

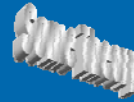
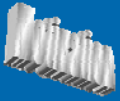
A Step Toward Barcoding Life II: A Model-Based, Decision-Theoretic Clustering Method

University of Idaho
Open Space. Open Mind.

Zaid Abdo¹ and Brian Golding²

¹Departments of Mathematics and Statistics, University of Idaho, Moscow, Idaho, USA

²Biology Department, McMaster University, Hamilton, Ontario, Canada



- One aspect of the barcoding of life is to cluster sampled individuals into well-supported groups that either conform with pre-identified species or highlight new ones for further evaluation.
- This involves evaluating the statistical evidence towards combining individuals into groups based on DNA evidence provided by DNA barcodes.

- The process of identifying groups of sequences or clusters is a decision making problem.
- In our approach clustering involves joining individuals or groups of individuals if such a merge results in a reduction in the posterior risk of clustering R_p .
- Posterior risk [1], R_p , is defined to be the expected loss; loss measures the error of merging two groups with different origin.

$$R_p = E[L(k, i) | D_j, \epsilon = i] = \sum_k L(k, i) Pr(D_j \in k | D_j, D_k, \theta_k)$$

where:

$$L(k, i) = \begin{cases} LS(1 - p_{\text{dist}}[\text{consensus}(D_j) - \text{consensus}(D_k)]) & |D_k| \geq 1, k \neq i \\ LS(1 - 0.25) & |D_k| = 0, k \neq i \\ 0 & k = i \end{cases}$$

loss resulting from merging group j with group i when group k is the true origin. LS, is the length of the sequence; and $\text{dist}()$ is the distance of the consensus of group D_j from that of group D_k .

$$Pr(D_j \in k | D_j, D_k, \theta_k) \approx \frac{Pr(D_j, D_k | D_j \in k, \theta_k) / Pr(D_k | \theta_k)}{\sum_l Pr(D_j, D_l | D_j \in l, \theta_l) / Pr(D_l | \theta_l)}$$

posterior probability that D_j belongs to group k given that we know all sequence data within it; that we have a sample, D_k , of individuals that belong to group k ; and that we know the true parameter governing group k 's evolution.

- We assumed that different groups evolve independently; that groups are fully defined and pre-specified; that each group forms a panmictic population that follows a Wright-Fisher, simple neutral model of evolution governed by a known parameter θ ; and that the model governing nucleotide evolution on each genealogy is F81. We also assume that the chance of observing only one individual from a group is q and that the probability of belonging to an outlier group is p .

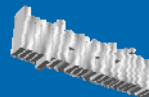
To efficiently use the available information in Barcoding data to objectively identify population groups in an accurate and efficient manner.



- To accomplish our goal we have developed a model based, decision theoretic framework based on the coalescent theory.
- We use both distance and the posterior probability of a group, given DNA sequences from members of this group and from other groups to evaluate the evidence for combining these groups to form clusters.
- We believe that this approach makes efficient use of the available information in the data.



- We used 2 simulated data, based on the phylogeny in Figure 1, with known cluster structure [1] and with two different phylogenetic divergence levels ($\nu = 1$ and 0.01 sub/site/gen); and one level of within group evolution ($\theta = 0.01$ per site) and one sample size ($n = 5$) to for a preliminary test of the power of our method under different conditions of cluster overlap.
- In the cases where the divergence is high ($\nu = 1$) our method managed to correctly retrieve the original clusters at each level of with-group evolution and at each sample size.
- A shallow phylogeny ($\nu = 0.01$) blurred differences between some of the deep groups (NUMT, CELT, and YESENN) resulting in combining them into one larger group. This is expected due to the closeness of the most recent common ancestors of these groups. Moreover, the sample size did not create enough signal to distinguish these groups.



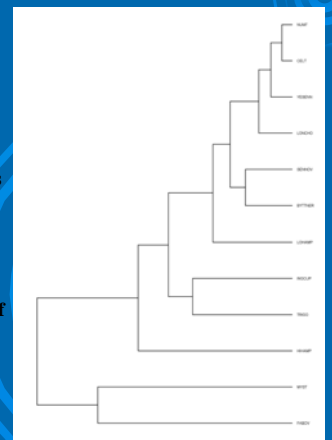
Two steps to implement our proposed model:

1. Progressive model-based hierarchical clustering.
 2. An EM (Expectation-Maximization) step to optimize the clusters after the stopping rule is attained for the first step.
- We stop the hierarchical clustering when posterior risk does not change for a full iteration.



- Decision theoretic, model-based approaches can improve clustering resolution by incorporating information about the evolutionary history, in addition to distance, in making grouping decisions.
- This approach is computationally intensive, though needs to be implemented occasionally to validate the results of the assignment approach we have developed earlier.
- Special attention is focused on developing other stopping rules for this algorithm; the current rule involves some judgment. Our aim is to maintain an objective stopping rule, some approaches to accomplish this include model comparison using Bayes factors [2].
- Further work is underway to compare this approach to the commonly used distance based approaches such as NJ, and to fine tune the implemented algorithm.

Figure 1: Data were simulated using the phylogeny of neotropical skipper butterfly *Astraptes flugerator* as a base. First, DNA sequences of the most recent common ancestors (MRCAs) of each of 12 groups were generated. These MRCA's were then used as seeds to generate within group DNA sequences based on a coalescent simulation. Depth of this tree varies depending on the rate of substitution [1].



REFERENCES

1. Abdo, Z. and G. B. Golding. 2007. A step toward barcoding life: A model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology*, 56:44-56.
2. Yeung, K. Y., C. Frisley, A. Murru, A. E. Rafferty and W. L. Ruzzo. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17: 977-987.

ACKNOWLEDGEMENTS

This research was funded by an NSERC discovery grant and an CRC award to GBG. ZA is also funded by a research program startup funding provided by the University of Idaho. This research was also supported through funding to the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, NSERC, and other sponsors listed at www.BOLNET.ca.