



## Data Sources

---

### Getting Protein Data from the Internet

Data Sources

1



## Introduction

---

- Structural bioinformatics deals with the representation of molecular structural data along and its subsequent storage, retrieval, analysis and visualization.
- Objectives:
  - Design geometric algorithms for manipulating structural data.
  - Design pattern analysis algorithms for detecting structural patterns in macromolecules.
  - Applying these algorithms to further the understanding of cellular processes and to accomplish therapeutic interventions.

Data Sources

2



## Structural Data

---

- The study of protein domains and their functionality requires a reliable source of structural data for both macromolecules and ligands.
- We will review the following sources of molecular data:
  - There are many others...
    - PDB
    - PDBsum
    - MSD
    - Relibase
    - GRASS
    - SCOP
    - CATH
    - PubChem
    - HIVSDB

Data Sources

3



## PDB: The Protein Data Bank

---

<http://www.rcsb.org/pdb>

- The PDB is a large protein public domain repository maintained by the RCSB
  - Research Collaboratory for Structural Bioinformatics
  - “a non-profit consortium dedicated to improving our understanding of the function of biological systems through the study of the 3-D structure of biological macromolecules”
  - Most of the structures have been determined by x-ray crystallography.
    - More recently, NMR data has been added.
  - Currently (Jan. 03, 2012) there are 78,304 entries.
    - This represents less than 10-15% of all known proteins.
    - Only deposited data is available to the public...
      - One year ago: Jan. 04 there were 70,303 entries.

Data Sources

4



Jan. 4, 2012:

| Exp.Method          | Proteins     | Nucleic Acids | Protein/NA Complexes | Other     | Total        |
|---------------------|--------------|---------------|----------------------|-----------|--------------|
| X-RAY               | 64014        | 1335          | 3134                 | 2         | 68485        |
| NMR                 | 8056         | 967           | 185                  | 7         | 9215         |
| ELECTRON MICROSCOPY | 273          | 22            | 101                  | 0         | 396          |
| HYBRID              | 42           | 3             | 2                    | 1         | 48           |
| other               | 138          | 4             | 5                    | 13        | 160          |
| <b>Total</b>        | <b>72523</b> | <b>2331</b>   | <b>3427</b>          | <b>23</b> | <b>78304</b> |

A 11.4% increase in one year.

Jan. 2011:

| Exp.Method          | Proteins     | Nucleic Acids | Protein/NA Comp | Other     | Total        |
|---------------------|--------------|---------------|-----------------|-----------|--------------|
| X-RAY               | 57073        | 1252          | 2736            | 17        | 61078        |
| NMR                 | 7599         | 933           | 167             | 7         | 8706         |
| ELECTRON MICROSCOPY | 233          | 22            | 85              | 0         | 340          |
| HYBRID              | 26           | 1             | 1               | 1         | 29           |
| other               | 129          | 4             | 4               | 13        | 150          |
| <b>Total</b>        | <b>65060</b> | <b>2212</b>   | <b>2993</b>     | <b>38</b> | <b>70303</b> |

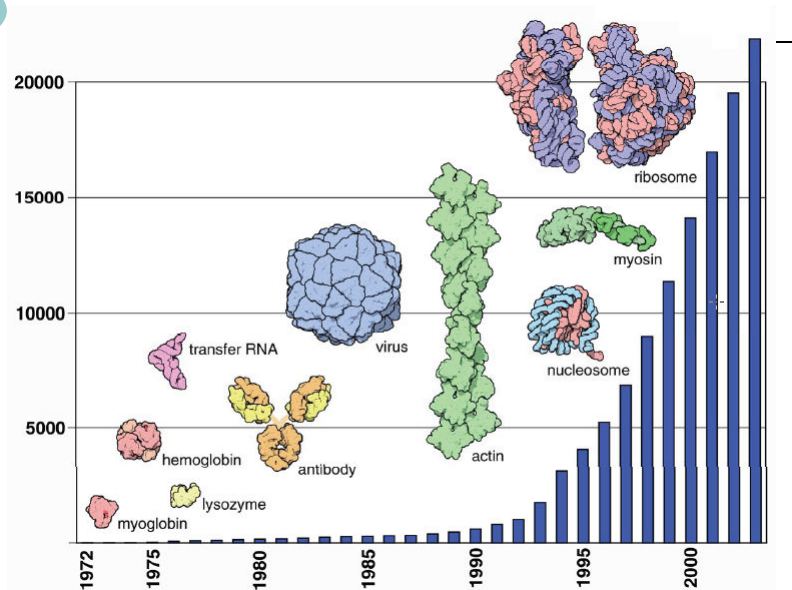
"Other" is mainly carbohydrates.

Data Sources

5



## Growth of the PDB



Data Sources

6





- In response to a search the PDB provides information related to a particular protein crystallization study:
  - Title of research paper
  - Authors
  - History
  - Experimental method
  - Parameters and Unit cell specification
  - Molecular description
  - Various classifications (SCOP, CATH, ... )
  - The full PDB file is available.

Data Sources

9



## PDB Limitations

---

- For X-ray structures, the resolution is not high enough to accurately position the hydrogen atoms.
  - So only the "heavy" atoms are given in the file.
- No connection and bond data.
  - There is a CONECT record that for each atom in the chemical component, lists how many and to which other atoms that atom is bonded.
  - Other bonds have to be inferred from the atom name.
- Many PDB files are inconsistent, informal and not fully checked for errors.


Data Sources

10



## PDBsum: The PDB Summary

---

-  :
- Summaries and analyses of PDB structures.
- <http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/>
- As noted on the site:
  - It provides schematic diagrams of the molecules in each structure and of the interactions between them. Entries are accessed by their PDB code, by simple text search, or through any of the browse options in the left column.
- There are many links to a variety of other sites dealing any particular protein that you want to specify using PDB ID.

Data Sources

11



## MSD: The Macromolecular Structure Group

---

- The Macromolecular Structure Group (MSD)
  - A European project for the collection, management and distribution of data describing macromolecular structures.
  - <http://www.ebi.ac.uk/msd-srv/docs/dbdoc/refaindex.html>
  - Macromolecules are stored in their database using a relational database approach.
  - Their homepage shows several examples of SQL for accessing various types of data related to proteins and ligands.

Data Sources

12



## SCOP: Structural Classification Of Proteins

---

- Introduction
  - Proteins have structural similarities
    - This may be due to common evolutionary origins or possibly convergent evolution from origins that are not really related.
  - SCOP: (<http://scop.mrc-lmb.cam.ac.uk/scop/> )
    - has been created by visual inspection
      - but with the help of various software tools
    - hierarchically organizes proteins according to their structure and evolutionary ancestry
    - also provides entry links for coordinates, structure images, literature references, etc.
    - June 2009: 38,221 PDB Entries, 110,800 Domains.

Data Sources

13



## Hierarchical Levels in SCOP

---

- Classes → Folds → Superfamilies → Families
- The unit of categorization is the protein *domain*.
  - Not the protein itself...
    - Small proteins have a single domain, so in this case categorization is by the protein.
    - Otherwise, large proteins may have more than one domain and these are categorized on an individual basis.
      - More precisely, when a protein is to be placed into the SCOP hierarchy it is first separated into domains.

Data Sources

14



## SCOP Levels: Classes

---

There are 11 classes:

- [All alpha proteins](#)
- [All beta proteins](#)
- [Alpha and beta proteins \( \$\alpha / \beta\$ \)](#)  
*Mainly parallel beta sheets (beta-alpha-beta units)*
- [Alpha and beta proteins \( \$\alpha + \beta\$ \)](#)  
*Mainly antiparallel beta sheets (segregated alpha and beta regions)*
- [Multi-domain proteins \( \$\alpha\$  and  \$\beta\$ \)](#)  
*Folds consisting of two or more domains belonging to different classes*
- [Membrane and cell surface proteins and peptides](#)  
*Does not include proteins in the immune system*
- [Small proteins](#)  
*Usually dominated by metal ligand, heme, and/or disulfide bridges*
- [Coiled coil proteins](#)  
*Not a true class*
- [Low resolution protein structures](#)  
*Not a true class*
- [Peptides](#)  
*Peptides and fragments. Not a true class*
- [Designed proteins](#)  
*Experimental structures of proteins with essentially non-natural sequences. Not a true class*

Data Sources

15



## SCOP Levels: Classes (cont.)

---

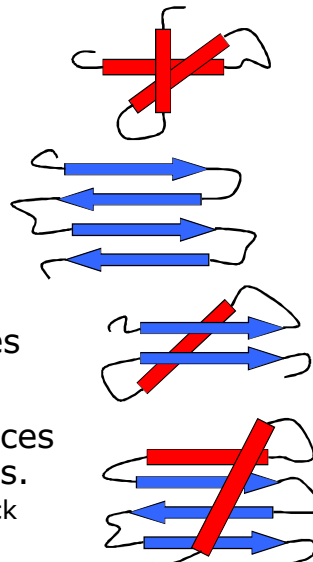
- A class contains domains with similar global characteristics
  - Not necessarily any evolutionary relation.
  - The next slide gives details for the first four classes.
  - Notes:
    - Membrane proteins are a separate class.
    - *Small* proteins are stabilized by disulfide bridges or by metal ligands in lieu of hydrophobic cores.

Data Sources

16

## SCOP Levels: Classes (cont.)

- Class  $\alpha$ :
  - A bundle of  $\alpha$  helices connected by loops
- Class  $\beta$ 
  - Anti-parallel  $\beta$  sheets
- Class  $\alpha / \beta$ 
  - Mainly parallel  $\beta$  sheets with intervening  $\alpha$  helices
- Class  $\alpha + \beta$ 
  - Mainly segregated  $\alpha$  helices and anti-parallel  $\beta$  sheets.
    - Note that the helices pack against the sheet.

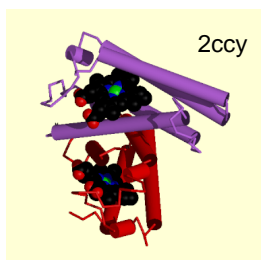


Data Sources

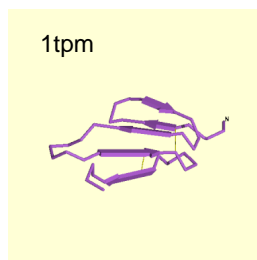
17

## SCOP Levels: Classes (Some examples)

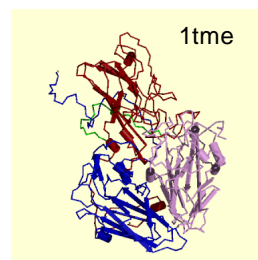
All alpha:



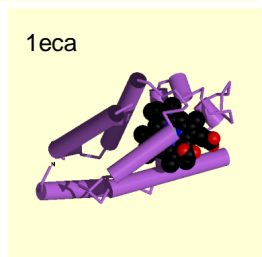
All beta:



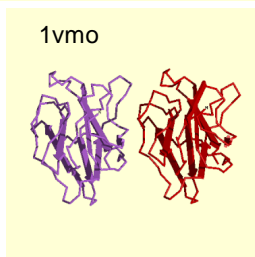
Alpha-beta:



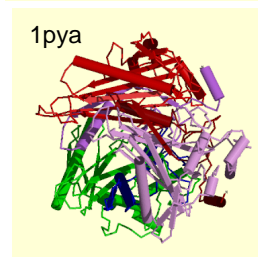
1eca



1vmo



1pya



Data Sources

18



## SCOP Levels: Folds

---

- Definition:
  - A particular topological arrangement of alpha helices, beta strands, and loops in 3D space is called a fold.
    - There are over 1,195 known folds (Feb. 2009).
    - Many different protein sequences can produce the same fold.
    - A short description of the most significant structural features is used as the name of the fold.

Data Sources

19



## SCOP Levels: Superfamilies

---

- A superfamily contains a clear *structural* similarity.
  - The stress on *structural* similarity is to emphasize the fact that they might have low sequence identity.
  - Release 1.75 contains 1,962 superfamilies.
  - There is a *probable* common evolutionary origin.
  - They share a common fold.
  - They usually perform similar functions.

Data Sources

20



## SCOP Levels: Families

---

- A family contains a clear sequence homology.
  - There is a clear evolutionary relationship.
  - The pair-wise residue identity (sequence similarity) is at least 30%.
    - In some cases, this figure may be lower (say 15%) but the proteins are still put in the same family because they have very similar functions and structures.
    - Release 1.75 contains 3,902 families.
      - An example of this is the family of globins.

Data Sources

21



## Classification Statistics

---

- **SCOP** release **1.75**  
38,221 PDB Entries (Feb. 2009). 110,800 Domains.

| Class                              | Number of folds | Number of super families | Number of families |
|------------------------------------|-----------------|--------------------------|--------------------|
| All alpha proteins                 | 284             | 507                      | 871                |
| All beta proteins                  | 174             | 354                      | 742                |
| Alpha and beta proteins (a/b)      | 147             | 244                      | 803                |
| Alpha and beta proteins (a+b)      | 376             | 552                      | 1055               |
| Multi-domain proteins              | 66              | 66                       | 89                 |
| Membrane and cell surface proteins | 58              | 110                      | 123                |
| Small proteins                     | 90              | 129                      | 219                |
| Total                              | 1195            | 1962                     | 3902               |

Data Sources

22



## Importance of SCOP

---

- Consider some protein P.
  - Structural bioinformatics query:
    - Is there a protein that is nearby to P in the *structure space*?
  - Molecular biologist query:
    - Is there a another protein with a fold that is similar to the fold of P and so might interact with the same ligand that binds with P?

Data Sources

23



## Entering SCOP at the Top of the Hierarchy

---

(This hierarchical descent was done with SCOP version 1.63)

**Root: scop**

○ **Classes:**

1. [All alpha proteins](#) (171)
2. [All beta proteins](#) (119)
3. [Alpha and beta proteins \(a/b\)](#) (117)  
*Mainly parallel beta sheets (beta-alpha-beta units)*
4. [Alpha and beta proteins \(a+b\)](#) (224)  
*Mainly antiparallel beta sheets (segregated alpha and beta regions)*
5. [Multi-domain proteins \(alpha and beta\)](#) (39)  
*Folds consisting of two or more domains belonging to different classes*
6. [Membrane and cell surface proteins and peptides](#) (34)  
*Does not include proteins in the immune system*
7. [Small proteins](#) (61)  
*Usually dominated by metal ligand, heme, and/or disulfide bridges*
8. [Coiled coil proteins](#) (6)  
*Not a true class*
9. [Low resolution protein structures](#) (18)  
*Not a true class*
10. [Peptides](#) (101)  
*Peptides and fragments. Not a true class*
11. [Designed proteins](#) (37)  
*Experimental structures of proteins with essentially non-natural sequences. Not a true class*

Data Sources

24



## Clicking on: [All alpha proteins](#)

### Class: All alpha proteins

#### Lineage:

1. Root: [scop](#)
2. Class: [All alpha proteins](#) [46456]

#### Folds:

1. [Globin-like](#) [46457] (2)  
*core: 6 helices; folded leaf, partly opened*
2. [Long alpha-hairpin](#) [46556] (11)  
*2 helices; antiparallel hairpin, left-handed twist*
3. [Type I dockerin domain](#) [63445] (1)  
*tandem repeat of two calcium-binding loop-helix motifs, distinct from the EF-hand*
4. [LEM/SAP HeH motif](#) [63450] (4)  
*helix-extended loop-helix; parallel helices*
5. [Cytochrome c](#) [46625] (1)  
*core: 3 helices; folded leaf, opened*
6. [DNA/RNA-binding 3-helical bundle](#) [46688] (12)  
*core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down*
7. [Another 3-helical bundle](#) [81602] (2)  
*topologically similar to the DNA/RNA-binding bundles; distinct packing*
8. [RuvA C-terminal domain-like](#) [46928] (6)  
*3 helices; bundle, right-handed twist*
9. ....

Data Sources

25



## Clicking on: [DNA/RNA-binding 3-helical bundle](#)

### Fold: DNA/RNA-binding 3-helical bundle

- *core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down*

#### Lineage:

1. Root: [scop](#)
2. Class: [All alpha proteins](#) [46456]
3. Fold: [DNA/RNA-binding 3-helical bundle](#) [46688]  
*core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down*

#### Superfamilies:

1. [Homeodomain-like](#) [46689] (11)  
*consists only of helices*
2. [Methylated DNA-protein cysteine methyltransferase, C-terminal domain](#) [46767] (1)
3. [ARID-like](#) [46774] (1)  
*contains extra helices at both N- and C-termini*
4. ["Winged helix" DNA-binding domain](#) [46785] (36)  
*contains a small beta-sheet (wing)*
5. [C-terminal effector domain of the bipartite response regulators](#) [46894] (3)  
*binds to DNA and RNA polymerase; the N-terminal, receiver domain belongs to the CheY family*
6. [TrpR-like](#) [48295] (2)  
*contains an extra shared helix after the HTH motif*
7. ....

Data Sources

26



Clicking on: [Homeodomain-like](#)

## Superfamily: Homeodomain-like

- *consists only of helices*

### Lineage:

1. Root: [scop](#)
2. Class: [All alpha proteins](#) [46456]
3. Fold: [DNA/RNA-binding 3-helical bundle](#) [46688]  
*core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down*
4. Superfamily: [Homeodomain-like](#) [46689]  
*consists only of helices*

### Families:

1. [Homeodomain](#) [46690] (23)
2. [Recombinase DNA-binding domain](#) [46728] (5)
3. [Myb](#) [46739] (4)
4. [GARP response regulators](#) [81683] (1)  
*plant myb-related DNA binding motif*
5. [DNA-binding domain of human telomeric protein, hTRF1](#) [46745] (1)
6. [Paired domain](#) [46748] (3)  
*duplication: consists of two domains of this fold*
7. .....

Data Sources

27



Clicking on: [Paired domain](#)

## Family: Paired domain

- *duplication: consists of two domains of this fold*

### Lineage:

1. Root: [scop](#)
2. Class: [All alpha proteins](#) [46456]
3. Fold: [DNA/RNA-binding 3-helical bundle](#) [46688]  
*core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down*
4. Superfamily: [Homeodomain-like](#) [46689]  
*consists only of helices*
5. Family: [Paired domain](#) [46748]  
*duplication: consists of two domains of this fold*

### Protein Domains:

1. Pax-5 [68962]
  1. [Human \(\*Homo sapiens\*\)](#) [68963] (2)
2. Pax-6 [68936]
  1. [Human \(\*Homo sapiens\*\)](#) [46750] (1)
3. Paired protein (prd) [46751]
  1. [Fruit fly \(\*Drosophila melanogaster\*\)](#) [46752] (1)

Data Sources

28



Clicking on: Pax-6 [68936]

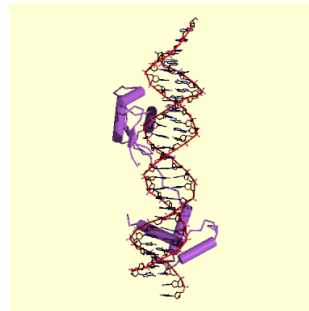
[Human \(\*Homo sapiens\*\)](#) [46750] (1)

---






## Protein: Pax-6 from Human (*Homo sapiens*)

### Lineage:

1. Root: [scop](#)
2. Class: [All alpha proteins](#) [46456]
3. Fold: [DNA/RNA-binding 3-helical bundle](#) [46688]  
*core: 3-helices; bundle, closed or partly opened, right-handed twist; up-and down*
4. Superfamily: [Homeodomain-like](#) [46689]  
*consists only of helices*
5. Family: [Paired domain](#) [46748]  
*duplication: consists of two domains of this fold*
6. Protein: Pax-6 [68936]
7. Species: [Human \(\*Homo sapiens\*\)](#) [46750]



### PDB Entry Domains:

1. [6pax](#) 
  1. [region a:1-68](#) [64758]  
  2. [region a:69-133](#) [64759]  

Data Sources

29



## CATH: The CATH Hierarchy

---

- CATH is another protein structure classification hierarchy:
  - **C**lass
  - **A**rchitecture
  - **T**opology
  - **H**omologous superfamily
  - **S**equence families
- <http://www.cathdb.info/>
- As with SCOP classification is for protein domains.

Data Sources

30



## CATH: Statistics

- Currently, v3.4:

| Class        | Architecture | Topology    | Homologous Superfamily | S35 Family   | S60 Family   | S95 Family   | S100 Family  | Domains       |
|--------------|--------------|-------------|------------------------|--------------|--------------|--------------|--------------|---------------|
| 1            | 5            | 376         | 839                    | 2763         | 3571         | 4679         | 9217         | 32396         |
| 2            | 20           | 228         | 514                    | 2514         | 3573         | 5668         | 9824         | 39140         |
| 3            | 14           | 577         | 1082                   | 5849         | 8381         | 10626        | 21900        | 79038         |
| 4            | 1            | 101         | 114                    | 204          | 253          | 352          | 547          | 2346          |
| <b>Total</b> | <b>40</b>    | <b>1282</b> | <b>2549</b>            | <b>11330</b> | <b>15778</b> | <b>21325</b> | <b>41488</b> | <b>152920</b> |

24,232 newly assigned domains since v3.3.

- Last version's stats:

| Class        | Architecture | Topology    | Homologous Superfamily | S35 Family   | S60 Family   | S95 Family   | S100 Family  | Domains       |
|--------------|--------------|-------------|------------------------|--------------|--------------|--------------|--------------|---------------|
| 1            | 5            | 360         | 773                    | 2400         | 3092         | 4040         | 7838         | 26880         |
| 2            | 14           | 558         | 1031                   | 5151         | 7294         | 9192         | 18679        | 66456         |
| 3            | 20           | 217         | 473                    | 2283         | 3225         | 5002         | 8589         | 33469         |
| 4            | 1            | 98          | 109                    | 185          | 233          | 323          | 477          | 1883          |
| <b>Total</b> | <b>40</b>    | <b>1233</b> | <b>2386</b>            | <b>10019</b> | <b>13844</b> | <b>18557</b> | <b>35583</b> | <b>128688</b> |

Data Sources

31



## CATH Levels: Class

- There are three major classes plus an extra minor class:
  - Mainly  $\alpha$
  - Mainly  $\beta$
  - Mixed  $\alpha$  &  $\beta$
  - Low secondary structure
- Most classification (90%) can be done automatically.
  - When this is in doubt, classification is done by visual inspection.

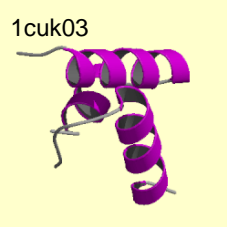
Data Sources

32

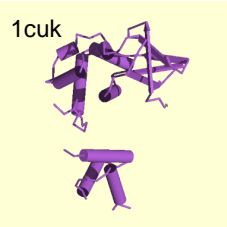


# Class Examples

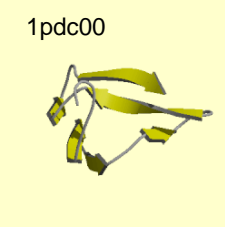
## Mainly Alpha



The full protein has two chains:

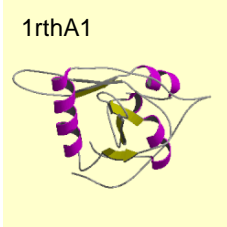


## Mainly Beta

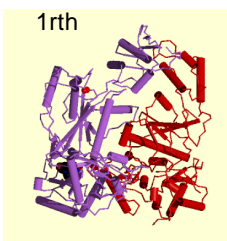


1pdc has only one domain

## Mixed Alpha-Beta



Several domains in two chains of the full protein:



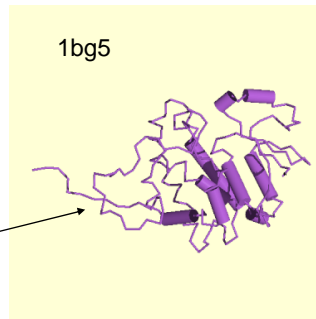
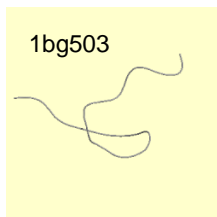
Data Sources

33



# Class Examples (cont.)

Few secondary structures:



The full protein has 9 helices. The domain 1bg503 is here:

Data Sources

34



## CATH Levels: Architecture

---

- The architecture level defines the overall shape of the domain structure as defined by the orientations of the secondary structure.
  - E.g.: barrel, sandwich, etc.
  - Connectivity between secondary structures is ignored.
  - Classification at this level is done by inspection appealing to the literature for descriptions of various architectures.

Data Sources

35



## CATH Levels: Topology

---

- These are essentially fold families.
  - Classification depends on the overall shape and connectivity of the secondary structures.
    - Done by using the structure comparison algorithm (SSAP) of Taylor and Orengo (1989).
    - SSAP assigns a score to a pair-wise comparison.
      - A score of 70 and a matching that indicates that at least 60% of the larger protein is contained in the smaller protein means that the two proteins are in the same fold family.
      - The score threshold is raised for some families if they are highly populated.

Data Sources

36



## CATH Levels: Homologous superfamily

---

- Homologous ==> a common ancestor
  - Similarities are determined by sequence comparison and then by structure comparison.
    - using SSAP
  - Two structures are in the same homologous superfamily if any of the following hold:
    - sequence identity  $\geq 35\%$
    - SSAP score  $\geq 80$  and sequence identity  $\geq 20\%$
    - SSAP score  $\geq 80$ , 60% of larger structure is equivalent to the smaller structure and the domains have related functions.

Data Sources

37



## CATH Levels: Sequence families

---

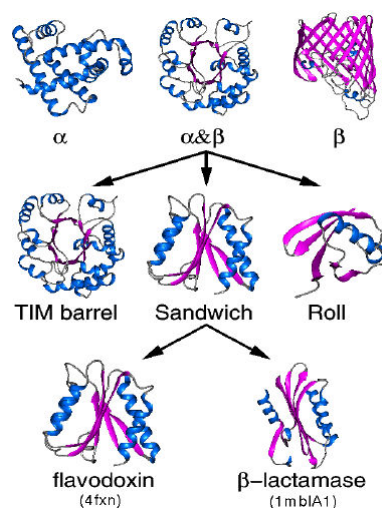
- Structures within an H level are clustered into smaller groups based on sequence identity.
  - Domains clustered in the same S level have sequence identities greater than 35% with at least 60% of the larger domain equivalent to the smaller domain.
    - indicates highly similar structure and function.

Data Sources

38

## A Portion of the Hierarchy

- Class:
- Architecture:
- Topology:



Data Sources

39

## PubChem:

<http://pubchem.ncbi.nlm.nih.gov/>

- PubChem is a database holding structural information for millions of molecules.
  - Maintained by the NCBI
    - National Center for Biotechnology Information
    - within the National Library of Medicine, NIH (National Institute of Health).
  - Data is free and can be downloaded via FTP.
    - The American Chemical Society (ACS) claims the database competes with its Chemical Abstracts Service (CAS).
      - A publishing operation that had \$400m/year revenues.
    - The ACS tried to get the U.S. Congress to restrict the operation of PubChem.
      - This was resolved in the summer of 2005.
      - Congress did not censure the NIH.

Data Sources

40



- Currently PubChem has over 30 million compounds. (About 25 million last year).
- Different searches:
  - PubChem Compound
    - Search for compounds using names, synonyms or keywords.
  - PubChem Substance
    - Search for chemical substances
  - PubChem BioAssay
    - Search on bioassay results
  - Structure Search
    - Search using chemical structure (eg. SMILES) as a query.

Data Sources

41



## PubChem Sample Search (1)

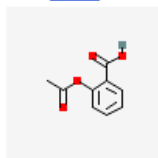
- Search for aspirin:

**PubChem Text Search**

PubChem Compound

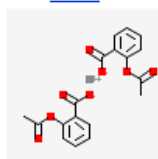
A partial list of results:

1: CID: [2244](#)



aspirin, Enterosarin ...  
IUPAC: 2-acetyloxybenzoic acid  
MW: 180.157 | MF: C9H8O4

2: CID: [6247](#)



Calcascorbin, Calscorbate ...  
IUPAC: calcium 2-acetyloxybenzoate  
MW: 398.377 | MF: C18H14CaO8

Data Sources

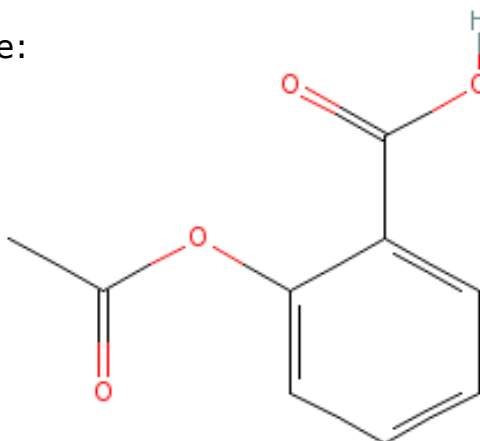
42



## PubChem Sample Search (2)

Aspirin 2D structure:

CID: 2244  
↑  
PubChem  
identifier



Data Sources

43



## PubChem Sample Search (3)

- As well as 2D structure several links are made available:

- [CID: 2244](#)
- [BioActivity: Summary](#)
  - Inactive: [90 Links](#)
  - Inconclusive: [3 Links](#)
- [Protein Structures: 3 Links](#)
- [Protein Sequences: 53 Links](#)
- [NLM Toxicology: Link](#)
- [Substances:](#)
  - All: [232 Links](#)
  - Same: [47 Links](#)
  - Mixture: [185 Links](#)
- [Related Compounds:](#)
  - Same, Connectivity: [2 Links](#)
- [Similar Compounds: 498 Links](#)
- [Structure Search](#)

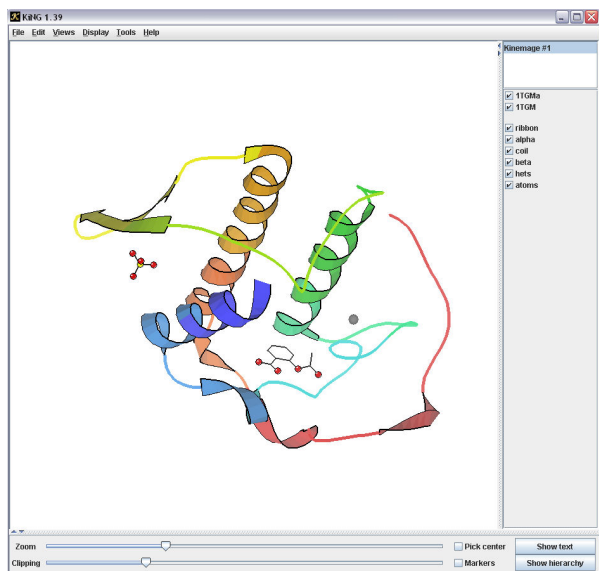
Data Sources

44



## PubChem Sample Search (4)

- Following the "Protein Structures" link eventually leads to proteins that bind with aspirin, in this case 1TGM:



45



## PubChem Sample Search (5)

- Description for **Aspirin**:  
The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)
- **Pharmacological Action:**  
[Anti-Inflammatory Agents, Non-Steroidal](#)  
[Fibrinolytic Agents](#)  
[Platelet Aggregation Inhibitors](#)  
[Cyclooxygenase Inhibitors](#)

Data Sources

46



## PubChem Sample Search (6)

---

- Research references for **Aspirin**:

### **PubMed via MeSH** Choose by Subheadings:

[administration and dosage](#)  
[analysis](#)  
[cerebrospinal fluid](#)  
[classification](#)  
[economics](#)  
[isolation and purification](#)  
[pharmacology](#)  
[radiation effects](#)  
[therapeutic use](#)  
[urine](#)

[adverse effects](#)  
[antagonists and inhibitors](#)  
[chemical synthesis](#)  
[contraindications](#)  
[history](#)  
[metabolism](#)  
[physiology](#)  
[standards](#)  
[therapy](#)

[analogs and derivatives](#)  
[blood](#)  
[chemistry](#)  
[diagnostic use](#)  
[immunology](#)  
[pharmacokinetics](#)  
[poisoning](#)  
[supply and distribution](#)  
[toxicity](#)

Data Sources

47



## PubChem Sample Search (7)

---

- **Properties Computed from Structure:**

- **Molecular Weight:** 180.157g/mol  
**Molecular Formula:** C<sub>9</sub>H<sub>8</sub>O<sub>4</sub>

**XLogP:** 1.4

**Hydrogen Bond Donor Count:** 1

**Hydrogen Bond Acceptor Count:** 4

**Rotatable Bond Count:** 3

**Topological Polar Surface Area:** 63.6

Data Sources

48



## PubChem Sample Search (8)

International Union of Pure and Applied Chemistry

### ○ Descriptors Computed from Structure:

- **IUPAC Name:** 2-acetyloxybenzoic acid
- **Canonical SMILES:** CC(=O)OC1=CC=CC=C1C(=O)O
- **InChI:** [InChI=1/C9H8O4/c1-6\(10\)13-8-5-3-2-4-7\(8\)9\(11\)12/h2-5H,1H3,\(H,11,12\)/f/h11H](#)

### ○ Substance Category:

- **Biological Properties:** [28 Links](#)
- **Journal Publishers:** [1 Link](#)
- **Metabolic Pathways:** [2 Links](#)
- **Molecular Libraries Screening Center Network:** [4 Links](#)
- **Physical Properties:** [4 Links](#)
- **Protein 3D Structures:** [4 Links](#)
- **Substance Vendors:** [1 Link](#)
- **Theoretical Properties:** [1 Link](#)
- **Toxicology:** [3 Links](#)

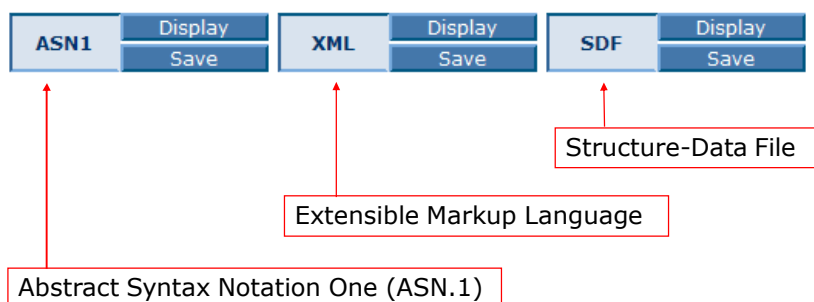
Data Sources

49



## PubChem Sample Search (9)

### ○ Links to structure files:



Data Sources

50

## Specialized Databases: HIVSDB (1)

### HIV Structural Database & Chem-BLAST

<http://xpdb.nist.gov/hivsdb/hivsdb.html>

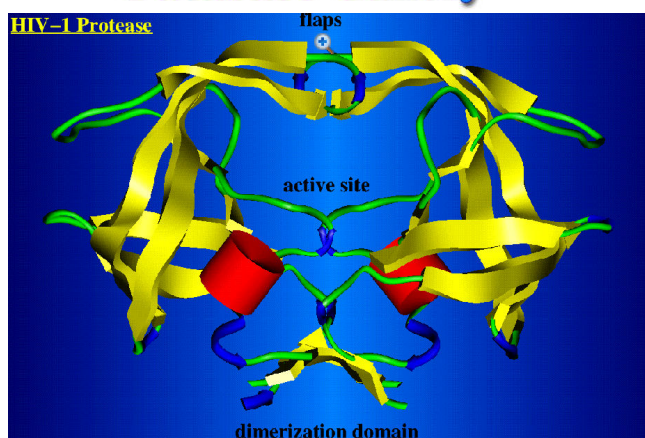
- NIST (National Institute of Standards and technology) provides this database as a central structural information resource for AIDS related molecules.
  - Proteins and ligands
- Related publications:
  - Prasanna, M.D., Vondrasek, J., Wlodawer, A., Bhat, T. N. Application of InChI to curate, index and query 3-D structures. *PROTEINS: Structure, Function, and Bioinformatics* **60**, 1-4 (2005).
  - 2.Prasanna M.D, Vondrasek J, Wlodawer A, Rodriguez H, Bhat T.N. Chemical compound navigator: a web-based Chem-BLAST, chemical taxonomy-based search engine for browsing compounds. *Proteins* **63**(4), 907-917(2006).

Data Sources

51

## Specialized Databases: HIVSDB (2)

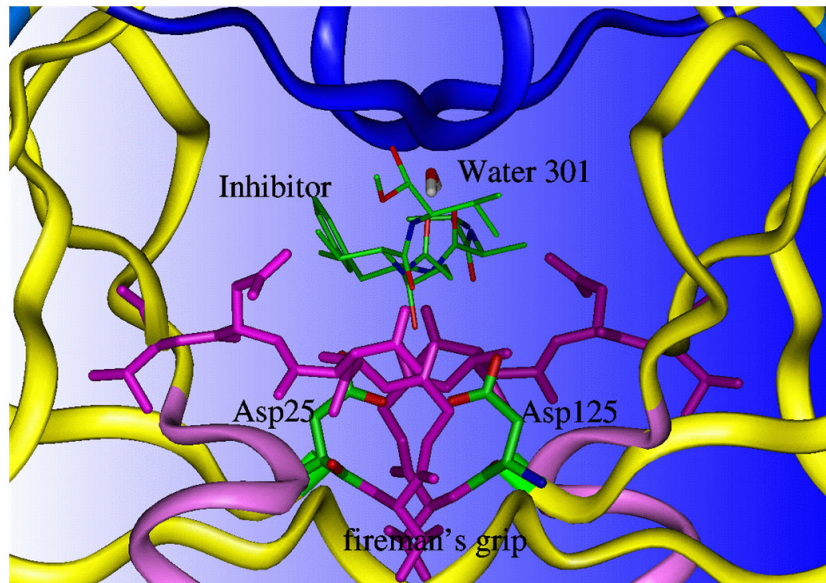
### HIV Protease/RT Database Gallery



Data Sources

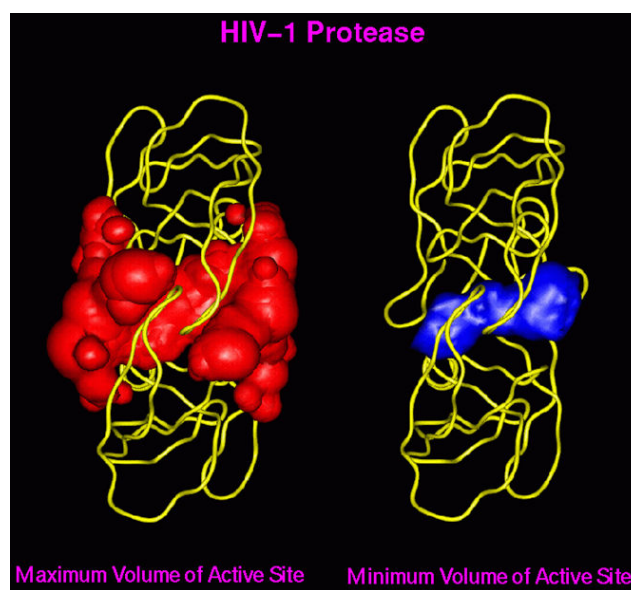
52

### Specialized Databases: HIVSDB (3)



53

### Specialized Databases: HIVSDB (4)



54



## Specialized Databases: HIVSDB (5)

