Prediction of structure and function of proteins

ELIXIR workshop
University of Thessaly
Margarita Theodoropoulou

- *Bsc* in Biology, University of Athens
- *Msc* in Bioinformatics, University of Athens
- *PhD* in Bioinformatics, University of Athens
- Post-doctoral researcher, University of Athens

Currently: post-doctoral researcher, University of Thessaly

Konstantinos Tsirigos

- *Bsc* studies in Biology, University of Athens
- *PhD* in Bioinformatics, Stockholm University

Currently: database manager, Technical University of Denmark

Lamia, 14-16/12/2018
Introduction to protein databases
Protein database development

- Protein databases (DBs) are the second largest biological DBs (after DNA DBs)
- Important because proteins exhibit large variability in their structure and function
- An important focus of modern bioinformatics is the analysis of protein sequences and functional data related to them, that are constantly being produced through wet-lab experiments

- **Atlas of protein sequence and structure** (1966) → the first protein sequence DB (before the advent of bioinformatics). Today known as Protein Information Resource (PIR)
- **Protein data bank** (PDB, 1971) → DB for structural data (1971), still remains the most widely used DB regarding macromolecular structures
- **United Protein Databases** (UniProt, 2003) → the largest protein sequence and protein function DB, created by the unification of SWISS-PROT, TrEMBL and PIR
PIR (Protein Information Resource)

PIR was developed in 1984 by National Biomedical Research Foundation (NBRF) in USA in order to provide researchers with information regarding protein sequences.

Between 1965-1978, NBRF had the first complete collection of macromolecular sequences (Atlas of Protein Sequence and Structure).

http://pir.georgetown.edu
PIR entry: IPPG

>P1;IPPG

insulin precursor - pig

C;Species: Sus scrofa domestica (domestic pig)

C;Accession: A01583; A94572; S16492; A60835; B60835

C;Keywords: hormone; pancreas

F;1-30/Domain: insulin chain B #status experimental

F;1-30,64-84/Product: insulin #status experimental

F;33-63/Domain: connecting peptide #status experimental

F;64-84/Domain: insulin chain A #status experimental

F;7-70,19-83,69-74/Disulfide bonds: #status experimental

>P1;IPPG

FVNQHLCGSH LVEALYLVCG ERGFFYTPKA RREAENPOAG AVELGGGLGG LQALALEGPP QKRGIVEQCC TSICSLYQLE NYCN*
EBI and SIB created SwissProt and TrEMBL DBs. SwissProt was the main project of Amos Bairoch Msc and PhD studies back in 1990s at SIB and then was further developed by Rolf Apweiler at EBI.

**Swissprot is different** than other protein DBs, because:

- Manual annotation – high quality data, including function, classification, post-translational modifications
- Minimal redundancy
- Cross-references to many other DBs
- Detailed manual
Prediction of Structure and Function of Proteins

ExPASy
Bioinformatics Resource Portal

Databases
- UniProtKB: functional information on proteins
- UniProtKB/Swiss-Prot: protein sequence database
- STRING: protein-protein interactions
- SWISS-MODEL Repository: protein structure homology models
- PDBSite: protein domains and families
- VinaZinc: portal to virtual UniProtKB entries
- nXProt: human proteins

Tools
- SWISS-MODEL Workbench
- SwissDock: protein-ligand docking
- 3ZP: prediction of Tertiary Structure
- 3Dj: find user-defined 3D structures
- AAComplete: protein sequence alignment
- AACompartSim: amino acid composition
- Agadir: Prediction of Tertiary Structure
- ALF: simulation of protein function
- Alignment tools: Dyna
- AAIAn: protein sequence analysis
- APSA: Advanced Protein Structure Analysis
- Ascalaph: Molecular similarity
- bgf: predict GFL motifs
- Biochemical Pathways
- BLAST: sequence similarity
- BLAST (UniProt): BLAST search
- BLAST: NCBI
- BLAST: PBLAST
- Blast2Fast: Blast to fastA
- Boxshade: MSA pretty
- GSIP: Protein sequence analysis
- Chorop: chloroplast annotation
- Click2Drug: Directory

Visual Guidance
- SideBar resources
- External resources (no support from the ExPASy Team)

Categories
- proteomics:
- protein structure and annotation
- mass spectrometry and SAGE data
- protein characterization and function
- families, patterns and paralogs
- post-translational modifications
- protein structure
- protein-protein interactions
- protein signaling

Genomics
- structural genomic tools
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A-Z

Link/Documentation
Swissprot text entry

<table>
<thead>
<tr>
<th>ID</th>
<th>INS_PIG</th>
<th>STANDARD;</th>
<th>PRT;</th>
<th>108 AA.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>P01315; Q9TSJ5;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>INSULIN PRECURSOR.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GN</td>
<td>INS.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Sus scrofa (Pig).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>-!- FUNCTION: INSULIN DECREASES BLOOD GLUCOSE CONCENTRATION. IT INCREASES CELL PERMEABILITY TO MONOSACCHARIDES, AMINO ACIDS AND</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>EMBL; AF064555; AAC77920.1; ALT_INIT. [EMBL / GenBank / DDBJ]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KW</td>
<td>Insulin family; Hormone; Glucose metabolism; Signal; 3D-structure.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>SIGNAL 1 24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>CHAIN 25 54 INSULIN B CHAIN.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQ</td>
<td>SEQUENCE 108 AA; 11671 MW; CB4491B429858EBE CRC64; MALWTRLLPL LALLALWAPA PAQAFVNQHL CGSHLVEALY LVCGERGFFY TPKARREAEN PQAGAVELGG GLGGLQALAL EGPPQKRGIV EQCCTSICSL YQLENYCNC //</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The UniProt consortium

European Bioinformatics Institute
European Molecular Biology Laboratory
Swiss Institute of Bioinformatics

Protein Information Resource

UniProt
the universal protein resource

Lamia, 14-16/12/2018 Prediction of Structure and Function of Proteins
UniProt (Uniprot Knowledge Base) is a collaborative project between 3 institutes, namely the European Bioinformatics Institute (EBI -UK), the Swiss Institute of Bioinformatics (SIB-CH), and the Protein Information Resource (PIR - USA).

- SIB contributed a well-annotated protein sequence DB
- EBI contributed TrEMBL, an automatic, not annotated translated nucleotide DB
- PIR contributed their own protein sequence DB, as well as a group of protein families (PSD)
- Uniprot contains 3 sub-DBs:
  - UniProtKB (Swiss-Prot + TrEMBL)
  - UniRef
  - UniParc
- Uniprot is updated monthly and has 3 main FTP servers, one in each institute
- Offers lots of functionalities, e.g. text or BLAST searches, as well as Multiple Sequence Alignment (MSA) tool (ClustalO) and retrieve/mapping of protein IDs to respective entries

http://www.uniprot.org
TrEMBL

- TrEMBL is comprised of two parts: SP-TrEMBL which contains entries that will be included in Swiss-Prot and REM-TrEMBL which contains entries that will not be part of Swiss-Prot. It can have very short protein sequences (from 8 amino-acids long) or sequences that are under patent.

- Contrary to Swiss-Prot, TrEMBL is based on automated annotation instead of manual curation.

- TrEMBL does not translate DNA sequences, nor does it use gene finding software. It provides only the coding sequence (CDS) that is recommended by the researchers that deposit it in genomic databases (EMBL/Genbank/DDBJ).

- The CDS and the respective protein sequence can have been experimentally verified or derived from prediction methods. This is not clear in a TrEMBL entry.

- TrEMBL does not validate any sequence. The quality of the data is solely dependent on the researcher that submits it.
EMBL

Gene/protein name
Taxonomy
Reference
CDS

Automated extraction of protein sequence – CDS, gene name and literature references. Automated annotation

TrEMBL

Manual curation of the sequence and other biologically-related data (literature, other DBs); addition of cross-references, alternative splicing

Swiss-Prot

Swiss
Prot
EMBL

CDS

Manual curation
of the sequence
and other
biologically-
related data
(literature,
other DBs);
addition of
cross-references,
alternative splicing

Gene/protein name
Taxonomy
Reference
# The Universal Protein resource components

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Entries</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProtKB/TrEMBL</td>
<td>Computer annotated protein sequences</td>
<td>137,213,158</td>
<td>832,433</td>
</tr>
<tr>
<td>UniProtKB/Swiss-Prot</td>
<td>Manually annotated protein sequences</td>
<td>558,898</td>
<td>~9,463</td>
</tr>
<tr>
<td>UniRef100</td>
<td>One entry = All identical sequences (including fragments). ≈ 170M entries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniRef90</td>
<td>One entry = Sequences that have at least 90% or more identity ≈ 85M entries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniRef50</td>
<td>One entry = Sequences that are at least 50% or more identity ≈ 32M entries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniProt Archives</td>
<td>Archived raw protein sequences, found in publicly accessible databases: Swiss-Prot, TrEMBL, PIR, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, Patent Offices etc</td>
<td>~244,490,125</td>
<td></td>
</tr>
</tbody>
</table>

- **Release 2018_11** (Dec 2018)

- **Produced by**
  - SIB and EBI (UniProtKB/Swiss-Prot)
  - PIR (UniRef90, UniRef50)
  - EBI (UniProt Archives)

- **Use with extreme caution:** Contains pseudogenes, incorrect CDS predictions, etc...
Uniprot search
Uniprot search results

Filter by:
- Reviewed (285)

Popular organisms
- Human (285)

View by:
- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

Your results in sequence

UniProtKB results

BLAST Align Retrieve/ID mapping Peptide search

1 to 25 of 285 Show 25
Non-redundant databases

• Definition: Repeated entries

• Cause: Identical or overlapping sequences originating from the same or different author(s)

• No redundancy in Swiss-Prot

• How? When different genes in the same species code for the same protein sequences, they are merged under the same entry in UniProtKB/Swiss-Prot and all gene names appear in one field (http://www.uniprot.org/uniprot/P68431).

• Non-redundancy in UniProtKB/Swiss-Prot means that identical sequences are presented in one entry. However, if the identical sequences derive from different species, then they are multiple entries, one per species.
Redundancy

- Contain **only sequences** / search can be done using a sequence

- Created by **combining** more than one DBs, e.g:
  - NR Nucleic (genbank+EMBL+DDBJ+PDB DNA)
  - NR Protein (SWISS-PROT+TrEMBL+GenPept+PDB protein)
Protein existence evidence

- Since most protein sequences are derived from translation of nucleotide sequences (i.e. predictions), the PE line in a Uniprot entry informs us regarding the existence of the protein.

- 'Protein existence evidence' has 5 confidence levels:
  1. Evidence at protein level
  2. Evidence at transcript level
  3. Inferred from homology
  4. Predicted
  5. Protein uncertain - Unassigned (used mostly in TrEMBL)
Annotation errors

- C. Hardley, EMBO reports, 4(9), 2003.
  “Sequences are rarely deposited in a “mature” state; as with all scientific research, DNA and protein annotation is a continual process of learning, revision and corrections.” ....
  “Sequencing error rates: ~1 base in 10,000” ....
  “Making people aware of errors is good and great; making people aware that they’re responsible also for correcting errors is even greater”

- Fixing sequence errors is a **key point** in the effort for providing the scientific community with reliable entries

- The **manually annotated** entries consist of information derived from literature, specialised DBs, expert researchers/curators, idea exchange and brainstorming

- Clear **distinction** from data/information obtained by computational analyses
P63284 - CLPB_ECOLI

Protein | Chaperone protein ClpB
--- | ---
Gene | clpB
Organism | Escherichia coli (strain K12)
Status | Reviewed - Annotation score: 5/5 - Experimental evidence at protein level

Display | None

Function

Part of a stress-induced multi-chaperone system, it is involved in the recovery of the cell from heat-induced damage, in cooperation with other proteins. Protein binding stimulates the ATPase activity; ATP hydrolysis unfolds the denatured protein aggregates, with ClpB-bound aggregates, contributing to the solubilization and refolding of denatured protein aggregates by DnaK.

Regions

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide binding</td>
<td>206 - 213</td>
<td>8</td>
<td>ATP 1</td>
</tr>
<tr>
<td>Nucleotide binding</td>
<td>605 - 612</td>
<td>8</td>
<td>ATP 2</td>
</tr>
</tbody>
</table>

GO - Molecular function

- ATP binding (Source: UniProtKB)

GO - Biological process

- protein processing (Source: InterPro)
- response to unfolded protein (Source: EcoCyc)

Keywords - Molecular function

- Chaperone

Keywords - Biological process

- Response to heat
Important fields in a Uniprot entry

- All **protein** and respective **gene names**
- Biological **origin** of the protein, with links to taxonomic DBs
- **Literature** references
- Summary of **what is known** regarding the protein, e.g. function, alternative splicing, post-translational modifications, tissue expression, 3D structure etc
- Multiple **cross-references** to other DBs
- Selected **keywords**
- Description of important **sequence features** of the protein, e.g. signal peptide, transmembrane segments, PTMs, sequence variations
Cross-references in Uniprot

- Swiss-Prot was the first DB that contained cross-references to other DBs

- A Uniprot Accession Number (AC) can be used by various other DBs (e.g. BLOCKS domain db) as an identifier for their entries, i.e. facilitating direct link to Uniprot, without being explicitly referenced in Uniprot (implicit cross-references)

- Currently (Dec 2018), there are >170 cross-referenced DBs in Uniprot
  - DNA (EMBL/GenBank/DDBJ)
  - 3D-structure (PDB)
  - Family and domain (InterPro, HAMAP, PROSITE, Pfam, etc.)
  - genomic (OMIM, MGI, FlyBase, SGD, SubtiList, etc.)
  - specialized DBs (e.g. GlycoSuiteDB, PhosSite, MEROPS)
  - literature (PubMed)
Use of keywords in UniprotKB/Swiss-Prot
Link to Gene Ontology DB for further analysis and information retrieval
Reference proteomes

- Are created for model organisms, for which is a demand for extensive and comprehensive information (> 16,000 as of Dec. 2018)
- They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny.

https://www.uniprot.org/proteomes/
Downloads and updates

- **New** release every month
- **Various formats** (flat file, XML file, FASTA file)
- **Always** cite the Accession number, not the entry name (ID)
- Information included in the entries can be **altered** by Uniprot curators if they deem necessary (**not possible** in genomic databases, where only the submitting authors are responsible for the information)
- **User manual** is frequently updated

Other important protein-related databases
What can we learn from structures?

- Secondary structure
- Function
- Similarity, evolutionary relationships
- Shape, size
- Folding
- Structural motifs
- Distances, angles
- Surface interactions
- Effect of mutations
- Transmembrane segments
Worldwide Protein Databank (wwPDB)

- It was established in 1971 at Brookhaven National Laboratories (BNL) in USA and it was moved to the Research Collaboratory for Structural Bioinformatics (RCSB) in 1998.

- Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

- The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.
The PDB database

Announcing the Protein Data Bank

Nature New Biology
Vol. 233 October 20 1971

https://www.rcsb.org/

9 structures
Vast majority come from X-ray crystallography

Most structures have been solved at a resolution between 1.8-2.2 Ångstrom
PDB entry

1BXW
OUTER MEMBRANE PROTEIN A (OMPA) TRANSMEMBRANE DOMAIN

DOI: 10.2210/pdb1.BXW/pdb

Classification: MEMBRANE PROTEIN
Organism(s): Escherichia coli (strain K12)
Expression System: Escherichia coli BL21(DE3)
Mutation(s): 1

Deposited: 1998-10-03 Released: 1998-10-14
Deposition Author(s): Schulz, G.E., Pautsch, A.

Experimental Data Snapshot
Method: X-RAY DIFFRACTION
Resolution: 2.5 Å
R-Value Free: 0.235
R-Value Work: 0.189

wwPDB Validation

Metric | Percentile Ranks | Value
--- | :---: | ---
Clashscore | 6.1% | 15
Ramachandran outliers | 18.0% | 6
Sidechain outliers | 13.3% | 3
RMSD outliers | 13.3% | 2

This is version 1.4 of the entry. See complete history.

Literature

Structure of the outer membrane protein A transmembrane domain.

Pautsch, A., Schulz, G.E.

PubMed: 9808047
DOI: 10.1038/23903
Also Cited By: 1QJP
PDB information

• Coordinates of atoms that make up the structure

• Literature references

• Details regarding structure determination (e.g. experimental procedures)

• Flat file with defined format

• Every structure, before being published, is checked for errors using a computer software. Subsequently, it obtains a unique code and is deposited in the database
# Selected Protein Data Bank Record Types

<table>
<thead>
<tr>
<th>Record Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>Atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in standard residues (amino acids and nucleic acids).</td>
</tr>
<tr>
<td>HETATM</td>
<td>Atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in nonstandard residues. Nonstandard residues include inhibitors, cofactors, ions, and solvent. The only functional difference from ATOM records is that HETATM residues are by default not connected to other residues. Note that water residues should be in HETATM records.</td>
</tr>
<tr>
<td>TER</td>
<td>Indicates the end of a chain of residues. For example, a hemoglobin molecule consists of four subunit chains which are not connected. TER indicates the end of a chain and prevents the display of a connection to the next chain.</td>
</tr>
<tr>
<td>SSBOBND</td>
<td>Defines disulfide bond linkages between cysteine residues.</td>
</tr>
<tr>
<td>HELIX</td>
<td>Indicates the location and type (right-handed alpha, etc.) of helices. One record per helix.</td>
</tr>
<tr>
<td>SHEET</td>
<td>Indicates the location, sense (anti-parallel, etc.) and registration with respect to the previous strand in the sheet (if any) of each strand in the model. One record per strand.</td>
</tr>
</tbody>
</table>

# Protein Data Bank Format

<table>
<thead>
<tr>
<th>Record Type</th>
<th>Columns</th>
<th>Data</th>
<th>Justification</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>1-4</td>
<td>“ATOM”</td>
<td>left</td>
<td>character</td>
</tr>
<tr>
<td>7-11</td>
<td>Atom serial number</td>
<td>right</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>13-16</td>
<td>Atom name</td>
<td>left*</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Alternate location indicator</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>18-20</td>
<td>Residue name</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Chain identifier</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>23-26</td>
<td>Residue sequence number</td>
<td>right</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Code for insertions of residues</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>31-38</td>
<td>X orthogonal Angstrom coordinate</td>
<td>right</td>
<td>floating</td>
<td></td>
</tr>
<tr>
<td>39-46</td>
<td>Y orthogonal Angstrom coordinate</td>
<td>right</td>
<td>floating</td>
<td></td>
</tr>
<tr>
<td>47-54</td>
<td>Z orthogonal Angstrom coordinate</td>
<td>right</td>
<td>floating</td>
<td></td>
</tr>
<tr>
<td>55-60</td>
<td>Occupancy</td>
<td>right</td>
<td>floating</td>
<td></td>
</tr>
<tr>
<td>61-66</td>
<td>Temperature factor</td>
<td>right</td>
<td>floating</td>
<td></td>
</tr>
<tr>
<td>73-76</td>
<td>Segment identifier (optional)</td>
<td>left</td>
<td>floating</td>
<td></td>
</tr>
<tr>
<td>77-78</td>
<td>Element symbol</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>79-80</td>
<td>Charge (optional)</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>HETATM</td>
<td>1-6</td>
<td>“HETATM”</td>
<td>same as ATOM records</td>
<td></td>
</tr>
<tr>
<td>7-80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TER</td>
<td>1-3</td>
<td>“TER”</td>
<td>right</td>
<td>character</td>
</tr>
<tr>
<td>7-11</td>
<td>Serial number</td>
<td>right</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>18-20</td>
<td>Residue name</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Chain identifier</td>
<td>right</td>
<td>character</td>
<td></td>
</tr>
</tbody>
</table>
The SCOP database

- Structural Classification of Proteins
- Manual hierarchical classification
- 38,221 PDB Entries (2009 release)
- 110,800 Domains
- Alexey Murzin (first release 1995)

SCOP aims at the analysis of structural and evolutionary relationships between all proteins with a known structure that are deposited in PDB. The core entity is a structural domain and its hierarchy represents secondary structure content.

http://scop.mrc-lmb.cam.ac.uk/scop/
The SCOP classification

- **Family**: Clear evolutionary relationship. It is defined by two criteria, namely very similar structures and functions and >30% identical amino acids.

- **Super-family**: Possible common evolutionary origin. Low sequence similarity but common structural features.

- **Fold**: Strong structural similarity (common secondary structure features, common orientation, topology and interactions). Here, the similarity is based on physico-chemical properties rather than evolutionary relationship.

- **Class**: It is mainly related to the secondary structure content:
  a. all-α → the secondary structure is formed by α-helices
  b. all-β → the secondary structure is formed by β-sheets
  c. α/β → α-helices and β-sheets alternate in the secondary structure
  d. α+β → α-helices and β-sheets are found in distinct regions of the secondary structure

[http://scop.mrc-lmb.cam.ac.uk/scop/](http://scop.mrc-lmb.cam.ac.uk/scop/)
Structural Classification Of Proteins

Root: scop

Classes:

1. All alpha proteins [46456] (218)
2. All beta proteins [48724] (141)
3. Alpha and beta proteins (a/b) [51349] (138)
   - Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (279)
   - Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (46)
   - Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [58635] (47)
   - Does not include proteins in the immune system
7. Small proteins [56992] (75)
   - Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (6)
   - Not a true class
9. Low resolution protein structures [58117] (24)
   - Not a true class
10. Peptides [58231] (116)
    - Peptides and fragments. Not a true class
11. Designed proteins [58789] (42)
    - Experimental structures of proteins with essentially non-natural sequences. Not a true class

Done

Class: All beta proteins

Lineage:

1. Root: scop
2. Class: All beta proteins [48724]

Folds:

1. Immunoglobulin-like beta-sandwich [48725] (23)
   - sandwich, 7 strands in 2 sheets; greek-key; some members of the fold have additional strands
2. Common fold of diphtheria toxin/transcription factors/cytochrome f [49379] (10)
   - sandwich, 9 strands in 2 sheet; greek-key; subclass of immunoglobulin-like fold
3. Prealbumin-like [49451] (6)
   - sandwich, 7 strands in 2 sheets, greek-key
   - variations: some members have additional 1-2 strands to common fold
4. HSP40/DnaJ/peptide-binding domain [49492] (1)
   - sandwich, 6 strands in 2 sheets
5. Penicillin-binding protein associated domain [69188] (1)
   - sandwich, 6 strands in 2 sheets
6. Alpha-Amylase inhibitor tandem motif [49497] (1)
   - sandwich, 6 strands in 2 sheets
7. C-terminal domain of mollusc hemocyanin [81278] (1)
   - sandwich, 6 strands in 2 sheets

Done
Superfamily: alpha-helical ferredoxin

contains two Fe4-S4 clusters

Lineage:
1. Root: scop
2. Class: All alpha proteins [46456]
3. Fold: Globin-like [46457]
   core: 6 helices, folded leaf, partly opened
4. Superfamily: alpha-helical ferredoxin [46548]
   contains two Fe4-S4 clusters

Families:
1. Fumarate reductase/Succinate dehydrogenase iron-sulfur protein, C-terminal domain [46549] (3)
2. Dihydropyrimidine dehydrogenase, N-terminal domain [46553] (1)
**Structural Classification of Proteins**

- **3HHB** - all alpha
- **1CD8** - all beta
- **1KFJ** – alpha/beta
The CATH database

- Classification of protein domain structures
- 124 folds
- 226 Superfamily
- 1148 Sequence family
- 14473 Domain
- Homologous superfamilies, Topology, Architecture, Class

CATH is a hierarchical classification of protein structures that are deposited in PDB, using their structural domains.

Protein structures included in CATH have been determined at a resolution higher than 3 Ångstroms.

Some degree of automation, unlike SCOP.

http://www.cathdb.info/
Domain families databases

- Proteins are made up of 1 or more **distinct functional regions**, that are called **domains**, and are often also **structurally independent** of each other.
- These domains can function and evolve independently of the rest of the protein. Different **combinations** of such regions are responsible for the great **variation** of proteins in nature. It is thus important to detect these domains in order to functionally classify proteins.
- Domain **detection** is done by looking for **sequence similarity** (local alignment) between proteins.
- Because structure is more conserved than the sequence, domain families databases help towards identification and classification of new proteins, as well as the detection of novel protein folds.
- These DBs **differ** mainly in the way that they **detect** and **model** the domains (local similarity, patterns, Hidden Markov Models etc) and in the way that they initially define the domains.
- **CATH** and **SCOP** are based solely on **structural criteria**, whereas **Pfam** and **PROSITE** mainly on the **sequence** (thus contain more proteins).
Detecting protein domains and families

• There are two non-excluding approaches regarding function determination of a protein sequence:
  ▪ Comparison to an existing protein database using e.g. BLAST
  ▪ Scan the sequence at hand using a database of patterns or profiles

• Most proteins can be classified into families. Proteins that belong to a specific family have common functional characteristics and a common ancestor

• Some regions of a protein sequence are more conserved than others during evolution because they are important for the function and/or structure of the protein

• The ‘fingerprints’ that are created with the use of patterns or profiles of protein sequences can be used to form hypotheses regarding the function of uncharacterized protein sequences
The PROSITE database

PROSITE is a database of classification of protein sequences and sequence domains into families. Family classification is based on the similarity of the domains. Proteins or domains in the same family have, most likely, same function and share common ancestor.

- It is derived from UniProt

- Sequence regions that are more conserved throughout evolution are related to the proteins’ functions and their structure. Analyzing protein sequences from the same family, through a multiple sequence alignment, results in a ‘fingerprint’, which is characteristic and unique for the family
Conserved regions can be classified into 5 different categories:

- **Families**: proteins with the same domain arrangement
- **Domains**: structurally independent regions – specific combination of secondary structures that result in characteristic 3D-structures or folds
- **Repeats**: structural units that are found in 2 or more copies and create a specific fold
- **Motifs**: short regions with conserved active or binding sites that usually adopt a conformation when they attach their ligands
- **Sites**: functionally active amino acid residues, e.g. active sites, disulphide bridges or post-translationally modified residues
Creating ‘fingerprints’ in PROSITE (I)

- Regular expressions or motifs (easier)
  - Similar to unix commands using wildcards, etc.
  - User-friendly, intuitive
  - Can be used to scan lots of sequences fast
  - Useful for detection of biologically important regions
  - Example motif: [AC]-x-V-x(4)-{ED} which means [Ala or Cys]-any-Val-any-any-any-any-any-
    any but Glu or Asp
  - As of Dec 2018, there are 1310 motifs in PROSITE
Pros and Cons of patterns

✓ There are intuitive to a human. By reading a pattern, one can easily deduct the information it contains
✓ They are comprehensive and condense the information of a possibly large multiple sequence alignment within a few characters/symbols
✓ They are computationally ‘cheap’ and ideal for practical reasons (fast searches in large databases)

✗ We lose large part of the information contained in a multiple sequence alignment
✗ There are patterns in PROSITE that cannot fully characterize all members of a protein family
✗ In families with lots of members, usually the conservation is related to physico-chemical properties, e.g. hydrophobic or charged amino acids) and not specific amino-acids
✗ Patterns cannot incorporate gaps of the multiple sequence alignment
✗ Patterns treat the various positions of multiple alignments as if they are independent observations
Creating ‘fingerprints’ in PROSITE (II)

- Profiles, i.e. matrices with position-specific amino acid probabilities (more accurate)
  - Designed to cover the whole protein or domain length
  - Can detect families with large variation or domains with very few conserved sites
  - In this approach, we create a $k \times p$ matrix, where $k$ is the size of our alphabet (e.g. 20 for proteins) and $p$ the size of the region that we model (the columns of the multiple sequence alignment). In each position $l$ of the MSA, we assign a vector with the probabilities of appearance $pb(i)$ of each symbol
  - As of Dec 2018, there are 1231 profiles in PROSITE
The insulin family of proteins [1] groups a number of active peptides which are evolutionary related. This family currently consists of:

---

[1] Lamia, 14-16/12/2018
Prediction of Structure and Function of Proteins
From MSA to Motif in Prosite

https://prosite.expasy.org/

Prediction of Structure and Function of Proteins
Protein domain databases

- **Pfam** ([https://pfam.xfam.org/](https://pfam.xfam.org/)) – Protein families database
  - Collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families

  - Identification and annotation of genetically mobile domains and the analysis of domain architectures

  - Combines SMART and Pfam databases for easier and quicker search
The Pfam database

- Pfam is a large collection of protein families

- Based on the same idea as PROSITE, but with profile Hidden Markov Models (more sensitive in remote homology detection) and fast at the same time

- Current version (32 – Sept 2018) contains ≈18,000 families and covers >77% of all protein sequences in Uniprot version 2018_4

- Main power of Pfam is the use of the HMMER package and defined scores over which a given protein is classified into a family

- Because proteins from different families can exhibit low similarity, there exists a higher level of classification, termed the clan
Pfam homepage
A Pfam family entry

Family: LYTB (PF02401)

Summary: LYTB protein

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

No Wikipedia article | Pfam | InterPro

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

LYTB protein

The mevalonate-independent 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway for isoprenoid biosynthesis is essential in many eubacteria, plants, and the malaria parasite. The LYTB gene is involved in the trunk line of the MEP pathway.

Literature references


External database links

- PANDIT: PF02401
- Pseudosac: PF02401
- SYSTERS: LYT2

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.

European Molecular Biology Laboratory
A representation of human tyrosine kinase HCK (Uniprot: P08631; PDB: 2HCK_A)