Biologists think bigger

Developments in academia and industry may encourage biologists to use large-scale computation

Biology is getting bigger. The science that once concerned itself with studying the intricate details of individual organisms and molecules is now taking a step back to get a larger picture of what life is: mechanisms, pathways and systems. This shift is a natural progression, but it is also a result of the huge amount of data that has been generated over the past decade from large-scale sequencing efforts such as the Human Genome Project. Making sense of this information requires entirely new ways of thinking, and an equivalent revolution in methodology. First and foremost, it calls for massive computing power to extract information from the raw data. This shift from reductionism to complexity in biology is now revealing problems of a scale that can only be solved using the computational power usually associated with the harder sciences, such as physics, astronomy and mathematics—disciplines that have many problems that require massive calculation to solve. What form this computational power takes depends on the specific problem at hand, but also on the biologists themselves. Industry may be useful, but in the past few years, academia has shown that it can often find its own solutions.

In 1999, computer scientists at the University of California at Berkeley (CA, USA), launched a project to search for extra-terrestrial intelligence (SETI) by analysing radio signals recorded on the Arecibo Observatory radio telescope in Puerto Rico. Realizing that available computing resources were not sufficient to process all the data received, they came up with an ingenious solution: distribute the task to members of the public, all over the world, who volunteer time on their own PCs. Users download software disguised as a screensaver, and are then sent a chunk of data via the Internet that is processed only when their PC is idle. The results are sent back to Berkeley, and a new set of data is received. This 'distributed computing' model connects thousands of PCs to create a virtual computer with more computational power than the most advanced supercomputer. With more than 500,000 active users, "in terms of total computing [power], we're the largest [distributed computing] project in history," claimed David Anderson, head of SETI@home.

It did not take long for biologists to copy this approach. In 2000, Vijay Pande and his colleagues at Stanford University (CA, USA) started the Folding@home project to elucidate the mechanisms behind protein folding. Like SETI@home, Pande's project uses distributed computing to harness the power of PCs throughout the world. The project currently has more than 100,000 regular users and over 750,000 total contributors. Unlike Anderson's project, Folding@home's volunteers have already achieved results, successfully simulating the folding of small specially designed polypeptides (*Nature* **420:** 102–106). "100,000 [volunteers' PCs] allows us to do work that really couldn't be done any other way," Pande said.

> This 'distributed computing' model connects thousands of PCs to create a virtual computer with more computational power than the most advanced supercomputer

But distributed computing is not necessarily suited to all biological problems. Pande likened the approach to speeding up a task by a factor of 1,000. "If someone gave you 1000 assistants, it's unclear whether that would really allow you to achieve that goal." Organizing and managing all those people would take up most of the time. Similarly, distributed computing is not simply a matter of dividing a complex problem into many smaller problems that can be calculated independently—the way in which the problem is approached is as important as the problem itself.

Nevertheless, other distributed computing projects have successfully tackled various biological problems. Arthur J. Olson's group at the Scripps Research Institute (La Jolla, CA, USA) uses the FightAIDS@home project to screen candidate drug compounds against detailed models of evolving AIDS viruses. Graham Richards, Chairman of the Chemistry Department at Oxford University, UK, has elicited support from more than 2 million volunteers for his cancer screening project. The main project aims to find new drugs for cancer therapy by screening a database of millions of small molecules against a selection of specific proteins thought to be involved in the development of the disease. Other smaller projects have already found potential drugs against smallpox and anthrax using the same approach.

By focusing on cancer and AIDS, these distributed computing projects also increase the chances of drawing public support. Their common denominator is an overwhelming public interest in being involved in and contributing to scientific research. In fact, few of the researchers make a concerted effort to encourage people to contribute; word of mouth is often enough. "In terms of public understanding of science, and this was not one of the original aims of the project, it has been remarkably successful," said Richards. "To get the general public involved is really very valuable." Pande explained, "there's also aspects of the way distributed computing works that is designed to try to encourage people to gather friends to get involved." Many projects keep track of who has volunteered the most computing time, or who has cracked a particular problem. "That aspect actually has a sort of a competition aspect to it that is in many ways a large driving force as well," Pande said. Referring to SETI, Anderson commented that "many people are motivated by the possibility of discovering life outside Earth, others are motivated by the competition aspect, and others like our high-tech screensaver graphics."

> …few academic researchers have the financial resources to buy the latest high-performance supercomputer

FURTHER INFORMATION ABOUT LARGE-SCALE COMPUTING

Types of technology	Description	Further reading
Supercomputer	Broad term describing the fastest computer available. Typically refers to many fast CPUs in one box, with high-speed interconnection	Twice a year, the University of Mannheim's Top500 project ranks the world's 500 fastest supercomputers: www.top500.org. Well-known manufacturers like NEC, Hewlett- Packard, Dell, IBM and Cray top the list. Cray pioneered supercomputing to dominate the market in the 1980s and 1990s, and remains one of the most recognized supercomputer manufacturers.
Cluster computing	A group of machines that act like a single system, such as rack-mounted CPUs connected by an Ethernet	Clusters are particularly popular in universities, where they provide maximum computing power at minimum cost. Donald Becker and Thomas Sterling designed the first 'Beowulf' cluster in 1993: www.beowulf.org. The top 500 can be found at http://clusters.top500.org
Grid computing	Sharing computing resources between organizations; similar to distributed computing, in that users can access computers around the world	The European DataGrid is at present preparing the infrastructure for a Grid that will service the scientific community across Europe: http://eu-datagrid.web.cern.ch/eu-datagrid. CERN (Geneva, Switzerland) will need a grid to handle data from its particle accelerator, the Large Hadron Collider (http://lcg.web.cern.ch/LCG). In the US, the Global Grid Forum aims to promote Grid technologies: www.gridforum.org. The Globus Alliance (www.globus.org) develops software and applications
Distributed computing	Large numbers of internet-connected PCs	Distributed computing is proving ideal for scientific research projects, such as SETI@home (http://setiathome.ssl.berkeley.edu), Folding@home (http://folding.stanford.edu), FightAIDS@home (http://fightaidsathome.scripps.edu), Graham Richards' cancer project (www.chem.ox.ac.uk/curecancer.html) and Evolution@home (www.evolutionary-research.net)

Although Pande and his group have obtained valuable results, distributed computing alone may not be sufficient to 'solve' protein folding. Computer giant IBM (Armonk, NY, USA) is ready to tackle the problem with brute force. The Blue Gene project, announced at the end of 1999, will build a petaflop supercomputer—capable of 1,000 trillion floating-point operations per second—by the end of 2004. Blue Gene will be 500 times faster than the current fastest supercomputer, and more than 2 million times faster than a standard desktop PC. The project is as much about computer engineering as it is about biology: IBM has revolutionized the construction of supercomputers in order to meet this challenge. Bob Germain, manager of the Biomolecular Dynamics and Scalable Modeling Group of the IBM Blue Gene team (Yorktown Heights, NY, USA) explained: "We think we can advance our understanding of the protein folding process through large-scale simulation … and also study other interesting biologically related systems."

IBM's foray into biology is indicative of the potential of this future market. "Biologists are not traditionally strong users [of supercomputers]," said Barry Utting, former General Manager and Vice President of Cray Europe and present director of BDUX Limited. This may be for economic reasons, as few academic researchers have the financial resources to buy the latest high-performance supercomputer. On the other hand, biologists may prefer to stick with what they know and leave supercomputing to the harder sciences. IBM's investment into Blue Gene may be an attempt to counter that attitude.

There are obvious differences between distributed computing and supercomputing that make the two approaches suitable for specific needs. "Each paradigm is good for some problems and not others," Anderson explained. "Public computing is useful only for problems that have public appeal (so you can get users), have a high computing/data ratio (so your Internet bill is reasonable) and don't involve secret data (since it will be visible to the world). There are quite a few problems that fall into this category. Because of the huge numbers of PCs, public computing will likely continue to outperform the other paradigms by some measures (such as total number of floating-point operations)." Supercomputers, however, are ideal "if your problem is big enough and communication-dependent enough," explained Utting. The difference lies in how each approach deals with the problem: distributed computing is ideal for large problems that can be split into many similar smaller problems that can be solved independently (so-called 'embarrassingly parallel' problems). Analysing the many different folding trajectories of one protein, or screening a database of molecules against one protein, are ideal applications for distributed computing. Supercomputers are more suited to non-linear complex problems that are not easily subdivided, but require high-speed interconnection. Understanding an entire system in which many factors influence many others, is one example.

> "Each paradigm is good for some problems and not others"

Blue Gene is therefore unlikely to tread on the toes of Folding@home. "The big difference between Folding@home and Blue Gene is that while Folding@home probably does have more raw power, the communication between the processors is a lot slower, whereas Blue Gene has state of the art communication," Pande explained. Continuing his analogy with the 1000 assistants, "it's kind of like having 1000 assistants where they can talk to each other extremely quickly such that they can help organize themselves better." Indeed, the two groups have collaborated to their mutual benefit, and both agree that their

approaches are complementary. Many of the algorithms used with Folding@home, in particular the models representing various physical and chemical laws, could have an impact on Blue Gene's design and software. Although the supercomputer will have more power than any other computer on earth, its success in determining how proteins fold still depends in part on the accuracy of these models. Germain confidently expects to go where no supercomputer has gone before: "Because we will have access to much larger computing resources than anyone has ever had, especially for this problem, we can do the larger size systems and study systems at longer time scales which gives us a better chance of connecting in a significant way with physical experiments."

With so much computing power available, it is not clear why biologists have yet to embrace this technology. Beside the obvious prohibitive cost of buying a supercomputer, what is stopping biologists from tackling larger problems with larger computers? "Certainly there is a discrepancy between the technology that might be available to solve these problems, and what is [used]," Utting observed. Distributed computing is similarly under-exploited but, according to Anderson, "there are at least 100 million Internet-connected PCs, and less than 1% of them are participating in distributed computing projects. There's no shortage of computers." It may be that biologists have not yet learned to think 'big' enough to use large-scale computation. "I think part of it is not the scale of the thoughts but trying to figure out a way to use computers to actually be useful," Pande said. So far, outside the realm of bioinformatics, computers have not done much for biologists, he believes. "Traditionally people think about large-scale simulation and calculation as the domain of physics but I think what's going to be happening now is that it's going to start to become very commonplace in biology," Pande predicted.

With so much computing power available, it is not clear why biologists have yet to embrace this technology

According to Utting, "I think the harder sciences naturally go to simulation because there's a certain level of numeracy required to do it, and biology in the past hasn't been a quantitative science and hasn't attracted quantitative people." Richards believes that often technology starts off with the physicists, subsequently moves into the field of chemistry, and will eventually be picked up by biologists: "On the whole, the biological community [are] not interested in the technology, they're interested in the answers." To promote distributed computing, Pande is thus planning to release the software behind Folding@home. "It takes a lot of work to create all the software yourself, and I think that barrier has kept people away," he said.

Ultimately, it may be that computing supply has so far outpaced demand. With all this technology and only a few really appropriate problems to tackle, there is the danger that biologists may fall into the trap of using more power and less thought when design experiments. As many in the bioinformatics field would agree, it is easy to produce a swathe of data without investing much brain power. Both Pande and Germain avoid this in their own groups by collaborating closely with experimentalists. "The end arbiter to anything is really its significance in terms of experiments and biomedical relevance," Pande said. "It always makes me very proud that experimentalists are excited about what we're doing…if they weren't excited about it, I think it would mean that there is something that's missing."

…the success of academic and industry projects may encourage more scientists to use large-scale computing

The use of computation in biology is still at the very early stages, but the success of academic and industry projects may encourage more scientists to use large-scale computing. The projects themselves will also evolve. According to Pande, the next step for distributed computing is to start looking at larger proteins and also proteins that are more biologically relevant. "If we really believe that this technology is useful, we should try to attack problems that are important," he said. Biologists would be able to model larger systems at much longer time scales, and concentrate on understanding complex systems previously intractable to current methodology. Germain therefore expects that scientists will eventually become more ambitious in the kinds of calculations that they want to do—regardless of whether these calculations are carried out on one expensive supercomputer, or millions of personal computers scattered around the globe.

Caroline Hadley

doi:10.1038/sj.embor.7400108