Database gives access to the latest findings about the tree of life

March 29 2010

If scientists have identified some two million species, where can you find the latest information about the tree of life that unites them all? A vastly improved database gives scientists and educators access to state-of-the-art knowledge about the evolutionary relationships among living things.

TreeBASE — a database designed to help scientists store, share, and study evolutionary trees — was first developed in the mid-1990s as way to archive the vast amounts of phylogenetic information accumulating in the literature.

"Phylogenies were being published at an explosive rate," said Bill Piel of Yale University. "What we needed was a database where we could compile them so people could use them later."

The database allows researchers to archive and retrieve published phylogenetic trees and data from different studies. "People can store sequence alignments, morphological character sets, and the resulting phylogenetic trees - all in digital form. They can also be recovered and reanalyzed or combined with other data," Piel said.

Since the first prototype was developed, researchers have contributed more than 6,500 trees from over 2400 articles, describing the relationships among well over 60,000 terminal taxa. A variety of journals now require their authors to deposit phylogenetic data in TreeBASE, and peer reviewers are given anonymous access to the data

prior to publication.

Years of work have gone into improving and upgrading the original version. "At some point we knew we had to make it bigger and better," said Michael Donoghue of Yale University. Now, a team of biologists and computer scientists is releasing a new version that is completely rebuilt. With this upgrade, the database is poised to become an increasingly valuable resource for a number of fields, including conservation biology, biogeography, and education, developers say.

"We have introduced a wide variety of features that didn't exist before," said Val Tannen at the University of Pennsylvania. "In terms of data deposition and how users interact with it, it has taken a huge leap forward," Donoghue added.

For one, TreeBASE can now store much richer information. "Trees can contain information such as the length of each branch, which is important for studying the timing of evolutionary events," Piel explained. The database also has an improved system for making sure that information such as taxonomic names and DNA sequence IDs match those found in other sources.

Researchers will also be able to take advantage of a more user-friendly interface and more advanced search techniques. "There are things you can query now that you couldn't before," said Piel. "For example, you can search for trees that share a certain topology."

"The visualization tools have also received a major upgrade," Piel added. "For example, now users can manipulate large trees and zoom in and out."

A number of advanced features have also been introduced that will allow bioinformaticians to do new and creative things with the data without

being blocked by the user interface, said Piel. These include support for new machine-readable phylogenetic data exchange and web service standards. In addition, the metadata in TreeBASE are being made available for harvesting en masse.

According to Rutger Vos of the University of Reading, "all these features basically mean that TreeBASE plays nice with other Linked Data resources on the web, allowing the next generation of web applications to automatically understand the connections among different biological data resources."

In addition to getting a major makeover, the database also has a new home. Most recently housed at the San Diego Supercomputer Center with support from the CIPRES project, TreeBASE is now being hosted by the National Evolutionary Synthesis Center (NESCent) in Durham, North Carolina.

NESCent has made an initial commitment to host TreeBASE for up to five years, explained Todd Vision, Associate Director of Informatics at NESCent. "This partnership enables TreeBASE to continue serving the scientific needs of the community and to keep pace with technological innovations," said Vision.

Looking to the future, the team has established a non-profit foundation to ensure the database's long-term sustainability. "The foundation will become a caretaker of TreeBASE and other phylogenetic resources, such as the Tree of Life Web (ToLWeb) project," said Piel.

To enable wider participation in TreeBASE's future development, the code has been made open source and is hosted by SourceForge. The developers now communicate on a public forum. "In essence, this allows anyone with the necessary skills to participate in TreeBASE development, whether small or large," says Hilmar Lapp, Assistant

Director for Informatics at NESCent.

"The really good news is that we now have a much better product - much more stable, much more industrial-strength - and we have an arrangement with NESCent that's going to be very successful," Donoghue added. "Now that we have a new home for it, we can service it, we can build it, and we can continue to modify it," he added. "This is a good place to be right now."

More information: TreeBASE II was released in March 2010 and is freely available online at www.treebase.org

Provided by National Evolutionary Synthesis Center