Reconstructing Reticulate Evolution in Species

LUAY NAKHLEH

Department of Computer Science
University of Texas at Austin
Austin, TX 78712 USA
http://www.cs.utexas.edu/users/nakhleh

RETICULATE EVOLUTION IN BIOLOGY

Project Members:

- UT CS: Tandy Warnow and Luay Nakhleh
- UT Biology: Randy Linder
- UNM CS: Bernard Moret

Reticulate Evolution

- Wouldn't it be nice if...

  Sexual creatures would just behave themselves

  then, we could stick with bifurcating trees to properly describe the evolutionary history of organismal lineages

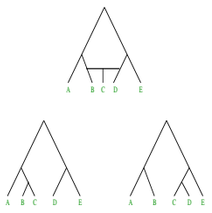
- However...
  - Unruly Nature: Whatever is not forbidden will occur...
  - Lateral gene transfer: Lawrence, Ochman estimated that 755 of 4,288 ORF's in E.coli were from at least 234 lateral gene transfer events (**PNAS USA 95, 9413-9417 (1988)**)
  - Hybridization: Plants (estimates that as many as 30% of all plant lineages are the products of hybridization), Fish, Some frogs, several lineages of invertebrates (e.g., corals)

Projects

- Simulation Tools for generating random networks and simulating sequence evolution down networks
- Error Measures to study the performance of network reconstruction methods
- Methods for detecting and reconstructing reticulate evolution

Phylogenetic Networks and Trees

- A phylogenetic network is a rooted directed acyclic graph
  - Tree nodes (indegree 1) and network nodes (indegree 2)
  - A set of time constraints
- A phylogenetic network induces a set of trees



- Each tree is the result of removing exactly one of the two edges incoming into each network node
- A network with p network node induces at most 2^p trees

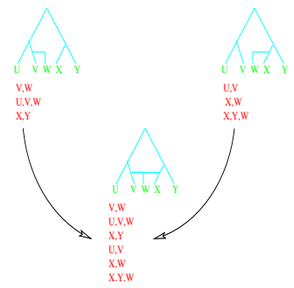
Simulation Tools for Networks

- GENNET: A tool for generating random networks
  - Follows a birth-death process
  - Handles various hybrid speciation types
- NETSEQGEN: A tool for simulating the evolution of DNA sequences on networks
  - An extension of Seq-Gen (Rambaut et al.)
  - Handles various models of evolution
- COMPNET: A tool for comparing networks
  - Tripartition-based: each edge in a network induces a tripartition of the set of taxa

"Towards the Development of Computational Tools for Evaluating Phylogenetic Network Reconstruction Methods", Nakhleh et al., **Pacific Symposium on Biocomputing 2003**.

Topological Accuracy of Networks

The splits of a network N, denoted by $C(N)$, is the union of the splits of the trees induced by the network



Based on the splits of networks, we define the false positives and false negatives between two networks.
Let N and N' be the model and inferred networks. Then:

- $FN(N, N') = |C(N) - C(N')|$
- $FP(N, N') = |C(N') - C(N)|$

Existing Reconstruction Tools

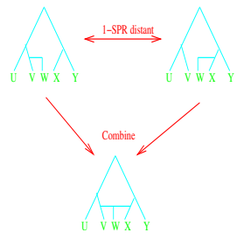
- SplitsTree (Huson)
- NeighborNet (Bryant and Moulton)
- T-REX (Makarenkov)
- NETWORK (Bandelt *et al.*)
- Lateral Gene Transfer Program (Hallett and Lagergren)
- Arlequin (Scheneider *et al.*)
- Pyramids (J.C. Aude *et al.*)
- STATGEOM (K. Nieselt-Struwe)
- GEOMETRY (Kuznetsov and Morozov)
- TCS (Clement *et al.*)

The Separate Analysis Approach

Maddison's Observation

Systematic Biology, 46(3):523-536, 1997.
"What is needed is a method that counts the minimal number of branch moves needed to convert one tree into another, where branch moves are restricted so as not to violate a linear time order."

Maddison's Method



Challenges and Solutions

Challenges

- Computational: computing SPR distances is computationally expensive
- Systematic: reconstructed gene trees almost always contain topological error

Solutions

- we gave an efficient algorithm for computing SPR distances for constrained reticulations
- instead of single gene trees, consider the strict consensus of a few "good" ones

Our Approach: SpNet

Rationale: The "wrong" edges in the individual trees will be contracted in the consensus tree
The method:

- Given two genes g_1 and g_2, construct a set U_1 of trees for gene g_1 and a set U_2 of trees for gene g_2
- Compute the strict consensus tree t_1 of a subset of trees from U_1 so as to achieve a certain level of resolution in t_1, and repeat the same for U_2 to obtain t_2
  - If t_1 and t_2 are compatible, construct a tree on the combined datasets
  - If t_1 and t_2 are not compatible, check if there exist two binary trees T_1 and T_2 such that (1) T_1 refines t_1, (2) T_2 refines t_2, and (3) T_1 and T_2 are 1-SPR apart.
    - If such T_1 and T_2 exist, construct a network N with one reticulation event, and return N
    - If no such pair of trees exist, determine that there is more than one reticulation event

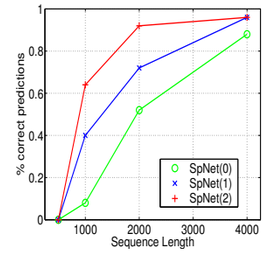
DETECTION QUALITY

How often does the method infer the correct number of reticulations in the model phylogeny?

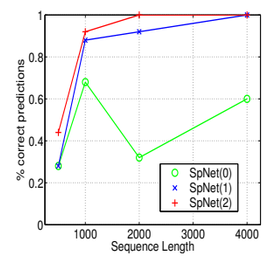
Method and Settings

- SpNet(i): Our method applied to strict consensus of ML trees such that the consensus trees are missing i edges. SpNet(0) amounts to Maddison's approach
- 20-taxon networks, 0.1 scaling factor, 2 deviation factor
- Sequences evolved under the GTR+Γ model with invariant sites

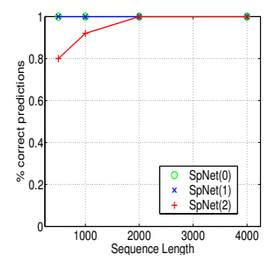
Model Phylogeny: Tree



Model Phylogeny: 1-hybrid Network



Model Phylogeny: 2-hybrid Network



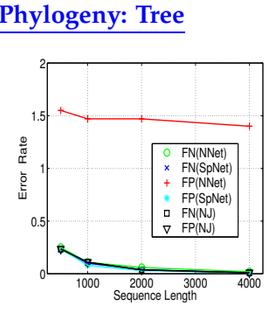
RECONSTRUCTION QUALITY

What is the topological accuracy of the phylogenies inferred by the various methods?

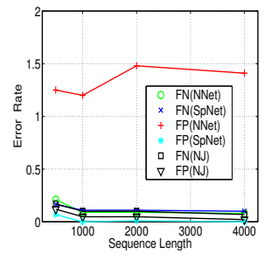
Method and Settings

- Methods:
  - SpNet: our method applied to SC of ML trees, where the SC tree is missing one edge
  - NNet: the method of Bryant and Moulton (NeighborNet)
  - NJ: Neighbor Joining (Saitou and Nei)
- 20-taxon networks, 0.1 scaling factor, 2 deviation factor
- Sequences evolved under the GTR+Γ model with invariant sites

Model Phylogeny: Tree



Model Phylogeny: 1-hybrid Network



Model Phylogeny: 1-hybrid Network