Defense

Application of the DIKW Model in Malaria Systems Biology: From NGS Data to Disease Progression Insight

Jung-Ting Chien
Emory University

Abstract: The data, information, knowledge and wisdom (DIKW) model has been widely used in data science fields to generate a comprehensive view of each domain. It provides a hierarchical representation of the understanding of the domain knowledge; the DIKW model can reveal insights in systems biology by integrating different types of omics data to form a comprehensive understanding.

The foundation of systems biology is mining genomics data with machine learning. As the use of high-throughput, next-generation sequencing (NGS) applications grows, research in genomics enters the big data era. NGS applications can be divided into two major categories, short-read and long-read techniques, which are based on the principle differences in generating reads. A read is the fundamental element of genomic information. Short-read applications have been widely applied in several fields of genomics research, while long-read applications just came to market in 2011. Long-read applications have shown the potential to handle several areas of genomic questions. However, obtaining a well-defined genome still has a number of challenges in malaria systems biology research, and these challenges block researchers understanding the mechanism of the malaria disease progression.

To tackle these challenges, we built a novel long-read NGS pipeline with third party modules and modified them to solve complicated Plasmodium genome assembly questions. These techniques provided a solution where traditional, short-read technologies could not because of the Plasmodium genomes highly repetitive nature. We also implemented infrastructure to solve data management difficulties and developed several novel and robust pipelines to process and analyze the data. We host this pipeline along with other third party applications for data quality control, generic data visualization and data management tools. Our pipeline is also scalable and flexible to combine different technologies (long reads and short reads) to assemble the Plasmodium genome and conduct downstream annotations.

This dissertation describes an overview of omics research in the big data era and reveals the possibility of applying DIKW models through mining genomics data. A detailed discussion on how to apply our platform to solve questions, including multiple Plasmodium genome assemblies and annotations, and an initial discussion of applying machine learning approaches in a host-pathogen transcriptome analysis and its data mining applications are also provided.

Friday, July 7, 2017, 10:00 am
Mathematics and Science Center: W306

Advisor: Mary Galinski, Co-advisor: Zhaohui Qin

Mathematics and Computer Science
Emory University