

# Access to Big Data in Bioinformatics

Brendan Ball  
Andrew van Rooyen

## 1. PROJECT DESCRIPTION

Collaboration between bioinformatics organizations involves shared access to large datasets. This project investigates two avenues for tackling the collaborative analysis of big data in bioinformatics: Efficient transport of large datasets across high speed WAN networks and implementation of community clouds that host and securely execute code close to the location where data is stored.

Historically, data has been exchanged using tools and protocols like SSH and FTP, but new protocols, such as GridFTP and HPN-SSH, offer more efficient use of high speed networks.

Community clouds are cloud computing services built on micro clouds hosted by collaborating organisations, as opposed to conventional cloud computing hosted by cloud vendors (such as Microsoft, Google and Amazon). Hosting micro clouds close to scientific data collections would facilitate scientific collaboration through moving code closer to data, rather than vice versa.

## 2. PROBLEM STATEMENT

The University of Cape Town (UCT) and the University of the Western Cape (UWC) are both part of the South African National Research Network (SANReN). This provides a 10Gb link to other academic institutions in South Africa. Currently, bioinformatics data in the Western Cape cannot be moved at these speeds, even though the infrastructure theoretically allows it. In practice, the software configuration at the endpoints is the bottleneck. In the scope of this project, the software configuration refers to the network protocol, but in practice includes more layers. This project aims to bring the transfer rate up to the actual capacity of the network by investigating different network protocols.

The second part of this project will survey existing cloud computing infrastructures towards their suitability for creating organisation-level micro clouds. The aim is to design and implement a micro cloud solution allowing universities to easily deploy their own micro cloud which can access and be accessed by other connected micro clouds. These connected micro clouds will form a community cloud. The micro cloud solution will include implementing a code migration framework which will allow users to easily migrate code between micro clouds, execute that code on the micro cloud which is storing the data and return any results to a specific micro cloud.

## 3. PROCEDURES AND METHODS

The first part of this project will compare and contrast the performance of GridFTP, HPN-SSH, FTP and SSH for transferring multi-gigabyte datasets. The comparison will examine throughput, delays, security and authentication features of these protocols and their implementations. This will be tested mostly on the local UCT network, and then between UCT and UWC once the nuances of the protocols and their configurations are understood.

For the second part, existing software will be surveyed, and a combination of solutions at various stack levels will be fitted together to form a complete solution. At least two micro clouds will be deployed in order to prototype a community cloud where code can be migrated between micro cloud installations. The functionality will be tested us-

ing given bioinformatics analysis code. Usability testing will be done by doing systematic observation. A limited number of experts will be observed to determine usability using client satisfaction measures.

#### 4. ETHICAL, PROFESSIONAL AND LEGAL ISSUES

Usually, there would be ethical issues when working with bioinformatics data. However, because this project is only a platform, testing doesn't rely on having data specific to a particular person or group, and it doesn't have to be recent. Any data we need can be obtained from public sources such as the National Center for Biotechnology Information (NCBI) database.

All experiments for the first section will involve quantitative testing of systems. No ethical issues are raised here. However, access to the UWC network (especially via SANReN) will need to be granted by the appropriate authorities. There may be user testing involved with the second part, but this will be expert testing with the collaborators of the project.

All work will be made publically available under the MIT licence.

#### 5. RELATED WORK

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 [?], 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 [?].

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 [?], depending on context.

There has been a lot of work on storing 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 [?], and BAM, the binary, compressed version of the Sequence Alignment Map (SAM) format [?].

There are also 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 [?].

There has been an explorative push towards cloud solutions from Amazon, Google etc, but there are very significant drawbacks. Because the sequencing happens in labs, researchers need to upload their raw results 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 [?].

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 [?].

Work in this area includes Cloud BioLinux, which is a community driven project focussed on next generation sequencing. It is a toolkit which makes it easy to deploy virtual machines with bioinformatics infrastructure to a cloud platform. It bundles specific packages used in next generation sequence analysis, thereby decreasing configuration time and increasing maintainability. Instances of Cloud BioLinux have been tested on the Amazon EC2 cloud platform and on a private Eucalyptus cloud. [?]

## 6. POSSIBLE FUTURE WORK

- Further optimising the data transfer pipeline on different levels of the stack. For example, tuning how data is read from disk to match patterns used by the network layer.
- Linking user identity to existing databases, and using this as a platform for access control. Users can, for example, have permission to execute code on remote micro clouds if they are trusted.
- Visualising genomic datasets with software such as Visual Molecular Dynamics (VMD), and linking this to the traditional linear view of the data. Possible advances could involve highlighting areas on a 3D model when a section of the linear data is interacted with. This process would need to run on a cloud and send relevant data back to the user.

## 7. ANTICIPATED OUTCOMES

- Choice of best transfer protocol for big data in our context of bioinformatics in South Africa.
- A micro cloud solution providing capability of creating a community cloud.

### 7.1. Impact

- UWC can use their dedicated 10Gbs line
- Collaboration between universities can increase given the new community cloud platform, particularly in the bioinformatics department.

## 8. PROJECT PLAN

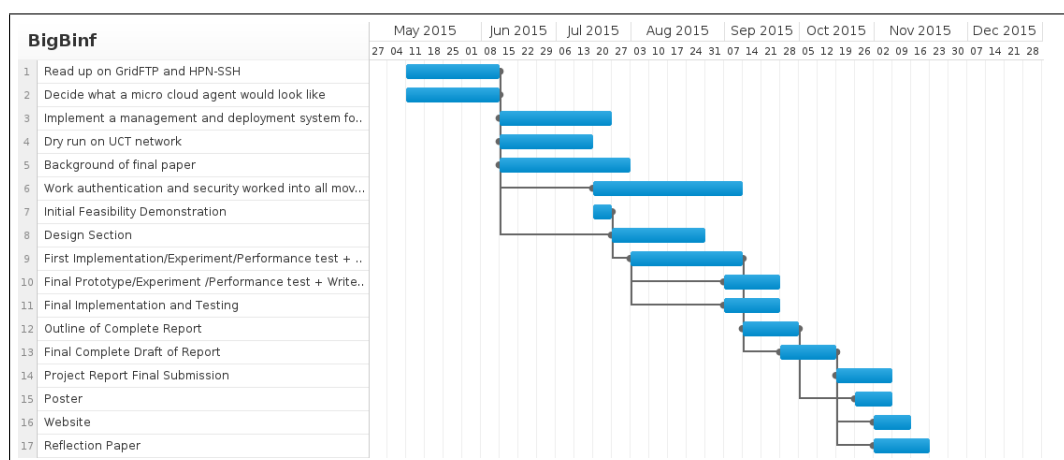


Fig. 1: Gantt Chart  
See Milestones

### 8.1. Resources Required

- Access to servers micro cloud solution
- Access to SANReN for testing transfer protocols

### 8.2. Deliverables

- |                                      |                      |
|--------------------------------------|----------------------|
| (1) Background of final paper        | (6) Final Report     |
| (2) Micro cloud first implementation | (7) Poster           |
| (3) Micro cloud final prototype      | (8) Website          |
| (4) Micro cloud final solution       | (9) Reflection paper |
| (5) Report Draft                     |                      |

### 8.3. Milestones

- |   |   |
|---|---|
| (1) Background of final paper                                   | (10) Authentication and security worked into all moving parts   |
| (2) Read up on GridFTP and HPN-SSH                              | (11) First Implementation/Experiment/Performance test + Writeup |
| (3) Dry run on UCT network                                      | (12) Final Prototype/Experiment/Performance test + Writeup      |
| (4) Best transfer protocol configured optimally.                | (13) Final Implementation and Testing                           |
| (5) Design Section  | (14) Outline of Complete Report                                 |
| (6) Decide what a micro cloud agent would look like             | (15) Final Complete Draft of Report                             |
| (7) Implement micro cloud                                       | (16) Project Report Final Submission                            |
| (8) Implement a management and deployment system for the agents | (17) Poster   |
| (9) Initial Feasibility Demonstration                           | (18) Website  |
|   | (19) Reflection Paper   |

### 8.4. Work Allocation

- Andrew will do the first section (comparing transfer protocols). Since this section is anticipated to be shorter than Brendans, he will then work on security and authentication for the cloud platform.
- Brendan will do the second section (micro cloud solution)

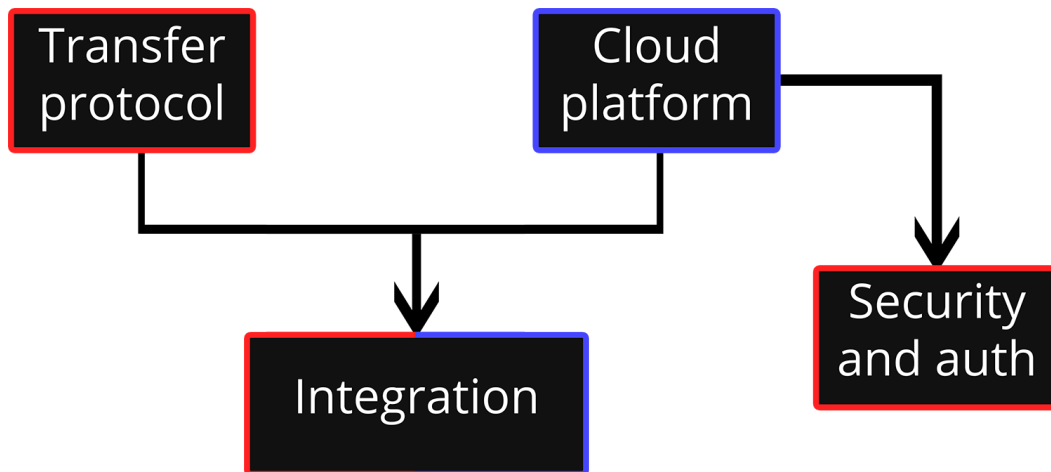


Fig. 2: Work split  
Blue: Brendan  
Red: Andrew

**8.5. Risks**

Risk	Mitigation
One of the team members is unable to complete his work on time due to unforeseen circumstances	The two sections of work are loosely coupled so it wont prevent the other team member from completing his work
Not getting access to key resources	Keep open communication with supervisors
Load shedding prevents completion of tasks in specified time	Account for load shedding in schedule
Project specification is too big, team is unable to complete all tasks on time.	Change project specification based on feedback from the presentation
Team conflict	The project is loosely coupled which should decrease chance of team conflict
Supervisors lose interest in project	Keep open communication with supervisors
Failure to integrate project components	Discuss design decisions to account for integration