GUI Interface to Biological Databases

A dissertation submitted to Jawaharlal Nehru University in partial fulfillment of the requirements for the award of the degree of

Master of Technology
in
Computer Science & Technology

By

B.DAMODAR

Under the guidance of

Prof. Parimala. N

School of Computer & Systems Sciences
Jawaharlal Nehru University
New Delhi 110067

JANUARY 2003

जवाहरलाल नेहरू विश्वविद्यालय

SCHOOL OF COMPUTER & SYSTEM SCIENCES
JAWAHARLAL NEHRU UNIVERSITY
NEW DELHI – 110067 (INDIA)

CERTIFICATE

This is to certify that the dissertation titled "**GUI Interface to Biological Databases**" which is being submitted by **Mr. B. Damodar** to the School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi, in partial fulfillment of the requirements for the award of **Master of Technology in Computer Science & Technology** is a bonafide work carried out by him under the supervision of **Prof. Parimala. N.** The matter embodied in the dissertation has not been submitted for the award of any other degree or diploma.

06-01-03

Prof. K.K. Bharadwaj
Dean, SC & SS
Jawaharlal Nehru University
New Delhi 110067

Professor K. K. Bharadwaj
Dean
School of Computer & Systems Sciences
Jawaharlal Nehru University
New Delhi - 110067

Parimala N.

Prof. Parimala. N.
SC & SS
Jawaharlal Nehru University
New Delhi 110067

ACKNOWLEDGEMENTS

I wish to convey my heartful gratitude and sincere acknowledgement to my supervisor Prof. Parimala. N for her constant encouragement, guidance and affection throughout my work.

I would like to thank Prof. K.K. Bharadwaj, Dean, School of Computers and Systems Sciences (SC&SS), JNU for providing excellent lab facilities.

I extend my thanks to Dr. Andrew, Bioinformatics Center (BIC) JNU for his cooperation and assistance in successful completion of the project.

I would like to thank all my faculty members for their help and useful suggestions during my work. I am also thankful to Mr. V. Praveen Kumar Reddy, Mr. V. Mallikarjuna Chetty, Mr. Binod Singh and my classmates specially Miss Neetu Rawat for their timely help and valuable suggestion.

B. DAMODAR

ABSTRACT

Designing a query Interface for Protein Database The project provides user-friendly query interface for Swiss_prot, Prosite & PDB databases. Data Dictionary is developed on user term to provide easy access to database, user can specify complex condition on user terms. Graphical Representation of query is presented as cross-reference between the databases. Option for saving the query, opening the saved query, printing results on printer, online help about project and databases is provided.

CONTENTS

CHAPTER 1

Introduction

In the recent past, splendid achievements and numerous scientific advancements have been made in the field of human genome and protein structure. The Biological sciences field has made it possible to solve and analyze biological problems of long bandwidth spectrum which has resulted in the rapid growth of Biological databases in Bioinformatics.

Mostly scientist groups and different research institutes have developed their own databases. For example Sequence database is maintained by National center for Biotechnology information, Protein database is maintained by Research Collaboratory for Structural Bioinformatics, Swiss Institute of Bioinformatics is maintains by Swiss_prot and Prosite. Some of the universities have their own information retrieval tools for Biological database.

The databases are now available on the web for facilitating the research work of a diverse biological field to access these different databases that are available on the web and web based user interface systems have become developed which are used by accessible to Researchers Institutes and Universities.

Software packages such as Basic Local Alignment Search Tool (BLAST) [23] which is used as a search tool to help in training people in universities, colleges, and technical institute has been developed.

1.2 Heterogeneous database

As mentioned above, as different research institute maintains the databases, they are heterogeneous in nature with different configuration and with different sets of properties. Some of the major types of biological database are discussed below,

1.2.1. Sequence database

The sequence database is widely used biological databases. It is divided into nucleic acid and protein sequence databases as there exist a relationship between a protein sequence and a nucleic acid sequence. This relationship is captured as database references between nucleic acid entries and protein entries. Some of the Nucleic Acid Sequence databases are EMBL [7]/ GenBank [8] / DDBJ [19] and Protein Sequence databases SWISSPROT [3] / TrEMBL [24] / PIR [11]

1.2.2. Structural Database

Structure database is applicable to only Proteins. The structural database is three-dimensional structure of protein. Sequence database has primary and secondary structure data obtained. Protein Database (PDB) [5] contains 3D-structure information of proteins.

1.2.3 Genome Database

Genome databases provides views for variety of genomes and complete chromosomes. Genome Database (GDB) [16] is example of Genome Databases.

1.3 Existing Databases Formats

The Biological databases have their own form of format to store data. As different research institutes are maintaining databases they retained their initial format in which they have been developed. Some of them are listed as follows.

1.3.1 Flat file format

Flat files are data files that contain records with no structured relationships. The entries are stored in text form. The text fields are labeled with identifiers. The index on identifier is used to search a text in the file for faster retrieval. For example GenBank [8] exists in flat file format.

1.3.2 Relational database

A Relation database is a predefined row/column format for storing information. Relations are equivalent to tables. The collection of tables represents the information represented in the database. For example Genomone Database (GDB) [16] stores in relational database.

1.3.3 Object oriented database

Data is organized into a hierarchy of concept or classes. Classes have set of attributes. Classes can inherit attributes from parents in the hierarchy. For example AceDB [24] is stored in Object oriented database form.

1.4 Existing data Retrieval Systems

There are graphical user interface retrieval systems available on web to fetch the data from multiple databases. Some of the strategies of existing retrieval system discussed below.

1.4.1 HyperText Navigation

This allows users to interactively navigate from one database entry to another database by transferring links between the databases. Searching within one database to find a starting entry and then requesting a linked entry from another database for example retrieving a GenBank [8] entry using a protein name. This approach information- retrieval system is implemented to provide fast indexes access to flat-file database. For example this system is implemented in SRS [22], the ExPAsy [26] web server and Genome Net [17]. We discuss about SRS [22] and ExPASy [26] retrieval system in next section.

1.4.1.1 SRS (Sequence retrieval systems)

SRS is a web-based integrated system that provides data retrieval and application for homogeneous data analysis to all flat file data banks. It provides search from multiple database by shared attributes and to query across database fast and efficiently. SRS is the easiest and simplest method available to quickly access data from multiple databases

SRS allows web based searching and retrieval of nucleotide and protein sequence. It also allows user to query most of the major bioinformatics databases and retrieve textual information.

The Sequence Retrieval System retrieve the data based on database indices. SRS contains indices for nearly 100 databases that can be searched. The search is specifies to a single index or a group of indices of one or more databanks simple and complex searches are easy to do. For example the database search by SRS [22] are Swiss_prot [3], TrEMBL [25], Trembl_new [25], Swiss2D [3], Prosite [4] and Enzyme [27].

1.4.1.2 ExPAsy

ExPAsy (Expert Protein Analysis System) is a World Wide Web server, which is provided as a service to the Life Sciences community. Its main focus is on proteins. It provides access to a variety of databases and analytical tools dedicated to what is now known as proteomics. It is developed at the Swiss Institute of Bioinformatics (SIB)[2].

A variety of access options are available for each database. These options allow the users to display and retrieve specified subsets of the database. For example, SWISS-PROT [3] and TrEMBL [25] options that allows searching by description, accession number, author, and citation or by full text search.

A large variety of documents (user's manual, release notes, indices, nomenclature documents, etc.) are available for each databases All the databases available on ExPASy [26] are extensively cross-referenced to other molecular biology databases or resources all over the world. For example SWISS-PROT [3] is cross-referenced to more than 50 different databases such as: EMBL [7]/GenBank []/DDBJ [19], PDB [5], MEDLINE/PubMed [15], EcoGene [17] etc. The databases are frequently updated The database SWISS-PROT [3] knowledgebase, SWISS-2DPAGE [3], Prosite [4], Enzyme [26], Swiss-3dImage [3], Swiss Model Repository [3] and CD40Lbase [28] are access through ExPASY [26].

Over the years an extensive collection of software tools has been developed most of which are either targeted toward the access and display of the databases which are used to analyze protein sequences and proteomics data.

These tools such as Compute pI/MW, Translate, SWISS-MODEL [3] can all be accessed from ExPASy [26] some of them are listed as follows.

Compute pI/MW: computes the theoretical isoelectric point (pI) and molecular weight (MW) from a SWISS-PROT [3] or TrEMBL [25] entry or for a user sequence.

Translate translates a nucleotide sequence to a protein.

Swiss-Model an automated knowledge-based protein modeling server. It is able to build models of the three-dimensional structure of proteins whose sequence is closely related to that of proteins with known 3D structure.

1.4.2 Unmediated MultiDatabase Queries

This approach allows user to construct complex queries that is evaluated against multiple, Physically distinct and heterogeneous database, A query explicitly identifies both member database that is applied to tables and attributes that are to be queried within each database, that is a single query can include reference to several database. This approach is included in Kleisli include a query language CPL.

1.4.2.1 Kleisli

A principal novelty of this system was the query and transformation language CPL. Based on the principle that database query languages can be constructed from some fundamental operations used in the types used in Specification of a database, CPL can be used against free combinations of tuple, Variant, set, mustiest, list and array types. It naturally extends the relational algebra to these types, based on a formal foundation grounded in the mathematical theory of categories. The language and optimization techniques have been implemented in the Kleisli system, which provides generic access to a wide variety of types of external data sources through functions registered within the Kleisli library.

Kleisli has the ability to specify transformations involving complex datatypes found throughout biomedical data applications and the ability to specify transformations in a partial, step-wise manner. The ability to partially specify transformations is very useful, as data sources are large and complex, and frequently difficult to understand in their entirety. The system has generic interfaces to different relational databases, such as Oracle, Sybase, ASN.1, object-oriented database.

Kleisli has been deployed with considerable success for bioinformatics support within the Human Genome Project. In particular, Kleisli has been used to answer a number of queries claimed to be unanswerable "until a fully relationalized sequence database is available" in a 1993 meeting report published by the Department of Energy. The Kleisli technology has been incorporated into commercial products.

1.5 Difficulty with Existing approach

As discussed in section 1.3 databases exist in different formats and the existing retrieval systems are database specific for example ExPAsy [26] is for only Protein Databases and SRS [22] are only for homogenous databases.

In system like ExPASy [26] and SRS [22] user has to explicitly select the database from which data is to be fetch .The user has to know about the database and their attributes, it is difficult to remember database attributes for each database. It is possible user want to fetch data from all the database for selected attributes in this case user has to select all the database or should know in which database required attribute exists.

User need to perform logical operation such as AND, OR and NOT to fetch information from several database this requirement does not met with the existing retrievals system like Sequence Retrieval Systems (SRS) [22], ENTREZ [] by NCBI, ExPAsy [21] WWW server by Bairoch. This places a burden on the user.

The hypertext navigation approach is implemented in systems like SRS [22], ENTREZ [29] where user must navigate from one database to other to know the information. The user has to know the cross-references that exist between the database.

No option for query refining is available i.e.; user cannot perform the previously used query by adding/removing attributes and adding /removing condition on the query.

1.6 Proposed Approach

In our approach the user is provided with query Interface to access Protein Databases The Protein databases are Swiss_Prot [3], Prosite [4], and Protein Database (PDB) [5] represent different information on protein. Swiss_Prot [3] and Prosite [4] are sequence database where as PDB [5] is structural database.

The database is implicitly selected for the query processing and execution. Data Dictionary is developed for database term this s help user to select data from any of databases. As database term are difficult to remember for the user.

After selecting user term and the condition specification, query is submitted for the processing. The query is decomposed into multiple sub query access information from individual databases. To make search faster index are created on each database.

The database are regularly updated on web so we are directly reading the database from the ftp site of the corresponding database rather than having a copy of database on local system.The data is fetched from different databases and output is presented to the user in three forms they are LIST, TABLE and STRUCTURAL form. He structural form is the 3D structure of protein.

Refinement of queries in terms of adding/removing an attribute and Adding/removing/changing the condition is required from the user point of view to enhance or narrow down the search.

The layout of this thesis is as follows. In Chapter 2 we describe the Protein databases and their structure. Here three Proteins database i.e. SWISS-PROT, Prosite and PDB are explained. In Chapter 3 we discuss Design to GUI Interface for the system and Data Dictionary for three databases. In Chapter 4 design of system is presented in the form of structure chart. In Chapter 5 implementation of GUI Query Interface is discussed. In Appendix A sample entry for Swiss_Prot, In Appendix B sample entry for PDB, In Appendix C sample entry for Prosite is given.

CHAPTER 2

Protein Database

In this project we are concerned with three Protein Databases, i.e., Protein database (PDB)[5], Swiss_Prot [3] and Prosite [4] to present list, table and structural information of Proteins. Each of these databases represents different information of the Proteins. Each database and their formats are discussed below

2.1 Swiss_Prot

SWISS-PROT is a protein knowledgebase established in 1986 and maintained collaboratively, since 1987, by the Department of Medical chemistry of the University of Geneva (now the Swiss Institute of Bioinformatics (SIB)) [2] and the EMBL [7] Data Library (now the EMBL [7] Outstation The European Bioinformatics Institute (EBI)) [1]. The SWISS-PROT [3] protein knowledgebase consists of sequence entries.

Sequence entries are composed of different line-types, each with their own format. For standardization purposes the format of SWISS-PROT [3] follows as closely as possible that of the EMBL [7] Nucleotide Sequence Database.

The SWISS-PROT [3] Protein Sequence Database is a database of protein sequences It distinguishes itself from other protein sequence databases by three distinct criteria. They are Annotation, Minimal redundancy, Integration with other databases. Each of them is discussed below.

2.1.1 Annotation

In SWISS-PROT [3], as in most other sequence databases, two classes of data can be distinguished the core data and the annotation. For each sequence entry the core data consists of the sequence data the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein) while the annotation consists of the description of the following.

Function(s) of the protein, Post-translational modification(s) for example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc, Domains and sites for example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc, Secondary structure, Quaternary structure for example homodimer, heterotrimer, etc, Similarities to other proteins, Diseases associated with deficiencies in the protein and Sequence conflicts, variants, etc.

2.1.2 Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. It is possible to merge all these data so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

2.1.3 Integration with other databases

It is important to provide the users of bimolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialized data collections. SWISS- PROT [3] is currently cross-referenced with about 60 different databases. Cross-references are provided in the form of pointers to information related to SWISS-PROT [3] entries and found in other databases.

Each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order, in which they appear in an entry, are shown in the table below. A sample sequence entry of this Swiss-prot is shown in Appendix A.

Line code	Content	Occurrence in an entry
ID	Identification	Once, starts the entry
AC	Accession number	Once or more
DT	Date	Three times
DE	Description	Once or more
GN	Gene name	Optional
OS	Organism species	Once or more
OG	Organelle	Optional
OC	Organism classification	Once or more
OX	Taxonomy	Once or more cross-reference(s)
RN	Reference number	Once or more
RP	Reference position	Once or more
RC	Reference comment	Optional
RX	Reference	Optional
RA	Reference authors	Once or more
RT	Reference title	Optional
RL	Reference location	Once or more
CC	Comments or notes	Optional
DR	Database	Optional
KW	Keywords	Optional
FT	Feature table data	Optional
SQ	Sequence header	Once
(Blanks)	Sequence data	Once or more
//	Termination lines	Once ends

2.2 Prosite

PROSITE [4] is a database of protein families and domains. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor.

PROSITE [4] is a method of determining what is the function of uncharacterized proteins translated from genomic or cDNA sequences. It consists of a database of biologically significant sites and patterns formulated in such a way that with appropriate computational tools it can rapidly and reliably identify to which known family of protein the new sequence belongs. The use of protein sequence patterns to determine the function of proteins is becoming very rapidly one of the essential tools of sequence analysis. PROSITE [4] contains patterns and profiles specific for more than a thousand protein families or domains

The entries in the database are structured so as to be usable by human readers as well as by computer programs. Each entry in the database is composed of lines. Different types of lines, each with its own format, are used to record the various types of data, which make up the entry. The general structure of a line is the following

Characters	Content
1 to 2	Two-character line code. Indicates the type of information contained in the line.
3 to 5	Blank
6 up to 128	Data

Each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order, in which they appear in an entry, are shown in the table below. A sample sequence entry of Prosite is shown in Appendix C.

LineCode	Content	Occurrence in an entry
ID	Identification	Once, start the entry
AC	Accession number	Once
DT	Date	Once
DE	Short description	Once
PA	Pattern	Optional
MA	Matrix/profile	Optional
RU	Rule	Optional
NR	Numerical results	Optional
CC	Comments	Optional
DR	Cross-references to SWISS-PROT	Optional
3D	Cross-references to PDB	Optional
DO	Pointer to the documentation file	Once
//	Termination line	Once endentry

2.3 Protein Database (PDB)

Protein Database is most prominent Protein Database. The PDB [5] is the largest repository for 3D-protein structure determined by X-ray crystallography or nuclear magnetic resonance (NMR) and contains examples of all known unique protein families.

The PDB [5] file may also be viewed as a collection of record types. Each record type consists of one or more lines. Each record type is further divided into fields. The description of each record type includes the sections Over View, Record Format, Details, Verification/Validation/Value Authority Control, Relationship to other Record Types, Example, and Known Problems.

The PDB [5] file has a number of lines terminated by an end-of-line indicator. Each line in the PDB [5] entry file consists of 80 columns. The last character in each PDB [5] entry should be an end-of-line indicator. Each line in the PDB [5] file is self-identifying. The first six columns of every line contain a record name, left justified and blank-filled. This must be an exact match to one of the stated record names.

The currently used line types, along with their respective line codes, are listed below: A sample sequence entry of PDB is shown in Appendix B.

LineCode	Content
CRYST1	Unit cell parameters, space group, and Z.
END	Last record in the file.
HEADER	First line of the entry, contains
PDB ID	code, classification, and date of deposition.
MASTER	Control records for bookkeeping.
ORIGXn	Transformation from orthogonal coordinates to the submitted coordinates (n = 1, 2, or 3).
SCALEn	Transformation from orthogonal coordinates to fractional crystallographic coordinates (n = 1, 2, or 3).
AUTHOR	List of contributors.
CAVEAT	Severe error indicator. Entries with this record must be used.
COMPND	Description of macromolecular contents of the entry.
EXPDTA	Experimental technique used for the structure determination.
KEYWDS	List of keywords describing the macromolecule.
OBSLTE	Statement that the entry has been removed from distribution and list of the ID code(s) which replaced it.
SOURCE	Biological source of macromolecules in the entry.
SPRSDE	List of entries withdrawn from release and replaced
TITLE	Description of the experiment represented in the entry.
ANISOU	Anisotropic temperature factors.
ATOM	Atomic coordinate records for standard groups.
CISPEP	Identification of peptide residues in cis conformation.

CONECT	Connectivity records.
DBREF	Reference to the entry in the sequence database(s).
HELIX	Identification of helical substructures.
HET	Identification of non-standard groups or residues (heterogens)
HETSYN	Synonymous compound names for heterogens.
HYDBND	Identification of hydrogen bonds.
LINK	Identification of inter-residue bonds.
MODRES	Identification of modifications to standard residues.
MTRIXn	Transformations expressing non-crystallographic symmetry (n = 1, 2, or 3). There may be multiple sets of these records.
REVDAT	Revision date and related information.
SEQADV	Identification of conflicts between PDB and the named sequence database.
SEQRES	Primary sequence of backbone residues.
SHEET	Identification of sheet substructures.
SIGATM	Standard deviations of atomic parameters.
SIGUIJ	Standard deviations of anisotropic temperature factors.
SITE	Identification of groups comprising important sites.
SLTBRG	Identification of salt bridges SSBOND Identification of disulfide bonds.
TURN	Identification of turns.
TVECT	Translation vector for infinite covalently connected structures.
FORMUL	Chemical formula of non-standard groups.
HETATM	Atomic coordinate records for heterogens.
HETNAM	Compound name of the heterogens.
ENDMDL	End-of-model record for multiple structures in a single coordinate
MODEL	Specification of model number for multiple structures in a single coordinate entry.

TER	Chain terminator.
JRNL	Literature citation that defines the coordinate set.
REMARK	General remarks, some are structured and some are free form.

For records that are fully described in fixed column format, columns not assigned to fields must be left blank.

CHAPTER 3

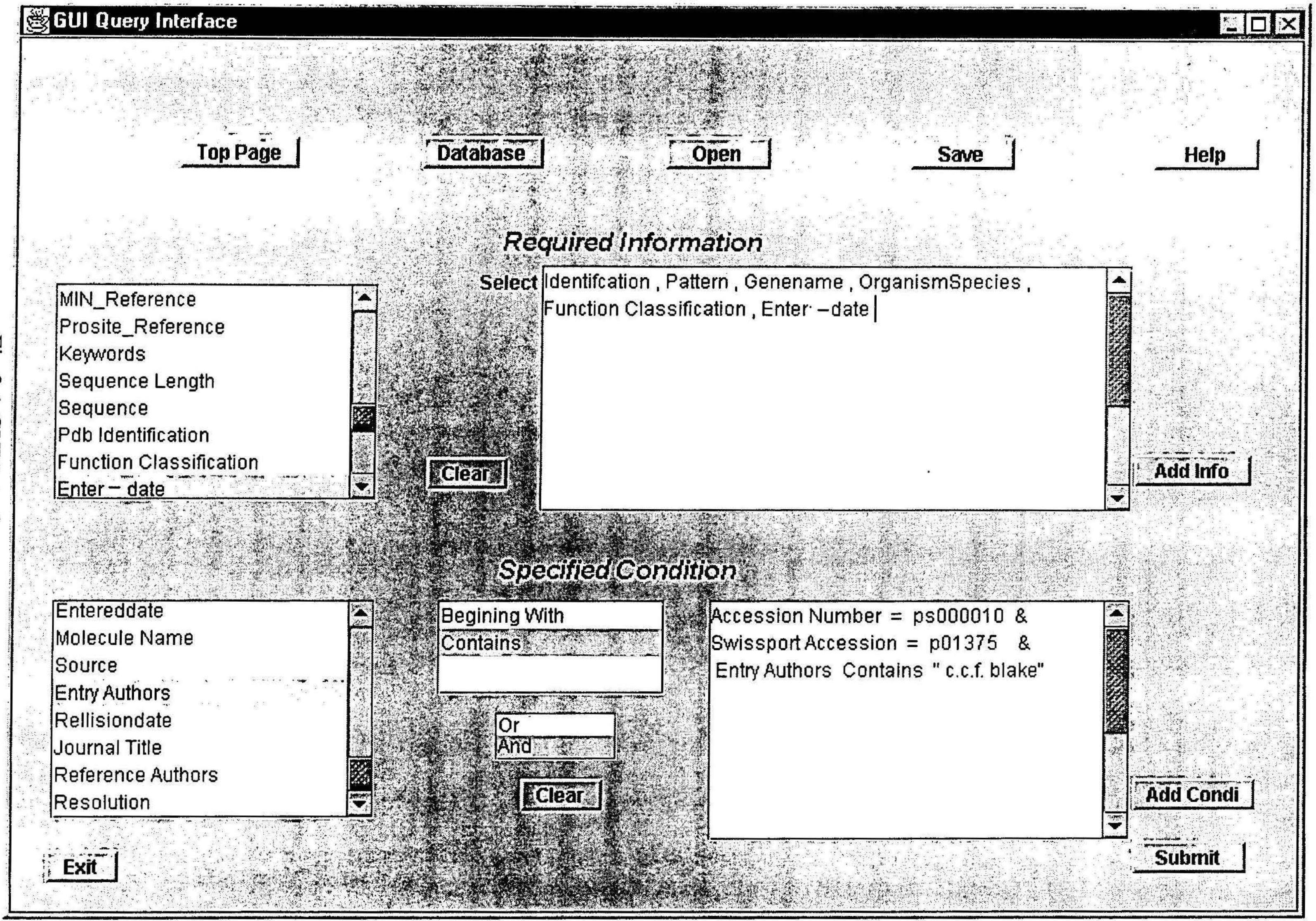
Design of GUI Interface

Our project on Protein Database provides GUI query Interface for accessing database from the web and displaying the result in structural and textual format. In this chapter we discuss about GUI Query Interface design.

GUI query Interface is developed. The basic requirement for accessing data from the database is GUI query interface must be user-friendly. Biological databases are very large database term are difficult to remember so database terms is provides with user-friendly terms. User needn't be aware of database terms. Data dictionary is developed with user terms for all the database terms.

The user can specify condition in two aspects one for numerical attribute with relational operators or string attribute with string matching. User can specify complex queries using logical operator. The graphical representation of the user query is provided so as to know from which database the attributes are selected and corresponding database and relationship between the database that is cross-reference.

An option for saving query with condition specification is provided for the user to save on local computer with extension *.doc file The save query can be retrieve from *.doc file with Open option. The three forms can be viewed on the printer with Printer option. The query is processed and executed in Query Processing Retrieval System.



MIN_Reference
Prosite_Reference
Keywords
Sequence Length
Sequence
Pdb Identification
Function Classification
Enter – date

Select Identifcation , Pattern , Genename , OrganismSpecies , Function Classification , Enter –date

Entereddate
Molecule Name
Source
Entry Authors
Rellisiondate
Journal Title
Reference Authors
Resolution

Begining With
Contains

Or
And

Accession Number = ps000010 &
Swissport Accession = p01375 &
Entry Authors Contains " c.c.f. blake"

Fig 3.1 GUI Interface

3.1 SYSTEM IN USE

Fig 3.1 is explained below.

TopPage is a Button. It refreshes the screen.

Save is a Button. The given query with attribute and condition specification can be saved in local computer with extension *.doc file It is shown in fig 3.3

Open is a Button. The saved query with condition specification store in file is opened. The data is fetch from the file and is displayed in attribute textbox and condition textbox. It is shown in Fig 3.4

Help is a Button. It gives online help go to through the project. It is shown in Fig 3.6

Database is a Button. It gives online help about databases and their format. It is shown in fig 3.5

Attribute ListBox is ListBox it has attributes that to be selected for performing a query.

Select is a Label

Clear is a Button. It clears the Attribute TextBox.

Attribute TextBox is a TextBox. The attribute selected in attribute ListBox are displayed.

Add Info is a button Default it is disabled, For refining of query it is enabled.
It allows appending the attribute to the existing Attribute TextBox.

Condition ListBox is a ListBox. The attributes are selected for the condition specification.

Operator ListBox is a ListBox. When user select numeric attribute from Condition ListBox relational operator (<, <=, >, >=, =) are displayed and string containing attribute is selected "Beginning with", "Contain" are displayed.

Logical ListBox is a ListBox. Logical operator &&, || are displayed.

Condition TextBox is a ListBox. The attribute selected from Condition ListBox is displayed.

Add Condition is a Button default it is disabled, For refining of query it is enabled. It allows appending the condition attribute to the existing Condition Attribute TextBox.

Submit is a Button the query is submitting for the processing.

Exit is Button It is end of program.

Example of the query:

Select Identification, Pattern, Genename, Organism Species, Function Classification, Enter-date.

Condition:

Accession Number = ps000010 & Swissport Accession = p01375 &
Entry Authors Contains " c.c.f. blake"

3.2 Developing Data Dictionary

The protein database is very large and it is difficult for the user to remember all the database terms. In all the database attribute used in the database are not unique and the same information is represented with different names in different database. It is expected that the user remember all the attribute names and their meaning while fetching data from these databases.

It provides all the data attribute in user friendly terms so that he can select data using terms with which user is familiar. In this interface user need not be aware of from which database user is extracting the data or which are the available databases.

To build the query user need not be aware of the data that is present in the existing databases. A data dictionary is built for this with user terms for all the available data so that user can select what he wants.

Data Dictionary is developed for column attribute and condition attribute. Both are discussed below.

3.2.1 Data Dictionary for Required Information.

User term corresponding to database for Prosite, Swiss_Prot and PDB in column attribute are given below.

PROSTITE

Database term	**User Term**
ID	Prosite_Identification
AC	Prosite_AccNumber
DT	Created_Date, Updated_Date
DE	Short description
PA	Pattern
MA	Matrix
RU	Rule
NR	Numerical results
DR	Swiss_Prot_Reference
3D	PDB_Reference
DO	Doc_Reference

SWISS_PROT

Database term	**User Term**
ID	Swissprot Identification
AC	Swissport Accession
DT	Swissprot_createddate, Swissprot_sequenceupdateddate, Swissprot_Annotationupdateddate
GN	Genename,
OS	Organism Species

OC	Organism Classification
RA	Author
DR	EMBL_Reference, PIR_Reference, SPdb_Reference,MIN_Reference, Prosite_Reference
KW	Keywords
SQ	Sequence Length Sequence

PDB

Database Term	**User Term**
HEADER	Pdb_Identification, Function Classification, Enter date
AUTHOR	Entry Authors
COMPND	Molecule name
HELIX	Helix
SOURCE	Source
SHEET	Sheet
REVDAT	Revision date
TURN	Identification of turns.
JRNL	Journal tile, Reference Author
REMARK	Resolution
SCALEX	Scaling Information

3.2.2 Data Dictionary for Condition Specification

User terms corresponding to database for Prosite, Swiss_Prot and PDB in condition specification are given below.

PROSTITE

Database term	User Term
ID	Identification
AC	AccNumber
DT	Created_Date, Updated_Date
DE	Short description
PA	Pattern
DR	Swiss_Prot_Reference
3D	Pdb_Reference
DO	Doc_Reference

SWISS_PROT

Database term	User Term
ID	Swissprot Identification
AC	Swissport Accession
DT	Swissprot_createddate, Swissprot_sequenceupdateddate, Swissprot_Annotationupdateddate
GN	Genename,
OS	Organism Species
OC	Organism Classification

RA	Author
DR	EMBL_Reference, PIR_Reference, SPdb_Reference,MIN_Reference, Prosite_Reference.
KW	Keywords
SQ	Sequence Length Sequence

PDB

Database Term	User Term
HEADER	Pdb_Identification, Function Classification, Enter-date
AUTHOR	Entry Authors
SOURCE	Source
COMPND	Molecule name
SOURCE	Source
REVDAT	Revision date
JRNL	Journal tile, Reference Author
REMARK	Resolution

3.3 Graphical Representation of query

After developing the query a graphical display of query is presented for the user to know from which database required information is to be extracted. Graphical Representation for example query in page 27 is shown Fig 3.2

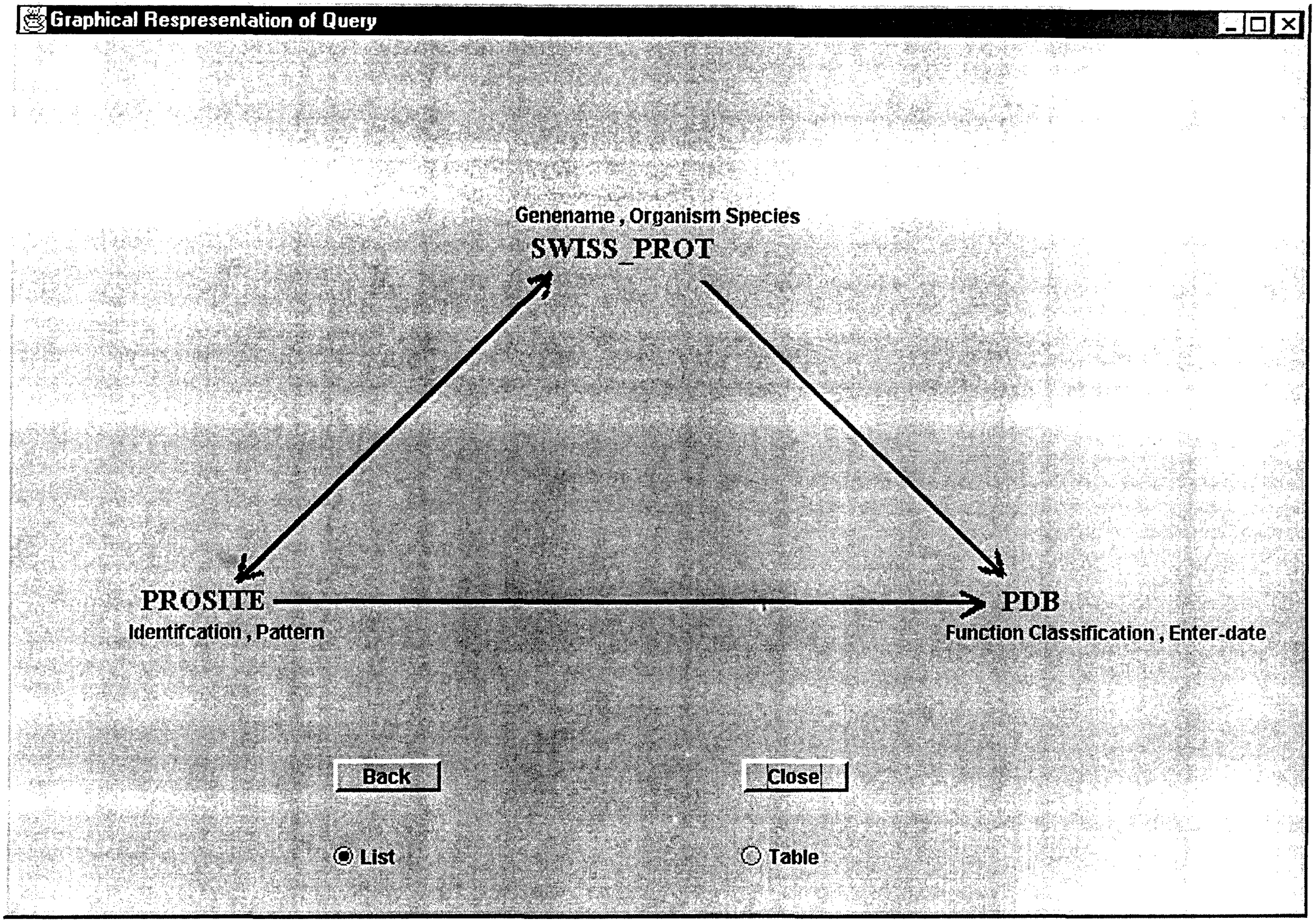

Graphical Respresentation of Query
Genename , Organism Species
SWISS_PROT
PROSITE
Identifcation , Pattern
PDB
Function Classification , Enter-date
Back
Close
● List
○ Table

Fig 3.2 Graphical Display of Query.

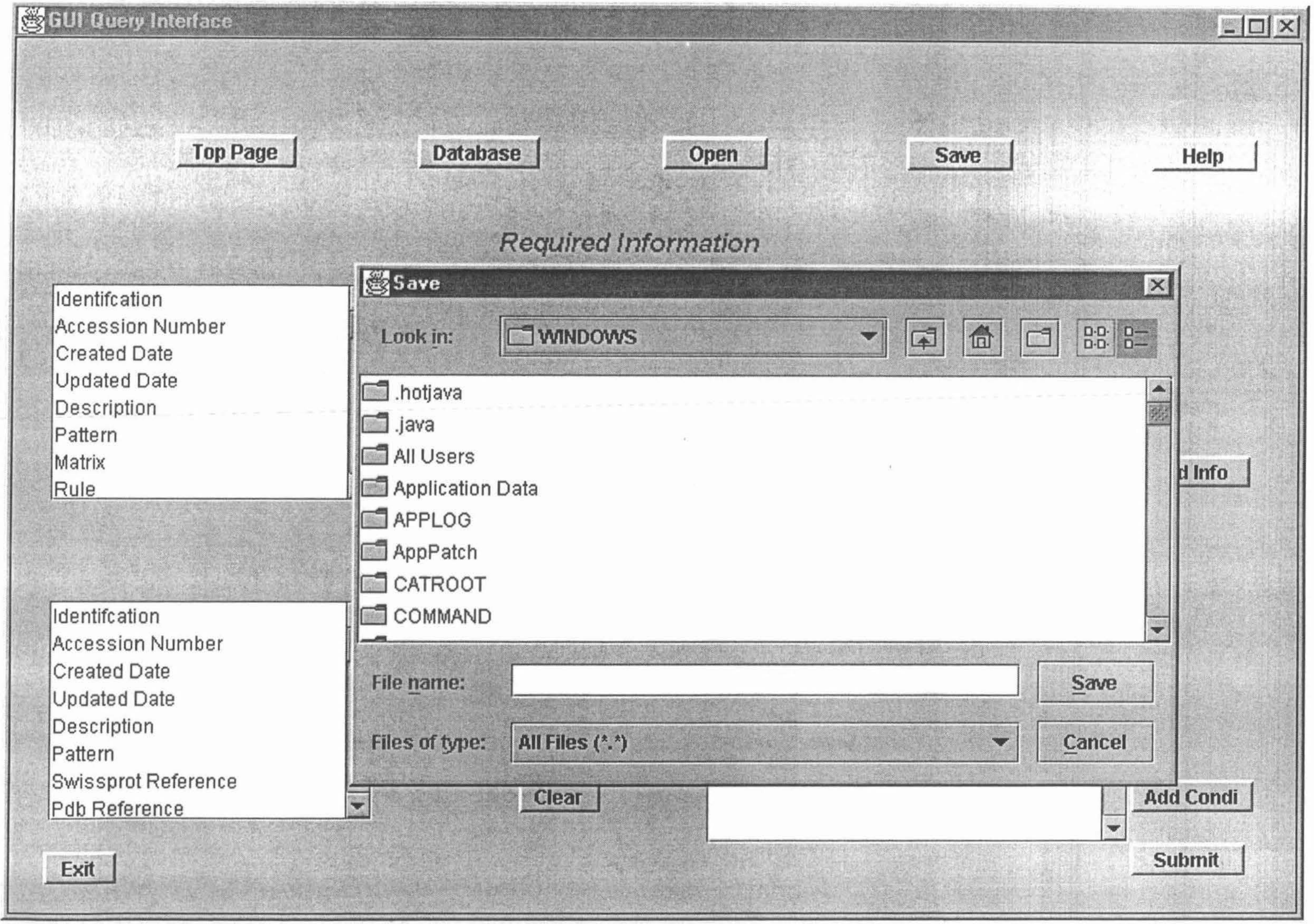


Fig 3.3 Saving a Query

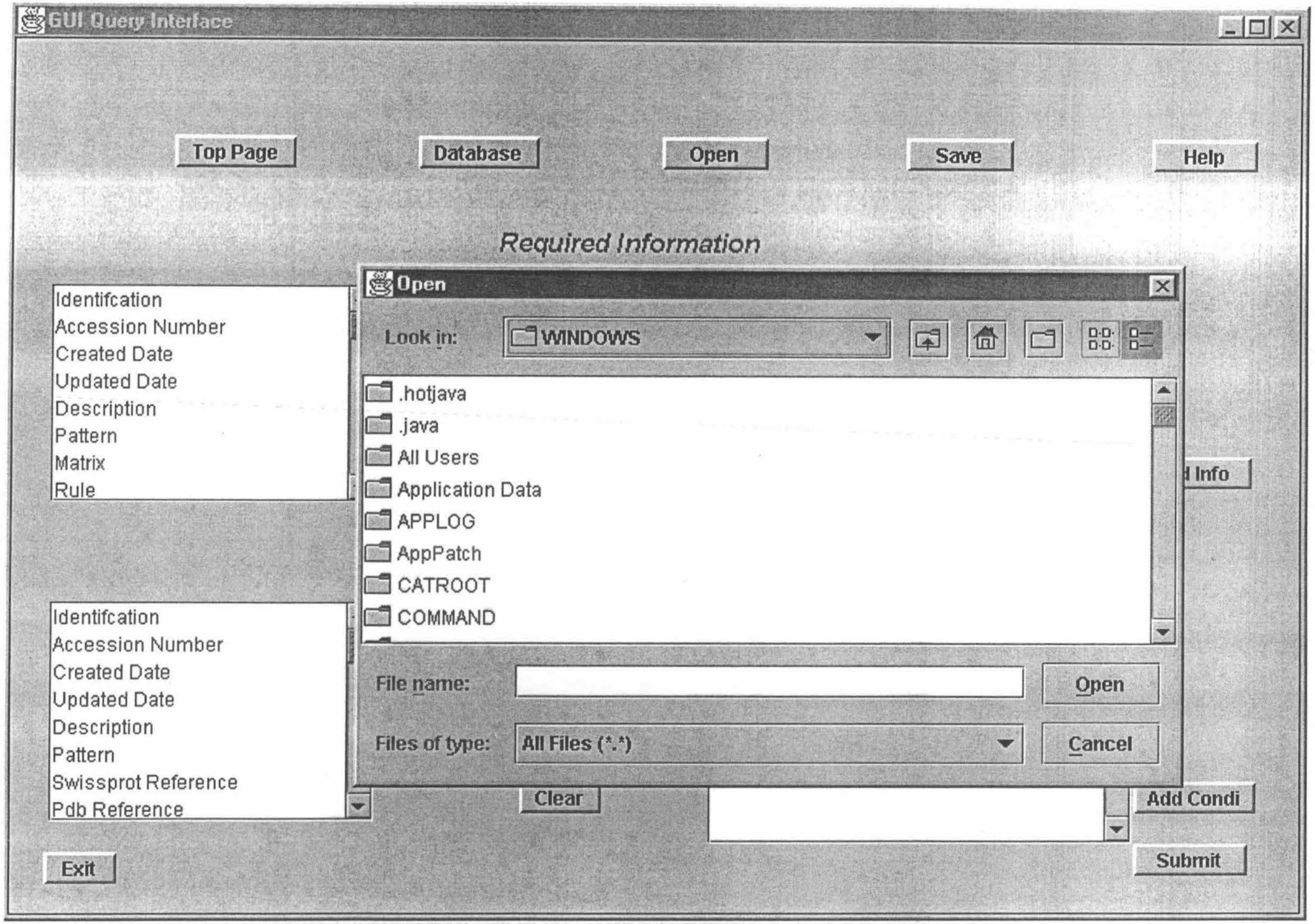
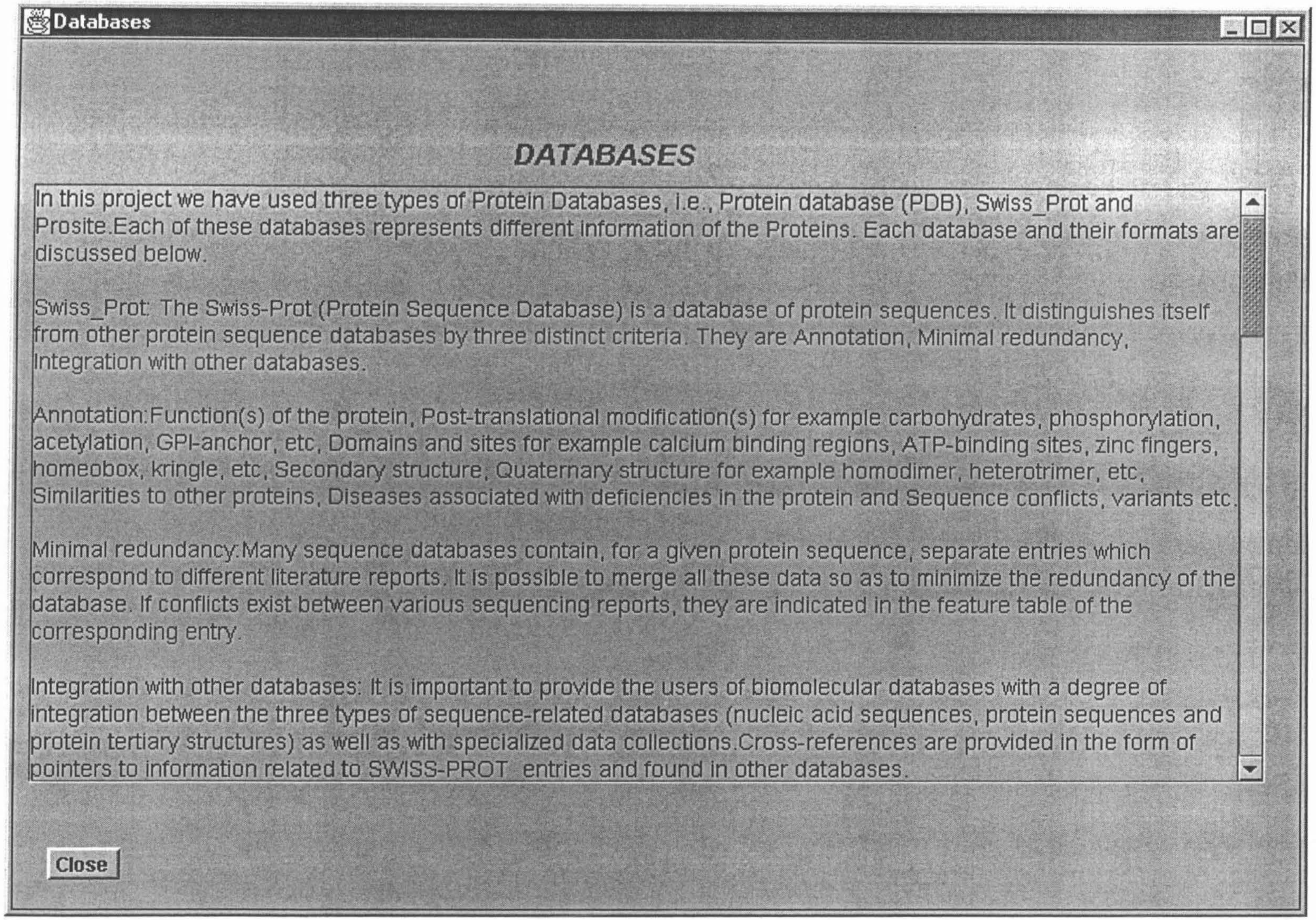


Fig 3.4 Opening a saved Query



DATABASES

In this project we have used three types of Protein Databases, i.e., Protein database (PDB), Swiss_Prot and Prosite.Each of these databases represents different information of the Proteins. Each database and their formats are discussed below.

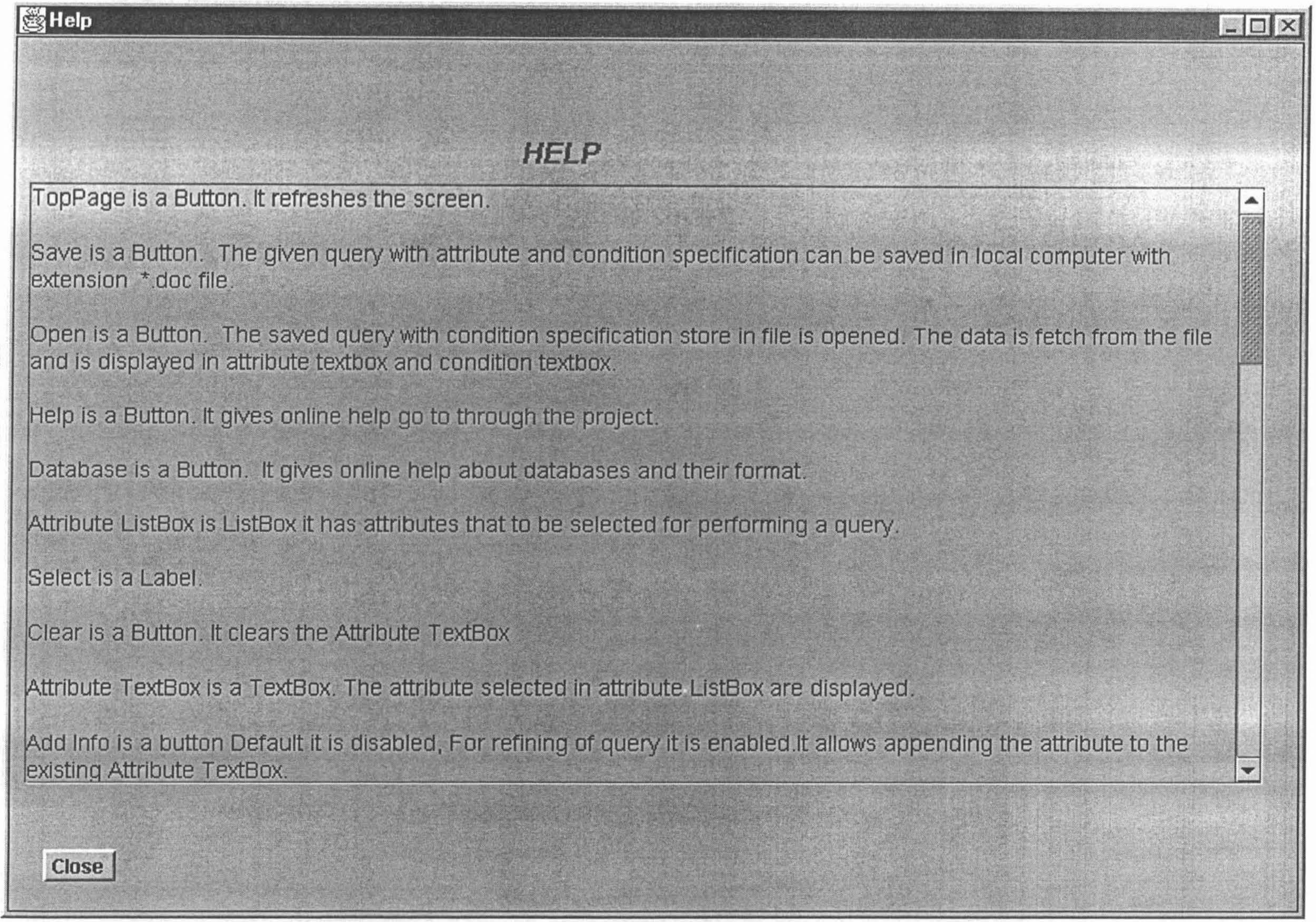
Swiss_Prot: The Swiss-Prot (Protein Sequence Database) is a database of protein sequences. It distinguishes itself from other protein sequence databases by three distinct criteria. They are Annotation, Minimal redundancy, Integration with other databases.

Annotation:Function(s) of the protein, Post-translational modification(s) for example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc, Domains and sites for example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc, Secondary structure, Quaternary structure for example homodimer, heterotrimer, etc, Similarities to other proteins, Diseases associated with deficiencies in the protein and Sequence conflicts, variants etc.

Minimal redundancy:Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. It is possible to merge all these data so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

Integration with other databases: It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialized data collections.Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in other databases.

Fig 3.5 Database Help



TopPage is a Button. It refreshes the screen.

Save is a Button. The given query with attribute and condition specification can be saved in local computer with extension *.doc file.

Open is a Button. The saved query with condition specification store in file is opened. The data is fetch from the file and is displayed in attribute textbox and condition textbox.

Help is a Button. It gives online help go to through the project.

Database is a Button. It gives online help about databases and their format.

Attribute ListBox is ListBox it has attributes that to be selected for performing a query.

Select is a Label.

Clear is a Button. It clears the Attribute TextBox

Attribute TextBox is a TextBox. The attribute selected in attribute ListBox are displayed.

Add Info is a button Default it is disabled, For refining of query it is enabled.It allows appending the attribute to the existing Attribute TextBox.

Fig 3.6 Help

CHAPTER 4
Design

In this chapter we discuss System Architecture of our project and Design of Structure chart for GUI Interface.

4.1 System Architecture

The system follows a Two-tier architecture. Fig 4.1 shows overall System Architecture.

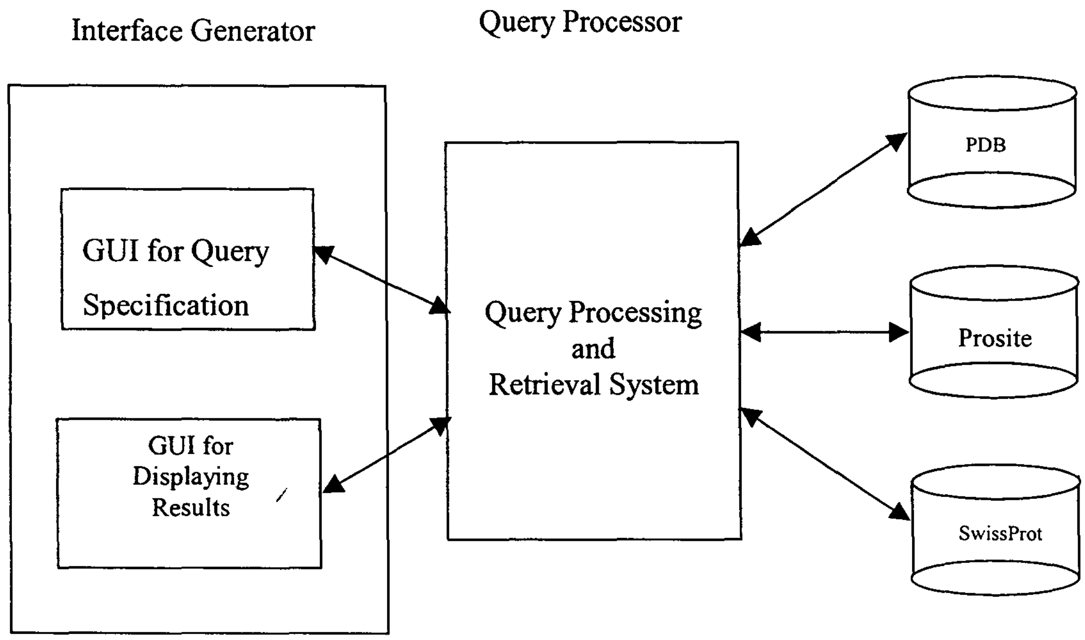

Interface Generator
Query Processor
GUI for Query Specification
GUI for Displaying Results
Query Processing and Retrieval System
PDB
Prosite
SwissProt

Fig 4.1 Two-tier Architecture

The GUI applications act as a client to the system. It generates the user query in user terms. Then the query is divided and presented to the database server. The Database server gives the required data to the client. The GUI application at the client displays the data in user friendly forms. The design details are explained in the following sections.

GUI for Query Specification is a user friendly GUI is provided for the user to develop the query for the system. The databases are implicitly selected Data Dictionary is developed for database terms. The user can perform complex queries using logical connective like conjunction and disjunction. Condition using arithmetic comparisons greater than (>), less than (<), equals to (=) can also be performed. The query and condition developed will be processed and will be executed in Query Processing Retrieval System.

In **Query Processing and Retrieval System** query given by the user through GUI will be processed and data is fetched from the databases. Required information that satisfies the given condition is fetched from the databases for which the query is applied to. To search information the query is decomposed into multiple sub queries, which access individual databases. The data from the various sources is collated. These aspects of the project are explained in accessing Multiple Biological Protein Databases.

In **Displaying Results,** results obtained from the retrieval system are presented in a user-friendly way. Two forms of output is presented to the user one is Textual form and Graphical form, Textual form is in List or table form, Graphical form is a 3-D structure of Protein. This provides the user to analyze the results obtained from the system. These aspects of the project are explained in Structural Information of Protein based on Multiple Databases.

In this thesis we are concerned with the design of GUI Interface module. Design of this module is explained below.

4.2 Design

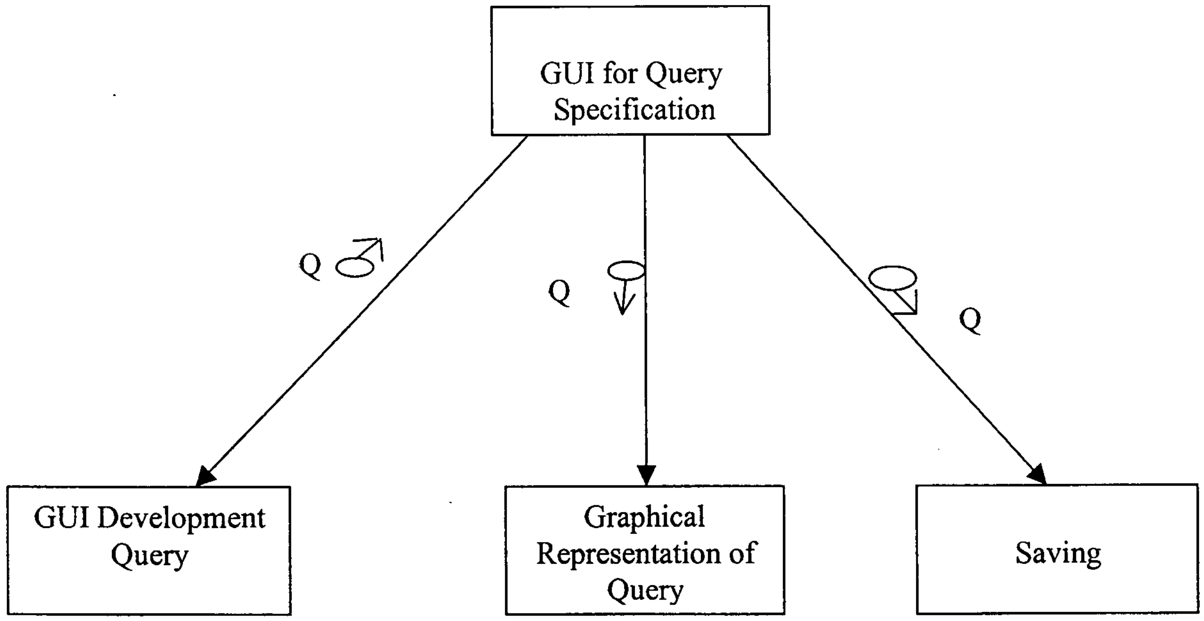

GUI for Query Specification
Q
Q
Q
GUI Development Query
Graphical Representation of Query
Saving

Fig: 4.2 Structure chart for GUI Interface

The above diagram represents the complete design of the GUI Interface that consists of three sub modules in it. These are discussed in detail below. In the above structure chart 'Q' stands for Query information.

GUI development query module gives the user-friendly interface to select the required information from the attribute column and condition specification from the condition attribute column, these columns are composed of Data Dictionary i.e., user terms. The required information with condition specification is send for processing with submit button on Fig 3.1. Sub module of development query is discussed in detail in following sections.

GUI Representation of Query is presented to the user as shown in Fig 3.3. Graphical display of query shows user from which database the user terms are selected to perform the query. It also gives user to know the cross-references between the databases through cross-reference keys.

Saving of query with required information and condition specification is saved on local computer with extension of *.doc file.

4.2.1 Structure chart for Query Development

Fig 4.3 represents structure chart for Query development above diagram represents the complete design of the GUI Development that consists of two-sub module Identification specification and condition specification. These are discussed in detail below.

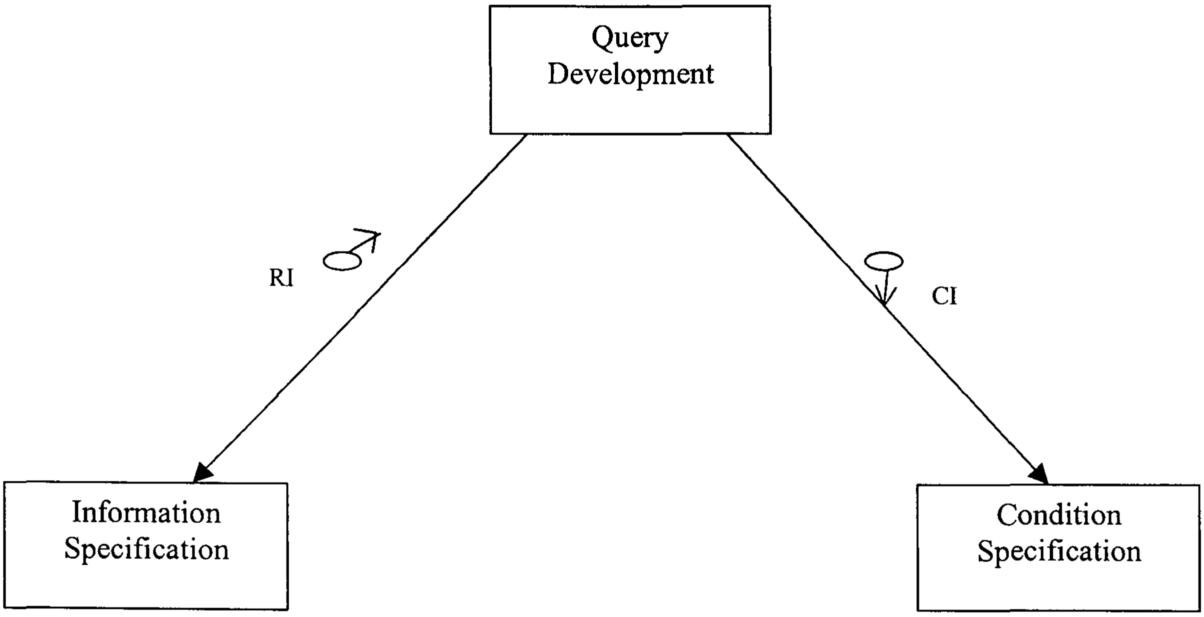


Fig 4.3 Structure Chart for Query Development

In the above structure chart 'RI' stands for Required Information and C stands for Condition information.

Information Specification gives the user to select the required information from the attribute column. Data Dictionary is developed for required information is discuss in detail in section 3.2.1

Condition specification gives the user to select the condition attribute from the condition attribute column The user can perform complex condition queries using logical OR, AND. Condition using relational operators (>, <, >=, <=) can also be performed.

4.2.2 Structure chart for Condition Development

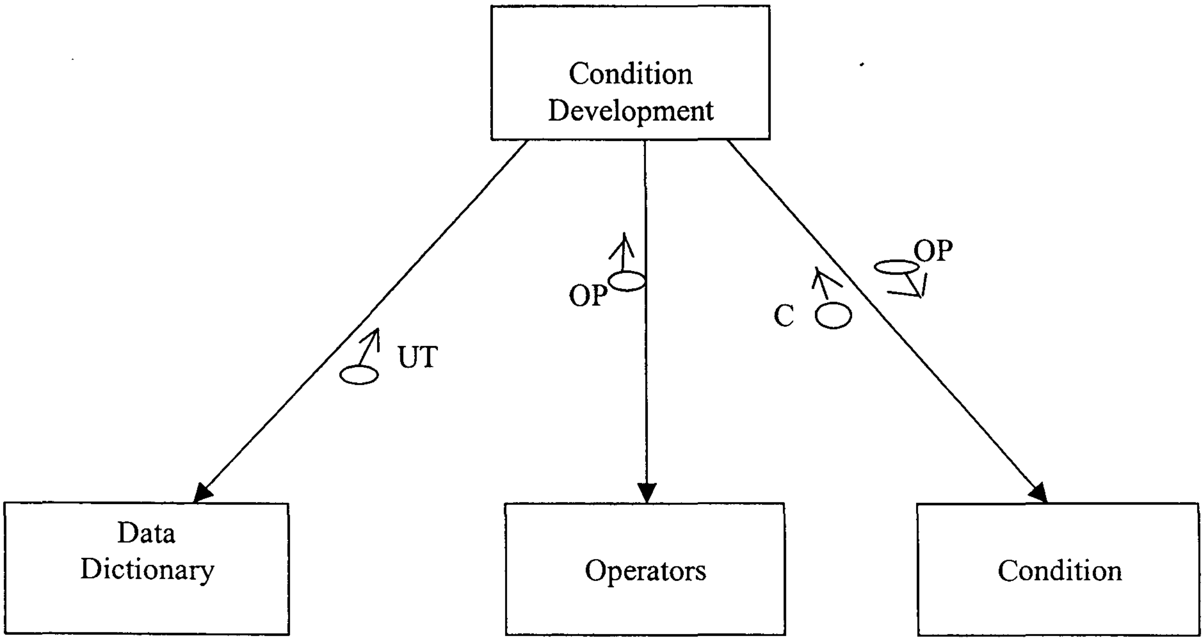


Fig: 4.4 Structures chart for Condition Development

The above diagram represents the design of the Condition Development that consists of three sub modules in it. These are discussed in detail below. In the above structure chart 'Q' stands for Query information.

Data Dictionary is developed for condition attribute column in which user can select the username.

Operator can be performed on the condition query AND, OR and relational opearator used to join the more than one condition

Condition is generated with relational and logical operators.

CHAPTER 5
Implementation

The proposed approach "GUI Interface to Biological Database" has implemented in Java. The aim of designing GUI is user-friendly. Java provides windows like GUI design with the concept of Swing We use Swing features JButton, Jlabel etc to implementation .The details of the implementation are explained in the following sections.

5.1 GUI Interface

The design details of GUI Interface are discussed in Fig 4.1. To achieve this user-friendly nature of GUI design we use Swing features. Swing is a set of classes that provides more powerful and flexible components for designing GUI with which user interacts via the mouse or the keyboard. Swing provides many standard GUI components such as buttons, lists, radiobutton, text area etc. It also includes containers such as windows and tool bars.

The swing component allows the programmer to specify a different look and feel for each platform or a uniform look and feel across all platforms or ever to change the look and feel while the program is running. Swing related classes are contained in package javax.swing.

The Swing package is part of the Java Foundation Classes (JFC). The JFC encompasses a group of features for designing look and feel GUIs; Swing package provides all the components from buttons to split panes and tables. Prior to the Swing package, the Abstract Window Toolkit (AWT) components provided all the UI components. Although the Java supports the AWT components, we strongly use Swing components instead. Swing components have their names start with J. The AWT button class, for example is named Button, whereas the Swing button class is named JButton.

Fig 5.1 shows the inheritance hierarchy of the classes that define attributes and behaviors that are common to meet Swing components.

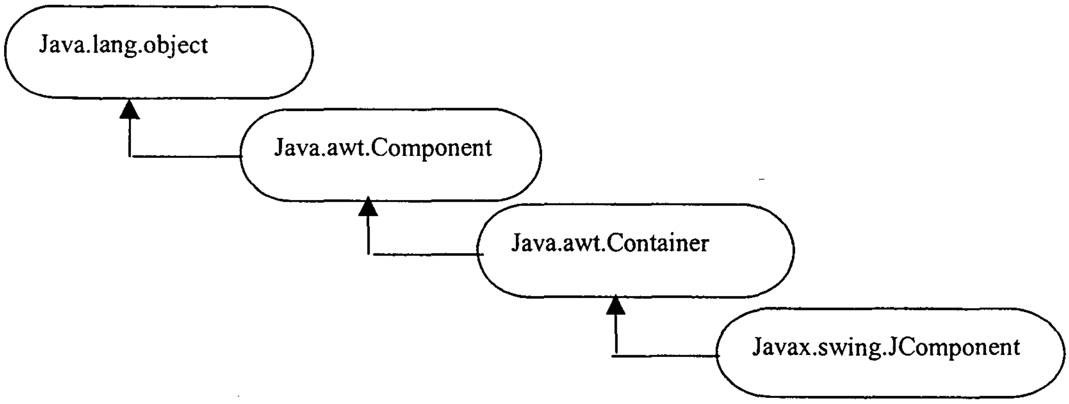


Fig 5.1 Swing hierarchy

Swing components in a GUI fit together into a containment hierarchy. Swing Application creates three commonly used Swings components:
a frame (JFrame)

a panel (JPanel)
a component (JComponent)

The frame is a top-level container. It exists mainly to provide a place for other Swing components. JFrame provides the basic attributes and behaviors of a window like title bar and buttons to minimize, maximize and close window.

The panel is an intermediate container. Its only purpose is to simplify the positioning of Components Following constructor is used in create JPanel object.
JPanel panel=new JPanel ();

In our GUI design each component requires be placed in an exact location for that we use Null layout, it is used to place components at specific position.
To set layout of panel we used
setLayout (name of layout) is used.
panel.setLayout(null);

The classes JTextArea Jlabel, JRadioButton etc are subclass of JComponents each of this class is explained below.

JButton is like a Push Button we use Button to provide windows like GUI Interface.

To create new JButton
JButton button=new JButton ("Caption");

For good look up Bevel Border is given to each button with following class
Border raisedbevel = BorderFactory.createRaisedBevelBorder();

To apply the Bevel Border following method is used

```
button.setBorder(raisedbevel);
```

User has to perform some action when button is click for example Help Button option in Fig 4.1 give the online help for project.

Event is generated when button is click each listener implements ActionListener Interface

ActionPerformed () method is used when an event occurs

example of code is shown below.

```
button.addActionListener( new ActionListener()
{
public void actionPerformed(ActionEvent e)
    {
    }
 }
);
```

To display name on the frame for example in Fig 4.1 Required Information we use JLabel for that.

JLable is a display area for a short text string. This is used for to display text on the frame. To create a Jlabel object we use following constructor

```
JLabel label=new JLabel (name);
```

5.2 Data Dictionary

As discussed in section 4.3 data dictionary is developed with user terms.

Data dictionary provides the user to select user term rather than database term. For Swiss_Prot [], PDB [] and Prosite [] data dictionary is developed. All user term should be kept at one place for that we use JList component of Swing.

JList allows selecting item with left click button of the mouse. Two JLists components are used, one for column attribute and other for condition attribute. The selected item is displayed in corresponding textarea.

The new JList is created with following constructor.

```
JList list=new List (name);
```

Sets the height of item in the list setFixedCellHeight (height) is used

```
list.setFixedCellHeight(height);
```

ScrollBar is provided for the list to see all items in the list

```
JScrollPane listScrollPane = new JScrollPane (list);
```

Selected Item in List, event is generated when item in the list is click each listener implements ListSelectionListener interface

valueChanged() method is used when an event occurs, sample code is shown below.

```
list.addListSelectionListener(new ListSelectionListener()
{  public void valueChanged(ListSelectionEvent e )
   {
   }
}
);
```

For example user select Swiss_prot_Identification in column attribute List, the selected value is displayed in column JTextArea.

A JTextArea is a multi-line area that displays plain text. This is used in displaying column attributes and condition attributes when they are selected in their respective JList.

JTextArea object is created using following constructor.

```
JTextArea textarea=new JTextArea (Text, rows, column);
```

Use is not allow to edit the TextArea setEditable() method is used.

```
 textarea.setEditable(false);
```

ScrollBar is provided for the textarea for line wrapping JScrollPane class is used

```
JScrollPane textScrollPane = new
JScrollPane (textarea, JScrollPane.VERTICAL_SCROLLBAR_ALWAYS,
JScrollPane.HORIZONTAL_SCROLLBAR_NEVER);
```

5.3 Graphical Display of Query

When query is submitted for processing first Graphical Representation of query is shown as the relationship between the database through the cross-reference keys. Fig 4.2 has shown as Graphical Display of Query.

In Fig 4.2 database name and their corresponding attribute are displayed so as to know user from which database attributes are selected.

JLable is used to display Database name and their corresponding database attribute. Link between the database is shown as the ImageIcon

To create icon we use following constructor

```
Icon image=new ImageIcon (name);
```

Icon is added to the panel as an object of Label.

```
JLable labelImage=new JLable
labelimage.setIcon(image);
```

After displaying the Graphical Display of query the next step is present the output in List form or Table form. Either of one is selected we use JRadioButton

JRadioButton component is either selected or not selected   Object of JRadioButton is created with following constructor.

```
JRadioButton rbutton=new JRadioButton (name,);
```

JRadioButton is used to see the output in Table form or List form. Event is generated when rbutton is click. Each listener implements ActionListener Interface

ActionPerformed() method is used when an event occurs. sample code is shown below

```
rbutton.addActionListener(  new ActionListener()
{
public void actionPerformed(ActionEvent e)
    {
    }
```

```
 }
 );
```

Only one output is seen at a time for that we have to choose their Table form or List form as shown in Fig 4.2 we have to group the JRadioButton we use ButtonGroup Class.

To create object of ButtonGroup, following constructor is used.

```
ButtonGroup radiogroup=new ButtonGroup ();
```

JRadioButton are added to the object of JButtonGroup, add () method is used.

```
radiogroup.add(rbutton);
```

5.4 Saving the Query

User is provided with option to execute the previous query, this can be done as saving query on local Computer. Fig 4.3 is shown as layout of saving a query in *.doc file.

JFileChooser is used to provide a mechanism to create a new file in the specified directory on local Computer. Constructs a JFileChooser pointing to the user's default directory, following constructor is used.

```
JFileChooser fc = new JFileChooser ();
```

Name of the file is stored in object of File Class,
getSelectedFile() method is used to get the file name from JFileChooser.

```
File file = fc.getSelectedFile ();
```

The selected file is opened in write mode with FileWriter Object.

```
FileWriter fout=new FileWriter (file);
```

Column Attribute and condition attributes are saved in file. getText() is used to get the data from TextArea

```
String s=text.getText ();
```

To write data into the file write () method is used.

```
fout.write(s,0,s.length());
```

“Condition “ phase is appended to the file.

```
s="\nCondition\n"; fout.write(s,0,s.length());
s=condition.getText(); fout.write(s,0,s.length());
```

JOptionPane object is used to display message on the screen when data is saved in the file or any error has occurred file storing in the file.

```
JOptionPane.showMessageDialog (frame, " Message”,
JOptionPane.PLAIN_MESSAGE);
```

5.5 Recalling the Query

Recalling the query will open the saved *.doc file and data will be displayed in column attribute and condition attribute TextArea. This layout is shown in Fig 4.4.

JFileChooser is used to open the file.

The desired file is open in read mode with RandomAccessFile

```
RandomAccessFile fin=new RandomAccessFile(file, "r");
```

StringBuffer is used to store sequence of characters.

```
StringBuffer buf=new StringBuffer ();
```

The line by line is read from the file. readLine() method is used.

```
buf=fin.readLine();
```

To append to the StringBuffer append () is used

```
buf.append(line+"\n");
```

JOptionPane object is used to display message on the screen when data is open from the file or any error has occurred while opening the file.

```
JOptionPane.showMessageDialog (frame, " Message”,
JOptionPane.PLAIN_MESSAGE);
```

5.6 Printing the Results

The Results of the Query can be printed on printer. Print Job object is used to controls printing. An application calls methods in this class to set up a job, optionally to invoke a print dialog with the user, and then to print the pages of the job.

```
PrinterJob job = PrinterJob.getPrinterJob();
```

PageFormat is used to print the frame in portrait/landscape .

```
PageFormat pf = job.pageDialog(job.defaultPage());
```

Conclusion

In our approach the user is provided with GUI query Interface to access Protein Databases and to present structural form and textual form of protein as output. The Protein databases are Swiss_Prot [3], Prosite [4], and Protein Database (PDB) [5] represent different information on protein. Swiss_Prot [3] and Prosite [4] are sequence database where as PDB [5] is structural database.

GUI query Interface is provided to the user for building the query. It provide all the data attribute in user-friendly terms so that he can select data using term with which he is familiar. After developed query graphical representation y of query is provided so as to know from which database attributes are selected. Option for saving, opening the saved query, printing the result on printer is also provided.

Biological databases are very large database term are difficult to remember so database terms is provides with user-friendly terms. User needn't be aware of database terms. Data dictionary is developed with user terms for all the database terms. The user can specify condition in two aspects one for numerical attribute with relational operators or string attribute with string matching. User can specify complex queries using logical operator.

After selecting attribute with condition specification Graphical Representation of query is provided so as to view the attribute in corresponding database and relationship between the database that is cross-reference.

An option for saving query with condition specification is provided for the user to save on local computer with extension *.doc file The save query can be retrieve from *.doc file with Open option. The three resulted forms can be viewed on the printer with Printer option.

In our project we are concerned with only three Protein Database, we can extent our project by considering other protein database as Nucleic Acid Sequence databases and Genome Databases. This project can be extended to other database format like relational model and object-oriented model etc.

References

[1] European Bioinformatics Institute (EBI) http://www.ebi.ac.uk/Information/index.html

[2] Swiss Institute of Bioinformatics http://www.isb-sib.ch/

[3] Swiss-Prot http:// www. expasy. ch/ sprot/

[4] Prosite http:// www. expasy. ch/ prosite/

[5] PDB http:// www. rcsb. org/ pdb/

[6] San Diego Super Computer Centre http://www.sdsc.edu/

[7] EMBL http:// www. ebi. ac. uk/ embl/

[8] Genbank http:// www. ncbi. nlm. nhi. gov/ Genbank

[9] cDNA http://www.cbc.umn.edu/ResearchProjects/Arabidopsis/

[10] EPD (Eukaryotic Promoter Database) http:// www. epd. isb- sib. ch/

[11] PIR http:// pir. georgetown. Edu

[12] Rebase http://rebase.neb.com/rebase/rebase.html

[13] HSC-2DPAGE 2-DE Gel Protein Databases at Harefiel http://www.harefield.nthames.nhs.uk/nhli/protein/

[14] Molecular Modelling databases
http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

[15] Pubmed http:// www. ncbi. nlm. nih. gov/ PubMed/

[16] GDB http:// www. gdb. Org

[17 Human genome database http://www.ncbi.nlm.nih.gov/genome/guide/human/

[18] Genome Sequence databases
http://inn.weizmann.ac.il/hg3m/databases/sequence.html

[19] DDBJ http://www.ddbj.nig.ac.jp/E-mail/homology.html

[20] AtDB http://www.arabidopsis.org/

[21 NCBI http://www.ncbi.nlm.nih.gov/

[22] SRS http://srs.embl-heidelberg.de:8000/srs5/\

[23] BLAST http://www.ncbi.nlm.nih.gov/BLAST/

[24] AcEDB www.acedb.org/

[25] TrEMBL http://www.ebi.ac.uk/trembl/

[26] ExPASy www.expansy.org

[27] Enzyme www.expasy.ch/enzyme/

[28] CD401bas www.us.expasy.org/cd40lbase/ -

[29] Entrez www.ncbi.nlm.nih.gov/Entrez

[30] A Molecular Biology Database Digest, Franc¸ois Bry and Peer Kr¨ oger

[31] TAMBIS - Transparent Access to Multiple Bioinformatics Information Sources, Patricia G. Baker a, Andy Brass a, Sean Bechhofer b, Carole Goble b, Norman Paton b, Robert Stevens b.

[32] Heterogeneous Data and Algorithm Integration in Bioinformatics, arbara Eckman, Julia Rice, William Swope

[33] Overview Of Selected Molecular Biological Databases ,Karen D.Rayal and Terry Gaasterland

[34] QUICK:Graphical User Interface to Multiple Databases,Wang Chiew Tan, Ke Wang, Limsoon Wong

[35] A Strategy for Database Interoperation, Peter D.Carp

Appendix A

Sample entry for Swiss_Prot

```
ID GRAA_HUMAN STANDARD; PRT;   262 AA.
AC  P12544;
DT  01-OCT-1989 (Rel. 12, Created)
DT  01-OCT-1989 (Rel. 12, Last sequence update)
DT  16-OCT-2001 (Rel. 40, Last annotation update)
DE  Granzyme A precursor (EC 3.4.21.78) (Cytotoxic T-lymphocyte proteinase
DE  1) (Hanukkah factor) (H factor) (HF) (Granzyme 1) (CTL tryptase)
DE  (Fragmentin 1).
GN  GZMA OR CTLA3 OR HFSP.
OS  Homo sapiens (Human).
OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC  Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX  NCBI_TaxID=9606;
RN  [1]
RP  SEQUENCE FROM N.A.
RC  TISSUE=T-cell;
RX  MEDLINE=88125000; PubMed=3257574;
RA  Gershenfeld H.K., Hershberger R.J., Shows T.B., Weissman I.L.;
RT  "Cloning and chromosomal assignment of a human cDNA encoding a T
RT  cell- and natural killer cell-specific trypsin-like serine
RT  protease.";
RL  Proc. Natl. Acad. Sci. U.S.A. 85:1184-1188(1988).
RN  [2]
RP  SEQUENCE OF 29-53.
RX  MEDLINE=88330824; PubMed=3047119;
RA  Poe M., Bennett C.D., Biddison W.E., Blake J.T., Norton G.P.,
RA  Rodkey J.A., Sigal N.H., Turner R.V., Wu J.K., Zweerink H.J.;
RT  "Human cytotoxic lymphocyte tryptase. Its purification from granules
RT  and the characterization of inhibitor and substrate specificity.";
RL  J. Biol. Chem. 263:13215-13222(1988).
RN  [3]
RP  SEQUENCE OF 29-40, AND CHARACTERIZATION.
RX  MEDLINE=89009866; PubMed=3262682;
RA  Hameed A., Lowrey D.M., Lichtenheld M., Podack E.R.;
RT  "Characterization of three serine esterases isolated from human IL-2
RT  activated killer cells.";
RL  J. Immunol. 141:3142-3147(1988).
RN  [4]
RP  SEQUENCE OF 29-39, AND CHARACTERIZATION.
RX  MEDLINE=89035468; PubMed=3263427;
RA  Kraehenbuhl O., Rey C., Jenne D.E., Lanzavecchia A., Groscurth P.,
RA   Carrel S., Tschopp J.;
RT   "Characterization of granzymes A and B isolated from granules of
RT   cloned human cytotoxic T lymphocytes.";
RL   J. Immunol. 141:3471-3477(1988).
RN   [5]
RP   3D-STRUCTURE MODELING.
RX   MEDLINE=89184501; PubMed=3237717;
RA   Murphy M.E.P., Moult J., Bleackley R.C., Gershenfeld H.,
RA   Weissman I.L., James M.N.G.;
RT   "Comparative molecular model building of two serine proteinases from
RT   cytotoxic T lymphocytes.";
RL   Proteins 4:190-204(1988).
CC   -!- FUNCTION: THIS ENZYME IS NECESSARY FOR TARGET CELL LYSIS IN CELL-
CC       MEDIATED IMMUNE RESPONSES. IT CLEAVES AFTER LYS OR ARG. MAY BE
CC       INVOLVED IN APOPTOSIS.
CC   -!- CATALYTIC ACTIVITY: HYDROLYSIS OF PROTEINS, INCLUDING
FIBRONECTIN,
CC       TYPE IV COLLAGEN AND NUCLEOLIN. PREFERENTIAL CLEAVAGE: ARG-|-
XAA,
CC       LYS-|-XAA >> PHE-|-XAA IN SMALL MOLECULE SUBSTRATES.
CC   -!- SUBUNIT: HOMODIMER; DISULFIDE-LINKED.
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC GRANULES.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY. STRONGEST TO OTHER GRANZYMES AND TO MAST CELL
CC       PROTEASES.
CC   --------------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use  by  non-profit  institutions as long  as its content  is  in  no  way
CC   modified and this statement is not removed.  Usage  by  and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   --------------------------------------------------------------------------
DR   EMBL; M18737; AAA52647.1; -.
DR   PIR; A28943; A28943.
DR   PIR; A30525; A30525.
DR   PIR; A30526; A30526.
DR   PIR; A31372; A31372.
DR   PDB; 1HF1; 15-OCT-94.
DR   MEROPS; S01.135; -.
DR   MIM; 140050; -.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00089; trypsin; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Serine protease; Zymogen; Signal; T-cell; Cytolysis;
KW   Apoptosis; 3D-structure.
```

```
FT   SIGNAL       1    26
FT   PROPEP      27    28       ACTIVATION PEPTIDE.
FT   CHAIN       29   262       GRANZYME A.
FT   ACT_SITE     69    69      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    114   114      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    212   212      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   DISULFID     54    70      BY SIMILARITY.
FT   DISULFID    148   218      BY SIMILARITY.
FT   DISULFID    179   197      BY SIMILARITY.
FT   DISULFID    208   234      BY SIMILARITY.
FT   CARBOHYD    170   170      N-LINKED (GLCNAC...) (POTENTIAL).
SQ   SEQUENCE   262 AA;  28968 MW;  DA87363A0D92BAF4 CRC64;
     MRNSYRFLAS SLSVVVSLLL IPEDVCEKII GGNEVTPHSR PYMVLLSLDR
KTICAGALIA
     KDWVLTAAHC NLNKRSQVIL GAHSITREEP TKQIMLVKKE FPYPCYDPAT
REGDLKLLQL
     TEKAKINKYV TILHLPKKGD DVKPGTMCQV AGWGRTHNSA SWSDTLREVN
ITIIDRKVCN
     DRNHYNFNPV IGMNMVCAGS LRGGRDSCNG DSGSPLLCEG VFRGVTSFGL
ENKCGDPRGP
     GVYILLSKKH LNWIIMTIKG AV
//
```

Appendix B

Sample entry of PDB

```
HEADER      HYDROLASE (O-GLYCOSYL)   16-SEP-77               8LYZ                 8LYZ 3
COMPND      LYSOZYME (E.C.3.2.1.17) IODINE-INACTIVATED                            8LYZ 4
SOURCE      HEN (GALLUS GALLUS) EGG WHITE                                         8LYZ 5
AUTHOR      C.R.BEDDELL,C.C.F.BLAKE,S.J.OATLEY                                    8LYZ 6
REVDAT 9     14-JUL-86 8LYZH  3    SEQRES TURN  ATOM                              8LYZH 1
REVDAT 8    22-OCT-84 8LYZG  1      SHEET                                         8LYZG 1
REVDAT 7     27-JAN-84 8LYZF  1    REMARK                                         8LYZF 1
REVDAT 6    30-SEP-83 8LYZE   1    REVDAT                                         8LYZE 1
REVDAT 5    01-MAR-82 8LYZD  1     REMARK                                         8LYZE 2
REVDAT 4    21-MAY-81 8LYZC  3     ATOM                                           8LYZE 3
REVDAT 3    25-MAY-78 8LYZB  1      SEQRES                                        8LYZE 4
REVDAT 2    01-NOV-77 8LYZA  1     SSBOND                                         8LYZE 5
REVDAT 1    24-OCT-77  8LYZ 0                                                     8LYZE 6
JRNL        AUTH   C.R.BEDDELL,C.C.F.BLAKE,S.J.OATLEY                             8LYZ 7
JRNL        TITL    AN X-RAY STUDY OF THE STRUCTURE AND BINDING                   8LYZ 8
JRNL        TITL  2  PROPERTIES OF IODINE-INACTIVATED LYSOZYME                    8LYZ 9
JRNL        REF      J.MOL.BIOL. V. 97 643 1975                                   8LYZ 10
JRNL        REFN     ASTM JMOBAK UK ISSN 0022-2836 070                            8LYZ 11
REMARK 1                                                                          8LYZ 12
REMARK 1    REFERENCE 1                                                           8LYZ 13
REMARK 1    AUTH R.DIAMOND                                                        8LYZ 14
REMARK 1    TITL     REAL-SPACE REFINEMENT OF THE STRUCTURE OF HEN                8LYZ 15
REMARK 1    TITL 2   EGG-WHITE LYSOZYME                                           8LYZ 16
REMARK 1    REF      J.MOL.BIOL. V. 82 371 1974                                   8LYZ 17
REMARK 1    REFN     ASTM JMOBAK UK ISSN 0022-2836 070                            8LYZ 18
REMARK 1    REFERENCE 2                                                           8LYZ 19
REMARK 1    AUTH D.C.PHILLIPS                                                     8LYZ 20
REMARK 1    TITL     CRYSTALLOGRAPHIC STUDIES OF LYSOZYME AND ITS                 8LYZ 21
REMARK 1    TITL 2  INTERACTIONS WITH INHIBITORS AND SUBSTRATES                8LYZ 22
REMARK 1    EDIT E.F.OSSERMAN,R.F.CANFIELD,S.BEYCHOK                           8LYZ 23
REMARK 1    REF     LYSOZYME 9 1974                                            8LYZ 24
REMARK 1    PUBL    ACADEMIC PRESS,NEW YORK                                    8LYZ 25
REMARK 1    REFN            ISBN 0-12-528950-2 977                             8LYZD 1
        [REF 3-12 deleted]
REMARK 2                                                                       8LYZ 95
REMARK 2    RESOLUTION. 2.5 ANGSTROMS.                                         8LYZ 96
REMARK 3                                                                       8LYZ 97
REMARK 3    REFINEMENT. BY THE MODEL-BUILDING AND REAL-SPACE                   8LYZ 98
REMARK 3    REFINEMENT PROCEDURES OF R. DIAMOND. REFER TO REFERENCE 1          8LYZ99
REMARK 3     ABOVE AND REMARK 4 BELOW.                                         8LYZ100
REMARK 4                                                                       8LYZ 101
REMARK 4    THE ONLY SIGNIFICANT FEATURES ON THE DIFFERENCE MAP ARE IN         8LYZ 102
REMARK 4    THE REGION OF GLU 35 AND TRP 108 SIDE CHAINS - THE OE2 ATOM        8LYZ 103
REMARK 4    OF GLU 35 FORMS A COVALENT BOND WITH THE CD1 ATOM OF TRP           8LYZ 104
REMARK 4    108. AN INTERACTIVE COMPUTER GRAPHICS SYSTEM WAS USED TO           8LYZ 105
REMARK 4     MANIPULATE THESE SIDE CHAINS IN THE RS5D COORDINATE SET OF        8LYZ 106
REMARK 4    R. DIAMOND (1974), ENTRY 2LYZ IN THE PROTEIN DATA BANK, SO         8LYZ 107
REMARK 4    THAT A FIT TO THE ELECTRON DENSITY MAP WAS OBTAINED.               8LYZ 108
REMARK 4    THESE COORDINATES, THEREFORE, ARE IDENTICAL TO THE RS5D            8LYZ 109
REMARK 4     ENTRY APART FROM PORTIONS OF THESE TWO SIDE CHAINS.               8LYZ 110
REMARK 5                                                                       8LYZA 1
REMARK 5    CORRECTION.                                                        8LYZA 2
REMARK 5    ADD SSBOND RECORDS.                                                8LYZA 3
REMARK 5    01-NOV-77.                                                         8LYZA 4
        [REMARKS 6-12 deleted]
SEQRES 1    129 LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS            8LYZ 111
SEQRES 2    129 ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY            8LYZ 112
SEQRES 3    129 ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN            8LYZ 113
SEQRES 4    129 THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP             8LYZB 3
SEQRES 5    129 TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN             8LYZB 4
SEQRES 6    129 ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE             8LYZB 5
SEQRES 7      129 PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER          8LYZ 117
SEQRES 8      129 VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASN GLY          8LYZH 7
SEQRES 9      129 MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY          8LYZ 119
SEQRES 10     129 THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU              8LYZ 120
HELIX          1       A ARG         5      HIS     15 1                      8LYZ 121
HELIX   2     B LEU           25      GLU    35 1                             8LYZ 122
HELIX          3      C CYS          80      LEU     84 5                      8LYZ 123
HELIX          4      D THR          89      LYS    96 1                       8LYZ 124
SHEET  1    S1 2  LYS          1        PHE  . 3 0                             8LYZ 125
SHEET  2    S1 2  PHE          38       THR    40  -1  N  THR  40 0  LYS   1   8LYZG 5
SHEET 1  S2 3  ALA  42   ASN   46 0                                            8LYZ 127
SHEET 2  S2 3  SER   50  GLY 54  -1  N  ASN  46  O SER  50                     8LYZ 128
SHEET 3  S2 3 GLN   57   SER 60  -1  N  TYR  53  O ILE  58                     8LYZ 129
TURN  1  T1   LYS   13   GLY 16         TYPE I.                                8LYZ 130
TURN  2  T2   LEU   17  TYR  20  NEARLY TYPE II CONFORMATION.                  8LYZ 131
TURN  3  T3   ASN  19   GLY 22  NEARLY TYPE II CONFORMATION.                   8LYZ 132
TURN  4  T4   TYR  20   TYR 23  NEARLY TYPE II CONFORMATION.                   8LYZ 133
TURN 5   T5   GLY 54  GLN 57 TYPE I,BETW STRNDS 2,3 SHT S2.                    8LYZ 134
TURN 6   T6   ASN 59    TRP 62  NEARLY TYPE I CONFORMATION.                    8LYZ 135
TURN 7   T7  THR  69    SER 72  NEARLY TYPE I CONFORMATION.                    8LYZ 136
TURN 8          T8   ASN  74   ASN 77  TYPE I.                                 8LYZ 137
TURN 9   T9   ASN  103 ASN 106 TYPE I.                                         8LYZH 8
TURN 10  T10 CYS  115 THR 118 TYPE II (IMPERFECT).                             8LYZ 139
TURN 11  T11 ILE   124  CYS 127 TYPE II (IMPERFECT).                           8LYZ 140
SSBOND        1 CYS   6        CYS    127                                      8LYZA 5
SSBOND        2 CYS   30       CYS    115                                      8LYZA 6
SSBOND        3 CYS   64       CYS    80                                       8LYZA 7
SSBOND        4 CYS   76       CYS    94                                       8LYZA 8
CRYST1        79.100   79.100   37.900   90.00    90.00    90.00    P  43  21  2        8   8LYZ 141
ORIGX1        1.000000          0.000000         0.000000         0.000000         8LYZ 142
ORIGX2         0.000000          1.000000         0.000000         0.000000        8LYZ 143
ORIGX3         0.000000         0.000000          1.000000         0.000000        8LYZ 144
SCALE1         .012642          0.000000         0.000000         0.000000         8LYZ 145
```

```
SCALE2      0.00000       0.012642      0.000000      0.000000                      8LYZ 146
SCALE3      0.000000      0.000000      .026385       0.000000                      8LYZ 147
ATOM               1 N LYS        1     3.240  10.040  10.380  1.00  0.00         8LYZ 148
ATOM               2 CA LYS       1     2.390  10.410   9.250  1.00  0.00         8LYZ 149
ATOM               3 C LYS        1     2.460  11.920  9.100   1.00  0.00         8LYZ 150
ATOM               4 O LYS        1     2.580  12.670 10.100   1.00  0.00         8LYZ 151
ATOM               5 CB LYS       1      .950     9.960 9.490  1.00  0.00         8LYZ 152
ATOM               6 CG LYS       1     -.050  10.450 8.450    1.00   0.00        8LYZ 153
ATOM        7 CD LYS       1      -1.470 10.060 8.820   1.00   0.00               8LYZ 154
ATOM               8 CE LYS       1     -2.350  9.920  7.590   1.00   0.00        8LYZ 155
ATOM               9 NZ LYS       1     -3.680 9.380  7.960    1.00   0.00        8LYZ 156
ATOM               10 N VAL       2     2.390 12.350 7.850     1.00    0.00       8LYZ 157
        [ATOM 11-998 deleted]
ATOM        999    CD1    LEU  129      -12.970  22.550  8.090  1.00 0.00        8LYZ1146
ATOM  1000  CD2    LEU 129       -13.000  20.080  8.010  1.00 0.00              8LYZ1147
TER    1002        LEU  129                                                     8LYZ1148
CONECT       48    47      981                                                  8LYZ1149
CONECT      238    237     889                                                  8LYZ1150
CONECT      277    275     820                                                  8LYZ1151
CONECT      513    512     630                                                  8LYZ1152
CONECT      601    600     724                                                  8LYZ1153
CONECT      630    513     629                                                  8LYZ1154
CONECT      724    601     723                                                  8LYZ1155
CONECT      820    277     819     822                                          8LYZ1156
CONECT      889    238     888                                                  8LYZ1157
CONECT      981    48      980                                                  8LYZ1158
MASTER      124  0  0  4  5  11 0 6  1000  1  10  10                            8LYZH 17
END                                                                            8LYZ1160
```

Appendix C

Sample entry of Prosite

```
ID  T4_DEIODINASE; PATTERN.
AC  PS01205;
DT  NOV-1997 (CREATED); JUL-1999 (DATA UPDATE); JUL-1999 (INFO UPDATE).
DE  Iodothyronine deiodinases active site.
PA  R-P-L-[IV]-x-[NS]-F-G-S-[CA]-T-C-P-x-F.
NR  /RELEASE=40.7,103373;
NR  /TOTAL=16(16); /POSITIVE=16(16); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR  /FALSE_NEG=0; /PARTIAL=0;
CC  /TAXO-RANGE=??E??; /MAX-REPEAT=1;
CC  /SITE=12,active_site;
DR  P49894, IOD1_CANFA, T; 042411, IOD1_CHICK, T; P49895, IOD1_HUMAN, T;
DR  Q61153, IOD1_MOUSE, T; 042449, IOD1_ORENI, T; P24389, IOD1_RAT  , T;
DR  P79747, IOD2_FUNHE, T; Q92813, IOD2_HUMAN, T; Q9Z1Y9, IOD2_MOUSE, T;
DR  P49896, IOD2_RANCA, T; P70551, IOD2_RAT  , T; 042412, IOD3_CHICK, T;
DR  P55073, IOD3_HUMAN, T; P49898, IOD3_RANCA, T; P49897, IOD3_RAT  , T;
DR  P49899, IOD3_XENLA, T;
DO  PDOC00925;
//
```