The digital side of biology

March 7 2011, By Lee Swee Heng

William Tjhi (left) and Rick Goh (right).

Revolutions in science come in waves. One of the epoch-making events in modern biology came in 1995 when J. Craig Venter, an American biologist, decoded the whole genome of the Haemohilus influenzae bacterium using a 'shotgun' sequencing technique that involved the computational assembly of data.

At that time, many molecular biologists, including Wayne Mitchell who is now with the A*STAR Experimental Therapeutic Centre (ETC), had experienced years of frustration seeing their work in gene cloning beaten to the finish line time and again by rival laboratories. Inspired by Venter's demonstration of how computers can be used to vastly accelerate biological research, Mitchell decided to self-train himself to acquire the computational skills needed to perform this type of analysis. At first, his new direction was not entirely welcomed by his colleagues,

who raised their eyebrows and asked why he would waste his time with such an endeavor. “I think they couldn’t get out of an old paradigm,” recalls Mitchell.

Fifteen years have passed since then, and Mitchell appears to have rightly captured the tide of the times as the computerization of biological research gains momentum. He has become one of the global pioneers who are armed with both biological and computational expertise. As the founding leader of the Informatics Group at the ETC, Mitchell took the initiative to build up information technology platforms such as electronic content management systems and new networks to transmit large amounts of data at faster speeds. On the research side, Mitchell and his colleagues are utilizing computerized processes—from the robotic screening of several thousands of chemical compounds with high-throughput computing machines, to the computer modeling of chemical structures—in a bid to develop therapeutic drug candidates for cancer and other diseases.

“Modern biology is all about automated machines churning out huge amounts of data, which then have to be managed, stored, analyzed and visualized,” says Mitchell. “None of these procedures amount to rocket science, but if you don’t do it, there is actually no point to conducting the experiment in the first place.”

Multiple data analysis

The advent of the information technology era has completely changed biological research. Network speed, storage capacity and computer clusters have seen continual improvements, and the development of new algorithms and knowledge management protocols has led to the introduction of novel computer-based methods such as sequence alignment, machine learning techniques for pattern identification and ontologies for formal data classification. Together these platforms have

made possible what are now widely used experimental technologies, including sequencers, microarrays, detectors for single-nucleotide polymorphisms and mass spectrometers. Large databases of sequencing results have also become publicly available, spurring informaticians to design tools that make the data readily available to experimental scientists. “Digital computing is the servant of non-digital, brain-based computing,” Mitchell says.

Biologists’ interest in research aided by cutting-edge computational tools, a field known as computational biology, has taken hold across A*STAR’s institutes. At the A*STAR Genome Institute of Singapore (GIS), almost all research involves computing, and one third of its investigators are either computer scientists or biologists equipped with strong computer skills. Now that sequencing has become routine, “there is more demand for computer researchers to analyze the massive quantities of data,” says Neil Clarke, deputy director of the GIS and a molecular biologist as well as a self-trained bioinformatician.

In a recent study from the GIS, Clarke’s team investigated one of the most important DNA regulatory systems—how transcriptional factors and nucleosomes compete with each other to obtain a position to bind to DNA. The researchers examined multiple publicly available, high-resolution datasets of genome-wide nucleosome position and performed computer analyses to compare positions in vivo and in vitro. They found that the key to this analysis is to know not the precise nucleosome location but rather its broader regional occupancy.

Ken Sung, a principal investigator and computer scientist at the GIS, and co-workers have recently designed a microarray to identify DNA mutations of the H1N1 influenza virus with high accuracy. In combination with software called EvolSTAR, the microarray chip can sequence the virus directly from blood samples without amplifying the virus beforehand. The kit has not only improved the efficiency of testing

compared with other chip-based methods, but also reduced research costs by a factor of ten. It enables researchers to perform large-scale biosurveillance studies to track the changes in influenza during a pandemic, which can help prevent the spread of the virus.

Elsewhere in the GIS, large-scale sequencing has brought great benefits to the emerging field of 'metagenomics', in which researchers directly study microbial samples from nature, bypassing a difficult culturing step in the lab. Niranjan Nagarajan, a principal investigator and computer scientist, is developing novel tools to accurately analyze metagenomic datasets for both marker gene sequencing and whole-genome sequencing. "New technologies are able to deeply profile these communities, but what researchers get is very fragmented information. They want to use the data to reconstruct a more comprehensive picture of a sample," says Nagarajan.

Finding out the correlation between variations and phenotypes for multi-marker studies is another computationally intensive field. "It is a highly complex, high-dimensional challenge to handle millions of variables and construct optimal mathematical models. We spend a lot of time building into a model analyses or predictions that reduce the complexity of the problem," says Anbupalam Thalamuthu, a principal investigator and statistician at the GIS. Thalamuthu is developing a model for the genetic analysis of polymorphisms of diseases including various infectious diseases, breast cancer and dengue.

Modelling a virus

At the A*STAR Institute for Infocomm Research, Victor Tong, an assistant department head, is developing three-dimensional (3D) models of immune cells and viruses to see which location of a virus is likely to activate strong immune responses. In the past, researchers used traditional wet lab procedures to analyze thousands or even tens of

thousands of different combinations of peptide sequences. But such an approach was not cost-effective and it took considerable time to pinpoint specific regions in pathogen proteins that can trigger effective immune responses. “This way, computational analysis can help accelerate the research,” says Tong, a biochemist and computer scientist. “When an antigenic peptide triggers an immune response, there must be a 3D fit between the peptide and the receptor binding site of the host immune cell. Hence, the use of 3D models is important for such studies,” he adds.

Tong’s recent studies include the analysis of chikungunya virus, a re-emerging tropical disease that causes flu-like symptoms in Singapore and elsewhere around the world[4]. His team has built a model and analyzed how the pathogen mutates across different times and geographical locations. They have found that the virus has accumulated different mutations within a specific region of a structural protein. Tong hopes his work on modeling the virus will lead to the design of new types of vaccines that contain precise fragments of the virus triggering immune responses. Using current technologies, vaccines contain randomly selected fragments of the virus or the entire attenuated dead virus.

Visualizing data

Although computers can produce mountains of interesting data, the data remain useless unless biologists are able to extract meaning from the data set. This is where they need the aid of computer scientists. “Biologists are interested in finding out what kinds of shapes cells are taking in reaction to treatments. Our work is to visualize the results of data analysis in a way that is understandable for biologists,” says William Tjhi, a research engineer at the A*STAR Institute of High Performance Computing (IHPC).

In collaboration with Frederic Bard, a biologist at the A*STAR Institute

of Molecular and Cell Biology, Tjhi is developing a methodology to transform millions of digitized cell images into numerical values or matrices. He then performs an analysis to find areas in which one coherent pattern can be separated from another. “Based on this cluster analysis, biologists make their own interpretation,” he says. Being able to develop an initial visual understanding of the data helps biologists plan detailed experiments. But in the past, the reliability of visual approaches was variable because biologists needed to categorize cells manually either under the microscope or by utilizing a tool called a classifier. Such methods are heavily dependent on human intervention in determining the initial definitions of interesting patterns. Tjhi is trying to reduce the level of human intervention in the data-creation process so that researchers can minimize subjectivity.

To expand the cluster analysis approach for full-scale operation, Tjhi is collaborating with Bard and Rick Goh, a senior research engineer at the IHPC, on a project to tackle the ‘millions of cells problem’. So far, Tjhi’s software is capable of performing analyses on 50,000 cells, but the team is trying to beef up the capability to a few million cells. Goh is currently assessing the kinds of high-performance computing techniques and tools that are needed for such a platform, which could include hybrid architectures such as multicore systems, accelerators and graphics processing units.

The computer trap

The new engineering solutions that have revolutionized biological research have revealed previously unimagined aspects of biology and refuted numerous biological myths. However, such technological development has leaped so far ahead of the information infrastructure that supports it that data are now being accumulated much faster than can be digested or synthesized. As Mitchell points out, the informatics aspect of the computational support system is becoming a bottleneck.

One of the issues faced by many researchers in this area is a shortage of data storage and memory capacity. High-throughput machines churn out terabytes of data every day, and many biologists would regard it unthinkable to delete any of the data in order to conserve storage capacity. “Our terabytes of storage can run out so easily,” says Goh. The large volumes of data are also affecting data mobility. “On our servers, we can move terabytes of data in less than a few hours. But moving data from one storage system to another can take days, or even weeks—and that is the time needed even before analysis starts,” Goh adds. One potential technique, instead of moving data, is to move the computing code and process the data on its original system, thus eliminating the time for data transfer.

The rapid generational change in experimental technologies is also a constant headache. “We spend a lot of time thinking about what applications and new technologies to work on, and how we can analyze the data acquired using them,” Clarke says.

Moreover, as researchers are sharing more and more data among the community, standardization is becoming another imminent issue requiring serious discussion. “Data could be inconsistent and not interpretable using other databases. There is no quality control for bioinformatics,” Tong points out.

Building a bridge

Aside from hardware issues, perhaps the biggest challenge for computational biology is the dialogue between computer scientists and biologists, many researchers say. Whenever building a model for analysis or predictions, or visualizing matrix data, computer scientists may not always be able to understand what it is biologists want to know. “If I just take data given by biologists and put them into my cluster analysis, the results would be poor,” says Tjhi. “I need to understand how the data are

generated by their experiments. I need to take into consideration this fact when incorporating data into the analysis, and then the results generally become better." Tjhi admits that understanding such different disciplines is far from an easy task, but there is no short cut, he says. "We have to communicate more with the biologists. More time needs to be invested in this kind of project in order for people to understand each other."

Many researchers involved in computational biology are hoping to employ a person who can understand the languages of both sides of this research. Mitchell, whose role is to talk to both biologists and computer scientists, agrees that the biggest issue is communication. But he is more optimistic about the future. "Younger scientists and undergraduate students feel more comfortable with computer-based methods because computers are already a natural extension of their daily lives. For the next generation, the computational–experimental dialogue will naturally become routine. I'll need to find another job."

More information: Mitchell, W., et al. Genomics as knowledge enterprise: Implementing an electronic research habitat at the Biopolis Experimental Therapeutics Center. *Biotechnology Journal* 3, 364–369 (2008). dx.doi.org/10.1002/biot.200700190
Goh, W. S., et al. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. *PLoS Computational Biology* 6, e1000649 (2010). dx.doi.org/10.1371/journal.pcbi.1000649

Provided by Agency for Science, Technology and Research (A*STAR)