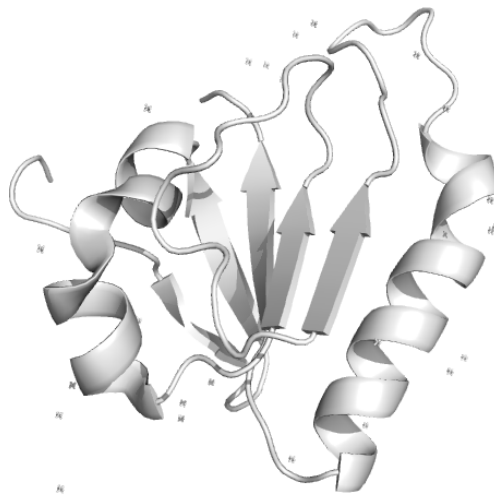


EMBO Practical Course: Computational Aspects of the Protein Target Selection, Protein Production Management and Structure Analysis Pipeline

Section: Searching Structure Databases
Hands-on Practical



Searching the PSI Structural Genomics Knowledgebase September 26, 2008

The PSI Structural Genomics Knowledgebase (PSI SGKB) contains structural and biochemical information in an effort to share knowledge that can be used to understand living systems and human disease.

Here are several exercises to introduce you to the PSI SGKB.

Questions that are indicated by an asterisk (*) will guide you through the site. Extra questions are available that will provide a deeper level of understanding the information presented to you. You can complete these questions if time allows.

Exercise 1: Searching the PSI-Nature Structural Genomics Knowledgebase by Amino Acid Sequence

Scenario: Your colleague has given you a novel gene from *C. elegans* to characterize. Since its genetic sequence was known, you have already converted it to the amino acid sequence of the gene product.

What sort of information is already available about this protein?

Step 1: Go to the PSI-Nature SGKB Gateway at <http://kb.psi-structuralgenomics.org/>

Step 2: Enter the target's one-letter-code protein sequence into the central Search box:

```
MPFRSMKDELIQKENVGGVVCCTEEFELKAAMNAMREVDWKNEGVEFFAVPMKDFTGTAP
RAEINEAVEFIESVASKGKTVYVHCKAGRTRSATVATCYLMKSRNWMSNVAWEFLLKDKRH
QVLLRNAAHWRTVNEYRRFLDSNSSSTGSSN
```

Step 3: Select the "by Sequence" radio button, and press Search. The Search Results Summary will be displayed.

Question 1.1:

* What types of information are available when you perform a sequence search?

Next: Click on the Database (DB) Reports link to display a detailed view of the query results.

We will now explore each tab and the information it presents in detail.

The Structures Tab

Using the DB report links will automatically open to the Structures tab. This first tab provides:

- links to view the 3D atomic coordinates (PDB file) and a link to download it
- the matching sequence identity (exact sequence similarity) percentage, with the ability to view the alignment
- any citation information with links to PubMed.

There is also a link to a glossary of terms in the top right corner which will explain many of the terms you will see with each entry. A glossary is available on all tabs.

Question 1.2:

- * a. How many protein structures did your sequence search find? _____
- * b. Look at their 'matching sequence identity values' listed in the report (I= x%). Compared to the sequence you provided, are any of these structures an identical match? _____

Step 3: Follow the PDB link on the first Structure result, 2G6Z. This will open a new window to the RCSB PDB's Structure Explorer page, which shows experimental details about this structure.

- * c. What kind of molecule is 2G6Z?
- * d. How similar in sequence is 2G6Z to your search sequence?

Extra questions (if time allows):

- d. Is this the structure of the entire protein?
- e. Look at the titles for the rest of the structures found to be similar to your query sequence. Do these proteins have anything in common? Switch to the Annotations tab to read more information.
- f. What is the source organism for these similar structures? How do they compare to your protein's source organism?

Next: The Annotations Tab

The Annotations Tab

The second tab displays the structural and functional descriptions and properties, or annotations, of the structures displayed in the Structures tab. Information is organized per protein chain since many structures contain multiple copies of protein chains.

The annotations search result is organized as follows:

- Links to general protein “encyclopedias” – PDBSUM, Proteopedia, TOPSAN
- Links to biochemical (functional) annotation databases – UniProt, EC, GO, ExpASy, KEGG, etc.
- Links to sequence-based annotation databases– Pfam, InterPro, Gene3D, etc.
- Links to structure-based annotation databases - CATH, SCOP, etc.

There is typically overlap between these databases. To see a complete list of Annotations available and what information they provide, see the [Annotation PSI Resource site](#) or the provided handout.

Question 1.3 – Functional Annotations:

* a. **Gene Ontology (GO)** annotations provide information on the cellular component, biological process and molecular function of gene products. Provide 1 “GO function” per category above for the first structure in the results summary, 2G6Z, chain C.

If time permits, click on the GO links you selected.

b. Provide brief definitions for the GO functions you listed in part (a) of this question:

Step 2: Click on the KEGG link for the first EC classification, EC 3.1.3.16, for chain C of 2G6Z.

KEGG, the Kyoto Encyclopedia for Genes and Genomes also catalogs biochemical information such as the reaction catalyzed.

* c. Give examples for the following; you may have to scroll down the page to see the information you need.

- Name any cofactors, metals, or ions commonly associated with this enzyme family 3.1.3.16.
- The names of 2 cellular pathways in which this family of enzymes is involved.

Step 3: Other structures contain additional annotations not available for 2G6Z. For the next few questions, return to the Annotations tab and scroll down to the annotation record for structure 1M3G, chain C.

Question 1.4 - Structural Annotations:

* a. CATH – a database which classifies the **Class, Architecture, Topology, and Homologous superfamilies of proteins**. What is the CATH classification for this domain, also found in 1M3G?

Class:

Architecture:

Topology:

Homologous superfamily:

* b. SCOP – a second database that catalogs the **Structural Classification Of Proteins**. Like CATH, SCOP also unites structure and function information. We have already identified a common sequence, and therefore a common domain among these proteins. Describe the 3-D “fold” of the domain that this sequence creates. (follow the SCOP link to find this detailed information).

Question 1.5 (Optional)

If time permits, return to the Pfam, CATH and SCOP links in the Annotations tab to answer these questions. Return to the Annotations tab after each one to then explore another link.

a. Proteins are made of domains linked together to create different functions. According to Pfam, in how many other protein arrangements has this domain been seen? (Also known as domain organizations or architecture... a shortcut is available from the left menu in Pfam) _____

b. In 1.4b, we list the CATH classification for this protein phosphatase. List 2 protein families that share the same Architecture? (level descriptions inside the CATH website are links)

d. List 2 protein superfamilies that also share the same fold?

In summary, by looking at the annotation results of experimentally-solved proteins similar in sequence to your query, we can make educated guesses about their structure or function.

Next, the Models Tab

The Models Tab

Theoretical homology models are created by fitting the sequence of a desired protein onto the structure of a similar (homologous) protein that is likely to fold into the same shape. This can be helpful in the absence of an experimentally determined structure.

The Models tab provides:

- a link to a searchable portal for such models, the Protein Models Portal resource
- displays how many homology models were found that are similar to your query sequence.

Within the Protein Models Portal summary, the following will be displayed:

- A summary of the matching sequence alignments against all sequences registered in UniProt, most comprehensive protein catalog. It appears as a series of red and blue lines. Experimental structure will appear in green.
- The UniProt ID for your search sequence
- List of the available homology models and where it came from (either PSI center or other models repository).

Showing the individual model displays more information in the Model Details page:

- View the alignment again on your search sequence (black line) with the template (blue overlay).
- Link to domain annotation, if applicable (InterPro)
- explore the 3-dimensional homology model of your search sequence with Astexviewer.
- view SCOP (if the structure was deposited before 2007) and CATH structural annotations of the structure whose shape your sequence has been fitted to, or “template structure”.

Step 1: Starting from the Models Summary page, click on the ‘view’ link. This will launch a new Query Result window that shows the multiple sequence alignments as blue lines against your red query sequence.

Question 1.6:

- * a. How many theoretical models have been created in total? By how many centers?
- * b. What experimentally determined structures were used as their templates, with their sequence identities?

Step 2: Look for the homology model that shares the highest sequence identity, and select it.

Explore this page freely if time permits. You can launch the interactive Astexviewer display by following the ‘display’ link to the right of the model image. Rotate the molecule with left click + drag.

Next, the Targets Tab

The Targets Tab

Proteins sequences that have been targeted by the PSI and other worldwide structural genomics efforts for structure determination are registered in a database called TargetDB. The PSI SGKB can compare the search sequence to the entire TargetDB library and report the protein production and structure solution progress for it and/or similar sequences.

Each protein sequence target has a name assigned by each center, along with its percent identity to the query sequence.

In the Targets view, there are direct links to the following:

- The TargetDB entry (by TargetID)
- View the matching sequence alignment
- Target status
- Any additional supporting annotations.

Question 1.8:

* a. How many TargetDB entries match your exact query sequence? _____

* b. List the protein properties relating to our search sequence.

Number of Residues:

Mol. Weight:

Avg. Hydropathy Score:

Charge:

pI Value:

* c. How much progress was made in obtaining the structure of this sequence? List the furthest experimental progress that was made.

Next, the Protocols Tab

The Protocols Tab

The Protocols Tab extends the Targets report by providing detailed protocols submitted by PSI scientists at every step of the target's structure determination pipeline, including reasons the work was stopped.

Step 1: Click on the Protocols Tab.

Question 1.9:

- * a. How many protocols are available for targeted proteins similar to your query sequence? _____
- * b. How many for your exact sequence of interest? _____

Step 2: Explore one of the trials, Target:4, by clicking on the TargetID value link.

Question 1.10:

- * a. Since they were trying to solve the structure of this protein by x-ray crystallography, what special method did they use to overexpress Target:4 in order to facilitate structural determination?

If time permits, answer the following question:

Step 3: Return to the Protocols Tab, and scroll down to entry NYSGXRC-8708a, where attempts at purifying a dual specificity protein phosphatase were more successful.

- b. What were the final yield, concentration, oligomeric state of the final protein sample?
- c. How many purification steps were required to purify this protein? _____

The Materials Tab

The Materials tab displays the related high-quality DNA clone samples, including empty DNA vectors like those used at the PSI centers, available for purchase through the PSI Materials Repository.

We will explore this feature in the next exercise.

Next, Exercise 2: searching by PDB ID.

Exercise 2: Searching the PSI-Nature Structural Genomics Knowledgebase by Structure (PDB ID)

If you already know that the structure for a particular protein sequence exists, you can search the PSI SGKB for information specific for that PDB entry alone.

Step 1: In this example, use the “example query” on the PSI SGKB search form. Be certain that you have selected ‘by Structure (PDB ID)’ as the search option.

Question 2.1:

* How many Structures did your sequence search find? _____

Step 2: Click on the Annotations tab.

Question 2.2:

* How many Annotation links are available for this structure? _____

There are no functional annotation links available here as in the previous exercise. This protein happens to be one of those solved as part of the PSI structural genomics effort, and the biochemical function, i.e. the reaction it catalyzes, is unknown.

Step 3: Since the experimentally determined structure is available, skip the Models and Targets tabs and select the Protocol tab.

Question 2.3:

* c. Which steps of the protein production pipeline have available protocols?

Step 4: Select the Materials Tab.

* d. How many Materials are available for this entry? _____

If time permits, answer the following question:

e. What is the difference between these two target DNA clones available for purchase?

By combining the information presented here, you can quickly obtain the necessary materials and protocols to start on such a project, minimizing a trial-and-error period.

Next, Exercise 3, searching the PSI SGKB by text query

Exercise 3: Searching by plain text

The final search option is to search by plain text. This is less restrictive than a “keyword search”, finding all instances of your search query from the following locations:

- The PSI Centers’ web sites and the Technologies Resource, which catalogs all technologies developed by the PSI
- Structural papers written by the PSI
- Methodology papers written by the PSI
- And the Annotations Resource again, if the term is a biological one.

For this exercise, return to the PSI SGKB homepage, and write “TEV Protease” in the search box. Be certain that you have selected ‘by plain text’ before pressing the Search button.

Step 1: First, a summary report will display how many hits result from this search. Select the ‘Keyword Search’ button to display these results in detail.

Question 3.1:

- * a. How many structures of TEV Protease were solved? _____
How many methodology publications were written about TEV Protease? _____
- * b. What annotations are presented for this particular search query?

Step 2: If time permits, explore the links that are listed in the Site Search.

Therefore, while a text search is not as specific as a sequence or structure search, it still provides a starting point to navigate the PSI SGKB effectively.

Exercise 4: Explore the site freely.

Now that you understand the various search options and what sorts of information they provide, we encourage you to search the PSI SGKB on your own.

Thank you, and enjoy! If you have any questions, feel free to contact the PSI SGKB at psi-sgkb@nature.com.

Annotation Resources linked within the PSI SGKB (listed alphabetically)

- [Astral](#) - Compendium that provides databases and tools useful for analyzing protein structures and their sequences. Mostly derived upon SCOP structural database (see below).
- [BMRB](#) (Biological Magnetic Resonance Data Bank) - A repository for data from NMR spectroscopy on proteins, peptides, nucleic acids, and other biomolecules
- [BRENDA](#) - Enzyme Information System which contains basic chemical, nomenclature, and biochemical/kinetic information.
- [CATH](#) – a hierarchical domain classification of protein structures in the Protein Data Bank. For any given structure classified in the database, CATH gives you information on the structure and function of that protein. The evolutionary relationships involving the structure of interest and other proteins in the database can also be determined.
- [DIP](#) - a database that catalogs experimentally determined interactions between proteins.
- [EC - Enzyme Nomenclature Committee of the IUBMB](#) - Recommendations on Biochemical & Organic Nomenclature, Symbols & Terminology; contains the recognized nomenclature of enzymes based on their function.
- [EC2PDB - Enzyme Structure Database](#) at the [European Bioinformatics Institute](#) (EBI). It contains the known enzyme structures that have been deposited in the [Protein Data Bank](#), organized by Enzyme Classification (EC) number.
- [Ensembl](#) - a joint project between EMBL/EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.
- [Entrez](#) - Life Sciences cross-database search Engine at NCBI/NIH. Many resources related to nucleic acid sequence, protein sequence, structure, and beyond.
- [Evolutionary Trace Viewer \(ETV\)](#) - a method to view and run Evolutionary Traces, elucidating evolutionarily conserved amino acids within protein families.
- [Evolutionary Trace Report Maker](#) - creates an integrated report about the evolutionary propensity of individual residues.
- [Expert Protein Analysis System](#) (ExpASy). – a proteomics server of the Swiss Institute of Bioinformatics (SIB) that provides databases and tools for the analysis of protein sequences and structures as well as 2-D PAGE.
- [Gene Ontology](#) (GO) - functional assignment for the proteins and [Gene Ontology Browser](#) at [European Bioinformatics Institute](#) (EBI). Data is organized into three categories, cellular component, biological process, and molecular function.
- [Gene3D](#) - Reliable Structural Markup of the Protein Universe; provides protein family, proteome, domain and features, functional annotation, and protein-protein interaction information.
- [GeneDB](#) - the Wellcome Trust Sanger Institute Pathogen Sequencing Unit (PSU) provides access to the latest sequence data and annotation/curation of 37 organisms sequenced by the Pathogen Sequencing Unit.
- [GPSS](#) - provides analysis of all MCSG functionally annotated surfaces.. It is a single resource to identify and explore geometrically defined surfaces and protein-hetero atom (ligands, metals, peptides and DNA) contact surfaces. Annotation can be viewed through the GPSS web interface or using a plug-in to the PyMOL molecular visualization program.
- [Integrated relational Enzyme database](#) (IntEnz)- enzyme nomenclature and classification of enzyme-catalyzed reactions.

- [InterPro](#) - a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.
- [iProClass](#) - contains links to over 90 biological databases, including databases for protein families, functions and pathways, interactions, structures and structural classifications, genes and genomes, ontologies, literature, and taxonomy.
- [Kyoto Encyclopedia of Genes and Genomes \(KEGG\)](#) - a database of biological systems, consisting of genetic building blocks of genes and proteins, chemical building blocks of both endogenous and exogenous substances, molecular wiring diagrams of interaction and reaction networks, and hierarchies and relationships of various biological objects.
- [NCBI - Taxonomy Browser](#). – complete listing of organism taxonomy.
- [PDBSUM](#) - provides a variety of structure and function annotation.
- [Pfam](#) – sequence-based collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
- [PlasmoDB](#) - genomic and proteomic data for different species of the parasitic eukaryote Plasmodium, the cause of Malaria.
- [PRINTS](#) - compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a SWISS-PROT/TrEMBL composite.
- [ProDom](#) - a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases.
- [ProFunc](#) - server that identifies likely biochemical function of a protein from its three-dimensional structure. This has been done for MCSG structures.
- [ProLinks](#) - A database of proteins having functional linkages to an input protein.
- [ProKnow](#) - A server to suggest functions and annotations for a protein of known structure.
- [ProSite](#) - describes protein domains, families and functional sites as well as associated patterns and profiles to identify them.
- [ProtoNet](#) - global classification of the proteins, from the SWISS-PROT (UNIPROT) database into hierarchical clusters.
- [SAVES](#) - Structure Analysis and Verification server. Provides links to online versions of crystallographic validation programs for real-time model quality assessment.
- [SCOP](#) - Structural Classification of Proteins resource. Provides a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- [TIGR](#) - The Institute for Genomic Research, now the Craig T Venter Institute. It contains a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data for microbes, plants and humans.
- [TOPSAN](#) - a wiki-based project where automated target annotations are integrated with structure determination summary reports for all PSI structures. TOPSAN provides access to outside collaborators to comment on and/or annotate a structure through an open mechanism similar to Wikipedia.
- [UniProt](#) (Universal Protein Resource) - the world's most comprehensive catalog of information on proteins.
- [WormBase](#) - The Biology and Genome of *C. elegans*