

ART

Victor Hanson-Smith & Katie Ray

victorhs@cs.uoregon.edu, kray@cs.uoregon.edu
The University of Oregon, Computer and Information Sciences

with advising from **Joe Thornton** and **Bryan Kolaczkowski**,
at the University of Oregon Center for Ecology and Evolutionary Biology

and special thanks to **John Conery**

Our **Ancestral Reconstruction Tool** streamlines phylogenetic data analysis tasks.

BACKGROUND (What is ancestral reconstruction?)

Phylogeneticists can resurrect ancient DNA sequences by a process called "ancestral reconstruction." Recreating extinct amino acid sequences allows scientists to better understand evolutionary pathways. Ancestral reconstruction begins with three pieces of information: a mathematical model describing the evolutionary process, a set of phylogenetic trees that relate ancestors to descendants, and the genetic sequences of known descendants. This information can be combined to determine the likelihood of each ancestral tree, and to recreate the genetic sequences for specific ancestors. The impetus lies on "wet-lab" scientists to experimentally validate resulting sequences.

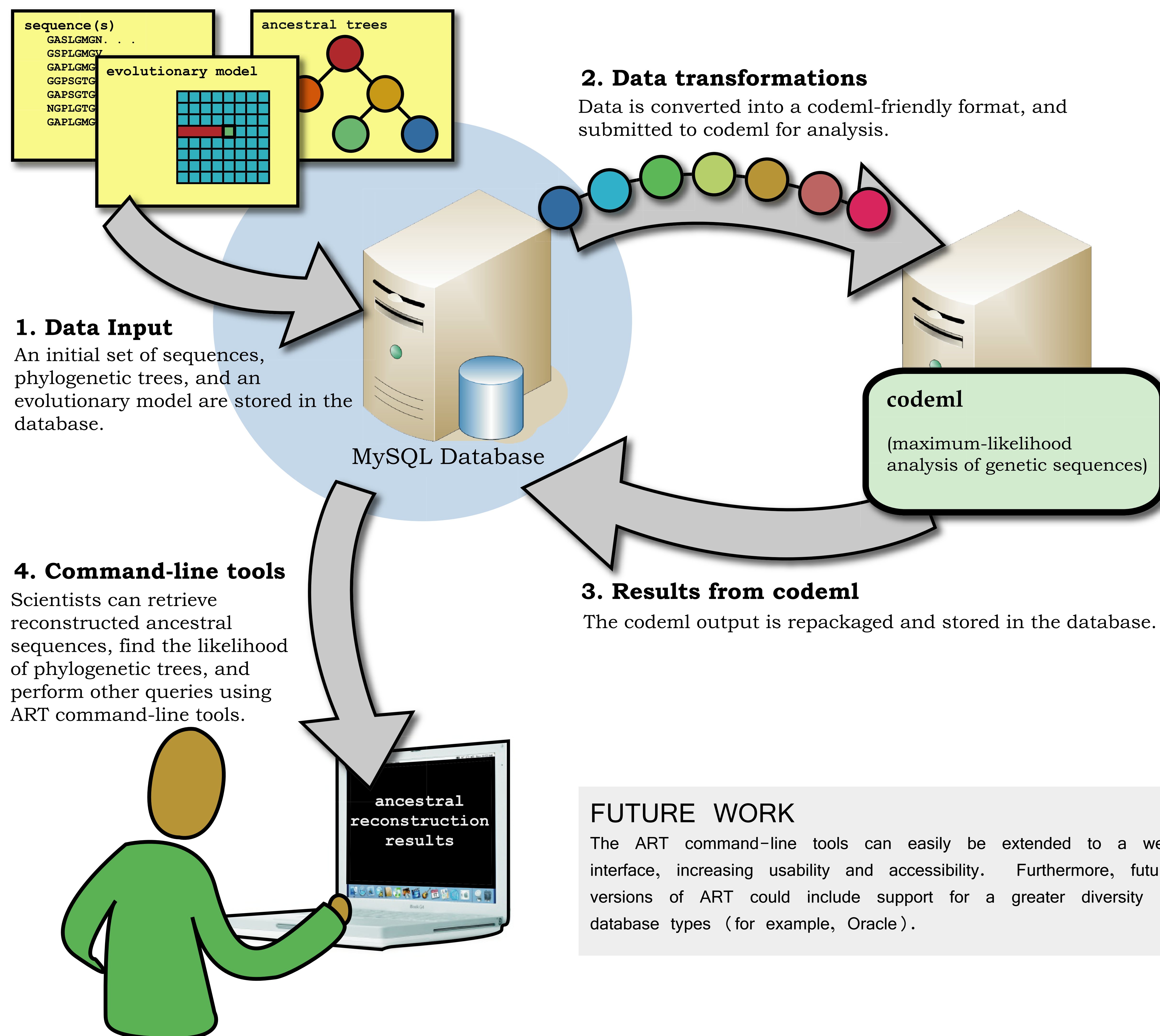
The most common methods for inferring ancient phylogenies are parsimony, maximum-likelihood (ML), and Bayesian inference. It is generally agreed that parsimony is the weakest of these three methods, while the debate continues over ML versus Bayesian inference. PAML (Phylogenetic Analysis by Maximum Likelihood) is a popular software package that uses the maximum-likelihood method.

MOTIVATION

CodeML is one of the tools embedded in the PAML suite. Unfortunately, using codeML can be laborious because input must be in a particular format. Furthermore, output arrives in a manner that is incompatible with databases and not easily understood by human readers. ART is a data wrapper around codeML, and simplifies the task of preparing input and retrieving output. ART helps phylogeneticists spend less time managing their data, and more time using it.

HOW ART WORKS

In the first stage, ART stores an evolutionary model, a set of hypothesized phylogenetic trees, and a list of genetic sequences for known species. In the second stage, ART transforms this data into a format that is convivial with codeML. ART launches codeML and saves the output in the third stage. In the final stage, scientists use the ART command-line tools to view reconstructed ancestral sequences, tree likelihoods, and other post-codeML information. Users can customize the output based on specific confidence intervals, probability cutoffs, and other statistical parameters. CodeML can produce huge output-sets, but ART keeps the results manageable.



FUTURE WORK

The ART command-line tools can easily be extended to a web interface, increasing usability and accessibility. Furthermore, future versions of ART could include support for a greater diversity of database types (for example, Oracle).