piikun: An Information Theoretic Toolkit for Analysis and Visualization of Species Delimitation Metric Space

Jeet Sukumaran[1*] and Marina Meila[2]

[1*]Biology, San Diego State University, Street, City, Postcode, State, Country.
[2]Statistics, University of Washington, Street, City, 10587, State, Country.

*Corresponding author(s). E-mail(s): jsukumaran@sdsu.edu;
Contributing authors: mmp2@uw.edu;

Abstract

Summary piikun is a Python package for analyzing and visualizing species delimitation models in an information theoretic framework that, in addition to classic measures of information such as the entropy and mutual information [1], provides for the calculation of the variation of information criterion [2], a true metric or distance function for species delimitation models that is aligned with the lattice of partitions. The metric as well as other more specialized statistical properties of the variation of information distance statistic calculated here that provide insightful interpretive benefits as well as computational ones when used to understand species delimitation model space.

Availability and implementation piikun is available under the MIT license from its public repository: https://github.com/jeetsukumaran/piikun. piikun is written in pure Python [3], and depends on the following libraries: NumPy [4], SciPy [5], PANDAS [6, 7], plotly [8], Matplotlib [9], Seaborn [10].

Keywords: evolutionary biology, species delimitation, information theory, metrics, distances

1 Introduction

The field of species delimitation has rapidly advanced in a number of ways recently, including machine learning approaches [11], incorporation of natural history [12], hierarchically modeling the speciation process over a primary population-level structuring [13] as well as other biological criteria [14]. Despite these advances, we still lack a true metric space to compare and relate various species delimitation models. Some useful indexes described previously, such as the taxonomic index of congruence (C_{tax}) of [15] or the match ratio of [16] (implemented in [17]), are not true metrics (e.g., having the property of triangle inequality [18]), while others are defined for cases of nested models [15]. These limitations make intuitive interpretation of the distances challenging when more than two species delimitation models are involved, as well as less ideal in supporting inference or optimization algorithms in machine learning [2].

2 `piikun`: Information Theoretic Species Delimitation Metrics

This paper describes `piikun`[1], a Python package for the analysis and visualization of species delimitation models in an information theoretic framework that provides a true distance or metric space for these models. The models being analyzed may be generated by any inference package, such as BPP [20], DELINEATE [13] etc., or taxonomies or classifications based on conceptual descriptions in literature, geography, speculation, etc. Regardless of source or basis, each of these ways to organizae or cluster a set of lineages into a set of higher-level units is a (set theoretic) *partition* of those lineages [13, 21], and can be described in numerous ways that `piikun` can read (e.g., a generic JSON dictionary, or the a species delimitation model data exchange format "`SPART-XML`" [21]). `piikun` further supports specialized input formats, such as the comprehensive results file from DELINEATE [13] or BPP [20] analysis, which allow for incorporation of additional information, such as support values, as shown below.

`piikun` provides a range of univariate information theory statistics for each individual model in the input set (e.g., the entropy [1]), as well as bivariate statistics (e.g., the mutual information, joint entropy, [1]) for each distinct pair of these models, as well as *true* metrics (distances) between every pair of species delimitation models based on these information theoretic measures: the variation of information [2] and the normalized joint variation of information distance [22].

3 The Variation of Information Partition Distance

Every species delimitation model is a *partition* of a set of lineages into a set of mutually exclusive and jointly comprehensive subsets [21]. As such, the *variation of information* criterion of [2], which provides a true distance function for partitions, also establishes a metric space for species delimitation models. Given two partitions, ψ^u, ψ^v, $VI(\psi^u, \psi^v)$, this is defined as [2]:

[1] "sparrowhawk" in the Kumeyaay language [19]. San Diego State University is built on Kumeyaay land.

$$VI(\psi^u, \psi^v) = H(\psi^u) + H(\psi^v) - 2I(\psi^u, \psi^v). \tag{1}$$

$H(\psi)$ is the entropy of partition ψ which divides n elements into K subsets, with the kth subset having n_k elements, and is given by [2]:

$$H(\psi) = -\sum_{k=1}^{K} p(k) \log p(k), \tag{2}$$

where $p(k)$ is the probability of subset or cluster k, which is given in this approach by the cardinality of the subset as a proportion of the entire set: $p(k) = \frac{n_k}{n}$. $I(\psi^u, \psi^v)$, on the other hand, is the mutual information of partitions ψ^u and ψ^v, and is given by:

$$I(\psi^u, \psi^v) = \sum_{k_i=1}^{K_i} \sum_{k_j=1}^{K_j} p(k_i, k_j) \log \frac{p(k_i, k_j)}{p(k_i)p(k_j)}, \tag{3}$$

where $p(k_i, k_j)$ is the joint probability of subset $\psi^u_{k_i}$ from partition ψ^u and subset $\psi^v_{k_j}$ from partition ψ^v. This is given by the proportion of the size of the intersection of subsets to the entire dataset: $p(k_i, k_j) = \frac{|\psi^u_i \cap \psi^v_j|}{n}$.

3.1 Usefulness as a Species Delimitation Partition Metric

The information theoretic-based distances provided by `piikun` are very flexible. For example, different organizations of different populations into sets of species may be analyzed together from multiple different inferences or publications, even if the numbers of individuals, populations, or genes sampled across these sources vary. Furthermore, there are no constraints on the relationships between the partitions considered, such as being nested or otherwise. Detailed discussion of the statistical characteristics and properties of this statistic are given in [2]. Here we present a conceptual description of the properties of this metric that make it useful for analysis of differences between species delimitation models.

A true metric or distance function, such as the variation of information, has the properties of non-negativity, symmetry, and the triangle equality [18]. These properties are useful in aligning with human intuition when interpreting values, as well as preferable in statistical or computational terms due to benefits in algorithm and data structure design or scaling up comparisons [2].

In addition to being a true metric, the variation of information is aligned with the lattice of the set of partitions. Formally, a partition lattice is an ordered set of partitions, where the order is defined by the refinement of the partitions. We can represent the set of all partitions of a particular set of elements as a *partition lattice* using the refinement order (a partition U is defined as a refinement of a partition V if every block in U is a subset of a block in V: we say that U is finer than V, or, equivalently,

U is coarser than V). Conceptually, a partition lattice provides geometrical representation of all possible divisions of a set of elements, from the most granular, where each element forms its own subset, to the most general, where all elements belong to a single subset. This allows for the understanding of relationships between different partitions, offering insights into how small changes in groupings can lead to new partitions, and how these partitions are (or are not) nested within each other. A metric on partitions that is aligned along the lattice of partition sets has numerous advantages for interpretation of the disagreements between different partitions as well as for computation as neighbors in the metric space correspond to refinements in the partition lattice [23]. Such a metric takes into account the hierarchical nature of partitions, recognizing that partitions related through a series of refinements are closer to each other than partitions that differ more fundamentally. That is, under this metric, the nearest neighbor of a species delimitation partition can be obtained by "lumping" the two smallest species blocks of that partition, and vice versa via "splitting". This alignment with the partition lattice is not only useful for intuition, but may also have potential to provide benefits in designing inference or stochastic sampling algorithms (e.g., MCMC moves in partition space).

The upper bound of this metric is given by $2 \log K$, where K is the maximum number of subsets in a partition, or by $log(n)$, with n the number of elements (populations or individuals, depending on the approach), if any partition is possible [2]. In some applications, it might be useful to have this normalized to the range $[0, 1]$. Unfortunately, while it is straight-forward enough to normalize by these known upper bounds, this will preclude comparison of distances across different datasets (in particular, datasets of different sizes). For this latter purpose, `piikun` provides for the calculation of the normalized variation of information following [22], which retains the metric properties of the variation of information.

4 Example: Exploratory Discovery Analysis

Here we focus on one of the visualizations offered by `piikun` which gives us insight into how the "disagreement" between two partitions, as measured by their distance, might vary with respect to some value, trait, or attribute of each of the respective partitions. The visualization in Figure 1 shows how different species delimitation partition distances might vary in relation to the support (probability) of each partition in the comparison. `piikun` was run on the 1000 most probable species delimitation models from a DELINEATE inference on a dataset of *Lionepha* beetles [24, available on the DELINEATE website]. The partitions with high probabilities (i.e., the upper-right area of Figure 1) are mutually close. In contrast, the less probable partitions are much more dissimilar, pairwise (i.e., the lower-left area of Figure 1), while remaining relatively similar to the group of probable partitions (i.e., the right edge of Figure 1). This strongly suggests the existence of a compact "core" formed of the most probable partitions, surrounded by a "halo" of diverse, less probable partitions. Further analysis (e.g., explicit enumeration and comparison of the subsets of the partitions) could then be used confirm that this is indeed because the species delimitation partitions with higher probabilities are disagreeing on the finer-scale splitting of species, while

the less probable models, conflict with each other in more fundamental ways. Other datasets might indicate other relationships, e.g. where more supported models show more fundamental differences from one another, serving as the motivation or bases for more directed statistical analysis at the identified regions of partition space of interest. This demonstrates the interpretive potential of the variation of information as a metric on species delimitation models.

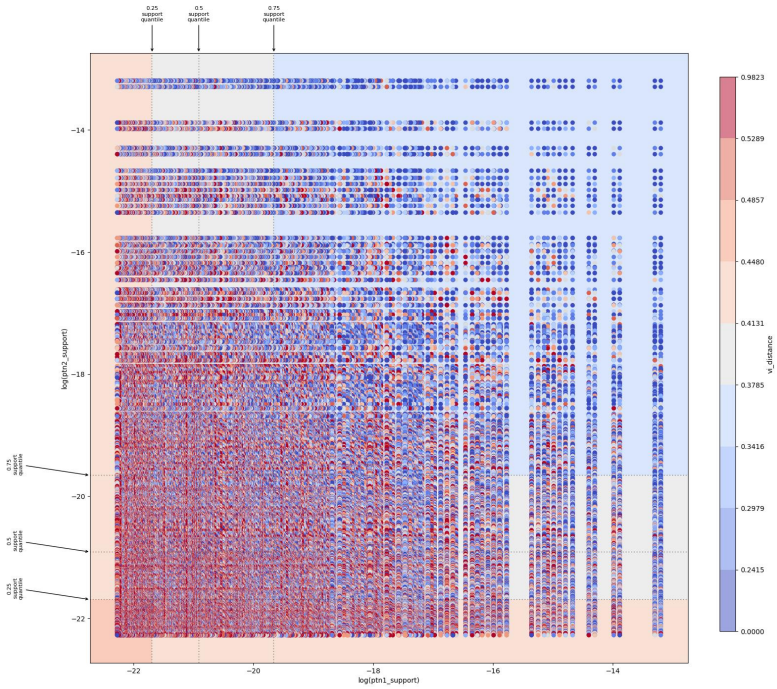

Fig. 1 Understanding how the pairwise similarity of competing species delimitation models varies with their probability (support). The log(ptn_1) and log(ptn_1) axes represent species delimitation models (partitions) ordered by their log-scaled probabilities. Each point represents a pair of models; its color encodes the variation of information distances between them. The color gradient goes from blue to red, with blue indicating smaller distances while red indicating greater distances. The plot background is regionalized into 0.25, 0.5, and 0.75 quantiles based on support, with mean distance for partitions compared within each region indicated with the same color scheme. Note how partitions with high probabilities (upper-right area) are mutually close (VI-distance almost 0), while partitions with lower probabilities (lower-left area) are much more different (VI-distance over 0.91).

Supplementary information.

Acknowledgements. This work was possible due to support from the National Science Foundation grant to author JS: NSF-DEB 1937725 “COLLABORATIVE

RESEARCH: Phylogenomics, spatial phylogenetics and conservation prioritization in trapdoor spiders (and kin) of the California Floristic Province".

References

[1] Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**(3), 379–423 (1948) https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[2] Meilă, M.: Comparing clusterings—an information based distance. Journal of Multivariate Analysis **98**(5), 873–895 (2007) https://doi.org/10/ftwhz3 . Accessed 2023-07-01

[3] Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)

[4] Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (2020) https://doi.org/10/ghbzf2

[5] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020) https://doi.org/10.1038/s41592-019-0686-2

[6] Team, T.P.D.: Pandas-Dev/Pandas: Pandas. Zenodo (2023). https://doi.org/10.5281/zenodo.8092754 . https://doi.org/10.5281/zenodo.8092754

[7] McKinney, W.: Data Structures for Statistical Computing in Python. In: van der Walt, S., Jarrod Millman (eds.) Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010). https://doi.org/10/ggr6q3

[8] Inc., P.T.: Collaborative Data Science. Plotly Technologies Inc., Montreal, QC (2015). https://plot.ly

[9] Hunter, J.D.: Matplotlib: A 2D graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007) https://doi.org/10/drbjhg

[10] Waskom, M.L.: Seaborn: Statistical data visualization. Journal of Open Source

Software **6**(60), 3021 (2021) https://doi.org/10/gjqn3g

[11] Pei, J., Chu, C., Li, X., Lu, B., Wu, Y.: CLADES: A classification-based machine learning method for species delimitation from population genetic data. Molecular ecology resources **18**(5), 1144–1156 (2018) https://doi.org/10/gpf33t

[12] Derkarabetian, S., Starrett, J., Hedin, M.: Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data. Frontiers in Zoology **19**(1), 8 (2022) https://doi.org/10/gshnx5 . Accessed 2023-07-21

[13] Sukumaran, J., Holder, M.T., Knowles, L.L.: Incorporating the speciation process into species delimitation. PLOS Computational Biology **17**(5), 1008924 (2021) https://doi.org/10/gpfx3r . Accessed 2023-07-18

[14] Douglas, J., Bouckaert, R.: Quantitatively defining species boundaries with more efficiency and more biological realism. Communications Biology **5**(1), 755 (2022) https://doi.org/10/gshbt7

[15] Miralles, A., Vences, M.: New Metrics for Comparison of Taxonomies Reveal Striking Discrepancies among Species Delimitation Methods in Madascincus Lizards. PLOS ONE **8**(7), 68242 (2013) https://doi.org/10/f5bzn7 . Accessed 2023-07-24

[16] Ahrens, D., Fujisawa, T., Krammer, H.-J., Eberle, J., Fabrizi, S., Vogler, A.P.: Rarity and Incomplete Sampling in DNA-Based Species Delimitation. Systematic Biology **65**(3), 478–494 (2016) https://doi.org/10/gfkh4v . Accessed 2023-07-24

[17] Ducasse, J., Ung, V., Lecointre, G., Miralles, A.: LIMES: A tool for comparing species partition. Bioinformatics (Oxford, England) **36**(7), 2282–2283 (2020) https://doi.org/10/ggfjdv

[18] Cover, T.M., Thomas, J.A.: Elements of Information Theory, (2006)

[19] Miller, A., Langdon, M.: Barona Inter-Tribal Dictionary: 'iipay Aa Tiipay Aa Uumall, (2008)

[20] Flouri, T., Jiao, X., Rannala, B., Yang, Z.: Species tree inference with BPP using genomic sequences and the multispecies coalescent. Molecular biology and evolution **35**(10), 2585–2593 (2018) https://doi.org/10/gfzpr5

[21] Miralles, A., Ducasse, J., Brouillet, S., Flouri, T., Fujisawa, T., Kapli, P., Knowles, L.L., Kumari, S., Stamatakis, A., Sukumaran, J., Lutteropp, S., Vences, M., Puillandre, N.: SPART: A versatile and standardized data exchange format for species partition information. Molecular Ecology Resources **22**(1), 430–438 (2022) https://doi.org/10/gshbt3 . Accessed 2023-07-18

[22] Vinh, N.X., Epps, J., Bailey, J.: Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. Journal of Machine Learning Research **11**(95), 2837–2854 (2010). Accessed 2023-07-01

[23] Meilă, M.: Comparing clusterings: An axiomatic view. In: Proceedings of the 22nd International Conference on Machine Learning. ICML '05, pp. 577–584. Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10/dkr8nb . https://doi.org/10.1145/1102351.1102424 Accessed 2023-07-20

[24] Maddison, D.R., Sproul, J.S.: Species delimitation, classical taxonomy and genome skimming: A review of the ground beetle genus Lionepha (Coleoptera: Carabidae). Zoological Journal of the Linnean Society **189**(4), 1313–1358 (2020) https://doi.org/10/gsr6r4