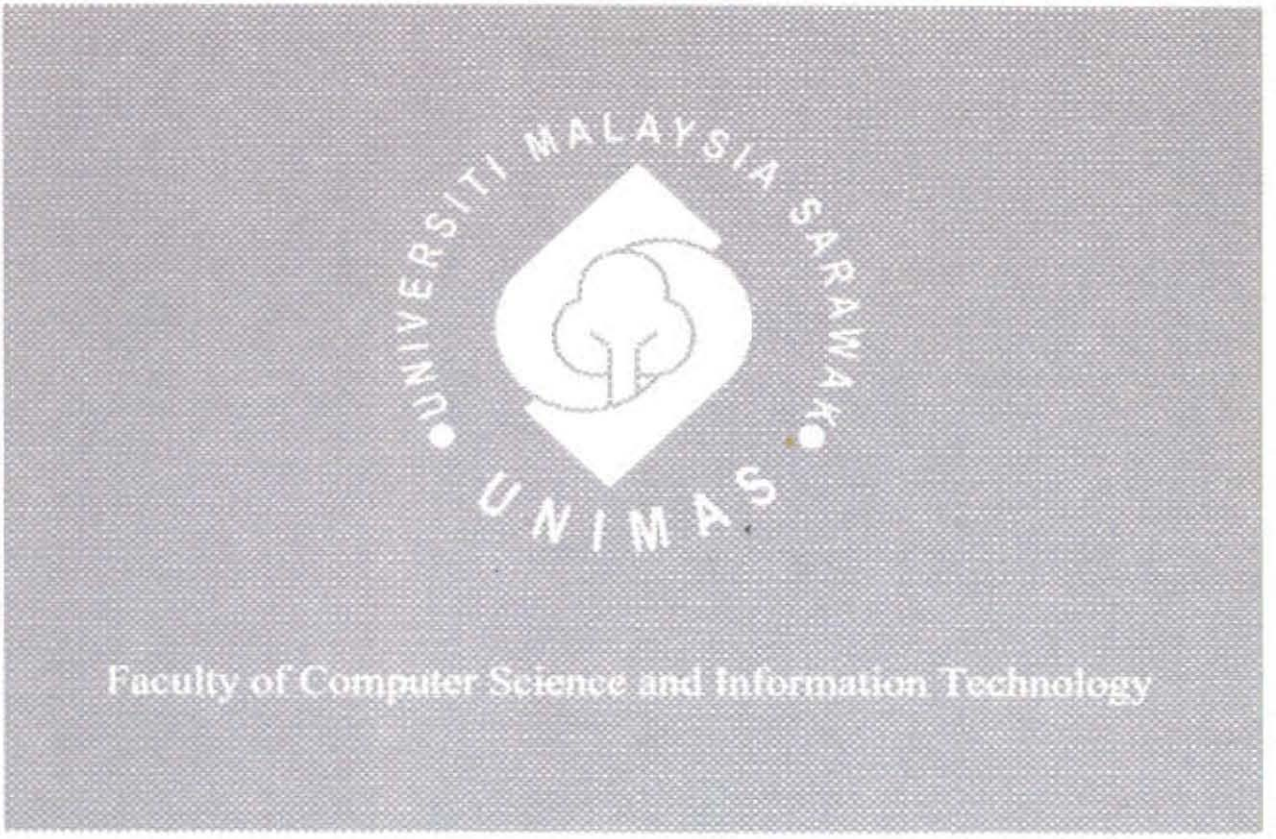

WEB BASED MICROBIAL GENOME ANNOTATION AUTOMATED SYSTEM

Tan Gui Ong

Bachelor of Computer Science with Honours
(Software Engineering)
2015

WEB BASED MICROBIAL GE[illegible]N AUTOMATED SYSTEM

TAN GUI ONG

This project is submitted in partial fulfilment of the
Requirements for the degree of
Bachelor of Computer Science with Honours
(Software Engineering)

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2015

WEB BASED MICROBIAL GE[illegible] AUTOMATED SYSTEM

TAN GUI ONG

This project is submitted in partial fulfilment of the
Requirements for the degree of
Bachelor of Computer Science with Honours
(Software Engineering)

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2015

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

WEB BASED MICROBIAL GENOME ANNOTATION AUTOMATED SYSTEM

ACADEMIC SESSION: 2014/2015

TAN GUI ONG

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

☐	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
☑	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
☐	UNRESTRICTED	

(AUTHOR'S SIGNATURE)

Permanent Address

66-C, JALAN NISBETH, PEKAN JABI, 85000 SEGAMAT, JOHOR

Date: 30/6/15

Validated by

(SUPERVISOR'S SIGNATURE)

Terrin Lim
Lecturer
Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK

Date: 30 June 2015

Note * Thesis refers to PhD, Master, and Bachelor Degree

** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak because they have been supportive of this project by providing the necessary guidelines and opportunities for me to develop this project. I would like to honorably acknowledge my supervisor Mr. Terrin Lim for his step by step guidance and advice throughout the period of this final year project. I would also like to express my gratefulness to my dynamic parents who by their encouragement and financial support made me strong throughout the period of this project. In light of the above, I would also like to appreciate the individual and collective effort of my friends and well-wishers and anyone who contributed to the completion of this project. I love them all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Nowadays, the demand for accurate and reliable automated genome annotation has increased significantly compared with the previous. This automated genome annotation system is able to help biologist to get the accurate gene predictions and distribute annotation data to the biosciences community effectively. Besides, this web-based application are allowed biologist to perform genome annotation at anywhere and anytime as long as the devices have a web browser and internet access. Furthermore, analyze and interpret the data into meaningful information. Lastly, present the information in graphical table or chart to gain insights on the organism of interest.

ABSTRAK

Pada zaman kini, permintaan automatik genom penjelasan yang tepat dan dipercayai mempunyai peningkatan yang ketara berbanding dengan sebelumnya. Automatik genom penjelasan sistem ini boleh membantu ahli biologi untuk mendapatkan ramalan gen dengan tepat dan mengedarkan data penjelasan kepada masyarakat biosains dengan berkesan. Selain itu, aplikasi berasaskan web ini adalah untuk membenarkan ahli biologi melaksanakan genom penjelasan pada bila-bila masa selagi peranti mempunyai pelayar web dan akses internet. Tambahan pula, menganalisis dan mentafsir data kepada maklumat yang bermakna. Akhir sekali, mengemukakan maklumat dalam grafik jadual atau carta untuk mendapatkan kepentingan pandangan organisma.

Chapter 1

Introduction

1.1 Introduction

Genome annotation is the process of identifying the location of genes and all of the coding regions in a genome, besides determining what those genes do. Therefore, gene finding is the most important step for the genome annotation. Gene finding is referring to the process of identifying the regions of genomic DNA that encode genes. Furthermore, protein-coding genes as well as RNA genes are also included, but it may also include prediction of other functional elements such as regulatory regions.

In the past decade, the quality of automated gene prediction in a microbial organism has improved gradually. Oak Ridge National Laboratory has cooperated with University of Tennessee at year 2007 to develop a microbial gene finding program name as Prodigal(Prokaryotic Dynamic Programming Genefinding Algorithm) in order to increase the number of correct identifications, both genes and all of the translation initiation sites for each gene, besides, reduce the overall number of false positives. At that moment, Prodigal has become the most popular microbial gene prediction algorithms throughout the world.

Prodigal software is able to improve gene structure prediction, improve translation initiation site recognition, and reduced false positives. Moreover, it is able to provide fast and accurate protein-coding gene prediction in GFF3, Genbank, or Sequin table format. Additionally, user

able to identify how Prodigal deal with gaps and has numerous options for allowing or forbidding genes to run into or span gaps.

BLAST (Basic Local Alignment Search Tool) program is used to compare primary biological sequence information such as amino-acid sequences of different proteins from the protein file that generated by the Prodigal program. Then, it uses protein GI query with the Blast2go database to get Gene Ontology (GO).

Therefore, the outcome of the project is to develop a web application name as Microbial Genome Annotation Automated System (MiGaSys). MiGaSys provides a complete set of methods for biologist to perform the microbial genome annotation. Furthermore, MiGaSys are combined Prodigal and BLAST software to go through gene prediction and gene annotation. The figure below is the pipeline of the system.

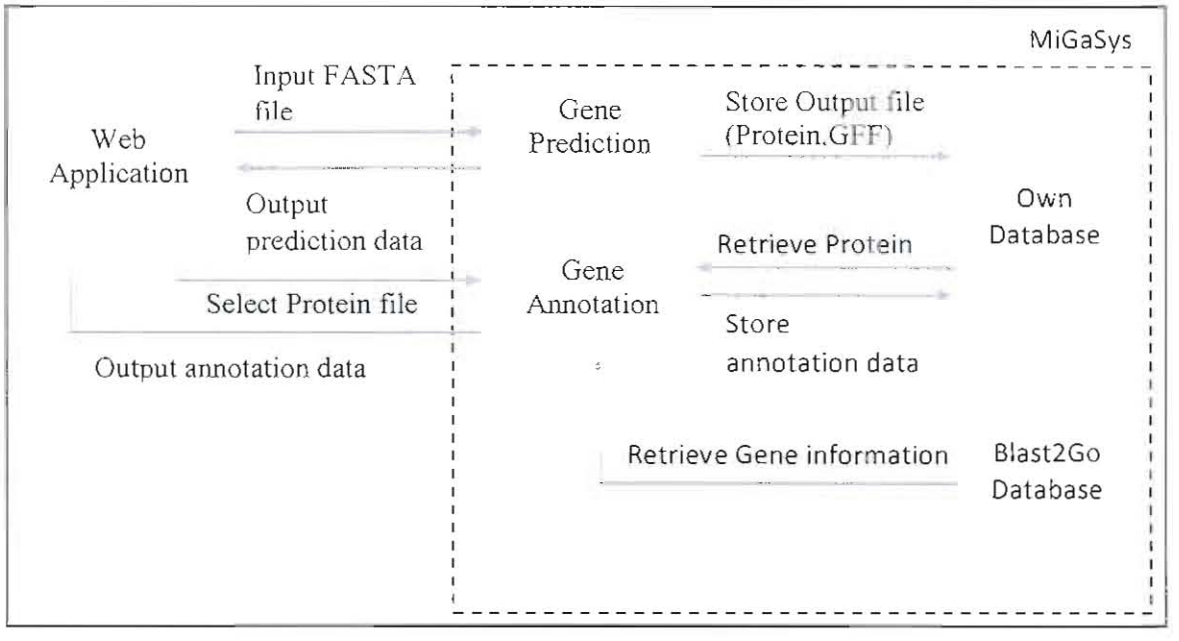

Figure 1.1 Pipeline of MiGaSys

1.2 Problem Statement

Most of the existing method to perform microbial genome annotation is in LINUX environment, mainly the biologists are not familiar with the environment. Therefore, there is a need to have a Bioinformatics pipeline in generating automated gene prediction and annotation for prokaryotic organisms that can execute automated genome annotation through a web browser. Besides that, there is no integrated/systematic tool to perform the tasks.

1.3 Objectives

There have three main objective need to achieve at the end of the project:

1. Develop an open-source gene prediction and gene annotation system using PHP.
2. Analyze and interpret the data into meaningful information, to be stored into database.
3. Retrieve data from database and present it in graphical table/chart to gain insights on the organisms of interest.

1.4 Procedures/Methodologies

Figure 1.2 Waterfall model

A software process model is an abstract representation of a process and presents a description of a process from some particular perspective. Software process models are general approaches

for organizing a project into activities. Many methodologies have been developed and also been introduced, hence, there are many software development methods nowadays such as waterfall model, spiral model, and component-based software engineering. This microbial genome annotation automated system will develop by using Waterfall model software development methodology.

- Requirement Analysis

  All the requirement will be identified and recorded. Besides, some of the possible solutions to the problem are determined.

- System Design

  When the requirement had been identified, a suitable system is designed. The proposed system will meet all the requirement.

- Implementation

  The system will be implemented.

- Identify defects & resolve bugs

  Once the system done developed, the system will test to reveal any bugs and errors. After that solve the bugs and errors that occur in the system.

- Operation and maintenance

  To improve and performance or other attributes.

1.5 Scope

The target of group in this project is Biologist. Looking at various open source Bioinformatics tools to analyze microbial genes. Benefit all aspects of biological research whereby identify the key features of the genome.

1.6 Project Significance

The demand for accurate and reliable automated genome annotation has increased significantly compared with the previous. This automated genome annotation system is able to help biologist to get the accurate gene predictions and distribute annotation data to the biosciences community effectively.

1.7 Project Schedule

Refer to appendix 1.

1.8 Expected Outcome

At the end of the project, user able to use this system at anywhere and anytime. Besides that, perform the microbial genome annotation without Linux environment. Thirdly, the user has user friendly interface to perform microbial genome annotation. Lastly, data easily exportable in various format.

Chapter 2

LITERATURE REVIEW

2.1. Introduction

This chapter focuses on explaining some of the general knowledge in biology. An overview about the process of microbial genome annotation. Review of several existing systems to understand the features, strengths and weaknesses of these systems. Follow by review the development tool that will apply in this web application.

2.2. General Biology Knowledge

2.2.1 Prokaryotes and Eukaryotes cell

In this planet, all the living organisms and microorganisms can be classified into two groups, prokaryotes and eukaryotes (Klappenbach, n.d.). These two groups have their own fundamental structure of cells. Prokaryotes is the single-celled organisms that without membrane-bound nucleus or any membrane-encased organelles. Most of the prokaryotes is made up of single cell (unicellular) but there also have few is made of collection of cells (multicellular). The Deoxyribonucleic Acid (DNA) in prokaryotes is less structure than eukaryotes is because the genetic material DNA is not bound within a nucleus. Scientist had divided the prokaryote into two groups, which is Bacteria and the Archaea. On the other hand, the eukaryotes is the organism which contains nucleus and membrane-encased organelles. The material DNA in eukaryotes is contained within a nucleus within the cell and DNA is organized into chromosomes. Eukaryotic organisms may be multicellular (Animal) or single-celled organisms

(Plants and fungi). Figure 2.1 shown below is the differences between eukaryotic and prokaryotic cell.

	Eukaryotic Cell	Prokaryotic Cell
Nucleus	Present	Absent
Number of chromosomes	More than one	One—but not true chromosome: Plasmids
Cell Type	Multicellular	Unicellular
True Membrane bound Nucleus	Present	Absent
Example	Animals and Plants	Bacteria and Archaea
Cell size	10-100um	1-10um

Table 2.1: The differences between eukaryotic aud prokaryotic cell

2.2.2 DNA, Genes, Genomes and Genome Annotation

Deoxyribonucleic acid (DNA) is the chemical compound that encodes the genetic instruction needed to develop and direct the activities of nearly all living organisms. DNA is a long polymer made from repeating units called nucleotides. According to the Saenger (1984), each DNA strand is made up of four chemical units, adenine (A), thymine (T), guanine (G), and cytosine (C) which called nucleotide bases. According to the opposite strands pair specifically, a A will always pair with a T while a C will always pairs with a G. The order of the nucleotide base has a different meaning of the information encoded in the part of the DNA molecule. The genome is an organism's complete set of DNA. Every single cell in our body contain approximately 3 billion DNA base pair, or letters, that make up the human genome.

DNA sequencing is the method to determine the exact order of the bases in a strand of DNA. Shotgun (Sanger) and high-throughput (Next-generation) sequencing had become the two broad categories to do the DNA sequencing. However, high-throughput sequencing had been overtaken the shotgun sequencing technology. According to Hall (2007), this is because high-throughput sequencing provide high demand for low-cost sequencing, besides that high-throughput sequencing technologies that parallelize the sequencing process, therefore able to produce thousands or millions of sequences at once.

According to Stein (2001), value of the genome is only as good as once it had been annotated. Annotation is the bridges that connect the sequence and the biological organism. The main objective of annotation is to identify the key features of the genome, the function about genes and their products. Followed by attaching information to the sequences. Genome annotation consists of three main steps. First step is to identify portions of the genome that do not code for proteins. Second step is to identify elements of the genome, the process called gene prediction and the last step is attaching biological information to the element.

Gene prediction or gene finding is one of the first and the most important process in computational biology. The main purpose of gene prediction is to identify the regions of genomic DNA that encode genes. This includes protein-coding genes as well as RNA genes, but also include prediction of other functional elements such as regulatory regions. Figure 2.2 shown the generic process for bacterial genome annotation.

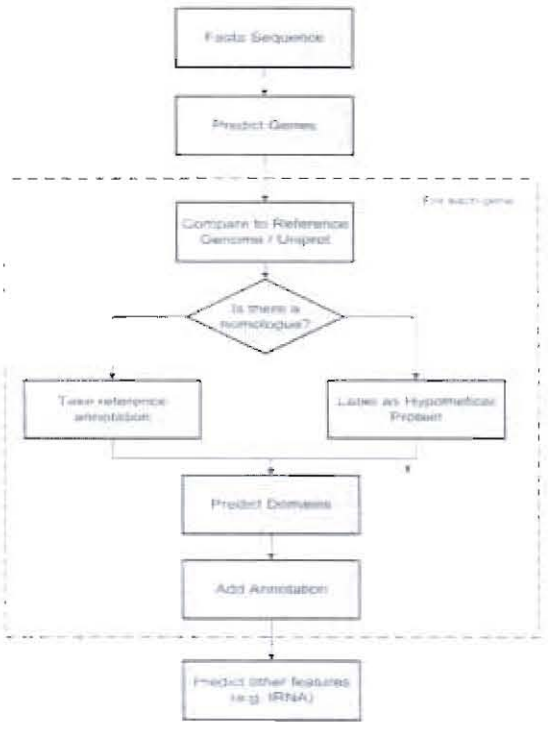

Figure 2.1: A generic process for bacterial genome annotation
Source:http://bib.oxfordjournals.org/content/early/2012/03/08/bib.bbs007.full

2.2.3 BLAST: Basic Local Alignment Search Tool

BLAST is an algorithm for finding regions of local similarity between sequences. BLAST will compare primary biological sequence information, such as nucleotide or protein sequences to sequence databases and identify the statistical significance of matches. Besides that, BLAST able to show functional and evolutionary relationships between sequences, and able to help identify members of gene families.

In addition, BLAST accept the sequence in FASTA or Genbank format or Accession Number (GI number). GI number is created by NCBI which is a series of digits that assigned consecutively to each sequence record. Where the output can be delivered in a variety of formats. These formats include HTML, plain text and XML formatting.

2.3 Existing Systems

2.3.1 AGeS: A Software System for Microbial Genome Sequence Annotation

AGeS is the Annotation of microbial genome sequences that developed as a standalone software application, which incorporates with the in-house Bioinformatics tools and database. User can install this software on Linux computer or a Linux cluster. The reason of developing into a standalone application is because annotation of genomes from next-generation sequencing required fast, high-throughput, and fully integrated and automated. Besides that, standalone application able to provide the best solution for researchers that need to annotate a large number of genomes and store the output locally for further analysis.

AGeS system was designed to support three main capahilities. The first is all the FASTA format and resulting annotation data will store in the central database and contain an easy-to-use graphical user interfaces (GUIs) for user to manipulation and performed visualization steps. The second is using the integrated software pipeline do the microbial genome annotation. Use Do-It-Yourself Annotation (DIYA) framework to analyze sequence contigs and locates genomic regions that code for proteins, RNAs, and other genomic elements. After that, using an in-house-developed high-throughput pipeline, the Pipeline for Protein Annotation (PIPA) to identify and annotate protein-coding regions. The third capability is using open-source genome browser GBrowse to do the visualization of annotated sequences.

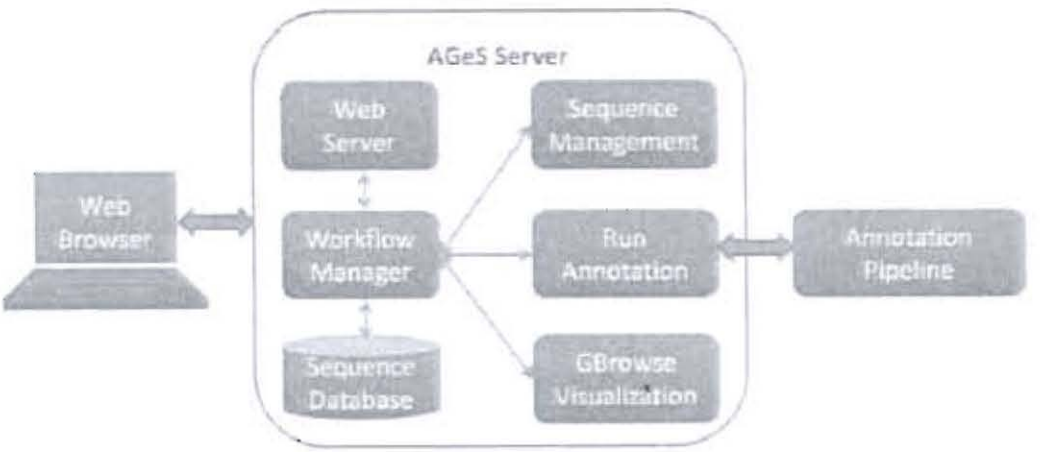

Figure 2.2: AGeS system architecture

Source: http://www.ncbi.nlm.nih.gov/pubmed/21408217

The figure 2.3 shown the system architecture of AGeS. AGeS server provide the easy-to-use GUI accessible via a web browser, an embedded relational database management system for storing sequences and other job-related data, and a high-throughput software pipeline for the annotation of input genomes. AGeS server allow multiple users to access the AGeS GUI by using the web browsers. The AGeS provide three functions for user.

1. Sequence management for uploading and manipulating genomic sequence and their properties.
2. Job submission for running the annotation pipeline.
3. Graphical visualization of the annotated sequence with GBrowse.

The workflow manager module in AGeS server is to guide the entire lifecycle of the user's job, starting from the upload of an input sequence and ending with the visualization of the annotated sequences. The annotation pipeline is the AGeS standalone application that initiated by the workflow manager when user's request and runs in batch mode on the Linux cluster to achieve high throughput. There have two options for user to obtain their annotation result. First is bookmarking the result page and loading it back at a later time. Second is providing an e-mail address for automated notification once the annotation completed. After the completion of annotation, the annotation result will stored within a user's session. Therefore, user able to view it by using GBrowse or download it as GenBank file.

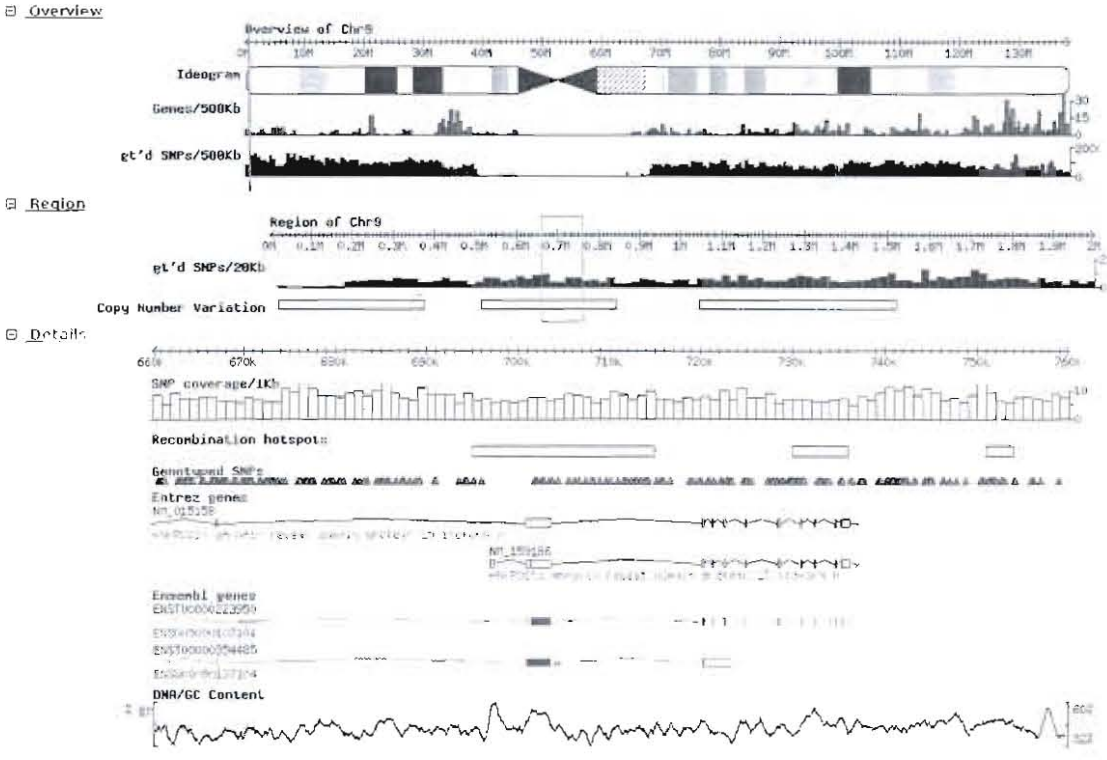

Figure 2.3: Annotation result in GBrowse

Source: http://gmod.org/wiki/GBrowse

2.3.2 BASys: a web server for automated bacterial genome annotation

BASys (Bacteria Annotation System) is the web application that perform automated, in-depth annotation of bacterial genomic example chromosomes and plasmid sequences. BASys not a single program or software that run the automated bacterial genome annotation. It is using more than 30 programs to determine approximately 60 annotation subfields for each gene, there including gene/protein name, GO function and so on. ,

BASys allow the anonymous and login user access, monitor and retrieval of genome annotation. The anonymous user is only allowed to summit single chromosome for annotation, the system will email the anonymous user a secure URL to monitor the progress and retrieve their annotation once completed. On the other hand, registered user can submit multiple chromosomes and plasmid annotation and monitor them at the same time. BASys provided front-end web interface for user submitting the raw genomic data in FASTA format that need go through the annotation process. After that the annotation engine will analyze the chromosome data and generating the annotations. Lastly, interpret the full annotation genome information into the various graphics, HTML and textual output.