The Disk Covering Method

Tandy Warnow

CIS, University of Pennsylvania

June 1998

Report on joint work with:

Daniel Huson
PACM, Princeton University

Scott Nettles
CIS, University of Pennsylvania

Kenneth Rice
Smith-Kline Beecham

Shibu Yooseph
DIMACS, Rutgers University

References

- D.H. Huson, S. Nettles, L. Parida, T.J. Warnow and S. Yooseph. The Disk-Covering Method for Tree Reconstruction. In: R. Battiti and A.A. Bertossi eds., Proceedings of *Algorithms and Experiments* (ALEX'98) (Trento, Italy, Feb. 9–11, 1998), 62-75, 1998.

- D.H. Huson and K. Rice and S. Nettles and T.J. Warnow and S. Yooseph. Hybrid Tree Construction Methods. Accepted: *Workshop on Algorithms Engineering* (WAE'98).

- D.H. Huson and T.J. Warnow. Obtaining highly accurate topology and evolutionary estimates of evolutionary trees from very short sequences. Submitted to: *Foundations of Computer Science* (FOCS'98).

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

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*.

Tree Reconstruction Methods

- Maximum Parsimony
  - Popular sequence-based method
  - Solve the Hamming distance Steiner tree problem to obtain the *most parsimonious tree*. (NP-*hard*)
- 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 (*low resolution*)

Maximum Parsimony

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

- For a given tree, find an optimal labeling (easy, using *Fitch's algorithm*)
- Look at *all possible* trees on given sequences, e.g. using branch-and-bound
- Use heuristics such as *branch-swapping*

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 $M(T)$:

One false positive: $\{S_1, S_3\}$ vs. $\{S_2, S_4, S_5\}$

One false negative: $\{S_1, S_2\}$ vs. $\{S_3, S_4, S_5\}$

Experimental Simulation Studies

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

- Choose model of evolution

  Jukes-Cantor model:

  - Markov model
  - Four state character sequences
  - edges have substitution probabilities $p(e)$
  - Root sequence drawn from uniform distribution

- Evolve sequences along the model tree

- Apply tree reconstruction method M to evolved sequences

- Compare estimation $M(T)$ with model tree T

(ecat, PAUP, Phylip. Our programs in C++, LEDA.)

Comparison of False Positive Rates

- sequence length vs. false positive rate
- 93 taxon tree (from 500 taxon *rbc*L dataset)
- maximum substitution probability $p(e)$ is 0.48
- 20 experiments per point

Comparison of False Negative Rates

- sequence length vs. false negative rate
- 93 taxon tree (from 500 taxon *rbc*L dataset)
- maximum substitution probability $p(e)$ is 0.48
- 20 experiments per point

Performance of Neighbor-Joining

- 93 taxon tree (from 500 taxon *rbc*L dataset)

- maximum mutation probability $p(e)$ vs. FP (=FN) rate

- 20 experiments per point

Big Trees are Hard to Infer

- Distance-based methods (e.g. neighbor-joining, 3-approximation, Buneman tree, split decomposition) are fast, but **degrade in accuracy** with high evolutionary divergence.

- Sequence-based methods (e.g. maximum Parsimony and maximum likelihood) do not degrade, but are **computationally expensive**.

- Parsimony does best if all branches are short, so that large numbers of taxa may be needed for accurate tree reconstruction using Parsimony.

- Year-long Parsimony analyses (Rice *et al.*) of large divergent datasets are infeasible for most researchers.

The Disk-Covering Method (DCM)

A divide-and-conquer approach based on the idea of covering given sequence data with small overlapping disks

- Each disk contains a small number of taxa.
- Taxa within a disk are very similar.
- Apply given *base-method* to subproblems.
- Use overlap to merge subtrees to obtain final tree.

The DCM Algorithm

- Input: distances and sequences
- Choose base-method (e.g. Parsimony or NJ)

  - **For a given threshold w:**
    - Compute threshold graph G
      - Vertices are taxa
      - Join two vertices if their distance $\leq$ threshold
    - Compute triangulation G^* of threshold graph
      - Produce perfect elimination scheme
      - Makes the following step easy:
    - Apply base-method to all maximal cliques in G^*
    - Merge trees guided by perfect elimination scheme

- Infer consensus of $\{T_w\}$.

Merging Two Trees

Given trees on two overlapping sets of taxa, e.g. $\{1, 2, 3, 4, 5, 6\}$ and $\{1, 2, 3, 4, 7\}$.

To merge the two trees together, first transform them (through edge contractions) so that they induce the same subtrees on their shared leaves and then combine them.

Neighbor-Joining vs. DCM-NJ

False Positives

- 93 taxon tree
- maximum mutation probability $p(e) = 0.48$
- 10 experiments per point
- *Greedy asymmetric median tree*, i.e. consensus over all trees $\{T_w\}$.

Neighbor-Joining vs. DCM-NJ

False Negatives

- 93 taxon tree
- maximum mutation probability $p(e) = 0.48$
- 10 experiments per point
- *Greedy asymmetric median tree*, i.e. consensus over all trees $\{T_w\}$.

Neighbor-Joining vs. DCM-NJ

- 135 taxon tree
- maximum mutation probability $p(e) = 0.64$
- 4-6 experiments per point
- *Greedy asymmetric median tree* of a small subset of $\{T_w\}$.

Choosing the Threshold for DCM-NJ

Choice of threshold is ruled by two factors:

- The accuracy of NJ degrades on subproblems with increasing threshold w.

- For small thresholds, the merger of subproblems is not uniquely defined.

- 135 taxon tree, $p(e) = 0.64$, sequence length 300

Threshold and Merge Step

Let T be a model tree and d an estimated distance matrix. A *short quartet* around an internal edge e is a set of four taxa a, a', b, b' that lie in the four subtrees induced by e, of minimal width.

Theorem If the threshold w is chosen large enough such that every short quartet induces a four-clique in G^*, then every merger is unique and a DCM method will recover the model tree T, if the base method is accurate on the base problems.

Sequence Lengths Required for Accuracy

The length of biological sequences obtainable for phylogenetic analysis is bounded by a *few thousand* base pairs, so the question how sequence length affects performance is critical.

The sequence lengths that suffice for accuracy of distance methods such as neighbor-joining or the Buneman Tree grow **exponentially** in the divergence of the model tree.
(Atteson 1997, Erdös et al. 1997)

For DCM-boosted distances methods we can show:

> For almost all trees, **polylogarithmic** length suffices for accuracy with high probability, and **polynomial** length suffices for all trees with high probability.

DCM vs. Short Quartet Method

P. Erdös, M. Steel, L. Székely and T. Warnow (1997) introduced the **Short Quartet Method** (SQM), the first method known to require only **polylogarithmic** length sequences for complete accuracy with high probability. Drawback: SQM returns **nothing**, if complete accuracy is unachievable.

- **Average performance** (5 experiments per point) of the SQM compared with DCM-Buneman, on a 35 taxon tree with maximum $p(e)$ equal to 0.04.

- For each dataset, SQM returns **either 0% or 100%** false negatives.

Conclusion and Future Research

By reduction to small and closely related-datasets, the DCM-method can substantially improve the accuracy and/or time requirements of phylogenetic tree reconstruction methods for large and divergent datasets.

Future research will focus on:

- a systematic study of the performance of DCM-NJ on many different simulated datasets
- how to determine a good threshold for DCM-Parsimony
- investigating DCM versions of other methods
- application to some really large problems, e.g. the 500 taxon *rbc*L dataset
- studying different recursive variants
- applying DCM-methods to real biological data sets
- developing a public version of the software.