Data–Informed Reconstruction of Single-Diploid-Hybrid Phylogenies

Monique M. Morin
University of New Mexico – Los Alamos
July 23, 2009
Evolutionary histories are complicated by reticulate events. A hybrid is one type of reticulate event. This work focuses on diploid hybridization. Parents each contribute half of their DNA to form the new species. Network specific reconstruction algorithms are required to reconstruct histories with reticulate events.
Confidence in reconstruction algorithms is improved, by comparing generated and reconstructed networks.

**Our Strategy:** generate new network, attempt reconstruction, and then compare the two topologies.
Network topology and sequences are generated in conjunction with each other.

Simultaneous Topology and Sequence Generation

- Assign Root Sequences and Start Topology
- Update Sequences in Preparation for Hybrid Event
- Execute Hybrid Event
- Continue Topology and Sequence Generation

simulator features include:

- extended birth/death model
- event queue driven
- independent tracking of evolutionary and clock branch lengths
- variable evolutionary rates
- diploid/polyploid hybrids
- sequence information can be used for hybrid creation
The tripartition score is used to assess the similarity between a generated and reconstructed network.

**Tripartition measure***:

- extension of the Robinson-Foulds distance for trees
- scores: “indistinguishable” (0) to “completely different” (1)

Investigation of reconstruction algorithms:

- multiple source/reconstructed network pairs are generated
- individual tripartition scores are calculated
- histograms of the scores are plotted along with the mean and standard deviation

Challenge: Can additional information be extracted from extant taxa to aid reconstruction?

One possible source of information – Reticulate Impact (RI).

\[
RI = \frac{\text{Taxa Descendant from Reticulate Node}}{\text{All Extant Taxa}}
\]

Acquiring reticulate impact information can pose another set of challenges, though there are some known hybrid indicators and proposed detection methods.*

Starting from the extant taxa, networks are reconstructed in four steps.

Network Reconstruction:
- hybrid subtree
- parental structure
- parental subtrees
- remaining subtree

Note: subtrees are created with PHYLIP*

Multiple algorithms have been explored for grouping extant taxa.

**Extreme Custom**
- all extant sets specified on input

**Closest Neighbor**
- find closest neighbor (Hamming distance) for all non-impacted extant taxa
- if a taxon's closest neighbor is impacted, make it a parental extant taxon

**Partial Impact Set**
- fraction of hybrid-impacted taxa
- closest neighbor identification for parental descendants

**Singleton Starter/Avg Size**
- one impacted chosen at random
- fill set to average size with closest neighbors (Hamming distance)
Accuracy of reconstructed networks improves when additional extant information is used.

- Tripartition scores computed for random, unrelated pairs of single-hybrid networks: avg = 0.995 +/- 0.010
- Closest neighbor average tripartition: 0.101 +/- 0.04
- Extreme custom all extant sets known: avg = 0.075 +/- 0.04

50 extant taxa, second parent = min hamming, birth=48, hybrid=12, >= 3000 runs
Reconstructions improve when more correct hybrid descendants are identified.

Initial knowledge loss of descendants from 100% to 75% has the largest effect on reconstruction efforts.

50 extant taxa, second parent = exponential hamming function, birth=4.85, hybrid=1, 500 runs

¼ data set requires further analysis due to rounding factors of set size
Preliminary data indicates that knowledge of a single hybrid descendant can improve the reconstruction.

Tripartition scores tend to be lower when at least one hybrid descendant has been correctly identified.

50 extant taxa, second parent = exponential hamming function, birth=4.85, hybrid=1, 500 runs

*Further analysis of the “double distribution” is needed.*
Conclusions

- We have developed an algorithm that uses reticulate impact knowledge to inform the reconstruction.

- Reconstructions were performed for simulated networks which had their topologies and sequences evolved simultaneously.

- The tripartition measure was used to assess the similarity of the source/reconstructed pairs.

- Analysis shows:
  - better reconstructions are achieved when reticulate impact information is incorporated,
  - similarity degrades as knowledge of the impact set drops, and
  - reconstruction accuracy is improved even with the knowledge of a single hybrid descendant.
Acknowledgements

Technical Discussions
John Burge
Steve Evans
Jens Lagergren
Randy Linder
Kris McGary
Bernard Moret

Technical Assistance
Gabriela Barrantes
Leigh Fanning
Eric Gottlieb
Edgar Leon
Nick Pattengale
Krister Swenson
Christy Warrender

Software available for download:
http://www.cs.unm.edu/~morin

This work was been supported in part by National Science Foundation grants IIS 01-21377, DEB 01-20709, EF-03-31654.