THE DEVELOPMENT OF
PERL SCRIPT DISTRIBUTION CENTER PORTAL
AT
MALAYSIAN GENOME INSTITUTE (MGI)

MOHD YUNUS BIN SHARUM

UNIVERSITI TEKNOLOGI MALAYSIA

THE DEVELOPMENT OF PERL SCRIPT DISTRIBUTION CENTER PORTAL AT MALAYSIAN GENOME INSTITUTE (MGI)

MOHD YUNUS BIN SHARUM

This TR is submitted in partial fulfillment of the
requirements for the award of
Masters of Science
(Computer Science - Real Time Software Engineering)

Centre for Advanced Software Engineering
Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

NOVEMBER 2006

ACKNOWLEDGEMENT

I praise gratefully to Allah S.W.T, because of His mercy, I manage to deliver this report within the specified time. Also solawat and salam to the Prophet Muhammad S.A.W., his family, relatives and friends.

I would like to dedicate my special thanks to the Malaysian Genome Institute (MGI), especially to the Director of MGI, Prof. Dr. Nor Muhammad Mahadi, Deputy Director of MGI, Prof. Dr. Rahmah Mohamed, and to the Lab Supervisor and Coordinator of MGI, Mr. Mohd Nor Mat Isa.

I also would like to thank my academic mentor at Centre for Advanced Software Engineering (CASE), Mr. Mohd Nazri bin Kama, for helping and guide me to complete this report. Special thanks to my wife and daughter, for showing a lot of patience and support during the completion of this report. Also to those who had provides a lot of help and support while I was in MGI – Mrs. Azurati, Raja Murzaferi, Ahmad Fuad, Khairil, Rashdi, Norazah, Aida Shazleen, Nor Haslinda, Hafiza Aida, Tajuddin and others.

A special thanks to Universiti Putra Malaysia, for sponsoring my Master studies, and to CASE and Universiti Teknologi Malaysia for providing me a great place to gain knowledge.

ABSTRACT

‘Perl Reference and Distribution Center’ portal system was developed as a web based system. The objective of this system is to help researchers in understanding Perl language, besides encouraging them to use Perl in their research. Perl is an acronym for *Practical Extraction and Report Language*, have the advantage in term of its syntax, since it was based on scripting and interpreted language. Perl script can be developed within a short period, plus with its open source feature, Perl had gained a lot of supports. Perl also have a huge library, and also used a lot in Bioinformatics’ field. Bioinformatics involved a lot use of software for data analysis tools. A lot of transactions and data manipulation has become a common requirement in bioinformatics, and they are better fulfilled by Perl. The main problem focused is the incapability of researchers in MGI to develop custom Perl application for their own purpose, since their attention is more on the biological research. This system tries to help researchers by providing easy access to the information regarding Perl through website system. Developed using WebE as development methodology, this system provides many applications, including tutorial section and a section for downloading application. This system would attract the researchers to Perl, and encourage them to learn the language because all the resources have been grouped in a single place. As a result, the researchers will find it is easier for them to learn Perl, and hopefully can increase the productivity and quality of their research.

ABSTRAK

Sistem Portal ‘Pusat Rujukan Dan Edaran Skrip Perl’, dibangunkan sebagai sistem berasaskan web. Objektif sistem ini adalah untuk membantu para penyelidik mendalami Perl, selain menggalakkan mereka menggunakan Perl dalam penyelidikan. Perl singkatan bagi *Practical Extraction and Report Language*, mempunyai kelebihan dari segi sintaks, kerana diasaskan kepada bahasa pengaturcaraan berbentuk skrip dan terjemahan (interpreted). Skrip Perl boleh dibangunkan dalam tempoh yang singkat, dan ditambah pula dengan ciri projek aturcara terbuka (open source), Perl menerima sokongan dari banyak pihak. Perl mempunyai sumber pustaka yang besar, dan juga digunakan dalam bidang bioinformatik. Bioinformatik banyak melibatkan penggunaan perisian sebagai alat dalam analisis data. Transaksi dan manipulasi data yang banyak adalah keperluan biasa dalam bioinformatik, dan ini dapat diselesaikan menggunakan Perl. Masalah yang ditumpukan ialah kelemahan penyelidik di MGI untuk membangunkan aplikasi Perl yang khusus untuk kegunaan mereka, kerana tumpuan penyelidik lebih kepada penyelidikan dalam bidang biologi. Sistem ini cuba membantu penyelidik dengan memudahkan capaian kepada maklumat berkaitan Perl melalui sistem laman web. Menggunakan WebE sebagai metodologi pembangunan, sistem ini menyediakan pelbagai aplikasi, selain ruangan tutorial dan ruang untuk memuat turun aplikasi. Sistem ini mampu menarik perhatian penyelidik untuk mempelajari Perl kerana semua sumber telah dihimpun dalam satu tempat. Hasilnya, penyelidik akan lebih mudah mempelajari dan menggunakan Perl, dan ini diharap dapat meningkatkan produktiviti dan kualiti penyelidikan mereka.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACRONYM	FULL NAME
ASP	Active Server Page
BLAST	Basic Local Alignment Search Tool
CGI	Common Gateway Interface
DFD	Data Flow Diagram
DNA	Deoxyribonucleic Acid
FastA	Fast Alignment
HTML	Hypertext Markup Language
MGI	Malaysian Genome Institute
NCBI	National Center for Biotechnology Information
PCR	Problem Change Request
Perl	Practical Extraction and Report Language
PHP	PHP Hypertext Preprocessor
Regex	Regular expression
SCM	Software Configuration Management
SDD	Software Design Document
SDP	Software Development Plan
SRS	Software Requirement Specification
STD	Software Test Description
STR	Software Test Report
WebE	Web Engineering

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Problem Background

The biology and biotechnology fields have a lot of division, and among of the fields focused by MGI is the research in molecular biology and structural molecular. Another new field of biology which MGI try to explore is the meta genomic. Besides doing research projects for universities and Malaysian government, MGI also provides facilities for researchers either from local or foreign countries who want to do research in Malaysia. Among the facilities that are provided by MGI are DNA sequencing and micro array machines, and also some lab facilities which are using the latest technology in related fields.

1.1.1 Bioinformatics

Bioinformatics is one of the fields in biology. Bioinformatics is said to be part of the computational biology fields, because it implements or applies a lot of mathematical techniques into biology. Some arguments indicates that bioinformatics is about using software and computer to analyze and process biological data, which include data prediction process. In MGI or even the around the world, the practice of microbiology and molecular structural biology is not just involving the use of machine and computer for lab works, but also involves bioinformatics field as one of the method to enhance the quality level of research's findings.

1.1.2 Software In Biology

In the beginning, the biological research work only focused on the lab works, because this field basically concentrates on research of living cells, and everything was done manually. But when research expands, and new findings are being discovered, the quality on the data that is collected gets questioned. Questions that aroused surrounding whether the data collected is reliable, and how the collected data can be expanded, which can be used as source for a new research. In fact, with the rapid development of new technology and the increased number of researches, the volume of data produced also started to increase.

Bioinformatics was born when biologist started to move their focus on the usage of software in doing analysis and research. Using software, researchers found that more reliable data can be collected, analyzed and stored effectively. Also, with the rapid development of software as a resource for computing and aiding tools, had help bioinformatics to expand. Software was used in variety of tasks, including software for databases, calculation software and spreadsheet, and also used for prediction and analysis tools etc.

Among the software categories mentioned above, there are lots of varieties in the aspect of usage. For example, there are lots of software for prediction and analysis tools. One of them is FastA or Fast Alignment. FastA is used to analyze the degree of similarity between amino acid sequences and also the similarity of nucleic acid sequences. These similarities are analyzed based on a selected formulae or algorithm, and comparison is done with one or multiple databases. Another tool, called BLAST (Basic Local Alignment Search Tool), is used to determine the percentage of similarities between sample sequence and sequences in database.

1.1.3 Software Used For Data Management

Nucleic acid is the basic molecule structure of Deoxyribonucleic Acid or DNA, while amino acid is the base for protein. DNA resides in nucleus, which is the central part of living creature's cells. DNA contains a genetic code. This code will be used to develop protein and specifies its function in organism. DNA will go through several phases of transitions (translation, mutation etc.), which is done by the cell to produce protein, and is basically a direct mapping between the protein and the DNA. In general, it is important to study about DNA so that researchers can understand more about the organisms, i.e. to predict the new protein that will be produced by a particular DNA, for the purpose of medical research and genetic engineering.

Microbiology focused highly on understanding the structure of DNA's molecule, while molecular structural study is about identifying the function and structure of particular protein. Because these fields mostly involve microbes, the research conducted will likely to produce a huge volume of data. The sequencing of single strands from DNA's helix chains normally produced thousands of character sequence or strings of C, G, A, T for DNA's code, where each alphabet of C, G, A or T currently can be represented by single byte of computer storage. Our body, or other multi cells creatures, is built by millions of cells. A single cell can contains several chains of DNA and this means billions of DNA molecules will likely have to be identified and analyzed to study living creatures. Just imagine how huge the data storage required, if all these data plus with their metadata (information about the data) needs to be stored.

With the use of software, the problem of keeping and managing this data can be solved using the database application. Database is highly needed in bioinformatics, because a great volumes of data needs to be stored and accessed. The database is not only required to be accessed locally, but also through internet. NCBI or National Center for Biotechnology Information in United States, for example, allows its genome database accessed using web service or web application. NCBI is one of the institutes that are conducting research on genome, and its database accepts submission from scientists worldwide. This allows other researchers from all over the world to access the result of the previous research faster and easier.

1.1.4 Software Used For Data Manipulation

Beside data management, the data manipulation process is also important in bioinformatics. There are many kinds of software used, and when we have to joint the input and output of this software, it will become a great problem. Normally, in bioinformatics, there are several phases of analysis and prediction processes need to be done to get complete information. The usage of software like BLAST, Phred, and FastA is common. Among the use of these software is to compare, analyze, predict, and stores data. Each of them sometimes require different kind of formats as input and will produce their own output. FastA for example, requires input in a format called FastA format, and will conduct analysis before producing output in its own format, while the FastA format in fact is not used by Phred. A research project might require to combines multiple software into single pipeline, and this process is cumbersome.

As a simple-but-tedious approach, this problem will be handled by manually editing the input or output file to joint them into single ‘pipeline’. These happened every time, and the researchers will suffer because editing process is time consuming and sometime is impossible because the huge size of the file. Beside that, there are also possibilities of doing a lot of mistakes during editing, and this will create a risk of unreliable data.

1.2 Problem Statement

A great amount of software use in bioinformatics has provided many advantages and improvement in doing biology research. Such advantages include providing database application to store information, and also many application and tools are being used to analyze and manipulate research data. There are also problems created along the way, such as the creation of incompatible and different formats between software. But these problems can be solved by creating custom application, using Perl, i.e. to handle format conversion, and the usage of software also can be enhanced by creating application to do automation in the research project. Perl is one of the best computer languages for bioinformatics, and have been chosen by many bioinformatics' experts. But biologists and researchers mostly are not computer expert. So we need to find a way to educate researchers and encourage them for using Perl.

1.3 Statement of Purpose

The purpose of this system is to provide a web portal system for the biologists and researchers in MGI to allow and attracts them to learn Perl. Through the web portal, the researchers can learn about Perl and learn how to implement a custom Perl application for their project, and also to provide other resources such as facilities for providing references, guides and downloads. The web portal system also going to provides some general-purpose applications that can be easily downloaded and used in research project.

1.4 System Objective

So, the objectives of this system are:

i. To provide a web portal system for biologists and researchers in MGI to learn Perl,

ii. To provide built-in applications through web that can be used by biologists and researchers in MGI,

iii. To provide a place to download general-purpose applications in Bioinformatics for researchers in MGI,

iv. To encourage researchers to learn Perl through actual system implementation as web portal.

1.5 System Scope

The scope for this system is only to educate researchers about Perl, and likely to provide reference and Perl application's download site for them. However, not all kinds of Perl's applications are available for download, except few applications that have been developed and used for bioinformatics before. The system is more like a skeleton system, assuming more custom Perl applications are added later. Also, the content and materials provided by the system are only focusing on Perl for bioinformatics, as to educate and attract researchers to learn and develop Perl application for their projects.

1.6 Report Organization

This report is divided into several main chapters. Table 1.1 explains about these chapters:

Table 1.1 : Chapters organization

Chapter	**Description**
1	Contains the introduction about the system development background, such as problem's background, statement of purpose, system's objective and system's scope.
2	Contains the literature studies that have been done prior to the system development. The literature studies include explanation about Perl language and its capability as scripting language, and also some studies about web, internet and Bioperl, a Perl's library for bioinformatics.
3	Explains about the methodology that was used in developing the system, comprising the methods, techniques and the processes done during system development phases.
4	Discuss about the results of system development, which include discussion about the effectiveness and some findings about the system.
5	Contains the conclusion that was made and also explains about some future works regarding the study and the system.

REFERENCES

1. P Aho, A.V., and Ullman, J.D. *Principles of Compiler Design*. Addison-Wesley. 1977
2. Aho, A.V., Sethi, R. and Ullman, J.D. *Compilers - Principles, Techniques and Tools*. Prentice Hall. 2003
3. Brown, M.C. *Perl: The Complete Reference. Second Edition.* Osborne/McGraw-Hill. 1998
4. Hall, J.N. and Schwartz, R.L. *Effective Perl Programming - Writing Better Programs with Perl*. Addison Wesley Longman. 1998
5. Harshawardhan, P.B. *Bioinformatics Principles and Applications*. Tata McGraw-Hill Publishing. 2005
6. Ignacimuthu, S. *Basic Bioinformatics*. Alpha Science. 2005
7. Jana, M.Z. *Panduan Analisis dan Rekabentuk Sistem*. Dewan Bahasa dan Pustaka. 1991
8. Pressman, R.S. *Software Engineering - A Practitioners Approach*. Fifth Edition. McGraw-Hill International. 2001
9. Roff, J.T. *UML - A Beginner's Guide*. McGraw-Hill/Osborne. 2003
10. Wagner, R. and Wyke R.A. *Javascript Unleashed*. Third Edition. Sams. 2000
11. Westhead, D.R., Parish, J.H. and Twyman, R.M. *Instant Notes - Bioinformatics*. BIOS Scientific Publishers. 2002

URL

12. http://genome.ukm.my/perldev/perlcenter/
13. http://genome.ukm.my/perldev/regex_builder/app/builder.html
14. http://genome.ukm.my/
15. http://www.cpan.org/
16. http://www.perl.org/
17. http://www.perl.org/community.html
18. http://www.perl.org/docs.html
19. http://www.bioperl.org/