PROTEIN YEAST CLASSIFICATION INFORMATION SYSTEM

SITI HAJAR BINTI AB RAHMAN

A report submitted in partial fulfillment
of the requirements for the award of the degree of
Bachelor of Computer Science (Computer Systems & Networking)

Faculty of Systems Computer & Software Engineering
Universiti Malaysia Pahang

MAY 2011

ABSTRACT

Protein Yeast Classification Information System is a system design for scaintific area for sceintist to use or view the protein of Saccharomyces Cerevisiae or yeast. For this system, the protein Saccharomyces Cerevisiae will be classified into structured data organizes manner replacing the existing system. Previously, the existing system contain a lot od data and there is unstructured data in it. As the result it consume a lot of space to store all data. Besides that, it is also time consuming when need to find particular of protein for viewing purpose. The system using Microsoft SQL to classified a data. In this system, the user can view The protein data easily and fast. This system can provide the scientific community with an integrated set for browsing and extracting information of protein yeast network for yeast. This system also use Spiral SDLC model as a methos to develop this system.

ABSTRAK

Protein Yeast Classification Information System adalah satu rekaan sistem yang di reka untuk bidang sains untuk saintis menggunakan atau melihat protein ragi atau nama saintific ialah Saccharomyces Cerevisiae. Untuk sistem ini. Ragi akan di bahagikan ke dalam bentuk yang lebih tersusun bagi menggantikan sistem yang telah sedia ada. Ianya kerana. Sistem yang sedia ada mengandungi banyak data yang tidak tersusun. Ini akan menyebabkan, penggunaan ruang yang lebih banyak untuk menyimpan data dan ini akan merugikan ruang. Selain itu, masa juga akan digunakan dengan lebih banyak untuk mencari nama protein-protein ragi. Sistem ini mengunakan Microsoft SQL untuk membahagikan data. Bagi sistem ini, pengguna boleh melihat protein ragi dengan mudah dan cepat. Sistem ini juga menggunakan kaedah SDLC model untuk mereka bentuk sistem.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS

PYCIS	- Protein Yeast Classification Information System
UniProt	- Universal Protein Resources
YPD	- Yeast Proteome Database
CYGD	- Comprehemsive Yeast Genome Database
SGD	- Sacchomarcyes Genome Database
PANTHER	- Protein Analysis Through Evolutionary Relationship
CATH	- Protein Structure Classification System
SQL	- Structured Query Language

CHAPTER 1

INTRODUCTION

1.0 Introduction

Classification

This chapter will provide a brief overview of the entire project including objective, scope and problem statement of the project. Classification of protein database is a largely manual classification of protein structural domains based on similarities of their amino acid sequences and three-dimensional structures. Based from the classification that have been state above, this project will develop the system that will classified proteins yeast into a specified groups. As we know, yeast or the scientific name is *Saccharomyes* usually used in the baker such as to make bread, wine, and beer. Yeast physiology can be either obligatory aerobic or facultative fermentative. There is no known obligatory anaerobic yeast. In the absence of oxygen, fermentative yeast produce their energy by converting sugars into carbon dioxide and ethanol (alcohol). In brewing, the ethanol is used.

As the technology growth, there are lot of data about protein yeast we can find out in the web browser now days such as YDP, MID and many more, but the data in that browser is unstructured in proper way. A lot of scientist and student find difficulties to differentiate which protein belong to which group because there a lot of protein in the yeast. Classification is a method generalization of minimal distance methods, which form the basis of several machine learning and pattern recognition methods. Classification success depends on adaptive parameters and procedures used in the contractions of the classification model.

Protein classification is a method to classified of protein domain structures. Each protein has been chopped into structural domains and assigned into homologous super families (group of domain that are related by evolution). From the record in the YDP, the total proteins in their database is 106021 and it will increase year after year because there a lot of experimental is done by the scientist and they will submit the result in the YDP as a references. Because of that all the data is unstructured very well in their database.

Therefore, by develop the protein yeast classification in the science field it could make the user more easier to use. It help to solve a problem unstructured data in database and makes the existing system more useful.

Figure 1.1 Yeast

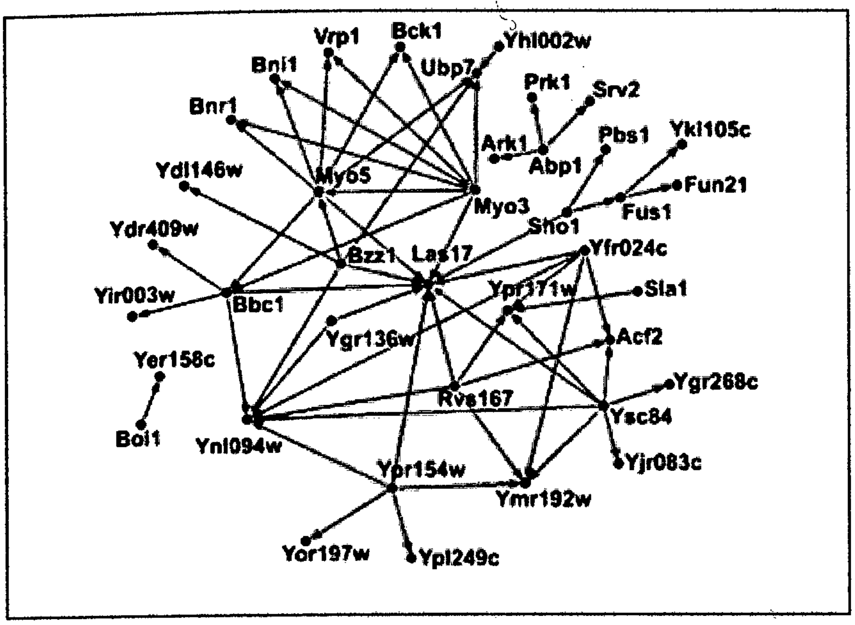


Figure 1.2 Protein Yeast Interaction

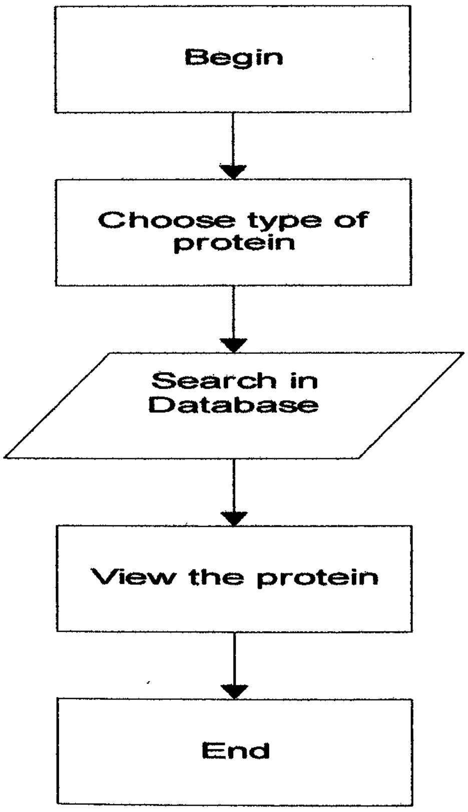
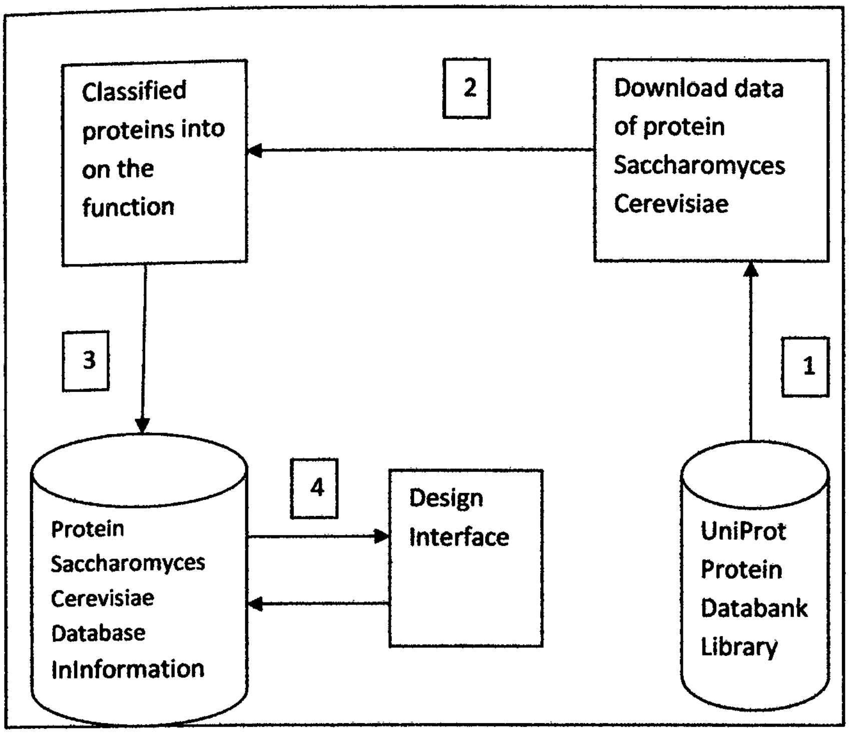


Figure 1.3 Flowchart of System


Classified proteins into on the function
2
Download data of protein Saccharomyces Cerevisiae
3
1
4
Protein Saccharomyces Cerevisiae Database Inlnformation
Design Interface
UniProt Protein Databank Library

Figures 1 .4 Flowchart of Data

1.1 Problem Statement

1. There are unstructured data in protein data bank that lead to the problem to find the correct protein of yeast. It is because there a lot of data in the protein database. Thus the data keep messy way. It took a long time when the users to browse the information about the protein yeast.
2. Unorganized and inefficiency data management makes the system not user friendly.
3. There are lot of space being used to store the data of protein. Because the data is unorganized, when the admin store the data, it will cause a lot of space been used to store the protein data in data bank. Maybe it will affect the performance of the system such as the system will slow and error.
4. The other problem are, the users cannot recognize which protein of yeast belong to correct group. As we know, in the data bank, there are lot of unorganized data and it mix along with others

1.2 Objective

1. To provide the scientific community with an integrated set for browsing and extracting information of protein yeast network for yeast.
2. For users to visualize and analyze yeast protein group network.
3. To identify protein classification in huge database in public database.

1.3 Scopes

1. There will be protein yeast of yeast to be used in this system
2. Classification protein yeast.

1.4 Thesis Organization

The thesis consists of four chapters. Chapter one is explanation of introduction to system and research. The topics in this chapter have proposed will discuss which are introduction, objectives, problem statements, and scope of the project.

Even though chapter two will discuss about the research for project that has been chosen. The researches divide into two that are for current system or case study and research for technique that will be used to develop current system.

For chapter three will be discuss on approach and overall work load to develop this system. This content consist of technique for implementation the projects.

In chapter four will be discuss on result that has been received and all data analysis. The content that must have in this chapter consists of analysis of result, difficulty of projects and improvement of project.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

This chapter will discussed on the critical points of current knowledge on a particular topic. The literature review usually precedes a research proposal, methodology and result section. Its ultimate goal is to bring the reader up to date with current literature on a topic and form the basis for another goal, such as the justification for future research in the area.

A literature review can be just a simple summary of the sources, but it usually has an organizational pattern and combines both summary and synthesis. A summary is a recap of the important information of the source, but a synthesis is a re-organization, or a reshuffling, of that information. It might give a new interpretation of old material or combine new with old interpretations. Or it might trace the intellectual progression of the field, including major debates. And depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant.

Nowadays, a variety of high throughput technologies have been developed[1]. For example, to discover how genes and their encoded protein function together in normal and disease states[2]. many of these technologies use the baker's yeast, Saccharomyes cerevisiae, either as a model organism to study biological processes that yeast have in common with humans, or as an assay system to test the function of genes and proteins from humans or other model organisms[3]. For examples, a technologies known as the "yeast two-hybrid system" uses yeast cells essentially as test tubes to assay interaction between specific protein from other organisms[4], [5]. To find the information of proteins that are required, a system for classifying the proteins is needed. Identify the protein needed based on protein "Relevant Biological Function" that have been divide into three section. It organizes the intrinsic and large volume of data in public database. Furthermore, it help reduce the time consuming to browse the information about protein needed.

2.1 A STUDY OF RELEVANT BIOLOGICAL FUNCTION CLASSIFICATION

As databases of protein yeast classified into relevant biological function become increasingly available, and as the number of sequenced protein grows exponentially, techniques to automatically classify unknown proteins into biological function become more important. In the relevant biological function, it will divide into three groups there are "Protein characterized by Genetic or Biochemistry", "Protein known by Homology to Characterized Protein", and "Protein of Unknown Function". Many approaches have been presented for protein classification. Classification of protein provides valuable clues to structure, activity, and metabolic role. Protein relevant biological function classification has several advantages as a basic approach for large-scale genomic annotation which is improves the identification into distinct classes, each of which shares a unique expression property and is presumably under same regulatory mechanism. And the function of well-characterized genes would give insight into uncharacterized ones in the same class[6]. Furthermore, it provides an effective means to retrieve relevant biological information from vast amount of data [6]. A lot of researcher have been performed to detect the association between gene expression pattern and biological function.

Most of those researches adopt an approach to find specific biological function associated with each class classified by gene expression, namely, mapping from expression to function.

2.2 A STUDY ON Universal Protein Resources (UniProt)

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Cluster (UniRef), and the UniProt Archive (UniParc). The UniProt Metagenomic and Enviromental Sequences (UniMes) database is a repository specifically developed for metagenomic and environmental data. UniProt is based on protein sequences, many of which are derived from genome sequencing project [7]. It contains a large amount of information about the biological function of protein derived from the research literature [7]. It also provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information [8]. The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data [8].

2.3 TYPES OF CLASSIFICATION

Classification of protein database is a largely manual classification of protein structural domains based on similarities of their amino acid sequences and three-dimensional structures [9]. Classification success depends on adaptive parameters and procedures used in the construction of the classification model [9]. Classification is a method generalization of minimal distance methods, which form the basis of several machine learning and pattern recognition methods. Protein classification is a method to classified domain structures [10]. Each protein has been chopped into structural domains and assigned into homologous super families (group of domain that are related by evolution) [10].

2.3.1 DYANAMIC CLASSIFICATION

The dynamic classification system does not use fixed classification method but use dynamic classification method. The user who uses this system can select classifier filter that he wants to in the classifier library and design a classifier which he wants to construct.

Table 1.1: comparison with previous system

	Previous system	Dynamic classifier system
Features	The user use a fixed classification method	There are various classifier filters and the user can design classifiers.
Strength	Static classification	Dynamic classification. The user can try constructing various classification systems.
Weakness	The user cannot construct various classification system	There are some possibilities of error in whole classification result by irrelevant classifiers.

2.3.2 P-tree CLASSIFICATION

P-trees are a lossless, compressed, and data-mining-ready data structure [11]. This data structure has been successfully applied in data mining applications ranging from Classification and Clustering with K-Nearest Neighbor, to Classification with Decision Tree Induction, to Association Rule Mining [11]. A basic P-tree represent one attribute bit that is reorganized into a tree structure by recursively sub-dividing, while recording the predicate truth value regarding purity for each division. Each level of the tree contains truth-bits that represent pure-trees and can then be used for fast computation of count. This construction is continued recursively down each tree path till until a pure sub-division is reached that is entirely pure. The root count of

2.3.3 COLLECTIVE CLASSIFICATION

Collective classification refers to the combined classification of a set of interlinked object using all three types of information. Note that, sometimes the phrase relational classification is used to denote an approach that concentrates on classifying network data by using only the first two type of correlation. However, in many applications that produces data with correlation between labels of interconnected objects (a phenomenon sometimes referred to as relational autocorrelation [12]) labels of the object in the neighborhood are often unknown as well. In such cases, it becomes necessary to simultaneously infer the labels for all the objects in the network.

2.4 A STUDY ON EXISTING SYSTEM

As a guide for this Protein Relevant Biological of *Saccharomyces cerevisiae* Classification System, some existing system were picked and were analyze to get methods and also how the protein relevant biological is classified. The lists of all the existing systems are:

1. The Yeast Proteome Database (YPD)
2. The MIPS Comprehensive Yeast Genome Database (CYGD)
3. The Mycobank Yeast Species Database
4. The Saccharomyces Genome Database (SGD)
5. InterPro
6. Protein Analysis through Evolutionary Relationships (PANTHER) Classification System.
7. Protein Structure Classification System (CATH)

2.4.1 THE YEAST PROTEOME DATABASE (YPD)

The Yeast Proteome Database (YPD) is a model for the organization and presentation of comprehensive protein information. Based on the detailed curation of the scientific literature for the yeast *Saccharomyces cerevisiae.* YPD contains more than 50 000 annotations lines derived from the review of 8500 research publications. The YPD is the first annotated proteome database for any organism [13].

YPD is annotated by in-depth cur ration of the research literature and it is a proteome database because it contain entries for each known or predicted protein of *Saccharomyces cerevisiae.* the information for each of the approximately 61000 yeast protein is presented in a convenient one-page format. In this system, users can display pop-up windows with more detailed information or description, such as the full protein sequence, protein-protein interactions, regulation of gene expression, protein modification and sequence alignments with protein from humans and model organism [14].

The annotation and properties contained in YPD are written by a staff of PhD level curators experienced in yeast research. The curatorial staff has read and annotated 85000 research articles. As an indicator, YPD tracks the number of yeast protein that have an assigned function, as determined by generic or biochemical experiment. YPD is now based in a relational (Oracle) format which affords major improvement in the structuring of search queries. YPD has expended the classification scheme for protein for proteins, to better define the protein for the reader and to allow more powerful searches. These data are displayed together in an expanded Properties table.

YPD now provides sequences alignment on the Related Genes pop-up window. It connect protein with common physical properties or common gene regulation. This past year YPD introduced the first presentation of functional genomic data integrated into the proteome database. The data, kindly provided by provided by Joseph DeRisi, Vishawanth Iyer and Patrick Brown, describe the effect of diauxic shift on transcript abundance, measured simultaneously for every gene in the genome.

YPD curates all newly published articles concerning yeast protein and is making a major effort to complete duration of the older literature. In the near future, YPD will complete the assignment of protein roles, functions, and pathways based on experimental evidence in the curates literature.