2004-09-01

Jumpstarting Phylogenetic Analysis

Mark J. Clement
clement@cs.byu.edu

Keith A. Crandall

See next page for additional authors

Follow this and additional works at: https://scholarsarchive.byu.edu/facpub

Part of the Computer Sciences Commons

Original Publication Citation
Jumpstarting Phylogenetic Analysis, Mark Clement, Quinn Snell, Keith Crandall, Kevin Seppi, Biotechnology and Bioinformatics Symposium (BIOT), September 24, page 17-25.

BYU ScholarsArchive Citation
Clement, Mark J.; Crandall, Keith A.; Seppi, Kevin; and Snell, Quinn O., "Jumpstarting Phylogenetic Analysis" (2004). All Faculty Publications. 420.
https://scholarsarchive.byu.edu/facpub/420
Jumpstarting Phylogenetic Analysis

Mark Clement, Quinn Snell, Kevin Seppi, Keith Crandall
3360 TMCB
Brigham Young University
Provo, Utah 84602
clement@cs.byu.edu

clement@cs.byu.edu

Abstract
When a new epidemic strikes, it is often important to determine the relationship between the current organism and others that have been successfully treated previously. The phylogenetic analysis problem generates the most likely family tree for a group of organisms based on DNA sequence data. This process can take a prohibitively long period of time with current algorithms. If trees resulting from prior searches are used to seed the search, correct trees can be found much more quickly. This jumpstarting algorithm can generate superior phylogenetic solutions much more quickly than existing algorithms.

1. Introduction
Phylogenetic analysis has become an integral part of many biological research programs. These include such diverse areas as human epidemiology [20,21], viral transmission [22], biogeography [23], and systematics [24]. With the advent of high speed sequencing equipment, an increasingly large volume of sequence data is becoming available. Scientists should be able to take advantage of this data and also of the research that others have performed. For example, when a new virus is detected, it should be possible to build a phylogenetic tree (an evolutionary history) containing all related viruses and the unknown variety in order to answer questions such as:
• Where did this virus come from?
• When did this virus arrive in the human population?
• What are the related species from which we might derive ideas about appropriate antibodies for testing and remedies for treatment?
• Has this virus been genetically modified through natural or human induced recombinant technology?
• How is this virus evolving and what genetic changes occurred to allow it to successfully enter the human population? This allows us to gain insights into how we can prevent future outbreaks.

Unfortunately, this kind of phylogenetic search is currently computationally infeasible. The time it takes to perform a complete search using maximum likelihood exceeds several months with even a small number of taxa. In the case of the SARS epidemic, and others like it, treatment information must be available in days or at most weeks in order for appropriate action to be taken. Much of the problem comes from the culture and software design for most phylogenetic software packages [14,18,27]. These packages require the user to start a search from scratch every time a new sequence is added to the search (this is exactly the situation when a new antigen is observed). The software packages also do not allow users to share partial trees that could speed up the phylogenetic search process. This creates a culture where investigators see little or no benefit to collaboration in phylogenetic research.

1.1 Phylogenetics

The branching pattern of ancestor/descendant relationships among species or their parts (e.g., genes) is a phylogeny. Researchers attempt to estimate these historical relationships by examining character evolution
using a tree — a mathematical structure used to model the actual evolutionary history of species or their parts [28]. These inferred trees (historical branching relationships) can be represented as cladograms, where branch lengths are arbitrary and only the branching order is significant, or as phylograms, where the branch lengths are proportional to the amount of evolutionary change along the branch.

Phylogenies were historically used to classify organisms into natural evolutionary groups based on these ancestor/descendant relationships. Indeed, great effort is currently being spent on estimating the “tree of life” to quantify the biodiversity of our planet. However, phylogenies have also spread in use as the utility of the evolutionary framework for numerous other disciplines becomes increasingly obvious. For example, phylogenies are now being extensively used in the biomedical sciences including developmental biology, genomic biology, infectious disease, virology, and human genetics.

Phylogenies have become essential tools in the study of the molecular epidemiology of disease agents [30]. A prime example of the troubles encountered when the phylogenetic approach is ignored comes from the outbreak of the West Nile Virus in New York City. This virus was responsible for multiple deaths in New York, yet the Centers for Disease Control and Prevention (CDC) initially misdiagnosed the causative agent as St. Louis encephalitis due to their lack of an appropriate phylogenetic comparison [25]. The study of origins, spread, and diversity of pathogens are clearly evolutionary questions. Only after the serological evidence was coupled with strong phylogenetic evidence was the etiological agent responsible for the encephalitis outbreak in New York correctly identified as the West Nile Virus [29].

Phylogenetic estimation is accomplished by optimizing character change relative to some criterion over a tree. The tree for which the character data show the best optimization is the preferred tree. Two of the principle optimization criteria used by researchers are maximum parsimony and maximum likelihood. The parsimony criterion attempts to minimize the number of changes among a tree for shared-derived characters, while likelihood attempts to maximize the probability of change for all characters relative to some model of evolution. Each criterion has its own strengths and weaknesses. For example, maximum parsimony can incorporate insertion/deletion (indel) events and have asymmetric changes (e.g., a change from character A to character B is not the same as a change from character B to character A), whereas current implementations of maximum likelihood cannot accommodate these biological realities. Likewise, maximum likelihood can account for heterogeneity in evolutionary rates and multiple changes at the same character position, whereas maximum parsimony cannot. Thus there is, often times heated, discussion about appropriate methods to use to estimate phylogenetic relationships.

Another reason there is such debate about phylogenetic methods is that their performance varies depending upon the type of data used, the number of sequences involved, and the depth of the evolutionary relationships to be inferred. Exact searches, those that explore every possible tree topology for a given optimality criterion, are only possible for a very small number of taxa (on the order of 20-30). This limited search is due to the rapidly increasing number of possible trees with a modest increase of taxa [26]. The total number of (unrooted, strictly bifurcating) trees for $T$ taxa is

$$B(T) = \prod_{i=3}^{T} (2i - 5).$$

So, for example, with only 50 sequences, there are $3 \times 10^{74}$ possible trees. For the tree of life, there are estimated to be well over 10 million species, yet for 10 million sequences there are $5 \times 10^{68,667,340}$ possible trees! Therefore, the phylogeny problem is a particularly tough one that is well suited for distributed technology (because one performs the same calculations over different, independent, tree topologies) such as web based systems that utilize distributed resources.
Phylogenetics has become an active field in and of itself [31]. It is an extremely exciting field where talents in mathematics, computer science, and biology can be brought together to work on the problem of inferring historical relationships. A survey of the recent literature in many fields in the biomedical literature will attest to the ever increasing applicability of phylogenetic analyses to these fields.

2. Jumpstarting

The jumpstart algorithm first searches a database for similar searches that have been performed previously. The best tree from the prior search with the maximum subset of common taxa is then used as a starting point for the next search.

Jumpstarting Algorithm:

1) Let $T = \{x \mid x \text{ is in the set of taxa involved in the new search}\}$.
2) Query the data base for prior searches with the set of taxa $S_i$ where at least one of the taxa in the prior search is the same as the new search $\exists x( x \in T \text{ and } x \in S_i)$.
3) For each of these prior searches on taxa $S_i$, determine the intersection $I_i = T \cap S_i = \{x \mid x \in T \text{ and } x \in S_i\}$.
4) Find the intersection $|I_i|$ with the largest number of common taxa.
5) Use the Newick parenthetical notation for the best tree from this maximal intersection as the base tree for the new search.
6) Add taxa $x \in T$ where $x \not\in S_i$ in positions where they minimize the length of the tree.
7) Begin a normal search with the tree from the previous search as a starting place.

Figure 1.1 provides a concrete example of the jumpstart algorithm. In this example, User A on Peer 1 has performed a search resulting in Tree X version 1.1. The following steps are included in the algorithm.

1) User B on Peer 2 prepares a set of taxa that will be used in a phylogenetic search and creates the data structure for Tree Y, Version 1.1. A query is sent to peer machines to determine if searches have already been performed with some of these taxa.
2) Peer 1 has Tree X Version 1.1 which matches the criteria in the query. This tree is returned to Peer 2.
3) Peer 2 uses Tree X Version 1.1 combined with other local taxa to jumpstart a phylogenetic search.
4) After expending significant computational resources, User B generates Tree Y Version 1.2 which refines the relationships between taxa in Tree X as well as Tree Y. This version of the tree is entered into the database.
5) User A has a reference to Tree Y since a subtree of Tree X was used as a jumpstart point for Tree Y. When Tree Y Version 1.2 is generated Peer 1 can send a query for derivative trees of Tree X Version 1.1.
6) Peer 2 will return Tree Y Version 1.2. User 2 may decide that all of the relationships contained in Tree Y Version 1.2 should not be made public. In this case, the subtree containing only the nodes originally found in Tree X would be returned.
7) Peer 1 receives the refined relationships and can create Tree X version 1.2. This tree can be used for future searches.

In this example interaction, both User A and User B have benefited from the collaboration. The tree returned from Peer 2 can be used, or discarded depending on the value that User A places on the results. User B has been able to cut months off of his search time because of the initial jumpstart tree he/she was able to derive from Tree X Version 1.1.
1) Query for trees including Taxa from Tree Y

2) Respond with Tree X containing a subtree that includes taxa from tree Y.

3) The relationships in Tree X are used to refine the starting point for the search on tree Y.

4) Search on Tree Y yields Tree Y V1.2.

5) Query for derivative Trees of Tree X V1.1.

6) Return Tree Y V1.2.

7) Use new relationships to refine Tree X Version 1.2 and as a starting point for future searches.

**Figure 1.1 Example interactions between jumpstarting peers**
3. Experimental Results

Typical phylogenetic research involves the collection of data, sequence alignment, and phylogenetic analysis. Regrettfully, the results obtained in the last two stages in this process are not generally able to be used in subsequent analyses. For example, when a new sequence is added, a new alignment must be computed, and the old trees that were generated in phylogenetic analysis cannot currently be used as a starting point. Both alignment and phylogenetic analysis are time-consuming processes (some analyses have been known to take months). When a researcher adds a sequence to the analysis, all the previous time spent in computation is essentially thrown away because the researcher must start over. This drastically slows the scientific process. Our research group has been studying ways to eliminate the wasted time and take advantage of previous analyses.

The most promising way to “jumpstart” phylogenetic analysis is to use previously computed trees. Basically, the idea is to start where you or another researcher left off rather than starting over. Typically, researchers a) don’t save previously computed trees, and b) don’t have easy means whereby they may use previously computed trees. To determine the benefit of jumpstarting, we randomly selected a sequence to remove from the target dataset. Phylogenetic analysis was then performed on the remaining data and the best trees were obtained. Finally, the removed sequence was added back into the dataset and new trees were generated using our software in conjunction with PAUP*[14]. The time to generate the most optimal tree using jumpstarting was compared with the time required to start from scratch (see Table 3.1). More optimal trees have shorter length values.

Table 3.1 Preliminary jumpstart results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Strategy</th>
<th>Paup* time (seconds)</th>
<th>Paup* length</th>
<th>Jumpstart time (seconds)</th>
<th>Jumpstart length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zilla</td>
<td>Parsimony</td>
<td>288054</td>
<td>16220</td>
<td>510</td>
<td>16218</td>
</tr>
<tr>
<td>HIV</td>
<td>Parsimony</td>
<td>2410</td>
<td>110709</td>
<td>1816</td>
<td>110658</td>
</tr>
<tr>
<td>SARS</td>
<td>Likelihood</td>
<td>30605</td>
<td>542771.9</td>
<td>127</td>
<td>542771.9</td>
</tr>
</tbody>
</table>

Clearly, jumpstarting is advantageous. In all cases, jumpstarting resulted in trees that could not be found by starting from scratch. In the case of the Likelihood search with SARS data, the collaborative search was performed 240 times faster than the search without jumpstart trees. In the experiments with parsimony, not only was the search time drastically reduced, but the quality of the trees (smaller tree lengths) was also improved. More research is needed with a larger number of data sets to determine average performance improvements, but initial results indicate that collaboration in this area can have a significant impact on the quality of and time taken for phylogenetic searches.

Jumpstarting is advantageous when adding sequences to an existing analysis. However, this is not the only use case. More typically, a researcher may request data from the database and wish to begin computation by taking advantage of these data. A user can request all trees that contain certain taxa or sequences from the database. However, these trees may also contain extraneous taxa. One approach is to simply strip out the extraneous samples and use the resulting trees as a starting point. A consensus tree could also be used.
Figure 3.1 Comparison of various jumpstart methodologies

Figure 3.1 demonstrates the different jumpstarting possibilities available. A researcher may choose to start computation based on one of the most optimal (Smallest) trees returned from the query. Optionally, a consensus tree may be created and used for jumpstarting the new search (Strict, 50%). Preliminary studies show that creating a majority rule consensus tree (50%) from the collection of most optimal trees returned by the jumpstart system seems to be the best option. Figure 3.1 shows that this method found the most optimal tree in 17 minutes, whereas the other choices took at least 57 hours. Increases in performance such as this are vital to advances and to apply the power of the phylogenetic approaches to studies in biomedical research.

4. Conclusions

Jumpstarting is an effective algorithm for improving search times and tree quality in phylogenetic analysis. It can be combined with various search algorithms to deal with important medical problems. Future work will investigate ways of mining the database and combining existing trees to provide the best starting point for future searches.
5. References

5. NIH Collaboratory Project, http://collaboratory.psc.edu


SourceForge (2003), http://sourceforge.net/

BugZilla (2003), http://bugzilla.mozilla.org/
