CloudButton

Serverless computing to scale up genomic alignment algorithms

Modern biology is more and more underpinned by massive datasets, such as those produced by high-throughput sequencing experiments. The inner programming of the cell is stored as DNA in the genome; the cell transcribes part of the genome into RNA and possibly translates RNA into protein, thus eventually producing all the constituents of living things. The information content of the cell is impressive: Just the genome for a human individual is more than 3 billion bases, corresponding to 3 billion 'A', 'C', 'G', 'T' letters, or 3 GB of uncompressed hard disk space. To that, one has to add many other products of the processing of the genome by the cellular machinery and inter-cellular variability.

Modern sequencing machines allow scientists to decode (or "sequence") part of this information as very large collections of short text strings (or "reads"). At the moment, we cannot reliably sequence molecules longer than a few hundred or a few thousand bases; on the other hand, current machines are able to decode tens of millions of those short strings in a single experiment and for an affordable price. Thus one single sample of biological material can originate several datasets, each one taking tens of GB of uncompressed hard disk space. Processing such data requires complex ad-hoc high-performance analysis workflows (called "pipelines") and large amounts of computing power. That is why high-performance computing platforms are now commonplace in biology departments around the world.

Access to this data has revolutionized biology, putting it on a firmer quantitative ground for the first time. Being able to sample with great precision the internal state of the cell at the molecular level allows scientists to shed light on complex phenomena, such as the development of cancer or the interaction between viruses and their hosts. Other examples are the re-sequencing of individuals to determine their genomic variants and ancestry, or the sequencing on a large number of viral samples to track the spread of an outbreak (for instance, by flu or the recent Covid-19) and the evolution of the virus while it moves from individual to individual. That often results in the generation of large datasets comprising hundreds or thousands of samples.

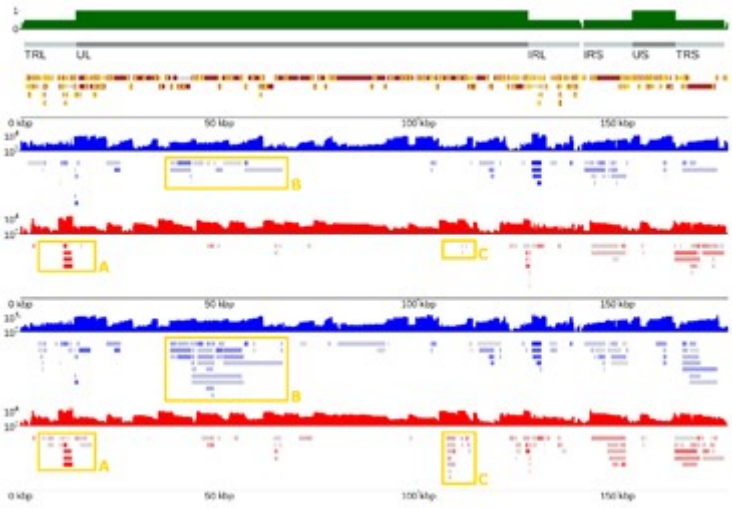

Figure 1: Transcriptional landscape of Marek's disease virus, a virus causing cancer in chicken and large economic losses to the poultry industry. The plot shows the results of several sequencing experiments as a genome browser. Source: doi.org/10.3390/v12030329

Biomathematics and Statistics Scotland (BioSS) is a research

institution specialised in developing sound mathematical and analytical methods for a range of problems, including biological and epidemiological ones. In particular, several BioSS scientists have a long-standing experience in the development of analysis pipelines for sequencing data. This activity ranges from low-level heavy-duty algorithmic tools to comprehensive processes that go all the way from raw data straight to high-level results directly interpretable by non-specialists, such as decision makers in the Scottish government.

Many use cases at BioSS require the manipulation and analysis of large biological datasets (see Figure 1). That is usually accomplished by accessing dedicated infrastructure, such as computer clusters offering a large amount of computer cores and storage. However, such resources are finite and need to be shared between a number of scientists working at different institutions. Situations requiring quick reanalysis of very large amounts of data – such as the processing of hundreds of sequencing samples obtained from individuals or animals infected by a viral disease to understand infection patterns and predict future spread – would be hard to tackle in this framework. The ability to offload peak demand to the cloud, in particular to agile serverless platforms, would be very desirable.

Unfortunately, migrating existing algorithms to the cloud is far from straightforward, due to a number of technical problems. The most expensive computational step is the alignment of sequencing reads to a reference genome, i.e. the operation of locating all the regions in the genome that the read might have come from. While one can in principle process sequencing reads independently of one another and hence leverage the elasticity of serverless platforms, the large amounts of memory required by most alignment algorithms make this difficult.

Thanks to serverless technologies, a widely used genomic alignment algorithm can now scale to thousands of cores in the Cloud

Thanks to the mutual participation of the two institutions in the CloudButton project, BioSS has teamed up with Large-Scale Data and Systems Group at Imperial College. The LSDS group at Imperial has a long-standing experience on new abstractions and infrastructures for building scalable, reliable, and secure distributed applications. The collaboration has resulted in a novel scheme to parallelise read alignment, which is able to split memory requirements among different nodes. Thanks to this design and serverless technologies, a widely used genomic alignment algorithm developed in the past by BioSS scientists can now scale to thousands of cores in the Cloud.

This work represents a first, fundamental ingredient towards migrating to serverless architectures arbitrarily complex analysis workflows for the analysis of sequencing data. Eventually the analyst will be able to choose between local execution, when available resources are sufficient, and remote processing in the Cloud, whenever the situation requires very fast processing of a very large amount of data. Unquestionably, such tools will prove useful in the future, as novel and increasingly aggressive threats challenge our biosecurity.