

T3

Biological databases and modern computer science: Toxygates and EzCatDB

Nozomi Nagano 1, Johan Nyström-Persson 2,5, Kazuyoshi Ikeda 1,2,5, Yoshinobu Igarashi 2, Kenji Mizuguchi 2, Kentaro Tomii 1 and Shinichi Honiden 3,4,5

1 Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)

2 National Institute of Biomedical Innovation (NIBIO);

3 National Institute of Informatics (NII)

4 University of Tokyo

5 LEVEL FIVE Co.,Ltd.

Abstract

To realize their full potential, new biological databases in the era of “omics” require modern computer science. In recent years, large amounts of biological data obtained by the newest experimental technology are being produced with a view to industrial aims, such as drug discovery and high-functional gene design. However, the unprecedented size and complexity of these datasets pose new challenges for high-speed access and processing, data integration of different omics datasets, and effective visualization. Modern computer science techniques, such as NoSQL, data mining, machine learning, linked data, concurrency and visualization techniques, can help address these challenges.

In this talk, we focus on two databases: Toxygates and EzCatDB. Toxygates is a user-friendly integrated analysis platform for toxicogenomics data. It combines a large microarray dataset with the ability to fetch semantic linked data, such as pathways, compound-protein interactions and orthologs. EzCatDB classifies enzyme reactions manually in terms of enzyme active-site structures and their catalytic mechanisms, as well as ligand types, based on literature information, amino acid sequences of enzymes (UniProtKB), and the corresponding tertiary structures from the Protein Data Bank (PDB). EzCat-FORTE is recently implemented and enables us to detect sequences of remote homologues of enzymes. In summary, this presentation will introduce the new functionality of these important biological databases and discuss our expectations from computer science for the future of omics data.