The Evolution of Bioinformatics

According to Oxford dictionary, bioinformatics is the science of collecting and analyzing complex biological data such as genetic codes. However, in the last four decades the development of new technology to collect biological data has accelerated, increasing data Velocity (rate of flow), Volume (size of the dataset), Variety (data from multiple repositories, domains or types), Veracity (origin of the data and its management), Longitudinal and Ontology-based data integration. We can call this new data world, “Longitudinal Interconnected Big Data” (LIBD) that has the potential to revolutionize our approach to understanding biological systems in their natural environment. One simple example is the integration of medical records which contains medication and lab results that can be linked to large scale scientific international research databases such as “Drug Bank” and “The Human Metabolome Database” respectively. This simple integration would allow a research team to study metabolic changes linked to unknown pathways given a known drug chemistry (over time) possibly leading to new discoveries in off-label drug use at the mechanism level. Another example is building disease networks where two or more diseases are connected by set of genes that may express different proteins depending on DNA and RNA methylation [3] leading to varying disease pathways.

In order to pursue these use cases and many others in a meaningful way, a bioinformatician must have a good grasp on a wide variety new technologies & techniques, available data and domain experience in other disciplines. The most important technologies in this list are HPC, Distributed Computing (i.e. big data), Graph Databases, Data Science Workflows, R, Python, and TensorFlow (and other AI-Based Libraries). There are also a number of emerging techniques that are critical for advancing bioinformatics including Bayesian techniques, deep learning (i.e. new neural network techniques 28+), new methods in machine learning, differential equations & multiagent systems (non-linear systems), graph theory, constraint satisfaction (e.g. solvers), natural language processing and various search techniques. New scientific datasets include Drug Bank, The Human Metabolome Database, Microbiome DB, Earth Microbiome Project, just to name a few but there are many more important very large datasets.