RDF friendly Chemical Taxonomies for Semantic Web (Using ORACLE/MySQL)

T.N.Bhat*, J. Barkley

NIST, Gaithersburg

USA

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.

Dictionaries & File Formats
Software Tools
Educational Resources
BioSync
General Information
Acknowledgements
Frequently Asked Questions
Known Problems
Report Bugs/Comments

Prior work

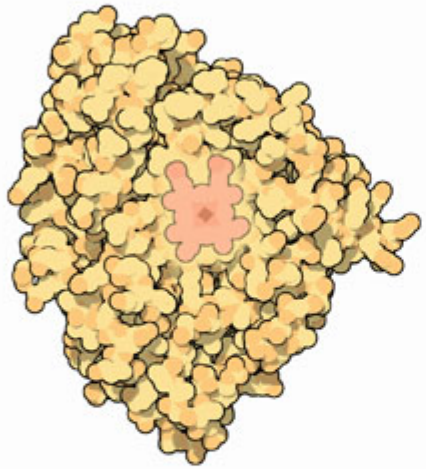
that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

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**

(at **RCSB** or **PDBj**) 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

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

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).

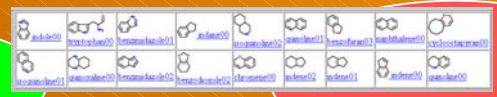
Information is available on the following topics:

- Show credits
- Accessibility Requirements

What is new in HIVSDB compared to what is out there?

- All the current AIDS drugs were developed based on structures of HIV proteins
  - There is no other database focused on structures of HIV proteins
- HIVSDB is a comprehensive resource of structures of AIDS proteins/complexes and of related info for drug design
  - 3-D data (X-ray or NMR)
  - Chemical structures (2-D, schematic drawing) of inhibitors obtained from literature
    - Synonyms, company names, IUPAC names
      - These are not ordinary structures that may or may not bind to a target enzyme
    - EC50s less than 1 micro Molar with a Therapeutic Index between 1 and 800
  - Biological, cellular, anti-viral data, references for further reading from published literature
    - Not info of the type ' he said, they said '. One can see the citation as well
  - Drug resistance data (over 10,000 citations have been weaned through this year alone)
    - Drug resistance mutant 3-D models
    - Semantic chemical fragments for (in silico) SAR studies
      - To design new drugs
      - To compare drugs
  - Illustration of enzyme drug interactions
    - To enable SAR modeling
    - Animation to illustrate mutant effects
- HIVSDB narrates (attempts) the (complete) story of drug design efforts for AIDS on which the world has spent billions of $ during the last two decades
- Database includes HIV protease, RT (incorporation of other proteins is planned)

Drug design by mix-match-method



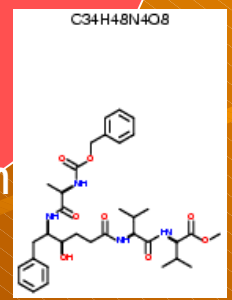
Cellular, toxicity,.. test

Get a list of all the fragments used in the past; add new fragments

Group the fragments based on sub-site properties

Wet lab

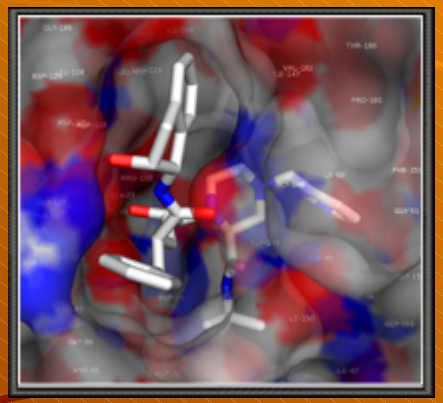
In silico



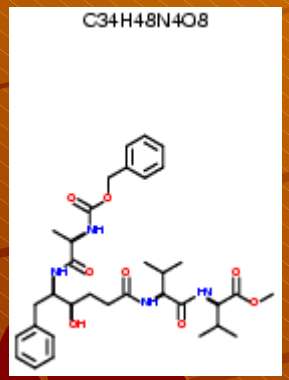
Enzyme inhibition test, Crystal structure

Mix-and-match the fragments to design a new drug

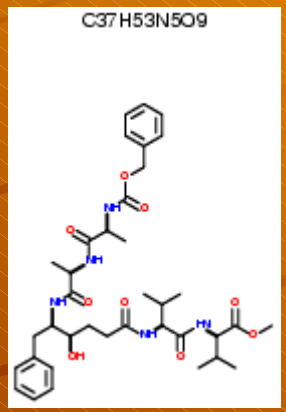
Data are chosen over a wide range of company, core type, and potency.



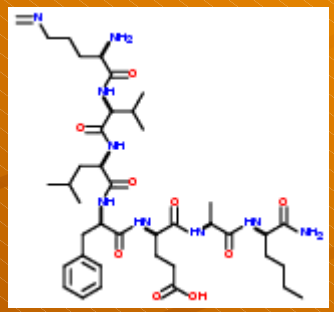
Inhibitors of HIV protease



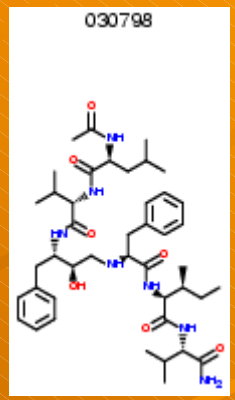
000503



000505

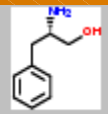


1A8K



030798



PDB Format

```
HETATM   1  C1  A7700 800     3.484   3.332  10.744  1.00 19.36
HETATM  10  C10 A7700 800     7.702   5.966  10.862  1.00 28.71
HETATM  11  C11 A7700 800     8.525   6.077   9.751  1.00 30.20
HETATM  54  C12 A7700 800     3.555  -8.734  19.771  1.00 26.67
HETATM  56  C13 A7700 800     2.317  -8.344  20.282  1.00 25.95
HETATM  57  C14 A7700 800     3.355  -6.296  20.983  1.00 31.02
HETATM  58  C15 A7700 800     2.211  -7.098  20.901  1.00 28.97
HETATM  51  C2  A7700 800     5.791  -8.392  16.839  1.00 22.59
HETATM  13  C22 A7700 800     2.786   1.298  11.960  1.00 17.23
HETATM  14  C23 A7700 800     3.932   0.378  12.380  1.00 17.53
HETATM  16  C25 A7700 800                                 18.10
```

>30 years

Mol format

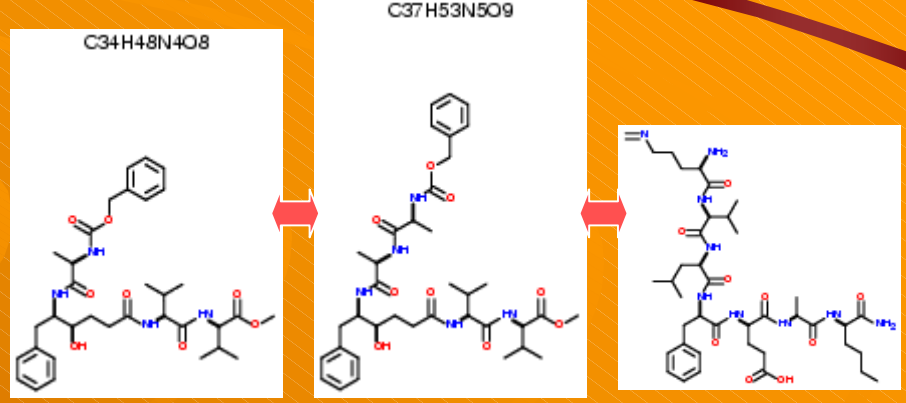
```
1HVC.mol
 -ISIS-          3D

116119  0  0  0  0  0  0  0  0  0
   3.4840   3.3320  10.7440 C  0  0  0  0  0
   7.7020   5.9660  10.8620 C  0  0  0  0  0
   8.5250   6.0770   9.7510 C  0  0  0  0  0
   3.5550  -8.7340  19.7710 C  0  0  0  0  0
   2.3170  -8.3440  20.2820 C  0  0  0  0  0
   3.3550  -6.2960  20.9830 C  0  0  0  0  0
   2.2110  -7.0980  20.9010 C  0  0  0  0
```

>15 years

CML

```
<molecule convention="MDLMol" id="ampa" title="AMPA">
<date day="22" month="11" year="1995" />
<atomArray>
<atom id="a1">
<string builtin="elementType">C</string>
<float builtin="x2">0.345</float>
<float builtin="y2">0.4329</float>
</atom>
<atom id="a2">
<string builtin="elementType">C</string>
<float builtin="x2">3.0607</float>
<float builtin="y2">0.4329</float>
</atom>
<float builtin="y2">0.9977</float>
</atom>
```

>5 years



PDB + PDB = ?
PDB – MOL = ?
CML + CML = ?
MOL – CML = ?

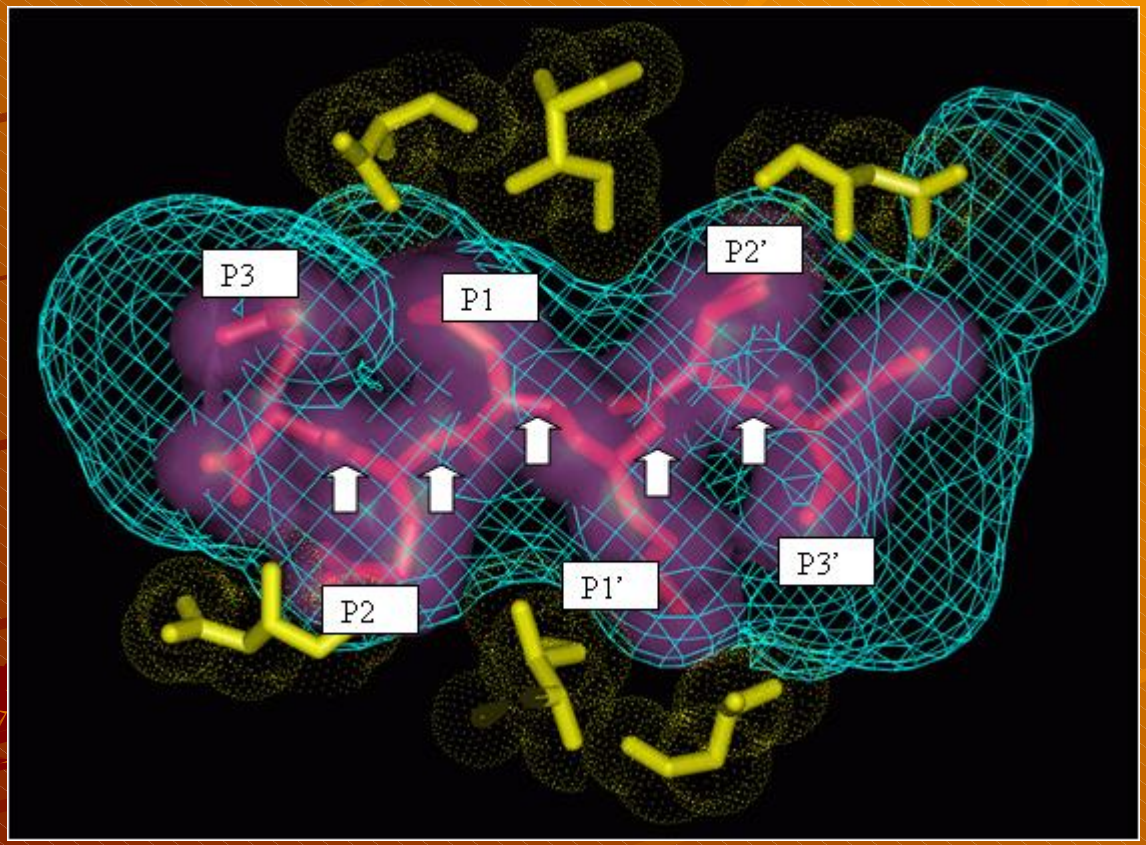
Problem

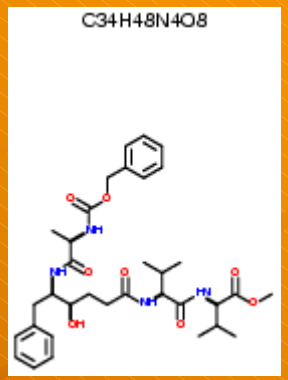
These formats need a wrapper to operate on them

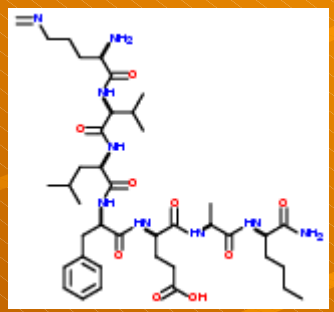
Goal

Eliminate the need of a wrapper and develop RDF statements that facilitate the above operations on structures in a database or Web environment to enable drug design

Elements of a RDF

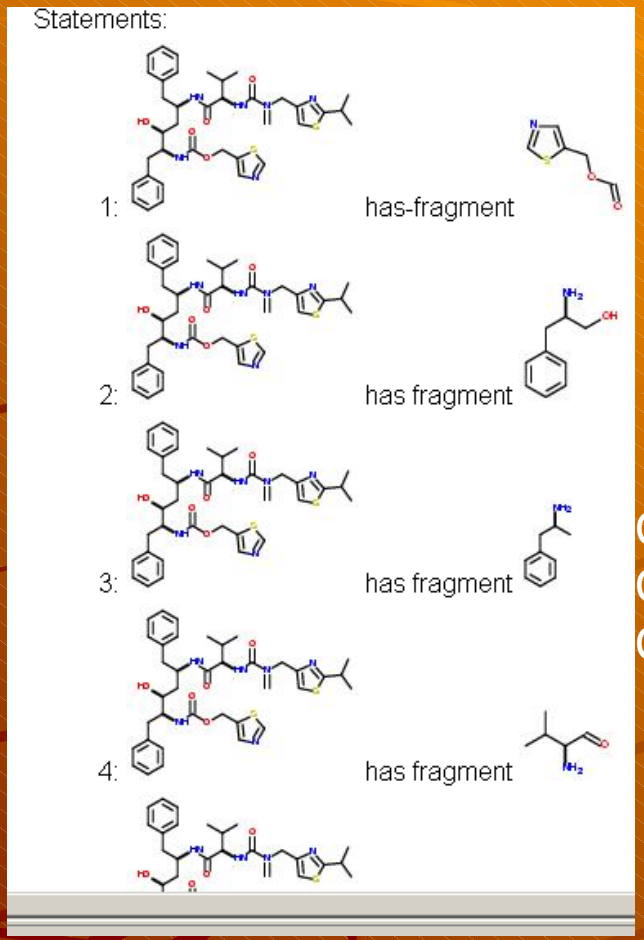






http://esw.w3.org/topic/HCLS/ChemicalTaxonomiesUseCase

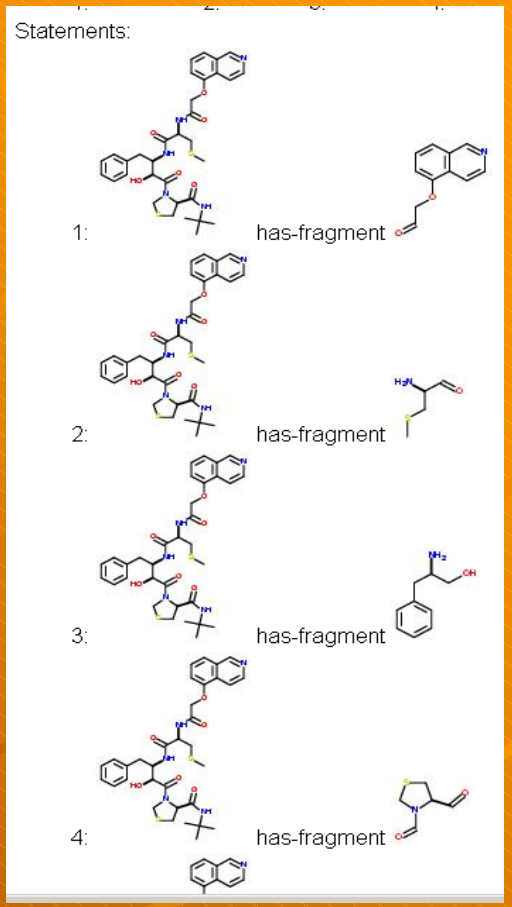
Each site such as P1, P2 is a building block for drug design
These sites are used as the elements of RDF
They are truly Semantic and mean 'some thing for drug design'



Chemical RDF

Can be added
Can be subtracted
Can be intersected

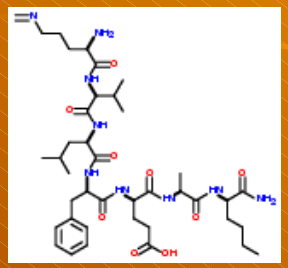
And, all, or,
at least, all most



From these statements computers are capable of automatically answering questions such as (using SQL in ORACLE/MySQL)

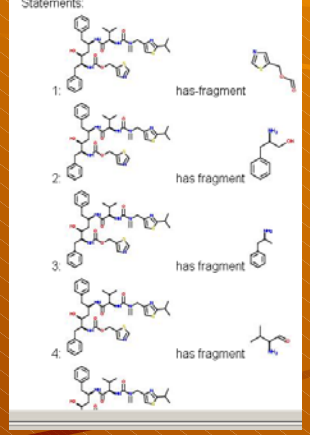
What fragments are common between a set of compounds? (intersection)
What fragments are unique for a compound? (not in)
What are best neighbors of a compounds? (max(intersection))

We are now building a set of rules and schema for a Semantic Web for chemical compounds to answer these questions dynamically in a Web environment



A Review of Indexing Methods

- Proper indexing (ontology) of data is key in developing a database to support Semantic Web
  - Data indices are the only means by which a Web tool may hold, organize, select and display data
- PDB (three-letter code, like naming of hurricanes), chemical abstracts (CAS number) are not suitable for machine reasoning
  - They are assigned by humans
  - They are not rule based
  - They have no structural rationale
    - 3TC and 4TC may not be structurally more similar when compared to 3TC and 6DA
    - One may not use such indices to infer relationships
- IUPAC International Chemical Identifier (InChI) is machine assigned structure based identifier
  - 1/C10H10S/c1-2-5-9(6-3-1)10-7-4-8-11-10/h1-3,5-7H,4,8H2
  - Prasanna M, Vondrasek, J., Wlodawer, A., Bhat, TN. Application of InChI to curate, index and query 3-D structures. Proteins, Structure, Function, and Bioinformatics 2005;60:1-4.
  - Murray-Rust P, Rzepa HS, Stewart JJ, Zhang Y. A global resource for computational chemistry. J Mol Model (Online) 2005;11(6):532-541.



Indexing of Structures

InChI is Good but not Good Enough

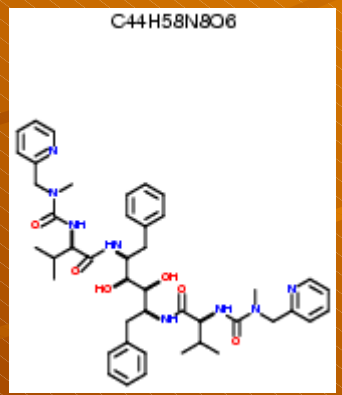
Indexing using the RDF elements solves the problem

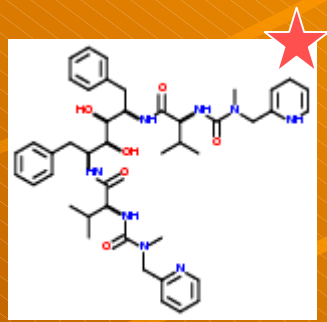
Use of InChI turned out to problematic for 3-D Structures

1. InChI is sensitive to observed variations in structures
2. InChI uses a data tree of depth =1

The problem is solved if indexed on elements of RDF statements

Example RDF finds new related structure

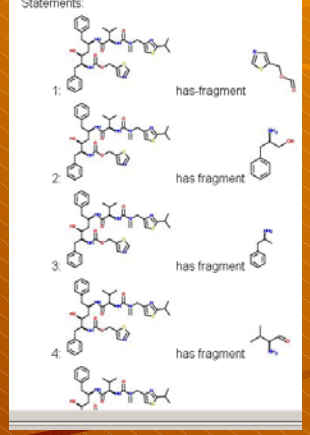




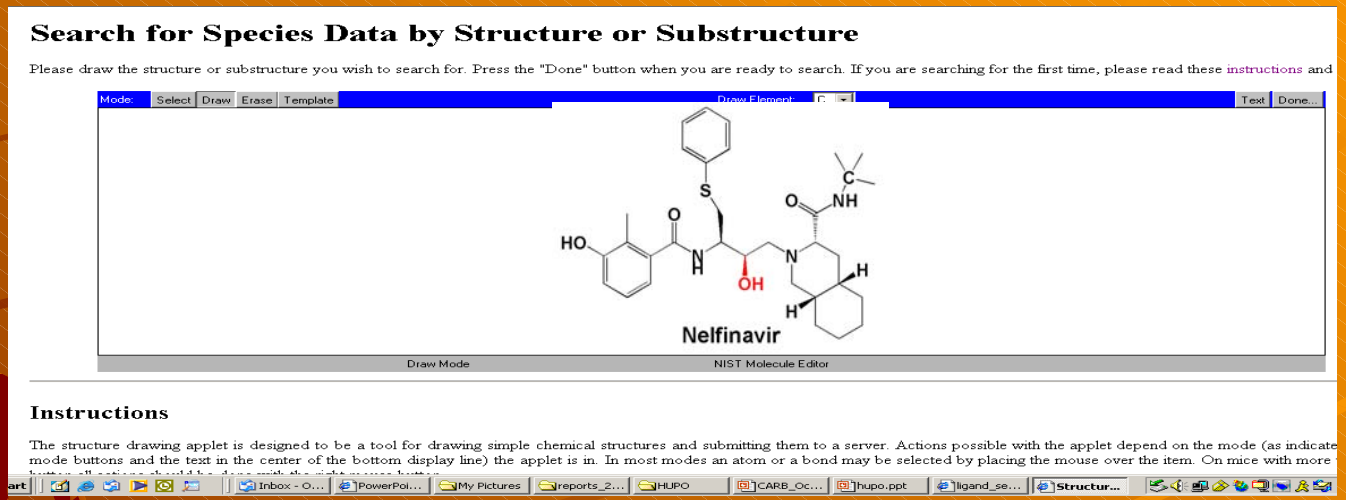
- Use InChI (Iupac International Chemical Identifier) Developed by NIST (Steve Stein)
  - **Application of InChI to curate, index, and query 3-D structures.** Prasanna MD, Vondrasek J, Wlodawer A, Bhat TN.

    Proteins. 2005 Jul 1;60(1):1-4.
- InChI is an identifier using chemical connectivity, chirality,… of the molecule
  - 1/C10H10S/c1-2-5-9(6-3-1)10-7-4-8-11-10/h1-3,5-7H,4,8H2

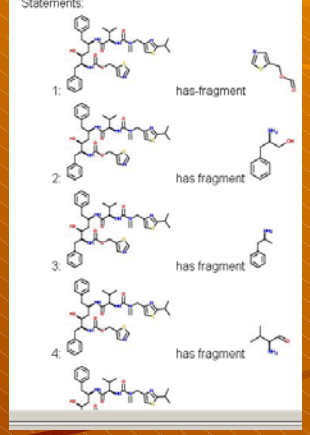
1HVC. Chemical & Crystallographic structures get different InChI



Current Query Technique



- These methods 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



Current Query Technique is Difficult to Use

- At present there is miss-mach between the way drugs are indexed and the way drugs are queried in many databases e.g., PDB, Pub-CHEM,...
  - Drugs are indexed as a whole, for instance using InChI
  - However, 99% of the queries are based on fragments specified either by a drawing window or by an input text chosen from an IUPAC name
    - This miss match results in missed hits or overwhelming hits
  - If we use the proposed RDF based index, the problem is solved

hivsdb-single-file-org Protégé 3.2 beta (file:\C:\from_jb\hivsdb-single-file-org.pprj, OWL / RDF Files)

File Edit Project OWL Code Tools Window Help

Metadata | OWLClasses | Properties | Individuals | Forms | Instance Tree | RQL

CLASS BROWSER

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)

Taxonomy is rule based

INSTANCE BROWSER

For Class: compound

Asserted | Inferred

Asserted Instances

_1AXA
_1B11
_1B6J
_1B6K
_1B6L
_1B6M
_1B6N
_1B6O
_1B6P
_1BDL
_1BDQ
_1BDR
_1BV7
_1BV9
_1BVE
_1BVG
_1BWA
_1BWB
_1C6X
_1C6Y
_1C6Z
_1C70
_1CPI
_1D4H
_1D4I
_1D4J
_1D4K

INDIVIDUAL EDITOR

For Individual: _1A30

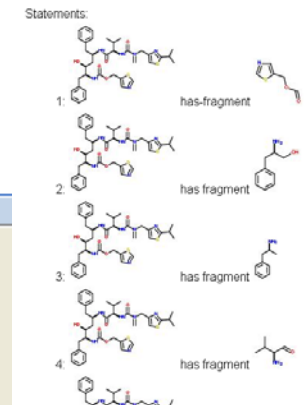
protege:inferredType

compound

has-fragment

_1_2FC4H7NO3_2Fc5-3_282-6_291-4_287_29
_1_2FC5H9NO3_2Fc6-4_283-7_291-2-5_288_2
_1_2FC6H13NO2_2Fc1-4_282_293-5_287_296

Replace
/ by _2F
‘ ‘, (,) by _



For Project: hivsdb-single-file-org

For Class: fragment

For Individual:

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)

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

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...

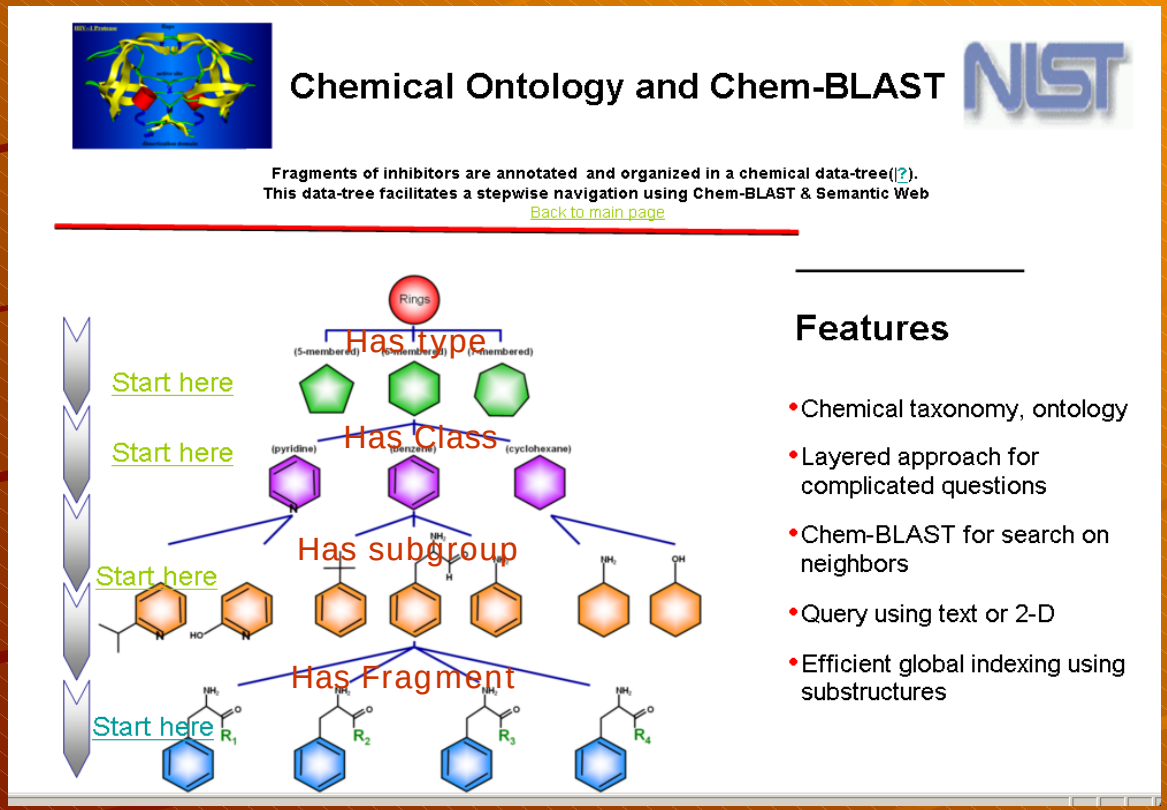
member-of-group

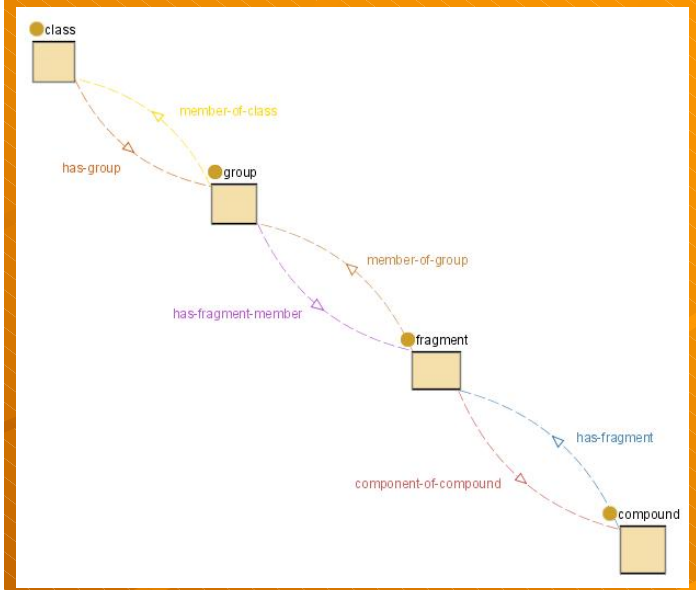
- cyclopentene00
- benzene00
- indane00

component-of-compound

- _1NPA
- _1NPW

Layered data indexing using elements of RDF solves the problem

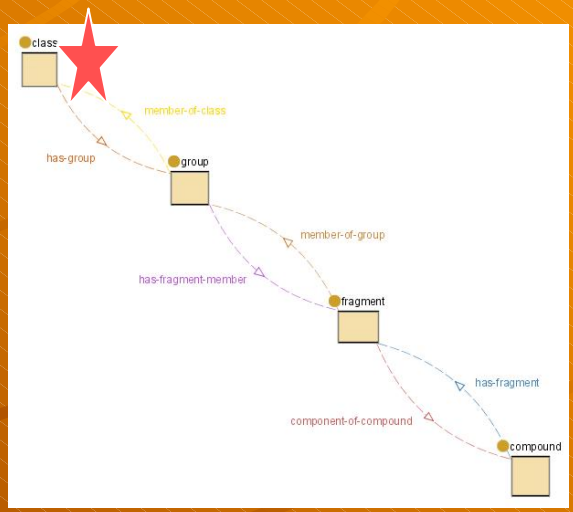


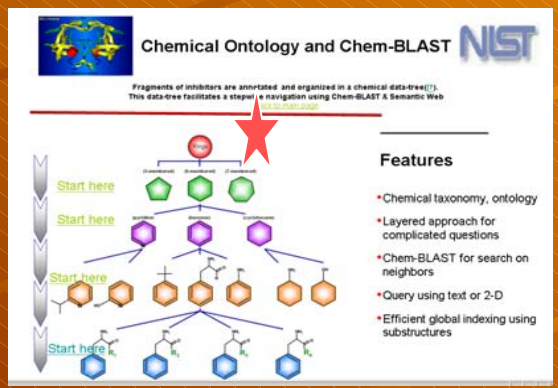


Layered query similar to Protégé or other Semantic Search engines
Each element is uniquely indexed by a rule that is based on its structure

Class layer

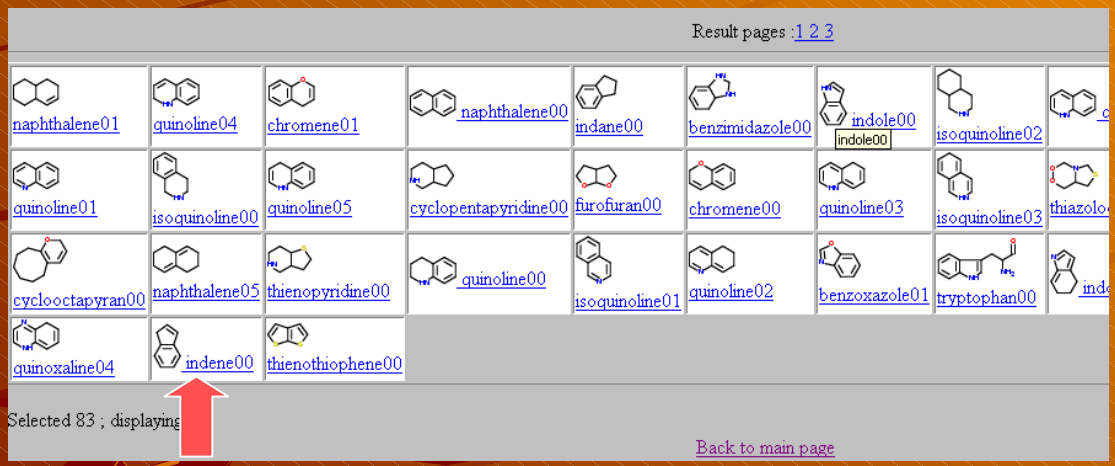
six-member ring	double fused ring
nitrogen containing	amino acid
oxygen containing	five-member ring
sulfur containing	seven-member ring
phosphorus containing	three-member ring

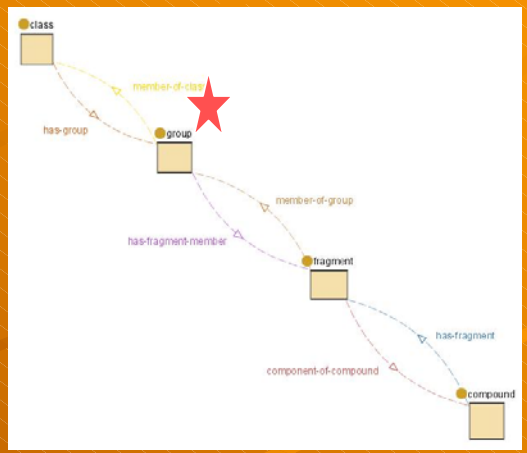


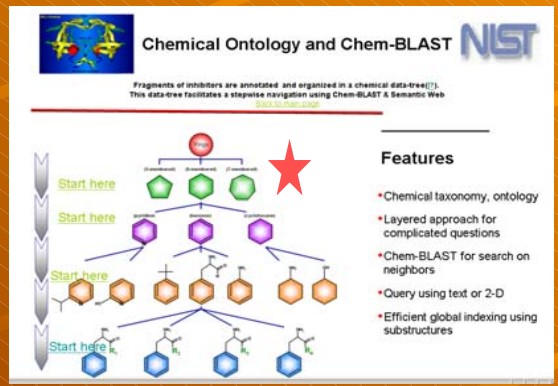


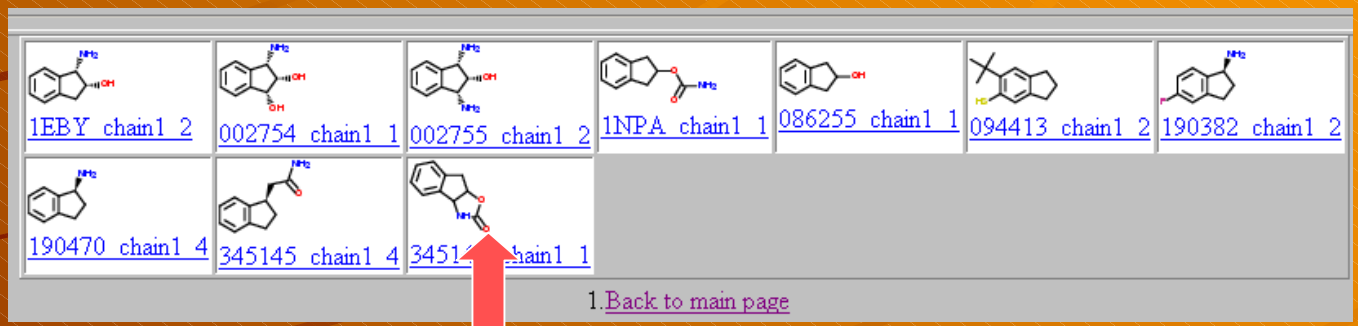
Fragment Layer for Fused Rings

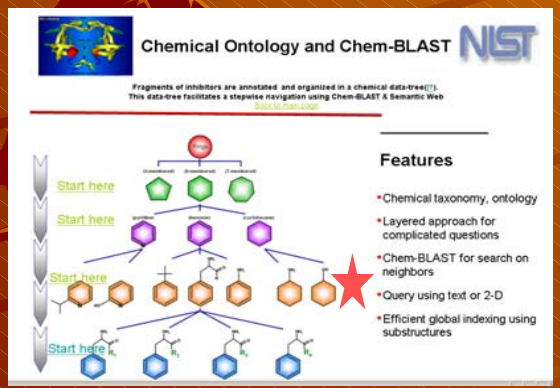
- A user gets to see all the rings of known drugs to query related compounds

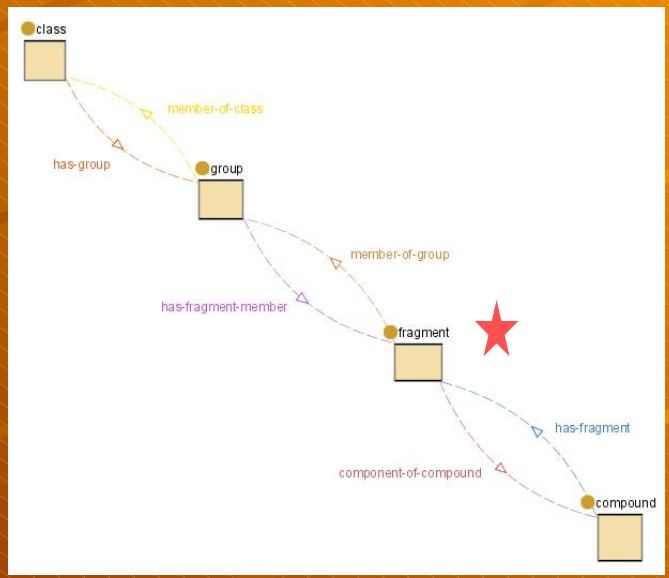


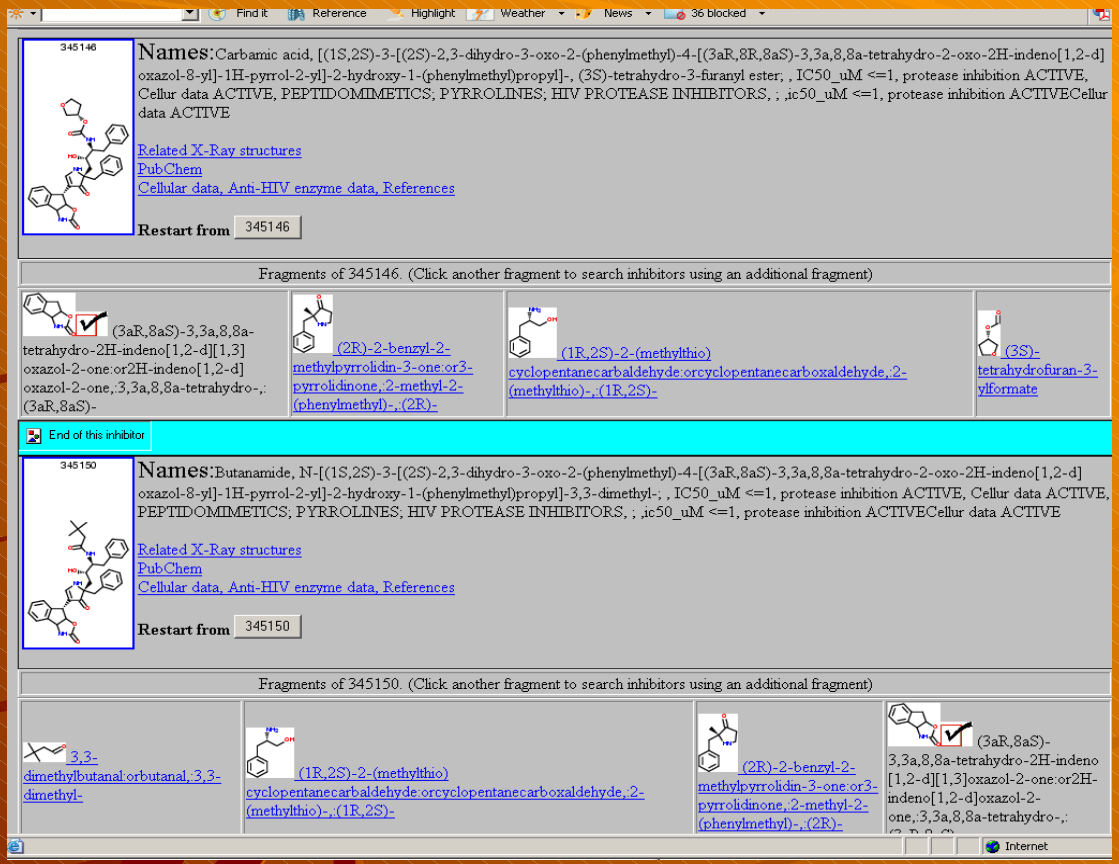


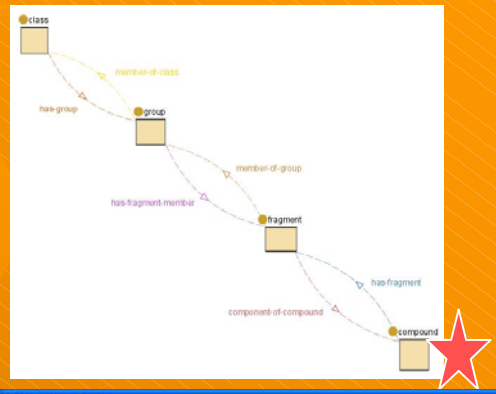


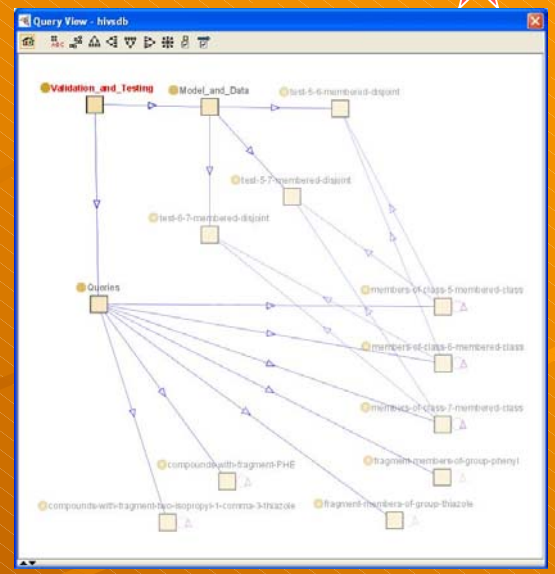

1EBY chain1 2
002754 chain1 1
002755 chain1 2
1NPA chain1 1
086255 chain1 1
094413 chain1 2
190382 chain1 2
190470 chain1 4
345145 chain1 4
1. Back to main page


Chemical Ontology and Chem-BLAST
NIST
Fragments of inhibitors are annotated and organized in a chemical data-tree(!).
This data-tree facilitates a stepwise navigation using Chem-BLAST & Semantic Web
Start here
Start here
Start here
Start here
Features
•Chemical taxonomy, ontology
•Layered approach for complicated questions
•Chem-BLAST for search on neighbors
•Query using text or 2-D
•Efficient global indexing using substructures

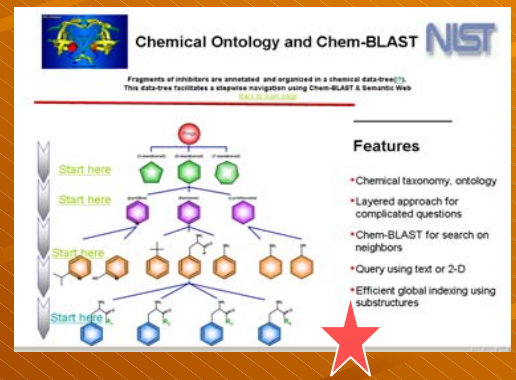

class
member-of-class
has-group
group
member-of-group
has-fragment-member
fragment
has-fragment
component-of-compound
compound



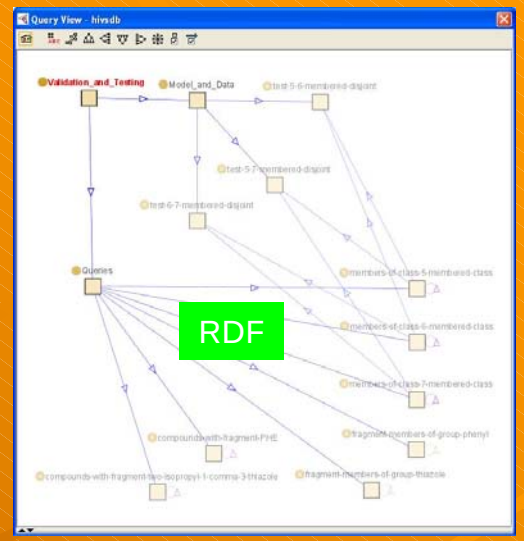


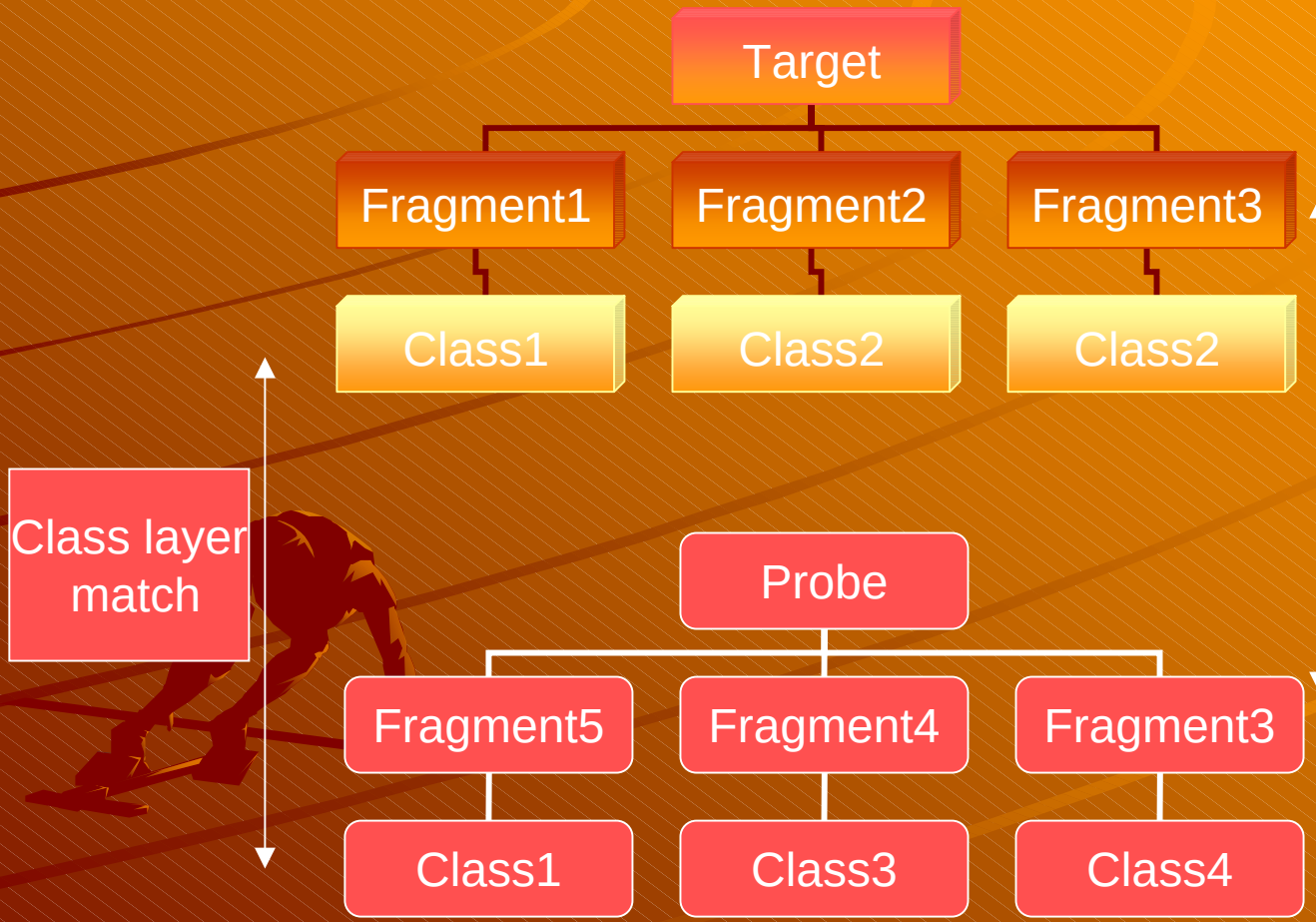


Elements of RDF can be used for complicated queries involving 'intersection', 'in', 'not in' type of searches in a database environment



Chem-BLAST
Chemical Block Layered Alignment of Substructure Technique





Chem-BLAST uses structure relationships (RDF) in a Oracle/ MySQL

Use XML Entities

Source Code

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://localhost/web/hiv-pdb/hivsdb-ind.owl#"
     xml:base="http://localhost/web/hiv-pdb/hivsdb-ind.owl"
     xmlns:pl="http://www.owl-ontologies.com/assert.owl#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:dc="http://purl.org/dc/elements/1.1/"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:owl="http://www.w3.org/2002/07/owl#">
    <owl:Ontology rdf:about="">
        <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >&lt;p style=&quot;margin-top: 0&quot;&gt;

    &lt;/p&gt;</rdfs:comment>
    </owl:Ontology>
    <fragment rdf:ID="_1_2FC12H1002_2Fc13-8-9-14-12-7-3-5-10-4-1-2-6-11_2810_2912_2Fh1-8H_2C9H2">
        <has-inchi-name rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >1/C12H1002/c13-8-9-14-12-7-3-5-10-4-1-2-6-11(10)12/h1-8H,9H2</has-inchi-name>
        <has-iupac-name rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >(1-naphthyloxy)acetaldehyde</has-iupac-name>
        <component-of-compound rdf:resource="#_1HIV"/>
        <component-of-compound rdf:resource="#_1IVP"/>
        <component-of-compound rdf:resource="#_HIV25NCI"/>
        <member-of-group rdf:resource="#benzene00"/>
        <member-of-group rdf:resource="#naphthalene00"/>
    </fragment>
    <fragment rdf:ID="_1_2FC12H17F02S_2Fc1-9_282_298-10_283_2916_2814_2C15_2912-6-4-11_2813_295-7-12_2Fh4-
        <has-inchi-name rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >1/C12H17F02S/c1-9(2)8-10(3)16(14,15)12-6-4-11(13)5-7-12/h4-7,9-10H,8H2,1-3H3</has-inchi-name>
        <has-iupac-name rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >(1R)-1,3-dimethylbutyl 4-fluorophenylsulfone</has-iupac-name>
        <component-of-compound rdf:resource="#_1XL5"/>
        <member-of-group rdf:resource="#benzene00"/>
        <member-of-group rdf:resource="#oxosulfane_oxide00"/>
        <member-of-group rdf:resource="#sulfur00"/>
    </fragment>
    <fragment rdf:ID="_1_2FC10H12O2_2Fc1-8-4-3-5-9_282_2910_288_2912-7-6-11_2Fh3-6H_2C7H2_2C1-2H3">
        <has-inchi-name rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >1/C10H12O2/c1-8-4-3-5-9(2)10(8)12-7-6-11/h3-6H,7H2,1-2H3</has-inchi-name>
        <has-iupac-name rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >(2,6-dimethylphenoxy)acetaldehyde</has-iupac-name>
        <component-of-compound rdf:resource="#_1XL5"/>
        <member-of-group rdf:resource="#benzene00"/>
    </fragment>
    <fragment rdf:ID="_1_2FC8H9NO_2Fc9-8-4-2-1-3-7_288_295-6-10_2Fh1-4_2C6H_2C5_2C9H2">
```

Summary

- Structure based drug design method is described
- Current methods of indexing and querying of structures are reviewed
- A novel chemical taxonomy for structure based drug design using Semantic Web concepts is proposed and illustrated for AIDS research
- Use of the rule based chemical taxonomy for developing RDF for Semantic Web is illustrated
- Implementation of this taxonomy is database (ORACLE & MySQL) friendly and it provides better on-line experience both for developers and users of the database
- 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 result set to get what he wants- Web page lays out the contents of the database in an orderly fashion for pick and choose
- 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
- http://xpdb.nist.gov/hivsdb/advanced_query_files/slide0002.htm