



# Large-Scale Co-Phylogenetic Analysis on the Grid

*Heinz Stockinger, Swiss Institute of Bioinformatics, Switzerland*

*Alexander F. Auch, University of Tübingen, Germany*

*Markus Göker, University of Tübingen, Germany*

*Jan Meier-Kolthoff, University of Tübingen, Germany*

*Alexandros Stamatakis, Ludwig-Maximilians-University Munich, Germany*

---

## ABSTRACT

*Phylogenetic data analysis represents an extremely compute-intensive area of Bioinformatics and thus requires high-performance technologies. Another compute- and memory-intensive problem is that of host-parasite co-phylogenetic analysis: given two phylogenetic trees, one for the hosts (e.g., mammals) and one for their respective parasites (e.g., lice) the question arises whether host and parasite trees are more similar to each other than expected by chance alone. CopyCat is an easy-to-use tool that allows biologists to conduct such co-phylogenetic studies within an elaborate statistical framework based on the highly optimized sequential and parallel AxParafit program. We have developed enhanced versions of these tools that efficiently exploit a Grid environment and therefore facilitate large-scale data analyses. Furthermore, we developed a freely accessible client tool that provides co-phylogenetic analysis capabilities. Since the computational bulk of the problem is embarrassingly parallel, it fits well to a computational Grid and reduces the response time of large scale analyses.*

*Keywords:* bioinformatics; co-phylogenetic analysis; Grid computing; phylogeny

---

## INTRODUCTION

The generation of novel insights in many scientific domains such as biology, physics, or chemistry increasingly relies on compute-intensive applications that require high-performance or large-scale, distributed high-throughput computing technology and infrastructure. In the discipline of bioinformatics, biological insight is typically generated via data analysis pipelines

that use a plethora of distinct and highly specialized tools. Most commonly, bioinformaticians and biologists collaborate to analyze data extracted from large databases containing DNA and/or protein data in order to study, e.g., the function of living beings, the effect and influence of diseases and defects, or their evolutionary history. Early “classic” bioinformatics tools, such as CLUSTALW (Thompson et al., 1994) or BLAST (Altschul et al., 1997) that have

been ported to Grid computing environments deal with biological sequence search, analysis, and comparison. Typically, these programs are embarrassingly parallel and therefore represent ideal candidate applications for Grid computing environments (Stockinger et al., 2006).

The study of the genome represents a way to obtain new insight and extract novel knowledge about living beings. In particular, stand-alone phylogenetic analyses have many important applications in biological and medical research. Applications range from predicting the development of emerging infectious diseases (Salzberg et al., 2007), over the study of Papillomavirus evolution that is associated with cervical cancer (Gottschling et al., 2007), to the determination of the common origin of Caribbean frogs (Heinicke et al., 2007).

Recent years have witnessed significant progress in the field of stand-alone phylogeny reconstruction algorithms, which represent an NP-complete optimization problem (Chor and Tuller, 2005), with the release of programs such as TNT (Goloboff, 1999), RAXML (Stamatakis, 2006), MrBayes (Ronquist and Huelsenbeck, 2003) or GARLI (Zwickl, 2006). Because of the continuous explosive accumulation and availability of molecular sequence data coupled with advances in phylogeny reconstruction methods, it has now become feasible to reconstruct and fully analyze large phylogenetic trees comprising hundreds or even thousands of sequences (organisms). However, current meta-analysis methods for phylogenetic trees such as programs that conduct co-phylogenetic tests can currently not handle such large datasets.

To alleviate this bottleneck in the meta-analysis pipeline, we recently parallelized, and released the highly optimized co-phylogenetic analysis program AxParafit (Axelerated Parafit - Stamatakis et al., 2007) that implements an elaborate statistical test of congruence between host and parasite trees (Legendre et al., 2002). AxParafit is a typical stand-alone Linux/Unix command line program. AxParafit has been integrated and can be invoked via a user-friendly graphical interface for co-phylogenetic analyses called CopyCat (Meier-Kolthoff et al.,

2007). In this article, we present an enhanced version of this tool suite (henceforth denoted as CopyCat(AxParafit)) for co-phylogenetic analyses, that is packaged into a client tool which makes use of a world-wide Grid environment and thereby allows for large-scale data analysis. In the current version, the underlying Grid middleware is gLite (Laure et al., 2006) that is coupled with an efficient submission and execution model called Run Time Sensitive (RTS) scheduling and execution (Stockinger et al., 2006).

The remainder of this article is organized as follows: initially, we provide a brief introduction to the field of phylogenetic inference, co-phylogenetic analyses, and related software packages in Section 2. Next, we discuss the implementation and architecture of our new approach for efficient adaptation of the CopyCat(AxParafit) tool-suite to a Grid environment. Finally, we provide detailed performance results on the EGEE (Enabling Grids for E-Science, <http://www.eu-egee.org>) Grid infrastructure (where the gLite middleware is deployed in production mode) and demonstrate the performance as well as scalability of our proposed bioinformatics tool.

## BACKGROUND

Phylogenetic (evolutionary) trees are used to represent the evolutionary history of a set of  $s$  currently living organisms, roughly comparable to a genealogical tree of species rather than individual organisms. Phylogenetic trees or simply phylogenies are typically unrooted binary trees. The  $s$  organisms, which are represented by their DNA or AA (Amino Acid/Protein) sequences that are used as input data for the computation, are located at the leave nodes (tips) of the tree while the inner nodes of the topology represent common extinct ancestors. There exist various methods and models to reconstruct such trees which differ in their computational complexity and also in the accuracy of the final results, i.e., there exists a “classic” trade-off between speed and accuracy. As already mentioned in

14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/article/large-scale-phylogenetic-analysis-grid/2167](http://www.igi-global.com/article/large-scale-phylogenetic-analysis-grid/2167)

## Related Content

---

### Performance Evaluation of WMN-GA Simulation System for Different Settings of Genetic Operators Considering Giant Component and Number of Covered Users

Admir Barolli, Makoto Takizawa, Tetsuya Oda, Evjola Spaho, Leonard Barolli, Kazunori Uchida and Fatos Xhafa (2012). *International Journal of Distributed Systems and Technologies* (pp. 1-14).

[www.irma-international.org/article/performance-evaluation-wmn-simulation-system/67554](http://www.irma-international.org/article/performance-evaluation-wmn-simulation-system/67554)

### Design and Implementation of Hybrid Time (HT) Group Communication Protocol for Homogeneous Broadcast Groups

Isamu Tsuneizumi, Ailixier Aikebaier, Makoto Ikeda, Tomoya Enokido and Makoto Takizawa (2013). *Development of Distributed Systems from Design to Application and Maintenance* (pp. 282-293).

[www.irma-international.org/chapter/design-implementation-hybrid-time-group/72259](http://www.irma-international.org/chapter/design-implementation-hybrid-time-group/72259)

### An Integrated Fuzzy-Based System for Cluster-Head Selection and Sensor Speed Control in Wireless Sensor Networks

Miralda Cuka, Donald Elmazi, Takaaki Inaba, Tetsuya Oda, Makoto Ikeda and Leonard Barolli (2017). *International Journal of Distributed Systems and Technologies* (pp. 1-14).

[www.irma-international.org/article/an-integrated-fuzzy-based-system-for-cluster-head-selection-and-sensor-speed-control-in-wireless-sensor-networks/179571](http://www.irma-international.org/article/an-integrated-fuzzy-based-system-for-cluster-head-selection-and-sensor-speed-control-in-wireless-sensor-networks/179571)

### Flexible MapReduce Workflows for Cloud Data Analytics

Carlos Goncalves, Luis Assuncao and Jose C. Cunha (2013). *International Journal of Grid and High Performance Computing* (pp. 48-64).

[www.irma-international.org/article/flexible-mapreduce-workflows-for-cloud-data-analytics/102756](http://www.irma-international.org/article/flexible-mapreduce-workflows-for-cloud-data-analytics/102756)

### Supercomputing in the Study and Stimulation of the Brain

Laura Dipietro, Seth Elkin-Frankston, Ciro Ramos-Estebanez and Timothy Wagner (2021). *Handbook of Research on Methodologies and Applications of Supercomputing* (pp. 290-300).

[www.irma-international.org/chapter/supercomputing-in-the-study-and-stimulation-of-the-brain/273408](http://www.irma-international.org/chapter/supercomputing-in-the-study-and-stimulation-of-the-brain/273408)