# Semantic Web for Chemical Images

T. N. Bhat[1], John Barkley[2], Anh Dao Nguyen[1]

[1]Biochemical Science Division,
[2]Software Diagnostics Division, NIST

http://bioinfo.nist.gov/SemanticWeb_hiv3d/chemblast.do
http://xpdb.nist.gov/hivsdb/advanced_query_files/slide0002.htm

bhat@nist.gov

Downloads  Query 3-D data  Query 2-D data

Prasanna MD, Vondrasek J, Wlodawer A, Rodriguez H, Bhat TN. Chemical compound navigator: a web-based chem-BLAST, chemical taxonomy-based search engine for browsing compounds. Proteins 2006;63(4):907-917.

PDB

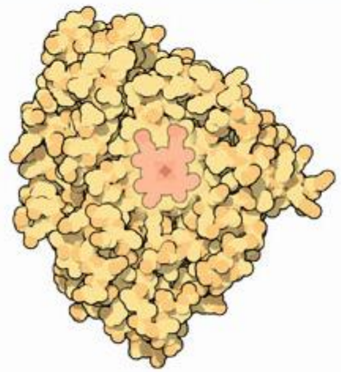
a more consistent and comprehensive archive.

Information about compatible browsers can be found **here**.

A **narrated tutorial** illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

**Comments? info@rcsb.org**

## Molecule of the Month: Cytochrome p450



If you have a headache and take a drug to block the pain, you'll notice that the effects of the drug wear off in a few hours. This happens because you have a powerful detoxification system that finds unusual chemicals, like drugs, and flushes them out of your body. This system fights all sorts of unpleasant chemicals that we eat and breathe, including drugs, poisonous compounds in plants, carcinogens formed during cooking, and environmental pollutants. The cytochrome p450 enzymes are our first line of defense in this chemical battle.

- **More ...**
- **Previous Features**

includes examples and definitions provided in the **PDB Exchange Dictionary** as guides for users depositing their structures.

- Full Story ...

17-October-2006

**RCSB PDB Focus: Exploring Domains in Protein Structure**

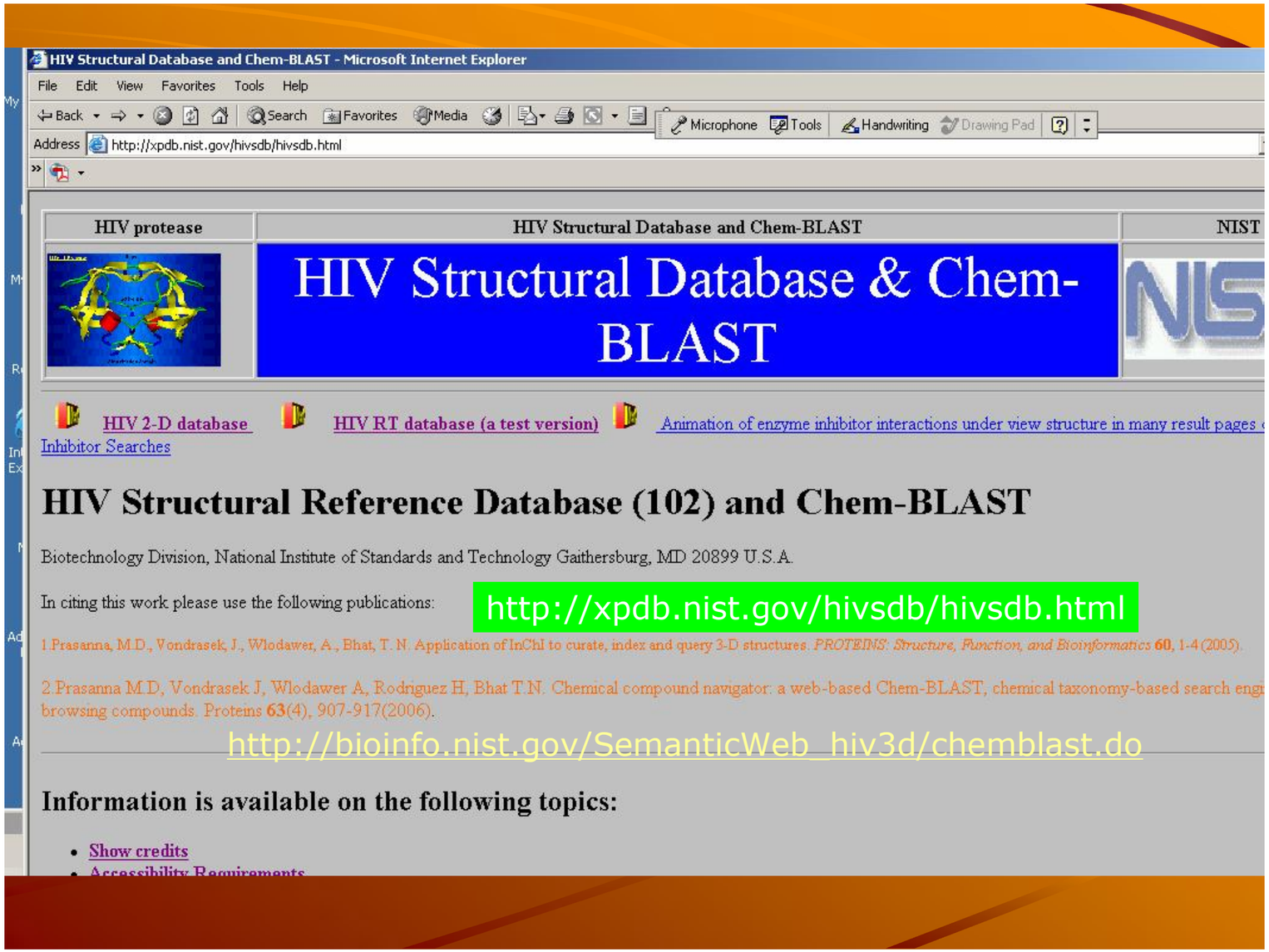
Domains can be thought of as the smallest structural units from which proteins are assembled that retain properties of the whole protein, such as a hydrophobic core.

The RCSB PDB offers various ways of exploring domains in protein structures.

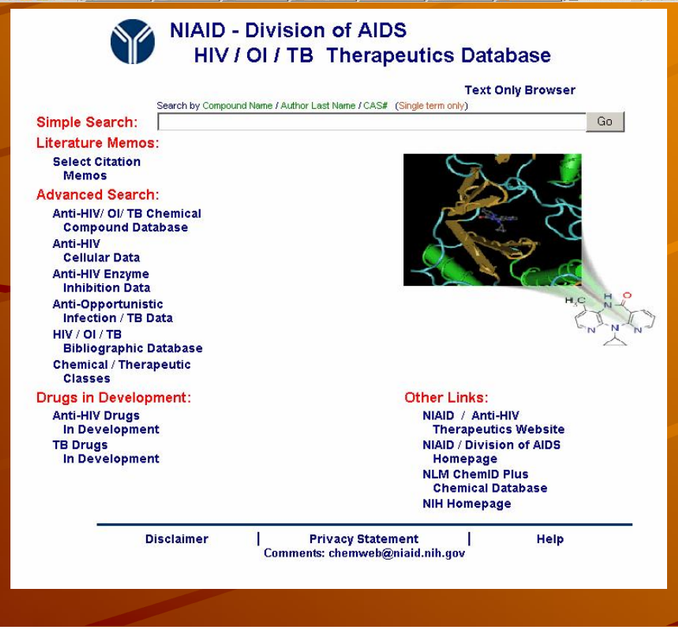
- Full Story ...


The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)


In citing the PDB please refer to: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).


HIV Structural Database and Chem-BLAST - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Microphone Tools Handwriting Drawing Pad
Address http://xpdb.nist.gov/hivsdb/hivsdb.html
HIV protease
HIV Structural Database and Chem-BLAST
NIST
HIV Structural Database & Chem-BLAST
HIV 2-D database
HIV RT database (a test version)
Animation of enzyme inhibitor interactions under view structure in many result pages
Inhibitor Searches
HIV Structural Reference Database (102) and Chem-BLAST
Biotechnology Division, National Institute of Standards and Technology Gaithersburg, MD 20899 U.S.A.
In citing this work please use the following publications:
http://xpdb.nist.gov/hivsdb/hivsdb.html
1.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 engi
browsing compounds. Proteins 63(4), 907-917(2006).
http://bioinfo.nist.gov/SemanticWeb_hiv3d/chemblast.do
Information is available on the following topics:
Show credits
Accessibility Requirements


The PubChem Project - Microsoft Internet Explorer
http://pubchem.ncbi.nlm.nih.gov/
NCBI
PubChem
National Library of Medicine
PubMed
Entrez
Structure
GenBank
PubChem
Help
PubChem Text Search
PubChem Compound
GO!
PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative. If you would like to learn more about how to use the PubChem resources, please go to our help page.
PubChem contains more than 400 BioAssays.
The PubChem BioAssay services now provide target links and bioassay neighboring information identified by activity overlap and target sequence similarity.
More PubChem announcements ...
PubChem Compound: Search unique chemical structures using names, synonyms or keywords. Links to available biological property information are provided for each compound.
PubChem Substance: Search deposited chemical substance records using names, synonyms or keywords. Links to biological property information and depositor web sites are provided.
PubChem BioAssay: Search bioassay records using terms from the bioassay description, for example "cancer cell line". Links to active compounds and bioassay results are provided.
Structure Search: Search PubChem's Compound database using a chemical structure as the query. Structures may be sketched or specified by SMILES, MOL files, or other formats.
PubChem is organized as three linked databases within the NCBI's Entrez information retrieval system. These are PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem also provides a fast chemical structure similarity search tool. More
1HTE
GLN
1HTE
LEU:GLN:GLU:2-amino-3-hydroxypropanoic_acid


NIAID - Division of AIDS
HIV / OI / TB Therapeutics Database
Text Only Browser
Search by Compound Name / Author Last Name / CAS# (Single term only)
Simple Search:
Go
Literature Memos:
Select Citation Memos
Advanced Search:
Anti-HIV/ OI/ TB Chemical Compound Database
Anti-HIV Cellular Data
Anti-HIV Enzyme Inhibition Data
Anti-Opportunistic Infection / TB Data
HIV / OI / TB Bibliographic Database
Chemical / Therapeutic Classes
Drugs in Development:
Anti-HIV Drugs In Development
TB Drugs In Development
Other Links:
NIAID / Anti-HIV Therapeutics Website
NIAID / Division of AIDS Homepage
NLM ChemID Plus Chemical Database
NIH Homepage
Disclaimer | Privacy Statement | Help
Comments: chemweb@niaid.nih.gov

# Why Select Data From These Databases?

- These databases have lots of chemical compounds that are in common
  - Databases have different types of data for a given chemical compound
    - Therefore users would like to collate information from all of these databases
  - Without a Semantic Web, queries for the databases are performed independently
  - Without a Semantic Web, results from different databases are manually collated
- Databases also have several compounds unique for each of them
  - Many of these unique compounds are structurally related to commonly held compounds
    - A drug design effort involves the study of structure and property relationships among compounds
  - Without a Semantic Web, a user manually collates data for related structures
    - Results in missed hits
    - Results in over-whelming hits
    - Time consuming
- Semantic Web promises to reduce the effect of some of these problems

AIDS research is still a work in progress and the compounds we have chosen are AIDS inhibitors

# Semantic Web

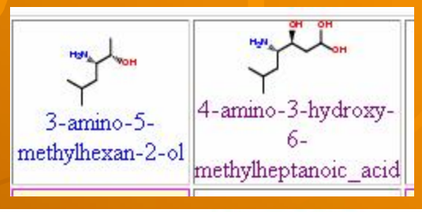
- There is considerable interest in Semantic Web as a vision for the future Web and for interoperability standards
- Standard names and ontology are critical components for Semantic Web
  - Without standard names, exchange of data is hard
  - Without ontology, exchange of knowledge is hard
- There are four important issues in developing a Semantic Web
  - Data identification
    - Uniform resource identifier –URI – a string of characters
  - Data presentation
    - Ontology – defines a context-dependent meaning for the data represented by a URI
  - User query and interface
    - Most of the current techniques focus on text-based query
    - Images pose special problems
  - Information exchange between databases
    - Common URI and/or ontology
- Ontology is easy to create but hard to sell
  - Most of us, use an ontology of some sort– but other may use a different ontology!
  - Software issues, annotation standards, focus of the use case of the ontology are some factors that work against an universal ontology
- As part of the efforts by HCLS of W3C we have developed a use case for Semantic Web for chemical images
  - Illustrate the use and implementation of URI and ontology for chemical images
  - Illustrate how the Semantic Web concepts may be implemented and used to exchange data between independently maintained databases without a common ontology

# Part 1: URI for chemical Images

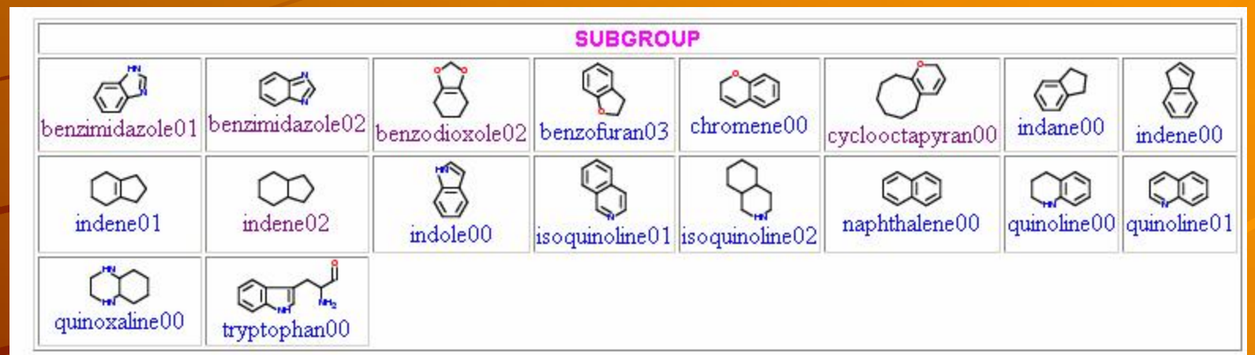
- For decades scientists have been developing names to identify structures and their images
  - Simple names
    - His
    - Ala
    - DNA
    - ATP
  - Not so simple names
  - 2-amino-3-methylpentanamide
  - 4-amino-3-hydroxy-6-methylheptanoic_acid
  - 1-[(Benzenesulfonyl-methyl-amino)-phenyl-butyl]-piperidin-4-yl}-propyl-carbamic acid, naphthalen-1-ylmethyl ester
- Names have limitations
  - They are not easily scalable – we may not know how to assign names for a new structures that we may come across
  - Not suitable for machine reasoning – without a lookup table machines may not create an image from its identifier or vice versa
- IUPAC International Chemical Identifier (InChI)
  - Uses chemical connectivity to assign identifier
    - **_1_2FC10H11NO2_2Fc11-10_2812_2913-9-5-7-3-1-2-4-8_287_296-9_2Fh1-4_2C9H_2C5-6H2_2C_28H2_2C11_2C12_29**
  - But not amenable to human interpretation

# A image is worth thousand words!

Why not we use images as URI?




Similarity between structures are explicit in images



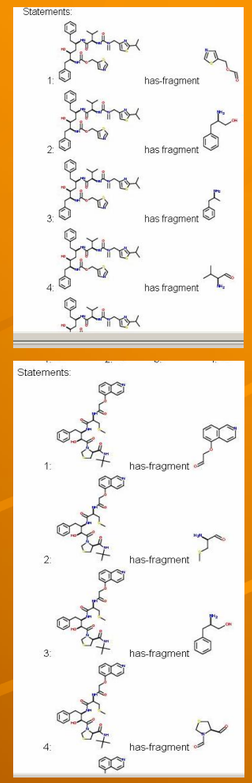

Basically the use of images as URI instantly transforms a name into knowledge!
(information on similarity, and substructures,…)

W3C defines URI as a set of characters.
Therefore we combine name and images to define URI

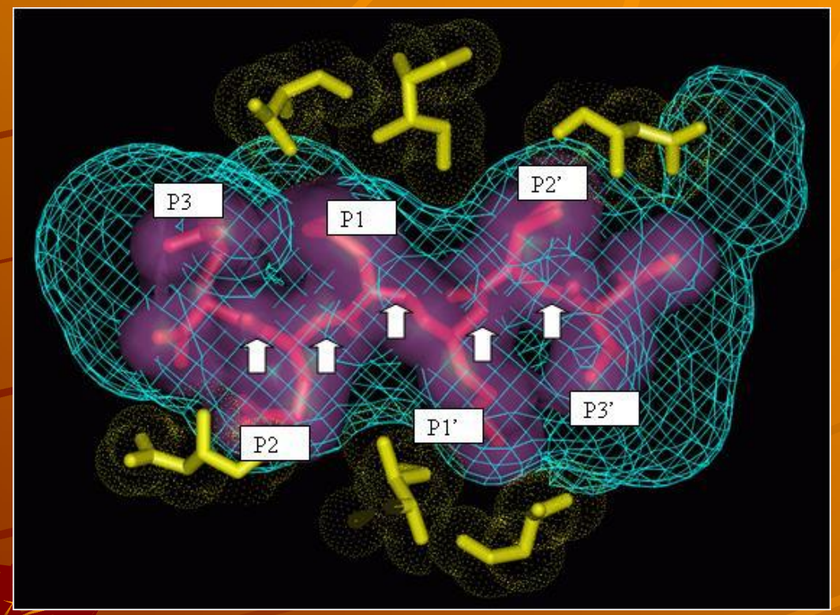
Proposal is to use images as URI

# Part II: Development of Ontology

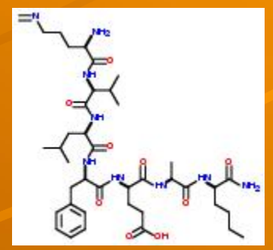
- Ontology defines the meaning of an image within the frame work of a given application
  - Ontology defines the meaning of a structure represented by an image for drug-discovery for AIDS



# Defining Ontology – a meaning that is relevant to the context of an image







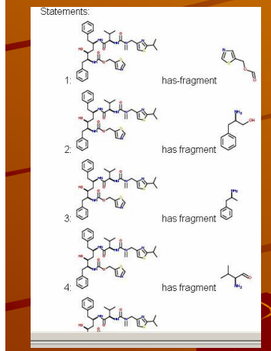
http://esw.w3.org/topic/HCLS/ChemicalTaxonomiesUseCase

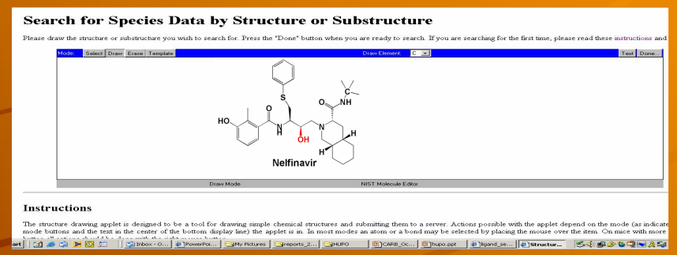
Each sub-site such as P1, P2 is a building block for drug design.
These sub-sites are used as the elements of RDF.
The fragments that bind to these sub-sites are truly Semantic and mean ‘some thing for drug design’

# Part III: User Interface & Query

- So far we talked about creating data and URI
- Third part of my talk is on presenting the URI and the knowledge to users using images
  - Construction of user interface and query by users

# Query an Image by Interactively Creating an Image



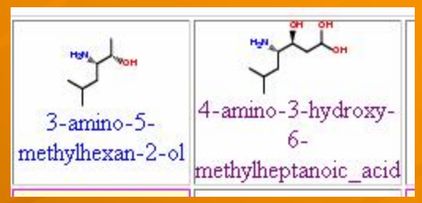




Thanks to Bill Gates.
We love to ‘click’.

- Chemical databases often provide an interface to create an image to query an image!
- These methods are hard to use, often lead to missed or overwhelming hits
  - One need to know what is there in the database to query effectively using such method
  - If one queries simple group such as on a phenyl group every compound will be hit
    - In HIVSDB every structure has a phenyl group
  - If one queries a piperazine you may not get any hit

# Text-based search on an Image?

URI on images comes with a price – how to query them?

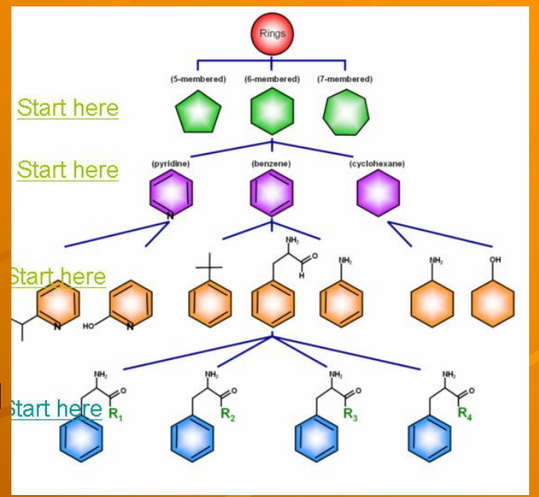



- People use vocabulary based identifiers (Phe, Val, Cas Numbers ..) because
  - Text based query on names can be made
- Text based query may not be made on images because
  - an image may not have a meaningful name
  - Names like Mike_April_1_2007 serve to glorify Mike, but they are not fun for others!
- We propose to use ontology and Semantic Web concepts to query on images of structures
  - The concept completely removes the need to use any name at any time by a user of the Web resource

If we chose to query images using names – then the novelty of using images is compromised by requiring a user to know its name to query an image

# Displaying Image Ontology

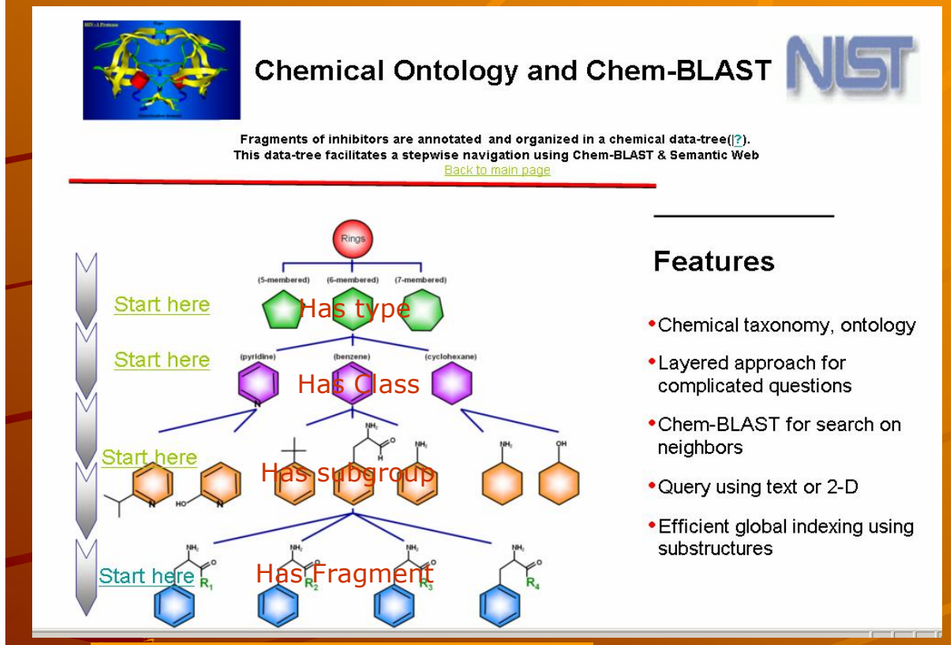
- Use of images and ontology sounds good
- But the question is how to use images and ontology?
  - Compared to vocabularies, images are huge and thus displaying images may require too many Web pages
    - A user may not have patience to go through all the images to find what he wants
  - Presenting the images in an orderly and predictable fashion is crucial to avoid missed and over-whelming hits
- We propose the use of ontological tree for presenting images over a Web



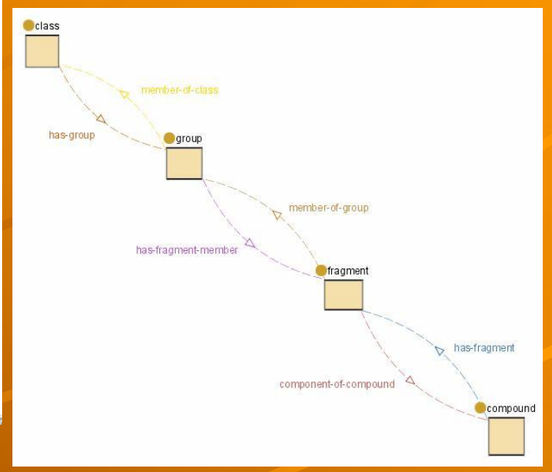

How to present ontological images?

Use ontological tree to present images

# Displaying Image Ontology



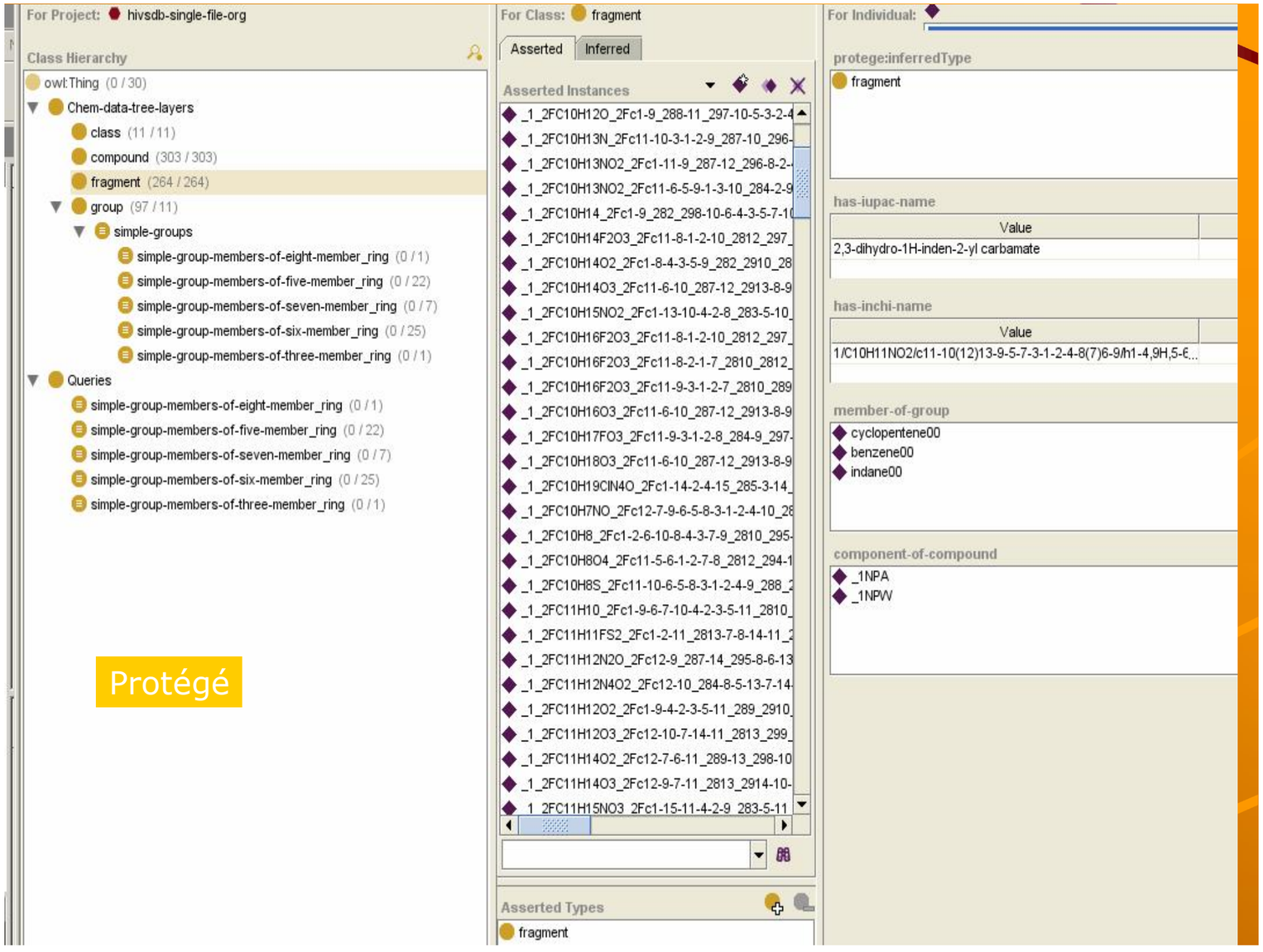


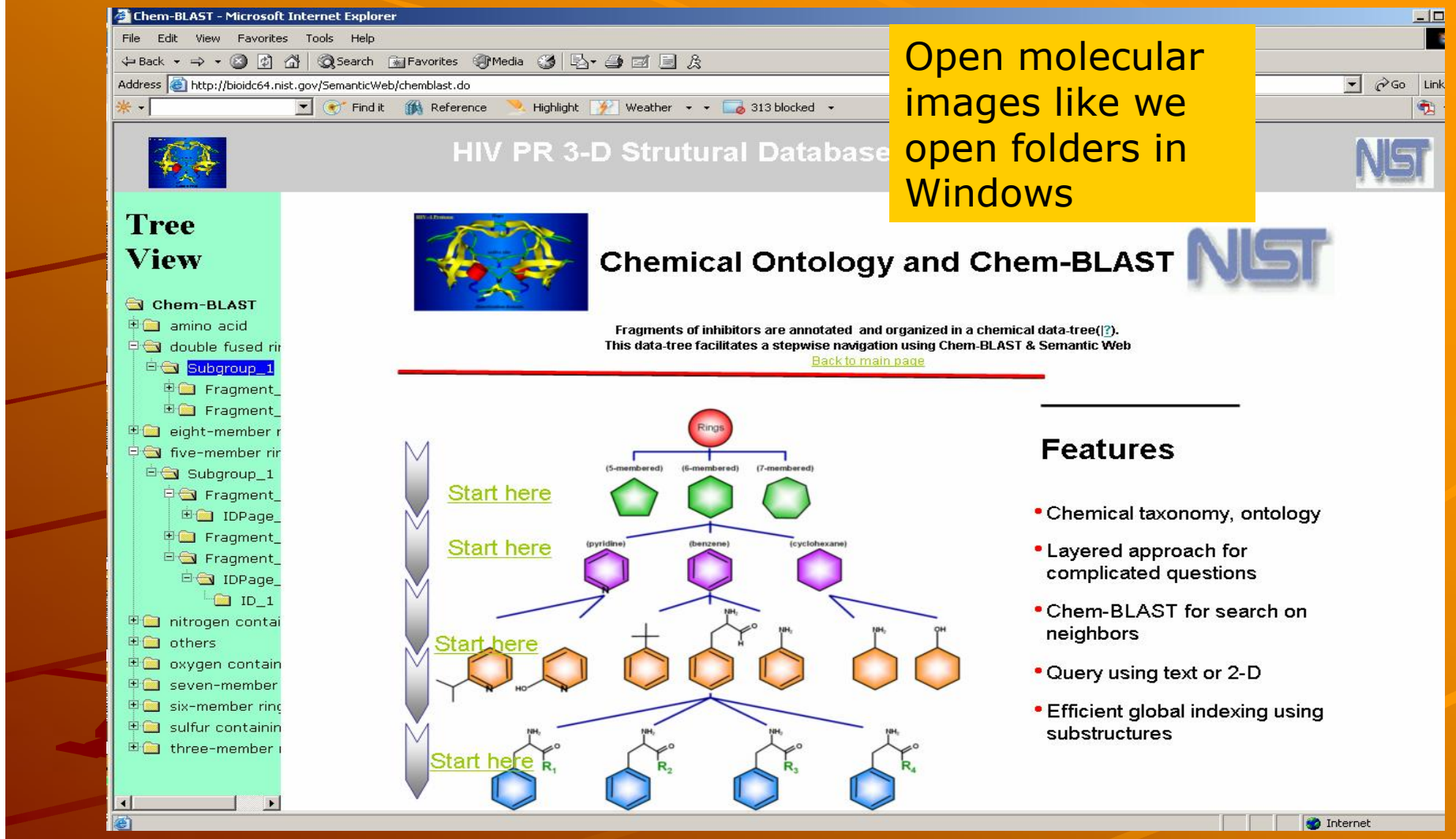



## What is ontological tree?

- Ontological is tree like a matrix that
  - Quantizes each data and its components
  - Defines relationships between each quantum of elements
  - Quantum elements are defined using the rules applicable for a particular use case of the Semantic Web

Information from RDF (resource description framework) and ontological tree can be interleaved

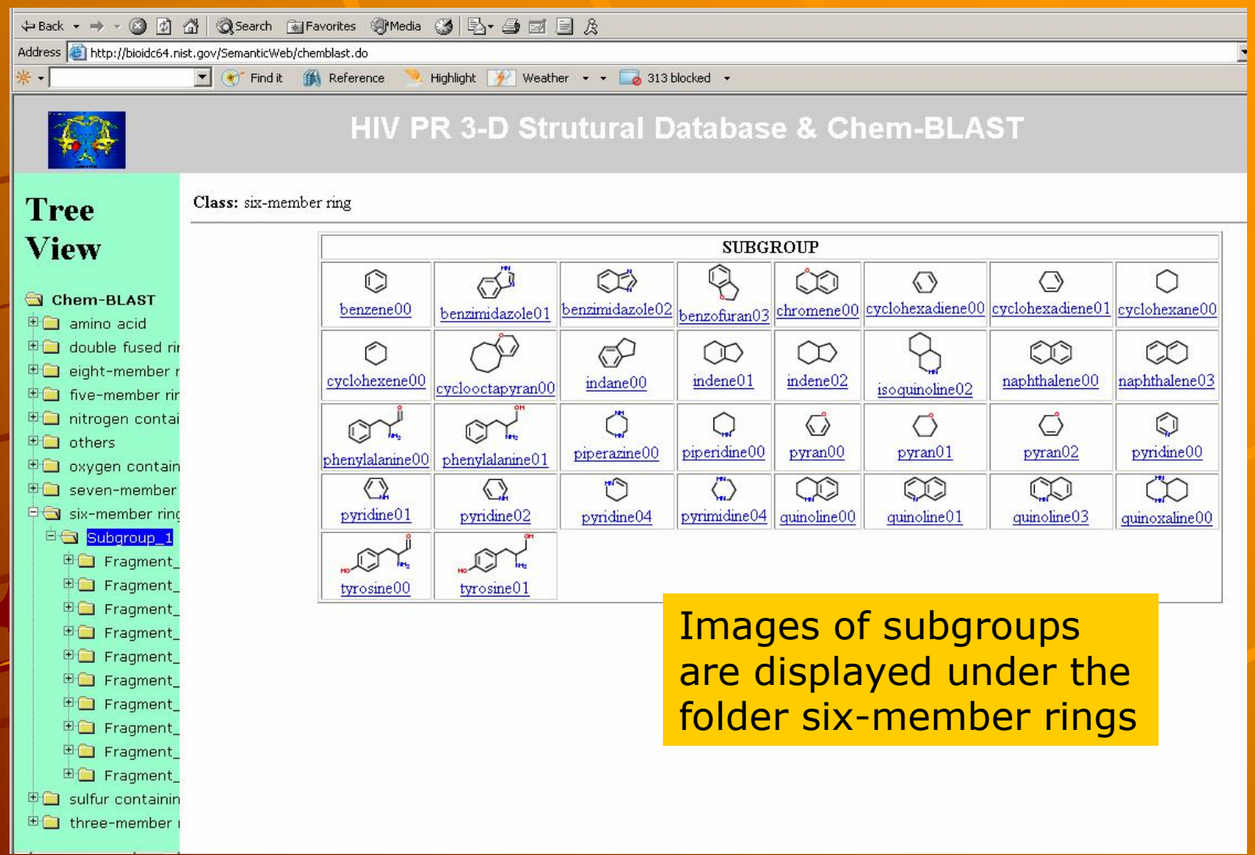

For Project: hivsdb-single-file-org
Class Hierarchy
owl:Thing (0 / 30)
Chem-data-tree-layers
class (11 / 11)
compound (303 / 303)
fragment (264 / 264)
group (97 / 11)
simple-groups
simple-group-members-of-eight-member_ring (0 / 1)
simple-group-members-of-five-member_ring (0 / 22)
simple-group-members-of-seven-member_ring (0 / 7)
simple-group-members-of-six-member_ring (0 / 25)
simple-group-members-of-three-member_ring (0 / 1)
Queries
simple-group-members-of-eight-member_ring (0 / 1)
simple-group-members-of-five-member_ring (0 / 22)
simple-group-members-of-seven-member_ring (0 / 7)
simple-group-members-of-six-member_ring (0 / 25)
simple-group-members-of-three-member_ring (0 / 1)
Protégé
For Class: fragment
Asserted
Inferred
Asserted Instances
_1_2FC10H12O_2Fc1-9_288-11_297-10-5-3-2-4
_1_2FC10H13N_2Fc11-10-3-1-2-9_287-10_296-
_1_2FC10H13NO2_2Fc1-11-9_287-12_296-8-2-
_1_2FC10H13NO2_2Fc11-6-5-9-1-3-10_284-2-9
_1_2FC10H14_2Fc1-9_282_298-10-6-4-3-5-7-1
_1_2FC10H14F2O3_2Fc11-8-1-2-10_2812_297_
_1_2FC10H14O2_2Fc1-8-4-3-5-9_282_2910_28
_1_2FC10H14O3_2Fc11-6-10_287-12_2913-8-9
_1_2FC10H15NO2_2Fc1-13-10-4-2-8_283-5-10_
_1_2FC10H16F2O3_2Fc11-8-1-2-10_2812_297_
_1_2FC10H16F2O3_2Fc11-8-2-1-7_2810_2812_
_1_2FC10H16F2O3_2Fc11-9-3-1-2-7_2810_289
_1_2FC10H16O3_2Fc11-6-10_287-12_2913-8-9
_1_2FC10H17FO3_2Fc11-9-3-1-2-8_284-9_297-
_1_2FC10H18O3_2Fc11-6-10_287-12_2913-8-9
_1_2FC10H19ClN4O_2Fc1-14-2-4-15_285-3-14_
_1_2FC10H7NO_2Fc12-7-9-6-5-8-3-1-2-4-10_28
_1_2FC10H8_2Fc1-2-6-10-8-4-3-7-9_2810_295-
_1_2FC10H8O4_2Fc11-5-6-1-2-7-8_2812_294-1
_1_2FC10H8S_2Fc11-10-6-5-8-3-1-2-4-9_288_2
_1_2FC11H10_2Fc1-9-6-7-10-4-2-3-5-11_2810_
_1_2FC11H11FS2_2Fc1-2-11_2813-7-8-14-11_2
_1_2FC11H12N2O_2Fc12-9_287-14_295-8-6-13
_1_2FC11H12N4O2_2Fc12-10_284-8-5-13-7-14
_1_2FC11H12O2_2Fc1-9-4-2-3-5-11_289_2910_
_1_2FC11H12O3_2Fc12-10-7-14-11_2813_299_
_1_2FC11H14O2_2Fc12-7-6-11_289-13_298-10
_1_2FC11H14O3_2Fc12-9-7-11_2813_2914-10-
_1_2FC11H15NO3_2Fc1-15-11-4-2-9_283-5-11
Asserted Types
fragment
For Individual:
protege:inferredType
fragment
has-iupac-name
Value
2,3-dihydro-1H-inden-2-yl carbamate
has-inchi-name
Value
1/C10H11NO2/c11-10(12)13-9-5-7-3-1-2-4-8(7)6-9/h1-4,9H,5-6...
member-of-group
cyclopentene00
benzene00
indane00
component-of-compound
_1NPA
_1NPW




Open molecular images like we open folders in Windows

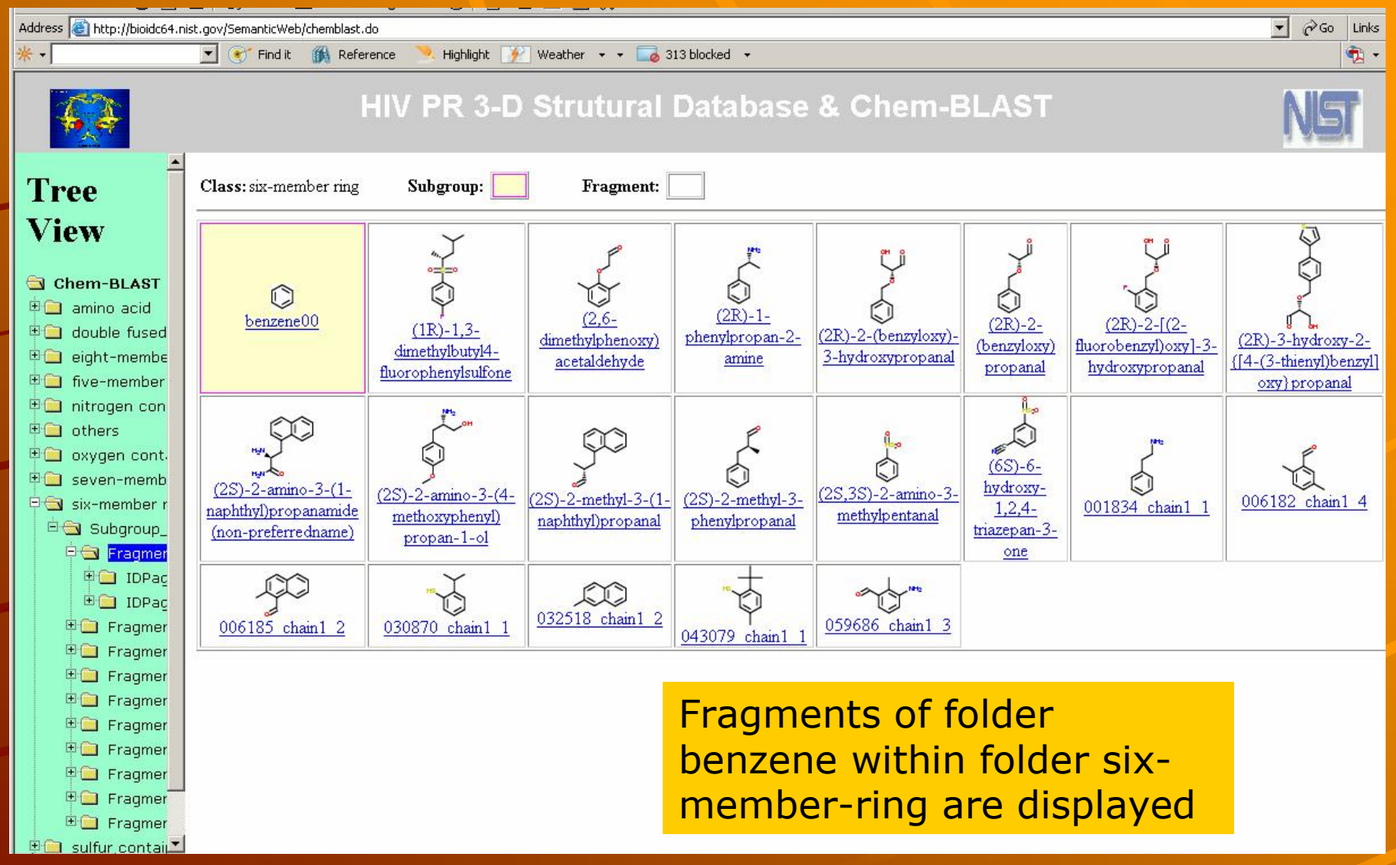
http://bioinfo.nist.gov/SemanticWeb_hiv3d/chemblast.do

http://bioinfo.nist.gov/SemanticWeb_hiv3d/chemblast.do



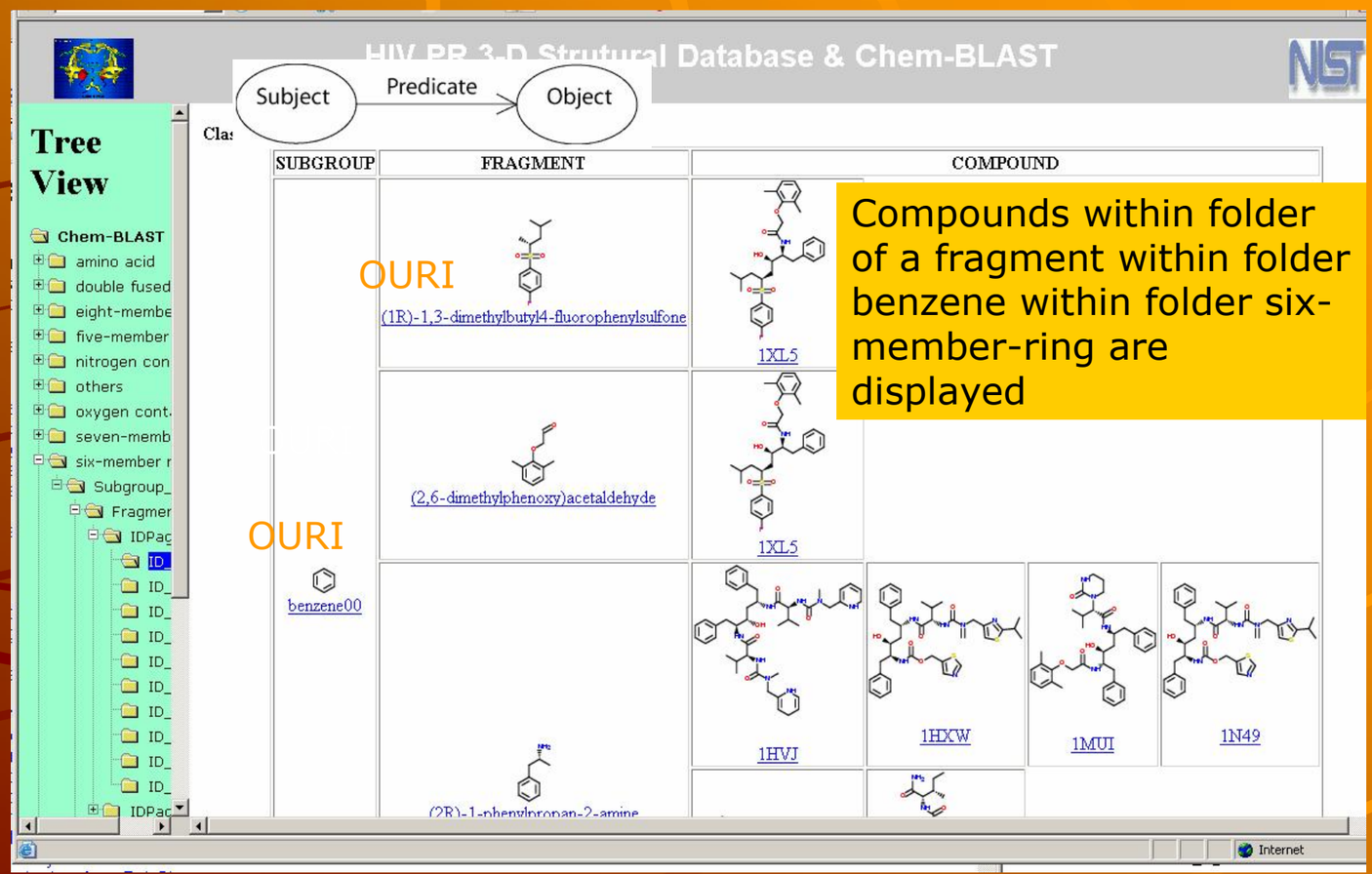

Images of subgroups are displayed under the folder six-member rings

http://bioinfo.nist.gov/SemanticWeb_hiv3d/chemblast.do




Fragments of folder benzene within folder six-member-ring are displayed

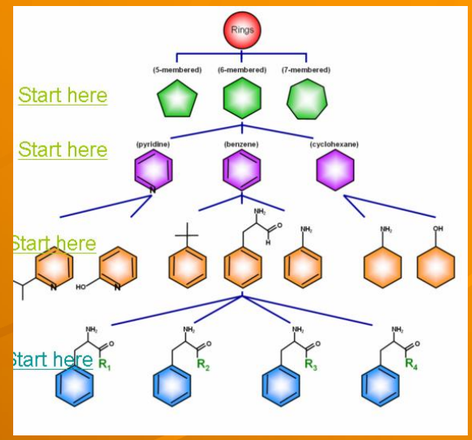
http://bioinfo.nist.gov/SemanticWeb_hiv3d/chemblast.do




Compounds within folder of a fragment within folder benzene within folder six-member-ring are displayed

# Part IV: Handshakes Between Web Browsers

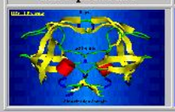
- Efficient exchange of information between Web browsers requires the use identical standards in URI by all participating Web browsers
- Often database providers use different standards in URI
  - Annotation standards, software needs, focus of the Web page are just a few reasons that work against the use of common ontology among independently maintained Web Browsers
- Therefore we sub-divide URIs into two types
  - Ontology independent, Ontology dependent (OURI)
    - OURI standards for ontological tree
    - Global URI standard for complete compounds
    - As more and more database providers use the same local standards in URI, a local URI can become a global URI




Once a query reaches the end of ontological tree control is passed on to other Web Browsers (such as the AIDSDB maintained by NIAID or PubChem) using URI

Find it | Reference | Highlight | Weather | 390 blocked

| HIV protease | HIVSDB HOMOLOGY MAPING FROM 2-D TO 3-D STRUCTURES | NIST |
|---|---|---|



# HIVSDB HOMOLOGY MAPING FROM 2-D TO 3-D STRUCTURES

This page shows the structural neighbors of a 2-D structure in the 3-D database. The listing of fragments that are common between 2-D and 3-D structures may not be complete. The purpose of this page is to facilitate a tentative mapping between 2-D and 3-D structures inside the active site of HIV 1 protease to better understand drug protein interactions and drug resistance when 3-D structure determined by X-Ray or NMR is not available. Result pages are presented in the descending order of the number of fragments common between the 2-D and 3-D data.

Submit | Reset

1HVC ID

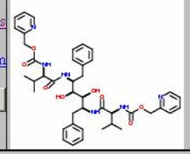
Current Page :1

Total Number of Pages: 63

Result pages :1 |2 |3 |4 |5 |6 |7 |8 |9 |10 |11 |12 |13 |14 |15 |16 |17 |18 |19 |20 |21 |22 |23 |24 |25 |26 |27 |28 |29 |30 |31 |32 |33 |34 |35 |36 |37 |38 |39 |40 |41 |42 |43 |44 |45 |46 |47 |48 |49 |50 |51

2-D structure of 003504

C42H52N6O8

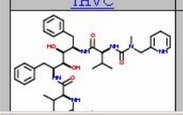


Cellular data, Anti-HIV enzyme data, References

PubChem

**Restart from** 1HVC

Interactive view of 3-D structural neighbor 1HVC



3-D fragments of 1HVC that are common to 2-D structure of 003504

Internet

Inbox - Outlook Ex... | papers-2007 | Microsoft PowerPoi... | PowerPoint Present... | ligand_searches ...

**Chemical Name:** 2,5-(S,S)-Bis(2-pyridylmethoxyvalyl)-1,6-diphenyl-3,4-(R,S)-dihydroxyhexane

Transfer Structure to MarvinSketch | Quick Structure Search

(2S, 3R, 4S, 5S)

C42 H52 N6 O8 **MW:** 768.91

**H-bond donors:** 6 **H-bond acceptors:** 14 **PHIA (Flexible Bonds):** 17.69

**Company:** ABBOTT

| | | | | |
|---|---|---|---|---|
| Anti-HIV Cellular data<br>Lines of Data: 4 | Anti-HIV Enzyme data<br>Lines of Data: 2 | Anti-OI data: 0 | **TB Min MIC** | **TB Min IC50** |
| HIV Cellular Lit. Refs<br>Number of References: 2 | HIV Enzyme Lit. Refs<br>Number of References: 2 | Number of References: 0 | | |

**Classes:** PEPTIDOMIMETICS, SYMMETRY-BASED DIHYDROXY; HIV PROTEASE INHIBITORS

**Synonyms**

- A-77212

**Calc. LogP** (MDL QSAR): 3.02

**Calc. LogP** (KowWin): 2.77

**Lipinski:** 1 (Score, out of 4)

**AIDS#** 003504

**Links to ChemID Plus by CAS#**

**Links to PubChem by AIDS#**

003504

**Links to PubMed by CAS#**

**Links to NIST by AIDS#**

003504

# Summary

- Semantic Web page for chemical images is described
  - Chemical images are organized into an ontological tree using the concepts used in drug-discovery
  - The images are indexed using ontological concepts that may add value for human use of the images
- An incremental approach for Semantic Web is described
  - A global URI and local ontological OURI for chemical compounds are presented
  - Concepts are implemented in a database (ORACLE & MySQL) environment
  - The implementation is Web friendly and it reduces missed hits without producing overwhelming result pages
  - In this approach a user need not wean through a large set of images
  - Web page lays out the images in an orderly fashion for pick and choose like searching files in a Windows environment
  - This approach makes the contents of the database transparent to the users using the pictures of the molecules that are indexed with the commonly used concepts
- Sample data from four major independently maintained databases (PDB, HIVSDB, AIDSDB, PubChem) are used to illustrate the concept
- http://xpdb.nist.gov/hivsdb/advanced_query_files/slide0002.htm
- Using the technology a user can go back and forth between the independently maintained HIVSD and AIDSDB
- http://bioinfo.nist.gov/SemanticWeb_hiv3d/chemblast.do