BioSpectrum the business of Bio & Health Sciences

Big Data, Big Challenges

04 February 2013 | Views | By BioSpectrum Bureau

Big Data, Big Challenges

Our ability to make simultaneous measurements on several tens of thousands of genes or even on the entire DNA in a biological system has improved many-fold over the last several years. The cost of making these measurements has also been reducing continually. This has enabled researchers to generate candidate genes/mutations of interest far more quickly than previous gene-by-gene methods. This has also brought about the need for researchers to add one more dimension to their research toolbox-the need to handle big data.

Big data could mean different things in different contexts. In the context of a research lab doing molecular biology research today, big data typically signifies a five GB file of raw data for every sample run; and a typical experiment can have tens or even hundreds of samples, reaching hundreds of GBs in size. These have to be processed by special algorithms before their sizes can be compressed to the scale of 10s to 100s of MBs, amenable for exploration and discovery. The algorithms are in a semi-mature state at the moment: algorithm developers across the world have built and made available open source and commercial tools that will allow researchers to do the same. However, as measurement technology continues to change, these algorithms are evolving as well. And with it, there is an increasing need for many more algorithms and tools.

While many research groups in India use available tools effectively to answer their research questions, few have contributed new tools and algorithms for use by the community at large. Though much of our effort at Strand Life Sciences does go towards this goal, through a commercial setting, via our GeneSpring and Avadis NGS products; I do know of few academic initiatives in this direction as well. This needs to change in the future via a multi-disciplinary involvement of computer scientists, statisticians and biologists, under key umbrella initiatives that could be undertaken in India towards understanding the Indian population (similar to, for example, the 1000 genomes project).

Note that the above paragraph defines big data largely from the perspective of a single research lab generating up to a few 100 GBs of data a year, which is the typical case in India. But there is a case to be made for aggregation and pooling of data across labs. For instance, mutation hunting requires knowing how often a particular mutation occurs in the population at

large, which could benefit from the pooling of normals or controls across labs and studies. In essence, we are talking about the creation of central genomic repositories with modern interfaces that are very easily accessible and usable by researchers at large. This is where the real problem of big data hits; data sizes can reach TB (terabyte) and PB (petabyte) scales. There are of course ways to reduce these sizes, and there are several groups across the world building such infrastructure, including our group at Strand. However, researchers in India haven't had an impressive history of pooling data or creating central bioinformatics resources; this is something that should change in the future.

There is a third definition of big data that might confront us in the future. A whole industry is working on bringing the cost of whole genome sequencing down. It currently stands at about $1000 (internal cost that is, the cost of selling and deploying usually multiplies this many-fold, but that is guaranteed to reduce with maturity). If and when the cost reduces by another factor of two-three, it may become feasible for individuals to get their genomes sequenced proactively, regardless of immediate medical need. There are several examples of genomic variants providing warnings for individuals to monitor aspects of their health more aggressively (for instance, the now famous Indian mutation in the MYBPC3 gene that 1 in 25 Indians have and that almost always causing ventricular wall thickening post age 45). A whole genome raw file is close to 100 GB, and this multiplied over potentially millions of individuals will yield the biggest of our big data problems.

In summary, biological and medical research will drive off big data to a very large extent in the future. With the big data getting bigger over time, multi-disciplinary groups will need to come together to create appropriate infrastructure so we can continue to make efficient use of this big data.