

Session-4

Sequence motif databases- Prosite, ProDom, Pfam, Interpro

Dr Sitaram Swain

Assistant Professor

Centre for genetics and Genomics

**Centurion University of Technology and
Management,**

Odisha

MOTIF

- A **sequence motif** is a nucleotide or amino acid sequence pattern that is widespread and has a biological significance.
- For proteins, a sequence motif is distinguished from a structural motif, a motif formed by the three-dimensional arrangement of amino acids which may or may not be adjacent.
- A sequence motif residing in the coding region may encode a structural motif.
- Some such databases include:
 - i. Prosite
 - ii. PFAM
 - iii. CSD

PROSITE

PROSITE is an annotated collection of motif descriptors. PROSITE is a good source of high quality annotation for protein domain families.

A PROSITE sequence family is represented as a pattern or profile. The profiles provide a means of sensitive detection of common protein domains in new protein sequences.

PROSITE release 16.46 contains signatures specific for 1,098 protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins.



ScanProsite tool

This form allows you to scan proteins for matches against the [PROSITE collection of motifs](#) as well as against your own patterns.

- **Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.**
- Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.
- Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

Reset

STEP 1 - Submit PROTEIN sequences [\[help\]](#)

- Submit PROTEIN sequences (max. 10) [Examples](#)
- Submit a PROTEIN database (max. 16MB) for repeated scans (The data will be stored on our server for 1 month).

Enter UniProtKB accessions or identifiers or PDB identifiers or sequences in FASTA format

Prosite



ExPASy
Bioinformatics Resource Portal

Query all databases ▼

MRIILLGAPGAGKGTQAQFIMAKFGI ✕

search help

Visual Guidance

Categories

proteomics

protein sequences and identification

mass spectrometry and 2-DE data

protein characterisation and function

families, patterns and profiles

post-translational modification

protein structure

protein-protein interaction

similarity search/alignment

genomics

structural bioinformatics

systems biology

phylogeny/evolution

population genetics

Examples: [\[hide\]](#)

UniProt AC: P12445, Q0B0R4, O75445

PDB-ID: 1ake, 1s7j

Full Text Search: Thymidylate kinase, Cyclohexadienyl dehydratase

Amino acid sequence: Sequence 1, Sequence 2

! Detected query type: AA

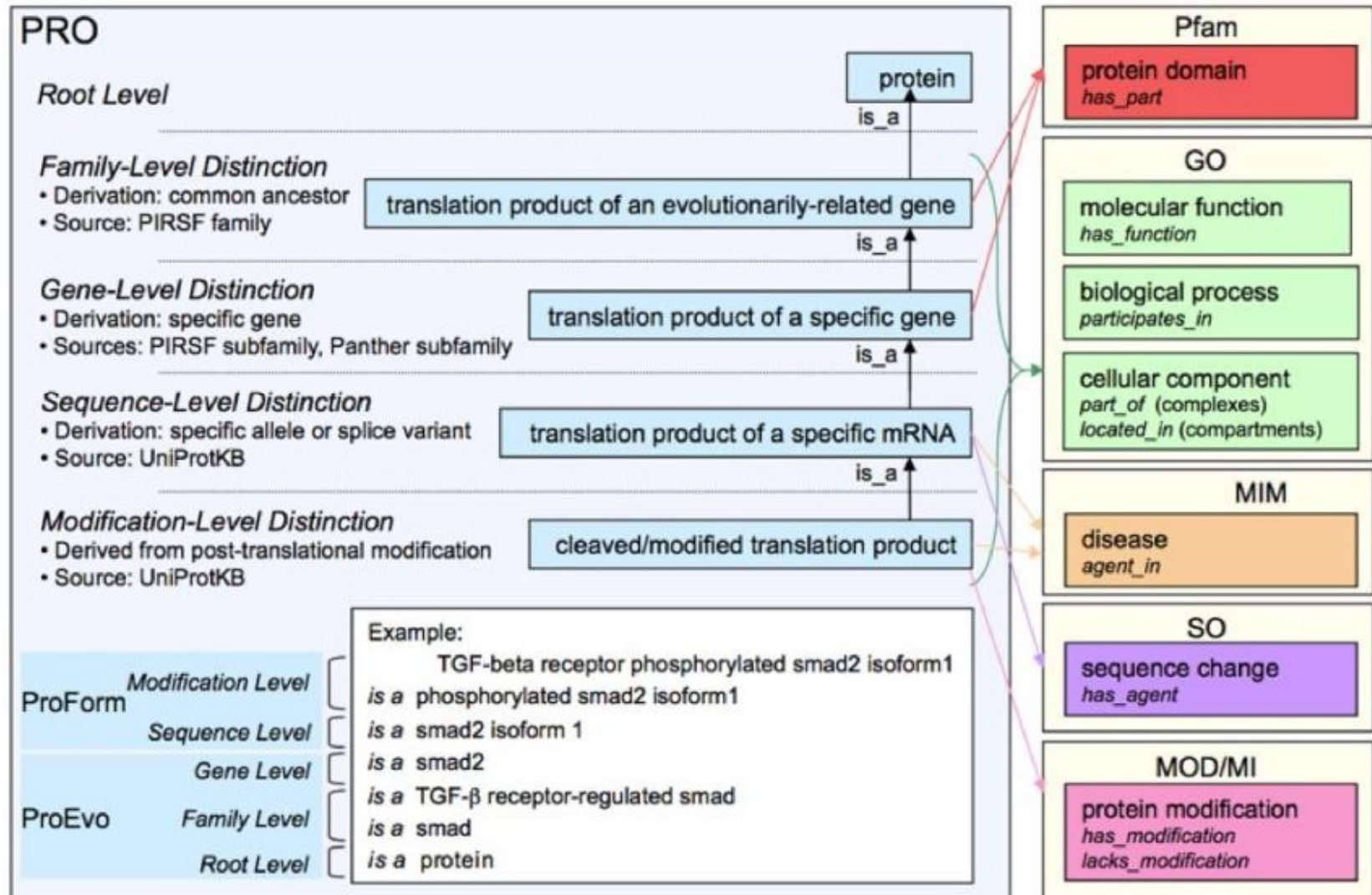
[Search again as text](#)

Resource	Hits	Category	Comment
PROSITE	4 hits	pr	PROSITE motif matches in submitted sequence
SWISS-MODEL Repository	1 hit	bi, pr, st	
UniProtKB	1 hit	pr	

[Search again as text](#)

PIR PRO Database

<http://pir.georgetown.edu/pro/pro.shtml>



PFAM

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families.

For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures
- PFAM database may search for similarity to a query protein sequence.
- PFAM may also be used to analyze proteomes and domain architectures.
- <https://pfam.sanger.ac.uk/>

Pfam is a database of two parts, the first is the curated part of Pfam containing over 5193 protein families (**Pfam-A**). Pfam-A comprises manually crafted multiple alignments and profile-HMMs .

To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called **Pfam-B**. This contains a large number of small families taken from the **PRODOM** database that do not overlap with Pfam-A.

Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

InterPro

“InterPro is an integrated documentation resource for protein families, domains & sites.

InterPro combines a number of databases that use different methodologies & a varying degree of biological information on well-characterised proteins to derive protein signatures.

By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful integrated database & diagnostic tool.

ProDom

- ProDom is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases.
- ProDom (Corpet et al., 2000) is a database of protein domain families automatically generated from SWISSPROT and TrEMBL sequence databases (Bairoch and Apweiler, 2000) using a novel procedure based on recursive PSI-BLAST searches (Altschul et al., 1997).



ExPASy Proteomics Server

<http://www.expasy.org/>



Swiss Institute of
Bioinformatics



Search for

ExPASy Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH

The ExPASy (**Ex**pert **P**rotein **A**nalysis **S**ystem) **proteomics** server of the **Swiss Institute of Bioinformatics** (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE ([Disclaimer](#) / [References](#) / [Linking to ExPASy](#)).

Databases

UniProtKB, PROSITE, HAMAP, SwissVar,
ViralZone, SWISS-MODEL Repository,
SWISS-2DPAGE, World-2DPAGE
Repository, MIAPEGelDB, ENZYME,
GlycoSuiteDB, UniPathway
[\[details\]](#) [\[full list\]](#)

Education & services

Downloads, Protein Spotlight,
Protéines à la «Une», e-proxemis,
Bioinformatics core facility for Proteomics
[\[full list\]](#)

Tools & Software

Proteomics tools, Blast, ScanProsite,
Melanie, MSight, Make2D-DB, SWISS-
MODEL, Swiss-PdbViewer
[\[full list\]](#)

Documentation

What's New?, E-mail alerts, UniProtKB
documentation, How to link to ExPASy,
Advanced search
[\[full list\]](#)

Latest News

Protein Spotlight - Dec 21, 2009 **String of intrusion**

When I was little, I used to wear small cotton shirts that were knitted by my grandmother. So? Well, onto them she sewed tiny nacre buttons you could never get hold of and which mesmerized me because of the different colours that shone off them depending on how you oriented them in the light. [\[more\]](#)

World-2DPAGE - Oct 23, 2009

New data uploaded into the [World-2DPAGE Repository](#). Currently, 113 maps for 16 species are available from the [World-2DPAGE Portal](#).

[\[more news\]](#) [\[SIB news\]](#)



Prosite Database

You are here: ExPASy CH > Databases > PROSITE



[Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)

Database of protein domains, families and functional sites

PROSITE consists of [documentation entries](#) describing protein domains, families and functional sites as well as associated [patterns](#) and [profiles](#) to identify them [[More details](#) / [References](#) / [Disclaimer](#) / [Commercial users](#)].

PROSITE is complemented by [ProRule](#), a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [[More details](#)].

Release 20.59, of 19-Jan-2010 (1567 documentation entries, 1308 patterns, 873 profiles and 874 ProRule)

PROSITE access

e.g: PDOC00022, PS50089, SH3, zinc finger

☐ add wildcard '*'

Browse:

- [by documentation entry](#)
- [by ProRule description](#)
- [by taxonomic scope](#)
- [by number of positive hit](#)

PROSITE tools

Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)



Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC
[[help](#)]:

☒ exclude [patterns with a high probability of occurrence](#)

- [ScanProsite](#) - advanced scan
- [PRATT](#) - allows to interactively generate conserved patterns from a series of unaligned proteins.
- [MyDomains - Image Creator](#) ^{new} - allows to generate custom domain figures.





ScanProsite Patterns and Profiles

<http://www.expasy.org/tools/scanprosite/>



Swiss Institute of
Bioinformatics



ExPASy Proteomics Server

Search for

[Databases](#) [Tools](#) [Services](#) [Mirrors](#) [About](#) [Contact](#)

You are here: [ExPASy CH](#) > [Tools](#) > [Pattern and profile searches](#) > [ScanProsite](#)



ScanProsite

[Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)

Sequence(s) to be scanned:	Motif(s) to scan for:
Enter: • UniProtKB(Swiss-Prot and TrEMBL) AC and/or ID (e.g. P00747, ENTK_HUMAN) • PDB identifier(s) • your own protein sequence(s) Clear ☒ Exclude motifs with a high probability of occurrence ☐ Do not scan profiles	Enter: • PROSITE AC and/or ID (e.g. PS50808, CHEB) • your own pattern(s) Clear Protein database(s): ☒ UniProtKB/Swiss-Prot ☒ including splice variants ☐ UniProtKB/TrEMBL ☐ PDB randomize databases no ☐ excluding fragments Filter(s): • On taxonomy: (e.g. Eukaryota; Escherichia coli;) • On description: (e.g. protease) Pattern option(s): • Allow at most 0 X sequence characters to match a conserved position in the pattern • Match mode greedy, overlaps, no includes
Output: • Format Graphical rich view • Show only sequences with at least 1 hit(s) Maximum of matched sequences 1000 START THE SCAN reset ☐ Show low level score ☐ Retrieve complete sequences Your e-mail:	

InterPro Scan

<http://www.ebi.ac.uk/InterProScan/>

The screenshot displays the InterPro Scan web interface. At the top, there is a navigation bar with links for Databases, Tools, Groups, Training, Industry, About Us, and Help. A search bar is also present with the text "Enter Text Here" and a "Go" button. Below the navigation bar, a sidebar on the left contains links for InterPro home, Text Search, InterProScan, Databases, Documentation, FTP Site, InterProScan Help (with sub-links for Help, FAQ, and README), InterProScan Programmatic Access, and Database Information (with sub-links for UniProt and UniParc).

The main content area is titled "InterProScan Sequence Search". It includes a paragraph explaining that the form allows querying a sequence against InterPro, with links to documentation, a README file, and FAQ's. A "Please Note" box states that job submissions should be limited to one sequence only as of Monday Feb 13, 2006, and to contact support for help with multiple sequences.

Below the note is a "Download Software" link. The main form is divided into several sections:

- YOUR EMAIL:** A text input field.
- RESULTS:** A dropdown menu set to "interactive".
- APPLICATIONS TO RUN:** A section with radio buttons for "Clear all" and "Check all". Below are checkboxes for various tools: BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, ScanRegExp, SuperFamily, SignalPHMM, TMHMM, HMMPanther, and Gene3D. All are currently checked.
- TRANSLATION TABLE (DNA/RNA only):** A dropdown menu set to "None".
- MIN. OPEN READING FRAME SIZE:** A dropdown menu set to "100".

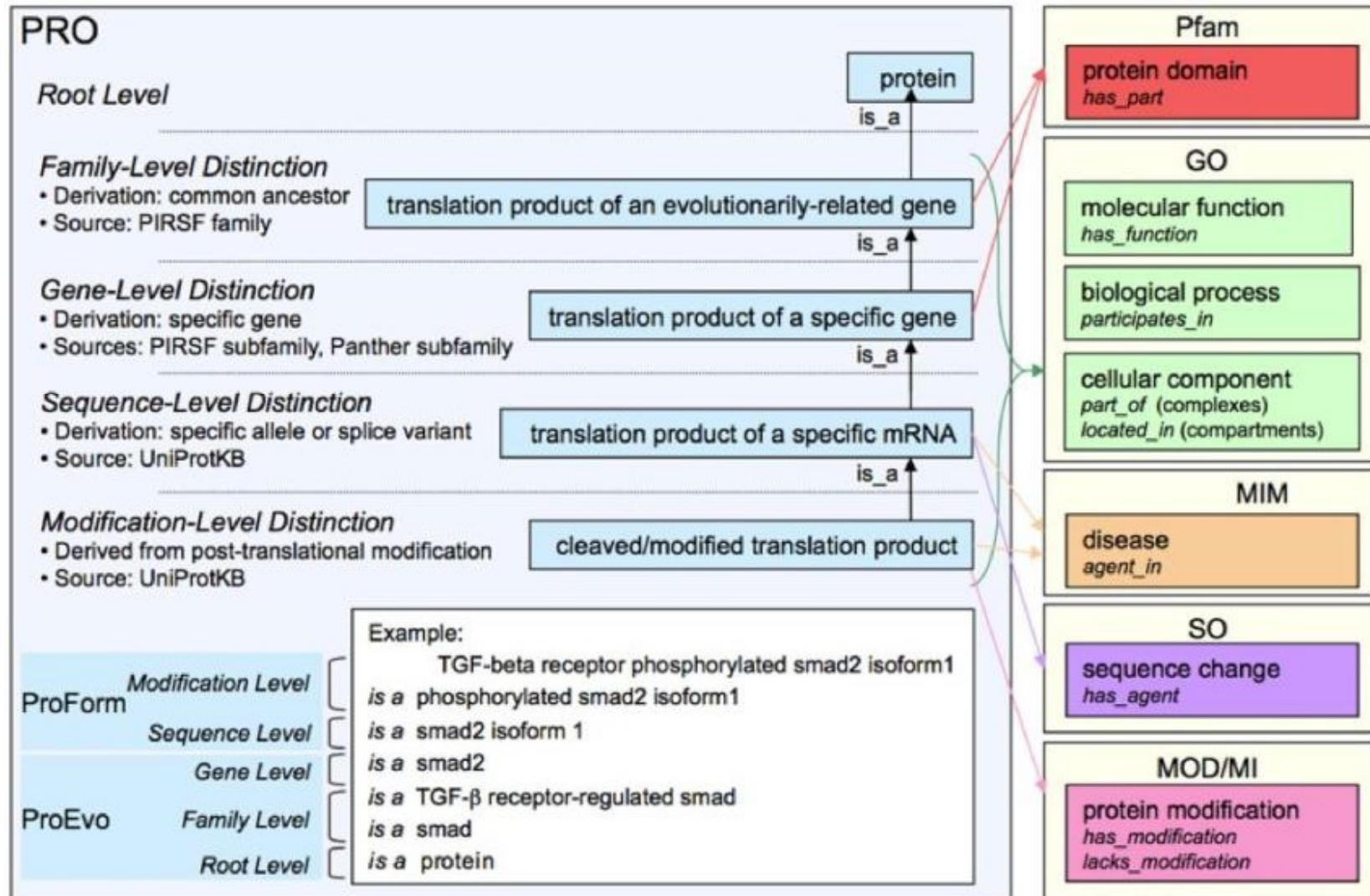
The bottom section is for entering the sequence. It has a label "Enter or Paste a" followed by a dropdown menu set to "PROTEIN" and the text "Sequence in any format:". A "Help" button is to the right. Below this is a large text area containing a protein sequence:

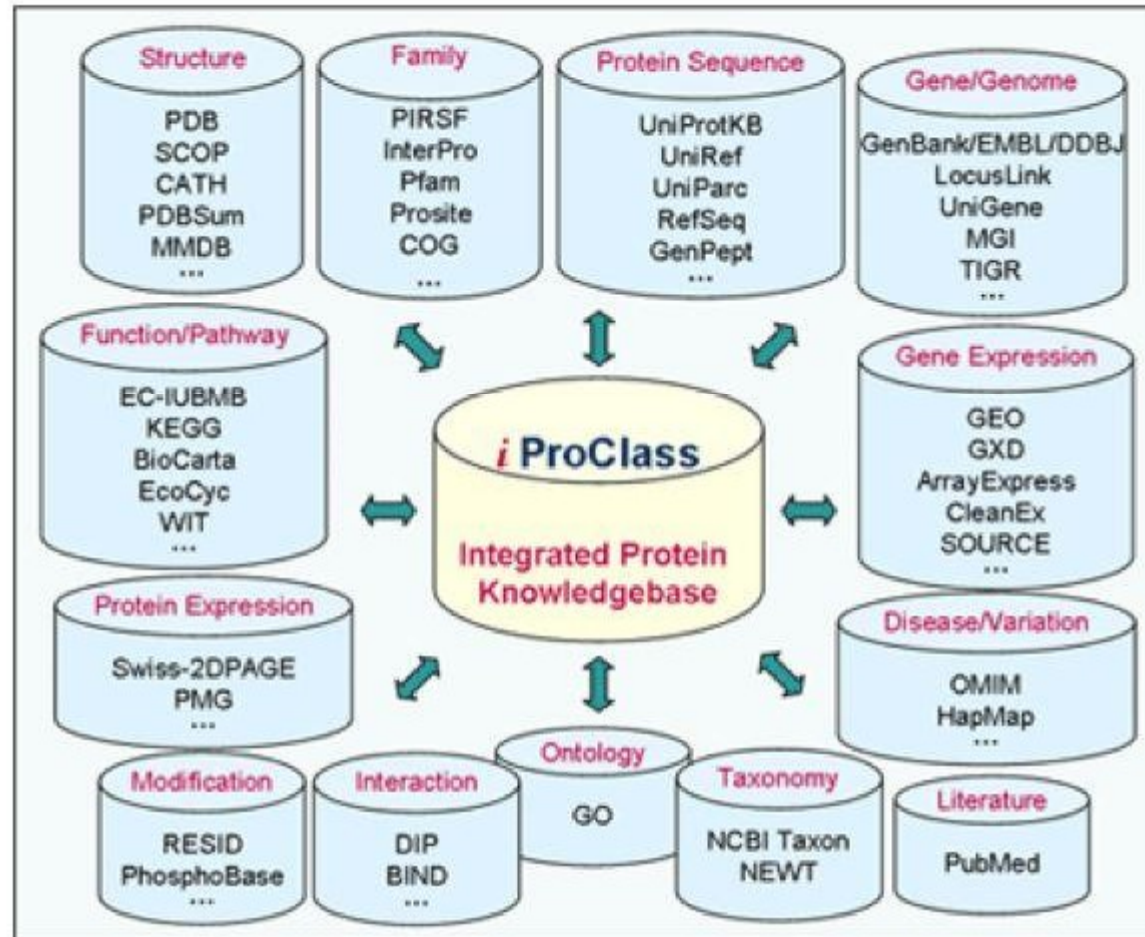
```
SFTLTNKNVIFVAGLGGIGLDTSKELLKRDKNLVILDRIENPAAIAELK
AINPKVTVTFYPYDVTVPPIAETTKLLKTIFAQLKTVDVINGAGILDDHQ
IERTIAVNYTGLVNTTTAILDFWDKRGGPGGIIICNIGSVTGFNAIYQVP
VYSGTAAAVVNFTSSLAKLAPITGVTAITVNPICITRTTLVHKFNSWLDV
E
PQVAEKLLAHPTQPSLACAENFVKAIELNQNGAIWKLDLGTLEAIQWTK
H
WDSGI
```

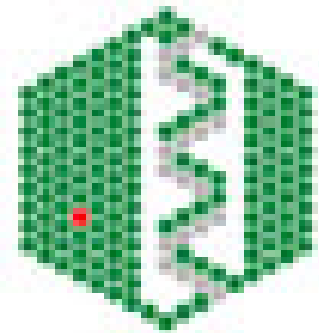
At the bottom, there is an "Upload a file:" section with a "Choose File" button and the text "no file selected". To the right are "Submit Job" and "Reset" buttons.

PIR PRO Database

<http://pir.georgetown.edu/pro/pro.shtml>







European Bioinformatics Institute

<http://www.ebi.ac.uk/>



EB-eye Search

All Databases

Enter Text Here

Go

Reset ?
Advanced Search

Give us feedback

DatabasesToolsEBI GroupsTrainingIndustryAbout UsHelpSite Index

Data Resources & Tools

- EMBL-BANK - UniProt - ArrayExpress - Ensembl - InterPro - PDB	- Genomes - Nucleotide Sequences - Protein Sequences - Macromolecular Structures - Small Molecules	- Gene Expression - Molecular Interactions - Reactions & Pathways - Protein Families - Enzymes	- Literature - Taxonomy - Ontologies - Patent Resources	- Sequence Similarity & Analysis - Pattern & Motif Searches - Structure Analysis - Text Mining - Downloads - Web Services
---	--	--	---	---



European Bioinformatics Institute

About the EBI

- Research - PhD Studies - Training - Industry Support - Group & Team Leaders - EBI Funders	- User Support - EBI Mission - People - Events at the EBI - Genome Campus Events - How to Find us - Jobs
---	--

EBI Hosted Project Websites

1000 Genomes - BioCatalogue - BioSapiens - E-MeP - EGA - ELIXIR - EMBRACE - EMERALD	- ENFIN - FELICS - IMPACT - INSDC - LRG - SPINE - SYMBIOmatics
---	--

Latest News

Open access drug discovery database launches with half a million compounds
18 January 2010
[ChEMBLdb](#), a vast online database of information on the properties and activities of drugs and drug-like small molecules and their targets, launches today with information on over half a million compounds. The data lie at the heart of translating information from the human genome into successful new drugs in the clinic... [more](#)

Research Highlights

EMBL-EBI articles are top of the list
20 November 2009
Articles on three resources hosted by EMBL-EBI ([PDB](#), [Ensembl Genomes](#) and [Gene Expression Atlas](#)) are highlighted as featured articles in the latest Database issue of Nucleic Acids Research. Featured articles are selected by the journal's Executive Editors based upon their originality, significance and scientific excellence ... [more](#)

Events



ExPASy Proteomics Server

<http://www.expasy.org/>



Swiss Institute of
Bioinformatics



Search for

ExPASy Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH

The ExPASy (**Ex**pert **P**rotein **A**nalysis **S**ystem) **proteomics** server of the **Swiss Institute of Bioinformatics** (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE ([Disclaimer](#) / [References](#) / [Linking to ExPASy](#)).

Databases

UniProtKB, PROSITE, HAMAP, SwissVar,
ViralZone, SWISS-MODEL Repository,
SWISS-2DPAGE, World-2DPAGE
Repository, MIAPEGelDB, ENZYME,
GlycoSuiteDB, UniPathway
[\[details\]](#) [\[full list\]](#)

Education & services

Downloads, Protein Spotlight,
Protéines à la «Une», e-proxemis,
Bioinformatics core facility for Proteomics
[\[full list\]](#)

Tools & Software

Proteomics tools, Blast, ScanProsite,
Melanie, MSight, Make2D-DB, SWISS-
MODEL, Swiss-PdbViewer
[\[full list\]](#)

Documentation

What's New?, E-mail alerts, UniProtKB
documentation, How to link to ExPASy,
Advanced search
[\[full list\]](#)

Latest News

Protein Spotlight - Dec 21, 2009 **String of intrusion**

When I was little, I used to wear small cotton shirts that were knitted by my grandmother. So? Well, onto them she sewed tiny nacre buttons you could never get hold of and which mesmerized me because of the different colours that shone off them depending on how you oriented them in the light. [\[more\]](#)

World-2DPAGE - Oct 23, 2009

New data uploaded into the **World-2DPAGE Repository**. Currently, 113 maps for 16 species are available from the **World-2DPAGE Portal**.

[\[more news\]](#) [\[SIB news\]](#)

InterPro Scan

<http://www.ebi.ac.uk/InterProScan/>

EMBL-EBI **EB-eye Search** All Databases Enter Text Here Go Reset Advanced Search

Databases Tools Groups Training Industry About Us Help Site Index

- InterPro home
- Text Search
- InterProScan
- Databases
- Documentation
- FTP Site

- InterProScan Help
 - Help
 - FAQ
 - README
- InterProScan Programmatic Access
- Database Information
 - UniProt
 - UniParc

EBI > Tools > Protein Functional Analysis

InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package ([Readme file](#) or [FAQ's](#)), or the InterPro [user manual](#) or [help pages](#).

Please Note: InterProScan job submissions should be limited to one sequence only. The system will no longer process 6 protein sequences simultaneously as of Monday Feb 13, 2006. Please contact [support](#) for help in submitting multiple sequences.

[Download Software](#)

YOUR EMAIL:

RESULTS: **interactive**

APPLICATIONS TO RUN ☐ Clear all ☒ Check all

☒ BlastProDom	☒ FPrintScan	☒ HMMPIR	☒ HMMPfam	☒ HMMSmart
☒ HMMTigr	☒ ProfileScan	☒ ScanRegExp	☒ SuperFamily	☒ SignalPHMM
☒ TMHMM	☒ HMMPanther	☒ Gene3D		

TRANSLATION TABLE (DNA/RNA only): **None**

MIN. OPEN READING FRAME SIZE: **100**

Enter or Paste a **PROTEIN** Sequence in any format: [Help](#)

```
SFTLTNKNVIFVAGLGGIGLDTSKELLKRDNLVILDRIENPAAIAELK
AINPKVTVTFFPYDVTVPPIAETTKLLKTIFAQLKTVDVLLINGAGILDDHQ
IERTIAVNYTGLVNTTTAILDFWDKRGGPGGIIICNIGSVTGFNAIYQVP
VYSGTAAAVVNFTSSLAKLAPITGVTAITVNPICITRTTLVHKFNWLDV
E
PQVAEKLLAHPTQPSLACAENFVKAIELNQNGAIWKLDLGTLEAIQWTK
H
WDSGI
```

Upload a file: no file selected

THANK YOU