Review of Crash Course Objectives and Arabidopsis Case Study

Christopher S. Henry

November 10th, 2022
Course Objectives

1. Teach those who are new to PDB or KBase a little more about these platforms
2. Highlight new tools that have been added to KBase to enhance connections to PDB
3. Demonstrate how these tools can be applied to discover new functions for genes in plant and microbial genomes
4. Generally improve audience knowledge in key tools in KBase (modeling, genomics, annotation) and PDB (mol*, structure query, structure alignment)
5. Obtain your feedback on the tools we have built so far and your suggestions on what we should build next
Course Outline

1. Burley - Introduction
2. Henry - Crash Course Objectives and Arabidopsis Case Study
3. Edirisinghe - Microbial Case Study Involving Pyridine
4. Piehl - Studying Experimental Structures in PDB - Arabidopsis resistosome
5. Vallat - Exploring tools and data in PDB to aid in function discovery
6. Dutta - Detailed tutorial on visualizing structure data in PDB
7. Zhang - Overview of all structure tools currently in KBase
8. Henry - Exploring other information to be gained from PDB tools in KBase and discussing future plans
9. Stephen - Future plans and concluding remarks
What is KBase?

KBase is an integrated platform for aggregating and sharing tools and data in order to collaboratively solve scientific problems.
What is KBase?

Data Management

• Unlimited data storage
• Import data by
  • Drag & drop
  • Globus
  • FTP, HTTP, Gdrive, Box/Dropbox
• File types
  • FASTQ, FASTA, SRA, GenBank, gff, expression matrix, media, phenotype, FBA models
What is KBase?

Data Analysis

• App Catalog
  • ~280 apps
  • Read processing
  • Genome assembly
  • Genome annotation
  • Sequence alignment
  • Comparative genomics
  • Metabolic modeling
  • Expression
  • Microbial communities

App Catalog

~280 apps
Read processing
Genome assembly
Genome annotation
Sequence alignment
Comparative genomics
Metabolic modeling
Expression
Microbial communities
What is KBase?

Collaboration

• As users conduct their work within notebooks called “narratives”, they can share those narratives with other users.

• Users can also gather their “narratives” into “Organizations”, or groups of users operating with a common mission in KBase.
What is KBase?

Narrative

- Object-oriented list of data entities
- Point and click app panel
- Notebook with analyses and markdown cells
- Buttons to share with users, make public, publish to HTML, write code cells
What is KBase?

Documentation

KBase Tutorial Narratives

These Narrative tutorials provide step-by-step examples that show how to use KBase tools and data to perform useful analyses. You can copy these Narratives and rerun the steps, modify the analyses, or try on your own data.

Multi-Omics Modeling Of Biochemical Pathways

Learn how to use visualization tools, analysis, and modeling of multi-omics data for understanding biochemical pathways in this virtual course from EMML.

Drafting Isolate Genomes

Genome Analysis in KBase Series focuses on how to use various tools to analyze genomic sequencing data in KBase. This first Narrative demonstrates a workflow for searching and identifying features within...

View Tutorial →

Searching For Features

Genome Analysis in KBase Series focuses on how to use various tools to analyze genomic sequencing data in KBase. This first Narrative demonstrates a workflow for searching and identifying features within...

View Tutorial →

Multi-Scale Microbial Dynamics Modeling

EMML Summer School 2023

Learn how to incorporate microbial metagenomic and environmental metatranscript data from watered ecosystems into metatranscriptome and community modeling using computational frameworks in this virtual course.

KBase For Educators

Modeled by Ellen Loew

Join the KBase community to learn best practices and computational biology!

Why KBase?

While KBase is routinely used for advanced analysis by scientists around

Assembly And Annotate Prokaryotic Genome Tutorial

Modeled by the KBase documentation team

KBase provides multiple Apps for de novo assembly of prokaryotic Next-Generation Sequencing (NGS) reads from various sequencing platforms. These assemblies can then be annotated with RAST or Prokka, enabling you to...

Many workflow specific tutorials and hundreds of videos on KBase youtube channel

U.S. DEPARTMENT OF

ENERGY

Office of Science

DOE KBase

774 subscribers

DOE KBase Quickstart Guide

Page last modified: 6 weeks ago

Please confirm! This is where to start learning about the basics of KBase and get on your way to doing advanced analysis with KBase. Follow us on QuikLab on Google+ Account: https://youtube/DOE KBase.
1. There is enormous and rapidly growing synergy between KBase and PDB

2. KBase contains extensive genomics data, particular non-reference user data

3. KBase has numerous workflows to integrate multi-omics data and comparative genomics approaches to predict and understand protein function

4. Advances in protein folding now make it possible to easily obtain a predicted structure for proteins of interest discovered in KBase, but what next?

5. Here PDB steps in offering a large database of experimental structures to aid in contextualizing new predicted structures generated for genes in KBase

6. PDB has a growing body of structure-related tools and data to enable users to actually gain functional insights from structure data
Case study in Arabidopsis...
Acknowledgements

We thanks DOE BER for funding this work!

KBase team

Special thanks to all those who attended and contributed to our KBase-PDB design workshops!
Identifying a Novel degradation Pathway with KBase Discovery Pipeline and PDB tools

Scientific Problem

- There are many microbial transformations yet to be characterized. How should we computationally predict potential gene candidates for novel enzymatic transformations?
- How can we effectively use RCSB data in aid in identifying predicted high confidence gene candidates?
Phenotypic data such as biology/growth experiments often show that microbes can degrade certain nutrients, but the degradation pathway is either poorly characterized or unexplained.

Scientists at ANL identified a *Micrococcus letus* strain that could degrade pyridine, but the pathway was unknown. The degradation pathway wasn’t explained until very recently.

We will apply KBase Functional Discovery Pipeline and knowledge that derives from RCSB data to identify potential pyridine degradation enzymes in *M.letus*.
Transcriptomics

MetaTranscriptomics

Metabolomics

4719 genes

1127 genes

1135 reactions

7320 novel reactions

Fill gaps with novel and known compounds

View key reactions and pathways

> 20 gene candidates (~15)

KBase Functional Discovery Pipeline

Link: Base Narrative Workflow
(https://narrative.kbase.us/narrative/127880)
Annotation of genomes and constructing metabolic models

Organisms / Microbiomes

Sequencing

Annotations

Metabolic Models

Documentation and a tutorial Narrative on Intro to Metabolic Models

https://docs.kbase.us/apps/analysis/metabolic-modeling

https://narrative.kbase.us/narrative/18302
Introducing new biochemistry in filling metabolic gaps

PickAxe app – Based on biochemical rules can be derived from known biochemistry is used to predict novel or promiscuous enzymatic reactions

210 enzymatic reaction rules been developed from the rules in the original BNICE approach by Hatzimanikatis et al.
Filling gaps in a metabolic network

- Extracellular compartment
- Metabolic Gaps
- Nutrients
- Genes
- Draft Model
- Gapfill on Pyridine media (No growth)
- Cheminformatics (Introduce Pyridine degradation reactions)
- Gapfill on Pyridine media (able to grow)
- Model Simulations - FBA
Discovering genes involved in pyridine degradation
Using cheminformatics to predict a novel pathway

Metabolic modeling workflow in KBase is used to predict the most likely pathway, then Escher is used to visualize with flux
Linking cheminformatics to genes in the genome
Structure prediction and analysis on selected gene candidates by the PDB team – example gene: CDS. 3483

• Query experimentally resolved structures that are corresponding to the potential gene candidates

• Query and learn from co-crystalized structures with the docking of the substrate

• Align experimental and computational structures to aid binding site identification and functional characterization
Acknowledgements

Special thanks to all those who attended and contributed to our KBase-PDB design workshops!

We thanks DOE BER for funding this work!

KBase team

Stephen Burley  John Westbrook  Ed O’loughlin
Qizhi Zhang  Kelly Skinner  Jack Gilbert
Sam Seaver  Claudia Lerma-Ortiz  Nidhi Gupta
Accessing Experimental Structures from the PDB

Dennis Piehl, Ph.D., RCSB PDB/Rutgers
Outline

● Introduction to the PDB Archive and RCSB PDB
● Overview of search tools
● Overview of download services and structure data format
● Tutorial: Accessing data from RCSB.org
  ○ Case study: Arabidopsis thaliana resistosome
  ○ Searching and browsing
  ○ Structure summary pages
  ○ Downloading data
  ○ Programmatic access (searching and downloading)
● Additional resources and documentation
History of the Protein Data Bank (PDB)

- Established in 1971 as the first open-access digital data resource in biology, with just seven protein structures
- Now hosts >190,000 experimental 3D structures of biomolecules, deposited by researchers worldwide
- The PDB Core Archive is jointly managed by the Worldwide PDB (wwPDB), comprised of the RCSB PDB (U.S.), PDBe (Europe), PDBj (Japan), BMRB (U.S./Japan), and EMDB (Europe)
- Committed to making PDB data “FAIR” (Findable, Accessible, Interoperable, & Reusable); all data is made freely available to the public

3D structural data from around the world

PDB data (curated and distributed by wwPDB members)

The RCSB PDB Web Portal (RCSB.org)

- **RCSB.org**: Tools for searching, visualizing, analyzing and downloading the contents of the PDB Archive
- Integration of PDB Archive data with annotations from ~50 external data resources (UniProt, SCOPe, CATH, …)
- Recently introduced tools and features:
  - Protein 1D-3D Feature View
  - Groups 1D-3D Alignment View
  - Integration of Computed Structure Models (CSMs)
- **PDB-101** (pdb101.rcsb.org): Educational resources and training

3D structural data from around the world

Search Tools at RCSB.org

Basic search

Advanced search
Search Tools at RCSB.org

- Basic search
- Advanced search
- Search API (search.rcsb.org)
Advanced Search Tools at RCSB.org

- Full-text “Google-like” search
- Structure or entry property (author, method, size, …)
- Chemical property (name, weight, …)
- Sequence identity search (MMseqs2)
- Amino acid regular expression (e.g., ST*G)
- Global structure similarity (in-house algorithm)
- Local structural motif (e.g., binding site residues)
- Chemical structure similarity (SMILES, InChI, sketch)
- Group structures by various criteria (sequence, deposit id, …)
Search API at RCSB.org
Download Services

● While browsing RCSB.org:
  ○ Structure summary pages (i.e., structure entry pages)
  ○ Search results page
  ○ Bulk download page (https://www.rcsb.org/downloads)

● PDB Archive:
  ○ Over HTTP (via direct navigation on browser or programmatic access):
    *Structure data under the directory: /pub/pdb/data/structures/divided/mmCIF/
    ■ https://s3.rcsb.org
    ■ https://files.rcsb.org (supports recursive directory retrieval)
    ■ Can use command-line/scripting tools (e.g., curl, wget, or Python)
  ○ Using rsync: rsync://rsync.rcsb.org
# Structure Data Format: PDBx/mmCIF

- **PDBx/mmCIF Data Standard** ([https://mmcif.wwpdb.org/](https://mmcif.wwpdb.org/))
  - Became the standard file format for data in the PDB Archive in 2014
  - Files are given the extension “.cif” (or “.cif.gz”)
  - Supersedes the now legacy “PDB” (“.pdb”) file format, which is limited by fixed-column width restrictions and is no longer extended to support new types of metadata

## FASTA Sequence

```plaintext
# data_6J5T
# _entry_id 6J5T
# _audit_conform.dict_name mmcif_pdbx_dic
# _audit_conform.dict_version 5.303
# _audit_conform.dict_location http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx_dic
#
# loop
# _database_2.database_id _database_2.database_code
# PDB 6J5T
# WWDPDB D_1300010525
# EMDB EMDB-0680
# _pdbx_database_related.db_name EMDB
# _pdbx_database_related.details 'Reconstitution and structure of a plant NLR resistosome conferring immunity'
# _pdbx_database_related.db_id EMDB-0680
# _pdbx_database_related.content_type 'associated EM volume'
```

## PDB Format

- **Legacy**
Case study: *Arabidopsis thaliana* resistosome

- *Arabidopsis thaliana*: the gold standard model organism in plant molecular biology
- In plants, the innate immune response is triggered via the sensing of pathogens by nucleotide-binding, leucine-rich repeat receptors (NLRs)
- Upon pathogen sensing, NLRs oligomerize to form the activated “resistosome”
- Ultimately, this process culminates in programmed cell death

Case study: *Arabidopsis thaliana* resistosome

Case study: *Arabidopsis thaliana* resistosome

**Indirect pathogen sensing**

Tutorial: Accessing structure data in the PDB

NLR (nucleotide-binding, leucine-rich repeat receptor)

Indirect pathogen sensing

Let’s try to find these structures

Tutorial: Accessing structure data in the PDB

NLR (nucleotide-binding, leucine-rich repeat receptor)

Indirect pathogen sensing

Let’s try to find these structures

Resistosome

Tutorial: Accessing structure data in the PDB

Tutorial time
Summary: Accessing structure data in the PDB

- **Interactive access (i.e., while browsing RCSB.org):**
  - Search data with “Basic” and “Advanced” search tools (https://www.rcsb.org/search/advanced)
  - Download data via:
    - Structure summary pages
    - Search results page
    - Direct navigation of the PDB Archive at https://s3.rcsb.org or https://files.rcsb.org
      - Structure data under the directory: /pub/pdb/data/structures/divided/mmCIF/

- **Programmatic access:**
  - Search data with our Search API (try the query editor: https://search.rcsb.org/query-editor.html)
  - Download data via command-line scripted access of the PDB Archive
    - Over HTTP (e.g., using curl, wget, or Python):
      - https://s3.rcsb.org
      - https://files.rcsb.org (supports recursive directory retrieval)
    - Using rsync: rsync://rsync.rcsb.org
Resources and Documentation

- **Basic and advanced searching:**
  - Infographic: [https://cdn.rcsb.org/rcsb-pdb/search/SearchnBrowse2go.pdf](https://cdn.rcsb.org/rcsb-pdb/search/SearchnBrowse2go.pdf)
  - [https://www.rcsb.org/docs/search-and-browse/overview-search-and-browse](https://www.rcsb.org/docs/search-and-browse/overview-search-and-browse)

- **Programmatic access:**
  - Search API: [https://search.rcsb.org/index.html#search-api](https://search.rcsb.org/index.html#search-api)
  - **Downloading structures:**
    - [https://www.rcsb.org/docs/programmatic-access/file-download-services](https://www.rcsb.org/docs/programmatic-access/file-download-services)
    - AWS S3: [https://www.rcsb.org/news/6266e0379c3931864b072861](https://www.rcsb.org/news/6266e0379c3931864b072861)
    - [https://www.rcsb.org/docs/programmatic-access/batch-downloads-with-shell-script](https://www.rcsb.org/docs/programmatic-access/batch-downloads-with-shell-script)

**Questions?**
RCSB PDB Team

RCSB.ORG
info@rcsb.org

Funding
National Science Foundation (DBI-1832184),
National Institute of General Medical Sciences,
National Institute of Allergy and Infectious Disease,
and National Cancer Institute (NIH R01GM133198),
and the
US Department of Energy (DE-SC0019749)

Management

Member of the
Worldwide Protein Data Bank
(wwPDB; wwpdb.org)

Follow us

John D. Westbrook
In memoriam
1957-2021
Accessing Computed Structure Models Generated Using AlphaFold2 and RoseTTAFold

DOE KBASE/RCSB PDB VIRTUAL CRASH COURSE
NOVEMBER 10, 2022

Brinda Vallat, Ph.D., RCSB PDB/Rutgers
brinda.vallat@rcsb.org
Outline

● What are computed structure models (CSMs)?
  ○ Protein structure prediction
  ○ AlphaFold2 and RoseTTAFold
  ○ AlphaFoldDB and ModelArchive
  ○ Model quality metrics
  ○ ModelCIF data standard

● CSMs in RCSB.org

● Accessing CSMs on RCSB.org
  ○ Live demo: RCSB.org tools and functionalities
  ○ Case studies: Disease resistance RPP13-like protein 4 from Arabidopsis thaliana resistosome and Micrococcus luteus aldehyde dehydrogenase
What are computed structure models?
Protein structure prediction

Template-based modeling

Template-free modeling

Computed structure models (CSMs) are predicted models obtained using template-based or template-free modeling methods.
AlphaFold2 and RoseTTAFold

- AlphaFold2 and RoseTTAFold are artificial intelligence / machine learning (AI/ML) methods that predict a protein’s 3D structure from its amino acid sequence
- AlphaFold2 is developed by Google Deepmind
  - Achieves accuracy comparable to low resolution experimental structures
- RoseTTAFold is developed by HHMI investigator David Baker and his team at the University of Washington
  - Prediction accuracies approach that of AlphaFold2
- These prediction methods were trained using thousands of sequences and structures of known proteins from the PDB
- **PDB data is the primary input for training these AI/ML prediction methods**

AlphaFold Protein Structure Database and ModelArchive

- **AlphaFold Protein Structure Database** (https://alphafold.ebi.ac.uk) provides open access to AlphaFold2 predictions
  - Hosted at the European Bioinformatics Institute (EBI)
  - Includes >200 million protein structure predictions, providing a broad structural coverage of sequences in UniProt

- **ModelArchive** (https://www.modelarchive.org) is a repository for CSMs referenced in publications
  - Hosted at the Swiss Institute of Bioinformatics (SIB)
  - Includes predictions of core eukaryotic heteromeric protein complexes modeled using a combination of RoseTTAFold and AlphaFold2

Model quality metrics

Two types of intrinsic model accuracy estimates:

- pLDDT: per-residue measure of local confidence on a scale from 0 - 100
- PAE: expected position error at residue x, when the predicted and true structures are aligned on residue y

ModelCIF Data Standard

- Extension of PDBx/mmCIF for CSMs
- Definitions retained from PDBx/mmCIF
  - Representation of small molecules, polymeric macromolecules, complexes, atomic coordinates and relevant metadata (e.g., authors, citations, software)
- New definitions in ModelCIF
  - Targets, templates, alignments, coevolution data, predicted contacts, model quality metrics
- Publicly available via GitHub: github.com/ihmwg/ModelCIF
- Managed by the wwPDB ModelCIF working group: wwpdb.org/task/modelcif
- Supported by AlphaFoldDB and the ModelArchive
CSMs in RCSB.org
CSMs in RCSB.org

● RCSB.org now delivers >1,000,000 CSMs together with ~200,000 experimental structures

● Made possible because ModelCIF extension of PDBx/mmCIF facilitates interoperation with PDB data

● Motivation in integrating CSMs:
  ○ Ameliorates the paucity of experimental 3D structures in the PDB
  ○ Improves structural coverage of proteomes of interest
  ○ Extensive set of tools and services provided by RCSB.org can be used to search, analyze, and visualize CSMs alongside experimentally-determined PDB structures

https://www.rcsb.org/docs/general-help/computed-structure-models-and-rcsborg
CSMs in RCSB.org

- CSMs are consistently distinguished from experimental structures through source-specific icons and unique coloring schemes.
- Provenance information about the CSMs are clearly provided.

https://www.rcsb.org/docs/general-help/computed-structure-models-and-rcsborg
CSMs in RCSB.org: AlphaFold2 and RoseTTAFold

- Pre-packaged collection of 999,255 CSMs from AlphaFoldDB (v3)
  - Model organism proteomes: Proteomes from 48 different model organisms e.g., Arabidopsis, *E. coli*, fruit fly, human, soybean, and zebrafish
  - Global health proteomes: Proteomes of disease-causing organisms, e.g., *H. pylori*, *K. pneumoniae*, *M. tuberculosis*, and *P. falciparum*
  - Swiss-Prot sequences
  - MANE (Matched Annotation from NCBI and EMBL-EBI) Select sequences

- RoseTTAFold CSMs from the ModelArchive
  - Computed structures of core eukaryotic protein complexes produced by the Baker lab computed using a combination of RoseTTAFold and AlphaFold2
  - Set of 1,106 heteromeric complexes archived in the ModelArchive (https://modelarchive.org/doi/10.5452/ma-bak-cepc)

https://www.modelarchive.org/

Accessing CSMs on RCSB.org
Live Demo
Summary

- RCSB.org now provides access to >1,000,000 CSMs from AlphaFold2 and RoseTTAFold alongside ~200,000 experimental structures of macromolecules

- Various search, analysis and visualization tools on RCSB.org facilitate the study of CSMs together with the experimental structures to obtain additional structural and functional insights

- Depending upon the scientific question being investigated and the data available to address the question, different search and analysis strategies can be devised
RCSB PDB Team

RCSB.ORG
info@rcsb.org

Funding
National Science Foundation (DBI-1832184),
National Institute of General Medical Sciences,
National Institute of Allergy and Infectious Disease, and
National Cancer Institute (NIH R01GM133198), and the
US Department of Energy (DE-SC0019749)

Management

Member of the
Worldwide Protein Data Bank
(wwPDB; wwpdb.org)

Follow us

John D. Westbrook
in memoriam
1957-2021
Thank you!
Introduction to the Mol* Molecular Graphics System

Shuchismita Dutta, Ph.D., RCSB PDB/Rutgers
shuchi.dutta@rcsb.org
Overview

- Exploring a 3D structure from RCSB.org using Mol*
- Standalone Mol*
- Exploring and Comparing multiple 3D structures using Mol*
Overview

● Exploring a 3D structure from RCSB.org using Mol*
  ○ Overall structure
  ○ Focus vs Select
  ○ Display and Color
  ○ Measure
  ○ Explore Annotations (1D-3D viewer)

● Standalone Mol*

● Exploring and Comparing multiple 3D structures using Mol*
Overall Structure (Access from Structure Summary Page → 3D View)

- 3D-Canvas
- Sequence Panel
- Controls Panel
- Components
- Quality Assessment
- Assembly
- Export Options
Focus vs Select
Select to Display and Color

- Atom/Coarse Element
  - Residue
- Chain
- Entity
- Model
- Operator
- Structure/Shape
- Atom/Coarse Element Instances
- Residue Instances
- Chain Instances

Add/Union Selection
- All
  - Polymer/Carbohydrate Entities
  - Ligand/Non-standard Residue
  - Type
  - Structure Property
  - Bond Property
  - Residue Property
  - Manipulate Selection
  - Amino Acid
  - Nucleic Base
  - Validation
  - Element Symbol
  - Helpers

Add Component
- Selection
  - Current Selection
- Representation
  - < Create Later >
  - Cartoon
  - Backbone
  - Ball & Stick
  - Gaussian Surface
  - Gaussian Volume
  - Label
  - Line
  - Molecular Surface
  - Orientation
  - Point
  - Putty
  - Options

Theme
- Selection
  - Current Selection
- Action
  - Color
- Color
  - RGB
    - Lighten
    - Darken
  - Represents...
    - All

Apply Theme
Measure

1. Select atom(s) on 3D Canvas
2. Measurements → +Add
3. Select options to measure distance, angle, torsion
Explore Annotations (1D-3D View)
Overview

- Exploring a 3D structure from RCSB.org using Mol*
- Standalone Mol* ([https://www.rcsb.org/3d-view](https://www.rcsb.org/3d-view))
  - Upload your coordinates
  - Compare structures
- Exploring and Comparing multiple 3D structures using Mol*
Upload Coordinates
Compare Structures
Overview

- Exploring a 3D structure from RCSB.org using Mol*
- Standalone Mol*
- Exploring and Comparing multiple 3D structures using Mol*
  - View 1D-3D Alignments
View 1D-3D Alignments
Case Study: *Arabidopsis thaliana* Resistosome

- Which is the Disease resistance RPP13-like protein 4?
- What proteins does it interact with?
- What ligands does it interact with?
- Mutations @ 297 vs 359?
- How do other (experimental/CSM) structures of this protein compare with this structure

PDB 6J5T (Wang 2019)
Exploring: *Arabidopsis thaliana* Resistosome

- Which is the Disease resistance RPP13-like protein 4?
  - Structure Summary Page
- What proteins does it interact with?
  - 1D-3D viewer
- What ligands does it interact with?
  - 1D-3D viewer, Mol*
- Mutations @ 297 vs 359?
  - 1D-3D viewer
- How do other (experimental/CSM) structures of this protein compare with this structure?
  - View 1D-3D Alignments
  - Pairwise alignment tool

PDB 6J5T (Wang 2019)
Summary

- Exploring a 3D structure from RCSB.org using Mol*
  - Overall structure
  - Focus vs Select
  - Display and Color
  - Measure
  - Explore Annotations (1D-3D viewer)
- Standalone Mol*
  - Upload your coordinates
  - Compare structures
- Exploring and Comparing multiple 3D structures using Mol*
  - View 1D-3D Alignments
RCSB PDB Team

RCSB.ORG
info@rcsb.org

Funding
National Science Foundation (DBI-1832184),
National Institute of General Medical Sciences,
National Institute of Allergy and Infectious Disease, and
National Cancer Institute (NIH R01GM133198), and the
US Department of Energy (DE-SC0019749)

Management

Member of the
Worldwide Protein Data Bank
(wwPDB; wwpdb.org)

Follow us

John D. Westbrook
In memoriam
1957-2021
KBase Apps for Protein Structure Data Communication and Integration with RCSB PDB

(DOE/KBase/RCSB PDB) Virtual Crash Course, Nov.10, 2022

Qizhi Zhang, PhD, KBase/Argonne National Laboratory
Introducing 4 KBase Apps interfacing with RCSB PDB

- Import ProteinStructures from a Metadata File in Staging Area
- Query RCSB databases for protein structures
- PDB - Import PDB Metadata into KBase Genome
- Import RCSB Structures (to associate with KBase genomic data objects)

Demo Narrative: https://narrative.kbase.us/narrative/130799
Purpose: With your own protein structure files, you can link them with KBase genome data that is in your narratives or shared with you. Given certain sequence similarity threshold, annotation data of structures and genes that satisfy the threshold will be uploaded (saved) to KBase as an object of the type of ProteinStructures.
Purpose: By specifying query filters, you can search the RCSB database to find structure hits that meet the criteria you give.
**PDB - Import PDB Metadata into KBase Genome**

Queries PDB API with genome proteins and annotates proteins with associated PDB metadata

Finished with success at Nov 8, 2022 at 5:41pm

<table>
<thead>
<tr>
<th>Input Objects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome</td>
<td>Athaliana_TAIR10_2012</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Suffix for annotated genomes</td>
<td>.pdb</td>
</tr>
<tr>
<td>Similarity threshold type</td>
<td>E-value</td>
</tr>
<tr>
<td>Similarity threshold</td>
<td></td>
</tr>
</tbody>
</table>
**Purpose:** Of the RCSB query structure hits, after examining the result metadata, the user can choose a subset of those structures (in RCSB IDs) and associate them with KBase genome data by using this app.
Demo in a KBase Narrative:

https://narrative.kbase.us/narrative/130799
Todo next…

1. Add viewer for the KBase ProteinStructures datatype
2. More RCSB PDB query features and tools (e.g., Docking/Alignment)
3. Better differentiation of experimental from computational structures (source-specific icon, provenance info, RCSB ids)
4. Search results into sets of KBase data objects
5. Other Suggestions?
Review of PDB-KBase Workflows

1. Using gene function discovery pipelines to identify candidate genes for functions of interest
   - Candidates could be proposed from gapfilling, pathway prediction, annotation apps, alignments, blast and many other algorithms in KBase

2. Pulling closest experimental structures associated with candidate genes from PDB

3. Exporting candidate genes as FASTA, generating structures in google collab, and importing computational structures back into KBase or query and analyze structures in PDB

4. Using PDB query and import app to also pull structures with desired functions or substrates cocrystalized

5. Aligning computational and experimental structures in the PDB site

6. Comparing PDB complexes and cocrystalized ligands with complexes and reaction substrates in models

7. Applying Import PDB Metadata app to determine extent of structure coverage of genome, identify gaps in structure coverage, test for existing and completeness of protein complexes, import PDB/uniprot data on annotation and co-crystalization, and identify structures of interest for deeper analysis.
Future plans for KBase PDB apps

1. Mol* viewer in KBase
   ○ Currently only available on upload - adding a widget to view any time
   ○ Enabling painting of other data in KBase on structure views (e.g. sequence alignments, variation, domains, docking, annotations, fold quality)
   ○ Integrating more mol* functionality into KBase

2. PDB Query app
   ○ Enabling query by feature ID in KBase objects (genome, feature set, metagenome) rather than copying and pasting sequence
   ○ Saving query output in KBase with annotations as protein sequence set
   ○ Enabling query by structure alignment

3. Structure import
   ○ Expanding import to include computational structures from PDB and other databases

4. Docking
   ○ Linking KBase autodock-VINA app to new structure datatype and enabling saving of docked poses for future viewing
Future plans for New PDB apps

1. Links to modeling
   a. Comparing all model complexes to complex data in structures
   b. Comparing reaction substrate to cocrystalized ligands and annotated models with co-enzyme data (e.g. PLP)

2. Structural alignment
   a. Mitchell and Sedova team at ORNL are adding tools for structure alignment

3. Experimental/computational structures
   a. Enabling automated linking of imported computational structures to closest experimental structures and facilitating rapid alignment of these pairs

4. Adding one or more computational structure prediction apps in KBase

5. Other ideas?
Acknowledgements

We thanks DOE BER for funding this work!

Special thanks to all those who attended and contributed to our KBase-PDB design workshops!