PEPDB CONSTRUCTION AND LARGE-SCALE ANALYSIS OF ESTS

PEPDB CONSTRUCTION

AND

LARGE-SCALE ANALYSIS OF ESTS

by

LING SHEN, B.Sc. (HONS.)

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2003) McMaster University

(Biology) Hamilton, Ontario

TITLE: PEPdb Construction and Large-scale Analysis of ESTs

AUTHOR: Ling Shen, B.Sc. Hons. (McMaster University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: [xi], 115

ABSTRACT

The Protist EST program (PEP) aims to explore the diversity of eukaryotic genomes, in a systematic and comprehensive way. A main element of the PEP initiative is to establish a database, the Protist EST Database (PEPdb), which is the centerpiece of the PEP collaboration. The major functions of the PEPdb are management of the data generated by PEP, analysis of these data, and to allow collected sequence information to be accessed via the Internet by PEP members or other users. In this project, a consistent and easy to use relational database was implemented. All information about PEP members, Publications, Libraries and ESTs can be stored in the database system. The operations are achieved by a friendly user interface. This database stores about 10000 records and is displayed on the web site "http://info.biology.mcmaster.ca/ling/estHome.html" for demonstration.

An analysis of ESTs from the ciliated protozoan *Tetrahymena thermophila* was undertaken. A total of 3740 non-redundant gene assemblies and singletons from TIGR were analyzed. These sequences have been compared against the NCBI non-redundant protein and nucleotide databases using BLASTX and BLASTN to identify putative genes. Of 850 highly significant matches with an expect value cut-off of 10^{-20}, 35.5% represent genes previously cloned from *T. thermophila*, and 64.5% had significant similarity to genes from other organisms deposited in the NCBI. There are 26 sequences (3.1%) that matched signal transduction proteins, including Rac, Ras, MAPK, ERK1, PKC, cAMP and 14-3-3 (a protein involved in signal transduction, exocytosis and cell cycle regulation). This result indicates that *T. thermophila* likely encodes the MAPK/ERK signaling pathway. About 53 sequences (6.2%) matched to cytoskeleton proteins which were divided into two groups. The first group matched genes coding for microtubules, especially to tubulin genes. The other group matched to microfilament genes including one actin, three actin-related and one profilin proteins. There were no sequences similar to intermediate filaments. Comparison of the EST counts from one gene provide absolute estimates of mRNA expression levels. The most abundant genes represented are enolase, SerH3 and Tubulin. Among 850 highly significant similarities, 196 were restricted to the ciliophora. GRL and SerH are ciliate-specific genes. There were 508 sequences that had highly significant matches (expect value $< 10^{-20}$) to human genes. Approximately 189 of them were present in humans but not found in the completely sequenced *Saccharomyces cerevisiae.* Based on Venn diagram analysis, *T. thermophila* contains abundant Eukaryotic specific proteins and many prokaryotic-like genes, and some metabolic enzymes in *T. thermophila* are also present in plants. These results support the fact that *T. thermophila* is an excellent unicellular model system for gene discovery and functional analysis.

ACKNOWLEGEMENTS

I would like to express my sincere appreciation to Dr. Brian Golding for leading me to the challenging area, bioinformatics, including his support and guidance the projects in this thesis. I am greatly thankful to Dr. Rama Shankar Singh and Dr. Elizabeth A. Weretilnyk for taking the time to be the members of my committee. Thank you very much to Graydon Hoare, Heidi Musters and to my family for their help. I would like to express thanks to Melanie Huntley, Brett Whitty, Frances Raftis, Weilong Hao, Paulo Niun and all other colleagues in our lab.

Contents

1 The Protist EST Database (PEPdb) Construction 1

1.1 Introduction 1

1.1.1 The function of PEPdb 2

1.1.2 Database implementation 4

1.2 System Structure 6

1.2.1 Environment 6

1.2.2 Database Architecture 11

1.3 Database Design 12

1.3.1 Use case phase 13

1.3.2 Entity-Relationship Model 16

1.3.3 Translating ER diagram into a relational database and refining the data structure 20

1.3.4 Normalization 36

1.3.5 Data Integrity 39

1.4 Graphic User Interface 40

1.4.1 Graphic User Interface Design 40

1.4.2 General Graphic User Interface Structure 41

1.4.3 Graphical User Interface 41

1.4.4 Techniques for User interface 54

1.4.5 User Interface Programming 59

1.5 Summary 63

2 Large-scale EST Analysis in the Ciliated Protozoan *Tetrahymena thermophila* 66

2.1 Introduction 66

2.2 Methods 69

2.3 Results and Discussion 71

2.3.1 Overall distribution of the general categories. 71

2.3.2 Sequences displaying no significant matches in public databases. . . 77

2.3.3 Classification of genes according to their functions. 77

2.3.4 Analysis of expression levels. 81

2.3.5 Identification of genes restricted to the Ciliophora and the Alveolata. 85

2.3.6 Sequences match proteins from human but not *Saccharomyces cerevisiae*. 87

2.3.7 Identification of MAPK/ERK signaling cascade in *T. thermophila* . 88

2.3.8 Phylogenetic comparison of gene homologies. 89

2.4 Summary 106

List of Figures

1.1 The general Web database structure and transaction 9

1.2 Use case diagram 14

1.3 The Entity-Relationship diagram 17

1.4 Refining the *EST* entity 23

1.5 Refining the *Consensus Sequence* entity 26

1.6 Refining the *Publication* and *Person* entities 29

1.7 Refining the *EST Library* entities 32

1.8 The Protist EST database (PEPdb) Schema 34

1.9 General Graphic User Interface Structure 42

1.10 Home page and Advanced search 44

1.11 Blast page and Batch search 46

1.12 Selected records displaying and Individual EST displaying 48

1.13 Cluster search, Contig displaying, Individual EST displaying and Library displaying 50

1.14 Member log on, User registration and Submit 52

1.15 PEPdb Submit System 55

2.1 Venn Diagram 98

2.2 Distribution of function classes within superkingdoms 100

2.3 Distribution of function classes within kingdoms 102

2.4 Distribution of function classes within phyla 104

List of Tables

2.1 Summary of EST sequence categories 72

2.2 The distribution of the sequences with significant matches to NCBI 75

2.3 Number of similarities within sixteen protein function classes 78

2.4 Relative mRNA abundance 82

2.5 Distribution of the sequences within kingdom groups 96

Chapter 1

The Protist EST Database (PEPdb) Construction

1.1 Introduction

The eukaryotic Kingdom Protista are mainly unicellular eukaryotes and are thought to include more evolutionary and biology diversity than the multicellular kingdoms of animals, plants and fungi combined (Gray 2001). Many protists are associated with human, animal and plant diseases. An EST (Expressed Sequence Tag) is a part of an expressed mRNA sequence. The mRNA is copied to cDNA and part of the cDNA is sequenced from one end or both ends. An EST is the gene fragment known to be expressed in a cell population. ESTs can provide a profile of the mRNA population and are a quick method for cloning a large number of genes.

Genomics has broad effects on all fields in the life sciences. Several large-scale genomics projects are well underway world-wide to determine the genetic make-up of different model organisms. However, such studies remain largely focused on higher animals, fungi and higher land plants. This bias provides only limited insights into eukaryotic diversity and origins. Genomic investigation of protists has only recently begun (Gray 2001).

The Protist EST program (PEP) aims to explore the diversity of eukaryotic genomes, in a systematic and comprehensive way. PEP has opted for an EST approach rather than whole genome sequencing as the most efficient and cost-effective strategy for generating the desired sequence database (http://megasun.bch.umontreal.ca/pepdb/pepdb_description.html). The PEP consortium includes members with knowledge of protistan diversity or the specific protistan groups, and members who are world leaders in comparative genomics. The principal investigator of PEP is Michael W. Gray (Dalhousie University). The co-investigators are Gertraud Burger (University of Montreal), G. Brian Golding (McMaster University), Dion G. Durnford (University of New Brunswick), Patrick J. Keeling (University of British Columbia), Ronald Pearlman (York University) and many others (Gray 2001).

A main element of the PEP initiative is to establish a database, the Protist EST Database (PEPdb), which will be a taxonomically broad eukaryotic database that organizes, compiles, integrates and stores cDNA sequences and related data from about 20 protistan species. The cDNAs and ESTs are being sequenced by the Canadian Protist EST program (PEP). In summary, the Protist EST database (PEPdb) is the centerpiece of the PEP collaboration. It will be accessible by the members of the PEP and other scientific users.

1.1.1 The function of PEPdb

PEPdb has three major functions: management of the data generated by PEP, analysis of these data, and to allow collected sequence information to be accessed via the Internet by PEP members or other users.

Data Management. Archiving, organization and dissemination of data is a main purpose for the PEPdb. A huge amount of sequence information will be generated by the PEP, estimated at larger than 50 Mbp coding DNA sequence in the first year, and increasing to 0.5-1 Gbp from about 30 different organisms by the fifth year (http://megasun.bch.umontreal.ca/pepdb/pepdb_description.html).

The EST information stored in the PEPdb not only contains bare DNA sequence but also includes library, clone information, submitter and publication information. Both trimmed sequences and raw sequences which include polyA or vector sequence are included in the database.

Data Analysis. Because most of the information in the EST databases is redundant and not of very high quality, reducing and removing the redundancy fragments is very necessary. Raw sequence reads are processed and good quality information is stored in the database. These valid data are trimmed of vector sequences and polyA tails. The primary annotation for sequence information can be stored in PEPdb.

An EST CLUSTER is a set of sequences which overlap or are copies of each other. This is determined by examining their sequence similarity. By finding all-against-all sequence similarities, a cluster is formed. The sequences of each individual cluster are then assembled into separate CONTIGs. Since the assembly procedure is quite strict and requires strict overlaps, the definition of individual contigs can change as new sequences are added. The cluster and consensus sequence information are stored in the database.

Access the database. The PEPdb with the web-interface will be publicly accessible via the Internet for queries, viewing and downloading of data. Text queries can be used to search the PEPdb. Querying can be done by searching several fields such as Keyword, Organism, Submitter, Library information or specific searching by ESTID and by citation information.

Batch searches can also be done using a list of ESTIDs. The PEPdb can be searched for sequence similarity using BLAST. Information on individual ESTID includes: identifiers, clone information, primers, sequence, comments, library expression, submitter, citation. Sequences can be displayed in fasta format. Information can be displayed by selecting specific records or by displaying all records. The organism record is linked to the NCBI taxonomy database.

PEPdb provides primarily cluster information. Information on individual clusters include: library information, consensus sequence, and the EST members in the cluster. Cluster information can be searched by keyword, organism or by a specific clusterID number. The searching can also be done by sequence similarity using BLAST.

Members can submit data to the database after registering. The submission process for EST sequence data involves four file input formats: SubmitID, Library, EST, and Publication. The submitter can also view a history of their submitted records.

1.1.2 Database implementation

The PEPdb system consists of a database back-end system and a front-end graphical user interface.

The database back-end system, information storage, was built as a relational database structure composed of Submitter, Publication, EST sequence, EST library and EST cluster information and the relationships among them. It was developed for archiving, storing and managing all information within the PEP. A relational database organizes these data into tables. Each table has a name, columns, and rows containing data for the records. A relational database provides retrieval operations that can generate new tables from existing ones and there are relationships among the tables. As a result, the entire database is in the form of tables (Yarger, Reese and King 1999). PEPdb is implemented in a relational database management system (RDBMS) MySQL. MySQL is not itself a database. It is software that enables a user to create, maintain, manage, access and retrieve electronic databases and acts as a broker between the physical database and the users of that database. MySQL is a lightweight, robust, fast and open source relational database management system (Welling and Thomson 2001).

The front-end graphical user interface is the web-interface that is publicly accessible via the Internet for queries and viewing of data. We have developed a custom Web interface with the PHP4 programming tool. PHP is a server side scripting language designed specifically for the web. The current major version of PHP is 4. PHP has many strengths like high performance and interfaces to many different database systems (Welling and Thomson 2001). A secure Web-based interface was also implemented, to enable individual members of the PEP consortium access to the PEPdb data that that member contributed. This requires user authentication via passwords.

Apache is the most popular server on the Internet. It provides a robust, commercial-grade and freely-available implementation of the HTTP protocol. We chose Apache as the web server, MySQL as the database management system and PHP as the programming tool. Our operating system is Linux. All of them are easy to configure. The system is functional, modifiable and extensible, and provides the detailed information that the user needs. The data is imported into the database using a number of perl scripts which can create suitable formats of sequence, library, submitter and publication information. The raw information is obtained from the submitter or NCBI.

1.2 System Structure

1.2.1 Environment

Several factors should be taken into consideration in the project design stage. The system should be easy to access and hosted on a convenient and popular platform, and this system must provide a user-friendly graphic interface, such as Linux. In the project development, a budget should be carefully planned, including the hardware components and the software packages.

We used Open Source Software (OSS) as the core of the application for reasons of flexibility and economy. Generically, open source refers to a program in which the source code is available to the general public for use and/or modification from its original design free of charge. Open source code is typically created as a collaborative effort in which programmers improve upon the code and share the changes within the community (http://www.webopedia.com).

Open Source softwares are used as following:

1) Workstation: Linux Mandrake 8.2.

2) A database management system: MySQL Distrib 3.23.47.

3) A programming language: PHP 4.1.2.

4) A Web Server Platform: Apache 2.0.

Hardware are used:

1) Process: Compaq 1.8 GHz.

2) Hard drive: 40 GB ATA/100.

3) RAM: 512 MB RAM server.

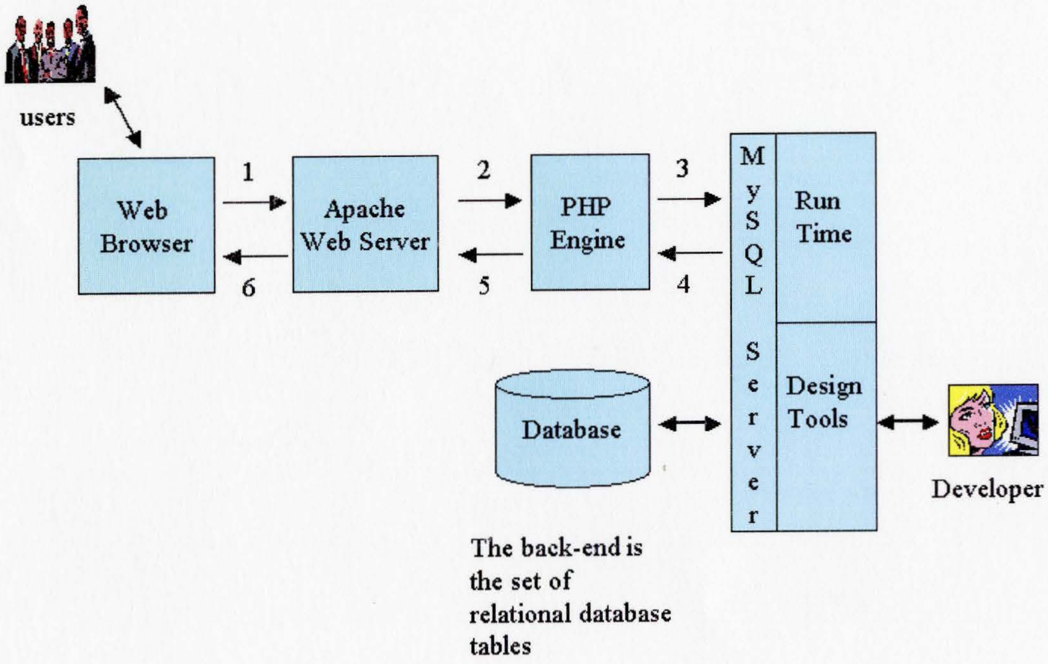
MySQL is the most popular Open Source relational database management system in the world. The MySQL software is a very fast, multi-user, multi-threaded, and robust SQL (Structured Query Language) database server. It ensures that multiple users can work with it concurrently, provide fast access to it, and ensures that only authorized users can obtain access to the data (Welling and Thomson 2001). The MySQL Server is intended for mission-critical, heavy-load production systems. Under constant development, the MySQL Server offers a rich and useful set of functions, and its connectivity, speed, and security make MySQL Server highly suited for accessing databases on the Internet (http://www.mysql.com).

PHP Hypertext Preprocessor is a server-side scripting language, which embedded with HTML. Using PHP, one can create dynamic web sites which display changing content, depending on many different factors. The syntax of the PHP script is very similar to that of the Perl or C language and the script is enclosed within special PHP tags (⟨? and ?⟩), so the programmer can mix the code between HTML and PHP. PHP script is first interpreted and executed on the web server and then the web page is sent to the users browser (Welling and Thomson 2001). That is why the user cannot view the script itself and the user's browser only receives the output of the page after the script was executed. PHP has some strengths in comparison to other products. PHP has efficient performance, and can directly connect to many different database systems including MySQL, PostgreSQL, mSQL, Oracle, Sybase etc, or connect to any database using the Open Database Connectivity Standard (ODBC) driver. PHP has many built-in libraries for performing many Web tasks including send email, work with cookies and sections, and generate PDF documents. PHP is portable and available for many operating system such as Linux, Unix, and Microsoft Windows without modification of the source code (Welling and Thomson 2001). We used a current version of PHP 4.1.2 which is much faster than previous versions.

Apache has been the most popular web server on the Internet since 1996. In 2002, Netcraft Web Server Survey found that 63% of the web sites on the Internet are using Apache, thus making it more widely used than all other web servers combined (http://www.apache.org/). Apache was originally written for UNIX. However there are versions that run under different platforms such as Windows. The Apache HTTP Server provides a secure, efficient and extensible function for the users.

Mandrake Linux was developed in 1998 and is a UNIX operating system. A typical Mandrake Linux system can run for months without a reboot and requires little maintenance since its software management system automatically handles dependency errors and avoids conflicts between applications. The system has stable, reliable, functional and easy-to-use environments. It provides a perfect integration of user-friendly graphical environments such as KDE and GNOME, and is easily upgradable, with excellent networking facilities and true multi-tasking (http://www.linux-mandrake.com/).

Figure 1.1: The general Web database structure and transaction


users
Web Browser
1
6
Apache Web Server
2
5
PHP Engine
3
4
MySQL Server
Run Time
Design Tools
Database
Developer
The back-end is the set of relational database tables

1.2.2 Database Architecture

The Figure 1.1 shows the general web database structure. Web database transactions consists of several stages:

1) The user issues an HTTP request for a particular Web page (eg. http://info.biology. mcmaster.ca/ling/estHome.html) through the Web browser.

2) The Apache Web server processes the HTTP protocol, receives the request for the Web browser, retrieves the file and passes it to the PHP engine for processing.

3) The PHP engine begins parsing the script. If there is a command for connection to the database and to execute a query in the PHP script, PHP opens a link to the MySQL server and a set of SQL Data Manipulation Language statements is sent to the MySQL Database Engine.

4) The MySQL server receives the query from the PHP engine and retrieves or manipulates data by executing the SQL statements, and sends the query results back to the PHP engine.

5) The PHP engine will format the query results nicely in HTML and finish running script. It then will return the formatted HTML result back to the Apache Web server.

6) The Apache Web server will pass the resulting HTML back to the Web browser and the user can see the list of query results.

The back-end is the set of relational database tables. The database DBMS serves as middle tier to connect the back-end database system and the front-end. Both developers and users can access the DBMS either directly or indirectly via application programs and the database is processed by the DBMS. We access the MySQL server directly to develop the back-end of the database.

1.3 Database Design

The task of developing this web database system was divided into two major parts. One was the database design and implementation using MySQL DBMS, the other was the graphical user interface design and implementation using the PHP language.

At the early stage of the development process, we used a top-down development strategy which proceeds from the general to the specific. The descriptions and models are worked downward toward more and more detail. By interviewing the users, studying some public databases, and analyzing the statement of the requirements, a use case diagram was built. A use case diagram provides a graphic description of who will use a system and what kinds of interactions to expect within that system (http://www.embarcadero.com/). From these requirements, we designed an ER model (The Entity Relationship diagram) so as to represent the structure of a database logically. While translating the data model into a relational database, all entities were normalized until a suitable model was developed. The model was refined until it contained well-structured relations which avoided storing redundant data, kept unique keys with few empty attributes in each table, and linked the tables with one-to-one or one-to-many relationships.

The major stages of this database design involved three phases:

1) A Data flow phase: We made a use case diagram which gathered all user requirements and represented the scenarios or actions.

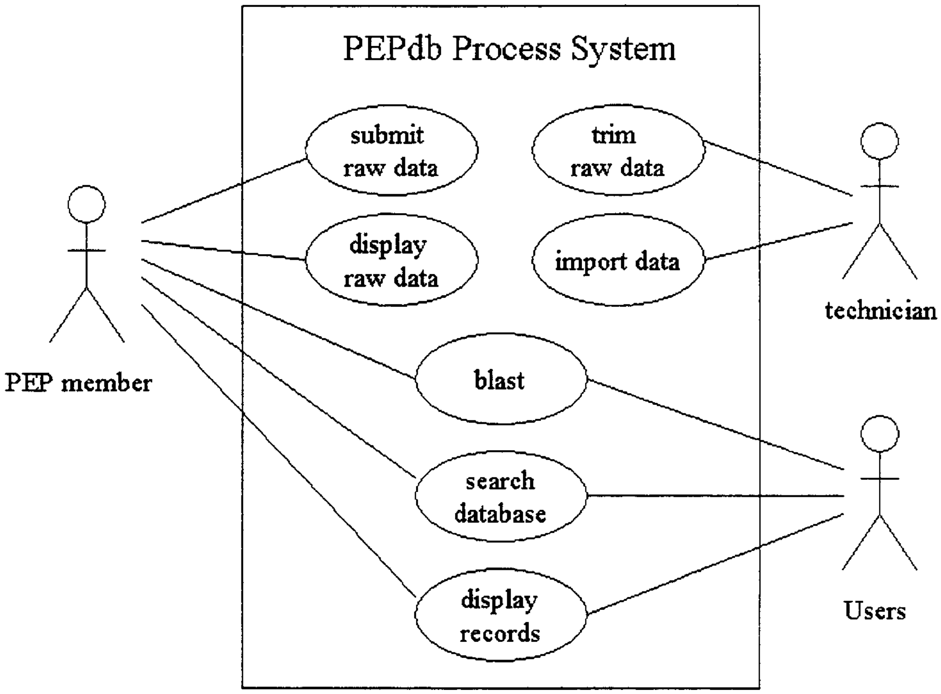
2) A Data modeling phase: We created an Entity Relationship (ER) diagram by collecting the information from the users, the public databases and the data flow diagram.

3) A Normalization phase: We normalized these relations by applying Normal Form rules to avoid most of the problems which may cause poorly structured relations.

1.3.1 Use case phase

In order to build an effective database and related applications, we must thoroughly understand the users' model of their activities. Since this project needs to work for multi-user work groups and organizational databases, the data modeling process is complicated. Many users envision many different logical data models such that the users may be using the same condition for different things or different conditions for the same things. The greater challenge is that no single user has a model of the complete structure and each user only understands some of the workgroup's data model. It is necessary to document the logical union of the different divisions in the data model. The Figure 1.2 shows the Use case diagram.

Figure 1.2: Use case diagram

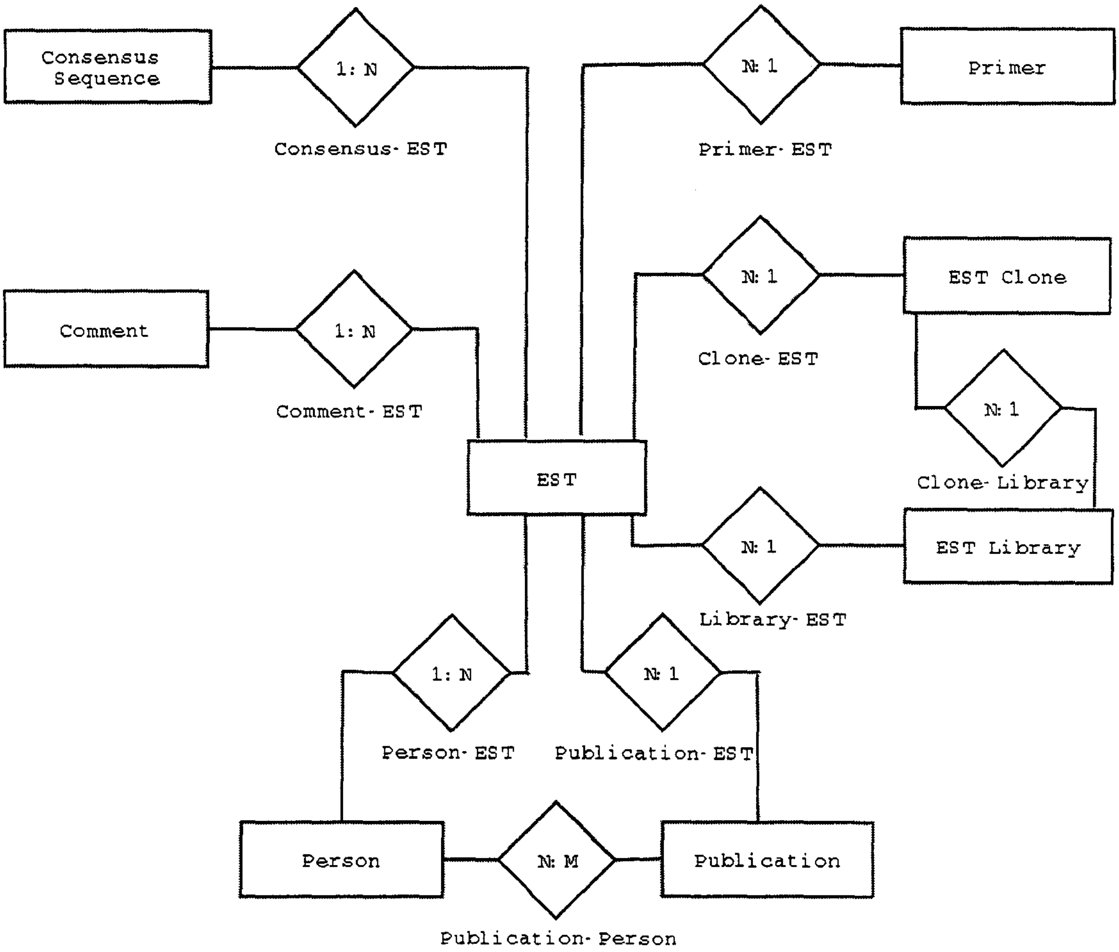

PEPdb Process System
submit raw data
trim raw data
display raw data
import data
blast
search database
display records
PEP member
technician
Users

1.3.2 Entity-Relationship Model

Data modeling is the most important task in the development of effective databases. These models provide a language to express the structure of the data and relationships in the users' environment. With this model, each entity is something that the users want to track and can be identified in the users' work environment. Entities have attributes that describe their characteristics, one or more attributes identify an entity, and relationships describe associations among the entities (Kroenke 2002). All entities of the E-R diagram are interrelated. There are three types of binary relationships, one-to-one relationship, one-to-many relationship and many-to-many relationship. A one-to-one relationship exists between a pair of tables if a single record in the first table is related to only one record in the second table, and a single record in the second table is related to only one record in the first table. A one-to-many relationship exists between a pair of tables if a single record in the first table can be related to one or more records in the second table, but a single record in the second table can be related to only one record in the first table. A many-to-many relationship exists between a pair of tables if a single record in the first table can be related to one or more records in the second table, and a single record in the second table can be related to one or more records in the first table (Hernandez 1997). The entities, the relationships between the entities, the attributes in each entity, and unique identifiers for each entity were developed for the PEPdb. There are eight categories in this ER model that describe a real-world object: *Consensus Sequence, EST, Primer, EST Clone, EST library, Publication, Person* and *Comment*. Figure 1.3 shows the entity-relationship diagram.

Figure 1.3: The Entity-Relationship diagram

The rectangles represent entities, the diamonds show the relationships, and the maximum cardinality of the relationship is shown inside the diamond. The name of the entity is shown inside the rectangle, and the name of relationship is shown near the diamond.


Consensus Sequence
1: N
Consensus- EST
Primer
N: 1
Primer- EST
Comment
1: N
Comment- EST
EST Clone
N: 1
Clone- EST
N: 1
Clone- Library
EST
EST Library
N: 1
Library- EST
1: N
Person- EST
N: 1
Publication- EST
Person
N: M
Publication- Person
Publication

Entities description.

1) *Consensus Sequence*: This entity represents the EST cluster. A *Consensus Sequence* is a set of ESTs which overlap or contain each other according to their similarity.

2) *EST*: The *EST* indicates the small pieces of cDNA. This entity includes the information about sequence content, base count, sequence length, submission date and definition.

3) *Primer*: The *Primer* describes the sequence of primer used to sequence the EST. This entity contains the sequences (5′ to 3′) used for the forward and backward PCR primers.

4) *EST Clone*: The *EST Clone* entity provides the cDNA clone information, such as clone source.

5) *EST Library*: The *EST library* entity describes the cDNA library, including the vector, restriction enzyme site, type of cell and tissue.

6) *Publication*: The *Publication* entity implements the research articles associated with the submitted ESTs. This entity contains the information of title, authors, journal, year and Medline unique identifier which can link to PubMed.

7) *Person*: The *Person* entity gives the contact information about members of the PEP project. This entity has person name, email, address etc.

8) *Comment*: The *Comment* entity stores any comments for the ESTs. The comments are written by the submitters.

Relationships and constraints of the entities.

1) *Consensus-EST*: *Consensus Sequence*s are formed by *EST*s. One *Consensus sequence* must have at least two *EST*s. An *EST* can be assigned to one *Consensus Sequence.* But not all *EST*s are assigned to a *Consensus Sequence.*

2) *Primer-EST*: *Primer*s are used for sequencing *EST*s. A *Primer* sequence can be used for from one to many *EST*s. An *EST* must be sequenced using only one *Primer* pair.

3) *Clone-EST*: *EST Clone*s contain *EST*s. An *EST Clone* stores from one to many *EST*s. An *EST* must come from only one *EST clone.*

4) *Library-EST*: *EST Librarie*s have *EST*s. Every *EST Library* has from one to many *EST*s. An *EST* is generated from only one *Library.*

5) *Publication-EST*: *Publication*s store references to *EST*s. One *Publication* refers to one or many *EST*s. An *EST* is referenced in one *Publication.* Not all *EST*s may be referenced by the *Publication.*

6) *Person-EST*: *Person*s provide *EST*s. One *Person* provides from one to many *EST*s. An *EST* must be provided by one *Person.*

7) *Clone-Library*: *EST Librarie*s include *EST Clone*s. One *EST Library* has many *EST Clone*s. One *EST Clone* must come from one *EST Library.*

8) *Comment-EST*: *Comment* stores comments on the *EST*s. One *Comment* can apply to many *EST*s. One *EST* can have zero or a *Comment.*

9) *Publication-Person*: *Person*s have *Publication*s. One *Person* has zero to many *Publication*s. One *Publication* record can reference one to many *Person*s records.

1.3.3 Translating ER diagram into a relational database and refining the data structure

After creating an E-R model which included eight entities, we refined the data structure, translated the ER diagram into a relational database, and then checked relations against the theoretical rules.

A relation is a two-dimensional table that has single-value entries. All entries in a given column are of the same kind; columns have a unique name; Columns are also called attributes (Kroenke 2002). If we try to put too many attributes into a single relation, anomalies will happen. There are several kinds of anomalies, such as redundancy where records may be repeated in several rows, and deletion anomalies where deleting a sets of values will lose other values in the row. Anomalies can be eliminated by splitting one relation into two or more relations which is called normalized. According to the rules of normalization, we broke up a relation and created referential integrity constraints. Every normalized relation has a single theme.

There are some entities and relationships among those entities in the ER model. For initial design, we defined a relation for each entity. The attributes of a relation are the attributes of the entity. The steps were as follows:

1) We created one table for each entity which has several attributes, and translated each attribute in the entity into a column in the corresponding table.

2) For each pair of entities related in one-to-one relationships, we added the primary key of one entity to its related entity as a foreign key.

3) For each pair of entities related in one-to-many relationships, we added the primary key of the parent entity (one side) to the child entity (many side) as a foreign key.

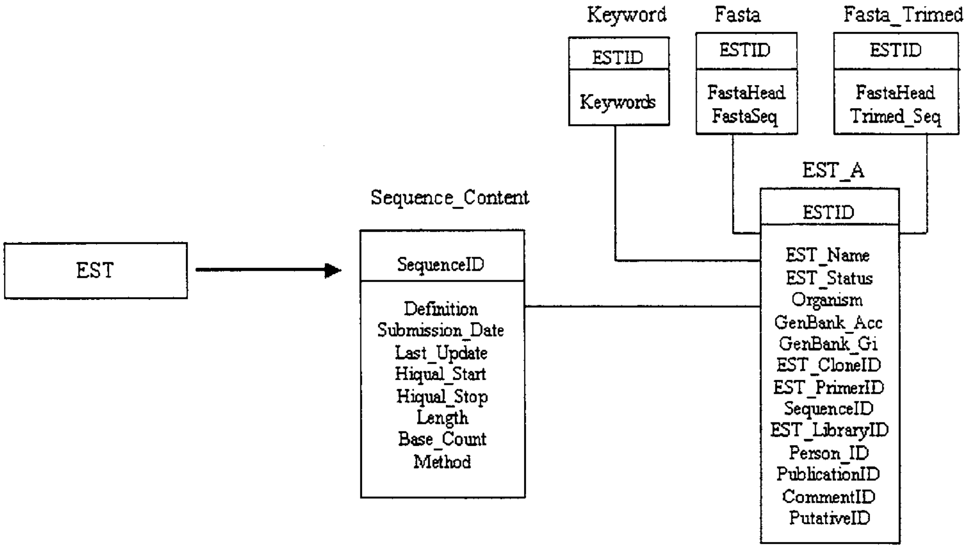
From the ER diagram, the main entity is the *EST* entity which includes EST sequence and definition. The other entities related to this entity are one-to-many relationships. The information in the *EST* entity will be updated and users will retrieve the information in this entity frequently. We split this entity into two entities: one is *Sequence_Content* which contains EST information, the other is *EST_A* which contains the foreign keys for linking to other entities. Since the *Sequence_Content* entity is separated from other entities,. it is easy to update and modify the data in this table and not adversely affect the values of the other fields or tables in the database.

The *Sequence_Content* entity contains EST information. It meets the rule that every normalized relation has a single theme. However, performance is very important. The sequence of EST is long text and users search the information according to the sequence similarity frequently. In these circumstances, it is appropriate to store the sequence in a separate table to decrease the time required to retrieve data. We created two tables to store the sequences, one for raw sequences which are not of high quality and submitted by the members, the other for trimmed sequences which do not contain vector sequences and polyA tails.

Most common queries can be done by searching keywords. We created a *Keyword* entity and separated it from the *Sequence_Content* entity so that searching for keywords can be processed faster. The *Keyword* entity includes the EST name, species name, definition, and publication information.

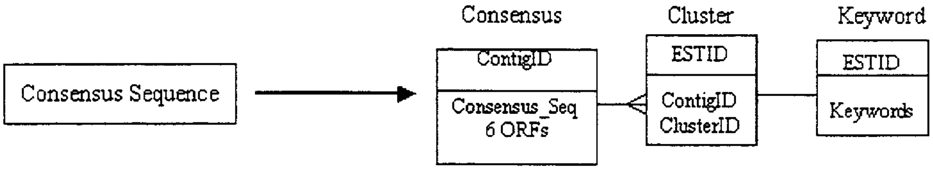
These entities to *EST_A* entity represent one-to-one relationships. In this case, the entities have the same primary key ESTID. The primary key ESTID in the *EST_A* entity can be the foreign key to link with other entities. Figure 1.4 shows these entities and relationships.

Figure 1.4: Refining the *EST* entity


Keyword
ESTID
Keywords
Fasta
ESTID
FastaHead
FastaSeq
Fasta_Trimed
ESTID
FastaHead
Trimed_Seq
EST_A
ESTID
EST_Name
EST_Status
Organism
GenBank_Acc
GenBank_Gi
EST_CloneID
EST_PrimerID
SequenceID
EST_LibraryID
Person_ID
PublicationID
CommentID
PutativeID
Sequence_Content
SequenceID
Definition
Submission_Date
Last_Update
Hiqual_Start
Hiqual_Stop
Length
Base_Count
Method
EST

A cluster sequence is formed by determining all-against-all sequence similarities. Because the sequences from PEP are submitted frequently, the number of sequences will keep expanding. The clusters and contigs will also change. In a well-defined structure, changes made to a specific value need only be made in one location (Hernandez 1997). We created a new entity of *Cluster* which stores an identity ID for each cluster. It is easy to modify the data in both *Consensus* and *Cluster* tables without affecting the other values in the database. *Consensus* entity to *Cluster* entity is a one-to-many relationship, since a cluster might be separated into several contigs during the assembly procedure. ContigID as a foreign key links these two entities. The *Cluster* entity is linked to the *Keyword* entity so that the user can access the consensus sequence by searching Keywords. The *Cluster* entity and the *Keyword* entity are a one-to-one relationship. Both of them have the same primary key and the primary key can also be the foreign key for linking each other. The figure 1.5 shows these entities and relationships.

Figure 1.5: Refining the *Consensus Sequence* entity

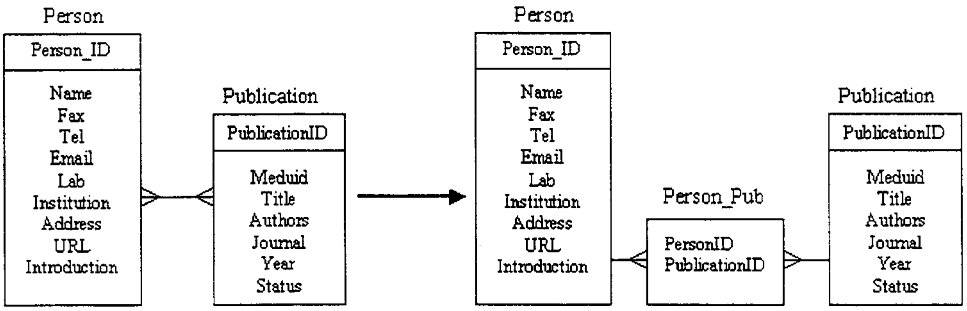

Consensus Sequence
Consensus
ContigID
Consensus_Seq
6 ORFs
Cluster
ESTID
ContigID
ClusterID
Keyword
ESTID
Keywords

The *Publication* entity and the *Person* entity is a many-to-many relationship, in which a Person can correspond to many Publications, and a Publication can correspond to many Persons. If a many-to-many relationship is directly represented by relations, a large amount of redundant data in the tables will be generated. An intersection relation was created for storing the intersection of a particular Person with a particular Publication in each row to avoid redundancy. The attributes in the intersection relation are composed of the primary keys in both *Person* and *Publication* entities. The *Person* entity and the *Publication* entity to the intersection relation are a one-to-many relationships. The figure 1.6 shows the entities and relationships.

Figure 1.6: Refining the *Publication* and *Person* entities

Referential integrity constraints:

PersonID in *Per_Pub* must exist as PersonID in *Person.* PublicationID in *Per_Pub* must exist as PublicationID in *Publication.*


Person
Person_ID
Name
Fax
Tel
Email
Lab
Institution
Address
URL
Introduction
Publication
PublicationID
Meduid
Title
Authors
Journal
Year
Status
Person
Person_ID
Name
Fax
Tel
Email
Lab
Institution
Address
URL
Introduction
Person_Pub
PersonID
PublicationID
Publication
PublicationID
Meduid
Title
Authors
Journal
Year
Status

In the *EST_Library* relation, the key is designated as EST_LibraryID, and functional dependencies are:

EST_LibraryID ⟶ Library_Name, Organism, Strain, Tissue_type, Cell_Type,
Cell_Line, Stage, Library_Host, Description, Obtained_Date,
Vector_Name

Vector_Name ⟶ Vector_Type, Rsite1, Rsite2

The attributes on the left side of the arrow are called determinants (Kroenke 2002). Since "EST_LIbraryID determines Vector_Name" and "Vector_Name determines Vector_Type, Rsite1 and Rsite2", the functional dependencies are:

EST_LibraryID ⟶ Vector_Type, Rsite1, Rsite2

An arrangement of functional dependencies like this is called a transitive dependency. The *EST_Library* relation has anomalies because of the transitive dependency. The *EST_Library* relation was therefore divided into two relations:

EST_Library (EST_LibraryID, Library_Name, Organism, Strain, Tissue_type,
Cell_Type, Cell_Line, Stage, Library_Host, Description, Obtained_Date,
Vector_Name)

Vector (Vector_Name, Vector_Type, Rsite1, Rsite2)

These two relations can be linked by the key Vector_Name. The figure 1.7 shows the relations and relationship. The figure 1.8 shows the entire PEPdb schema.

Figure 1.7: Refining the *EST Library* entities

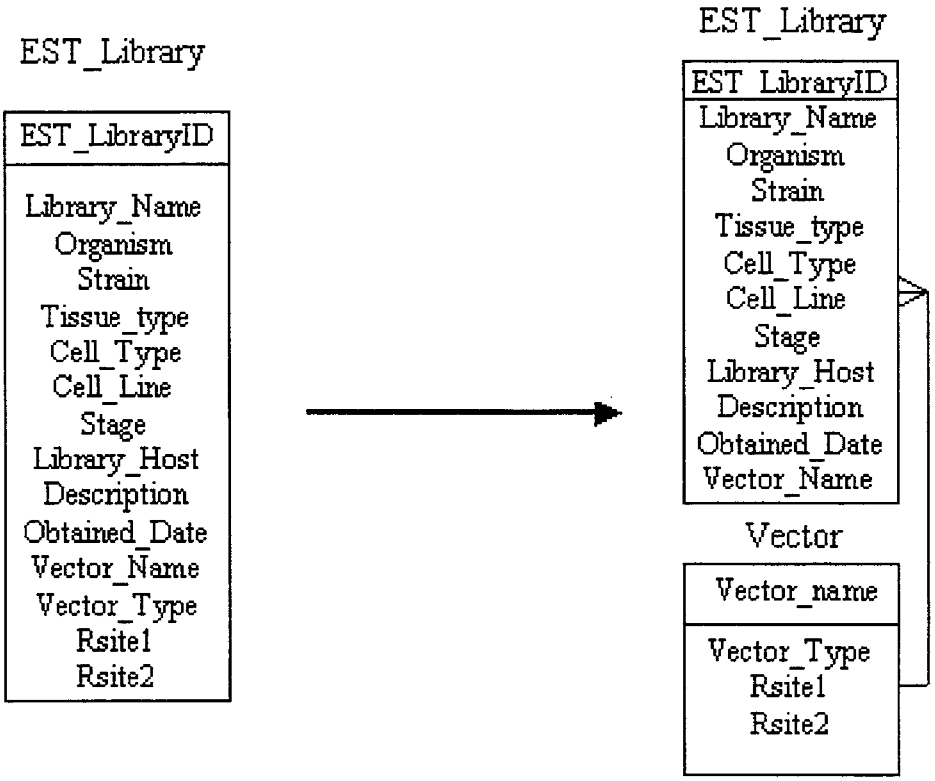
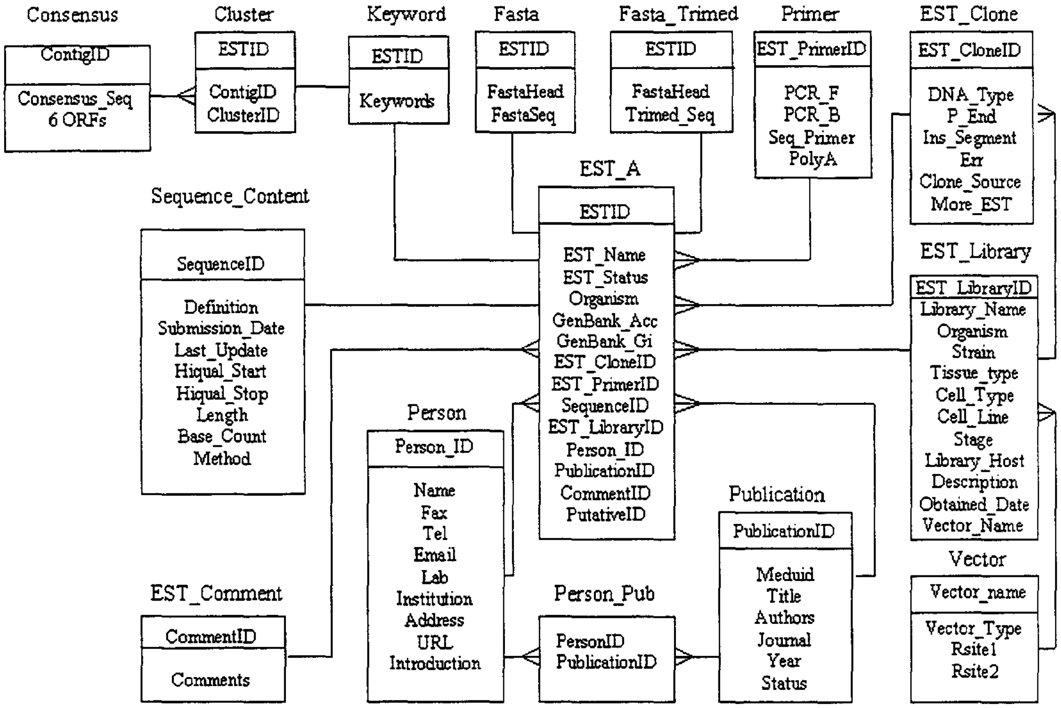

EST_Library
EST_LibraryID
Library_Name
Organism
Strain
Tissue_type
Cell_Type
Cell_Line
Stage
Library_Host
Description
Obtained_Date
Vector_Name
Vector_Type
Rsite1
Rsite2
EST_Library
EST_LibraryID
Library_Name
Organism
Strain
Tissue_type
Cell_Type
Cell_Line
Stage
Library_Host
Description
Obtained_Date
Vector_Name
Vector
Vector_name
Vector_Type
Rsite1
Rsite2

Figure 1.8: The Protist EST database (PEPdb) Schema

Each entity is represented by a relation. There are fifteen tables in the PEPdb schema. Eleven relations involved in one-to-many relationships. The *EST_A* table is a "many" side relation which stores the foreign keys for linking with other tables.


Consensus
ContigID
Consensus_Seq
6 ORFs
Cluster
ESTID
ContigID
ClusterID
Keyword
ESTID
Keywords
Fasta
ESTID
FastaHead
FastaSeq
Fasta_Trimed
ESTID
FastaHead
Trimed_Seq
Primer
EST_PrimerID
PCR_F
PCR_B
Seq_Primer
PolyA
EST_Clone
EST_CloneID
DNA_Type
P_End
Ins_Segment
Err
Clone_Source
More_EST
Sequence_Content
SequenceID
Definition
Submission_Date
Last_Update
Hiqual_Start
Hiqual_Stop
Length
Base_Count
Method
EST_A
ESTID
EST_Name
EST_Status
Organism
GenBank_Acc
GenBank_Gi
EST_CloneID
EST_PrimerID
SequenceID
EST_LibraryID
Person_ID
PublicationID
CommentID
PutativeID
EST_Library
EST_LibraryID
Library_Name
Organism
Strain
Tissue_type
Cell_Type
Cell_Line
Stage
Library_Host
Description
Obtained_Date
Vector_Name
Person
Person_ID
Name
Fax
Tel
Email
Lab
Institution
Address
URL
Introduction
Publication
PublicationID
Meduid
Title
Authors
Journal
Year
Status
Vector
Vector_name
Vector_Type
Rsite1
Rsite2
EST_Comment
CommentID
Comments
Person_Pub
PersonID
PublicationID

1.3.4 Normalization

All relations were checked against the normalization criteria. Normalization can be used as a guideline for checking the desirability and correctness of relations. Relations can be classified by the types of anomaly that they eliminate. Such classifications are called normal forms (Kroenke 2002). Theoretically, the higher the Normal Form the better design of the relation.

First Normal Form: The data meets the definition of a relation. A relation is not allowed repeating groups, each column has a unique name, and all of the entries in any column must be the same kind.

Second Normal Form: A relation is in First Normal Form and all its nonkey attributes are dependent on all of the key (Kroenke 2002).

Third Normal Form: A relation is in Second Normal Form and has no transitive dependencies (Kroenke 2002).

Boyce-Codd normal form (BCNF): A relation is in BCNF if every determinant is a candidate key (Kroenke 2002).

In the most cases, if we place the tables in third Normal Form, a poor relational design will be avoided. However, sometimes even relations in third normal form still have anomalies. Relations in BCNF may have no anomalies. The resulting schemas should be checked in BCNF.

The functional dependencies for the results of the decomposition are as follows:
The functional dependency in *EST_Clone* relation is:

EST_CloneID ⟶ *DNA_Type, P_End, Ins_Segment, Err, Clone_Source, More_EST*

The functional dependency in *Primer* relation is:

EST_PrimerID ⟶ PCR_F, PCR_B, Seq_Primer, PolyA

The functional dependency in *Publication* relation is:

PublicationID ⟶ Meduid, Title, Authors, Journal, Year, Status

The functional dependency in *Person* relation is:

Person_ID ⟶ Name, Fax, Tel, Email, Lab, Institution, Address, URL, Introduction

The functional dependency in *Sequence_Content* relation is:

SequenceID ⟶ Definition, Submission_Date, Last_Update, Hiqual_Start,
Hiqual_Stop, Method, Length, Base_Count

The functional dependency in *Fasta* relation is:

ESTID ⟶ FastaHead, FastaSeq

The functional dependency in *Fasta_Trimed* relation is:

ESTID ⟶ FastaHead, Trimed_Seq

The functional dependency in *EST_Comment* relation is:

CommentID ⟶ Comments

The functional dependency in *Keyword* relation is:

ESTID ⟶ Keywords

The functional dependency in *Consensus* relation is:

ContigID ⟶ Consensus_Seq, 6ORFs

The functional dependency in *EST_A* relation is:

ESTID ⟶ *EST_Name, EST_Status, Organism, GenBank_Acc, GenBank_Gi,*
EST_CloneID, EST_PrimerID, SequenceID, EST_LibraryID,
Person_ID, PublicationID, CommentID, PutativeID

The functional dependency in *Cluster* relation is:

ESTID ⟶ *ContigID, ClusterID*

The functional dependency in *EST_Library* relation is:

EST_LibraryID ⟶ *Library_Name, Organism, Strain, Tissue_type, Cell_Type,*
Cell_Line, Stage, Library_Host, Description, Obtained_Date,
Vector_Name

The functional dependency in *Vector* relation is:

Vector_Name ⟶ *Vector_Type, Rsite1, Rsite2*

By analyzing the functional dependencies on all relations, there are no relations that contain repeating groups and hence they are in 1NF. Since all the nonkey attributes in each relation are functionally dependent on a primary key, all relations are in 2NF. Each relation contains only one determinant, so there are no transitive dependencies. All relations are in 3NF. Because each relation has only one determinant which is a primary key, all relations are in BCNF.

1.3.5 Data Integrity

Data integrity refers to the validity, consistency, and accuracy of the data in a database (Hernandez 1997). It is one of the most important aspects of the database design process, and we can not ignore it. Mistakes in data integrity would result in a high risk of undetected errors. Based on inaccurate information, the users would make poor strategic decisions.

Table-level integrity ensures that the field that identifies each record within the table, is unique and is never missing its value (Hernandez 1997). In this case, the primary keys in the tables cannot contain null data. We created constraints for all primary keys to ensure that they could not be the null value.

Field-level integrity ensures that the structure of every field is sound, that the values in each field are valid, consistent, and accurate, and that fields of the same type are consistently defined throughout the database (Hernandez 1997). The fields and values in each of the tables were checked by this rule.

Relationship-level integrity (the referential integrity rule) ensures that the relationship between a pair of tables is sound and that there is synchronization between the two tables whenever data is entered, updated, or deleted (Hernandez 1997). The database design must not contain any unmatched foreign key values. One constraint cascade for each foreign key should be created. For updates or deletions, the change is cascaded to all dependent tables. Since MySQL DBMS does not have the function of the foreign key, we were careful in the design of the tables to link by the correct foreign keys so that the design will satisfy the referential integrity rule.

1.4 Graphic User Interface

1.4.1 Graphic User Interface Design

Users can enter input, issue commands, request specific actions, and receive responses through the Graphic User Interface. The interface was designed and implemented as serial pages. All information was categorized into serial pages which were interrelated to each other. The users can navigate the information from record to record and the pages are linked dynamically.

The creation of an effective Graphical User Interface (GUI) involves building a technical framework, including the implementation of text box, check box, drop-down list and creation and placement of buttons. The actual front end (graphic user interface) is separated into two approaches: decision support applications and transaction-processing applications. Decision support applications enable users to view and query information in the database such as "Search" and "Go" buttons, but do not allow them to add or modify information. These are the most common types of interface that designers choose for protecting data. The entire text boxes, check boxes and other various GUI representations are locked to prevent modification. Transaction-processing applications include the capability to add data, delete or edit existing data. In this project, we used the "ADD" application to allow PEP members to submit data into the database.

Tables are used to speed page download. The main structure of the web page is a table divided into three parts: top, right, and left. The top part contains the PEPdb name and logo. The left part contains hyperlinks to other sections. The right side contains text, text boxes and a submit button. Text boxes allow users to enter a single line of text for queries. To submit the query to the database, we use a submit button which sends data from the form to the PHP page that will process the data.

Graphics were chosen to be clean, clear, and fast to download. The color and style of the interfaces are consistent. The site is easy to use. Each page provides landmarks and navigation cues to help users and prevent them from getting lost.

1.4.2 General Graphic User Interface Structure

PEPdb contains cDNA sequence, library information, submitter, publication information, and also includes EST cluster information. The GUI structure was designed to manage the interaction of the relevant information. The basic hierarchy for the user interface typically had a main page, and some following pages for different categories of the linked pages.

The Figure 1.9 illustrates the general graphic user interface and relationships between the relevant windows. The web site contains eight sections including Home, Blast, Search Database, Batch Search, Cluster, Submit Data, About PEPdb and Contact Us. Each section is linked to the home page directly and independently. When starting from the main page, the file estHome.html is loaded. This page displays the key information of the PEPdb. Users can choose to view any window by clicking one of the eight linked items on this page. In each section, users can retrieve and view the relevant information from the database. The search pages and the serial result pages are dynamically linked and may interact.

1.4.3 Graphical User Interface

The Figure 1.10 - 1.14 show the graphical user interfaces.

Figure 1.9: General Graphic User Interface Structure

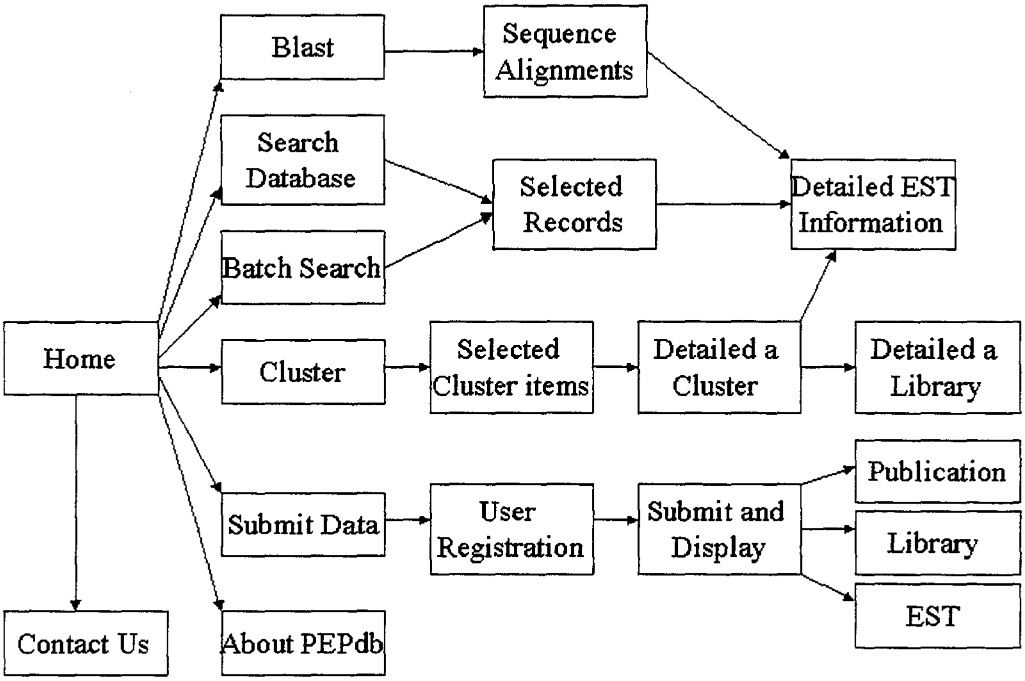
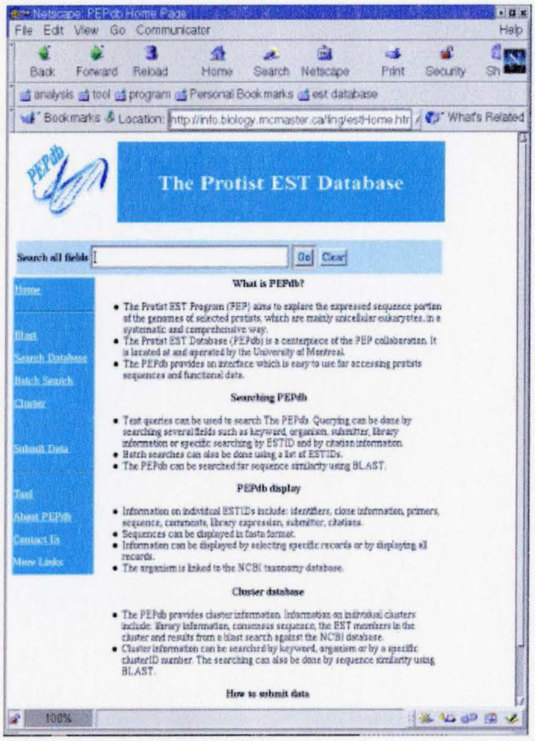

Blast
Sequence Alignments
Search Database
Selected Records
Detailed EST Information
Batch Search
Home
Cluster
Selected Cluster items
Detailed a Cluster
Detailed a Library
Publication
Submit Data
User Registration
Submit and Display
Library
EST
Contact Us
About PEPdb

Figure 1.10: Home page and Advanced search

The text on the Home page gives the main description about the PEPdb, including "What is PEPdb", "Searching PEPdb", "PEPdb display" and "How to Submit Data". The left side of this page links to four pages for querying the database: Blast, Search Database, Batch Search and Cluster. There is a link to Submit Data for PEP members submitting data into the database. Text queries can be used to search PEPdb in the text box of "Search all fields" in this page.

The form element is text box in the Search PEPdb more detailed page. Ten text boxes are listed for text searching, key searching and citation searching. In the "Search for text" area, querying can be done by searching several fields such as "Keyword", "Organism", "Submitter", "Library Name" and "Description" of the library. A user can enter the string in one or more of the text boxes in this area. In the "Search for key" area, specific searching can be done by "ESTID" or "EST Name". A user can enter a single word of the complete EST name and ESTID in either the "EST Name" field or the "ESTID field". In the "Search for citation" area, a user can enter the string in one or more of the text boxes to search for a citation.


Netscape: PEPdb Home Page
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Sh
analysis tool program Personal Bookmarks est database
Bookmarks Location: http://info.biology.mcmaster.ca/ling/estHome.htr
What's Related
PEPdb
100%

The Protist EST Database

Search all fields [ ] Go Clear

Home
Blast
Search Database
Batch Search
Cluster
Submit Data
Tool
About PEPdb
Contact Us
More Links

What is PEPdb?

- The Protist EST Program (PEP) aims to explore the expressed sequence portion of the genomes of selected protists, which are mainly unicellular eukaryotes, in a systematic and comprehensive way.
- The Protist EST Database (PEPdb) is a centerpiece of the PEP collaboration. It is located at and operated by the University of Montreal.
- The PEPdb provides an interface which is easy to use for accessing protists sequences and functional data.

Searching PEPdb

- Text queries can be used to search The PEPdb. Querying can be done by searching several fields such as keyword, organism, submitter, library information or specific searching by ESTID and by citation information.
- Batch searches can also be done using a list of ESTIDs.
- The PEPdb can be searched for sequence similarity using BLAST.

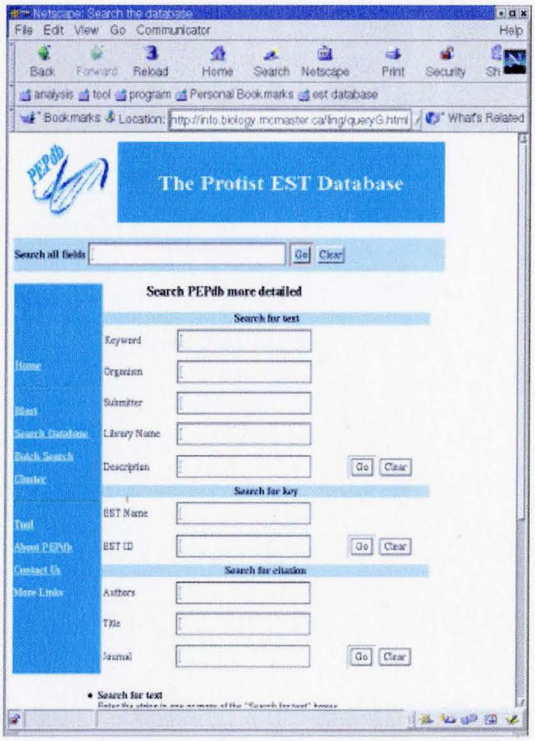
PEPdb display

- Information on individual ESTIDs include: identifiers, clone information, primers, sequence, comments, library expression, submitter, citations.
- Sequences can be displayed in fasta format.
- Information can be displayed by selecting specific records or by displaying all records.
- The organism is linked to the NCBI taxonomy database.

Cluster database

- The PEPdb provides cluster information. Information on individual clusters include: library information, consensus sequence, the EST members in the cluster and results from a blast search against the NCBI database.
- Cluster information can be searched by keyword, organism or by a specific clusterID number. The searching can also be done by sequence similarity using BLAST.

How to submit data


Netscape: Search the database
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Sh
analysis tool program Personal Bookmarks est database
Bookmarks Location: http://info.biology.mcmaster.ca/ling/queryG.html
What's Related
PEPdb

The Protist EST Database

Search all fields [ ] Go Clear

Home
Blast
Search Database
Batch Search
Cluster
Tool
About PEPdb
Contact Us
More Links

Search PEPdb more detailed

Search for text

Keyword [ ]
Organism [ ]
Submitter [ ]
Library Name [ ]
Description [ ] Go Clear

Search for key

EST Name [ ]
EST ID [ ] Go Clear

Search for citation

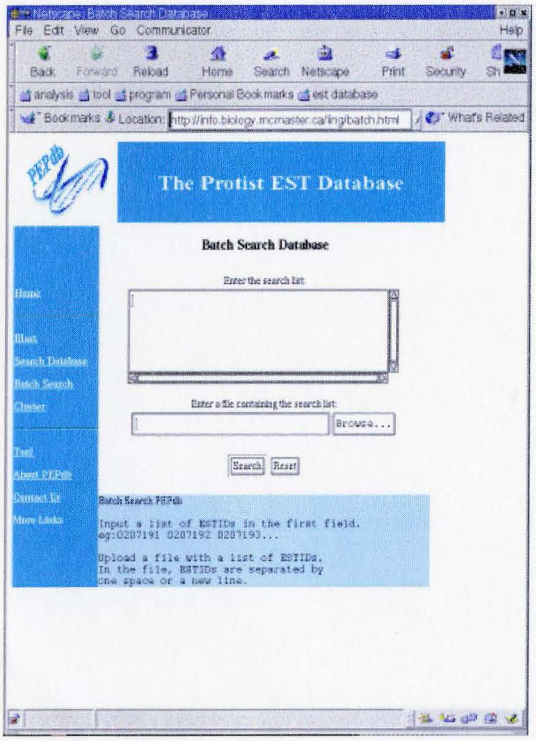
Authors [ ]
Title [ ]
Journal [ ] Go Clear

- **Search for text**
  Enter the string in one or more of the "Search for text" boxes

Figure 1.11: Blast page and Batch search

The Blast page contains two text boxes for user entered data, one for pasting the DNA or protein sequences, the other for to upload a file which contains DNA or protein sequences. Two drop-down lists provide list so that a user can select an available program and database. Three options of the programs include blastn, blastp and blastx. Two options of the blasted database are the PEPdb and the Clustered PEPdb. The output page generates the matched ESTIDs, the definitions of these ESTIDs, and the alignment of the sequences. Each ESTID is linked to a page to display its details.

The textarea element and the text box element as the interface component are in a form of the Batch Search Database page. A user can enter a list of ESTIDs in the "Enter the search list" box for searching these EST records and can also upload a file with a list of ESTIDs in the "Enter a file containing the search list" box for searching EST records in this file.


Netscape: Batch Search Database
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Sh
analysis tool program Personal Bookmarks est database
Bookmarks Location: http://info.biology.mcmaster.ca/ling/batch.html
What's Related
PEPdb
Home
Blast
Search Database
Batch Search
Cluster
Tool
About PEPdb
Contact Us
More Links
Browse...
Search
Reset

The Protist EST Database

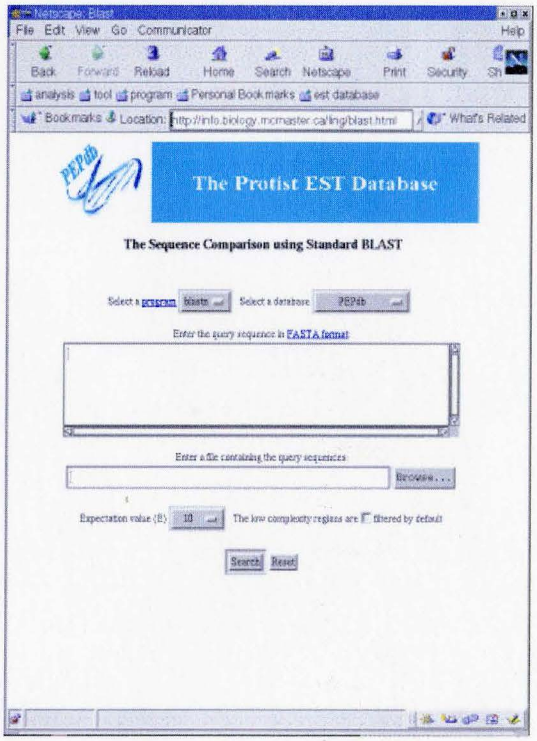
Batch Search Database

Enter the search list:

Enter a file containing the search list:

Batch Search PEPdb

```
Input a list of ESTIDs in the first field.
eg:0207191 0207192 0207193...

Upload a file with a list of ESTIDs.
In the file, ESTIDs are separated by
one space or a new line.
```


Netscape: Blast
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Sh
analysis tool program Personal Bookmarks est database
Bookmarks Location: http://info.biology.mcmaster.ca/ling/blast.html
What's Related
PEPdb
Browse...
Search
Reset

The Protist EST Database

The Sequence Comparison using Standard BLAST

Select a program: blastn Select a database: PEPdb

Enter the query sequence in FASTA format

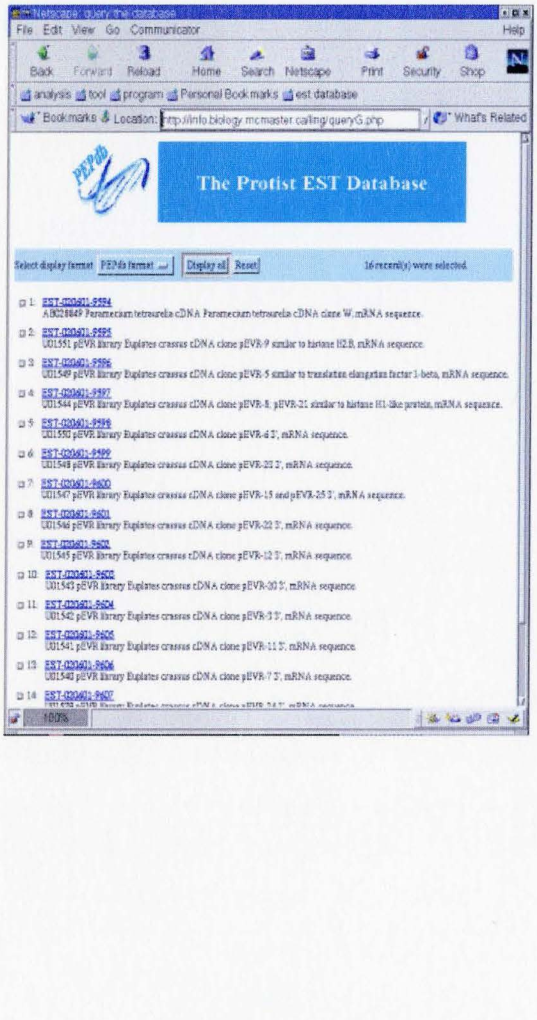
Enter a file containing the query sequences:

Expectation value (E): 10 The low complexity regions are ☐ filtered by default

Figure 1.12: Selected records displaying and Individual EST displaying

Both detailed searching and batch searching give the result page that displays a number of the selected records. Two form elements are used. One is a drop-down list which contains two options: "PEPdb format" and "Fasta format". The other is check boxes which are listed with these selected records. The check boxes allow users to specify which states they have visited and select one or more options by check marks. If users do not check any boxes, after clicking "Display all" button, the result page will display details for all ESTIDs listed in this page.

Information on individual ESTID includes eight categories: IDENTIFIERS, CLONE INFO, PRIMERS, SEQUENCE, COMMENT, LIBRARY, SUBMITTER and CITATIONS. In the LIBRARY part, the "Organism" name is linked to the NCBI taxonomy database.


Netscape: query the database
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop
analysis tool program Personal Bookmarks est database
Bookmarks Location: http://info.biology.mcmaster.ca/ling/queryG.php
What's Related
PEPdb
100%

The Protist EST Database

Select display format PEPdb format Display all Reset 16 record(s) were selected.

□ 1: EST-020601-9594
AB028849 Paramecium tetraurelia cDNA Paramecium tetraurelia cDNA clone W, mRNA sequence.

□ 2: EST-020601-9595
U01551 pEVR library Euplotes crassus cDNA clone pEVR-9 similar to histone H2B, mRNA sequence.

□ 3: EST-020601-9596
U01549 pEVR library Euplotes crassus cDNA clone pEVR-5 similar to translation elongation factor 1-beta, mRNA sequence.

□ 4: EST-020601-9597
U01544 pEVR library Euplotes crassus cDNA clone pEVR-8; pEVR-21 similar to histone H1-like protein, mRNA sequence.

□ 5: EST-020601-9598
U01550 pEVR library Euplotes crassus cDNA clone pEVR-6 3', mRNA sequence.

□ 6: EST-020601-9599
U01548 pEVR library Euplotes crassus cDNA clone pEVR-23 3', mRNA sequence.

□ 7: EST-020601-9600
U01547 pEVR library Euplotes crassus cDNA clone pEVR-15 and pEVR-25 3', mRNA sequence.

□ 8: EST-020601-9601
U01546 pEVR library Euplotes crassus cDNA clone pEVR-22 3', mRNA sequence.

□ 9: EST-020601-9602
U01545 pEVR library Euplotes crassus cDNA clone pEVR-12 3', mRNA sequence.

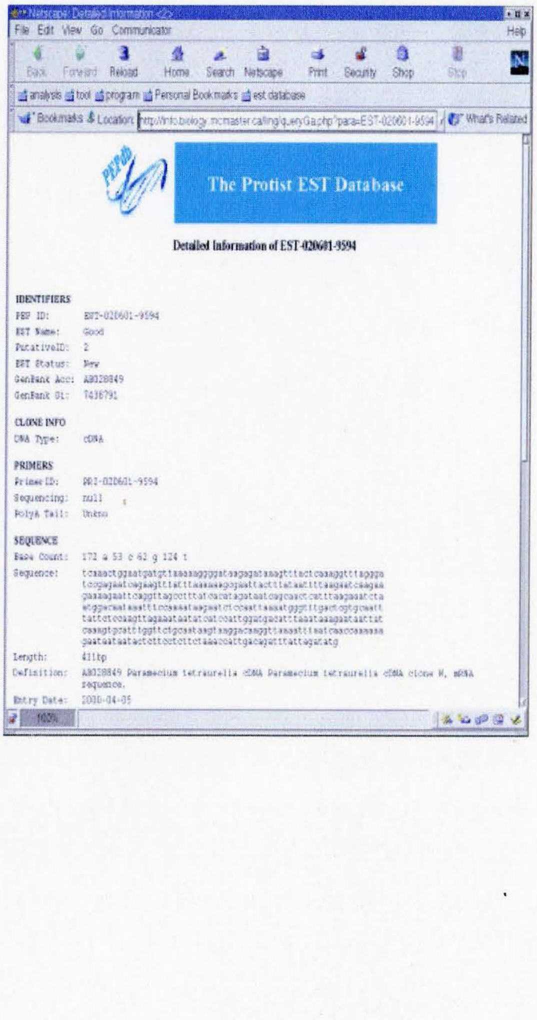
□ 10: EST-020601-9603
U01543 pEVR library Euplotes crassus cDNA clone pEVR-20 3', mRNA sequence.

□ 11: EST-020601-9604
U01542 pEVR library Euplotes crassus cDNA clone pEVR-3 3', mRNA sequence.

□ 12: EST-020601-9605
U01541 pEVR library Euplotes crassus cDNA clone pEVR-11 3', mRNA sequence.

□ 13: EST-020601-9606
U01540 pEVR library Euplotes crassus cDNA clone pEVR-7 3', mRNA sequence.

□ 14: EST-020601-9607


Netscape: Detailed Information <2>
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
analysis tool program Personal Bookmarks est database
Bookmarks Location: http://info.biology.mcmaster.ca/ling/queryGa.php?para=EST-020601-9594
What's Related
PEPdb
100%

The Protist EST Database

Detailed Information of EST-020601-9594

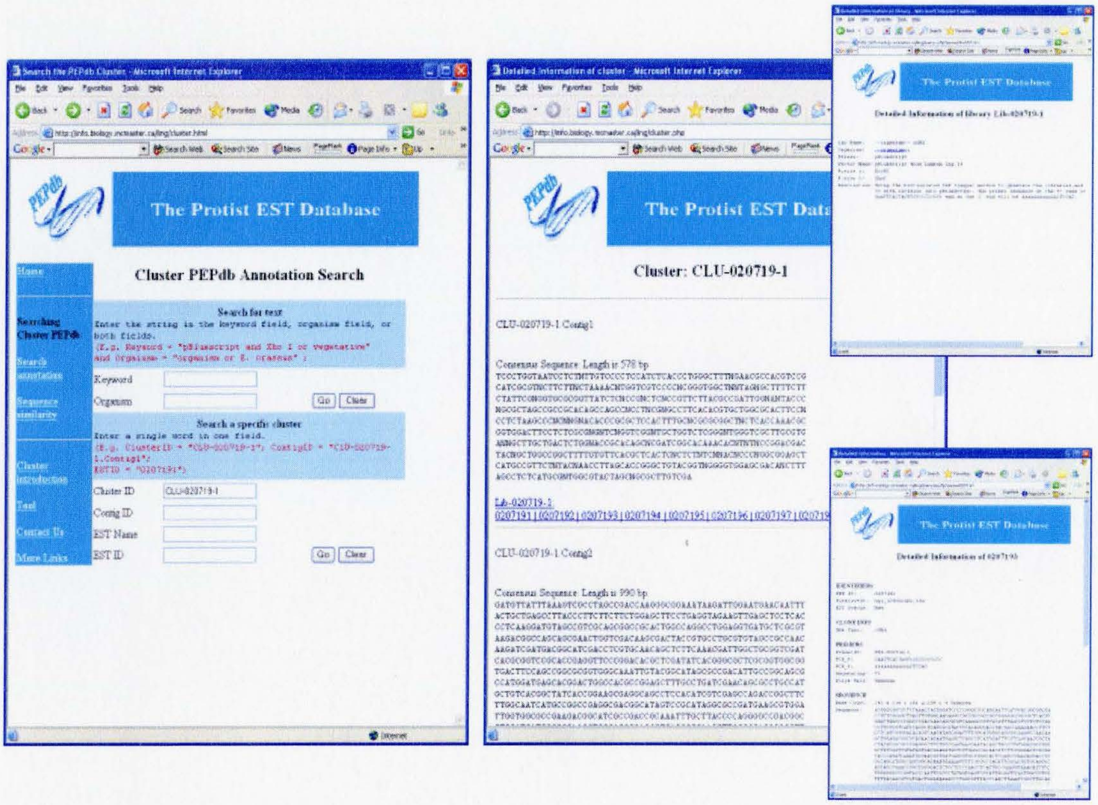
```
IDENTIFIERS
PEP ID:       EST-020601-9594
EST Name:     Good
PutativeID:   2
EST Status:   New
GenBank Acc:  AB028849
GenBank GI:   7436791

CLONE INFO
DNA Type:     cDNA

PRIMERS
Primer ID:    PRI-020601-9594
Sequencing:   null
PolyA Tail:   Unkno

SEQUENCE
Base Count:   172 a 53 c 62 g 124 t
Sequence:     tcaaactggaatgatgttaaaaaggggataagagataaagtttactcaaaggtttaggga
              tccgagaatcagaagtttatttaaaaaagcgaattactttataatttaagaatcaagaa
              gaaaagaattcaggttagcctttatcacatagataatcagcaactcatttaagaaatcta
              atggacaataaatttccaaaataagaatctccaattaaaatgggtttgactcgtgcaatt
              tattctccaagttagaaataatatcatccattggatgacatttaaataaagaataattat
              caagtgcatttggttctgcaataagtaaggacaaggttaaatttaatcaaccaaaaaa
              gaataataatactcttcctcttctaaacaattgacagatttattagatatg
Length:       411bp
Definition:   AB028849 Paramecium tetraurelia cDNA Paramecium tetraurelia cDNA clone W, mRNA
              sequence.
Entry Date:   2000-04-05
```

Figure 1.13: Cluster search, Contig displaying, Individual EST displaying and Library displaying

The Cluster PEPdb Annotation Search page has two areas of text boxes for searching cluster information, one is the "Search for text" box where the user can query a string in the "keyword" field, "organism" field, or both fields. The other is the "Search a specific cluster" box where the user can enter a single word in one field for specific searching. When searching for text such as "Keyword" or "Organism", the output page displays the selected ClusterID and check boxes with these ClusterID. The user can select one or more options from the listed items by the check marks. When searching for a specific cluster by entering Cluster ID, Contig ID, EST Name and EST ID, the output page displays the individual cluster information. There is a link of Sequence similarity to "The Sequence Comparison using Standard BLAST" page, so that searching will also be done by sequence similarity using BLAST against Clustered PEPdb.

The individual cluster information includes the clusterID, contigID, consensus sequence length, consensus sequence, the libraryID from which the EST sequences are generated, and ESTIDs which form the consensus sequences.

Each LibraryID is linked to a page to display this library information. Each ESTID is linked to a page to display this EST detailed information.



Search the PEPdb Cluster - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://info.biology.mcmaster.ca/ling/cluster.html
Go
Google Search Web Search Site News PageRank Page Info Up
Home
Searching Cluster PEPdb
Search annotation
Sequence similarity
Cluster introduction
Tool
Contact Us
More Links
Go
Clear
Go
Clear

The Protist EST Database

Cluster PEPdb Annotation Search

Search for text

Enter the string in the keyword field, organism field, or both fields.
(E.g. Keyword = "pBluescript and Xho I or vegetative" and Organism = "organism or E. gracilis")

Keyword

Organism

Search a specific cluster

Enter a single word in one field.
(E.g. ClusterID = "CLU-020719-1"; ContigID = "CLU-020719-1.Contig1";
ESTID = "0207191")

Cluster ID: CLU-020719-1

Contig ID

EST Name

EST ID

Detailed Information of cluster - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://info.biology.mcmaster.ca/ling/cluster.php
Google Search Web Search Site News PageRank

The Protist EST Data

Cluster: CLU-020719-1

CLU-020719-1.Contig1

Consensus Sequence: Length is 578 bp
[illegible]

Lib-020719-1
0207191 | 0207192 | 0207193 | 0207194 | 0207195 | 0207196 | 0207197 | 02071

CLU-020719-1.Contig2

Consensus Sequence: Length is 990 bp
[illegible]

The Protist EST Database

Detailed Information of library Lib-020719-1

[illegible]

The Protist EST Database

Detailed Information of 0207193

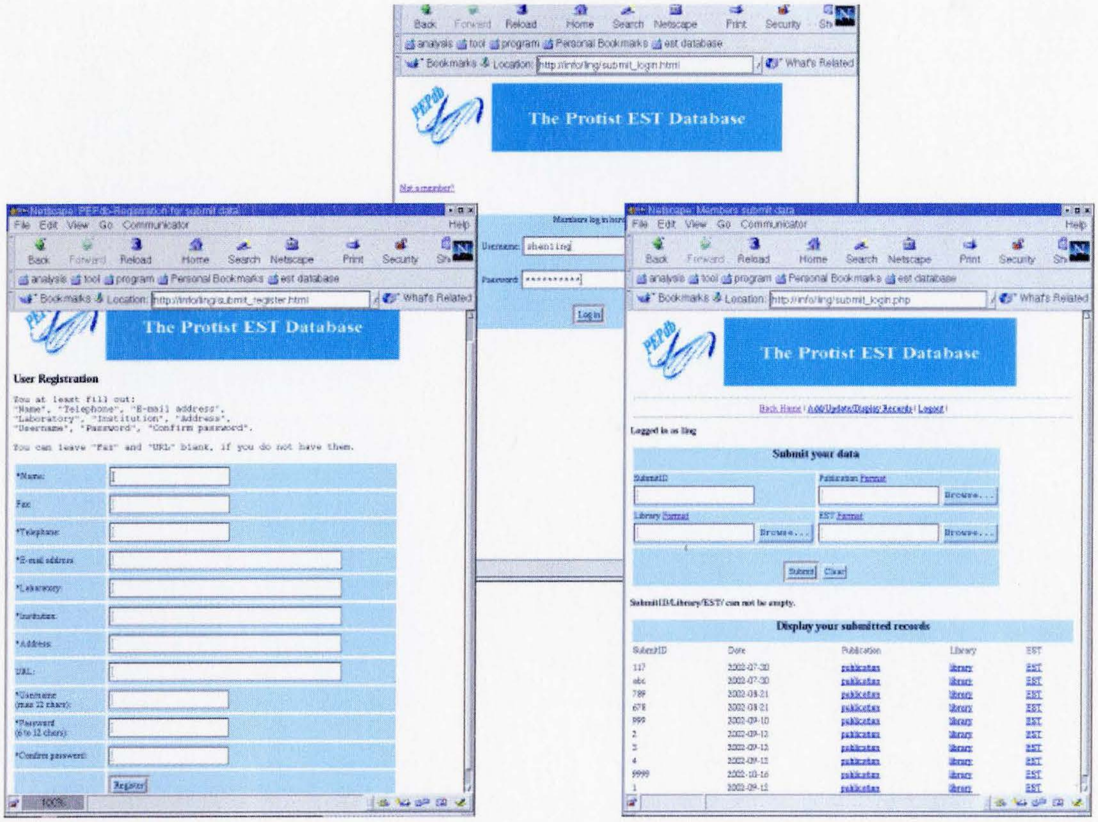
[illegible]

Figure 1.14: Member log on, User registration and Submit

There are two options on member log on page: the PEP member can register or log in if they have already registered. When a user clicks on "Not a member", the next page will display a registration form.

The registration form enters information about the PEP members. The script can display different error messages if a member gives a wrong password or a wrong username. After entering the valid username and password, the member will be logged in his own page.

The PEP member can submit data to the database in his own page. The member can upload the data files through the three text boxes, "Publication", "Library", and "EST". "SubmitID" is specified by the member. From the four text boxes, upload files and SubmitID number are imported into the database directly. The page can display the message whether the submission was successful or not. The site also displays a history of the member's submitted records. The history records include the SubmitID, submit Data, Publication records, Library information, and EST records. There is a link to "Logout" at top of this page.



Location: http://info/ling/submit_login.html

Not a member?

Members log in here

Username: shenling

Password: **********

Log in

Netscape: PEPdb-Registration for submit data

Location: http://info/ling/submit_register.html

User Registration

```
You at least fill out:
"Name", "Telephone", "E-mail address",
"Laboratory", "Institution", "Address",
"Username", "Password", "Confirm password".

You can leave "Fax" and "URL" blank, if you do not have them.
```

*Name:	
Fax:	
*Telephone:	
*E-mail address:	
*Laboratory:	
*Institution:	
*Address:	
URL:	
*Username (max 12 chars):	
*Password (6 to 12 chars):	
*Confirm password:	
	Register

100%

Netscape: Members submit data

Location: http://info/ling/submit_login.php

Back Home | Add/Update/Display Records | Logout |

Logged in as ling

Submit your data

SubmitID

Publication Format Browse...

Library Format Browse...

EST Format Browse...

Submit Clear

SubmitID/Library/EST/ can not be empty.

Display your submitted records

SubmitID	Date	Publication	Library	EST
117	2002-07-30	publication	library	EST
abc	2002-07-30	publication	library	EST
789	2002-08-21	publication	library	EST
678	2002-08-21	publication	library	EST
999	2002-09-10	publication	library	EST
2	2002-09-12	publication	library	EST
3	2002-09-12	publication	library	EST
4	2002-09-12	publication	library	EST
9999	2002-10-16	publication	library	EST
1	2002-09-12	publication	library	EST

1.4.4 Techniques for User interface

Users and Privileges.

The MySQL system can be used by users and PEP members. One of the best features of MySQL is that it supports a sophisticated privilege system. A privilege is the right to perform a particular action on a particular object, and is associated with a particular user. The "principle of least privilege" is that a user (or process) should have the lowest level of privilege required in order to perform his assigned task (Welling and Thomson 2001). According to this principle, in this case, for each member who needs to submit into the database, the root user should set up an account and password. These do not need to be the same as usernames and passwords outside of MySQL. The Grant command is used to create the PEP member's account and give them privileges. Since PEP members can submit data into the table and display their history records, they are granted the privileges of "SELECT" and "INSERT" on the "EST2" database, which contains the submitted data and submitter's information. Common users are granted the privilege of "SELECT" on the "EST1" database, which is the main database of the detailed and categorized cDNA information, in order that they can query and view the data.

System Security and Data Submission.

The PEPdb submit system obtains a security level by preventing unauthorized programs or persons from accessing objects directly. To achieve the secure system, members must identify themselves before access to the system granted. This is accomplished by entering a unique name and password, which will produce a unique identity. The system collects and manages member resources, and also ensures that any data on the network is not misused or destroyed. A account allows the member to participate his own domain or resource. All accounts have certain access rights and activities in common and share the same profile.

Figure 1.15: PEPdb Submit System

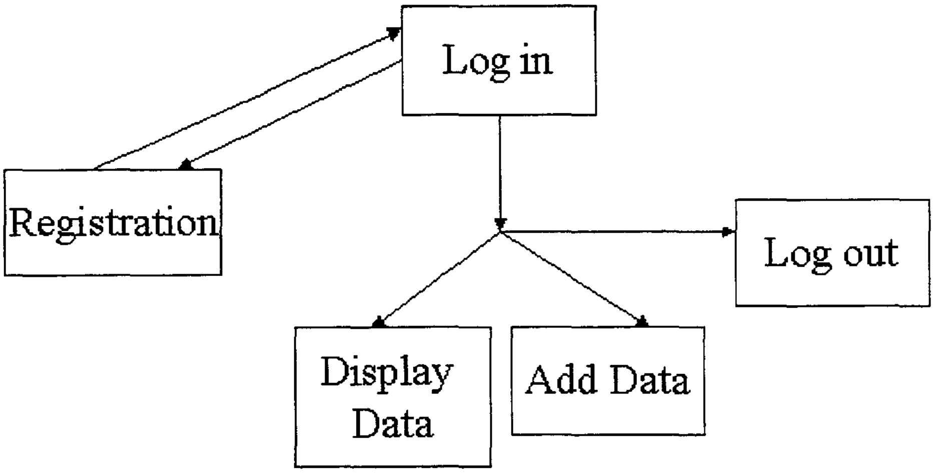

Log in
Registration
Log out
Display Data
Add Data

Figure 1.15 shows the PEPdb submit system. There are three main elements to the member authentication module: user registration, login and logout.

User registration: To tie a user to some personalization information, the database stores the member's details including the username and password in a table with ten columns. After the member fills in the text boxes in the User Registration page, and clicks the Register button, the script submit_register.php will execute and check if the form is filled in:

```
if ((!$name) || (!$lab) || (!$institution) || (!$address) || (!$username)|| (!$password) ||
    (!$confirm) || (!$tel) || (!$email))
```

Check if the email address supplied is valid:

```
if (!ereg(“[a-zA-Z0-9_]+@[a-zA-Z0-9\-]+\.[a-zA-Z0-9\-\.]+$”, $email))
```

Check if the telephone number is valid:

```
if (!ereg(“[0-9_\-\(\)]”, $tel))
```

Check if the password matchs the confirm password:

```
if ($password != $confirm)
```

Check that the password is the appropriate length:

```
if (strlen($password)< 6|| strlen($password) > 16)
```

Check if username is unique by comparing with username in the database:

```
select * from Person where Username=‘$username’
```

After checking is done, if there is no error message, the information is inserted into the table using SQL.

Login: If the members type their details into the form at submit_login.html and submit it, they will be taken to the script called submit_login.php. This script starts a session by calling session_start(). The script checks if the username and password are filled in:

```
if (empty($username) || empty($password))
```

After connecting to the database which stores the member's information, the script checks the password that the member has provided is the same as the stored one:

```
if($dbPassword == $password)
```

When script has made sure that the user provided a valid password and username, the script will create his username as a session variable:

```
session_register("valid_user");
$valid_user=$username;
```

The script will also display the "Submit your data" page by calling require("submit_add_up date.php").

Logout: Members can log out when they have finished using the site. This is very important for security. After clicking the link marked "Logout" at the top of the "Submit your data" page, the submit_logout.php script will be executed. The script logs out by using the functions:

```
$result=session_unregister("valid_user");
session_destroy();
```

The script then confirms that member is logged out:

```
if($result)
```

1.4.5 User Interface Programming

Processing Data From a Form.

The data is submitted in a form, which is part of a web page. It is easy to process data from a form using PHP. When a PHP page gets form information, the name attribute of a form element is converted to a PHP variable automatically. The data entered in the element is also assigned to that PHP variable. The variable is named by prefixing the name of the form element with a dollar sign ($).

To transfer the entered data to the PHP page, "get" and "post" methods of the form are used. The code for the Web page that contains the form should have an "action" attribute which contains the URL of the PHP page. For each form, there is a Submit button that allows users to send the entered data to the Web server, and then the Web server transfers the data to the PHP page that was located by the "action" attribute. The PHP page can perform an action with the data, such as query the data or displaying information in the Web browser.

In this project, text boxes and text areas are used to pass single values to a PHP page, and PHP code processes the entered data as a single variable. By using drop-down lists and check boxes as form elements, multiple selections can be passed for one element to a PHP page and PHP processes the data as an array of values.

Uploading a File.

The HTML form interface such as the "Batch Search" page allows users to upload a file, to process the data in the file. PHP supports HTTP upload, and writing the PHP to receive the file is very straightforward. The uploaded file comes from the browser to the Web server and is written into a temporary location on the server. The temporary file will be deleted if the PHP script does not move or rename the file. Since the "name" attribute is "uploadFile", the value stored in variable $uploadFile is the location where the file has been temporarily stored on the Web server. A function fopen($uploadFile, "r") is used to open the uploaded file for reading and processing the file. In the batch.php script, the code is as following:

```
$fp=fopen($uploadFile, "r");
$contents=fread($fp,filesize($uploadFile));
fclose($fp);
```

After reading the content of the uploaded file to a variable $contents, PHP code processes this variable and uses SQL to query the records from the database.

Session Control.

The idea of session control is to be able to track a user during a single session on a web site (Welling and Thomson 2001). HTTP does not provide a way to tell when the same user requests one page, followed by another. PHP includes session control functions so that using PHP, one can track the requests of one user from one transaction to another.

In this project, PHP session control functions are used to support logging in a member, and then displaying history records of that member and allowing that member to submit new data into the database. By assigning a session ID for each current user, the Web server keeps track of each session. The basic steps of using sessions in PHP are as follows:

Starting a session: When a session is started, the Web server stores a random session ID as a cookie on the user's computer. When the user requests another page from the site, the user's Web browser sends this session ID to the Web server to identify the same user. The function session_start() is used at the start of the PHP scripts.

Registering session variables: A session variable should be registered with a function, so that it can be tracked from one PHP script to another. The PHP script for the member submitting and viewing his records, calls function session_register("valid_user") so that the member who logs in is recorded as a single variable to pass to another script, and that member can add and display his own records in the page.

Using session variables: After the variable is carried to another script, the PHP script will deal with that variable according to the requirements of the user. For adding and viewing the data of the member, the PHP script checks if the user is valid, and then SELECTs his data from the database or INSERTs his data into the database.

Deregistering variables and destroying the session: When a user is finished with a session variable, the script calls session_unregister() to deregister it, and then calls session_destroy() to clean up the session ID. The functions session_unregister("valid_user") and session_destroy() will logout of a member's account.

Dealing with entered string.

Most searches are text searching in a database. Users access the PEPdb by entering a string into the box for text searching. The entered string allows the user to use logical operators such as AND, OR and NOT, so that query results can be specified. The entered string should be checked and filtered. The function searchterm($searchtype, $table) processes the string coming from the user. The function trims and filters the entered string, splits the string into an array, deals with each cell in the array using For loops, and then returns a query command which is concatenated by the cells of the array.

Building SQL.

Several PHP built-in functions were used for connecting, querying the database and retrieving the results.

To set up a connection to the MySQL database: *$linkID = mysql_pconnect("info0", "ling", "shen").*

The function returns a link identifier to "info0" MySQL true on success, or false on failure.

To choose a database to use: *mysql_select_db("est1", $linkID).*

To query the database: *$resultID = mysql_query($query, $linkID).*

The variable $query is SQL for running the query. One example in the queryG.php script for "Search for text" of the "Search PEPdb more detailed" page:

```
$query = "select EST_A.ESTID, Sequence_Content.Definition
from EST_A, Keyword, EST_Library, Sequence_Content, Person
where EST_A.SequenceID = Sequence_Content.SequenceID AND
Keyword.ESTID = EST_A.ESTID AND
EST_A.EST_LibraryID = EST_Library.EST_LibraryID AND
EST_A.Person_ID = Person.Person_ID AND
($queryOrganism) AND
($queryLibrary) AND
($queryLibDescrip) AND
($queryKeyword) AND
($queryPerson) ";
```

The variables in this query are trimmed strings that are entered through five text boxes.

To retrieve the query results: The results are retrieved in an enumerated array using function *$row = mysql_fetch_row($resultID).*

To disconnect from the database we call the function *mysql_close($linkID).*

Formatting output.

After queries have been done, the PHP script formats query results and displays them in the web page. A function display_query() is used for formatting. If the selected field is null in the database, it will be skipped and not be printed out. Non-null fields are printed out with titles.

1.5 Summary

The primary objectives of this project were to design a database to store various information from the PEP, to process components of the PEPdb, and to develop a friendly user interface.

Throwaway prototype: At the beginning, an ER model was designed with eight portions. In the throwaway version, there were no one-to-one relationships between entities. All attributes in the entities related by strict one-to-one relationships were organized in one *EST* entity. However since some attributes were modified or retrieved more frequently, and some attributes such as EST sequence are long text, they were taken out of the *EST* entity and categorized into different entities. By doing this way, it is easy to import the data into the table and also speed up the database system.

A throwaway user interface was created using PHP. It accessed the database using the SQL data manipulation language to perform insert, retrieval, update, and delete operations. The interface prototype was used to find problems in the database and refine the database until a suitable one, with maximal efficiency and ease of use, was developed which would encapsulate all entities and represent correct relationships.

Recognizing the importance of thorough testing to ensure high quality of the product, each script was tested as it was developed. Each procedure and function was tested in each script as well.

Analysis of results: A consistent, reliable and easy to use relational database was implemented in this project. The web pages possess features of speed, simplicity, consistence and clarity. All information about PEP members, Publications, Libraries, ESTs and Clusters are stored in the database system. The operations can be achieved by a friendly user interface.

Normalizing relations in the database separates one entity into two or more relations. This makes it possible to avoid modification anomalies. When relations were split, relationships were represented by matching primary and foreign keys. The database management system performs matching operations between relations. Queries allow data searches from more than one table. This relational database structure allows efficient retrieval of individual or groups of records, as well as adding records in a very simple manner. The strict enforcement of database constraints helps to maintain data consistency and accuracy.

Built-in functions of PHP4 are used to connect the interface to the database. The entered text string for querying is trimmed and filtered to guarantee that valid data is entered. When the database is queried, a set of data is returned and formatted by the PHP script and sent back to the interface.

The completed Requirements are listed as following:

1) Hardware setup and configuration.

2) Software installation and configuration.

3) Network system design and implementation.

4) Database design and implementation.

5) Web interface design and implementation.

During the design of the database system, future modifications were considered. The developed database system is very flexible and efficient. It is possible to add and remove parts of the system without altering the basic structure of the system.

Suggestions for future work and improvement: All PEP members have the same permission and right to log on the submit section with their own accounts in the current system. The members can modify only data at their classification level. However, some PEP members should have the right to access other member's information so as to supervise, manage and collect one part or several parts of the project. We should set up more than one account on the system in the future for highly privileged members who can view or modify data at several lower classification levels.

In the future, it may be possible to expand the existing database in order to generate more information. The current database, provides cluster information which includes consensus sequence, library and EST details. However sometimes, the user's concern is how a consensus sequence is formed. A method to display the alignment of a consensus sequence using graphics should be developed in the future. It can be added to the database by designing more relations and displaying the graphics with JAVA.

It is possible to add more function controls in the "Submit data" part to increase the member's interaction with the database. The current system provides a GUI for the PEP members to add their data into the database. For updating and modifying the EST sequences or records by the submitter, PEPdb should provide a Web-based editor which allows the member to update his own data On-line. So an "On-line modify" function in the database system needs to be developed.

Chapter 2

Large-scale EST Analysis in the Ciliated Protozoan *Tetrahymena thermophila*

2.1 Introduction

Tetrahymena thermophila belongs to the Alveolates. Alveolates are one of the largest and most diverse protist assemblages presently recognized and a major evolutionary branch of eukaryotic protists and have three primary lineages: Dinoflagellates are abundant components of the vast aquatic phytoplankton suspended near the water surface and provide the foundation of most marine and many freshwater food webs; Apicomplexans are parasites of animals and some cause serious human diseases (Campbell and Reece 2002); and Ciliates. Most Ciliates live as solitary cells in fresh water. In contrast to most flagella, cilia are relatively short. They are associated with a submembrane system of microtubules that coordinates the movement of the thousands of cilia. A unique feature of ciliate genetics is the presence of two types of nuclei, a large macronucleus and usually several tiny micronuclei

(Campbell and Reece 2002). Ciliates are completely non-photosynthetic (Taylor 1987).

The unicellular eukaryote *T. thermophila* is a member of the Ciliates, and is a free-living, fresh-water organism. *T. thermophila* cells have a striking variety of highly complex and specialized cell structures. Its metazoan-like cellular complexity occurs within a single large cell, starting from physiologically homogeneous clonal cultures (Orias 1998).

Over the past decade, researchers have manipulated the unique biology of *T. thermophila* to generate a premier experimental organism for functional genomic analysis (Turkewitz, Orias and Kapler 2002). Main discoveries made in *T. thermophila* include the discovery of the first cytoskeletal motor, dynein and its directional activity (Gibbons and Rowe 1965), identification of the molecular structure of telomeres (Blackburn and Gall 1978) and telomerase (Greider and Blackburn 1985), the first descriptions of the mechanism and structure of the *T. thermophila* self-splicing RNA and RNA catalysis (Cech 1990), and the role of nuclear histone acetyl transferase in transcription regulation (Brownell *et al.* 1996).

T. thermophila is an excellent unicellular model system for studying many aspects of animal biology that combines diverse experimental advantages with powerful capabilities for genetic manipulation (Wickert and Orias 2000). With a minimum doubling time below two hours, it is one of the fastest replicating eukaryotes making it an ideal organism to study (Orias, Hamilton and Orias 2000). Also, a simple efficient procedure for long term freezing of *T. thermophila* in liquid nitrogen have been developed (Cassidy-Hanley, Smith and Bruns 1995). Efficient mutant isolation and genetic analysis methods have been developed in order to better study this organism (Hamilton and Orias 2000). The gene knockouts technique is performed by exact gene replacement (Shen *et al.* 1995). These unicellular organisms maintain two functionally distinct nuclei within the same cytoplasm

- the silent 'germline' micronucleus and the transcriptionally active macronucleus - provides a powerful means for controlling the expression of transgenes (Turkewitz, Orias and Kapler 2002). This organism is used for quick, reliable, sensitive and inexpensive bioassays (Orias, Hamilton and Orias 2000).

Comparative genomics between human and model organisms can provide important clues as to gene interaction and function. Comparative analysis of domain structures, even between human and *E. coli* can still provide information on gene function (Clark 1999). With the complete genome sequence of *S. cerevisiae* , a finite number of about 6000 genes was shown to be sufficient to encode all the proteins from a eukaryotic cell (Goffeau 2000). The *T. thermophila* gene and protein number is estimated range from 20,000-40,000 (Calzone, Angerer and Gorovsky 1983). This number is of the same order of magnitude as that of Drosophila (14,000 genes), one order larger than yeast (*Saccharomyces*) and the same order as human (32,000 genes) (Modrek and Lee 2002). *T. thermophila* is likely to become an important model organism for genome-scale gene discovery and functional analysis for comparative genomics study (Fillingham *et al.* 2002).

ESTs or 'expressed sequence tags' are the part of the cDNA sequences which are read from both ends of expressed gene fragments. The ESTs have a variety of uses, including the discovery of novel genes, identification of homologous genes, analysis of alternative splicing, chromosomal localization of genes, and detection of polymorphisms (Pandey and Lewitter 1999). Underlying digital expression profiling can be performed by counting the ESTs and relating them to the total sequenced population of ESTs. This provides absolute estimates of mRNA expression levels (Kozian and Kirschbaum 1999; Prade *et al.* 2001). EST analysis allows not only the rapid identification of abundantly expressed genes, it also provides data sets for informing phylogenetic analyses (Li *et al.* 2003). BLAST comparison (Altschul *et al.* 1990) with ESTs against public databases can enable researchers to rapidly identify the putative genes. ESTs are a very effective tool for gene discovery in human and other organisms when combined with comparative genomics.

The gene discovery via large-scale EST projects, in addition to the sequencing and annotation of the *Tetrahymena* genome, will be of general scientific benefit (Fillingham *et al.* 2002). To further the process of gene discovery in the *T. thermophila* genome, analysis of a large scale ESTs was undertaken. A total of 3740 non-redundant *T. thermophila* assemblies and singletons from TIGR were analyzed in this project. The objectives are to study *T. thermophila* as an important model system for gene discovery and functional analysis using large scale ESTs analysis, and to study genes with restricted phylogenetic distributions versus genes with wide phylogenetic distributions by comparision of gene homologies using *T. thermophila* ESTs.

2.2 Methods

Sequences obtained. In total, 3740 *T. thermophila* sequences which are in Fasta format have been sent to us by TIGR. From a total 3740 unique sequences, there are 1166 Tentative Consensus sequences (TCs), 139 singleton ETs which are non-redundant sets of nucleotide sequences that represent mature transcripts, and 2435 singleton ESTs which are not contained in any assembly. TCs are created by assembling ESTs into the virtual contigs. TCs contain information on the source library and abundance of ESTs and in many cases represent full-length transcripts (http://www.tigr.org/tdb/tgi_info.html). ETs are sequences which were either loaded directly from GenBank (cDNAs) or were derived from genomic sequences when cDNAs were not available (http://www.tigr.org/tdb/tgi_info.html). ESTs are partial, single-pass sequences from either end of a cDNA clone. Singleton ESTs went through the assembly process but did not meet the match criteria (see below) to be assembled with any other EST in the collection of ESTs and other GenBank sequences used to create the consensus sequences (http://www.tigr.org/tdb/tgi_info.html).

Protocol for assembly of ESTs in TIGR. The 7,419 EST sequences were extracted from dbEST and these were trimmed to remove vector sequences, poly(A/T) tails, and contaminating bacterial sequences. The minimum length of ESTs used is 100bp. The 245 non-redundant transcript (ET) sequences were extracted from GenBank. Non-coding sequences were discarded and cDNAs and coding sequences from genomic entries were saved. Redundant entries for the same gene were removed (http://www.tigr.org/tdb/tgi_info.html).

Cleaned EST sequences and non-redundant transcript (ET) sequences were combined. Using the CAP3 Sequence Assembly Program (Huang and Madan 1999) sequences were assembled into contigs. TCs are consensus sequences based on two or more ESTs that overlap for at least 40 bases with at least 95% sequence identity. These strict criteria help minimize the creation of chimeric contigs. These contigs are assigned a Tentative Consensus (TC) number by TIGR (http://www.tigr.org/tdb/tgi_info.html).

TIGR constructed tentative consensus sequences that represent the underlying mRNA transcripts. It has demonstrated several advantages to this protocol. It separates closely related genes into distinct consensus sequences; it separates splice variants; and it produces longer representations of the underlying gene sequences (Liang *et al.* 2000).

Sequence analysis. The 3740 sequences were compared against the NCBI non-redundant protein database using BLASTX. It compares a nucleotide query sequence translated in all reading frames against a protein sequence database (Gish and States 1993). Sequences that did not match sequences in the protein databases were further analyzed by searching NCBI for non-redudant nucleotide database using BLASTN which compares a nucleotide query sequence against a nucleotide sequence database (States and Agarwal 1996). The assemblies were compared using BLASTX with a cutoff of expect value less than 10^{-4} and BLASTN with an expect value cutoff 10^{-5}. Searches were performed using BLASTCL3 program with a low complexity filter (SEG with BLASTX, DUST with BLASTN), BLOSUM62 matrix, and other default arguments. The resulting total matches and best matches are listed in the website "http://life.biology.mcmaster.ca/~ling/result.html" with links to the queried results of NCBI. Those records also were imported into the local database for storing and retrieving the results. A number of programs were written to help run the blast program and generate a summary of the PERL results.

2.3 Results and Discussion

2.3.1 Overall distribution of the general categories.

The 3740 sequences were compared against the NCBI non-redundant database using BLASTX and BLASTN. The total clusters and singletons are divided by different matches (putative homologs) when the non-redundant EST groups were searched for similarity using the non-redundant protein or nucleotide database. These were divided into three general categories: (1) highly significant (e value $< 10^{-20}$); (2) weakly significant ($10^{-20} \leq$ e value $< 10^{-4}$); (3) no significant match (e value $\geq 10^{-4}$) (Table 2.1).

Table 2.1: Summary of EST sequence categories

Category	No.of sequences	%
Highly significant (e value $< 10^{-20}$)	850	22.7%
Strong matches with *T. thermophila*	302 (35.5%)	
Strong matches with other organisms	548 (64.5%)	
Weakly significant ($10^{-20} \leq$ e value $< 10^{-4}$)	631	16.9%
Weakly matches with *T. thermophila*	64 (10.1%)	
Weakly matches with other organisms	567 (89.9%)	
No significant match (e value $\geq 10^{-4}$)	2259	60.4%
Total	3740	100%

From Table 2.1, 850 of the total 3740 sequences are in the highly significant category, of these 850 sequences, 302 (about 35.5%) strongly match with known proteins of *T. thermophila*, and 548 (about 64.5%) strongly match with putative, probable or hypothetical proteins other than from *T. thermophila.* Some of these putative proteins are derived from genome sequencing projects of model organisms (human, mouse, *Drosophila*, yeast, *Arabidopsis*, etc). 631 of the total 3740 sequences are weakly significant (e value $< 10^{-4}$). In the weakly significant category, approximately 10.1% of the sequences matched with known genes of *T. thermophila*, and approximately 89.9% of the sequences matched with proteins other than from *T. thermophila.* Among the 3740 sequences, 2259 sequences (about 60.4%) did not have a significant match and thus might represent either *T. thermophila* genes or transcripts that have not yet been isolated from other organisms.

Among the 3740 sequences, 1166 sequences are tentative sequences which are formed by ESTs overlap or contained within these sequences. These tentative sequences represent the underlying mRNA transcripts. 479 of the 1166 (about 41.08%) tentative sequences (TCs) had matches by searching the database (Table 2.2).

Table 2.2: The distribution of the sequences with significant matches to NCBI

Total unique sequences	TCs	Singletons	Total
No. of sequences matches with NCBI	479	1002	1481
No. of sequences without match	687	1572	2259
% of sequences with matches	41.08%	38.93%	39.60%
Total unique sequences	1166	2574	3740

2.3.2 Sequences displaying no significant matches in public databases.

Many of the sequences in this study did not give any information about their function. Approximately 60.4% of total ESTs and tentative sequences were classified into the class of no significant match with any sequence (show in Table 2.1). In addition, among the 1166 clusters, 687 tentative sequences (about 58.9% of total clusters) displayed no significant match to any known proteins (Table 2.2), indicating that a large majority of the genes expressed in *T. thermophila* are unknown and also these genes are unknown in other model organisms. These full-length transcripts may be unidentified in other organisms, possibly because they are novel Tetrahymena-specific, ciliate-specific or alveolate-specific genes (Fillingham *et al.* 2002).

2.3.3 Classification of genes according to their functions.

Functions of genes predicted or known, from the similarity of expect value less than 10^{-20}, were categorized into sixteen classes based on their biological roles (Fillingham *et al.* 2002). The number of sequences within the sixteen function classes are summarized in Table 2.3.

Table 2.3: Number of similarities within sixteen protein function classes

Functional classes	Similarities
energy metabolism and lipid metabolism	149
protein synthesis	97
DNA replication, recombination and repair, and chromotin functions	61
transport and binding proteins	57
cytoskeleton protein	53
amino acid metabolism	46
stress response, detoxification, and cell defence proteins	32
protein degradation and processing, proteases	30
signal transduction	26
regulatory functions	21
antigen	20
regulated secretion	19
transcription	17
protein folding	15
mitochondrial genome	21
hypothetical and not enough information to classify	186
Total	850

The 850 non-redundant sequences which were previously found to be in highly significant were further analyzed. The largest number (149) are related to energy metabolism functions. Other classes included sequences related to protein synthesis (97), DNA replication (61), transport and binding proteins (57), cytoskeleton protein (53), amino acid metabolism (46), stress response (32), protein degradation and processing (30), signal transduction (26), regulatory functions (21), antigen (20), regulated secretion (19), transcription (17), and protein folding (15).

Of the 97 ESTs matched to protein synthesis, 67 of them had similarities to ribosomal proteins. Two ESTs matched the translation initiation factor IF-2 protein that binds to the ribosome interacting with both 30S and 50S ribosomal subunits (La Teana, Gualerzi and Dahlberg 2001). Eight matched Elongation Factor1-alpha from *T. pyriformis* (Kurasawa *et al.* 1992).

53 ESTs had similarities to cytoskeleton proteins. The cytoskeleton is a dynamic three-dimensional structure that fills the cytoplasm, and is unique to eukaryotic cells. There are three types of cytoskeletal elements, microtubule, microfilament and intermediate filaments. In this study, the sequences that matched cytoskeleton proteins were divided into two groups. The first group had similarities to proteins for microtubules, especially to tubulin proteins, which act as a scaffold to determine cell shape, and also form the spindle fibers for separating chromosomes during mitosis (McKean, Vaughan and Gull 2001). The other group matched to microfilament proteins which the most abundant cellular protein and is associated with cellular movements including gliding, contraction, and cytokinesis (Bretscher 1991). In this project, microfilament sequence products included one actin protein, three actin-related, and one profilin proteins that encourages the polymerization of actin onto the barbed ends of actin filaments (Dos *et al.* 2003). There are no sequences similar to intermediate filaments, one of three types of cytoskeletal elements, which are

abundant in mammalian cells. This probably indicates that the intermediate filaments are highly expressed proteins in mammals, while microtubules and microfilaments are likely more conserved proteins which exist in *Tetrahymena thermophila*.

2.3.4 Analysis of expression levels.

To find particularly important biological pathways for an organism, identifying highly expressed genes can be useful. By counting ESTs and relating them to the total sequenced population of ESTs, absolute estimates of mRNA expression levels can be provided (Kozian and Kirschbaum 1999).

With the large numbers of ESTs available in this study, the number ESTs present in each contig should correlate with the expression level of a particular gene. To identify expressed genes, we summarized the top 40 assemblies ranked by the numbers of ESTs contained in each tentative sequence (Table 2.4).

Table 2.4: Relative mRNA abundance

Redundancy	TC	Gene
196	TC1	enolase [*Paramecium tetraurelia*].
182	TC587	SerH3 immobilization antigen [*T. thermophila*].
106	TC2	TUBULIN BETA CHAIN.
85	TC591	multifunctional beta-oxidation protein 2, peroxisomal - rat.
82	TC592	TUBULIN ALPHA CHAIN.
49	TC5	lysosomal acid phosphatase [*T. thermophila*].
38	TC593	histone H4, minor - *Tetrahymena pyriformis.*
37	TC594	ELONGATION FACTOR 1-ALPHA (EF-1-ALPHA).
32	TC9	similar to Succinyl-CoA ligase [GDP-forming] [*Mus*].
32	TC599	granule lattice protein 7 precursor; Grl7p.
32	TC598	Ndc1 protein [*T. thermophila*].
31	TC600	probable Zinc-binding dehydrogenases.
12	TC30	probable Zinc-binding dehydrogenases.
31	TC11	hypothetical protein [*Nostoc sp.* PCC 7120].
25	TC601	glutamate dehydrogenase [*Paramecium tetraurelia*].
24	TC603	granule lattice protein 4 precursor; Grl4p.
23	TC597	selenophosphate synthetase [*T. thermophila*].
12	TC596	selenophosphate synthetase [*T. thermophila*].
22	TC602	60S ribosomal protein L44 (12.4 kD) [*C. elegans*].
20	TC606	fructose-bisphosphate aldolase-like protein [*A. thaliana*].
20	TC17	4-hydroxyphenylpyruvate dioxygenase (4HPPD).
19	TC608	polyubiquitin 5 - *T. thermophila.*

Redundancy	TC	Gene
18	TC20	putative cytochrome c1 precursor [textitOryza sativa].
16	TC609	Tat-binding protein-1;
14	TC611	20k cyclophilin - Toxoplasma gondii (fragment).
14	TC26	Seryl-tRNA synthetase.
14	TC25	granule lattice protein 3 precursor[*T. thermophila*].
14	TC24	Triacylglycerol lipase, pregastric precursor.
14	TC22	ADP,ATP carrier protein 2, mitochondrial precursor.
13	TC616	acetyl-coa acetyltransferase (EC 2.3.1.9).
13	TC3	TUBULIN BETA-1 CHAIN.
13	TC3	*T. thermophila* beta-tubulin (BTU1) gene.
13	TC28	Proteasome subunit alpha type 7.
12	TC620	tyrosine aminotransferase [*Mustela vison*].
11	TC623	HIGH-MOBILITY-GROUP PROTEIN B.
11	TC622	*T. pyriformis* gene for 26S large subunit ribosomal RNA.
11	TC621	ATP synthase, H+ transporting mitochondrial F1complex.
11	TC35	Probable cystathionine gamma-synthase (CGS).
11	TC34	Citrate synthase, mitochondrial precursor.
11	TC31	nuclear RNA helicase (DEAD family) homolog - rat.

The most abundant genes represented in the data set are the enolase genes with 196 ESTs of TC1. The enolase superfamily contains a conserved binding site for a catalytically essential divalent metal ion and enolase will allow a better understanding of the strategy feature uses to evolve "new" enzymes from "old" enzymes (Gulick *et al.* 2001). The second most represented ESTs matched *T. thermophila* SerH3 with 182 ESTs of TC587. The SerH3 of *T. thermophila* is a temperature specific surface protein with numerous alleles encoding variants of the cell surface immobilization antigen (LaCrosse and Doerder 1994). The microtubule proteins, tubulin, had a total of 188 ESTs from both alpha and beta chains. Tubulin determines cell shape and provides a set of motor functions, and cannot be investigated in yeast. Multifunctional beta-oxidation protein 2 catalyzes the second (hydration) and third (dehydrogenation) reactions of the latter pathway (Baes *et al.* 2000). Lysosomal acid phosphatase is associated with phagocytosis, endocytosis and/or autophagy (Fillingham *et al.* 2002). The other proteins receiving the most ESTs (Table 2.4) are protein synthesis and secreted proteins, including a collection of ribosomal proteins, elongation factors and granule lattice proteins.

2.3.5 Identification of genes restricted to the Ciliophora and the Alveolata.

T. thermophila is a member of the Ciliates, and has a striking variety of highly complex and specialized cell structures. To identify how many *T. thermophila* ESTs are restricted to Ciliophora will improve our understanding about genes conserved in *T. thermophila*. Identifying genes that are restricted to the ciliophora may help us to understand their unique biology.

Assemblies that did not have similarity to non-alveolata proteins with an expect value less than 10^{-20}, but which did have similarity (e value $< 10^{-20}$) to alveolata proteins were identified as alveolata-specific. The best matched protein was used to annotate their putative protein.

Among of 850 highly significant sequences, 201 sequences were restricted to the Alveolata kingdom. Only five sequences had similarities to Apicomplexa. While the other 196 sequences were specifically restricted to the ciliophora. 182 of the 196 sequences matched with *Tetrahymena thermophila*, 11 of them with *T. pyriformis*, and only three sequences matched with *Paramecium*. Only 1.6% of all highly significant sequences have best non-self (not from *T. thermophila*) similarity to another ciliophora. This likely displays the relatively small coverage of their genomes and the lack of proteins from these organisms in the database.

Of the 196 ciliophora-specific matches, 17 sequences matched the granule lattice genes (GRL1, 3,4,5 and 7). The GRL1 gene product has been shown to have a role in regulating secretion in the mucocyst, a secretory granule (Chilcoat *et al.* 1996). Also belonging to the ciliophora-specific matches, eight sequences matched *T. thermophila* SerH. Three other sequences matched to *Tetrahymena* tetrin A, B and C gene products. Tetrin, the cytoskeletal proteins of the oral filament, are a cluster of polypeptides with molecular masses ranging from 79 to 89 KD. These filaments are clearly distinguishable from intermediate filaments of metazoan (Brimmer and Weber 2000; Dress *et al.* 1992).

In the five sequences which had matches to Apicomplexa, TC802 matched to both the Ciliophora and the Apicomplexa gene, Cytochrome c. TC1092 matched to dynamin-like protein of *Plasmodium yoelii yoelii* in Apicomplexa. A few sequences matched to Apicomlpexa, but there were no sequences that matched to other Alevolata. It seems that ciliophora shares more genes with neighbor apicomplexa than with other taxon in Alveolata.

2.3.6 Sequences match proteins from human but not *Saccharomyces cerevisiae.*

World wide databases contain complete sequences of important multicellular eukaryotes, such as human, mouse, fruit fly, maize and *Arabidopsis thaliana*, however only a few unicellular eukaryotes, such as *S. cerevisiae*, has been completely sequenced. Experimental biology at the molecular and cellular level are easier in *Tetrahymena* than in multicellular eukaryotes because of rapid growth rate and clonal homogeneity of cell cultures. Evidence that humans share a high degree of functional conservation with ciliates is seen through better matches of *Paramecium* coding sequences to humans, than to non-ciliate microbial genetic model organisms (Dessen *et al.* 2001). Some of the genes shared by humans and *Tetrahymena* are missing in yeast. This shows that *Tetrahymena* is a well-established model organism for the study of fundamental molecular, cellular and developmental biology.

In this study, the sequences of *T. thermophila* genes and functions that give high significant matches with expect value less than 10^{-20} to human sequences but not to *S.cerevisiae* sequences (e value $< 10^{-20}$) were analyzed. The best protein similarity found, when comparing *T. thermophila* to human, was used to annotate the putative protein to *T. thermophila*.

During this comparison, a total of 508 sequences gave high confidence matches (e value $< 10^{-20}$) to human in a variety of gene functions. Among of them, 189 sequences matched to proteins from human but not *S. cerevisiae*. Genes from these sequences were identified to belong to a variety of functional categories. Four contigs TC926, TC185, TC190, TC192 and one EST M401028 matched Cathepsin H,L and S. Cathepsin L is the most active among lysosomal cysteine endopeptidases and is involved in several intra- and extra-cellular degradation processes in normal and pathological states (Reinheckel *et al.* 2001). Cathepsin-mediated diseases include: Alzheimer's, numerous types of cancer, autoimmune related diseases such as arthritis, and the accelerated breakdown of bone structure seen with osteoporosis (Buhling *et al.* 2000). Contig TC597 matched selenophosphate synthetase which is an enzyme and disturbance of selenoprotein expression or function is associated with some deficiency syndromes (Keshan and Kashin-Beck disease) (Kohrl *et al.* 2000). TC47 and M401317 matched only to the human specific cDNA KIAA 1181 and KIAA 0643 with unknown functions. EST BM396869 encodes a calmodulin 2 gene, which is a central regulatory element in gene regulation, protein synthesis, secretion, ion channel function, cell motility & chemotaxis (Van Eldik and Watterston 1998). BM394664 EST best matched to the human dynein light chain which is involved in ATPase activity, constitutes the side arms of the outer microtubule doublets in the ciliary axoneme and is responsible for the sliding (Fang *et al.* 1997). Microtubule areas represent important parts of human biology that cannot be investigated in *S. cerevisiae.*

2.3.7 Identification of MAPK/ERK signaling cascade in *T. thermophila*

Mitogen-activated protein kinases (MAPK) are a family of serine/threonine protein kinases widely conserved among eukaryotes. The MAPK/ERK signaling cascade is activated by a wide variety of receptors involved in growth and differentiation. *Tetrahymena* is an unicellular model system and MAPK/ERK plays a critical role in the regulation of cell growth and differentiation. New regulatory signal transduction pathways can be determined by using EST analysis in *Tetrahymena thermophila.*

In this study, sequence similarities represent an exciting avenue which explores many signal transduction functions. Of 850 sequences with high similarity (expect value <

10^{-20}) that have been analysed, 3.1% matched signal transduction proteins including, Rac, Ras, MAPK, ERK1, PKC, cAMP and 14-3-3 (a protein involved in signal transduction, exocytosis and cell cycle regulation). Sequence alignments confirm the selective sequences have a low expect value and are highly homologous to target proteins. MAPK signaling cascades are divided into three-tiered modules. MAPKs are phosphorylated and activated by MAPK-kinases (MAPKKs), which in turn are phosphorylated and activated by MAPKK-kinases (MAPKKKs). The MAPKKK is in turn activated by interaction with a family of small GTPases and/or other protein kinases connecting the MAPK module to the cell surface receptor or external stimuli (Cobb 1999; Garrington and Johnson 1999). The MAPK/Erk signaling cascade is activated by a wide variety of receptors involved in growth and differentiation including receptor tyrosine kinases (RTKs), integrins, and ion channels. The specific components of the cascade vary greatly among different mitogenic and external stimuli and transduct the signal to small GTP binding proteins (Ras, Rap1), which in turn activate the core unit of the cascade composed of a MAPKKK (Raf), a MAPKK (MEK1/2) and MAPK (Erk). An activated Erk dimer can regulate targets in the cytosol and also translocate to the nucleus where it phosphorylates a variety of transcription factors regulating gene expression (Giancotti and Ruoslahti 1999; Lewis, Shapiro and Ahn 1998). Our results indicate that *T. thermophila* likely contains the MAPK/ERK signaling pathway and provides an excellent unicellular model system for studying signal transduction function.

2.3.8 Phylogenetic comparison of gene homologies.

It is interesting to determine the phylogenetic associations of different taxa. This study could be useful not only for establishing gene origins according to the closer neighbors, but also for identifying genes with restricted phylogenetic distributions versus genes with wide phylogenetic distributions.

The sequences were divided into different categories according to the similarities to different taxonomic group proteins with expect value less than 10^{-20}. The best match proteins were used to annotate their putative function. The results from three taxonomic levels are summarized. The first level, the superkingdom, were divided into three domains, including Eukaryota, Archaea and Bacteria (Dacks and Doolittle 2001). The second level, kingdom, includes Alveolata, Fungi, Metazoa and Viridiplantae. The third level is phylum which includes Apicomplexa, Ciliophora and Dinophyceae. The number of *T. thermophila* sequences that had strong matches to different taxonomic groups were displayed in Figure 2.1 A B C and Table 2.5. In Figure 2.2, Figure 2.3 and Figure 2.4 show the distributions of protein classes in different Taxonomic groups.

Distribution of significant similarities in Eurkaryota, Archaea and Bacteria.

In the superkingdom level, 850 strong match sequences were separated into the seven groups (shown in Figure 2.1 A).

Based on Venn diagram A analysis, 585 sequences (about 68.8%) had similarities only to Eukaryota proteins. 24 sequences matched only to Bacteria proteins. These results show that *T. thermophila* contains prokaryotic-like genes, implicating that it provides a connection between the sequences of prokaryotic organisms and the Eukaryotic organisms (eg. Human). *T. thermophila* is a model organism which exists with abundant Eukaryota proteins and Prokaryotic proteins in the same metabolic pathway. There was no match only to Archaea proteins. Just three sequences matched with both Archaea and Bacteria proteins. And 26 sequences matched to Eukaryota-Archaea group. Of the genes in Ciliophora that are present in Archaea, the majority are ancestral to all life forms.

Based on Figure 2.2 analysis, in the 585 sequences which matched only to Eukaryota, these sequences exist with obviously high representation in all the sixteen protein classes summarized previously. Among these, 19 had similarities to regulated secretion proteins which only exist in Eukaryota. Of the 24 sequences which only matched to Bacteria, none had a similarity to cytoskeleton, antigen, regulatory, degradation, amino acid metatolism, protein folding, or signal transduction function proteins. However, in this Bacteria group, there were seven sequences with similarities to energy metabolism genes, such as 3-oxoadipate CoA-transferase, and succinyl-diaminopimelate desuccinylase. It further indicates that some energy metabolic enzymes in *T. thermophila* may have similar functions with Prokaryotic metabolic proteins. Since there are no cytoskeleton, antigen, or regulatory function proteins existing in both Bacteria and Eukaryota-Bacteria group, these proteins are associated with Eukaryotic features in *T. thermophila*.

A total of 85 sequences had similarities common to three domains: Eukaryota, Archaea and Bacteria. These similarities distribute in different protein classes. These ancestral genes in all three groups include: (1) 2 (2.3%) matched to protein synthesis genes; (2) 29 (34.1%) are energy metabolism proteins; (3) 5 (about 5.9%) matched to genes of amino acid metabolism; (4) 6 (7%) matched to DNA replication proteins; (5) 7 (8.2%) matched to protein degradation proteins; (6) 3 (3.5%) matched to transport and binding proteins; (7) 3 (3.5%) matched to protein folding; (8) 4 (4.7%) matched to stress response proteins; (9) 2 (about 2.6%) sequences matched to transcription genes; (10) 1 (1.1%) matched to antigen protein. There are no cytoskeleton, signal transduction or regulatory proteins that are common to these three domains.

Figure 2.2 shows distribution of protein synthesis function class in the three surperkingdoms. Of 97 sequences which matched to protein synthesis genes, 79 (81%) exist in Eukaryota, 5 (5.2%) in Eukaryota-Bacteria group, 2 (2.1%) in Eukaryota-Archaea-Bacteria group, and 10 (about 10.3%) in Eukaryota-Archaea group. Only one sequence (about 1.0%) had similarity to Bacteria. No homology exists in Archaea-Bacteria group. A relatively large set of proteins are confined to Eukaryota and Archaea in the translation process (protein synthesis function class). From Figure 2.2, the number of the sequences which are involved in transcription functions are less than sequences in other functional classes. In this class, it was found that 10 (about 58.8%) of the 17 had similarities to Eukaryota, 2 (about 11.76%) to Eukaryota-Bacteria group, 2 (about 11.76%) common to three superkingdoms, 2 (11.76%) to Eukaryota-Archaea group, and 1 (about 5.9%) to Bacteria. Although no transcription gene is distributed in the Archaea-Bacteria group, some Archaea and Bacteria genes have similarities to Eukaryota proteins, such as ribonucleotide reductase, RNA helicase, and RNA polymerase II subunit 2. In Figure 2.2, the distribution of DNA replication functions shows that of the 61 similarities, 48 (78.7%) sequences have similarities to Eukaryota, 3 (4.9%) to Eukaryota-Bacteria group, 6 (9.8%) to Eukaryota-Archaea-Bacteria group, 2 (3.3%) to Eukaryota-Archaea group, and 2 (3.3%) to Bacteria alone. Again, no homology exists in the Archaea-Bacteria goup. From the limited number of similarities in this function class, some genes have similarities common to Eukaryota and Bacteria or Archaea, such as endonuclease SceI, nucleoside diphosphate kinase, and telomerase component P80. Strikingly, from the three main cellular information processing systems, translation, transcription and DNA replication, there are a large number of proteins which are uniquely Eucarya. This number is significantly larger than the number of similarities common to the three domains and other taxonomic groups. And so these three basic functions of a cell are likely to be highly diverged in each of the three domains of organisms.

From Figure 2.2, all of the 53 sequences which matched to the cytoskeleton class are uniquely Eukaryota. Regulatory function proteins present that 20 of the 21 sequences of this class are restricted in Eukaryotes, one of the 21 sequences matched to Eukaryota and Archaea. In the 26 sequences which matched to the signal transduction function class, 24 sequences have similarities to Eukaryota. This confirms that cytoskeleton proteins are uniquely Eucarya. The regulatory function and signal transduction genes that are present in *T. thermophila*, are very similar to Eucarya. It seems that Eucarya and Archaea or Bacteria versions of these proteins are very different from each other.

The biggest number of matched sequences (149) are in energy metabolism function class. Of these sequences, there is 43.6% in Eukaryota, 31.5% in Eukaryota-Bacteria group, and 19.5% in the group of Eukaryota-Archaea-Bacteria. Many of the energy metabolism proteins are conserved among the organisms of the three domains.

Distribution of significant similarities in Alevolata, Fungi, Metazoa and Viridiplantae.

Among the 823 sequences which had similarities to Eukaryotic proteins, 822 had similarities to proteins of the four kingdoms: Alevolata, Fungi, Metazoa and Viridiplantae (one sequence was excluded due to the possibly it matched to a kingdom which was not studied). These sequences were divided into fifteen groups (showed in Table 2.5). Fungi and Metazoa kingdoms are combined in Figure 2.1 B.

Based on the analysis in Table 2.5, many of the sequences (300) are distributed among all of the four kingdoms, and less are restricted to the Alveolata (206). It seems that a large number of protein sequences within the Alveolata are ancestral and these genes are conserved among Alveolata, Fungi, Metazoa and Viridiplantae. Based on the analysis in Figure 2.3, there are 15 protein function classes within these four kingdoms, while only regulated secretion proteins which restricted in Alveolata kingdom, is not included in this group. Among the similarities in Alveolata exist all the 16 protein classes.

Approximately 3.3% of the sequences show high significant similarity to Viridiplantae.

Only 0.85% had matches to Fungi, whereas 8.6% were most similar to Metazoa (Table 2.5). In Figure 2.3, three protein classes, energy metabolism, DNA replication and signal transduction proteins, are included in the Fungi group. Eight protein classes exist in the Viridiplantae group. It seems that Alveolata shares more genes with Viridiplantae than with Fungi because there are more genes common to both Alveolata and Viridiplantae than to Alveolata and Fungi.

There are only three protein classes of the nine sequences in the Alveolata-Viridiplantae group, while there are eight protein classes of the 27 sequences within Viridiplantae group (Figure 2.3). It suggests that many plant-like genes in *T. thermophila* are putative and not annotated.

Among the Alveolata-Viridiplantae group there are some metabolic enzymes such as the Vacuolar-type H(+)-translocating inorganic pyrophosphatases (AAK38076) that have long been considered to be restricted to plants and to a few species of phototrophic bacteria. However, in recent investigations these pyrophosphatases have been found in parasitic protists (Drozdowicz and Rea 2001). These results indicate that some metabolic enzymes in *T. thermophila* are plant-like. These genes have likely originated from a common Prokaryotic ancestor. The metabolic pathways in this organism may have some similarities to plants although it is a non-photosynthetic organism.

Distribution of significant similarities in Ciliophora, Apicomplexa and Dinophyceae.

All of the 603 sequences which had similarities to Alveolata proteins were further separated into three phyla including Ciliophora, Apicomplexa and Dinophyceae with seven groups (shown in Figure 2.1 C).

From the analysis in Figure 2.1 C, many sequences (256) matched only to Ciliophora proteins. There are not any which matched only to Dinophyceae sequences. 182 sequences matched solely to Apicomplexa proteins. The 127 sequences in the common area of Apicomplexa and Ciliophora proteins are much more numerous than the sequences in the common area of the three phyla (35). Among these three phyla, there are more common genes between Ciliophora and Apicomplexa than either has to Dinophyceae. This indicates that Ciliophora shares more genes with Apicomplexa than with Dinophyceae. Other studies have shown that Apicomplexa and Dinoflagellates are sister groups, although the relationship is not strong by combined gene phylogenies (Fast *et al.* 2002). Since a large number of genes have been annotated for both ciliates and apicomplexa but only a limited number of molecular sequences are known from dinoflagellates, Alveolate relationships need further study.

Based on the analysis in Figure 2.4, regulated secretion proteins only exist within the Ciliophora group among the seven groups. This further confirms that regulated section (eg. GRL) proteins are Ciliophora-specific genes. Most of the antigen proteins (eg. SerH) that exist in *T. thermophila* are Ciliophora-specific genes. The mitochondrial genome of *T. thermophila* probably evolved rapidly, since the sequences that matched to mitochondrial genome are Ciliophora-specific. The similarities within Apicomplexa-Ciliophora-Dinaphyceae group include protein synthesis (5), cytoskeleton (5), energy metabolism (2), transport (6), DNA replication (1), protein degradation (3), protein folding (8) and signal transduction (3) proteins. These common genes of the three phyla are useful to further analyze the phylogeny of these groups.

Table 2.5: Distribution of the sequences within kingdom groups

Taxonomic rank	No. of sequences	Total of matches
Alveolata-Fungi-Metazoa-Viridiplantae	300	
Alveolata	206	
Fungi	7	
Metazoa	71	
Viridiplantae	27	Kingdom 822
Alveolata-Fungi	5	
Alveolata-Metazoa	20	
Alveolata-Viridiplantae	9	
Fungi-Metazoa	9	
Fungi-Viridiplantae	7	
Metazoa-Viridiplantae	35	Alveolata 603
Alveolata-Fungi-Metazoa	24	
Alveolata-Metazoa-Viridiplantae	34	
Alveolata-Fungi-Viridiplantae	5	
Fungi-Metazoa-Viridiplantae	63	

Figure 2.1: Venn Diagram

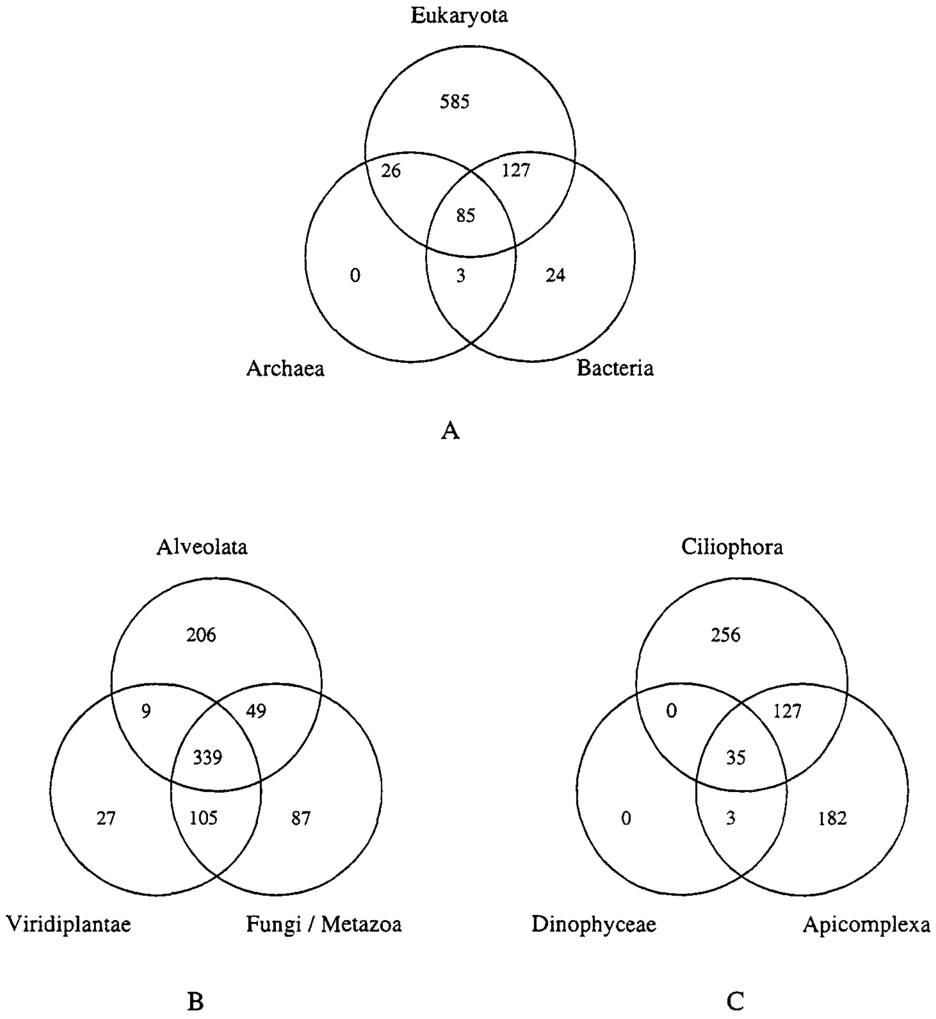

Eukaryota
585
26
127
85
0
3
24
Archaea
Bacteria
A
Alveolata
206
9
49
339
27
105
87
Viridiplantae
Fungi / Metazoa
B
Ciliophora
256
0
127
35
0
3
182
Dinophyceae
Apicomplexa
C

Figure 2.2: Distribution of function classes within superkingdoms

1.protein synthesis

2.cytoskeleton protein

3.energy metabolism and lipid metabolism

4.transport and binding proteins

5.DNA replication, recombination and repair, and chromotin functions

6.protein degradation and processing, proteases

7.amino acid metabolism

8.protein folding

9.antigen

10.regulated secretion

11.signal transduction

12.regulatory functions

13.stress response, detoxification, and cell defence proteins

14.transcription

15.mitochondrial genome

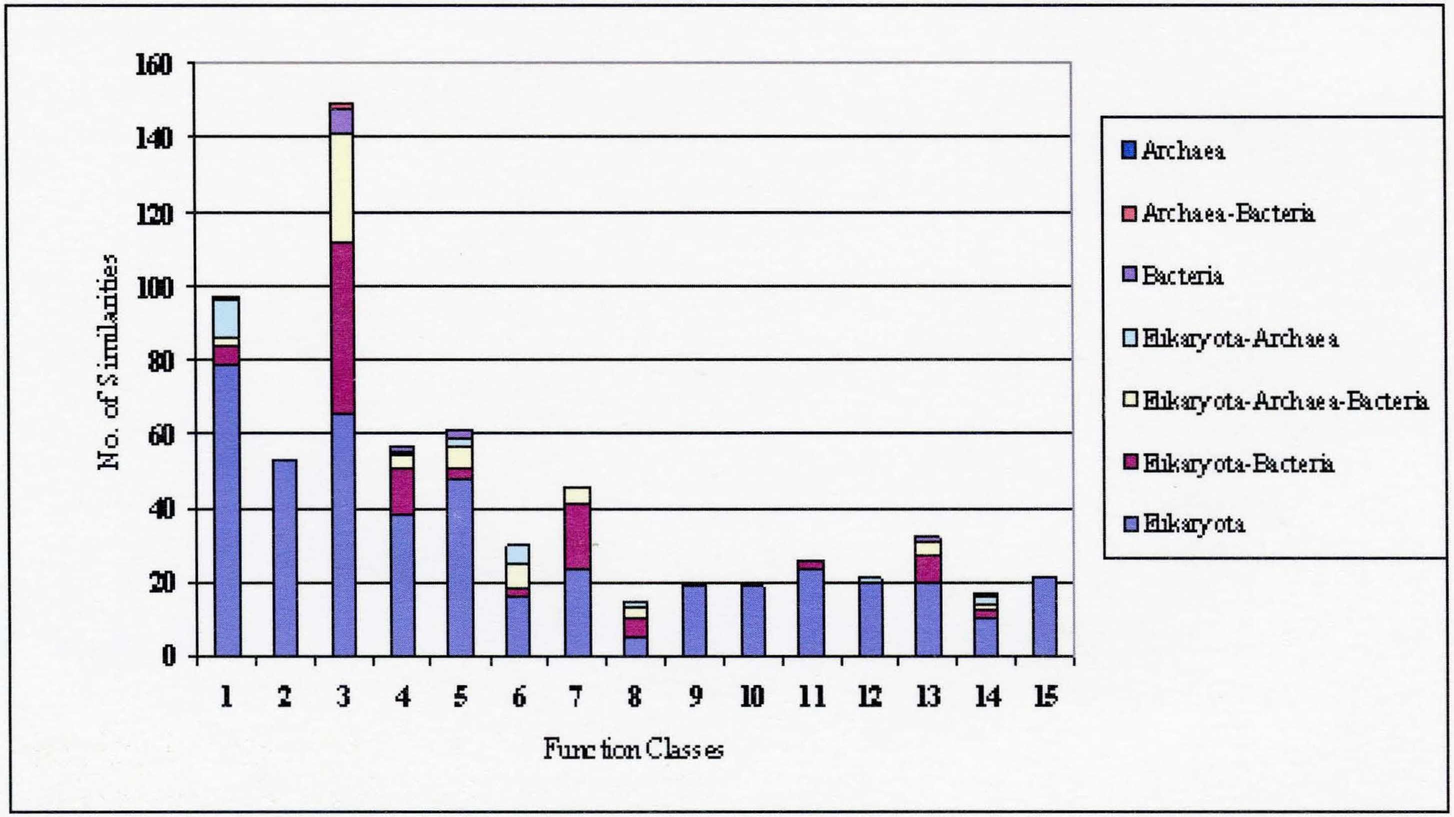

No. of Similarities
0
20
40
60
80
100
120
140
160
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Function Classes
Archaea
Archaea-Bacteria
Bacteria
Eukaryota-Archaea
Eukaryota-Archaea-Bacteria
Eukaryota-Bacteria
Eukaryota

Figure 2.3: Distribution of function classes within kingdoms

1.protein synthesis

2.cytoskeleton protein

3.energy metabolism and lipid metabolism

4.transport and binding proteins

5.DNA replication, recombination and repair, and chromotin functions

6.protein degradation and processing, proteases

7.amino acid metabolism

8.protein folding

9.antigen

10.regulated secretion

11.signal transduction

12.regulatory functions

13.stress response, detoxification, and cell defence proteins

14.transcription

15.mitochondrial genome

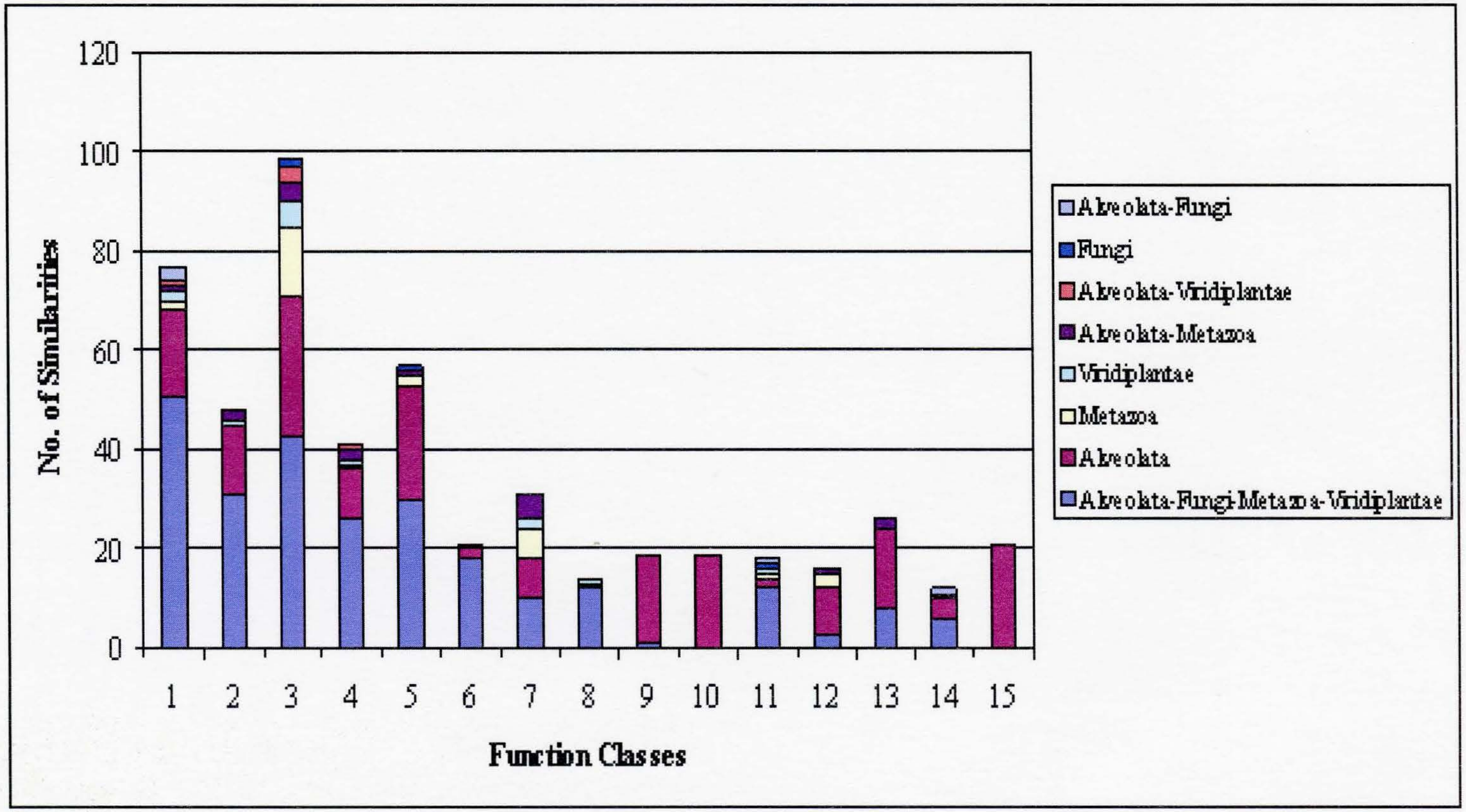

120
100
80
60
40
20
0
No. of Similarities
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Function Classes
Alveolata-Fungi
Fungi
Alveolata-Viridiplantae
Alveolata-Metazoa
Viridiplantae
Metazoa
Alveolata
Alveolata-Fungi-Metazoa-Viridiplantae

Figure 2.4: Distribution of function classes within phyla

1.protein synthesis

2.cytoskeleton protein

3.energy metabolism and lipid metabolism

4.transport and binding proteins

5.DNA replication, recombination and repair, and chromotin functions

6.protein degradation and processing, proteases

7.amino acid metabolism

8.protein folding

9.antigen

10.regulated secretion

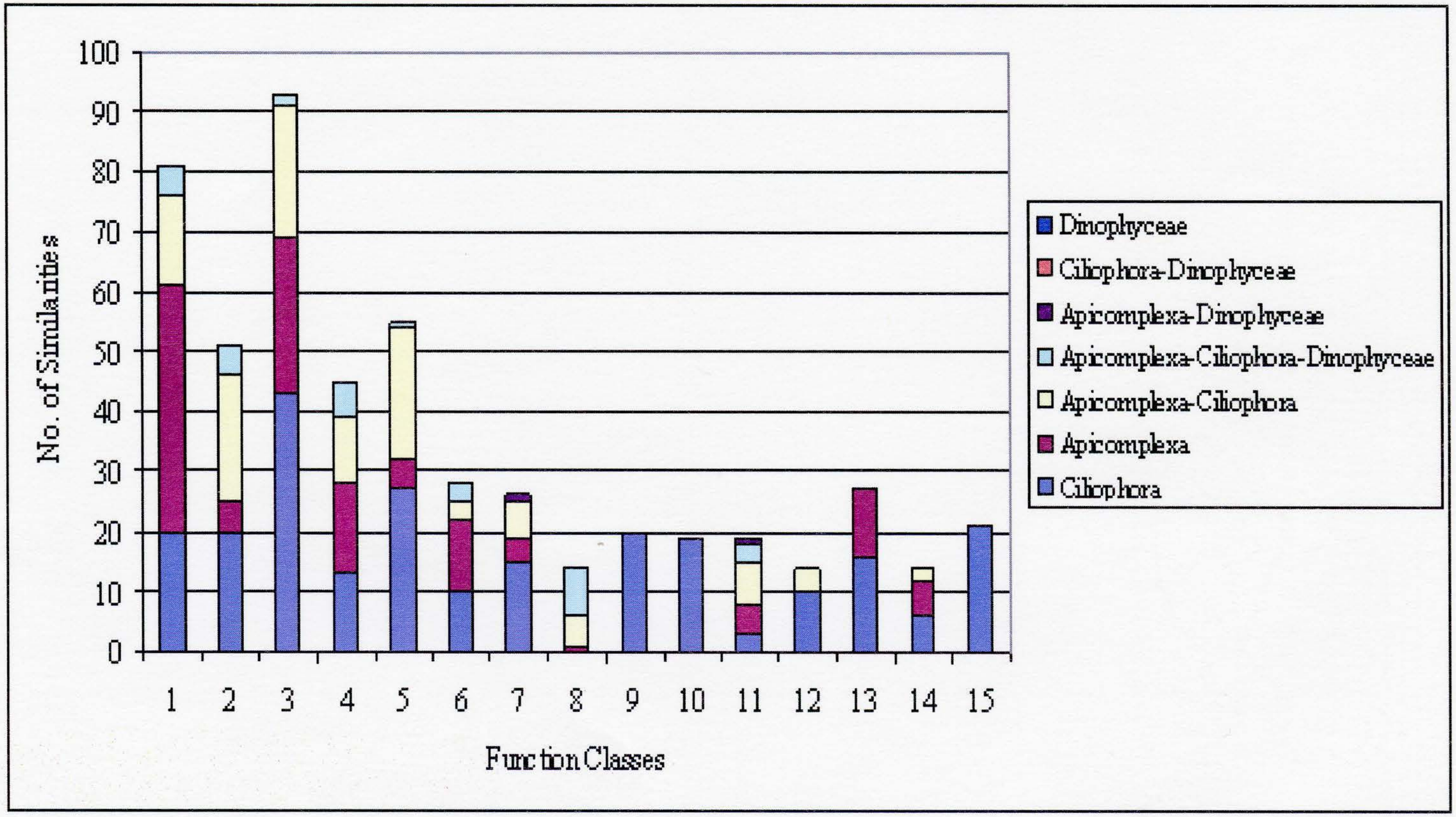
11.signal transduction

12.regulatory functions

13.stress response, detoxification, and cell defence proteins

14.transcription

15.mitochondrial genome


100
90
80
70
60
50
40
30
20
10
0
No. of Similarities
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Function Classes
Dinophyceae
Ciliophora-Dinophyceae
Apicomplexa-Dinophyceae
Apicomplexa-Ciliophora-Dinophyceae
Apicomplexa-Ciliophora
Apicomplexa
Ciliophora

2.4 Summary

Of 850 highly significant matches with an expect value cut-off of less than 10^{-20}, 35.5% represent genes previously cloned from *T. thermophila*, and 64.5% had significant similarity to genes from other organisms deposited in the NCBI.

Among 3740 EST sequences, 1166 sequences are tentatively unique. Of these 479 (41.08%) had significant matches, and 687 (58.92%) displayed no significant match to known proteins. This indicates that a large majority of the genes expressed in *T. thermophila* are unknown in other model organisms and these genes are possibly novel *Tetrahy mena*-specific, ciliate-specific or alveolate-specific genes.

About 53 sequences (6.2%) matched to cytoskeleton proteins which were divided into two groups. The first group matched proteins coding for microtubules, especially to tubulin proteins. The other group matched to microfilament proteins including one actin, three actin-related and one profilin protein. There is no sequence similar to intermediate filaments indicating that intermediate filaments are highly expressed proteins in mammals while microtubules and microfilaments are likely more conserved proteins that exist in *T. thermophila*.

There are 26 sequences (3.1%) that matched signal transduction proteins including Rac, Ras, MAPK, ERK1, PKC, cAMP and 14-3-3 (a protein involved in signal transduction, exocytosis and cell cycle regulation). This result indicates that the *T. thermophila* genome likely encodes the MAPK/ERK signaling pathway.

Comparison of the EST counts from a certain gene provide absolute estimates of mRNA expression levels. The most abundant genes in *T. thermophila* represented are enolase, SerH3 and tubulin.

There are 201 sequences restricted to the Alveolata, of which 196 were specifically restricted to the ciliophora. Only 1.6% of the 850 sequences has best non-self (not from *T. thermophila*) similarities to another ciliophora. Granule lattice genes, SerH and tetrin genes are ciliate-specific genes.

A total of 508 sequences gave highly confident matches to humans for a variety of proteins. Among them, 189 sequences matched proteins from humans but not *S. cerevisiae.* These include: cathepsin H, L and S; selenophosphate synthetase; calmodulin 2; dynein light chain; KIAA 1181 and KIAA 0643, which are human-specific cDNA with unknown functions.

Based on the distribution of function classes within superkingdoms, *T. thermophila* is a model organism which has abundant Eukaryotic-specific proteins and Prokaryotic-like proteins in the same metabolic pathway. Cytoskeleton, antigen and regulatory function proteins that are present in *T. thermophila* are associated with Eukaryotic features. The three main information processing systems of a cell, translation, transcription and DNA replication, are highly diverged in each of the three domains of organisms.

From the distribution of function classes within kingdoms, a large number of protein sequences in Alveolata are ancestral. These genes are conserved among Alveolata, Fungi, Metazoa and Viridiplantae. Alveolata share more genes with Viridiplantae than with Fungi. Many plant-like genes in *T. thermophila* are not annotated. The metabolic pathways in *T. thermophila* may have some similarities to plants although it is a non-photosynthetic organism.

Based on the analysis of the distribution of significant similarities in Ciliophora, Apicomplexa and Dinophyceae, Ciliophora likely shares more genes with Apicomplexa than with Dinophyceae. The regulated secretion proteins and mitochondrial genome of *T. ther-*

mophila are Ciliophora-specific. The similarities common to Apicomplexa, Ciliophora and Dinophyceae are with different protein classes. These common genes of the three phyla will be useful to further analyze the phylogeny of these groups.

Bibliography

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman (1990). Basic local alignment search tool. *J Mol Biol 215*, 403–10.

Baes, M., S. Huyghe, P. Carmeliet, P. Declercq, D. Collen, G. Mannaerts, and P. Van Veldhoven (2000). Inactivation of the peroxisomal multifunctional protein-2 in mice impedes the degradation of not only 2-methyl-branched fatty acids and bile acid intermediates but also of very long chain fatty acids. *J Biol Chem 275*, 16329–36.

Blackburn, E. and J. Gall (1978). A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in *Tetrahymena* . *J Mol Biol 120*, 33–53.

Bretscher, A. (1991). Microfilament structure and function in the cortical cytoskeleton. *Annu Rev Cell Biol 7*, 337–74.

Brimmer, A. and K. Weber (2000). The cDNA sequences of three tetrins, the structural proteins of the *Tetrahymena* oral filaments, show that they are novel cytoskeletal proteins. *Protist 151*, 171–80.

Brownell, J., J. Zhou, T. Ranalli, R. Kobayashi, D. Edmondson, S. Roth, and C. Allis (1996). *Tetrahymena* histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell 84*, 843–851.

Buhling, F., A. Fengler, W. Brandt, T. Welte, S. Ansorge, and D. Nagler (2000). Review: novel cysteine proteases of the papain family. *Adv Exp Med Biol 477*, 241–54.

Calzone, F., R. Angerer, and M. Gorovsky (1983). Regulation of protein synthesis in Tetrahymena. Quantitative estimates of the parameters determining the rates of protein synthesis in growing, starved, and starved-deciliated cells. *J Biol Chem 258*, 6887–98.

Campbell, N. and J. Reece (2002). *Biology, sixth edition.* Pearson Education, Inc.

Cassidy-Hanley, D., H. Smith, and P. Bruns (1995). A simple, efficient technique for freezing *Tetrahymena thermophila* . *J Eukaryot Microbiol 42*, 510–5.

Cech, T. (1990). Self-splicing and enzymatic activity of an intervening sequence RNA from *Tetrahymena* . *Biosci Rep, Nobel lecture 10*, 239–61.

Chilcoat, N., S. Melia, A. Haddad, and A. Turkewitz (1996). Granule lattice protein 1 (Grl1p), an acidic, calcium-binding protein in *Tetrahymena thermophila* dense-core secretory granules, influences granule size, shape, content organization, and release but not protein sorting or condensation. *J Cell Biol 135*, 1775–87.

Clark, M. (1999). Comparative genomics: the key to understanding the Human Genome Project. *Bioessays 21*, 121–30.

Cobb, M. (1999). MAP kinase pathways. *Prog Biophys Mol Biol 71*, 479–500.

Dacks, J. and W. Doolittle (2001). Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell 107*, 419–25.

Dessen, P., M. Zagulski, R. Gromadka, H. Plattner, R. Kissmehl, E. Meyer, M. Betermier, J. Schultz, J. Linder, R. Pearlman, C. Kung, J. Forney, B. Satir, J. Van Houten, A. Keller, M. Froissard, L. Sperling, and J. Cohen (2001). Paramecium genome survey: a pilot project. *Trends Genet 17*, 306–8.

Dos, R. C., D. Chhabra, M. Kekic, I. Dedova, M. Tsubakihara, D. Berry, and N. Nosworthy (2003). Actin binding proteins: regulation of cytoskeletal microfilaments. *Physiol Rev. 83*, 433–73.

Dress, V., H. Yi, M. Musal, and N. Williams (1992). Tetrin polypeptides are colocalized in the cortex of *Tetrahymena* . *J Struct Biol 108*, 187–94.

Drozdowicz, Y. and P. Rea (2001). Vacuolar H(+) pyrophosphatases: from the evolutionary backwaters into the mainstream. *Trends Plant Sci 6*, 206–11.

Fang, Y., E. Yokota, I. Mabuchi, H. Nakamura, and Y. Ohizumi (1997). Purealin blocks the sliding movement of sea urchin flagellar axonemes by selective inhibition of half the ATPase activity of axonemal dyneins. *Biochemistry 36*, 15561–7.

Fast, N., L. Xue, S. Bingham, and P. Keeling (2002). Re-examining Alveolate Evolution Using Multiple Protein Molecular Phylogenies. *J Eukaryot. Microbiol 49*, 30–37.

Fillingham, J., N. Chilcoat, A. Turkewitz, E. Orias, M. Reith, and R. Pearlman (2002). Analysis of expressed sequence tags (ESTs) in the ciliated protozoan *Tetrahymena thermophila* . *J Eukaryot Microbiol 49*, 99–107.

Garrington, T. and G. Johnson (1999). Organization and regulation of mitogen-activated protein kinase signaling pathways. *Curr Opin Cell Biol 11*, 211–8.

Giancotti, F. and E. Ruoslahti (1999). Integrin signaling. *Science 285*, 1028–32.

Gibbons, I. and A. Rowe (1965). Dynein: a protein with adenosine triphosphatase activity from cilia. *Science 149*, 424–426.

Gish, W. and D. States (1993). Identification of protein coding regions by database similarity search. *Nat Genet 3*, 266–72.

Goffeau, A. (2000). Four years of post-genomic life with 6,000 yeast genes. *FEBS Lett 480*, 37–41.

Gray, M. W. (2001). Genome Atlantic Proposal for a Large-Scale Project Protist EST Program (PEP).

Greider, C. and E. Blackburn (1985). identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell 43*, 405–413.

Gulick, A., B. Hubbard, J. Gerlt, and I. Rayment (2001). Evolution of enzymatic activities in the enolase superfamily: identification of the general acid catalyst in the active site of D-glucarate dehydratase from *Escherichia coli* . *Biochemistry 40*, 10054–62.

Hamilton, E. and E. Orias (2000). Genetically mapping new mutants and cloned genes. *Methods Cell Biol 62*, 265–80.

Hernandez, M. J. (1997). *Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design*. Indianapolis: Pearson Education Corporate Sales Division.

Huang, X. and A. Madan (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research 9*, 868–877.

Kohrl, J., R. Brigelius-Flohe, A. Bock, R. Gartner, O. Meyer, and L. Flohe (2000). Selenium in biology: facts and medical perspectives. *Biol Chem 381*, 849–64.

Kozian, D. and B. Kirschbaum (1999). Comparative gene-expression analysis. *Trends Biotechnol 17*, 73–8.

Kroenke, D. (2002). *Database processing: fundamentals, design & implementation*. New Jersey: Natalie E. Anderson.

Kurasawa, Y., O. Numata, M. Katoh, H. Hirano, J. Chiba, and Y. Watanabe (1992). Identification of *Tetrahymena* 14-nm filament-associated protein as elongation factor 1 alpha. *Exp Cell Res 203*, 251–8.

La Teana, A., C. Gualerzi, and A. Dahlberg (2001). Initiation factor IF 2 binds to the alpha-sarcin loop and helix 89 of *Escherichia coli* 23S ribosomal RNA. *Rna 7*, 1173–9.

LaCrosse, G. and F. Doerder (1994). A temperature-sensitive mutation of the temperature-regulated SerH3 i-antigen gene of *Tetrahymena thermophila* : implications for regulation of mutual exclusion. *Genetics 138*, 297–301.

Lewis, T., P. Shapiro, and N. Ahn (1998). Signal transduction through MAP kinase cascades. *Adv Cancer Res 74*, 49–139.

Li, L., B. Brunk, J. Kissinger, D. Pape, K. Tang, R. Cole, J. Martin, T. Wylie, M. Dante, S. Fogarty, D. Howe, P. Liberator, C. Diaz, J. Anderson, M. White, M. Jerome, E. Johnson, J. Radke, C. J. Stoeckert, . Waterston RH, S. Clifton, D. Roos, and L. Sibley (2003). Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Res 13*, 443–54.

Liang, F., I. Holt, G. Pertea, S. Karamycheva, S. Salzberg, and J. Quackenbush (2000). An optimized protocol for analysis of EST sequences. *Nucleic Acids Res 28*, 3657–65.

McKean, P., S. Vaughan, and K. Gull (2001). The extended tubulin superfamily. *J Cell Sci 114*, 2723–33.

Modrek, B. and C. Lee (2002). A genomic view of alternative splicing. *Nat Genet 30*, 13–9.

Orias, E. (1998). Mapping the germ-line and somatic genomes of a ciliated protozoan, *Tetrahymena thermophila* . *Genome Res 8*, 91–9.

Orias, E., E. Hamilton, and J. Orias (2000). *Tetrahymena* as a laboratory organism: useful strains, cell culture, and cell line maintenance. *Methods Cell Biol 62*, 189–

211.

Pandey, A. and F. Lewitter (1999). Nucleotide sequence database: a gold mine for biologists. *Trends Biochem Sci 24*, 276–80.

Prade, R., P. Ayoubi, S. Krishnan, S. Macwana, and H. Russell (2001). Accumulation of stress and inducer-dependent plant-cell-wall-degrading enzymes during asexual development in *Aspergillus nidulans* . *Genetics 157*, 957–67.

Reinheckel, T., J. Deussing, W. Roth, and C. Peters (2001). Towards specific functions of lysosomal cysteine peptidases: phenotypes of mice deficient for cathepsin B or cathepsin L. *Biol Chem 382*, 735–41.

Shen, X., L. Yu, J. Weir, and M. Gorovsky (1995). Linker histones are not essential and affect chromatin condensation in vivo. *Cell 82*, 47–56.

States, D. and P. Agarwal (1996). Compact encoding strategies for DNA sequence similarity search. *Proc Int Conf Intell Syst Mol Biol 4*, 211–7.

Taylor, F. (1987). The biology of dinoflagellates. *Botanical Monographs 21*, Blackwell Scientific Publicatons, Oxford.

Turkewitz, A., E. Orias, and G. Kapler (2002). Functional genomics: the coming of age for *Tetrahymena thermophila* . *Trends Genet 18*, 35–40.

Van Eldik, L. J. and D. M. Watterston (1998). *Calmodulin and Signal Transduction*. Academic Press.

Welling, L. and L. Thomson (2001). *PHP and MySQL Web Development*. Indianapolis, Indiana: Sams.

Wickert, S. and E. Orias (2000). *Tetrahymena* micronuclear genome mapping. a high-resolution meiotic map of chromosome 1l. *Genetics 154*, 1141–53.

Yarger, R. J., G. Reese, and T. King (1999). *MySQL and mSQL.* Sebastopol: O'Reilly & Associates, Inc.