

The Encyclopedia of Life: A Novel Toolkit for High-Throughput Proteome Annotation, Biological Data Federation and Analysis

Miller, Mark A. ^{*1}, Baldrige, Kim¹, Shindyalov, Ilya¹, Li, Wilfred¹, Quinn, Greg¹, Pekurovsky¹, Dmitry, Byrnes, Robert W. ¹, Reyes, Vicente¹, Birnbaum, Adam¹, Mosley, Coleman¹, Potier, Yohan¹, Amoreira, Celine¹, Veretnik, Stella¹, Bourne, Philip E. ^{1,2}

¹San Diego Supercomputer Center and ²Department of Pharmacology, University of California at San Diego, La Jolla, CA, USA

The explosion in the technologies associated with genome and proteome analysis networking (particularly wireless), and data storage capabilities makes it possible for biologists to generate and store vast amounts of data in the pursuit of biomedical discoveries. To realize the full potential of a data-driven biology new cyberinfrastructure is needed that is readily accessible to all scientists, and that facilitates the movement, storage and analysis of large amounts of data. To meet this need, a variety of novel tools and utilities must be created including: 1) automated pipelining tools that allow users to analyze vast quantities of data across diverse compute environments; 2) access to highly integrated database resources, so collected data can be instantly linked to existing knowledge; 3) a software “workbench” where federated data can be manipulated and visualized in a user-friendly environment; 4) access to computational resources to drive the calculations through grid computing; and 5) tools to store and share the results of individual investigations. The design philosophy for these tools must be to require minimum input by the user - computations must be carried out on the server, graphics must be lightweight and run on the client, data must be provided in a form that permits interoperability; and access to computational resources must be transparent.

The Encyclopedia of Life (EOL) is a large scale project aimed at creating precisely this type of cyberinfrastructure for the proteomics community. The EOL consists of three elements. The first is a software pipeline that automatically assigns protein sequences with putative structural and functional annotation [W.W. Li, G.B. Quinn, N. N. Alexandrov, P.E. Bourne and I.N. Shindyalov (2003) “Proteins of *Arabidopsis thaliana* (PAT) database: A resource for comparative proteomics” *Genome Biology* 4(8), R51]. This pipeline includes a workflow system that maps these calculations onto distributed resources at partner institutions throughout the world. The second is a reference database; annotations derived from the pipeline are stored in a normalized reference database that is federated with seven other major biological databases, allowing direct queries across several areas of specialization. The third element of EOL (under development) is focused on use and distribution of the data: all data generated, stored, and federated by the EOL project will be presented to the user for analysis and distribution using innovative WEB services based data sharing tools, including a web browser-based “encyclopedia” of annotated proteomes, and a virtual user “notebook” that allows for data storage, preservation of workflow information, and peer-to-peer data sharing.

The cyberinfrastructure created by the EOL project is designed with the intent of creating a significant number of “generic” software tools and middleware that is not confined to proteomics but can be applied more generally within the biomedical community.