
CIPRES software development

All Hands Meeting
Berkeley, CA
July 2009

CIPRES software

- CIPRES portal
- interoperability tools
- advances to standalone software tools

Talk outline

- history of CIPRES software development
- brief discussion of the CIPRES portal
- brief run-through of software products that will not be covered in other talks later today

CIPRES software development history

The broad goals of the software effort were to develop tools and strategies that enable large scale phylogenetic inference.

The software strategy was to focus on integration of tools from the community of phylogenetic software developers.

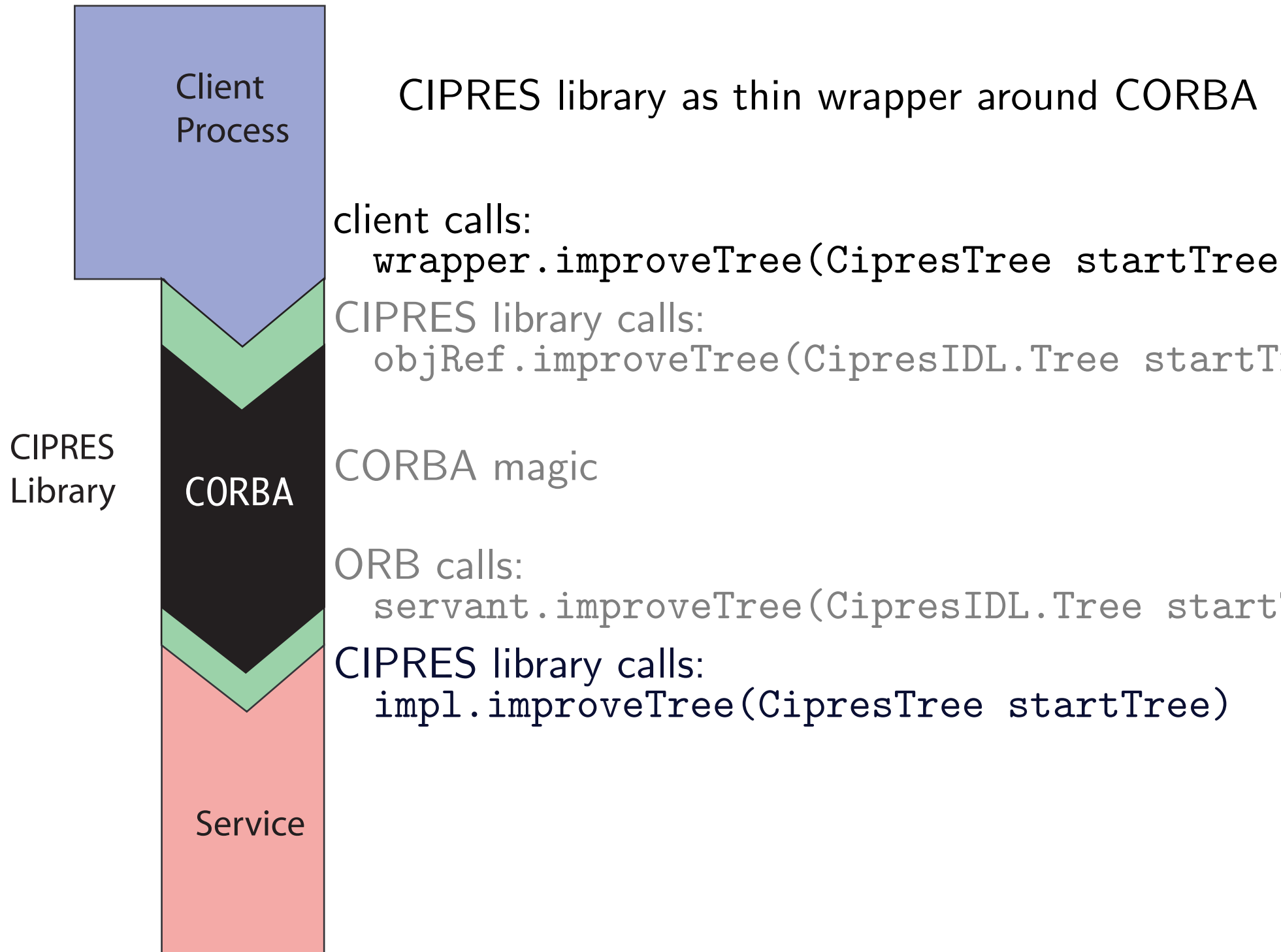
CIPRES software participants

- The core development team at SDSC, FSU and UBC worked on integrated CIPRES software;
- Postdocs and graduate students at other institutions worked fairly autonomously on their tools.

CIPRES library architecture

- Coarse-grained, modular design.
- Modules implement one of several defined interfaces (e.g. TreeInfer, TreeMerge, ReadNexus, ...).
- Inter-module communication occurs via a CIPRES library built upon CORBA, a mature protocol for inter-process and across-the-network communication.
- Multiple front-ends
- Generic tools for rendering tool-specific options for a tool to the user (based on transformation of XML).

CIPRES library as thin wrapper around CORBA



Hurdles

The CIPRES library has been operational (and in use) since 2006, however the original architecture had several drawbacks:

- slow and complex development cycle required a programmer to be conversant in several languages;
- complex deployment (excessive time spent developing configuration tools to make the software installation robust and easy for novice users); and
- complexity of the library was intimidating to external programmers

Portal development

- Starting in early 2007, the SDSC team focused on supporting a central portal at SDSC.
- Cipres Portal version 1 went online in May, 2007. It was built open the CIPRES libraries (communication tools and GUI generation from XML).
- GARLI, RAxML, and PAUP-ratchet searches were original tree inference services
- Rec-I-DCM3 boosting of tree construction methods added in the summer of 2007
- RAxML rapid bootstrapping algorithm (Stamatakis et al., 2008)
- In late 2007 we began to move to the Next Generation Biology Workbench architecture:
 - the largest number of user requests were for greater tool configurability;
 - thinner wrapping of tools make it easier to add new tools;
 - GUI-generation from PISE schema allows the portal to use tools wrapped by others.

Interoperability work

Problem: Current text formats in phylogenetics are not rich enough to express the connections between the relevant entities.

The result is redundant interfaces and shearing of useful metadata when results are returned.

Rutger Vos has led efforts to improve interoperability:

- Arlin Stoltzfus and Rutger organized the evolutionary informatics working group at NESCent
- NeXML file format (Vos and Midford: Java support; Vos: Perl support; Sukumaran: Python support; Holder: NeXML output from NCL)
- PhyloWS API (with Hilmar Lapp)
- extensions to the Nexus Class Library (Lewis, 2003)

Other software products and related research

Packages developed or expanded with the help of CIPRES include:

Phycas	RAxML	Phyutility
POY	GARLI	TASPI
TreeBase II	Rec-I-DCM3	SATé
RNASim	Mesquite	Probalign
Crimson	DendroPy	Superfine

Rapid bootstrap algorithm of RAxML (Stamatakis et al., 2008)

- Uses rapid search settings and starts a pseudo-replicate dataset's search from the final tree of the previous replicate;
- yields bootstrap proportions almost identical to standard bootstrapping;
- at least an order of magnitude faster;
- tested on trees with over 7000 leaves;
- freely available on the CIPRES portal (and a portal in Switzerland).

GARLI metapopulation parallelization scheme

(This work is still in-progress). Derrick Zwickl is implementing a new parallelization scheme in the current version of GARLI (Zwickl, 2006)

- Metapopulation structure;
- Subpopulations assigned to different nodes in a parallel environment;
- Recombination of topologies on the master node;
- An implementation of this scheme in a previous GARLI version was able to return better trees than found by the normal GARLI or RAxML.

Rec-I-DCM3

A divide-and-conquer method by Roshan et al. (2004) that can boost the performance of other search strategies on large datasets.

- Accelerates RAxML and PAUP searches on large datasets;
- Available on the CIPRES portal for use with RAxML, and PAUP-ratchet

Mesquite

The most significant additions to Mesquite from CIPRES funds were support for NeXML and the BiSSE module (Maddison et al., 2007):

- Binary-State Speciation Extinction model;
- Takes a tree and tips scored for a binary character;
- Estimates the speciation and extinction rates associated with each state of the character;
- Can be used to test whether or not an observable character is influencing divergence rates.

DendroPy

Python package by Jeet Sukumaran and Mark Holder

- NeXML parsing in Python;
- tree manipulation with efficient splits representations;
- in use by Ginkgo (a biogeography simulator by Sukumaran), SATé implementation (the implementation by Jiaye Yu and Mark Holder), SumTrees (tree summarization tool by Sukumaran), and incremental phylogenetics routines by Holder.

Phyutility

Java package by Smith and Dunn (2008) for manipulating trees and alignments.

Actively developed as an open source tool:

<http://code.google.com/p/phyutility/>

TASPI

Efficient storage of large collections of trees (Boyer et al., 2005)

- Uses shared references to common subtrees to compress collections of trees;
- makes dramatic speedups of post-tree analysis feasible by enabling on memoization;

written in ACL2 !

SATè

Simultaneous alignment and tree estimation by Liu et al. (2009)

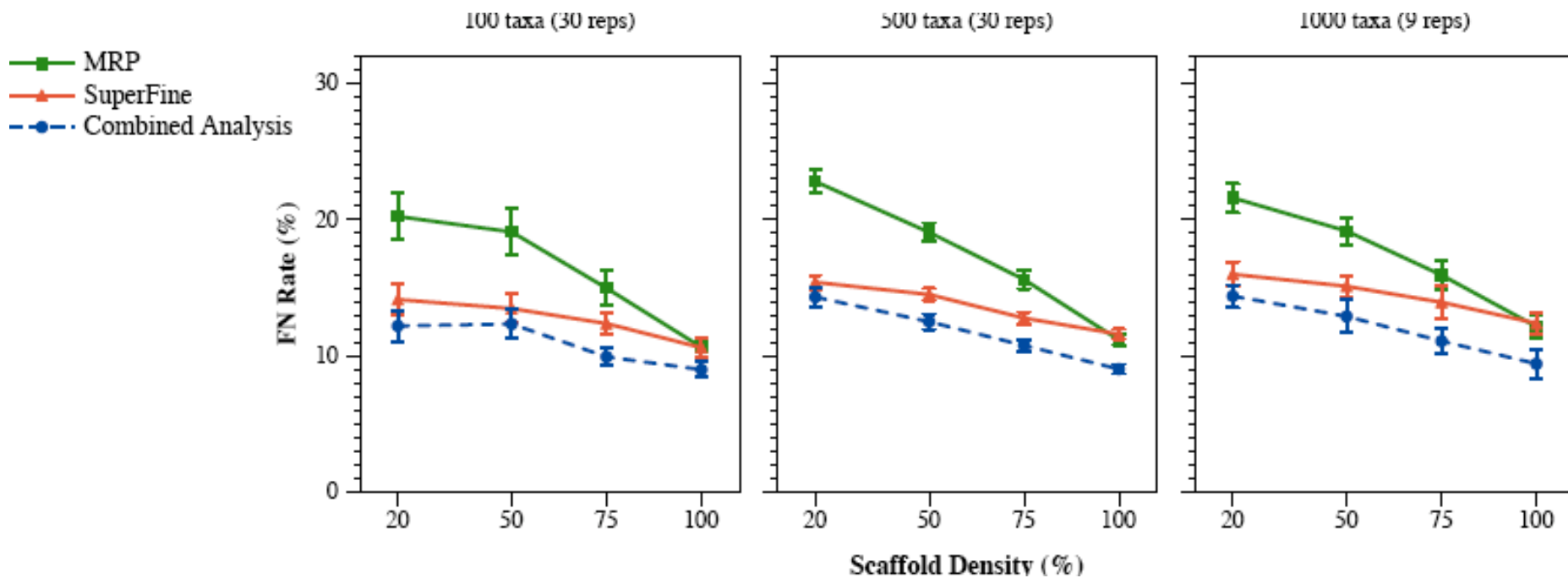
- Also funded by an AToL award to Warnow and collaborators;
- Iterative procedure of
 - tree-based decomposition of an alignment,
 - re-alignment of sequences in sub-problems,
 - merger of alignments, and
 - tree searching;
- When used in conjunction with RAxML and MAFFT, Liu et al. (2009) report more accurate tree inference than with other approaches.
- Alpha-testing version available from links on:
<http://phylo.bio.ku.edu:5000>

Probalign

- Multiple sequence alignment tool using pairwise posterior probabilities as calculated by an alignment partition function Roshan and Livesay (2006);
- Accuracy tested against BALiBASE, HOMSTRAD, and OXBENCH and found to outperform other methods on protein with very heterogeneous lengths.

Superfine

A supertree method by Swenson *et al* (2009) that uses strict consensus merger and quartet maxcut approaches.



References

- Boyer, R., Hunt, W., and Nelesen, S. (2005). A compressed format for collections of phylogenetic trees and improved consensus performance. *Lecture Notes In Computer Science*, 3692:353–364.
- Lewis, P. O. (2003). NCL: a C++ class library for interpreting data files in NEXUS format. *Bioinformatics*, 18:2330–2331.
- Liu, K., Raghavan, S., Nelesen, S., Linder, C., and Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564.
- Maddison, W., Midford, P. E., and Otto, S. E. (2007). Estimating a binary character's effect on speciation and extinction. *systematic biology*. *Systematic Biology*, 56(5):701–710.
- Roshan, U. and Livesay, D. (2006). Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, 22(22):2715–2721.

Roshan, U., Moret, B. M. E., Williams, T. L., and Warnow, T. (2004). Rec-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. *Proceedings of IEEE Computer Society Bioinformatics Conference (CSB 2004)*.

Smith, S. and Dunn, C. (2008). Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics*, 24(5):715.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, 57(5):758–771.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, The University of Texas at Austin.