Construction of the Application Environment for Phylogenetic Analysis of Land Plants

Zhen MENG, Xiao-Guang LIN, Yuan-Chun ZHOU, Xing HE, Yan-Ping GAO, Yong LIU, Qi LIU, Zhi-Duan CHEN, Shou-Zhou ZHANG, Yong LI & Jian-Hui LI

Chinese Academy of China, CN

Tree of Life (ToL) is a phylogenetic tree which compiles information about biodiversity and the evolutionary relationships of all organisms on the earth (living and extinct). With such a framework, we can present information about every species and significant group of organisms on earth, learn about and appreciate biological diversity and the evolutionary tree of life, and more significantly is we can also share information with other databases and analytical tools, and to phylogenetically link information from other databases.

However, how to construct the TOL and fully exploit its huge information is still a big challenge.

Phylogenetic Analysis of Land Plants Platform (PALPP) is initiated and compiled collaboratively by Computer Network Information Centre, Chinese Academy of Sciences (CNIC, CAS), Institute of Botany, Chinese Academy of Sciences (IB, CAS) and ShenZhen Fairylake Botanical Garden, Chinese Academy of Sciences (FBG,CAS). Core scientific component of PALPP is from biologists, experts in particular groups of organisms, informatics and computer researchers. The goals of PALPP are to provide automatically work environment for data gathering and phylogenetic analysis for scientific research, and furthermore, to share information with other databases, and to link information from other databases through phylogeny.

At present, PALPP extracts molecular data of rbcL, atpB and other genes or sequences, from the international public databases (GeneBank, DDBJ, and EMBL) in batch, which is automated acquisition, cleaning function for users to understand the situation of peer data. It also has the function of private data management to ensure that the data mining before publishing. The process of phylogenetic reconstruction includes several public modes and tools, such as batch extraction, multiple sequence alignment, cleaning & editing, tree reconstruction, phylogeny evaluation and visualiza-
tion. All these procedures demand a number of interactive interfaces for phylogenetic tree automatic generation and decision-making aids experiment.

Grid computing is applied in PALPP. It integrates several parallel modes due to the forms and requirement of calculation, such as MPI, MapReduce, GPU and so on. The unified approach is applied to manage the parallel modes in the computing workflow. CloudBlast (combining MapReduce and virtualization on Distributed resources) is adopted in batch extraction; MPI (clustalw-mpi) of the cluster computing is adopted in multiple sequence alignment; (mrbayes-mpi, TNT, etc.) of the cluster computing is adopted in tree reconstruction. At the same time, PALPP is gradually expanding GPU acceleration of multiple sequence alignment in the computing environment.

Based on the data and computing environment in CNIC, phylogenetic users only need to visit the platform and to complete all these works.