Technological Development in Biological Data Management and Analytics Platforms

As biological research has increasingly turned into a data-rich area, the need for storing and communicating large datasets has grown tremendously. In order to make these heterogeneous data sources easy to use, several efforts at data management are currently being undertaken.

Big Data is the new paradigm in biological research. With the advent of genomics, proteomics, metabolomics, clinical data, and electronic health records, life scientists are struggling with enormous data sets, largely unstructured. Biological Big Data refers to high-volume, high-velocity, and high-variety complex data sets requiring technology to collect, store, process, manage, organise, and analyse information to glean insights and consequently help in decision making.

It all started with the completion of the Human Genome Project, more than a decade ago. Genetic mapping is generating huge data sets in form “omics” data banks. The sheer size of these data banks or data libraries can be conceptualised by a simple fact that a single sequenced human genome is around 140 gigabytes. With the development of targeted therapies and personalised medicine, one can only imagine the volume of “omics” data in the world.

Drug development is also undergoing changes in recent times. Drug development firms are leaving no stone unturned and are putting every piece of data, structured or unstructured, under scrutiny. One can easily imagine that the cost of development of new drugs, price pressures from the regulatory and health technology assessment bodies, and economic situation across the globe put enormous pressure on innovators to adopt effective and efficient processes for drug development. There is a radical change in the way companies want clinical data analysed. It also builds pressure for market research firms such as ours and provides opportunity to constantly innovate and improvise the way we analyse data sets.

Another attribute of the big data is its complexity. Data sets in biology are far more heterogeneous than those in physics. Data from different experiments lead to many types of information, such as genetic sequences, protein chemistry, molecular interactions, and patient records, just to name a few. This highly variable information needs to be analysed and interpreted in light of prior knowledge. In any clinical trial, patient data needs to be compared with each other and also across prior information available for longitudinal analysis. If data sets contain genetic sequencing or protein interactions to be compared for pattern recognition, the complexity is daunting.

Several data management platforms have been developed to facilitate data analysis in healthcare settings. Companies such as Oracle, Accenture, and IBM have been instrumental in providing tools to life scientists to digitize published or non-published literature, enable text search, visualise outputs, and develop models and frameworks. Platforms such as electronic laboratory notebook (ELN), laboratory information management system (LIMS), sequencing data banks, omics libraries, clinical data collection (OpenClinica), electronic case report forms (eCRF), tranSMART, and electronic patient health records developed by these firms are playing pivotal roles in modern day biological research. A whole array of omic libraries and data banks such as COSMIC, dbGaP, GEO, ArrayExpress, or TCGA have been created to facilitate research and manage data sets.

It is very reassuring to see the academia and industry worldwide collaborating in advancing the big data biology. It indeed needs concentric efforts from all. Toward attaining this, much of the construction in big data biology is focused on cloud computing, which enables remote...
access of information to all, saving the huge investments on hardware at local establishments. The idea is to facilitate research, particularly in an era of reduced funding. In addition to Helix Nebula, a cloud-based infrastructure ideated by European Bioinformatics Institute (EBI) in collaboration with European Organisation for Nuclear Research (CERN) and European Space Agency (ESA) with IT partners, private companies such as Accenture and Oracle (The Life Sciences Cloud) are entering the domain.

However, cloud computing has its own challenges; one of the biggest being safety and security of the data stored and monitoring the access rights, which is why most drug developers are hesitant of using the cloud. But I firmly believe that early hesitation will move toward total acceptance as the advantage of using cloud-based data management systems is overwhelming.

The cost–benefits and potential to overcome logistical barriers will give cloud-based system the thrust it needs to propel. Moreover, the formation of international consortia for efforts such as eTRIKS Deliver and BioMart only reinforces the idea of collaborative research.

However, cloud computing has its own challenges; one of the biggest being safety and security of the data stored and monitoring the access rights, which is why most drug developers are hesitant of using the cloud. But I firmly believe that early hesitation will move toward total acceptance as the advantage of using cloud-based data management systems is overwhelming.

The cost–benefits and potential to overcome logistical barriers will give cloud-based system the thrust it needs to propel. Moreover, the formation of international consortia for efforts such as eTRIKS Deliver and BioMart only reinforces the idea of collaborative research.

The world of big data in biology is moving at an immense pace and changing the way biological research is done. The traditional “wet labs” are now changing to “dry in silico labs. A bright future lies ahead for big data analytics in biology, where IT platforms bring multiple internal and external data sources across functions (clinical, safety, regulatory, and operational), into a single analytics platform, thereby creating actionable insights to accelerate drug development, contain development costs, and improve patient outcomes. However, we still need high computational capacity to analyse heterogeneous data at high speed and skilled data scientists to undertake such high-end analytical research. Nevertheless, big data biology presents ample opportunities for efficient use of information in effective personalized healthcare.

Contact: rezni.dsouza@adfactorspr.com