Cyber Infrastructure for Phylogenetic Research

Fran Berman[1], Bernard Moret[2], Satish Rao[3], David Swofford[4], Tandy Warnow[5]

Keywords: phylogenetics, phyloinformatics, algorithms, biological dataset analysis

1 Introduction.

CIPRes (Cyber Infrastructure for Phylogenetic Research) is an NSF-funded community effort to design and build an integrated environment for large-scale phylogenetic analysis.

The environment will integrate high-performance computing platforms, large databases, biological datasets and their analyses, benchmark datasets, optimization software, and a flexible user interface. It will serve both biologists carrying out analyses of biological data and algorithm designers developing and testing new phylogenetic reconstruction methods.

The collaboration involves directly 13 universities and museums and 33 researchers in North America and indirectly many more institutions and individuals worldwide. CIPRes researchers coordinate closely with national and international initiatives for the reconstruction Tree of Life -- an ambitious project to reconstruct the evolutionary history of all living species (in the tens of millions).

2 Project Overview

CIPRes will be composed of a large computational platform, a collection of interoperable high-performance software for phylogenetic analysis, and a large database of datasets (both real and simulated) and their analyses; it will be accessible through any web browser by developers, researchers, and educators. The software, freely available in source form, will be usable on scales varying from laptops to high-performance, Grid-enabled, compute engines such as our platform, and will be packaged to be compatible with current popular tools. In order to build this resource, CIPRes will support research programs in phyloinformatics (databases to store multilevel data with detailed annotations and to support complex, tree-oriented queries), in optimization algorithms, Bayesian inference, and symbolic manipulation for phylogeny reconstruction, and in simulation of branching evolution at the genomic level, all within the context of a virtual collaborative center.

Biology, and phylogeny in particular, has been almost completely redefined by modern information technology, both in terms of data acquisition (new genomic data accumulates at a rate exceeding

[1] Department of Computer Science/San Diego Supercomputer Center, University of California at San Diego. La Jolla, California, USA. E-mail: berman@cs.ucsd.edu

[2] Department of Computer Science, University of New Mexico, Albuquerque, New Mexico, USA. E-mail: moret@cs.unm.edu

[3] Computer Science Division, University of California at Berkeley, Berkeley, California, USA. E-mail: satishr@cs.berkeley.edu

[4] Department of Biological Science, Florida State University, Tallahassee, Florida, USA. E-mail: swofford@csit.fsu.edu

[5] Department of Computer Sciences, University of Texas at Austin, Austin, Texas, USA. E-mail: tandy@cs.utexas.edu

Moore's law) and in terms of analysis (the literature shows over 10,000 citations to the top three phylogenetic software packages). Phylogeneticists have formulated specific models and questions that can now be addressed using recent advances in database technology and optimization algorithms. The time is thus exactly right for a close collaboration of biologists and computer scientists to address the IT issues in phylogenetics, many of which call for novel approaches, due to a combination of combinatorial difficulty and overall scale. CIPRes includes computer scientists working in databases, algorithm design, algorithm engineering, and high-performance computing, evolutionary biologists and systematists, bioinformaticians, and biostatisticians, with a history of successful collaboration and a record of fundamental contributions, to provide the required breadth and depth.

CIPRes brings together researchers from many areas and foster new types of collaborations and new styles of research in computational biology; moreover, the interaction of algorithms, databases, modeling, and biology will give new impetus and new directions in each area. CIPRes will help create the computational infrastructure that the research community will use over the next decades, as more whole genomes are sequenced and enough data is collected to attempt the inference of the Tree of Life. It will help evolutionary biologists understand the mechanisms of evolution, the relationship between evolution, structure, and function of biomolecules, and a host of other research problems in biology, eventually leading to major progress in ecology, pharmaceutics, forensics, and security (including computer security). The project will publicize evolution, genomics, and bioinformatics through informal education programs at our museum partners and will motivate high school students and college undergraduates to pursue careers in bioinformatics. CIPRes provides an extraordinary opportunity to train students, both undergraduate and graduate, as well as postdoctoral researchers, in one of the most exciting interdisciplinary areas in science; our institutions serve a large number of underrepresented groups and are committed to increase their participation in research.

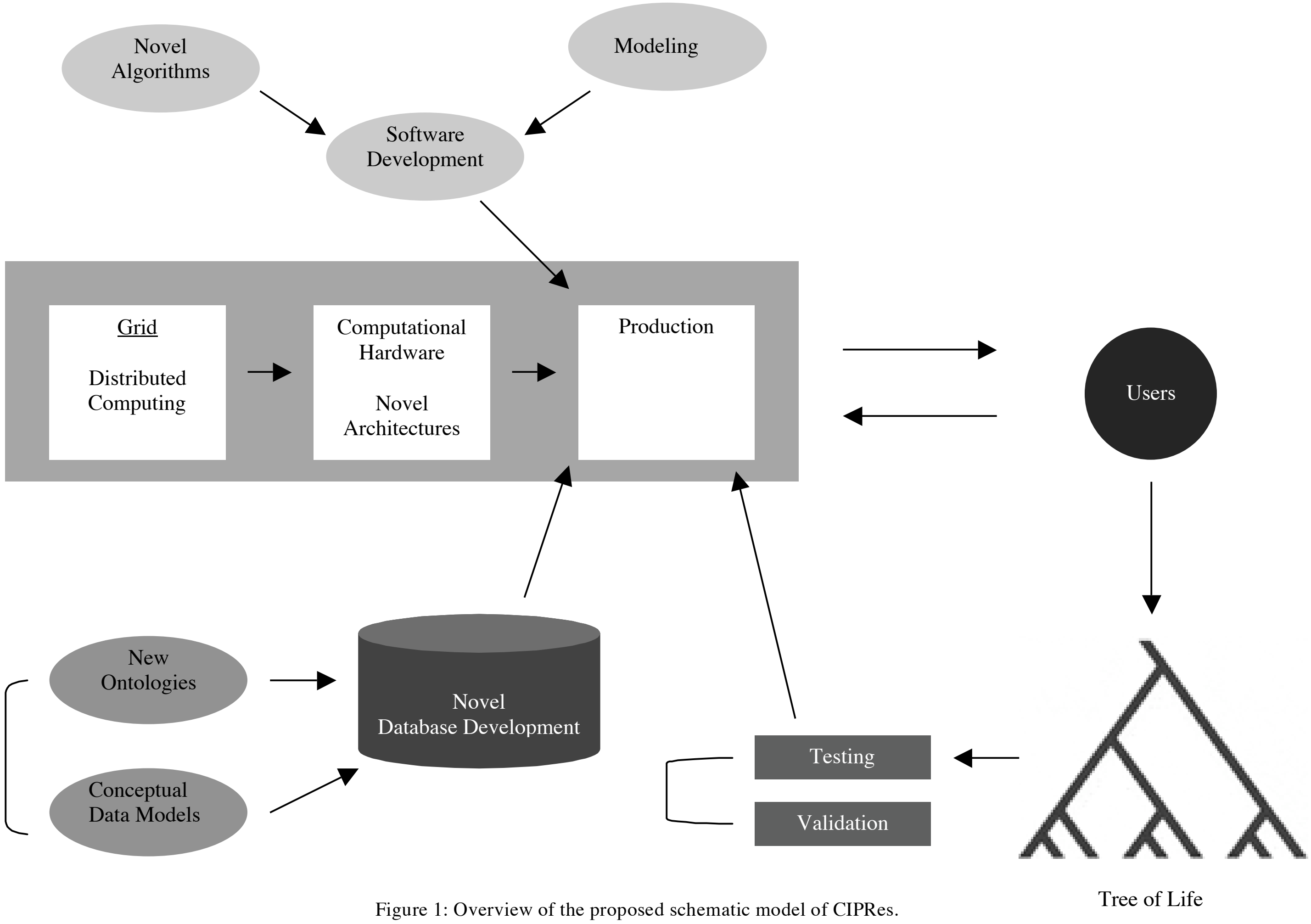

Figure 1: Overview of the proposed schematic model of CIPRes.