

A Step Toward Barcoding Life: A Model-Based,  
Decision-Theoretic Method to Assign Genes to  
Preexisting Species Groups  
Zaid Abdo and G. Brian Golding

Presented by Mat Kelly

December 8, 2011





# What We Need

- ▶ Statistical method to determine appropriate assignment
  - ▶ Would give us a degree of confidence of our classification
- ▶ Assign new individuals to existing groups, previously established by taxonomists
- ▶ Taking a sequence and determining appropriate group of origin is:
  - ▶ A clustering problem (unsupervised), not tackled in paper
  - ▶ An assignment problem (supervised), assumes pre-existing groups defined elsewhere

# The Method

- ▶ Assuming a correct group  $k$ , assigning individual to wrong group  $j$  results in error utilized to reflect assignment uncertainty
- ▶ Loss function  $L(k, j)$  quantifies error
- ▶  $k$  is likely not known
- ▶ Instead, find group  $i$  that minimizes posterior risk of assigning to a group

$$R_i = E[L(k, i)] = \sum_k L(k, i)Pr(x \in k | x)$$

# Independence and Transformation

$$R_i = E[L(k, i)] = \sum_k L(k, i)Pr(x \in k | x)$$

...assumes that individuals within a group are *i.i.d*

This is not the case when they share a common history, so we write:

$$R_i = E[L(k, i)] = \sum_k L(k, i)Pr(x \in k | x, D)$$

...where D represents preexisting data.

Also, using Bayes rule, we can calculate the posterior with:

$$Pr(x \in k | x) = \frac{Pr(x | x \in k)Pr(x \in k)}{\sum_j Pr(x | x \in j)Pr(x \in j)}$$

# The Coalescent

- ▶ Describes the behavior of a sample of individuals, backward in time, until they find their most recent common ancestor (MRCA)
- ▶ Eliminates need to consider unobserved and extinct individuals
- ▶ Model to be used combines aforementioned classifier and coalescent to form the “Coalescent Assigner”

# The Coalescent Assigner

- ▶ Avoids increased complexity that comes with considering phylogenetic history (pre-existing data in past equation)
- ▶ We assume that the ancestors of each population have been evolving long enough that it is now evolving independent of other groups
  - ▶ The stationarity of the nucleotide substitution process has been attained.
- ▶ **Assume:** Different groups that are potential targets of assignment are fully defined and prespecified
- ▶ **Assume:** Each group forms a panmictic population that follows Wright-Fisher neutral model
  - ▶ i.e. does not allow recombination, selection or migration
- ▶ ∴ Evolutionary process within each group is governed by one parameter
  - ▶  $\theta$  - expected number of mutational steps between any two individuals within that group
- ▶ Because groups are fully defined,  $\theta$  known with certainty

# Formalizing the Loss Function

$$L(k, i) = \begin{cases} LS - dist[x - consensus(d_k)] & : k \neq i \\ 0 & : k = i \end{cases}$$

$LS$  = length of sequence under study

$dist()$  = # nucleotides diff betw. individual and consensus of potential group

- ▶ This function is not unique, any function can be created to reflect importance of other factors.

# Implicit Doubt

- ▶ We set a limit above which assigning an individual to the group with minimal risk is questionable
  - ▶ Highlights need to evaluate origin of that individual and possibility of it being misplaced
- ▶ Proportion of nucleotides that match by chance = 0.25
- ▶ Each group is likely to obtain assignment
- ▶ Doubt threshold calculated as  $\frac{0.25}{M}$  where  $M$  is the total # of available groups

# Posterior Probability of Membership

$$Pr(x \in k | k, D, \theta) = \frac{Pr(x, x \in k | D, \theta)}{Pr(x | D, \theta)}$$

... where  $\theta$  is a known vector of parameters  $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_M$

Using Bayes rule:

$$Pr(x \in k | x, D, \theta) = \frac{Pr(x | x \in k, D, \theta) Pr(x \in k | D, \theta)}{\sum_j Pr(x | x \in j, D, \theta) Pr(x \in j | D, \theta)}$$

... where  $P(x | x \in k, D, \theta) =$  likelihood of newly added sequence,  $x$  given that:

- ▶ It originated from group  $k$
- ▶ The known, preassigned data  $D$
- ▶ The known evolutionary param  $\theta$

## Posterior Probability of Membership (continued)

The likelihood  $Pr(x|x \in k, D, \theta)$  can be rewritten as...

$$Pr(x|x \in k, D, \theta) = \frac{Pr(x, D_k | x \in k, D_1, \dots, D_{k-1}, D_{k+1}, \dots, D_M, \theta_1, \dots, \theta_k, \dots, \theta_M)}{Pr(D_k | x \in k, D_1, \dots, D_{k-1}, D_{k+1}, \dots, D_M, \theta_1, \dots, \theta_k, \dots, \theta_M)}$$

... assuming independence:

$$Pr(x|x \in k, D, \theta) = \frac{Pr(x, D_k | x \in k, \theta_k)}{Pr(D_k | x \in k, \theta_k)}$$

... assuming  $x$  belonging to group  $k$  will not signif. alter likelihood of data  $D_k$  given  $\theta_k$ :

$$Pr(x \in k | x, D, \theta) = \frac{\frac{Pr(x, D_k | x \in k, \theta_k)}{Pr(D_k | x \in k, \theta_k)} Pr(x \in k | D, \theta)}{\sum_j \frac{Pr(x, D_j | x \in j, \theta_j)}{Pr(D_j | x \in j, \theta_j)} Pr(x \in j | D, \theta)}$$

## Posterior Probability of Membership (continued)

... assuming presampled  $D$  and true parameter vector  $\theta$  will not where assigned new individual and **assume** a uniform prior on where the new individual might belong:

$$Pr(x \in k | x, D, \theta) = \frac{\frac{Pr(x, D_k | x \in k, \theta_k)}{Pr(D_k | \theta_k)}}{\sum_j \frac{Pr(x, D_j | x \in k, \theta_j)}{Pr(D_j | \theta_j)}}$$

Above has calculated likelihood of an individual belonging to a certain group rather than calculating the likelihood of the group once the individual is added!

We can put the result into the risk equation from before

$$R_i = \sum_k L(k, i) Pr(x \in k | x, D, \theta) = \sum_k L(k, i) \frac{\frac{Pr(x, D_k | x \in k, \theta_k)}{Pr(D_k | \theta_k)}}{\sum_j \frac{Pr(x, D_j | x \in k, \theta_j)}{Pr(D_j | \theta_j)}}$$

# Implementing the Coalescent Assigner

1. Establish true parameter  $\theta$  for each target group
  - ▶ This calculates the likelihood of the presampled data given its group of origin
  - ▶ Not possible in practice, we instead estimate the parameter for each group using available data
  - ▶ Produces Maximum Likelihood Estimates (MLEs), which are “close to the truth”
2. Use MLEs to calculate likelihood of data after adding individual to each available group.
  - ▶ No direct method exists to evaluate likelihood after adding individual but new likelihood function would be  $Pr(x, D_k | x \in k, \theta_k) = \sum_G Pr(x, D_k | x \in k, G) Pr(G | \theta_k) \dots$  where  $G$  is all possible genealogies.
  - ▶ Both a naïve method (sampling from posterior  $Pr(G | \theta_k)$ ) and a Markov Chain Monte Carlo approach were implemented to calculate the sum with an adjustment that allowed the direct calculation of the likelihood
  - ▶ Remember: This assumes full knowledge of  $\theta_k$ .
  - ▶ Naïve method works best for small within group sample sizes - used to show effectiveness vs. distance method

# Implementing the Coalescent Assigner (cont)

3. Finding the consensus sequence
  - ▶ Used a 50% majority rule
4. Evaluate the risk using the risk formula
  - ▶ Straightforward once all likelihoods are evaluated

# Validation

- ▶ Simulated and real data used to evaluate performance of Coalescent Assigner
- ▶ Compared against distance approach that does not take understanding of evolutionary process into account.
- ▶ The “correct” group of assignment,  $k$  was one with  $\min \text{dist}[x - \text{consensus}(D_k)]$
- ▶ Assignment is doubtful if  $\exists$  Group  $j$  where  $j \neq k$  with  $\text{dist}[x - \text{consensus}(D_k)]/LS < 3\%^*$  or equal distance from the newly sampled individual

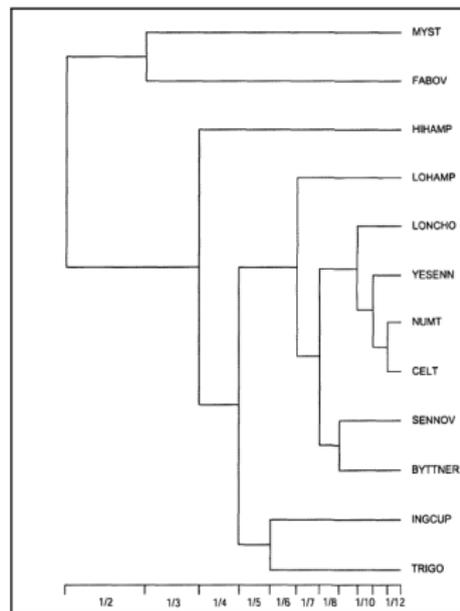
\* This threshold is derived from Herbert et al (2003b) and was shown to highlight the range of appropriate use of the distance measure

# Simulated Data

- ▶ We want to evaluate ability of proposed method to assign an individual to its true group of origin under various degrees of divergence and overlap
- ▶ Utilized *Asrptes flugerator* (neotropical skipper butterfly) data, assuming existence of 12 predefined groups
- ▶ Using PAUP\*4b10, we estimate phylogeny and parameters of model based on consensus sequences
  - ▶ Model of choice was General Time-Reversible (GTR) model
  - ▶ Unresolved nodes were resolved arbitrarily to create bifurcating tree
  - ▶ Only topology of tree and est. params of model were used for next steps of simulation

# A Generated Tree Example

- ▶ Branch lengths represent the expected number of substitutions, on average, based on Yule model with parameter  $\frac{1}{v}$
- ▶ Depth varied based on rate of substitution
  - ▶  $v = 1$  divergent groups
  - ▶  $v = 0.1$  moderately divergent groups
  - ▶  $v = 0.01$  short tree with moderate group overlap
  - ▶  $v = 0.001$  short tree with much overlap
- ▶ Utilized neutral coalescent to model within-group evolution
- ▶ An extra individual was generated within each group, 1/12 reassigned.
  - ▶ Done to maintain equal group sizes prior to the assignment
- ▶ In the case of no within group variation, assume the group's diversity was very small and set  $\theta$  to  $10^{-8}$



Tree used to simulate DNS sequences of ancestors (MRCAs) of each of the 12 groups

## Applying the Method

| Name              | Size |
|-------------------|------|
| Astraptes-MYST    | 3    |
| Astraptes-FABOV   | 31   |
| Astraptes-YESENN  | 78   |
| Astraptes-SENNOV  | 103  |
| Astraptes-BYTTNER | 4    |
| Astraptes-INGCUP  | 65   |
| Astraptes-TRIGO   | 51   |
| Astraptes-NUMT    | 4    |
| Astraptes-HIHAMP  | 16   |
| Astraptes-LOHAMF  | 47   |
| Astraptes-CELT    | 23   |
| Astraptes-LONCHO  | 41   |

- ▶ Neotropical skipper butterflies have traditionally been classified as a single species
- ▶ On the basis of barcode data, class represents 12 sympatric species
- ▶ In evaluating method, draw one query sequence at random from all sequences
- ▶ Reconstruct groups without this sequence
- ▶ Perform assignment on excluded sequence

# Performance Evaluation

- ▶ Evaluated change in power of coalescent assigner when used to assign presampled individual to groups with
  - ▶ high overlap (short phylogeny, high  $\theta$ )
  - ▶ high divergence and group delimitation (deep phylogeny, low  $\theta$ )
  - ▶ other scenarios between these 2 extremes
- ▶ Compared to change in power with scenarios using simple distance method

# Coalescent Assigner vs. Distance method

|        |       | Coalescent assigner |       |          |        |       |      | Distance |       |          |     |
|--------|-------|---------------------|-------|----------|--------|-------|------|----------|-------|----------|-----|
|        |       |                     |       | $\theta$ |        |       |      |          |       | $\theta$ |     |
|        |       | V                   | 0.001 | 0.01     | 0.1    |       |      | V        | 0.001 | 0.01     | 0.1 |
| n = 5  | 0.001 | 93%                 | 89%   | 82%      | n = 5  | 0.001 | 92%  | 61%      | 56%   |          |     |
|        | 0.01  | 99%                 | 97%   | 92%      |        | 0.01  | 99%  | 93%      | 60%   |          |     |
|        | 0.1   | 100%                | 100%  | 99%      |        | 0.1   | 100% | 100%     | 92%   |          |     |
|        | 1     | 100%                | 100%  | 100%     |        | 1     | 100% | 100%     | 99%   |          |     |
| n = 10 | 0.001 | 99%                 | 93%   | 82%      | n = 10 | 0.001 | 88%  | 59%      | 55%   |          |     |
|        | 0.01  | 99%                 | 93%   | 82%      |        | 0.01  | 95%  | 96%      | 71%   |          |     |
|        | 0.1   | 100%                | 100%  | 95%      |        | 0.1   | 100% | 100%     | 97%   |          |     |
|        | 1     | 100%                | 100%  | 100%     |        | 1     | 100% | 100%     | 100%  |          |     |
| n = 25 | 0.001 | 98%                 | 89%   | 52%      | n = 25 | 0.001 | 86%  | 78%      | 72%   |          |     |
|        | 0.01  | 100%                | 95%   | 68%      |        | 0.01  | 98%  | 90%      | 79%   |          |     |
|        | 0.1   | 100%                | 100%  | 91%      |        | 0.1   | 100% | 100%     | 94%   |          |     |
|        | 1     | 100%                | 100%  | 97%      |        | 1     | 100% | 100%     | 100%  |          |     |

Table: % Correct with 50,000 chain length. Divergence  $v$  measured in substitutions/site;  $\theta$  is exp. # evolutionary steps between two individuals;  $n$  is group size

- ▶ Table shows clear improvement of methods as divergence of phylogeny,  $v$ , increases compared to within group evolution
- ▶ The more divergent the groups are, the greater the distinction between their MRCA and the more likely that the within group evolutionary process are independent and easier to distinguish
- ▶ Coalescent assigner is more powerful than distance method in distinguishing newly assigned samples

# Conclusions

- ▶ Evolutionary history is critical in accurately assigning individuals to the most appropriate group
- ▶ Large # of taxa combined with larger # of nucleotide patterns can result in an extensive search requiring a much larger Markov Chain to guarantee a good sample, which would be drawn from the tail of the genealogical distribution.
- ▶ The coalescent assigner is much more powerful than the distance method when overlap might mask the correct assignment.
- ▶ Coalescent Assigner is simpler, faster than fully incorporating phylogenetic and population genetic history.

# References

- ▶ Z. Abdo and G.B. Golding, A step toward barcoding life: a model-based, decision theoretic method to assign genes to pre-existing species groups. *Systematic Biology*, 56 (2007), pp. 113.