblers, *Acrocephalus palustris* (Darolová *et al.* 2012); zebra finches, *Taeniopygia guttata* (Holveck *et al.* 2008; Neubauer 1999; Searcy & Andersson 1986); and song sparrows, *Melospiza melodia* (Pfaff *et al.* 2007). In other cases, researchers have proposed that sequences contain detailed communicative information such as individual identity, e.g., bottlenose dolphins, *Tursiops truncatus* (Sayigh *et al.* 1999). It is also possible that in some species, acoustic sequences are essentially stochastic with little significance to their precise composition.

Identifying the role of acoustic sequences in a particular species often involves comparing sequences within and between individuals, as well as within and between populations, so that the nature of the variation can be quantified and potentially correlated to ecological or behavioural factors. The task of comparing acoustic sequences presumes an unequivocal and globally relevant measure of sequence similarity, or difference. However, in practice, no such metric exists. It could be postulated that a measure of sequence similarity should reflect the proximal processes taking place in the brains of intended conspecific signal receivers; i.e., the best measure of sequence similarity is the one used by the animal itself (Kershenbaum *et al.* 2014). Given that such knowledge is essentially hidden in practice, various mathematical techniques have been used to form a realistic approximation of signal similarity (Ashby & Perrin 1988; Navarro 2001; Ranjard 2010; Young & Hamer 1994). It is possible to categorise similarity measures into two distinct approaches. Firstly, it is usually possible to characterise a sequence by measuring a small number of metrics that are inherent to the sequence itself; examples of this include length, or entropy (Freeberg & Lucas 2012). Sequences can then be compared by calculating the sum of square differences between each of these metrics. This is

equivalent to representing each sequence as a “feature vector” in some relatively compact feature space, and measuring the distance between two sequences as the Euclidean distance between their two feature vectors. While this method is straightforward, there is an assumption that it is possible to represent every sequence in a compact way, i.e., that some sufficiently large combination of metrics can "summarise" the properties of a sequence in a biologically meaningful way. However, it is far from clear that there exists a compact, yet exact, mathematical representation of a sequence, short of the trivial task of writing down the entire sequence of elements and attempting to measure the Euclidean distance between the full representations of two sequences, which is unlikely to produce the desired results. An alternative approach is to use aggregate techniques that measure properties of a large number of sequences, and summarise the characteristics of a corpus. For example, sequence transition tables and element frequency histograms have been used in previous studies (Jin & Kozhevnikov 2011). In these cases, each vector in feature space represents a collection of sequences, and the Euclidean distance between vectors measures the difference between the sequences from two sets of vocalisations, rather than between individual sequences. However, it is questionable whether any of these techniques, individual or aggregate, can represent the nature of the sequences with adequate fidelity. Since we do not know what cognitive processes an animal uses to interpret such sequences, we cannot be sure that any particular summary metric accurately reflects the interpretation of the sequence by the receiving individual. We refer to all of these above metrics as “unary”, as they are derived from measurements on each string sequence in isolation, even if distances are eventually calculated on an aggregate of sequences.

Secondly, it is possible to measure the difference between a pair of sequences directly (Levenshtein 1966), thereby bypassing the construction of a feature space, and generating a series of pairwise comparisons between sequences. Analysing the sequence of elements in animal vocalisations can be considered analogous to analysing the sequence of nucleotides in DNA, and some non-aggregate techniques have been borrowed from the field of bioinformatics to capture the similarity or difference between two sequences. This approach provides a direct measure of pairwise differences, in the form of a distance matrix, but without a Euclidean feature space. We refer to these metrics as “binary”, as

they can only be calculated as a pairwise comparison between exactly two sequences. Binary difference measures are attractive, as they do not rely on the fidelity of a particular unary metric in representing the properties of a sequence. Rather, binary metrics are an unequivocal measure of the similarity/difference between two sequences; although it cannot be assumed that this measure of similarity is the same as that used by the animal itself in distinguishing between sequences. Such metrics have long been proposed for the analysis of birdsong (Bradley & Bradly 1983; Ranjard *et al.* 2010), but have not been widely adopted. One disadvantage of binary metrics is that a number of common machine learning algorithms often used for clustering the results of similarity analyses (e.g., k-means, neural networks) rely on data presented as a Euclidean feature space, although there are exceptions, e.g. Ranjard & Ross (2008). To use such clustering techniques, it would be necessary to derive a series of feature vectors from the binary metric distance matrix. This can be done using techniques such as multidimensional scaling or principal component analysis to convert a distance matrix to feature vectors.

Here, we compare the performance of eight different methods for analysing animal vocal sequences, using both aggregate statistical metrics and a direct pairwise distance measure. We use simulated and empirical sequences to test which approach is more likely to reconstruct the information encoded in the sequences, and to test the fidelity of estimation of model parameters when the sequences are assumed to conform to particular statistical models. This direct comparison of a number of commonly employed analytical algorithms provides a comprehensive evaluation of the utility of these approaches to real-world data sets, and demonstrates the utility of comparing at least two different methods when assessing novel algorithms to ensure that results are robust under a range of analytical approaches.

METHODS

We performed two sets of tests (viz. artificial and empirical) to evaluate the performance of each metric. In the first tests, we generated artificial random sequences and used the different similarity

metrics to reconstruct the parameters used to generate these sequences, and the stochastic model types. In the second set of tests, we analysed recordings of animal vocalisations and used both unary and binary difference metrics to determine contextual information known to exist in these sequences. We used the signature whistles of the bottlenose dolphin (Kershenbaum, Sayigh & Janik 2013; Sayigh *et al.* 2007; Sayigh *et al.* 2007), to reconstruct individual identity, and the songs of the rock hyrax, *Procavia capensis* (Kershenbaum *et al.* 2012), the humpback whale, *Megaptera novaeangliae* (Garland *et al.* 2012), and the calls of the Carolina chickadee, *Poecile carolinensis* (Freeberg 2012), to reconstruct geographical dialect. In the case of the hyrax, humpback whale, and chickadee, the calls consisted of a sequence of discrete acoustic elements. In contrast, bottlenose dolphin whistles are often produced in isolation (rather than as a sequence of whistles); therefore we analysed the sequence of frequency modulation components (e.g., up, down, constant) within whistles, taking these modulation components as the acoustic elements (for more details see Kershenbaum, Sayigh & Janik 2013). In both our analysis of artificial sequences, and empirical animal vocal sequences, we evaluate a number of similarity metrics, both binary and unary. Before providing details of the simulation experiments and empirical data analysis, we describe each of the metrics used.

Binary metric

Levenshtein distance (LD)

The Levenshtein distance (Levenshtein 1966) is a type of *string edit distance metric*, as it provides a quantitative measurement of the difference between two string sequences regardless of string length. Specifically, the Levenshtein distance measures the minimum number of point operations (additions, deletions, and substitutions) needed to convert one string into another (Levenshtein 1966). By comparing the position of elements within a string and calculating the number of changes that it takes to change one string into the other, this metric relies more on the sequence of elements and less on the overall structural pattern. It has been used extensively in other fields, e.g., bioinformatics (Likic 2008) and text search/retrieve (Reis *et al.* 2004), and in a small number of previous studies of animal

sequences (e.g., Garland *et al.* 2012; Garland *et al.* 2013; Kershenbaum *et al.* 2012; Krull *et al.* 2012), and is related to the better known dynamic time warping algorithm (Buck & Tyack 1993). However, LD itself remains somewhat unknown in the field of animal acoustic communication. In practice, string edit distances are often paired with string alignment algorithms or additional standardisations, particularly when the strings being compared are of different lengths: Figure 1; see Kershenbaum *et al.* (2012) and Garland *et al.* (2012) for additional information on metric calculation. Importantly, the Levenshtein distance forms the basis of the Needleman-Wunsch string alignment (Likic 2008; Needleman & Wunsch 1970) that is used extensively in bioinformatics research to compare sections of DNA. In our implementation of the LD algorithm, we assign an equal cost (of 1) to any correction operation (addition, deletion, substitution), no cost (0) for a matching element, and no cost for differences in sequence lengths after optimal alignment.

Although other binary metrics exist apart from LD, they are in general unsuitable for the task at hand. For example, the Hamming distance requires sequences of the same length, and the most frequent *k* characters simply provides a count of the most common symbol/element. These therefore provide less information than the Levenshtein distance metric.

Unary metrics

Transition table (TT)

Acoustic sequences have often been modelled as a Markov chain (Berwick *et al.* 2011; Briefer *et al.* 2010; Briefer *et al.* 2010), in which the probability of a particular element occurring depends only on the preceding element (or sometimes, more than one preceding element). These conditional probabilities of each element, given the preceding element(s), can be expressed as a transition matrix *T*, in which the element $T_{i,j}$ represents the probability of the element *j* occurring after the element *i*. For a sequence consisting of *C* distinct element types, a *C* x *C* transition matrix can be estimated from empirical data. When comparing two sequences A and B, the similarity between the transition matrices T_A and T_B is an indication of the similarity between the sequences (Jin & Kozhevnikov

2011). To calculate a difference metric $D_{TT} = f(T_A,T_B)$, we can express each matrix as a C^2 dimensional feature vector V, where the elements of the vector are equal to the elements of the transition matrix T, i.e., $V = T(\cdot)$. We then calculate the Euclidean distance between the two vectors derived from sequences A and B:

$$D_{TT}(A,B) = \sqrt{\sum (V_A - V_B)^2}$$

However, such a metric would not be expected to produce a meaningful measure for sequences composed of non-overlapping element types (e.g. ABCABC, and DEFDEF). Therefore we sort vectors V_A and V_B in order of transition probability before comparison. This allows a comparison of transition probability distributions, independent of element type.

N-gram distribution (NG)

Researchers have previously proposed that an important property of animal sequences is the nature of repeating units within the sequence (Cane 1959; Kershenbaum *et al.* 2014; Pruscha & Maurus 1979). A sequence of length L consists of L-n+1 sub-sequences of length n. Thus, the five-element sequence ABBAC consists of 5-2+1=4 two-element sub-sequences: AB, BB, BA, and AC. For a sequence consisting of C distinct element types, there are a total of C^n distinct n-element possible sub-sequences. The vector of sub-sequence frequencies, $P(i \in C^n)$ can be considered a feature vector, and the distance between two strings calculated in a similar way to that shown above:

$$D_{NG}(A,B) = \sqrt{\sum (P_A - P_B)^2}$$

In the following analyses, we chose the n-gram distribution for n = 3, as this provides a good balance between coverage and diversity. For a comparison of different length n-grams in analysing birdsong, see Jin & Kozhevnikov (2011).

Shannon entropy (SE)

Information theory approaches to analysing animal vocal communication have become popular in recent years. One metric that is simple to understand and easy to apply is the Shannon entropy (Shannon *et al.* 1949), and this has been used in a number of studies to measure the complexity of animal vocal sequences (Da Silva, Piqueira & Vielliard 2000; Doyle *et al.* 2008; McCowan, Hanser & Doyle 1999; McCowan, Hanser & Doyle 1999; Suzuki, Buck & Tyack 2006). Shannon entropy measures the unpredictability of a sequence, or the lack of uniformity of a sequence, so that a completely predictable sequence (e.g., consisting of the same element repeated over and over) would have an entropy of zero, whereas a completely unpredictable (random) sequence would have an entropy of one. The equation for Shannon entropy H is as follows:

$$H = -\sum_{i \in 1 \ldots C} P_i \log_C P_i$$

where P_i is the probability of element i, drawn from a set of the C elements occurring in the union of all sequences.

Our SE metric compares two sequences by taking the ratio of the Shannon entropies of the sequences A and B:

$$D_{SE}(A, B) = H_A / H_B \text{ where } H_A < H_B$$

Although SE is calculated as a single comparison between single measurements on two sequences (in contrast to the TT and NG metrics described above, both of which result in multiple measurements on a single sequence), SE should still be considered a unary metric, because it does not directly measure the distance between two sequences, but rather the difference in a derived metric from each.

Entropy rate (ER)

Entropy rate has been shown to be a useful metric for measuring vocal sequence complexity (Kershenbaum 2013). Entropy rate is derived from the transition table of a sequence, and can be

thought of as a measure of transition table diversity, i.e., the extent to which different transitions between notes are of uniform or non-uniform probability. Given a transition table $T_{i,j}$ as described above, entropy rate ER is defined as:

$$ER = -\sum_{i \in 1 \ldots C} \pi_i \sum_{j \in 1 \ldots C} T_{i,j} \log T_{i,j}$$

where π_i is the stationary probability of element *i*, i.e., the overall probability of *i* occurring in the sequence; see Kershenbaum (2013) for additional information on metric calculation. As with Shannon entropy, we define a metric D_{ER} for the difference between sequences A and B:

$$D_{ER}(A,B) = ER_A / ER_B \text{ where } ER_A < ER_B$$

Repeat distribution (RD)

The repeat number distribution was used in a recent study to compare the similarity between natural and synthetic songs of Bengalese finches, *Lonchura striata var. domestica* (Jin & Kozhevnikov 2011). It is an aggregate measure, calculated on a corpus of sequences. For each set of sequences a histogram is generated showing the probabilities P_n that any element occurred in isolation (n = 1), was repeated twice (n = 2), three times (n = 3), and so on. As with the n-gram distribution, we define a metric that measures the difference between two such histograms, generated from sequences *A* and *B*, where P_A and P_B are the feature vectors of sequences *A* and *B*, comprising the repeat distributions for all the elements:

$$D_{RD}(A,B) = \sqrt{\sum (P_A - P_B)^2}$$

Mutual information (MI)

Mutual information is an information theory measure that can be applied easily to quantify the similarity of two sequences. MI combines both measures of the inherent complexity in a sequence (via Shannon entropy), and the joint entropy of the sequences, which measures the probability that a particular pair of elements will occur at the same point in two sequences; see Kershenbaum *et al.* (2012) for additional information on metric calculation. MI is defined as follows:

$$MI = H(A) + H(B) - \sum_i \sum_j p_{i,j} \log p_{i,j}$$

where *H(A)* is the Shannon entropy of sequence A, *H(B)* is the Shannon entropy of sequence B, and $p_{i,j}$ is the probability that elements *i* and *j* occur at the same point in sequences A and B. As with Shannon entropy, we define a metric D_{MI} for the difference between sequences A and B:

$$D_{MI} = MI_A / MI_B \text{ where } MI_A < MI_B$$

Lempel-Ziv (LZ)

The Lempel-Ziv complexity (Lempel & Ziv 1976) is an important algorithm used for data compression, as it is a measure of the number of distinct patterns in a sequence. As a metric of sequence complexity and an approximation to Kolmogorov complexity (Evans & Barnett 2002), it is potentially a useful indicator of the diversity of an animal vocal sequence. Although it has not been widely used in animal studies, Suzuki, Buck & Tyack (2006) suggested the use of the LZ metric for the analysis of humpback whale song, and Kershenbaum (2013) showed that the LZ metric outperformed Shannon entropy (SE) in quantifying realistic length acoustic sequences. LZ complexity was calculated using the Applied Nonlinear Time Series Analysis library for Matlab (Small 2005).

$$LZ = \frac{c \log L}{L \log K}$$

where *c* is the number of distinct substrings in a sequence of length *L*, and *K* is the maximum number of possible distinct substrings.

Sequences for analysis

Artificial sequences

In the first test, we evaluated the utility of each of the similarity metrics by their ability to identify correctly the stochastic process model from which artificial sequences were generated. We generated artificial sequences using three different stochastic processes, often used to model animal vocal sequences (Kershenbaum *et al.* 2014) : the zero-order Markov process (ZOMP), the first-order Markov process (FOMP), and the semi-Markov renewal process (RP). The ZOMP is an independent stochastic process, in which the probability of any particular element occurring at a particular point in a sequence is determined solely by the prior probability of that element. In the FOMP, element probabilities are determined by a transition table, where the probability of a particular element depends on the immediately preceding element. The RP has been shown to be a more realistic model of animal vocal sequence production (Kershenbaum *et al.* 2014) in which the number of repeated elements is drawn from a Poisson distribution, rather than being determined by the diagonal of a transition table. In each case, we examined 10 sequences of 10 elements each, drawn from five possible elements (A-E). We generated 30 sequences, 10 from each of the stochastic processes, ZOMP, FOMP, and RP. The ZOMP was modelled by selecting five random prior probabilities, one for each element type, and renormalising to sum to unity. We then generated the sequences by selecting elements according to these prior probabilities. The FOMP was modelled by generating a random 5 x 5 transition table in a similar way to the ZOMP prior probabilities, so that the rows of the transition matrix summed to unity. A random initial element was chosen for each 10-element sequence, and the remaining nine elements in each sequence were chosen randomly according to the probabilities in the transition table. The RP was modelled in a similar way to the FOMP, except that for each element generated, a random number of repeats were drawn from a Poisson distribution with mean five (to give 95% confidence of ≤ 9 repeats). Having generated 30 sequences of 10 elements, we then calculated a 30 x 30 distance matrix for each of the similarity metrics. We then used an Adaptive Resonance Theory (ART) artificial neural network to cluster these 30 points into natural groupings, setting a maximum of 100 possible clusters. ART networks have been used in a number of previous

studies to cluster data derived from animal vocalisations (Deecke & Janik 2006; Janik 1999; Quick & Janik 2012). We then calculated the normalised mutual information (NMI) as a metric of goodness of clustering (Zhong & Ghosh 2005), by comparing the composition of the generated clusters $H(Y)$ with the true generation process of each $H(\hat{Y})$. Thus, NMI indicates the proportion of uncertainty predicted by the metric. We then repeated this process 100 times using new random transition matrices, generating 3000 sequences in total.

In the second test using artificial sequences, we simulated "individuals" by generating 100 random RP transition matrices, and from each of them producing a set of 10 sequences of 10 elements each. We used the RP generation process, rather than a Markovian ZOMP or FOMP, as the RP more reliably describes many types of animal vocal sequences (Kershenbaum *et al.* 2014). Each sequence generated from a single transition matrix would be expected to be more similar to other sequences from the same transition matrix, than sequences generated by a different random transition matrix, therefore we used a similar clustering approach as in the stochastic process analysis above. We calculated the 100 x 100 distance matrix for each similarity metric, obtained by comparing the sequences from each of the 100 transition matrices, and clustered the results as before, measuring the NMI as an indication of clustering success.

For a final test using artificial sequences, we examined the effect of typical sample sizes (number of sequences) on each of the similarity metrics. Using the sequences generated in the individual simulation above, we varied the number of sequences analysed from one to ten, recalculated the distance matrices and clustering, and measured the NMI.

Animal sequences

We tested the performance of the above metrics using empirical sequences of animal vocalisations, where those sequences are thought to contain information that is known *a priori*. Very few examples exist where contextual information is objectively known to exist in animal vocal sequences. However, the signature whistles of bottlenose dolphins have been shown to encode individual identity in the

sequence of up-down frequency shifts, known as a Parsons code (Kershenbaum, Sayigh & Janik 2013). We used a data set consisting of 400 signature whistles, 20 from each of 20 individual dolphins, recorded during capture-release events; see Sayigh *et al.* (2007) and Kershenbaum, Sayigh & Janik (2013) for additional details. We converted each whistle into a 9-element Parsons code, with seven possible element values (‘‘large drop’’, ‘‘medium drop’’, ‘‘small drop’’, ‘‘no change’’, ‘‘small rise’’, ‘‘medium rise’’, and ‘‘large rise’’). We then calculated distance matrices using each of the similarity metrics described above, and clustered using an ART network. For the calculation of NMI, we compared the generated clusters to the known clusters of individual identity. As empirical data do not allow the generation of unlimited data sets as with artificial sequences, we estimated confidence intervals for each of the empirical data sets by randomly selecting 80% of the calls for clustering and calculation of NMI, and repeated this process 100 times.

We analysed three further empirical data sets for which contextual information in vocal sequences has been proposed. The first data set used recordings of humpback whales (for details see Garland *et al.* 2012), the second data set used recordings of rock hyraxes (see Kershenbaum *et al.* 2012), and the third set Carolina chickadees (see Freeberg 2012). Previous studies have shown that in the humpback whale, rock hyrax, and Carolina chickadee, song syntax varies according to the geographical origin of the population. For example, not only does chickadee song syntax vary between locations, but there appear to be different functional use of certain sequences in the different populations (Freeberg 2012).

The humpback whale data set consisted of 202 songs composed of 20 different element types (themes), recorded from 42 individuals. Humpback whale song is a complex, stereotyped, repetitive, long, male display that has multiple levels of hierarchy in its organisation (Herman & Tavolga 1980; Payne & Payne 1985; Payne & McVay 1971). A few sounds (units) are arranged in a stereotyped phrase which is repeated multiple times to make a theme (Payne & McVay 1971). A number of themes, sung in a particular order, are combined to form a song. The order and content of the themes are highly stereotyped, and all males within a population adhere to the same arrangement and content of the song at any given time as the display is constantly changing (Frumhoff 1983; Payne, Tyack & Payne 1983; Payne & Payne 1985). This analysis focused on the theme level in the hierarchical

arrangement of humpback whale song. Each string therefore represented the sequence of themes (elements) that comprised a song; e.g., theme 1, theme 2, theme 3, theme 4, theme 5; see Garland *et al.* (2012) for further information and example sequences. This level within the hierarchy takes into account information on the sequence of units and the repetition of phrases at a higher level, but does not examine these lower levels explicitly. Strings were classified according to their geographical location: New Caledonia, Vanuatu, or eastern Australia, and this geographical origin was compared to the clusters generated by the ART network. Humpback whale song is constantly changing, and has been shown to undergo complete song revolutions in this region (Garland *et al.* 2011; Garland *et al.* 2011; Noad *et al.* 2000). The current analysis incorporates two different song types (lineages) that contain different themes (vocabulary), and are present in these populations at various points over the four years of recording. Therefore, each metric must be robust to the underlying transmission dynamics of this display.

The hyrax data consisted of 1130 song sequences composed of five different element types, recorded from a single individual at each of 18 different locations in Israel. The Carolina chickadee data consisted of 1184 sequences of calls, recorded from 60 sites in the states of Tennessee and Indiana, USA. Links to these data sets are available in the supplemental information.

RESULTS

Artificial sequences

For sequences generated by different stochastic processes, the entropy rate (ER) metric provided the best clustering, with a NMI value of 0.518 ± 0.005 (standard error) (Figure 2a), while the binary Levenshtein distance (LD) metric gave a NMI of 0.476 ± 0.006. A post-hoc Tukey test following ANOVA showed significant differences between the NMI scores of these two metrics. All other metrics produced significantly lower NMI values.

Results from clustering sequences of simulated "individuals" (sequences generated by stochastic processes with similar parameters), indicated that NG produced the highest NMI score 0.751 ± 0.001, while the LD, RD, and TT metrics all produced high but slightly lower NMI scores (greater than 0.7; Figure 2b), with no significant differences among the NMI values of these three metrics.

Both the LD and NG metrics that performed well on the above clustering tasks were also robust to sample size (Figure 3). Most other metrics were also relatively unaffected by sample size. However, the RD performed poorly at smaller sample sizes (≤ 4), and the MI declined with increasing corpus size (> 2).

Animal sequences

When clustering to reconstruct the individual identity from bottlenose dolphin signature whistles, the Levenshtein distance (LD) performed significantly better than all other tested metrics, with an NMI of 0.661 ± 0.001 (Figure 4a). The n-gram distribution (NG) also performed well, with an NMI of 0.63p ± 0.001. Clustering of the humpback whale song data to indicate population (geographic) origin, showed the LD again performed significantly better than all other tested metrics (NMI of 0.491 ± 0.005; Figure 4b). The NG provided the second best, although significantly poorer, metric (NMI of 0.367 ± 0.005). All metrics performed poorly in clustering the geographical origin of hyrax songs; however, the LD metric was again significantly better than all others tested (NMI 0.1684 ± 0.001, compared to the next best NMI of 0.130 ± 0.001 for TT; Figure 4c). Clustering of the chickadee data to distinguish between birds recorded in Tennessee and those recorded in Indiana, showed the LD performed significantly better than all other metrics (NMI of 0.450 ± 0.001; Figure 4d), followed by NG (NMI 0.369 ± 0.001).

DISCUSSION

We analysed the performance of eight different techniques from two broad approaches, to investigate the utility of each approach in the comparison of animal sequences. The unary and binary metrics performed similarly well in the artificial sequence tests, with the entropy rate (ER) metric slightly better than the Levenshtein distance binary metric (LD), in distinguishing between data generated by different stochastic processes, and n-gram (NG) slightly better in distinguishing simulated individuals. However, the LD metric performed significantly better than all other tested metrics when presented with empirical animal sequences. This result emphasises that caution should be used when using artificially generated sequences based on simple stochastic models to simulate animal vocal sequences. Recent work has shown that assumptions of simple models for animal vocal production are likely to be inaccurate (Kershenbaum *et al.* 2014), and similar conclusions have been indicated for cetacean song (Miksis-Olds *et al.* 2008). The difference between metric performance on artificial and on empirical data is striking. Little is known of the cognitive mechanisms by which animals encode and decode information in vocalisations (Thornton, Clayton & Grodzinski 2012); researchers must rely on isolated examples where information content is known *a priori* to draw conclusions about which analytical techniques are best suited for vocal sequence data. Our results clearly show that the LD metric outperforms other metrics on empirical data, despite performing less effectively on simulated data. This indicates that the sequential order of the sequences varied across location/individual while the level of complexity is similar. The Levenshtein distance was the metric of choice for clustering dolphin signature whistles into individuals, humpback whale song into populations, hyrax songs into geographical region, and chickadee calls into state of origin. Analysis of the sensitivity of the different metrics to sample size showed that most of the metrics that performed well across the data sets (LD, NG, LZ), were also robust to sample size.

Results from the current paper in combination with previous work (Eriksen *et al.* 2005; Garland *et al.* 2012; Garland *et al.* 2013; Helweg *et al.* 1998; Tougaard & Eriksen 2006), highlight the success of the Levenshtein distance (LD) metric in the analysis of sequence content and comparison of humpback whale song. A large body of work has previously shown that song differences among humpback whale populations can indicate geographic origin of a singer (e.g., Garland *et al.* 2015;

Helweg *et al.* 1998; Payne & Guinee 1983). Despite dynamic song transmission in the South Pacific region, fine-scale song differences allow the identification of population origin (Garland *et al.* 2011; Garland *et al.* 2012; Garland *et al.* 2013; Garland *et al.* 2015). The current paper examined the theme sequences (i.e., a set of phrases under a single label) as part of the largest analysis to date of sequence comparison algorithms for humpback whale song (Garland *et al.* 2013), which indicated the LD out performed all other tested metrics. We suggest when *comparing* song sequences, the LD metric should be employed preferentially, while if the complexity or information content of each song is the focus of study, the researcher should employ other techniques such as entropy.

Previous studies of sequence comparison in hyrax song (Kershenbaum *et al.* 2012) have shown geographical variation in sequence structure using the LD metric, as these findings were supported by application of an unrelated (unary) metric, mutual information (MI). In the current study, MI performed very poorly on both simulated and empirical data, although MI performance was somewhat better on the hyrax data than on the other data sets. This implies that the aspect of the sequences that is measured by MI does not vary in correlation with geographic location or individual. While not all studies can compare large numbers of analytical algorithms, this emphasises the utility of comparing at least two different techniques when assessing novel algorithms, to ensure that results are robust under a range of analytical approaches.

Despite all tested metrics performing poorly in the assessment of geographic origin in hyrax song, the LD metric was significantly better than all others. In previous work, (Kershenbaum *et al.* 2012) measured the correlation between sequence similarity and the distance between populations, rather than classification success, and the latter suggests that distinct dialects are not present in the hyrax. Rather, small but significant differences are present between all pairs of populations, depending on geographic isolation. In contrast, humpback whales, chickadees, and bottlenose dolphins show strong discrimination between in-group and out-group sequences, indicating that the differences between the vocal sequences of different individuals or populations are much more marked. This may indicate an adaptive role to distinctive vocalisations in dolphins and whales, such as individual identification (Janik & Slater 1998; Janik, Sayigh & Wells 2006; Quick & Janik 2012), while in chickadees

adaptive, developmental, and phylogenetic explanations for regional dialects have been suggested (Freeberg 2012). Humpback whale song is hypothesised to contain information about the reproductive fitness and population origin of the signaller (Helweg *et al.* 1992; Helweg *et al.* 1992; Payne & Guinee 1983). Hyrax song complexity is not thought to contain contextual information beyond male fitness (Demartsev *et al.* 2014; Koren & Geffen 2009), although this assumption is currently untested. In contrast, dolphin signature whistles are known to be individually distinctive whistles that can be identified by the unique pattern of frequency modulations (Janik, Sayigh & Wells 2006). The characterisation of signature whistles based on a 7-element Parsons code in a previous study (Kershenbaum, Sayigh & Janik 2013) allows individual identification of the whistler. The LD significantly outperformed all other models in clustering to reconstruct not only the individual identity from signature whistles, but the geographic origin for humpback whale song, chickadee calls, and hyrax song, highlighting the importance of evaluating different metrics with *a priori* information.

One likely explanation for the higher performance of the LD metric is that it alone among the metrics analysed uses a direct comparison of the vocal sequences between samples, thereby using more information about the sequences than the other metrics. The LD metric by design can solely be employed to *compare* two strings and it excels at this task; it does not provide an understanding of the information content within each string, or the sequence structure. By necessity this means that LD also compares the vocabularies of a pair of sequences, and therefore two sequences that are based on the same set of sequence elements are likely to have a lower LD value than two sequences that are composed of different elements, but have similar sequence structure. Regional differences in the vocabulary (e.g., humpback song themes) provide important information on the connectivity of populations at a broad-scale despite an overall similarity in song structure (hierarchical arrangement). To establish the influence of overlapping vocabulary is beyond the scope of this paper (although two of the three humpback populations switched between two vocabularies – song types – over the course of this study), but we present as supplemental information (Figure S1) the element distributions of the different data sets, which in most cases were quite consistent.

Sample sizes can be constrained in the study of wild animals and particularly in marine mammal studies. Samples may be collected infrequently and with a patchy distribution due to the challenging conditions presented in collecting such data. Understanding how a metric reacts to a small sample size is invaluable in metric choice. The robust nature of the LD and NG to smaller sample sizes and their high performance in the comparison task makes them appealing for analysis. The data presented here indicated that LD and NG performed well with a sample size of three or less, while TT and RD should not be considered as a metric for analysis until a sample size of four or more is available.

Here, we have presented a robust understanding of which metric should be preferentially employed in studies involving the comparison of individual- or group-specific vocalisations, such as signature whistles. The success in identifying individual/geographic variations in vocal sequences has implications for assessing population structure, song transmission, and dialect similarity, particularly for populations where rapid song changes occur. For example, the analysis of humpback whale song presented here was able to identify population origin despite rapid song dynamics (Garland *et al.* 2011; Garland *et al.* 2012; Garland *et al.* 2013). We suggest that the LD can be applied to any level within a complex display, but suggest future studies strive for the lowest level sequence within the hierarchy (i.e., sequence of units or phrases), to increase the amount of information directly compared and thus encapsulated by the sequence.

The LD method provides a metric to compare sequence content and organisation (and thus songs) within and among multiple individuals, populations, years, and locations. In particular, transmission of humpback whale song is largely cultural, and the level and rate of change remains unparalleled in any other non-human animal as complete population-wide changes are replicated in multiple populations at a vast geographic scale (Garland *et al.* 2011). Thus, fundamental questions in animal culture, vocal learning, and cultural evolution can be explored using humpback whale song as a model, and with the help of the LD metric. Further, the evolution of complex vocal labels (i.e., signature whistles) and the underlying cognitive abilities required for such evolution, are extremely important in understanding the evolution of vocal complexity (Janik 2014). Robust metrics that capture the information encoded in the sequences with the highest fidelity are thus required to address

these far-reaching evolutionary questions. We suggest the LD should be utilised in such comparison studies in preference to Markov and information theory based models.

Conclusions

The Levenshtein distance (LD; binary metric) significantly outperformed all other tested metrics in our comparative analysis of animal acoustic sequences. It provides a direct measure of pairwise differences among sequences, instead of a comparison of aggregate similarity. N-grams (Markov chains) were the second most successful metric; the underlying issue that the tested species' vocalisations may be governed by non-Markovian dynamics and the consistent success of the LD metric, suggests n-grams should always be a second choice. Given the inherent interest in the origins of human language and the evolution of signalling complexity, robust and reliable metrics that can capture the content and arrangement of the signal are essential to address these fundamental questions in animal communication and cultural evolution.

Acknowledgements

We thank Melinda Rekdahl, Todd Freeberg and his graduate students, Amiyaal Ilany, Elizabeth Hobson, and Jessica Crance for providing comments of on a previous version of this manuscript. We thank Mike Noad, Melinda Rekdahl, and Claire Garrigue for assistance with humpback whale song collection and initial categorisation of the song, Vincent Janik and Laela Sayigh for assistance with signature whistle collection, Todd Freeberg with chickadee recordings, and Eli Geffen and Amiyaal Ilany for assistance with hyrax song collection and analysis. E.C.G is supported by a Newton International Fellowship. Part of this work was conducted while E.C.G. was supported by a National Research Council (National Academy of Sciences) Postdoctoral Fellowship at the National Marine Mammal Laboratory, AFSC, NMFS, NOAA. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service. We

would also like to thank Randall Wells and the Sarasota Dolphin Research Program for the opportunity to record the Sarasota dolphins, where data were collected under a series of National Marine Fisheries Service Scientific Research Permits issued to Randall Wells. A.K. is supported by the Herchel Smith Postdoctoral Fellowship Fund. Part of this work was conducted while A.K. was a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional support from The University of Tennessee, Knoxville.

DATA ACCESSIBILITY

As authors we do not own all of the data, and we have not been granted permission to archive it.

Hyrax and chickadee data:
http://rspb.royalsocietypublishing.org/highwire/filestream/47311/field_highwire_adjunct_files/1/rspb20141370supp2.xls

Humpback whale song data: Humpback whale song data are held by Dr. Noad (University of Queensland, Australia; mnoad@uq.edu.au) and Dr. Garrigue (Operation Cetaces, New Caledonia; op.cetaces@lagoon.nc). Dolphin whistle data: Dolphin whistle data are held by Dr Sayigh (Woods Hole Oceanographic Institution; lsayigh@whoi.edu) and Dr Janik (University of St Andrews; vj@st-andrews.ac.uk). Please contact each PI directly for access to their recordings.

535 REFERENCES

536

537

538 Ashby, F.G. & Perrin, N.A. (1988). Toward a unified theory of similarity and recognition.
539 *Psychological review*, **95**, 124.

540 Berwick, R.C., Okanoya, K., Beckers, G.J.L. & Bolhuis, J.J. (2011). Songs to syntax: the linguistics
541 of birdsong. *Trends in Cognitive Sciences*, **15**, 113-121.

542 Bradley, D.W. & Bradly, R. (1983). Application of sequence comparison to the study of bird songs.
543 *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison/edited*
544 *by David Sankoff and Joseph B.Krustal*,.

545 Briefer, E., Osiejuk, T.S., Rybak, F. & Aubin, T. (2010). Are bird song complexity and song sharing
546 shaped by habitat structure? An information theory and statistical approach. *Journal of Theoretical*
547 *Biology*, **262**, 151-164.

548 Buck, J.R. & Tyack, P.L. (1993). A quantitative measure of similarity for *Tursiops truncatus*
549 signature whistles. *The Journal of the Acoustical Society of America*, **94**, 2497-2506.

550 Cane, V.R. (1959). Behaviour sequences as semi-Markov chains. *Journal of the Royal Statistical*
551 *Society.Series B (Methodological)*, **21**, 36-58.

552 Catchpole, C.K. & Slater, P.J.B. (2003). *Bird song: biological themes and variations*. Cambridge
553 Univ Press, Cambridge.

554 Da Silva, M.L., Piqueira, J.R.C. & Vielliard, J.M.E. (2000). Using Shannon entropy on measuring the
555 individual variability in the rufous-bellied thrush *Turdus rufiventris* vocal communication. *Journal of*
556 *Theoretical Biology*, **207**, 57-64.

557 Darolová, A., Krištofík, J., Hoi, H. & Wink, M. (2012). Song complexity in male marsh warblers:
558 does it reflect male quality? *Journal of Ornithology*, **153**, 431-439.

559 Deecke, V.B. & Janik, V.M. (2006). Automated categorization of bioacoustic signals: avoiding
560 perceptual pitfalls. *The Journal of the Acoustical Society of America*, **119**, 645-653.

561 Demartsev, V., Kershenbaum, A., Ilany, A., Barocas, A., Ziv, E.B., Koren, L. *et al.* (2014). Male
562 hyraxes increase song complexity and duration in the presence of alert individuals. *Behavioral*
563 *Ecology*.

564 Doyle, L.R., McCowan, B., Hanser, S.F., Chyba, C., Bucci, T. & Blue, J.E. (2008). Applicability of
565 information theory to the quantification of responses to anthropogenic noise by southeast Alaskan
566 humpback whales. *Entropy*, **10**, 33-46.

567 Eriksen, N., Miller, L.A., Tougaard, J. & Helweg, D.A. (2005). Cultural change in the songs of
568 humpback whales (*Megaptera novaeangliae*) from Tonga. *Behaviour*, **142**, 305-328.

569 Evans, S.C. & Barnett, B. (2002). *Network security through conservation of complexity*. Proceedings
570 of IEEE MILCOM 2002.

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

573 Freeberg, T.M. & Lucas, J.R. (2012). Information theoretical approaches to chick-a-dee calls of
574 Carolina chickadees (*Poecile carolinensis*). *Journal of Comparative Psychology*, **126**, 68-81.

575 Frumhoff, P. (1983). Aberrant songs of humpback whales (*Megaptera novaeangliae*): Clues to the
576 structure of humpback songs. *Communication and Behavior of Whales* (ed R. Payne), pp. 81-127.
577 Westview Press, Boulder, Colorado.

578 Garland, E.C., Goldizen, A.W., Lilley, M.S., Rekdahl, M.L., Garrigue, C., Constantine, R. *et al.*
579 (2015). Population structure of humpback whales in the western and central South Pacific Ocean as
580 determined by vocal exchange among populations. *Conservation Biology*,.

581 Garland, E.C., Goldizen, A.W., Rekdahl, M.L., Constantine, R., Garrigue, C., Hauser, N.D. *et al.*
582 (2011). Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale.
583 *Current Biology,* **21**, 687-691.

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

587 Garland, E.C., Noad, M.J., Goldizen, A.W., Lilley, M.S., Rekdahl, M.L., Garrigue, C. *et al.* (2013).
588 Quantifying humpback whale song sequences to understand the dynamics of song exchange at the
589 ocean basin scale. *The Journal of the Acoustical Society of America,* **133**, 560-569.

590 Helweg, D.A., Cato, D.H., Jenkins, P.F., Garrigue, C. & McCauley, R.D. (1998). Geographic
591 variation in South Pacific humpback whale songs. *Behaviour,* **135**, 1-27.

592 Helweg, D.A., Frankel, A.S., Mobley Jr, J.R. & Herman, L.M. (1992). Humpback whale song: Our
593 current understanding. *Marine Mammal Sensory Systems*, pp. 459-483. Springer.

594 Herman, L.M. & Tavolga, W.N. (1980). The communication systems of cetaceans. *Cetacean*
595 *behavior: Mechanisms and functions*,, 149-209.

596 Holveck, M., de Castro, Ana Catarina Vieira, Lachlan, R.F., ten Cate, C. & Riebel, K. (2008).
597 Accuracy of song syntax learning and singing consistency signal early condition in zebra finches.
598 *Behavioral Ecology,* **19**, 1267-1281.

599 Janik, V.M. (1999). Pitfalls in the categorization of behaviour: a comparison of dolphin whistle
600 classification methods. *Animal Behaviour,* **57**, 133-143.

601 Janik, V.M., Sayigh, L. & Wells, R. (2006). Signature whistle shape conveys identity information to
602 bottlenose dolphins. *Proceedings of the National Academy of Sciences,* **103**, 8293-8297.

603 Janik, V.M. & Slater, P.J.B. (1998). Context-specific use suggests that bottlenose dolphin signature
604 whistles are cohesion calls. *Animal Behaviour,* **56**, 829-838.

605 Janik, V.M. (2014). Cetacean vocal learning and communication. *Current Opinion in Neurobiology,*
606 **28**, 60-65.

607 Jin, D.Z. & Kozhevnikov, A.A. (2011). A compact statistical model of the song syntax in Bengalese
608 finch. *PLoS Computational Biology,* **7**, e1001108.

609 Kershenbaum, A. (2013). Entropy rate as a measure of animal vocal complexity. *Bioacoustics,* **23**,
610 195-208.

611 Kershenbaum, A., Ilany, A., Blaustein, L. & Geffen, E. (2012). Syntactic structure and geographical
612 dialects in the songs of male rock hyraxes. *Proceedings of the Royal Society B: Biological Sciences,*
613 **279**, 2974-2981.

614 Kershenbaum, A., Sayigh, L.S. & Janik, V.M. (2013). The encoding of individual identity in dolphin
615 signature whistles: how much information is needed? *PLoS One,* **8**, e77671.

616 Kershenbaum, A., Blumstein, D.T., Roch, M.A., Akçay, Ç, Backus, G., Bee, M.A. *et al.* (2014).
617 Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews,*.

618 Kershenbaum, A., Bowles, A.E., Freeberg, T.M., Jin, D.Z., Lameira, A.R. & Bohn, K. (2014). Animal
619 vocal sequences: not the Markov chains we thought they were. *Proceedings of the Royal Society B:*
620 *Biological Sciences,* **281**, 20141370.

621 Koren, L. & Geffen, E. (2009). Complex call in male rock hyrax (*Procavia capensis*): a multi-
622 information distributing channel. *Behavioral Ecology and Sociobiology,* **63**, 581-590.

623 Krull, C., Ranjard, L., Landers, T., Ismar, S., Matthews, J. & Hauber, M. (2012). Analyses of sex and
624 individual differences in vocalizations of Australasian gannets using a dynamic time warping
625 algorithm. *The Journal of the Acoustical Society of America*, **132**, 1189.

626 Lempel, A. & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on*
627 *Information Theory*, **22**, 75-81.

628 Levenshtein, V.I. (1966). *Binary codes capable of correcting deletions, insertions and reversals*.
629 *Soviet Physics Doklady*, **10**, 707-710

630 Likic, V. (2008). *The Needleman-Wunsch algorithm for sequence alignment*,
631 http://www.ludwig.edu.au/course/lectures2005/likic.pdf edn. Lecture given at the 7th Melbourne
632 Bioinformatics Course, University of Melbourne.

633 McCowan, B., Hanser, S.F. & Doyle, L.R. (1999). Quantitative tools for comparing animal
634 communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal*
635 *Behaviour*, **57**, 409-419.

636 Miksis-Olds, J.L., Buck, J.R., Noad, M.J., Cato, D.H. & Stokes, M.D. (2008). Information theory
637 analysis of Australian humpback whale song. *The Journal of the Acoustical Society of America*, **124**,
638 2385-2393.

639 Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*,
640 **33**, 31-88.

641 Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities
642 in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.

643 Neubauer, R.L. (1999). Super-normal length song preferences of female zebra finches (*Taeniopygia*
644 *guttata*) and a theory of the evolution of bird song. *Evolutionary Ecology*, **13**, 365-380.

645 Noad, M.J., Cato, D.H., Bryden, M., Jenner, M. & Jenner, K.C.S. (2000). Cultural revolution in whale
646 songs. *Nature*, **408**, 537-537.

647 Payne, K. & Payne, R. (1985). Large scale changes over 19 years in songs of humpback whales in
648 Bermuda. *Zeitschrift für Tierpsychologie*, **68**, 89-114.

649 Payne, K., Tyack, P. & Payne, R. (1983). Progressive changes in the songs of humpback whales
650 (*Megaptera novaeangliae*): a detailed analysis of two seasons in Hawaii. *Communication and*
651 *Behavior of Whales*, (ed. R. Payne), Westview Press, Boulder, pp 9-57.

652 Payne, R.S. & McVay, S. (1971). Songs of humpback whales. *Science*, **173**, 585-597.

653 Payne, R. & Guinee, L.N. (1983). Humpback whale (*Megaptera novaeangliae*) songs as an indicator
654 of "stocks". *Communication and Behavior of Whales*, (ed. R. Payne), Westview Press, Boulder, pp,
655 333-358.

656 Pfaff, J.A., Zanette, L., MacDougall-Shackleton, S.A. & MacDougall-Shackleton, E.A. (2007). Song
657 repertoire size varies with HVC volume and is indicative of male quality in song sparrows (*Melospiza*
658 *melodia*). *Proceedings of the Royal Society B: Biological Sciences*, **274**, 2035-2040.

659 Pruscha, H. & Maurus, M. (1979). Analysis of the temporal structure of primate communication.
660 *Behaviour*, **69**, 118-134.

661 Quick, N.J. & Janik, V.M. (2012). Bottlenose dolphins exchange signature whistles when meeting at
662 sea. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 2539-2545.

663 Ranjard, L. (2010). *Computational biology of bird song evolution.* PhD, University of Auckland.

664 Ranjard, L., Anderson, M.G., Rayner, M.J., Payne, R.B., McLean, I., Briskie, J.V. *et al.* (2010).
665 Bioacoustic distances between the begging calls of brood parasites and their host species: a
666 comparison of metrics and techniques. *Behavioral Ecology and Sociobiology*, **64**, 1915-1926.

667 Ranjard, L. & Ross, H.A. (2008). Unsupervised bird song syllable classification using evolving neural
668 networks. *The Journal of the Acoustical Society of America,* **123**, 4358-4368.

669 Reis, D.d.C., Golgher, P.B., Silva, A. & Laender, A. (2004). *Automatic web news extraction using*
670 *tree edit distance*. ACM.

671 Sayigh, L.S., Esch, H.C., Wells, R.S. & Janik, V.M. (2007). Facts about signature whistles of
672 bottlenose dolphins, *Tursiops truncatus*. *Animal Behaviour,* **74**, 1631-1642.

673 Sayigh, L.S., Tyack, P.L., Wells, R.S., Solow, A.R., Scott, M.D. & Irvine, A.B. (1999). Individual
674 recognition in wild bottlenose dolphins: a field test using playback experiments. *Animal Behaviour,*
675 **57**, 41-50.

676 Searcy, W.A. & Andersson, M. (1986). Sexual selection and the evolution of song. *Annual Review of*
677 *Ecology and Systematics,* **17**, 507-533.

678 Shannon, C.E., Weaver, W., Blahut, R.E. & Hajek, B. (1949). *The Mathematical Theory of*
679 *Communication*. University of Illinois Press, Urbana.

680 Small, M. (2005). *Applied nonlinear time series analysis: applications in physics, physiology and*
681 *finance*. World Scientific Publishing Company Incorporated, Singapore.

682 Suzuki, R., Buck, J.R. & Tyack, P.L. (2006). Information entropy of humpback whale songs. *The*
683 *Journal of the Acoustical Society of America,* **119**, 1849-1866.

684 Thornton, A., Clayton, N.S. & Grodzinski, U. (2012). Animal minds: from computation to evolution.
685 *Philosophical Transactions of the Royal Society B: Biological Sciences,* **367**, 2670-2676.

686 Tougaard, J. & Eriksen, N. (2006). Analysing differences among animal songs quantitatively by
687 means of the Levenshtein distance measure. *Behaviour,* **143**, 239-252.

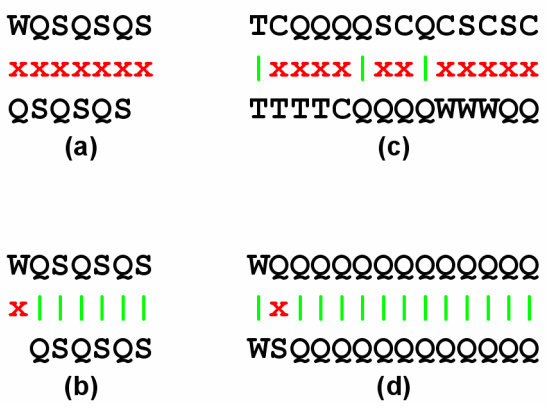
688 Young, F. & Hamer, R. (1994). Theory and applications of multidimensional scaling. *Hillsdale, NJ:*
689 *Eribaum Associates,*.

690 Zhong, S. & Ghosh, J. (2005). Generative model-based document clustering: a comparative study.
691 *Knowledge and Information Systems,* **8**, 374-384.

692

693 FIGURES

694



695

696 Figure 1. Examples of string alignment and edit distance. (a) Two unaligned strings with a LD of 7.
697 (b) After aligning the strings to minimise the difference, LD = 1. (c) Two hyrax bouts which are
698 highly different, LD = 11. (d) Two bouts which are very similar, LD = 1. Reproduced from
699 (Kershenbaum *et al.* 2012).

700

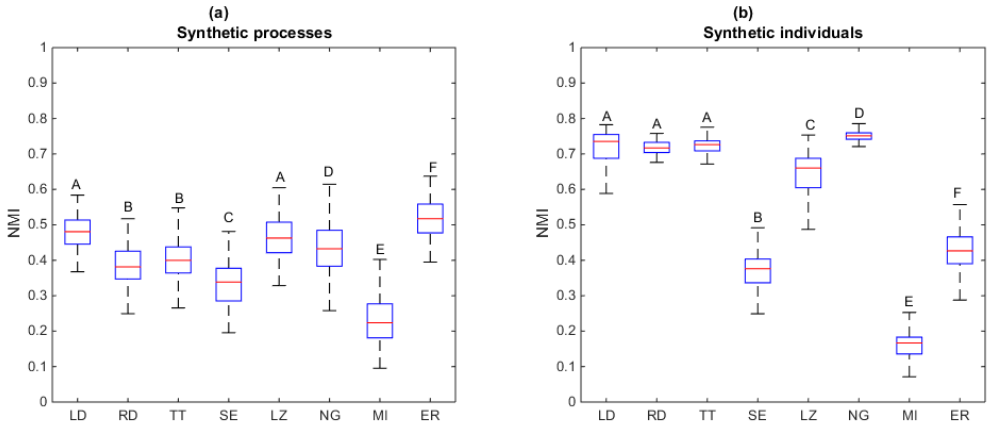


Figure 2. Results of the normalised mutual information (NMI) scores for each metric using a) synthetic processes, and b) synthetic individuals. Metric labels: Levenshtein distance (LD), Repeat distribution (RD), Transition table (TT), Shannon entropy (SE), Lempel-Ziv (LZ), N-gram (NG), Mutual information (MI), and entropy rate (ER). A-F indicate post-hoc Tukey groupings.

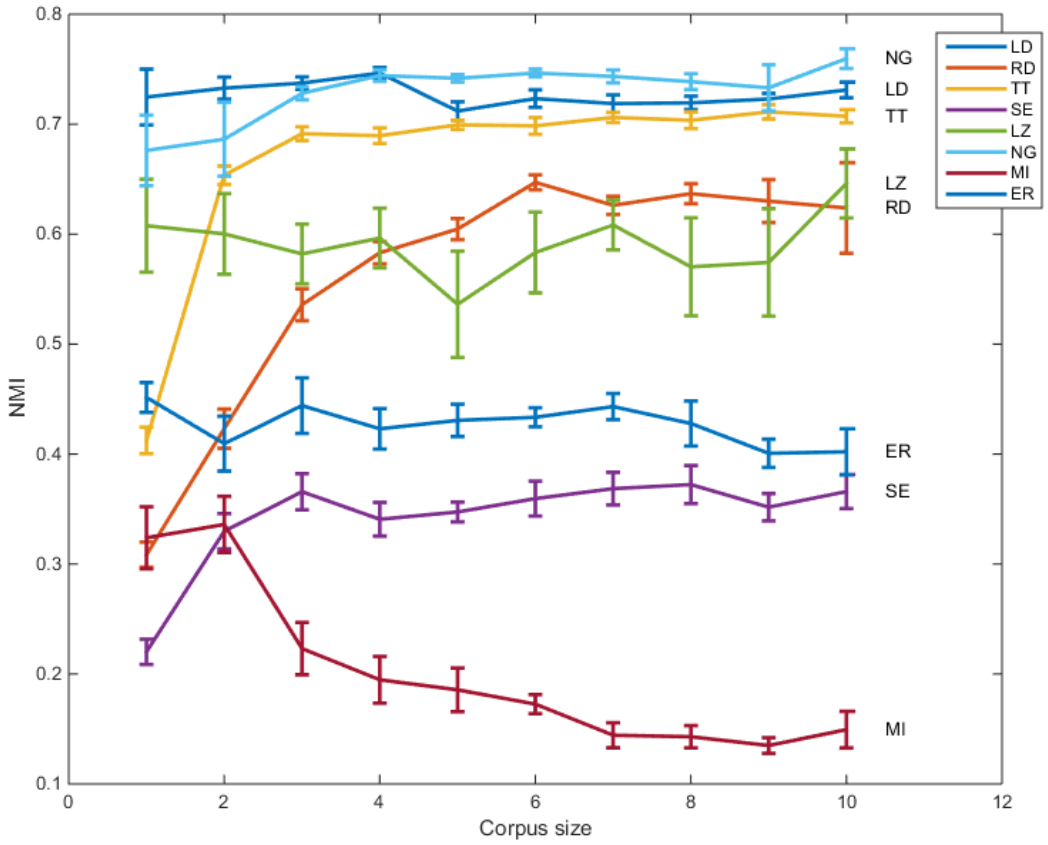


Figure 3. Results of the effect of sample (corpus) size on the NMI scores (± standard error) for each similarity metric. Metric labels are the same as Figure 2.

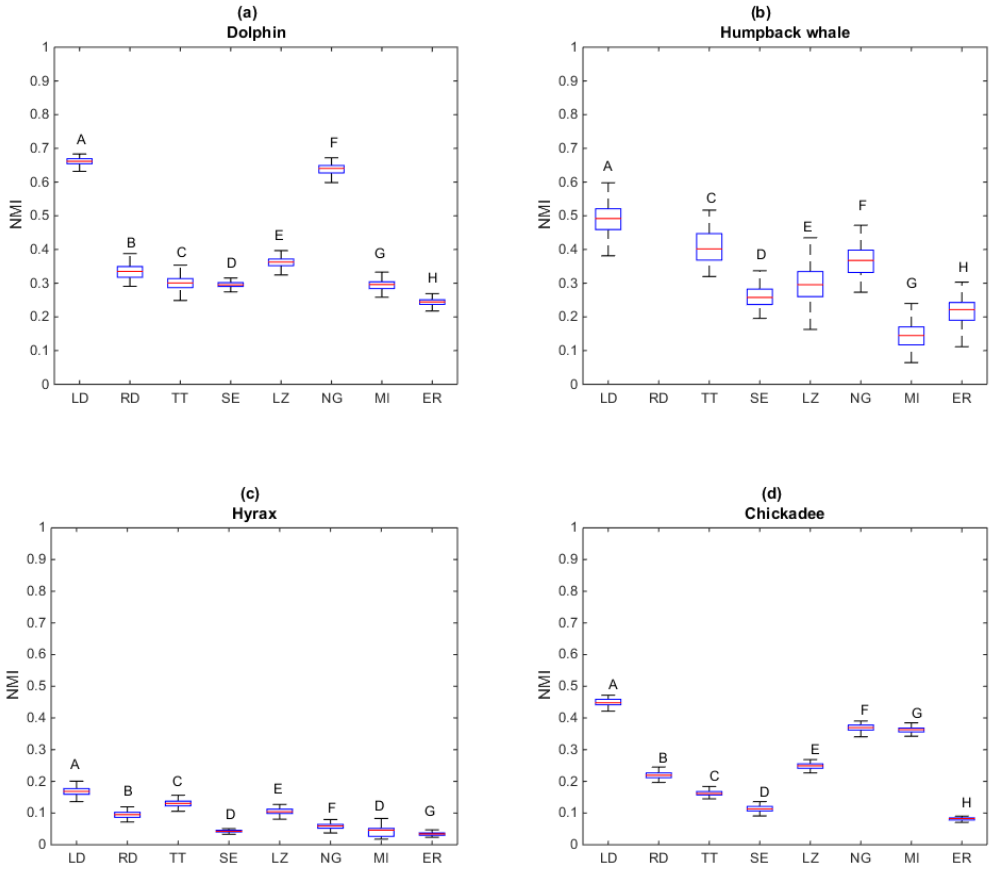


Figure 4. Results of the NMI (normalised mutual information) scores for each metric using a) bottlenose dolphin signature whistles, b) humpback whale songs, c) rock hyrax songs, and d) Carolina chickadee calls. Metric labels are the same as Figure 2. A-F indicate post-hoc Tukey groupings.