Bioclouds

Understanding how cloud computing can serve the scientific community is a research question in its own right and one that researchers—with the help of funders—should address.

In their struggle to help data processing keep up with data production, members of genomics laboratories are venturing into cloud computing (p. 495). First launched commercially about four years ago by Amazon, cloud computing allows users to rent storage and computing services from companies with spare server capacity. Cloud services running on specialized operating systems spread computational tasks over thousands of otherwise idle central processing units, completing analyses in hours rather than days or weeks. Cloud-compatible scientific tools can be used by researchers without installing or configuring the software themselves.

How best to do a scientific analysis in the cloud is literally a million-dollar question or actually two: how to use the cloud now, and how to make it better. The time is ripe for clarity on both questions.

In its current form, cloud computing caters mostly to business applications. Analyzing data in the cloud requires algorithms that break the analysis into separate tasks—an acceptable requirement for many genomics problems—but writing or converting tools for cloud computing requires considerable expertise.

One clear need in the community is for tools that allow regular biologists to analyze data in the cloud, and Galaxy is a great example. This project, by a group of academic researchers, consolidates genomics software into a common interface on the Amazon cloud, allowing researchers to upload data and run analyses with no programming or command line requirements. To encourage the continuation of such projects, their impact should be assessed, and those involved should be clearly credited.

Evaluation of such tools is key, as is the standardization of tools and pipelines for specific tasks and their design for easy scale-up. Ultimately though, the goal should be not only to share tools designed for large datasets but also to integrate data from different research groups so that individual groups are not forced to reanalyze data.

One area worth exploring is academic clouds built with the requirements of the research community in mind. The University of Maryland maintains clouds to allow researchers to test new programs and to accommodate fluctuating demand, and the Free Factory platform at Harvard University uses virtual machines, commodity hardware and free software to provide web-based services capable of coping with large datasets.

The Open Cloud Consortium is supporting frameworks to allow cloud interoperability such as in Bionimbus, a US National Institutes of Health–supported effort that hosts some genomic datasets and provides tools for researchers to run analyses in several clouds. The National Human Genome Research Institute is exploring a variety of cloud-computing pilot programs with both commercial and academic groups. And some specialized commercial vendors are building clouds and offering proprietary cloud-based genomics software. Researchers' options are set to expand greatly in the near future, but many researchers are wary that the open-source cloud operating systems now available are less robust and comprehensive than Amazon's.

The immediate question for researchers at most institutions, however, is where they should store and analyze their data. Although big cloud vendors can buy equipment at bulk discounts and rent out capacity at reasonable rates, institutions that have constant computing needs can still save money by building their own capacity. In contrast, institutions with fluctuating needs may benefit from the flexibility of renting cloud capacity as needed.

It is not always straightforward to assess whether using the cloud or adding internal capacity is more expensive; computing tasks are often built into overhead, and increases in capacity are counted as one-time costs. In contrast, computing in the cloud has small but recurring costs. The costs of data transfer must be considered as well because cloud providers also charge for this.

A big shift triggered by cloud computing could be putting a clear price on storage and analysis costs for individual projects. Researchers should be able to include such costs in their funding requests, but funding bodies could provide more fundamental support. Individually, researchers have little bargaining power with commercial cloud providers, but if large funding bodies purchased access to the cloud on behalf of grantees, costs could come down substantially.

Ultimately, the pace of progress may depend less on costs than on comfort and judgment. The rare researchers with cloud computing and genomics expertise are feeling fatigue as they attend workshops for hammering out best practices or find themselves repeatedly explaining the same concepts to newcomers. Answers are not only in flux but vary widely by application and institution. As the field matures, researchers will have to learn how to cobble together and coordinate both cloud-based and local solutions. But the future is looking increasingly cloudy, and learning how to best exploit this new domain will lead to a much clearer view of the genome.