Hybrid Tree Reconstruction Methods

Daniel Huson
PACM, Princeton University

Scott Nettles
CIS, University of Pennsylvania

Kenneth Rice
SmithKline Beecham

Tandy Warnow
CIS, University of Pennsylvania

Shibu Yooseph
DIMACS, Rutgers University

WAE98

The Tree of Life

The first phylogenetic tree of life, Ernst Haeckel (1866)

The Tree of Life

A modern view of the tree of life

Daniel Huson, 1998

Overview

- Experimental study
  - Performance of different tree reconstruction methods
  - Shows different behaviour of methods on *big* trees
- The Hybrid Method
  - Uses best features of each method
  - Improves performance for a range of parameters

Tree Reconstruction Methods

- Sequence methods
  - Maximum Parsimony
    - Popular sequence-based method
    - Solve the Hamming distance Steiner tree problem to obtain the *most parsimonious tree*. (NP-*hard*)
    - We use *heuristic search maximum parsimony with strict consensus*

- Polynomial time distance methods
  - Neighbor-Joining
    - Popular distance-based method
    - Successively "join" close pairs of *taxa* to infer tree. (*fast*)
  - Buneman Tree
    - Distance-based method with nice mathematical properties (*but low resolution*)

– Single Pivot 3-Approximation

Daniel Huson, 1998

Maximum Parsimony

Find tree that explains data using a minimal number of *mutations*.

```
A: 0 1 0 1 1

B: 0 0 0 1 0

C: 0 1 0 1 1
D: 1 0 0 1 0
```

Jukes-Cantor Model of Evolution

- Given a *model tree* T, with edge weights $P(e)$

- Interpret $P(e)$ as the probability of change at *any* given Position in a sequence along the edge e, (*i.i.d.* model)

- Fix a sequence length k. Choose a root r and a random start sequence S at r

- Evolve sequences along the tree T, *Markov tree*.

Experimental Simulation Studies

- Choose Jukes-Cantor model tree T (e.g. inspired by biology)

- Evolve sequences along the model tree

- Apply tree reconstruction methods M to evolved sequences S

- Compare $M(S)$ with model tree T for topological accuracy

- Software: ecat, PAUP, Phylip. Our programs in C++ & LEDA.

Objective: Topological Accuracy

The main goal in biology is to correctly infer the *order* of speciation events, hence the objective is to minimize:

- *False Positives*: wrongly inferred edges
- *False Negatives*: missing edges

Model tree T:

Estimation $M(S)$:

One False Positive: $\{S_1, S_3\}$ vs. $\{S_2, S_4, S_5\}$

One False Negative: $\{S_1, S_2\}$ vs. $\{S_3, S_4, S_5\}$

Statistical Consistency

Given longer and longer sequences (i.e. better and better approximations of the tree distances), does the result of the method converge to the correct tree?

- Distance methods: yes

- Maximum Parsimony: no, e.g.:

In the *Felsenstein zone*

If the distance between c and d is small, and all other distances are large, then c and d will only differ at a few sites, whereas any other pair will differ at many sites. Parsimony will be misled into making c and d siblings.

Comparison of False Negatives

False Negatives

- 35 taxon tree (from 500 taxon *rbc*L dataset)
- Sequence length vs. FN
- Max. $p(e) = 0.56$
- 100 experiments per point

Comparison of False Positives

False Positives

- 35 taxon tree (from 500 taxon *rbc*L dataset)
- Sequence length vs. FP
- Max. $p(e) = 0.56$
- 100 experiments per point

stuff on 93 taxon tree here...

Comparsion on 93 Taxon Tree

FP Parsimony FN Parsimony

FP=FN Neighbor-Joining FP≅FN 3-Approx.

- 93 taxon tree (from 500 taxon *rbc*L dataset)
- Max. $p(e)$ vs. FP or FN, by sequence length
- 100 experiments per point

Conclusion from Experiments

- Buneman Tree is conservative: very low FP rate, high FN rate

- Single Pivot 3-Approximation lies between Buneman and NJ, so it's not competitive

- New results on **HS-Parsimony vs. NJ**:
  - Given low divergence, both do well:
    - HS-Parsimony slightly better on short sequences
    - NJ slightly better on longer sequences
  - High divergence, both can do poorly:
    - NJ converges to true tree, but very slowly (*statistical consistency*)
    - Parsimony may not converge to true tree, but has lower errors than NJ on sequences of realistic length

The Hybrid Method

Picture explaining the idea!

Comparison of False Positives

False Positives

- 93 taxon tree (from 500 taxon *rbc*L dataset)
- Max. $p(e)$ vs. FN, by sequence length
- 100 experiments per point

Comparison of False Negatives

False Negatives

- 93 taxon tree (from 500 taxon *rbc*L dataset)
- Max. $p(e)$ vs. FN, by sequence length
- 100 experiments per point

Conclusions

Our experiments indicate that different tree reconstruction methods have different strengths. We proposed a hybrid method to make use of them.

The hybrid method produced

- the same or fewer False Negatives than the best of its components in more than 98% of our experiments

- distinctly fewer such False Negatives on over 60% of the datasets generated on the 93-taxon tree

- distinctly fewer such False Negatives on over 30% of the datasets generated on the 35-taxon tree

Moreover, it is a statistically consistent method, because the Buneman tree method is.