Higher Fidelity Subtree Merging for Disk-Covering Methods

Nicholas D. Pattengale[1], Krister M. Swenson[1,2]
Monique M. Morin[1], Bernard M.E. Moret[1,2,3]

[1]Department of Computer Science, University of New Mexico, USA
[2]Laboratory for Computational Biology and Bioinformatics, School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland
[3] Swiss Institute of Bioinformatics, Lausanne, Switzerland

Abstract

Disk-Covering methods (DCMs) are a family of divide-and-conquer strategies for large-scale phylogeny reconstruction. Merging subproblems is a critical aspect of all DCMs. In order to avoid introducing unnecessary error, the traditional DCM subtree merging technique is highly conservative. We present a framework for better discerning troublesome components in subproblems. We in turn use our framework as a basis for a new subtree merging algorithm.

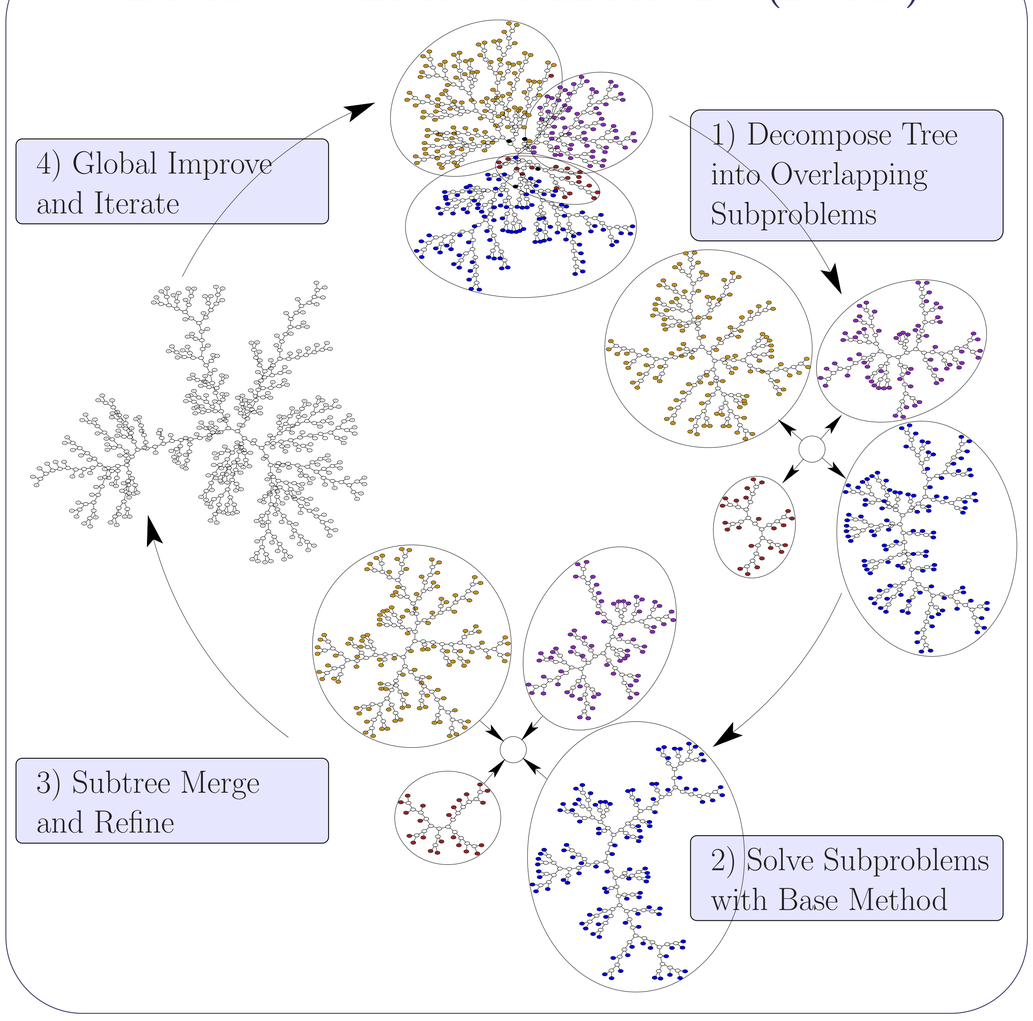
Motivation and Results

- Phylogenetic reconstruction is $\mathcal{NP}$-hard
- Disk Covering Methods (DCMs) use divide and conquer to effectively handle large datasets
- Standard DCM merger induces many polytomies
- A few misplaced ("rogue") taxa are often to blame
- We define an *impact* measure for taxa subsets
- Our new impact-based merger reduces polytomies

Summary

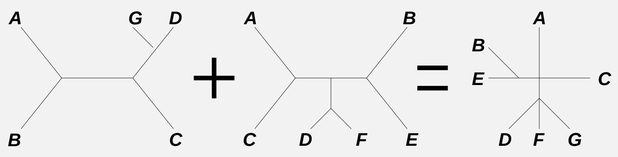
- *We define the* impact *of a subset of taxa informally as the improvement in resolution between the consensus of a tree set $\mathcal{T}$ and the consensus of the induced subtrees by removing U from each tree in $\mathcal{T}$.*
- *We consider a new subtree merger for DCMs that uses* impact *as a basis for temporarily ignoring troublesome components.*

Disk-Covering Methods (DCMs)



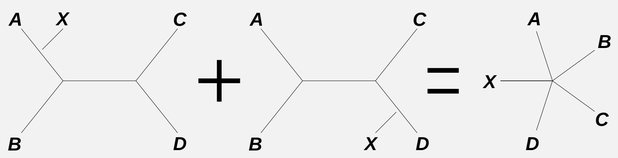
Strict Consensus Subtree Merger

DCM subproblems are merged using a conservative approach that discards all edges (in the subproblem overlap) that are not common to all subproblems. This leads to polytomies.

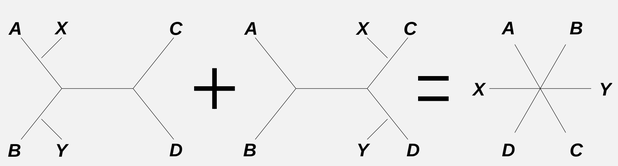


Rogue Taxa

We are interested in detecting situations when removing a small subset (i.e. taxon x) of taxa uncovers strong structural similarities between the induced subtrees.



In the next figure, removing any *single* taxon is not advantageous, whereas removing $\{x, y\}$ together reveals similarity.



Formalizing Taxon Impact

For a set of trees $\mathcal{T}$ (over taxa S) and a consensus method C we define the *impact* of a subset of the taxa $U \subset S$ formally as

$$I(U) = |\Sigma(C(\cup_{T \in \mathcal{T}}(T - U)))| - |\Sigma(C(\mathcal{T}))| + r - \sum_{V \subset U} I(V)$$

where r is a corrective factor to account for non-trivial bipartitions in $C(\mathcal{T})$ becoming trivial when removing U. Important considerations include:

- adding a penalty proportional to $|U|$ to avoid losing too much information
- adding normalization
- the impact function could use the score of an exponential number of subsets of U
- scoring subtrees instead of subsets reduces number of terms to $O(|\mathcal{T}| \cdot |S| \log |S|)$

Removable Taxa Consensus Merger

We now propose a new subtree merger procedure (which is polynomial time, assuming a fixed k):

Given a threshold parameter p and a constant k, calculate the taxa impact for all subsets of size less than or equal to k and of all subtrees. Whenever subsets are identified whose impact exceeds p, exclude the subset from the source trees, and then apply the strict consensus merger as before. In an application regime such as Rec-I-DCM3 the implication is that the reassembled (and subsequently decomposed) tree is missing taxa. However, our aim is that by guiding the search with strong components, we will better accommodate reincorporating rogue taxa.

Future Work

There are many ways to define taxa impact. Here we have presented one definition, however we plan to investigate others. In addition, we plan to implement and perform experimental evaluation of the proposed subtree merger routine in the context of a new disk-covering method.