pharmaphorum
bringing healthcare together

Monday 18 September 2017

Fixing the problems in bioinformatics


Linda Banks

January 25, 2016

Data management and analysis is increasingly important as the speed of genome sequencing increases and costs reduce. Kalina Cetnar outlines the issues bioinformaticians face and ways to address them.

Kalina Cetnar

The evolution of Next Generation Sequencing (NGS) technologies has revolutionised the world of genomics, enabling fast, cost-effective and accurate creation of sequencing data. These technologies are widely used in a variety of industries, including pharma and biotech, where they are helping the push towards personalised medicine.

However, as nobody predicted the speed of the evolution of sequencing technologies, computational biology is facing some major challenges in data storage, management, security and analysis.

Since the birth of DNA sequencing in 1977, the technology has seen a drastic decrease in sequencing costs. Nowadays a human genome can be sequenced for as little as $1,000. Unfortunately, computers are not evolving as quickly. Lower cost, and hence greater scale of genomic sequencing, is producing enormous amounts of data, resulting in major central processing unit (CPU) and storage problems. Since manipulating such enormous data sets requires computational resources beyond the power of a standard computer, there are two ways to solve the problem.

The traditional answer is to use a computer cluster. This is not cheap and requires investment in hardware, software, physical storage space and costs for electricity and cooling of the cluster.

A newer, affordable alternative is to move the computations to the cloud. Cloud infrastructures are flexible and dynamic, allowing users to scale the allocated resources up and down according to their needs. Since an outside provider maintains the cloud, there is no need to worry about costs of hardware or related costs.

Among the biggest problems faced by bioinformaticians are the search for, and making use of, publicly-available data, as well as dealing with multiple formatting types. A vast amount of data is publicly available, but using it can be a challenge. After searching for and

downloading the data, it is essential to analyse it, check its quality and suitability, and not lose all the metadata in the process. These are time-consuming tasks that more often than not lead to errors.

Researchers estimate that about 80% of their time is spent on data grooming and only 20% on actual data analysis. This stems from a lack of standardised file types and inconsistent data formatting, meaning that every new program results in a new data format.

The solution is to automate these tedious and time-consuming data-grooming tasks giving researchers more time to focus on data analysis. Operating on a 'format-free' data analysis platform means that when data is uploaded to it in any format, it 'loses' the format and becomes a meaningful biological object, with all objects of the same kind acting identically, regardless of underlying formatting differences.

Other common issues concern reproducibility and organisational problems, such as incorrectly annotated genes or even lack of any data annotation whatsoever. Keeping track of data provenance is essential and details such as scripts or specific versions of data used must be carefully recorded, so that the analysis could reproduced by someone else or in the future. Since reproducibility is a necessity for cumulative science, researchers should pay a lot of attention to such matters.

Another aspect that researchers complain about is losing track of their own computations, so repeating the methodology on another set of data becomes incredibly hard.

The way to ensure reproducibility of data is to keep track of data provenance. However, noting down all scripts and parameters is time-consuming. Therefore automation here would be a great advantage and save researchers significant time and effort.

Scientists working in bioinformatics who came from other research areas, for example biologists who previously worked in a laboratory, commonly mention lack of appropriate skills or coding experience as limiting their abilities to perform analysis. This problem is increasingly important, as there is a big discrepancy between the numbers of lab researchers and bioinformaticians in any research facility.

There is a growing need for tools that enable scientists to analyse their sequencing data without expert knowledge of programming languages such as Python or R. Further, lack of communication between bioinformaticians and biologists leads to a lack of understanding of the science behind an experiment or statistical analysis that makes results of the experiment relevant.

The solution here is to use platforms where all the coding has already been done by someone else with a user-friendly interface that allows researchers themselves to analyse their data and draw correct conclusions.

The world of genomics is rapidly changing the landscape of healthcare. With the prices of genome sequencing dropping below $1,000, personalised medicine and treatment plans based on an individual's genetic makeup will become an everyday reality.

What are the challenges of using NGS tools in the clinic? The most important ones include data security, storage, analysis and interpretation. Raw sequencing runs generate hundreds of gigabytes of data from a single measurement, and this means current clinical data management infrastructure is not enough to manage it. Use of cloud computing for storing and managing data is likely to become more and more common in the clinical setting. However, many remain uncertain about whether it will meet data security and archiving standards and how it will comply with regulatory requirements. New, integrated systems and methods are required to help unleash the full potential of genomics.

About the author:

Kalina Cetnar works at **Genestack**, an innovative bioinformatics cloud platform at the interface of science and business. Contact her at: **kalina@genestack.com**

Read more on genomes:

Genome project puts England at cutting edge of precision medicine