DAS searches

a work in progress report...

Andreas Prlić

UCSD

Standardizing searches is difficult

- Many possible questions
- Performance (in DAS < 10 sec...)
- Server stability

Some requirements

- Perform a complex search in a standardized but simple way
- Get a list of IDs, genome coordinates

A use case

- PDB advanced search

my developmental server: http://chili.rcsb.org/pdb/rest/

MyPDB

A MEMBER OF THE PDB

RCSB PDB PROTEIN DATA BANK

An Information Portal to Biological Macromolecular Structures

As of **Thursday Jan 15, 2009** there are 56217 Structures | **PDB Statistics**

Contact Us | **Feedback** | **Help** | **Print** | **PDB ID or keyword** | **Author** | Site Search | *Advanced Search*

Home | Search

- Search Database
  - **Advanced Search**
  - Latest Release
  - Sequence
  - Chemical Structure
  - Chemical Name
  - Chemical ID
  - Models
  - Unreleased Entries
  - Structural Genomics Targets
  - How to Search
- Browse Database
- Histograms

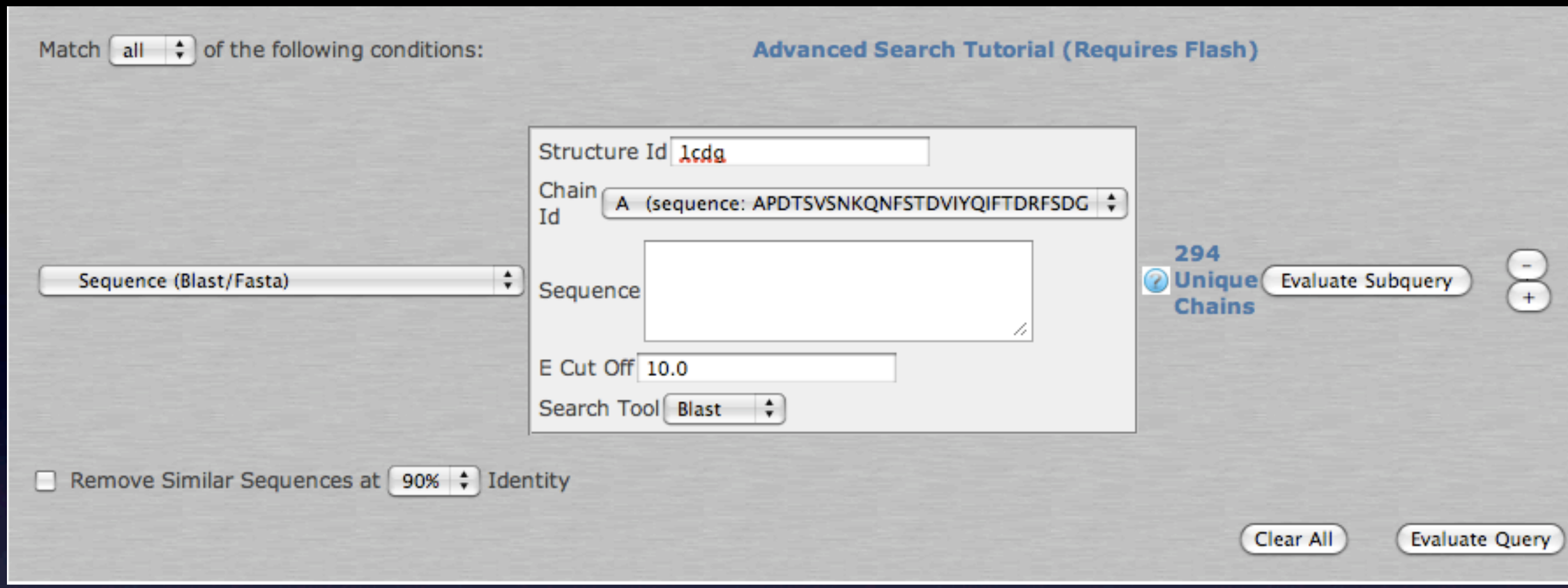
Match all of the following conditions: **Advanced Search Tutorial (Requires Flash)**

Evaluate Subquery

Clear All | Evaluate Query

- ✓ Choose a Query Type:
- ID Search for Structures
  - PDB ID(s)
  - PubMed ID(s)
  - UniProt Accession ID(s)
  - Genbank ID(s)
  - PIR ID(s)
  - PDB Entity ID(s)
- Structural Genomics
  - Structural Genomics Project
- Structure Summary
  - Structure Title
  - Structure Description
  - Molecule Name
  - Author Name
  - Deposit Date
  - Release Date
  - Latest Release
  - Experimental Method
  - Molecule / Chain Type
  - Oligomeric state (number of chains)
  - PDB Holdings
  - Model Count
- Keyword
  - Advanced
  - in PubMed
  - Medical Subject Headings (opens popup)
  - Author Assigned
- Structure Features
  - Secondary Structure
  - Secondary Structure Length
  - Disulphides
  - SCOP Classification (opens popup)
  - CATH Classification (opens popup)
- Sequence Features

Rutgers ... ers of the **RCSB**: ... **NLM**, **NCI**, **NINDS**, and **NIDDK**.


Match all of the following conditions:
Advanced Search Tutorial (Requires Flash)
Structure Id 1cdg
Chain Id A (sequence: APDTSVSNKQNFSTDVIYQIFTDRFSDG
Sequence (Blast/Fasta)
Sequence
294 Unique Chains
Evaluate Subquery
E Cut Off 10.0
Search Tool Blast
Remove Similar Sequences at 90% Identity
Clear All
Evaluate Query

Match all of the following conditions:

Advanced Search Tutorial (Requires Flash)

Sequence (Blast/Fasta)

Structure Id 1cdg

Chain Id A (sequence: APDTSVSNKQNFSTDVIYQIFTDRFSDG

Sequence

E Cut Off 10.0

Search Tool Blast

294 Unique Chains

Evaluate Subquery

Structure Title

Contains: transferase

2008 Structures

Evaluate Subquery

Remove Similar Sequences at 90% Identity

Clear All

Evaluate Query

Help | 42 Structure Hits | 111 Unreleased Structures | 22 Citations | 17 Ligand Hits | GO Hits | SCOP Hits | CATH Hits

Sequence Search (Structure:Chain = 1CDG:A, Expectation Value = 10.0, Search Tool = blast)
and
StructTitleQuery: struct.title.comparator=contains struct.title.value=transferase

1 2 3 4 5

1BPL

GLYCOSYLTRANSFERASE

Chains A

Characteristics Release Date: 17-Aug-1996 Exp. Method: X Ray Diffraction

Resolution: 2.20 Å

Classification Glycosyltransferase

Compound

Polymer: 1

Molecule: ALPHA-1,4-GLUCAN-4-GLUCANOHYDROLASE Chains: A type: polypeptide(L) length: 189 EC no.: 3.2.1.1

Polymer: 2

Molecule: ALPHA-1,4-GLUCAN-4-GLUCANOHYDROLASE Chains: B type: polypeptide(L) length: 294 EC no.: 3.2.1.1

Authors Machius, M., Wiegand, G., Huber, R.

1BPL

GLYCOSYLTRANSFERASE

Chains B

Characteristics Release Date: 17-Aug-1996 Exp. Method: X Ray Diffraction

Resolution: 2.20 Å

Classification Glycosyltransferase

Compound

Polymer: 1

Sequence Search (Structure:Chain = 1CDG:A, Expectation Value = 10.0, Search Tool = blast)

StructTitleQuery: struct.title.comparator=contains struct.title.value=transferase

```
<orgPdbCompositeQuery version="1.0">
 <queryRefinement>
  <queryRefinementLevel>0</queryRefinementLevel>
  <orgPdbQuery>
    <version>head</version>
    <queryType>org.pdb.query.simple.SequenceQuery</queryType>
    <description><![CDATA[Sequence Search (Structure:Chain = 1CDG:A, Expectation Value = 10.0, Search Tool = blast)]]></description>
    <runtimeMilliseconds>6660</runtimeMilliseconds>
    <structureId><![CDATA[1CDG]]></structureId>
    <chainId><![CDATA[A]]></chainId>
    <sequence><![CDATA[APDTSVSNKQNFSTDVIYQIFTDRFSDGNPANNPTGAAFDGTCTNLRLYCGGDWQGIINKINDGYLTGMGVTAIWISQPVENIYSIINYSGVNNTAYHGYWARDFKKTNPAYGTIADFQ
    <eCutOff><![CDATA[10.0]]></eCutOff>
    <searchTool><![CDATA[blast]]></searchTool>
  </orgPdbQuery>
 </queryRefinement>
 <queryRefinement>
  <queryRefinementLevel>1</queryRefinementLevel>
  <conjunctionType>and</conjunctionType>
  <orgPdbQuery>
    <version>head</version>
    <queryType>org.pdb.query.simple.StructTitleQuery</queryType>
    <description><![CDATA[StructTitleQuery: struct.title.comparator=contains struct.title.value=transferase ]]></description>
    <runtimeMilliseconds>371</runtimeMilliseconds>
    <struct.title.comparator><![CDATA[contains]]></struct.title.comparator>
    <struct.title.value><![CDATA[transferase]]></struct.title.value>
  </orgPdbQuery>
 </queryRefinement>
</orgPdbCompositeQuery>
```

Home Search

- Home
- Getting Started
- Structural Genomics
- Download Files
- Deposit and Validate
- Dictionaries & File Formats
- Software Tools
- General Education
  - RCSB PDB Newsletters
  - **Educational Resources**
  - Molecule of the Month
  - Education Corner
  - Looking at Structures
  - Proteins in the Sea
- Site Tutorials
- BioSync
- General Information
- Acknowledgements
- Frequently Asked

The RCSB RESTful web service interface

At the present the following RESTful web services are supported by RCSB:

- A generic **SEARCH service** allowing to POST advanced queries
- A **DAS service** to obtain PDB to UniProt mappings

SEARCH service:

At the present this interface exposes the RCSB **advanced search** interface as a XML web service.

In order to utilize this service, POST a XML representation of an advanced search to **/pdb/rest/search**.

XML representation of advanced search

Every **advanced search** can be represented by XML. To access this representation simply execute an advanced search query and then click on the **Result tab**. One of the links on the left hand menu is **Show Query Details**. Clicking on this link will provide you of the XML representation of your query.

Every query is described by two data items:

- **queryType:** the name of the class that is implementing the query
- **arguments:** depending on the type of query that is being executed one or more differently named arguments need to be provided.

Example:

```
<orgPdbQuery>
    <queryType>org.pdb.query.simple.UpAccessionIdQuery</queryType>
    <description>Simple query for a list of Uniprot Accession IDs: P50225</description>
    <accessionIdList>P50225</accessionIdList>
</orgPdbQuery>
```

- Name of query
- Parameters as key/value pairs
- Join operation between multiple queries

- also available via SOAP...
- also PDB - UniProt alignment DAS server (SIFTS)

- client API/library?
- other result types (sequences/ mmCif items)
- bi-directional web services (data exchange with other bio websites - Pfam/Hmmer)

Acknowledgements

- RCSB-PDB web site:
- Phil Bourne, Peter Rose
- Wolgang Bluhm, Gregg Quinn, Bojan Beran, Ben Yukich
- Viewers: Rick Berger, John Beaver

XML query:
Wayne Townsend Merin

We are hiring...