Project Proposal

Luke Darlow

February 2013

1 Project itle

Design Trade-offs in Web Services for Bioinformatics

2 Project title

Luke Darlow (BSc. Computer Science). 306 Allen Place, Allen Street, Grahamstown, 6139. g10d0410@campus.ru.ac.za.

3 A Statement of the problem

Although the term 'bioinformatics' is fairly new, the idea and application of merging computer science with biology has been around for decades. Although the biology side of bioinformatics may seem puzzling to computer and information scientists, the number and effectiveness of the tools developed is impressive [5]. The merging of computer science with various peripheral disciplines is an aspect of computer science contributing to its uniqueness and effectiveness. To provide context to what computer science is presently, its relationship to other fields needs to be considered. The history and conception of bioinformatics lays a good foundation for this consideration. It is clear that the evolution of bioinformatics has had great impact on both biology and computer science. [7]

Although many tools for processing DNA exist and work very well, the access to and use of these is much simpler for a computer scientist then for a biologist. Since the field of bioinformatics is essentially a merge of biology and computer science, this fact becomes a problem. The combined orchestration of a web service, well developed tools such as the MEME suite [3], and a biologist

friendly human computer interaction component may provide more structure to the metaphorical bridge for the gap existing between computer scientists and biologists.

4 Objective

Bioinformatics approaches biology, using computer and information science, to create models of *in vitro* and *in vivo* experiments. Data gathered on DNA is noisy. Bioinformatics approaches this huge amount of incomplete and noisy data to extract usable information regarding the evolution and relationships of organisms. This rapidly growing field has had many successes. There has also been much research on the visualisation and interpretation of this data. Although computational power increases at an astounding rate, processing DNA in order to arrive at certain conclusions is yet to become a rapid process. Providing a web based front end for biologists to abstract themselves from this seeming mundane step in the process, and a back end to abstract away the complexity of the analysis will provide a fundamental stepping stone to help bridge the interdisciplinary gap.

Cloud computing provides a means of implementing a solution for this gap. By using cloud computing and considering human computer ineraction will allow us to implement a web service to face this problem.

We will analyse the role of human computer interaction within the field of bioinformatics, explore the possible role of cloud computing in this field and implement a web service to be used by those within bioinformatics.

5 Background

5.1 What is DNA?

DNA (deoxyribonucleic acid) is a molecule consisting of a combination of paired bases bonded to a sugar phosphate.

There are four possible bases, namely:

1. Adenine
2. Thymine
3. Cytosine

4. Guanine

Most DNA is found in the nucleus of the cell. However, there is a small portion residing in the mitochondrial DNA. This mitochondrial DNA is found within the mitochondria of cells. A role of Mitochondria is the energy production of cells. [6]

Approximately 99% of DNA is almost identical between two human beings. The remaining 1% of dissimilar DNA is what is used in paternity tests and the like. [5]

It is DNA that is the determining factor of heredity and consequently how an organism self-replicates in the monumental fashion resulting in the flow and balance of life. Most of DNA processing has been concerned with DNA in terms of protein construction. Although this is the main coding function of DNA, as understood presently, it is not the only function.

Another concept worth understanding is that of the **transcription factor**. A transcription factor is a protein acting on DNA to influence the flow of genetic information. This influences the production of RNA and therefore the proteins that make up an organism.

Bayat [4, p. 1008] defines bioinformatics as :

> "The application of tools of computation and analysis to the capture and interpretation of biological data."

Computer scientists are able to develop the tools needed to do the large scale processing and management of raw data (produced by molecular biologists) into a manageable, useful and user-friendly form. This hastens the process of understanding in appropriate fields. Not only are the techniques and algorithms an important aspect of this, but so is the interpretation of data and the implementation of tools that make this information more available and user friendly.

5.2 Cloud computing

By the very nature of the problem of processing DNA data (this is very computationally heavy) utilizing cloud computing may be a very viable option. Cloud computing has been around for some time and is beginning to redefine software from that of an object to that of a service. The concept of cloud computing includes the hardware and software on the backend servers and infrastructure involved. [2]

Amazon's elastic compute cloud is an existing and possibly integral tool regarding bioinformatics cloud computing. The cost of computing using the cloud is variable and depends on the scale of the operation; Amazon does offer a free Amazon EC2 option (which will be used in testing). [1] Amazon EC2 allows users to control almost the entire software stack. This gives a low level access that is crucial with processing problems present in bioinformatics. [2]

5.3 Human Computer Interaction

The role of human computer interaction has increased dramatically with the advancement of computation and its availability. This is a sphere that requires the hard sciences such as computer science and software engineering as well as the soft sciences such as psycology and design. Making use of hard, qualitative, and technical sciences such as computer science alongside the soft and qualitative sciences such as psychology is necessary to design and implement software with the end user in mind.

6 Approach

Before embarking on the task of implementing a web service specifically aimed at the field of bioinformatics, many topics need to be explored and reasoned through. This includes knowledge of the popular tools used in bioinformatics, cloud and distributed computing, human computer interaction, and web server setup and implementation.

Once this knowledge base has been acquired the implementation of the web service will begin. If all goes according to plan this will make use of a cloud computing service (possibly Amazon EC2), incorporate a well designed human interaction component and be able to perform the back end tasks in a scalable fashion.

7 Resources/Requirements

With respect to hardware we require a system that can run Linux. The use of a second machine will be ideal in order to test distribution of the web service. Amazon EC2 has a free option and would be a good place to get started.

We have decided to use open source software; free tools for development in bioinformatics are available (the MEME suite)

Frameworks such as Catalyst and Dancer (using Perl), and Django (Python) will be reviewed in order to find the most viable option for this project.

8 Progression time-line

Deadline	Activity
01 March 2013	Project Proposal
19 March 2013	1st Seminar - project presentation.
08 April 2013	Grounding in bioinformatics established
08 May 2013	Peripheral topics researched and explored
27 May 2013	Literature survey completed
01 June 2013	Investigation on human computer interaction
21 July 2013	Setup of preliminary web service skeleton complete
16 September 2013	Short paper due
30 July 2013	2nd Seminar
1 October 2013	1st Draft paper handed in
15 October 2013	2nd Draft paper handed in
28 October 2013	Final seminar
1 November	Porject hand in
4 November	Project website finalised

References

[1] Amazon. AWS Marketplace. Online, 2013. Available from: `http://aws.amazon.com/`.

[2] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. A view of cloud computing. *Commun. ACM 53*, 4 (Apr. 2010), 50–58.

[3] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res 37*, Web Server issue (Jul 2009), W202–W208.

[4] Bayat, A. Science, medicine, and the future: Bioinformatics. *BMJ: British Medical Journal 324*, 7344 (2002), pp. 1018–1022.

[5] Benoït, G. Bioinformatics. ann. rev. info. sci. tech. *Annual Review of Information Science and Technology 39* (2005), 176–218.

[6] Mandal, A. Online. Available from: `http://www.news-medical.net/health/What-is-DNA.aspx`.

[7] Ouzounis, C. A., and Valencia, A. Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics 19*, 17 (2003), 2176–2190.