Background Chapter

Andrew van Rooyen

2. Background

Data sets are a big part of bioinformatics, and have introduced many new challenges with the rise of next generation sequencing. Sequencing technologies like SOLiD provide much higher data output at a cheaper cost [Shendure and Ji 2008], which is good news for research, but troubling for data storage, transfer and access. In fact, the cost of storing a byte has been more expensive than sequencing a base pair since before 2010 [Baker 2010].

This makes it difficult for researchers in different locations to manipulate and run processes on the data, because it will be stored in only one location. These files could be tens of gigabytes in size [Deorowicz and Grabowski 2011], depending on context.

2.1 Data storage

Generally, once the sequencing machine has generated the raw information on each base pair, this data will be stored in a data warehouse. Storing this information for long periods of time requires the data to be structured efficiently in order to save space, and allow it to be transferred efficiently. There has been a lot of work on how to structure this data. There are a plethora of file formats whose efficiency depends on the kind of data which needs to be stored. Two of the most popular formats are FASTQ, which stores aggregated reads along with the quality of each base pair [Cock et al. 2010], and BAM, the binary, compressed version of the Sequence Alignment Map (SAM) format [SAMTools 2015].

2.2 Data transfer

When researchers require specific information for their projects, they need to be able to access the data warehouse and transfer whichever sequences they need. Luckily, these locations are often connected massive data pipes like National Research and Education Networks (NRENs). Unfortunately, standard protocols like FTP and SSH were never designed for use on high-throughput networks, and alternate protocols need to be used to avoid bottlenecking.

There are some proprietary transfer protocols which are widely used in practice. For example, the fasp protocol by the US based company AsperaSoft. Based on UDP, the protocol eliminates the latency issues seen with TCP, and provides bandwidth up to 10 gigabits per second to transfer data [Beloslyudtsev 2014].

2.3 Alternate models

There have been some attempts to do data processing remotely, and there has been an explorative push towards cloud solutions from Amazon, Google etc. Unfortunately, even though these cloud data centres have plenty of cheap storage, there are very significant drawbacks.

Because the sequencing happens in labs, researchers need to upload their raw data to the cloud data centres every time they run a new experiment. This leads back to the original problem, as researchers resort to mailing hard drives [Baker 2010]. There are also security, privacy and ethical concerns with outsourcing this processing power to other companies, as sequenced DNA data is often highly sensitive information [Marx 2013].

REFERENCES

Monya Baker. 2010. Next-generation sequencing: adjusting to data overload. *nature methods* 7, 7 (2010), 495–499.

Dima Beloslyudtsev. 2014. Aspera transfer guide. (2014).

Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* 38, 6 (2010), 1767–1771.

Sebastian Deorowicz and Szymon Grabowski. 2011. Compression of DNA sequence reads in FASTQ format. *Bioinformatics* 27, 6 (2011), 860–862.

Vivien Marx. 2013. Biology: The big challenges of big data. *Nature* 498, 7453 (2013), 255–260.

SAMTools. 2015. Sequence Alignment/Map Format Specification. https://samtools.github.io/hts-specs/SAMv1.pdf. (2015). Accessed: 2015-04-27.

Jay Shendure and Hanlee Ji. 2008. Next-generation DNA sequencing. *Nature biotechnology* 26, 10 (2008), 1135–1145.