Eight Questions and Answers Defining Cloud Computing for the Digital Biology Era

Konstantinos Krampis*[1]

[1]Informatics Department, J. Craig Venter Institute, 9704 Medical Center Dr. ,Rockville, MD 20850, USA

Email: Konstantinos Krampis*- agbiotec@gmail.com;

*Corresponding author

Abstract

Background: Text for this section of the abstract. [**?**]

Results: Text for this section of the abstract . . .

Conclusions: Text for this section of the abstract . . .

Background: Sequencing Technologies and The New Era of Digital Biology

Digital Biology Research Enabled by High-Throughput Sequencing is defined by...

Sequencing technologies continue to move in a direction where throughput per run is increasing while the cost per basepair is decreasing (review in [1]). Several technologies available on the market today produce massive volumes of sequence data per run; for instance, one of the most widely used instruments in the field for the past few years and currently, Illumina's GAIIx system can produce up to 95 Giga-base of sequence (Gb) per run [2] while the also broadly used SOLiD sequencer has yields of a similar range up to 90 Gb [3]. With the latest generation of instruments including Illumina's HiSeq systems the yield per run

has reached 600 Gb [2], and with the the Pacific BioSciences instrument yields of 90 Gb can be achieved [].

Recently, small-factor, benchtop sequencers became available including GS Junior by 454, MiSeq by Illumina and Ion Proton by Life Technologies, all of which can be acquired at a fraction of the cost and be affordable for independent researchers running smaller laboratories. Nonetheless these sequencers still provide enough capacity at 0.035Gb, 1Gb and 1.5Gb respectively for GS Junior, Ion Proton and MiSeq review in [4]) for sequencing bacterial, small fungal or viral genomes. This fact in combination of the low cost per run (US $225 -$1100 depending on which benchtop sequencer is used and required throughput), can establish sequencing as standard technique for basic biological research. Examples include Single Nucleotide Polymorphism (SNP) variation discovery , gene expression analysis (RNAseq), DNA–protein interaction analysis (ChiPseq), (review in [5]).The new generation of sequencing technologies is also being used in the area of metagenomics, for large-scale studies of uncultivated microbial communities. The J. Craig Venter Institute (JCVI) for example has been involved in several such metagenomic projects, including the Sorcerer II Global Ocean Sampling (GOS, [**?**]) expedition to study marine microbial diversity, and also the National Institutes of Health funded Human Microbiome Project to study human associated microbial communities [**?**].

Discusion: Eight Questions and Answers to Define Cloud Computing Applications on Digital Biology

What is the Role of Cloud Computing Play in The Digital Biology Research Era?

While large datasets are generated during sequencing runs, sequencers are typically bundled with only minimal computational and storage capacity for data capture during the run. For example, the un-assembled reads returned from a single lane of the Illumina GAIIx instrument after base calling are approximately 100 GigaByte (GB) in size. Given the scale of datasets, scientific value cannot necessarily be obtained from the investment in a sequencing instrument, unless it is accompanied by an equal investment in a large-scale bioinformatics infrastructure.

For small laboratories acquiring a sequencing instrument, the currently available online software tools are not and option for downstream sequence analysis, since they cannot provide the required compute capacity. The NCBI website for example (Johnson et al. 2008), cannot accept input sequence data files of 0.5 GB size for BLAST sequence similarity search. With this as an example, we see that scientific value cannot be obtained from an investment in sequencing instruments, unless it is accompanied by an almost equal or

greater investment in informatics hardware infrastructure. Besides large capacity compute servers, also required are trained bioinformaticians competent to install, configure and use specific software to analyzethe generated data, and store the data in appropriate formats for future use (Richer et al. 2009).

Renting Cloud Computers Versus Building Local Clusters?

Even if bioinformatics computing clusters are accessible by researchers, a problem is related to the sub-optimal utilization of the hardware, and its associated maintenance costs. This is due to the nature of bioinformatics for next-generation sequencing, where computationally intensive tasks of genome assembly or whole genome alignments require extensive compute resources, while tasks such as genome annotation and browsing are less computationally demanding. For smaller laboratories this can become a hurdle, as in addition to getting funds for building a cluster with capacity to handle the large computations, they need to come up with the money for maintaining a system that is not utilized at its full capacity most of the time. A second problem arises from the fact that the public sequence databases are constantly growing in size. These databases must be downloaded to local storage systems, in order to be used for example when conducting comparative genome annotations. As these databases get larger the process becomes more time consuming, incurring in higher bandwidth and storage costs for replicating the data locally. Finally, building a bioinformatics infrastructure for next-generation sequencing also involves hiring trained bioinformaticians competent to install, configure and use specialized software tools and data analysis pipelines, which can present a higher expense than that of acquiring the computing cluster.
ur system will not be limited only to the Amazon EC2 cloud platform, since researchers that have access to a local cluster at their home institution will have the option to download the VM and run OSMF, without being required to perform any software installation. The only dependency will be a virtualization layer that can run the VM on the cluster such as for example the OpenStack open-source cloud (http://www.openstack.org). OpenStack is available as part of widely used Ubuntu Linux (http://www.ubuntu.com/cloud) and included by default on a compute cluster set-up to run this operating system, while it can be easily installed as a package on clusters running other Linux versions. Alternatively, for researchers that do not have access to local compute clouds, OpenStack installations can be accessed through the government-funded Argonne National Lab Magellan Cloud http://www.alcf.anl.gov/magellan) that provides compute allocations to researchers, in addition to a number of academic computing centers in both the US and abroad (http://openstack.org/user-stories/), or commercial cloud providers such as RackSpace (http://www.rackspace.com/cloud/private$_e$*dition*/*openstack*/)

Are Cloud-Based Bioinformatics Software Suites Available on the Cloud ?

For the sequence analysis suites currently available on the cloud, the common pattern is packaging existing bioinformatics applications within virtual compute servers. This is a great solution for smaller labs that lack informatics infrastructure, since it makes available pre-configured software andon-demand computing using virtualized infrastructures. However, a problem exists with this approach in regards to the monolithic design of existing bioinformatics applications, which are ported to the cloud. These applications usually process data serially, and are not designed to leverage specific characteristicsof cloud computing platforms, such as highly parallelism and distributed computing. Therefore, they cannot scale efficiently as the amount of data increases (Wilkening et al. 2009). According to the same authors, it is difficult to transfer on a cloud infrastructure specialized bioinformatics applications such as the MG-RAST metagenomics analysis pipeline for example (Glass et al. 2010). Despite being designed to scale in parallel for large sequences datasets, MG-RAST's design makes it tightly coupled with specific cluster computing hardware and the SunGrid Engine (online ref.10) scheduling framework.

Can Non-Computationally Savvy Researchers Easily Access the Cloud?

Text for this sub-section. More results ...

What the Public, Private, Open-Source or Commercial Cloud Available to Biologists Today ?

Text for this sub-section. More results ...

Available Options to Buying Software as A Service on the Cloud ?

During the past year, both public and commercial offerings of pre-configured sequence analysis applications on the cloud have become available. On the commercial side, DNAnexus (online ref. 3) currently includes tools for ChiPseq, RNAseq, 3'-end sequencing for expression quantification (3SEQ) and enzyme restriction analysis. DNAnexus runs on the Amazon Elastic Compute Cloud (EC2, online ref. 4), which provides on-demand virtual servers with various compute capacities. Another offering is iNquiry, which is a port to the cloud of the bioinformatics software suite that used to be bundled together with the computer clusters built by the BioTeam (online ref. 5), but is no longer maintained. This platform is essentially a web-server for pre-installed open-source tools such as EMBOSS, HMMER, BLAST and the R statistical package, also on the EC2 cloud platform.In regards to public offerings, the Galaxy bioinformatics workbench (online ref. 6), includes a range of software, from scripts that extract entries from FASTA files, to tools for processing

next-generation sequence data. It is a self-contained platform including a web server along with the bioinformatics tools, and users can easily add more software packages by editing simple configuration scripts. Galaxy has recently been ported to run on the Amazon EC2 compute cloud. Another public offering for computing on the cloud has come through our own work on JCVI Cloud Biolinux (online ref. 7), which is a virtual high performance computing server on Amazon EC2. Our offering bundles a set of sequence analysis tools similar to those offered by iNquiry, with the difference being that the virtual server is available for download (online ref. 8), and users can run it on the open-source Eucalyptus (online ref. 9) or Science Clouds platforms (Keahey et al. 2009). In addition, Cloud Biolinux also includes the Celera genome assembler and a set of scripts that allow for push-button creation of virtual computing clusters for parallel BLAST, geared towards researchers that intend to perform large scale genomic sequence analysis.For the sequence analysis suites currently available on the cloud, the common pattern is packaging existing bioinformatics applications within virtual compute servers. This is a great solution for smaller labs that lack informatics infrastructure, since it makes available pre-configured software andon-demand computing using virtualized infrastructures. However, a problem exists with this approach in regards to the monolithic design of existing bioinformatics applications, which are ported to the cloud. These applications usually process data serially, and are not designed to leverage specific characteristicsof cloud computing platforms, such as highly parallelism and distributed computing. Therefore, they cannot scale efficiently as the amount of data increases (Wilkening et al. 2009). According to the same authors, it is difficult to transfer on a cloud infrastructure specialized bioinformatics applications such as the MG-RAST metagenomics analysis pipeline for example (Glass et al. 2010). Despite being designed to scale in parallel for large sequences datasets, MG-RAST's design makes it tightly coupled with specific cluster computing hardware and the SunGrid Engine (online ref.10) scheduling framework.

Which Factors Challenge Adoption of Cloud-Based Solutions for Bioinformatics

Another important concern for cloud-based bioinformatic tools is related to the data transfer bottleneck from the local sequencing machines to the cloud servers. According to the data published for the Amazon cloud platform (online ref.11), 600GB of data would require approximately one week to upload on to the remote cloud servers, when using an average broadband connection of 10Mbps. With a faster T3 connection which is usually easily obtainable even at small research institutions, within one week 2TB of data can be uploaded or approximately 600GB in 2 days. Solutions addressing this issue are available both as software that maximizes data transfer over the network compared to traditional File Transfer Protocol

(FTP), or physical disk drive import/export services offered by the cloud provider to its customers. Aspera's server (online ref. 12) has been recently integrated to NCBI's infrastructure, and researchers can download a free client that allows increased upload speeds to the Short Read Archive (online ref. 13). Through the Aspera software, transfer bandwidth between NCBI and the European Bioinformatics Institute for data sharing in the 1000 Genomes Project, has been increased from 20Mbps to 1000Mbps (see online ref. 14).Finally, the Amazon offers the option for its users to physically ship disk drives to the company's offices and have the data copied to their servers (online ref. 11). With only 80 import cost for disk drives up to 4TB of data (4000GB), this is the most efficient method if we take into account the charge by Amazon for 0.10 per GB of bandwith consumed, which would add up to 60 for a 600GB data upload. In addition to that cost, the expense for obtaining a high-bandwidth internet connection for the data upload should be taken into account. We expect the Microsoft Azure cloud platform to offer a similar service in the near future, given the requests on the Azure developer forums and the immediate consideration of the matter by Microsoft (online ref. 15).

Yet another results sub-heading

Text for this sub-section. More results ...

Conclusions

Text for this section ...

Methods

Methods sub-heading for this section

Text for this sub-section ...

Another methods sub-heading for this section

Text for this sub-section ...

Yet another sub-heading for this section

Text for this sub-section ...

Authors contributions

Text for this section ...

Acknowledgements

Text for this section ...

References

1. Mason C, Elemento O: **Faster sequencers, larger datasets, new challenges**. *Genome Biology* 2012.
2. Illumina I: **Technical Specifications of Illumina Sequencers**.
3. Inc AB: **The SOLiD 5500 sequencing system.**
4. Loman N, Misra R, Dallman T: **Performance comparison of benchtop high-throughput sequencing platforms**. *Nature ...* 2012.
5. Mardis E: **Next-generation DNA sequencing methods**. *Annu. Rev. Genomics Hum. Genet.* 2008, **9**:387–402.

Figures

Figure 1 - Sample figure title

A short description of the figure content should go here.

Figure 2 - Sample figure title

Figure legend text.

Tables

Table 1 - Sample table title

Here is an example of a *small* table in LaTeX using `\tabular{...}`. This is where the description of the table should go.

My Table		
A1	B2	C3
A2	...	..
A3	..	.

Table 2 - Sample table title

Large tables are attached as separate files but should still be described here.

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.