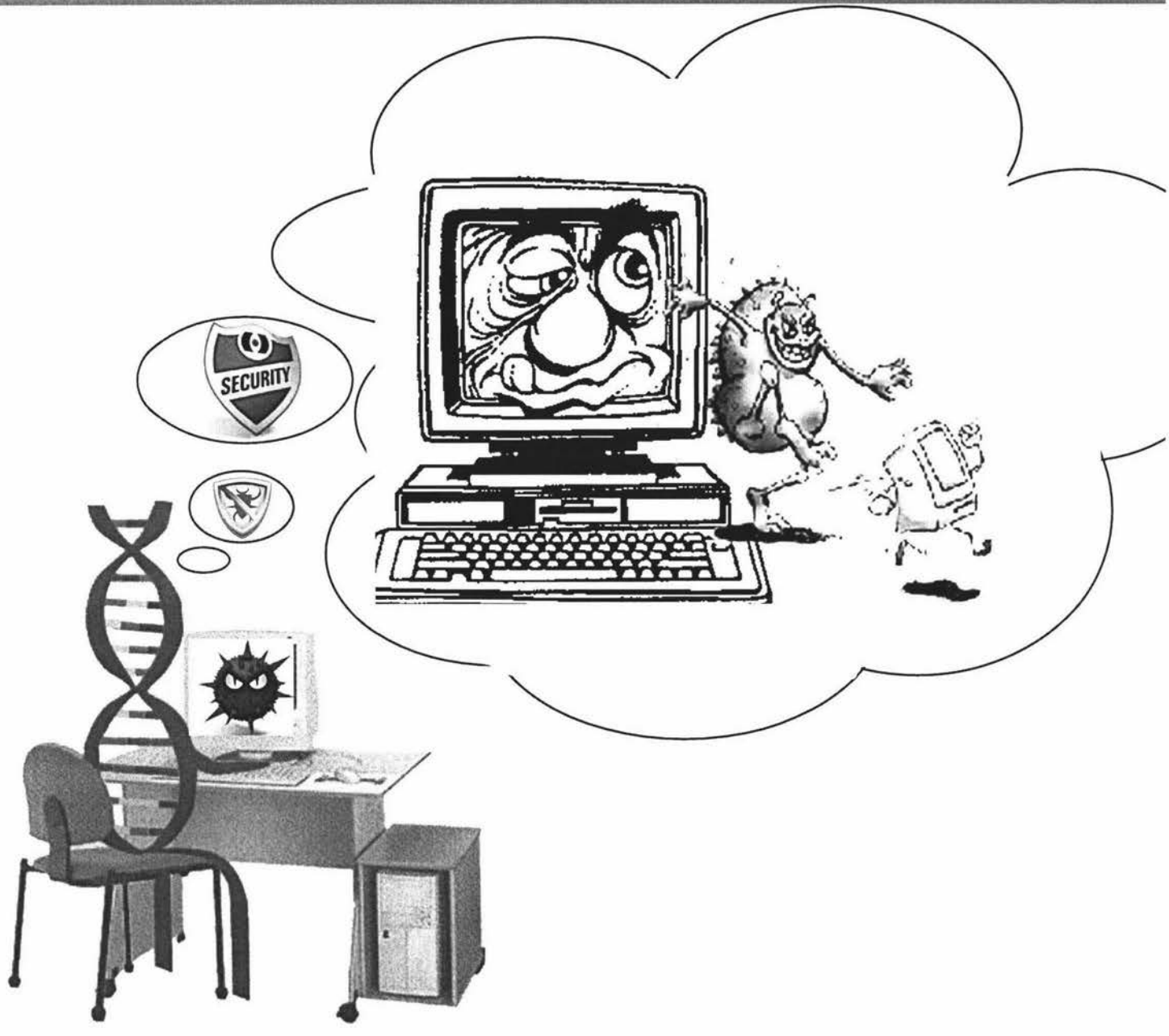
Massey University

Department of
Information Systems

157.899 Masters Thesis

Bio-mirrors and Networking Security



Prepared by **Douglas Mubayiwa**
Student 02430479

For the partial fulfilment of Masters of Information Sciences,
(Information Systems major), 2006

Abstract

Bioinformatics databanks have been the source of data to bioscience researchers over the years. They need this information especially in the analysis of raw data. When this data is needed, it has to be readily available. This thesis seeks to address the current problems of unavailable data at a critical time. Continued retrieval of data from far away sites is expensive in both time and network resources. Care must also be taken to secure this data otherwise by the time it reaches the researcher, it will be useless. In response to this problem being addressed, this thesis describes a way to move data securely so that the necessary data is stored nearest to whoever requires it.
A proposed initial prototype has been implemented with capacity to grow. The overall architecture of the system, the prototype and other related issues are also discussed in this thesis.

Acknowledgements

Research work of this magnitude cannot be created in isolation, especially when one is a family man, and I am grateful to many people for the help and support they rendered to me throughout the duration of this project. I would like to thank everyone who has contributed to the completion of this work, to come to mind pronto are:

• My supervisor, Sven Hartmann, for the much needed nudge when I needed wind to continue sailing, and for allowing me space when I needed it. Sven was my radar and I enjoyed his open door style. I learnt excellent management skills just by interacting with Sven. Thank you Sven.

• The staff of Massey University's Information Systems department, in particular, Markus and Madre , for providing advice and insight into how the systems at Massey University function, and for putting up with some of my more "interesting" ideas. "Oops! Did I say that?"

• Bryden for making sure I had no reason to complain about the department computers.

• My daughters for loving me even though I seemed to disappear when they needed me.

• Finally I would like to thank my wonderful wife, Rose, who has been my battle cry and primary source of inspiration over the years. Together through tight cooperation and a lot of sacrifices, we managed to come up academically, better than our starting point. We have shown that only through hard work and dedication can one's goals be achieved. Thank you Rose.

— Douglas Mubayiwa, May 2007

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background:

The objective of this research was to gain an in-depth understanding of the bioinformatics field and related topics. After this, I intended to develop a system that is able to recognize data movements between bio-mirror site and users. The amount of biological data is accumulating at an alarming rate, faster than improvements in computer hardware and current internetworking speeds. To make sense of complex biological systems, more data, often heterogeneous, should be included in increasingly complex computations. This data should be readily available otherwise the cost of retrieval will be too high. This thesis reviews the literature in the fields of Bioinformatics, Distributed Systems, Internetworking Security, and XML. A discussion on XML is featured prominently in this thesis because it is the vehicle chosen for biological data storage and distribution. It also has considerable impact on how a complete system should run. First this report will demystify bioinformatics issues before showing the proposed solution. It is hoped that the introduced system will inspire other developers to improve on it and make it an even better solution. The design and implementation details of the project are discussed to describe the developed prototype system. An in-depth analysis of the project's evaluation data is then given, concluding with a discussion of future work that has been identified.

1.2 Motivation:

In the biotechnology area, there is an ongoing need to share data that is stored and managed in publicly available web sites. The overall volume of the data stored in molecular biology databases has been growing tremendously, so much that transporting query results is hampered by existing relatively low Internet speeds. Currently, a major international project (called Bio-Mirror) is in place to provide for high-speed access to up-to-date molecular biology databases.

The idea is to establish local, regularly updated mirror sites that maintain local copies of molecular biology databases of interest for a country, region or community. Local bio-mirrors raise a number of security concerns for organisations running or using this service. I propose to investigate these concerns and to find pathways to address them. I will then describe the challenges I have encountered, particularly those relating to the long transaction times of bioinformatics resources and the difficulties of maintaining state, together with the solutions that I have developed to cope with these. High performance back-end computational infrastructure is essential for bioinformatics, but the cost of having the data to work with is equally critical.

1.3 Significance

Although the developments of relatively new Bio-Mirror (www.bio-mirror.net) archive services, such as to greatly reduce the bandwidth needs, while increasing their performance and usability to users, are well documented, this project should also have good impact, as a good solution would be adopted by most of these bio-mirror sites. This would reduce costs such as administrative by a great deal.

I believe also that adoption of this solution, albeit some changes here and there would result in a higher efficiency for repetitive high-throughput searches that result from the processing of large data sets. Much consideration has dwelt on such factors as the cost for storing and moving bioinformatics data. The subsequent cost of moving data from its new location to the end user is critical. We will discuss this in detail later on. In summary, we hope:

1. It will significantly reduce the time that a user must wait for requested data.
2. It considerably reduces the overall network traffic and servers processing power.
3. It will greatly reduces the load on the remote parent server, leaving it do handle other more critical tasks.

Web user interfaces for these sequence databases are well developed, we shall discus them, but wont direct much focus on these. For the purpose of the proposed system, it is sufficient to use the command line interface to issue commands and view outcome.

1.4 Related work

1.4.1 Hypothesis Research

The problem of searching huge biological databases on a large scale has been ongoing for several years. Biological databases are notorious for their massive volumes. Moving a file from its original location to your computer might require much in terms of storage space. Absolute care must be taken when selecting which data to move from a remote location and store locally. For the purpose of this research report, we shall not worry too much on the quality of data stored in such databases as GENBANK. We shall assume that this data is as accurate as it can be due to the measures kept in place at the point of entry. For our purpose, we shall also assume that all we are required to do is transport it as efficiently and as securely as possible. Primary focus will be on how this data gets to the final end user. In this day and age of disruptive man-in-the-middle, distributed computing data can reach the end user as useless piece of information that can seriously affect intended purpose like direction of study. Worse still, this could be data for medical purposes and might spell disaster to the end recipient (Kim, Choi, Hong, Kim, & Lee, 2003).

The bio-mirror project is doing a great job by trying to bring data as close to the end user as possible. More work however, needs to be done, in order to bring the 'right' data closer. The hypothesis of this research is that with careful and informed planning, we satisfy the customer more. We allow the customers to indirectly specify the data that they want and then fill a local mirror database with this invaluable data. We could start of with a near empty database that only contains information known to be important to the local community. Such information could be basic as competing organizations might choose to keep their need secretive for fear of giving away their cash cows. This will reduce the administration of the mirror site to low leaving the administrator to do other more demanding tasks. How this data is acquired is a subject of discussion on its own but for now, we must assume there is an aspect of secrecy where customers are kept from knowing each other.

To avoid filling up the database with information that is no longer required, we will assume we put checking measures to tell the administrator that particular data is now redundant. This data can then be cleaned up thereby freeing up invaluable space. Much related work to this is taking place in the research circles and a section dedicated to this will provide more detail.

1.4.2 Distributed Systems

The main ideas behind both bioinformatics and distributed systems computing are not new, and several successful bio-mirror systems and distributed systems have been developed on the basis of effectively distributing biological information in a timely manner (Strizh, 2006) (Bar-Or, Keren, Schuster, & Wolff, 2005). However, the current systems seem to focus only on the distribution of data. Much assumption is on the fact that this data reaches the end user as it was in the original state. Unfortunately, nowadays this is not the case anymore. In a dedicated section on Networking Security, we shall discuss the causes and cures of challenges.

1.5 Report outline

Following this introduction and quick review of related work, Chapter 2 of the thesis presents some essential bioinformatics and XML technology. Here is where we investigate the field of bioinformatics in a more detailed way. We also delve into XML schema extraction and illustrate some examples as to why it suits storage and in general the handling of bioinformatics data. Chapter 3 discusses bio-mirrors in detail and gives an example of one such site visited by this author during the study period. It starts off by exploring the major bioinformatics databanks and then look at the bio-mirror project itself. This bio-mirror project is, for the record, what we seek to enhance. Chapter 4 will focus on internetworking security. Issues such as the denial of service attack (DoS) and the address resolution attack (ARP) are covered to a great extend. We will also look at how security is provided at the physical layer and at the presentation layer of the OSI model, which itself would have been discussed, albeit, quickly. Chapter 5 will discus the implementation of the proposed system. Here, we will talk about the development language, the underlying physical and logical architectures as well as design decisions. Chapter 6 is basically about the test results. It is then followed by a discussion of future work in chapter 7 before the conclusion in the final chapter, chapter 8.