

**Subject: Bioinformatics**

**Lesson: Biological Sequence Databases Protein Information  
Resource (PIR)**

**Lesson Developer: Suman Sharma**

**College/ Department: Department of Botany, Ramjas College,  
University of Delhi**

## Table of Contents

### Chapter 1: Protein Information Resource (PIR)

- **Introduction**
  - **Features of PIR**
    - **Classification**
    - **Non Redundancy**
    - **Standardized Annotation**
    - **Cross Reference**
    - **Comprehensiveness**
    - **Regular releases with free accessibility**
    - **Retrieval of information from the site**
  - **Database organization and annotation**
    - **PIR – international sequences and auxillary database**
  - **PIR Resources**
    - **Data Retrieval system**
    - **Databases in PIR**
- **Summary**
- **Exercises**
- **Glossary**
- **Suggested Reading**

## Introduction

The rapid increase in number of genome sequencing projects has generated enormous amount of molecular data. In order to fully understand this huge genome base data, computational tools are required which can help in identification of structure, function and biologically relevant features in the sequences. In order to serve this purpose Protein Information Resource (PIR) was established to generate tools and resources for data storage and analysis of protein sequence for scientific community.

In year 1984, National Biomedical Research Foundation (NBRF) developed PIR (Protein Information Resource) for identification and interpretation of information on protein sequences (<http://www.nbrf.georgetown.edu/pir/find.html>). This database was actually derived from 'Atlas of Protein Sequence and Structure', which was developed by Margaret O. Dayhoff in the year 1964. Four years later in 1988, PIR along with NBRF, Munich Information Centre for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID), developed an organization referred as PIR – international with four main aims:

- (1) to create an organized, non redundant, comprehensive protein database to study structural, functional and evolutionary relationships
- (2) to generate information on biological origin of protein sequences
- (3) to make database easily accessible in public domain
- (4) to enable cross reference with other databases for presenting structural information of biomolecules.

The Protein Information Resource (PIR) is one of the most well established databases for annotated protein sequences in public domain. The expanded PIR website allows not only sequence similarity search but also other features like text based search for protein sequences and cross talk with auxillary databases, annotation – sorted search, domain search, combined global and domain search and interactive text searches.



**Figure:** PIR Homepage

Source: <http://pir.georgetown.edu/pirwww/>

## Features of PIR Database

**Classification :** In PIR on the basis of similarity, sequences are classified into families, superfamilies and homology domains. These families are organized and aligned so that database can be searched easily by the name of the gene family.

**Cross Reference:** In PIR all entries are cross-referred to reference and molecular databases like Medline, Genbank, EMBL, DDBJ, Protein Data Bank, Human Genome Database etc so that information retrieval can be optimized. Cross referenced database entries are represented in form of Hypertext-links.

**Non-Redundancy:** PIR is a non-redundant database; sequences from a species with very high identity and similarity value are merged as single entry. Even on merging identity of independently reported sequence is not lost and can be discretely observed from the canonical sequence so that the reported sequence can be reconstructed on PIR site.

**Annotation standardized:** Unlike other databases original submission entries are annotated at PIR. All entries have complete citations, which includes article titles, genetic information, mapped genes, position of introns. For high consistency and accuracy conserved and standardized terminologies and annotations are provided in the database

**Comprehensiveness:** PIR along with other databases, which are maintained by it, presents the most comprehensive repository of protein sequences.

**Regular releases and free accessibility:** the database is updated and released quarterly. Weekly updates can also be searched on PIR website. Unlike other database, sequences in PIR can be accessed in public domain as soon as they are received by the resource.

**Retrieval of information from the site:** retrieval of data and knowledge is supported by various options like superfamilies, features, authors, keywords, and sequence similarity. Multiple sequence alignments and family classification supported by hypertext links, facilitates fast retrieval of information on related sequences either in PIR or in other molecular databases.

**Table :** PIR web site URLs

Tools	URLs
PIR Home page	<a href="http://www.nbrf.georgetown.edu/pir/">http://www.nbrf.georgetown.edu/pir/</a>
MIPS Home Page	<a href="http://www.mips.biochem.mpg.de/">http://www.mips.biochem.mpg.de/</a>
Text Search	<a href="http://www.nbrf.georgetown.edu/pir/find.html">http://www.nbrf.georgetown.edu/pir/find.html</a>
Sequence Scan	<a href="http://www.nbrf.georgetown.edu/nbrf/scan.html">http://www.nbrf.georgetown.edu/nbrf/scan.html</a>
Sequence Search	<a href="http://www.nbrf.georgetown.edu/nbrf/search.html">http://www.nbrf.georgetown.edu/nbrf/search.html</a>
Complete Genome	<a href="http://www.nbrf.georgetown.edu/pir/genome.html">http://www.nbrf.georgetown.edu/pir/genome.html</a>
PIR Alignment search	<a href="http://www.nbrf.georgetown.edu/nbrf/getaln.html">http://www.nbrf.georgetown.edu/nbrf/getaln.html</a>

MIPS Protfam Project	<a href="http://www.speedy.mips.biochem.mpg.de/mips/programs/classification.html">http://www.speedy.mips.biochem.mpg.de/mips/programs/classification.html</a>
MIPS FASTA Database	<a href="http://www.speedy.mips.biochem.mpg.de/mips/programs/fasta.html">http://www.speedy.mips.biochem.mpg.de/mips/programs/fasta.html</a>
Atlas CD Rom	<a href="http://www.georgetown.edu/pir/atcd.html">http://www.georgetown.edu/pir/atcd.html</a>
PIR Documents (word lists, annotation)	<a href="http://www.georgetown.edu/pir/doc/index.html">http://www.georgetown.edu/pir/doc/index.html</a>
Editorial Board	<a href="http://www.georgetown.edu/pir/eb/edbd.html">http://www.georgetown.edu/pir/eb/edbd.html</a>

Source: Author

## Database organization and annotation

PIR – An International consortium that was constituted by PIR, MIPS and JIPD, maintains the PIR – International protein sequences Database (PSD) which is one of the largest protein sequence database in public domain. In addition to this it also provides access to auxillary databases and collection of data to be utilized for integrity checks and annotation of sequences.

### PIR – international sequences and auxillary database

- PTCHX – It is an assemblage of publicly available protein sequences which are not in PIR – International PSD. A collection of approximately 300,000 sequences, represents one of the largest repository of protein sequences available in public domain which is maintained by PATCHX along with PSD
- ARCHIVE – Its database for published or submitted protein sequences.
- NRL\_3D – It's a PDB database which provides annotation and three-dimensional structure of protein sequences.

- FAMBASE – In order to reduce time and increase sensitivity for similarity search while identifying distantly related families a small set of sequences are selected from protein families. This small collection of representative sequences is referred as FAMBASE.PIR- ALN – this database performs sequence alignment of homology domains, superfamilies and families along with annotations for the information derived from PSD and then calculates consensus alignments patterns.
- RESID – This database provides information on structure, chemistry, post translational modification and bibliography on the basis of the features available in the PSD.
- ProClass – It is a protein family database which consider PIR superfamilies and PROSITE patterns and organize PIR –international, PSD and SWISS- PROT sequences accordingly.
- Profam – Homology cluster and multiple sequence alignments which are automatically generated for homology domains, families and subfamilies are curated in this database.
- Superfamily and family classification-Margaret Dayhoff was the first one to classify proteins on the basis of superfamily. This work was alter refined by PIR – International for better evolutionary studies and database organization. According to this refinement the protein – family relationship can be used to organize database in three ways (1) On the basis of full length sequence similarities (2) homology domains (3) motifs.

## PIR Resources

The resources available at PIR can be classified into two categories, 1) data retrieval system 2) databases

### Data Retrieval System:

Data retrieval systems has three search engines

1. **Text based search engines** : They allow Boolean queries of text fields. Following steps are taken for the search
  - a. Select a database: Selection is done for UniProt KB and UniParc proteins from iProclass database which includes PIRSF families. The output is given in form of a

table showing protein families or protein entries singly for iProclass or PIRSF, respectively.

- b. **Field Selection:** It is done through a drop down menu. Since different database contains different types of information so entries will vary depending on database selected.
  - c. **Enter query:** Enter a query which can be accession number, keyword or amino acid residue for peptide search. Once query is entered results can be retrieved by pressing arrow button.
  - d. **Input Box Button selection:** It is used for multiple field search in one attempt. Multiple searches commands are connected together by certain operator words like AND, OR, NOT.
  - e. **Number of Results:** maximum of fifty entries can be shown on one page for faster results.
2. **Sequence similarity search engines:** BLAST, FASTA, Pairwise and Multiple alignments, Peptide Match, Pattern Match
- BLAST and FASTA are the two sequence similarity search tools available on all databases. The output of these search engines is generated in form of graphs showing position of hits in the query sequence. SSEARCH is the tool used for complete sequence alignment. Pairwise and Multiple sequence alignment of user provided and PSD sequences can be done by SSEARCH or ClustalW. In PIR peptide or pattern matching can work in three ways (1) by matching query sequence with sequences in database (2) by comparing pattern provided by user with database sequence (3) finding out absolute match for user – specified peptide sequence with any of the sequence database like ARCHIVE database.
3. **Advanced search engine:** It is a combination of similarity search programme with other advance features like annotation, determining gene family relationships, annotation sorted similarity search, Domain and Global similarity search, domain search etc. Option of annotation – sorted similarity search displays results by comparing BLAST or FASTA generated matches along with the user – selected annotation. The resulting matched entries can be further subjected for multiple alignments against the query using ClustalW and results can be viewed by Mview. For domain similarity search FASTA programme is used which searches and compile sequences from PIR- international PSD, and displays result in graphical form where matched region along with their links to domain alignments in PIR – ALN can also be seen. BLAST programme is used for global similarity search in PSD and FASTA for



local similarity search in collection of domain sequences and results are arranged according to global score showing extent of matches at global and domain levels.

The PIR Integrated Environment is an analysis system which execute multiple tasks like sequence similarity search, peptide match, pattern match, multiple sequence alignment PIR advance similarity search and entry retrieval by options like superfamily, family, species, title, taxonomic group, keywords or domains.

GeneFIND utilizes multiple tools like BLAST search, SSEARCH, Motif pattern matching, MOTIFIND neural networks, ClustalW multiple motif alignment and hidden Markov motif modeling.

Flow chart for text search from PIR webpage

1. Click on search/analysis option in PIR homepage
2. A dropdown menu will display all search option available
3. Click on text search option

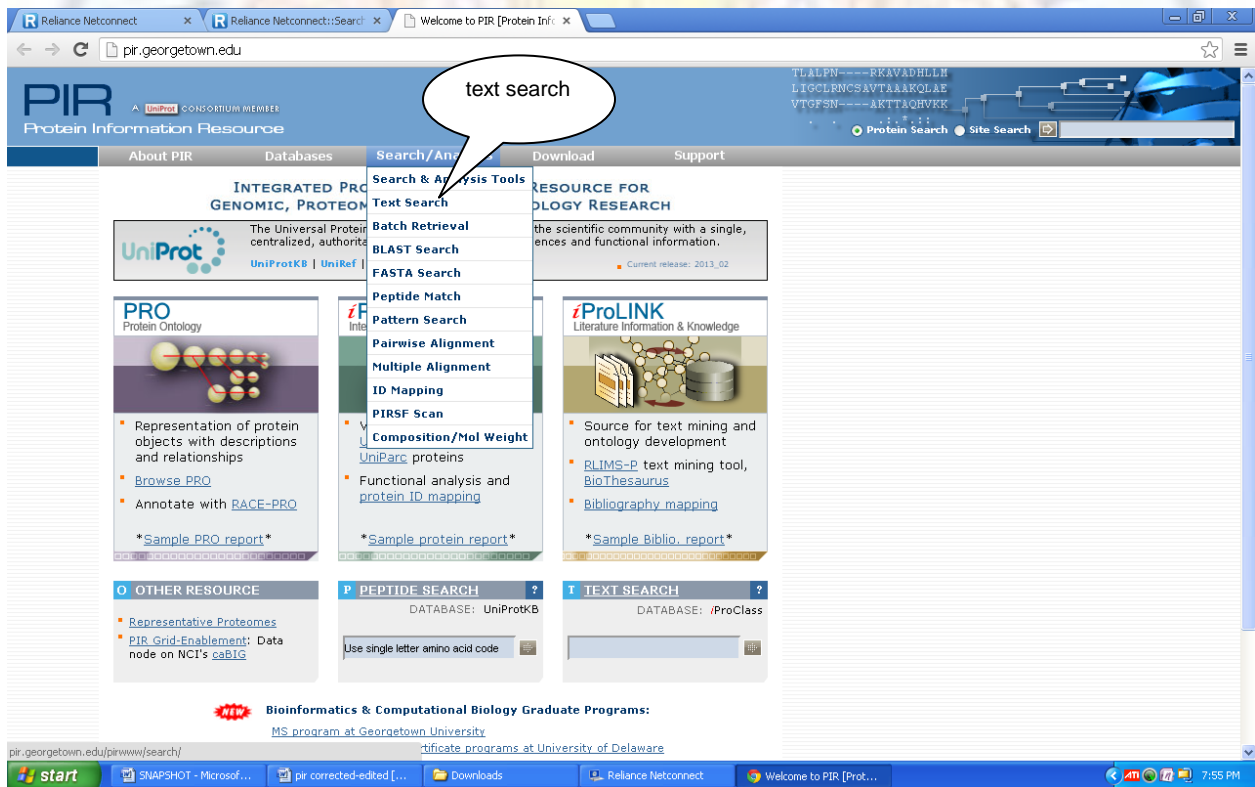


Figure: PIR homepage

Source: <http://pir.georgetown.edu/pirwww/>

4. Text search form will be displayed

## Protein Information Resource

Reliance Netconnect x Reliance Netconnect:Search x Text Search [PIR - Protein I x

pir.georgetown.edu/pirwww/search/textsearch.shtml

PIR Protein Information Resource

Uniprot CONORTIUM MEMBER

TL&LPN---RRKAVDHLLH  
LIGQLRNCASVTA&AKQLAE  
VTGFSN---AKTTAQHVKK

Protein Search Site Search

About PIR Databases Search/Analysis Download Support

HOME/Search/Text Search

Text Search Form

Retrieve sequences and reports matching your search string

1. Select a database:  iProClass  PIRSF

2. Select a field and insert query below:

Any Field  Add input box

Search Reset

Example: UniProtKB P04637 (sample report/annotated report)

Home | About PIR | Databases | Search/Analysis | Download | Support

SITE MAP | TERMS OF USE

©2009 Protein Information Resource

University of Delaware  
15 Innovation Way, Suite 205  
Newark, DE 19711, USA

Georgetown University Medical Center  
3300 Whitehaven Street, NW, Suite 1200  
Washington, DC 20007, USA

start Reliance Netconnect Text Search [PIR - Pr... Document1 - Microsof... 9:43 PM

### Figure: Text search form


Source: <http://pir.georgetown.edu/pirwww/search/textsearch.shtml>

5. Insert your query and click on search option

### Flow chart for blast search

1. Click on blast search in search/analysis option
2. Blast search form will be displayed
3. Select a database from the form
4. Enter query sequence and click on submit option

## Protein Information Resource



The image shows a screenshot of the Protein Information Resource (PIR) BLAST Search Form. The browser window title is "BLAST Search [PIR - Protein]". The URL is "pir.georgetown.edu/pirwww/search/blast.shtml". The page header includes the PIR logo, "A UniProt Consortium Member", and "Protein Information Resource". The navigation menu includes "About PIR", "Databases", "Search/Analysis", "Download", and "Support". The main content area is titled "BLAST Search Form" and contains the following instructions:

Retrieve sequences similar to your query

1. Select a database:  UniProtKB (or restricted by [organism/taxon group](#))  
 UniRef100
2. Insert the query sequence below (FASTA format or sequence only)  
or enter ">" followed by a [UniProtKB identifier](#):

Options

Submit Reset

Example: P53039 ([annotated output](#))

**Note: DO NOT** repeat search within a short period without waiting for results. Delays may be experienced due to heavy loads on our server or network traffic.

Query sequence should be in single letter [amino acid code](#).

The footer includes "Home | About PIR | Databases | Search/Analysis | Download | Support", "©2009 Protein Information Resource", "SITE MAP | TERMS OF USE", and a Windows taskbar with the time 9:42 PM.

Figure: Blast search page

Source: <http://pir.georgetown.edu/pirwww/search/blast.shtml>

### Flow chart for multiple alignment of sequences

1. Select multiple alignment option from search/analysis drop down menu
2. Choose a suitable programme
3. Insert sequences in fasta format and click on submit

## Protein Information Resource

**Figure:** Multiple alignment tool page

Source: <http://pir.georgetown.edu/pirwww/search/multialn.shtml>

Table : PIR search and analysis system

Search Engines	Description
Text/Entry	Interactive search of text fields for multiple queries
BLAST	Sequence similarity search and analysis tool
FASTA	Sequence similarity search and analysis tool
Pattern /Peptide	Tool for pattern or peptide matching
Pair-wise alignments	Alignment of PIR or user provided

	sequences using SSEARCH
Multiple alignment	Alignment of PIR or user provided sequences using Clustal W
PIR Annotation-sorted Search	Tool for displaying BLAST or FASTA matches sorted by user selected annotation
Domain search	Domain sequence search using FASTA
Global and Domain Search	BLAST and FASTA search of PSD for global and domain similarity
Integrated Environment for Sequence Analysis	Interface for entry retrieval and sequence annotation search
GeneFIND	Protein family classification by combining several search and alignment tools and the ProClass database

Source: Author

## Databases of PIR

Databases of PIR can be classified into three categories

1. **UniProt** – Universal protein Resource was established for providing a high quality, comprehensive tool in public domain for determining protein sequences and their function. The database can be further divided in three categories
  - a. UniProt Archive (UniParc) UniParc is a repository of non redundant, comprehensive database which contains almost all publicly available protein sequences. Unlike other databases where one type of protein may be present in multiple copies; UniParc avoid such repeatability by storing each unique sequence only once by tagging it to a unique identifier (UPI) so that it can be easily identified from different databases sources. Since UniParc contains only protein sequences therefore any other information about the protein only be retrieved

through cross references. UniParc maintains record of any changes that occur in sequences and can provide history of all the changes.

- b. UniProt Knowledgebase (UniProtKB): It provides stable and correct information on proteins for core data or for annotation. It also includes widely accepted biological ontologies, cross references, classifications and quality annotation of experimental and computational data. UniProt KB has two components. First component, referred as UniProtKB/ Swiss – Prot, contain manually annotated records which are collected from literature and the second component is UniProtKB/TrEMBL which is computationally analysed. The knowledgebase database provides information about all the protein products which may be obtained from diverse sources like species, accession number, alternative splicing, enzymatic cleavage and post translation variants.

- c. UniProt Reference (UniRef)

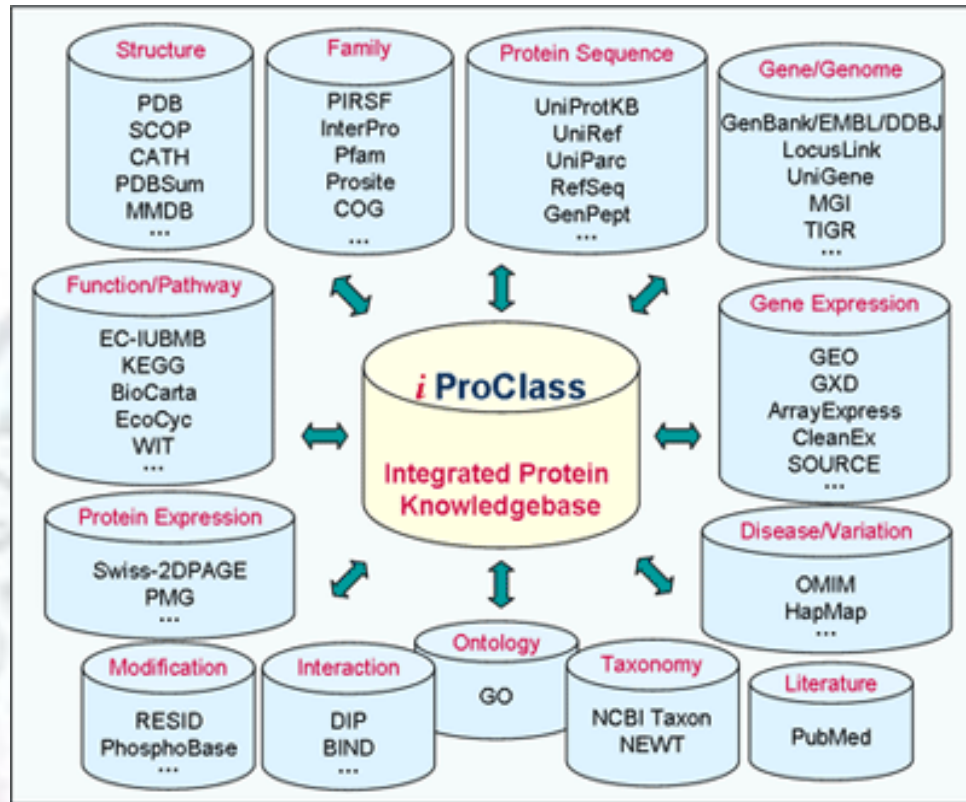
The UniRef provides more comprehensive information as it has cluster of sequences from UniProt knowledgebase and UniParc records so that user get complete coverage of sequences at several resolution after masking the redundant sequences. In UniRef 100 database sequences / subfragments which have 11 or more identical residues are merged into single UniRef entry. Despite that sequences, accession number and links to corresponding UniProtKB and UniParc records of a representative protein in the merge can be retrieved. In UniRef 90 and UniRef 50 sequences with 90 and 50 percent identity can be clustered. Entries in UniRef database are given a prefix UniRef 100, 90 and 50 depending on the identity percentage.

## 2. **iProClass: Integrated Protein knowledgebase**

This database is connected to almost 160 biological databases like database for structure and function, protein families, information on gene and genomes, ontologies, taxonomy and literature. iProClass provides a comprehensive information on protein properties which can be exploited for determining function which are uncharacterized / hypothetical. It also provides value addition to protein sequences.

Protein sequence reports can be obtained in two different patterns in iProClass. In first pattern complete information on gene, family, genetics, function, taxonomy, disease and literature supplemented with reference to other databases. This pattern also provide graphical representation of motif and domain and a link to related sequences in non analysed FASTA format. In the second pattern information about PIR superfamily

membership, length of sequence taxonomy, family relationship at domain and motif levels. It also provides links to multiple sequence alignments and phylogenetic trees generated for superfamilies with curated seed members.



**Figure :** Input sources for iProClass database

Source: <http://pir.georgetown.edu/pirwww/about/brochure.pdf>

### 3. PIRSF – Protein Family classification System

PIRSF system works not only on specific domains but also on whole protein by annotating genetic, biochemical and other specific biological function. It also classify proteins which do not have well defined domains.

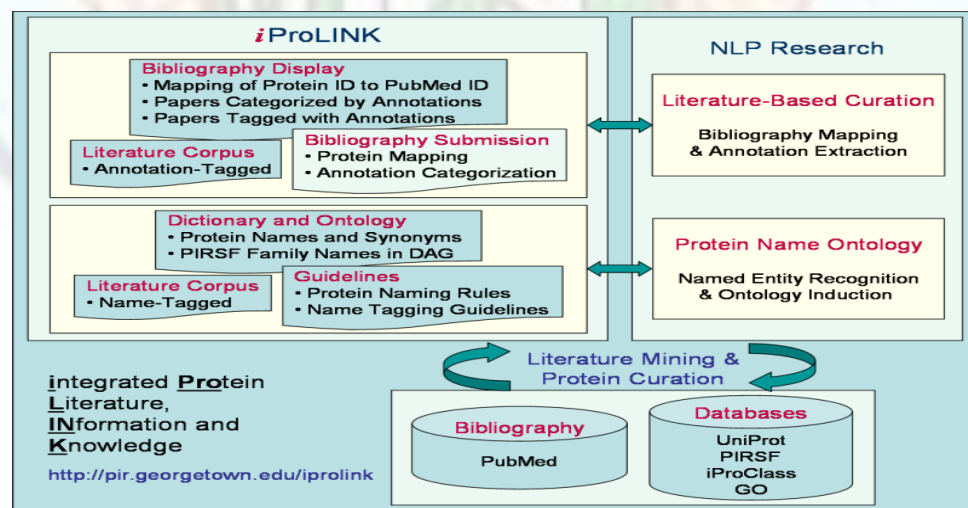
In PIRSF curation is done for homeomorphic family containing homologous proteins (descended from a common ancestor and show sequence similarity) and homomorphic protein having complete sequence similarity and common domain structure. Common domain structure is conserved for type, number and order in core domain varying only in repeating and auxiliary domains. Variable domains are called so because they can be easily lost, acquired or replaced during evolution. PIRSF database contains two kinds of data (i)

primary cluster families which are grouped on the basis of pairwise similarity or on cluster based parameters. (ii) curated families where grouping is done at two levels. In the first level a group of regular members are selected and referred as seed members which are then used to generate phylogenetic trees, multiple sequence alignments and Hidden Markov Models (HMM) for the respective families. In the second level additional features like name of the family, parent-child relationship, bibliography and description are provided. Several PIRSF families which are curated at second level are grouped into IneterPro. This grouping helps in checking the validity and integrity of existing families and also ensures stability and accuracy in UniProt classification and annotation.

Functional convergence and functional divergence in phylogenetic profiles can be easily deduced by PIRSF and this information can be further utilized for finding out protein structure, function and evolution.

#### 4. IProLINK : Integrated Protein Literature Information and knowledge

Tremendous increase in accumulation of information in form of active research for biological data mining in genomics/proteomics to improve their quality has posed a challenge before the scientific community. In order to meet this challenge PIR has developed a new resource iProLINK. Aim of iProLINK is to generate a data source which can be used for text mining, bibliography search, annotation retrieval by protein name and protein ontology. Special feature of iProLINK is annotation tagged literature corpora which includes many extracts full text articles supplemented with post translation modifications which are annotated in PIR.



**Figure** : Features and function of iProLINK

Source: <http://pir.georgetown.edu/pirwww/about/brochure.pdf>



## Summary

PIR is one of the most comprehensive resource for functional analysis of protein sequences derived from diverse sources like prokaryotes, eukaryotes, viruses, phages and archaea. The time when PIR is integrated with various other databases since then it has its own centralized data retrieval system, which allows users to find answers to complex biological queries that requires simultaneous involvement of multiple sources. The data in PIR is stored in well-organized manner according to protein – family relationship. According to this relationship the database is structured at three levels (1) family and superfamily (2) homology domain (3) motifs. Database is regularly updated and redundancy is checked by giving unique identifier to each unique sequence. PIR has well established data retrieval system, which can retrieve text based fields, perform sequence similarity search and also annotate the sequence. All protein data are merged in PIR which can be utilized at four level (1) UniProt (2) iPro Class (3) PIRSF and (4) iProLINK. IProLINK is a new resource developed by PIR, which can facilitate biological researchers to explore more information on protein and their properties.

## Exercise

1. Give function of following databases of PIR  
PATCHX , PIR-ALN, ProFam, RESID
2. Expand the following terms  
NBRF, PSD, UniProt KB, PIRSF
3. Write short notes on the following
  - a) Universal Protein Resource
  - b) Features of PIR
  - c) Give an account of data retrieval system in PIR

## Glossary

**Accession number** – it's a unique number given when a novel entry is submitted in any biological database.

**Annotation** – additional information given to a raw DNA sequence in form of start and stop codons, ORFs and coding gene for amino acid sequence

**BLAST** – Basic Local Alignment Search Tool. It's a sequence alignment tool which compare the query with the sequence in database for local regions of similarity

**Data Mining** – the process of questioning large database to answer a hypothesis or to produce a new hypothesis using statistical parameters

**Domain** – an separate folded unit in a protein which has a special function. The overall functioning of a protein is determined collectively by all its domains

**Gene** – a short sequence of nucleotide present on a chromosome which can form a protein or RNA molecule.

**Identity** – the degree to which two protein or amino acid sequence are similar

**Intron**- small intervening regions in DNA which can form a RNA but can not form code for a protein

**Molecular Phylogeny** – a method to find evolutionary relationship among organisms using DNA, RNA and proteins

**Query** – a sequence or term with which all entries of database are to be compared

**Similarity**- the extent to which DNA or amino acid sequence are related. It is expressed in percentage.

## Suggested Reading

1. Dayhoff M.O., Eck,R.V., Chang,M.A. and Sochard,M.R. (1965) *Atlas of Protein Sequence and Structure*, Vol. 1. National Biomedical Research Foundation, Silver Spring, MD.
2. Dayhoff M.O. (1979) *Atlas of Protein Sequence and Structure*, Vol. 5, Supplement 3. National Biomedical Research Foundation, Washington, DC.
3. Barker W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res.*, 21, 3089–3092. [[PMC free article](#)] [[PubMed](#)]
4. Pattabiraman N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) *Protein Seq. Data Anal.*, 3, 387–405. [[PubMed](#)]
5. Abola E.E., Manning,N.O., Prilusky,J., Stampf,D.R. and Sussman,J.L. (1996) *Res. Natl Stand. Technol.*, 101, 231–241.
6. Srinivasarao G.Y., Yeh,L.-S., Marzec,C.R., Orcutt,B.C. and Barker,W.C. (1999) *Bioinformatics*, 15, 382–390. [[PubMed](#)]
7. Ghosh Z., Mallick, B. (2008) *Bioinformatics Principles and Application*, Oxford University Press.