KBase
PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Review of Crash Course Objectives and Arabidopsis Case Study

Christopher S. Henry

November 10th, 2022

INTEGRATION and MODELING *for* PREDICTIVE BIOLOGY

RUTGERS

Argonne NATIONAL LABORATORY

RCSB PDB PROTEIN DATA BANK

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

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

U.S. DEPARTMENT OF ENERGY
Office of Science

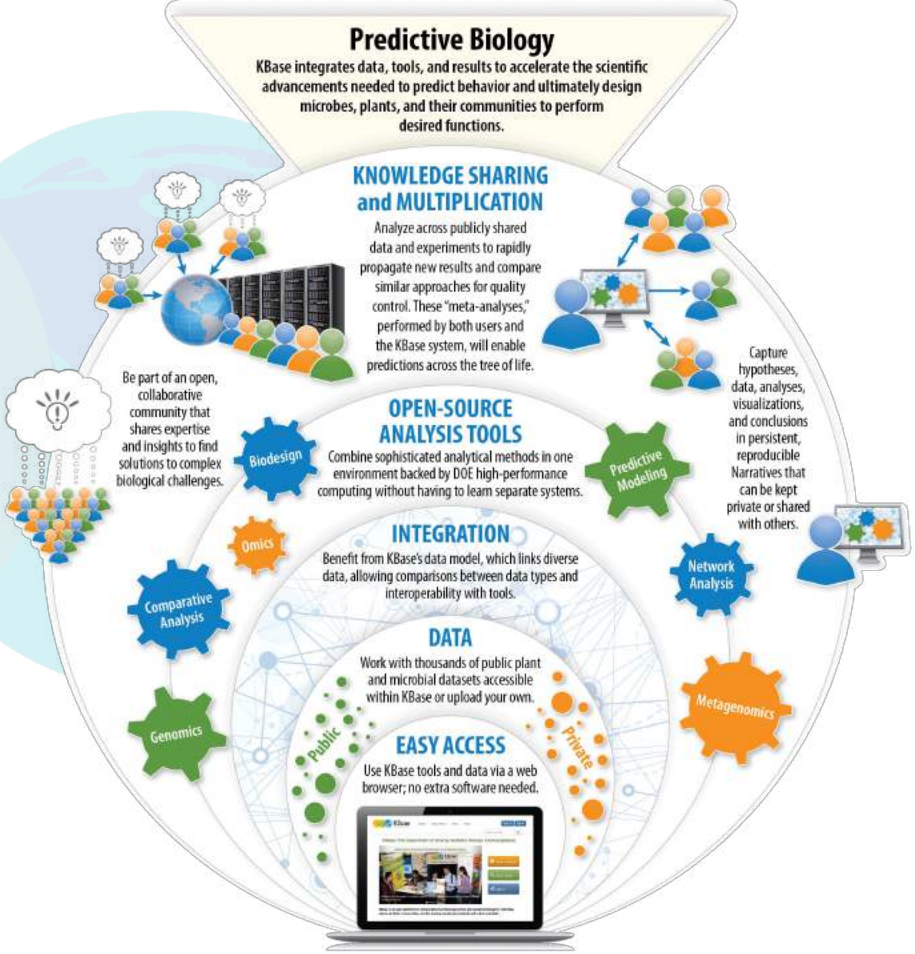
KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

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



U.S. DEPARTMENT OF ENERGY

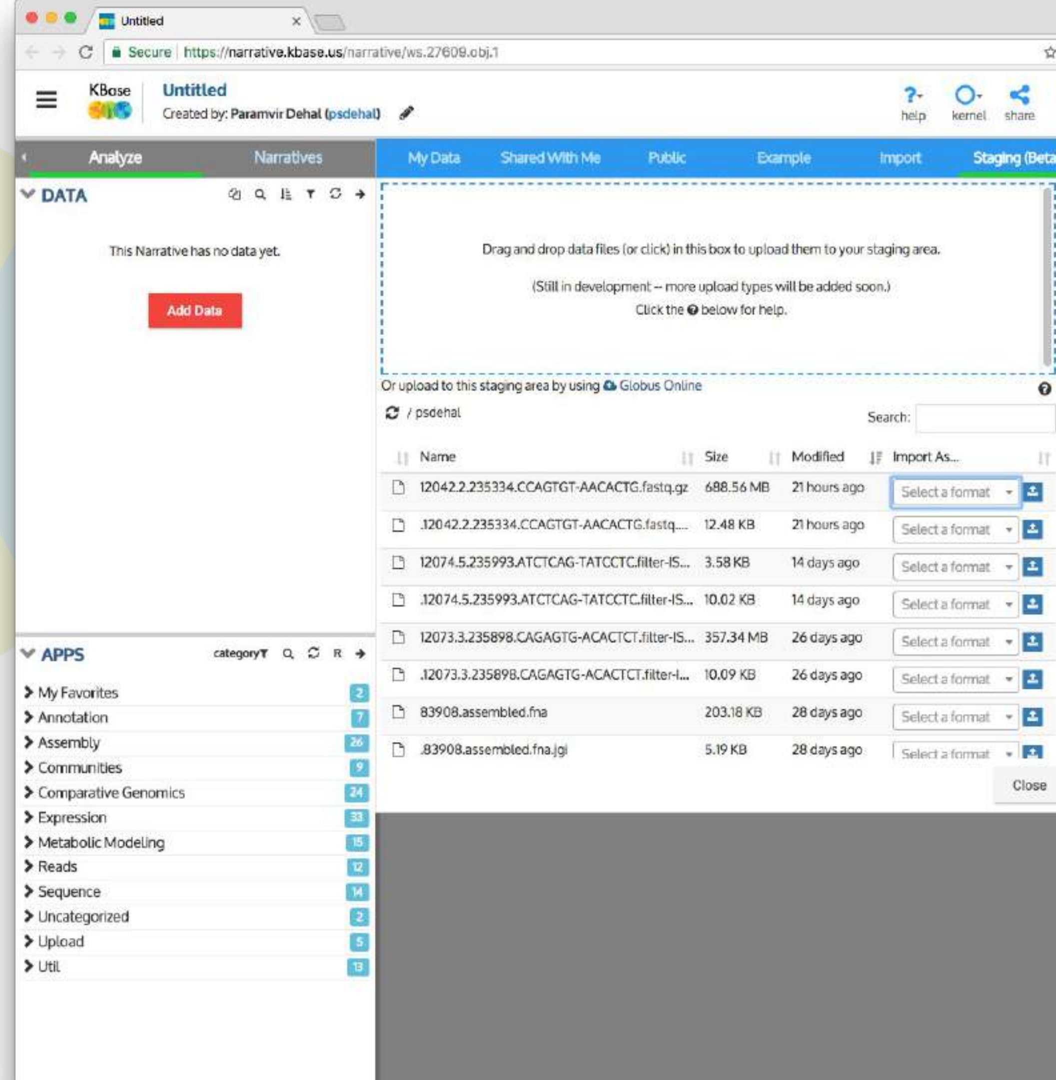
Office of Science

KBase PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

What is KBase?
Data Management

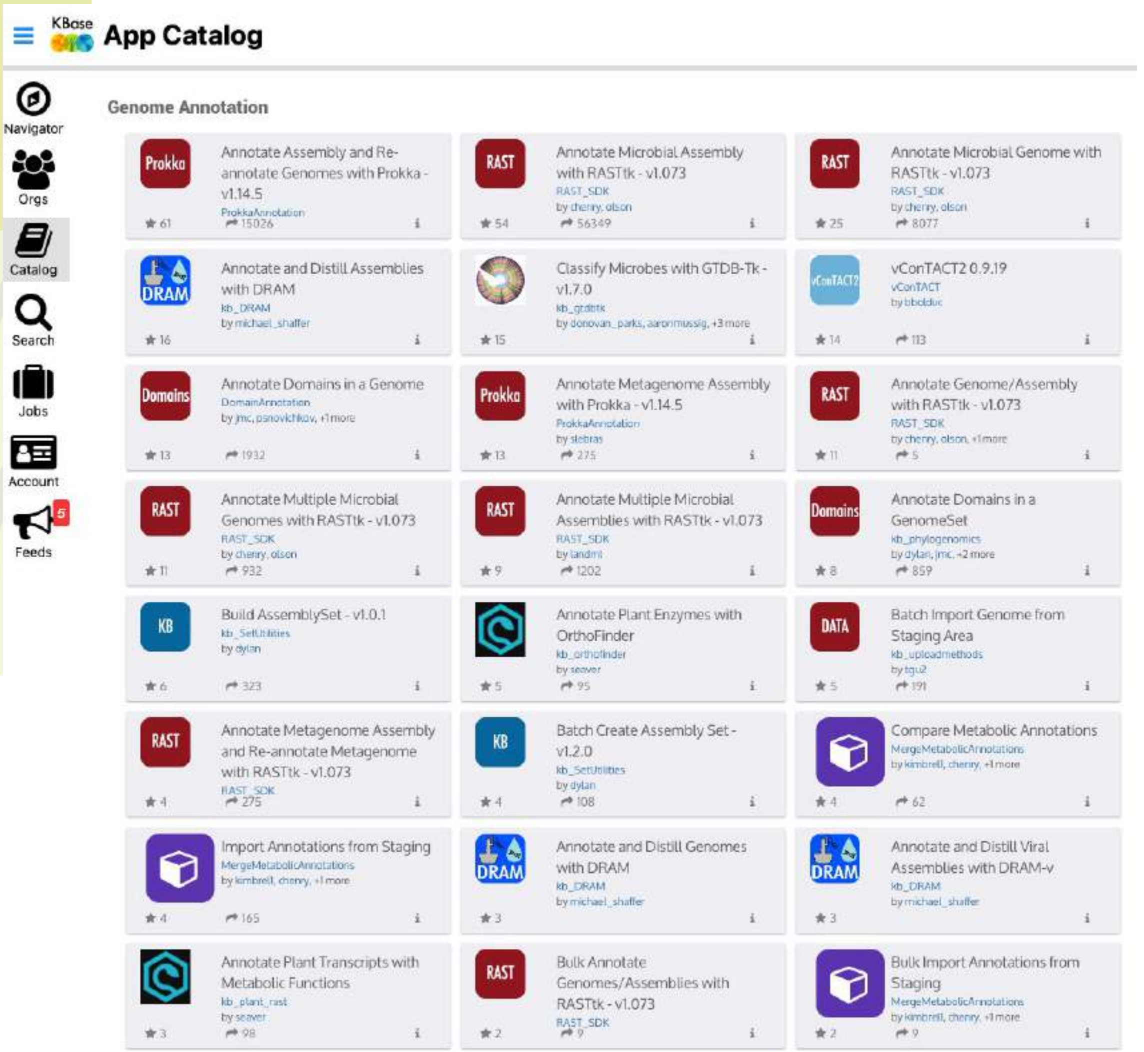
- Unlimited data storage
- Import data by
  - Drag & drop
  - Globus
  - FTP, HTTP, Gdrive, Box/Dropbox
- File types
  - FASTQ, FASTA, SRA, GenBank, gff, expression matrix, media, phenotype, FBA models



What is KBase?

Data Analysis

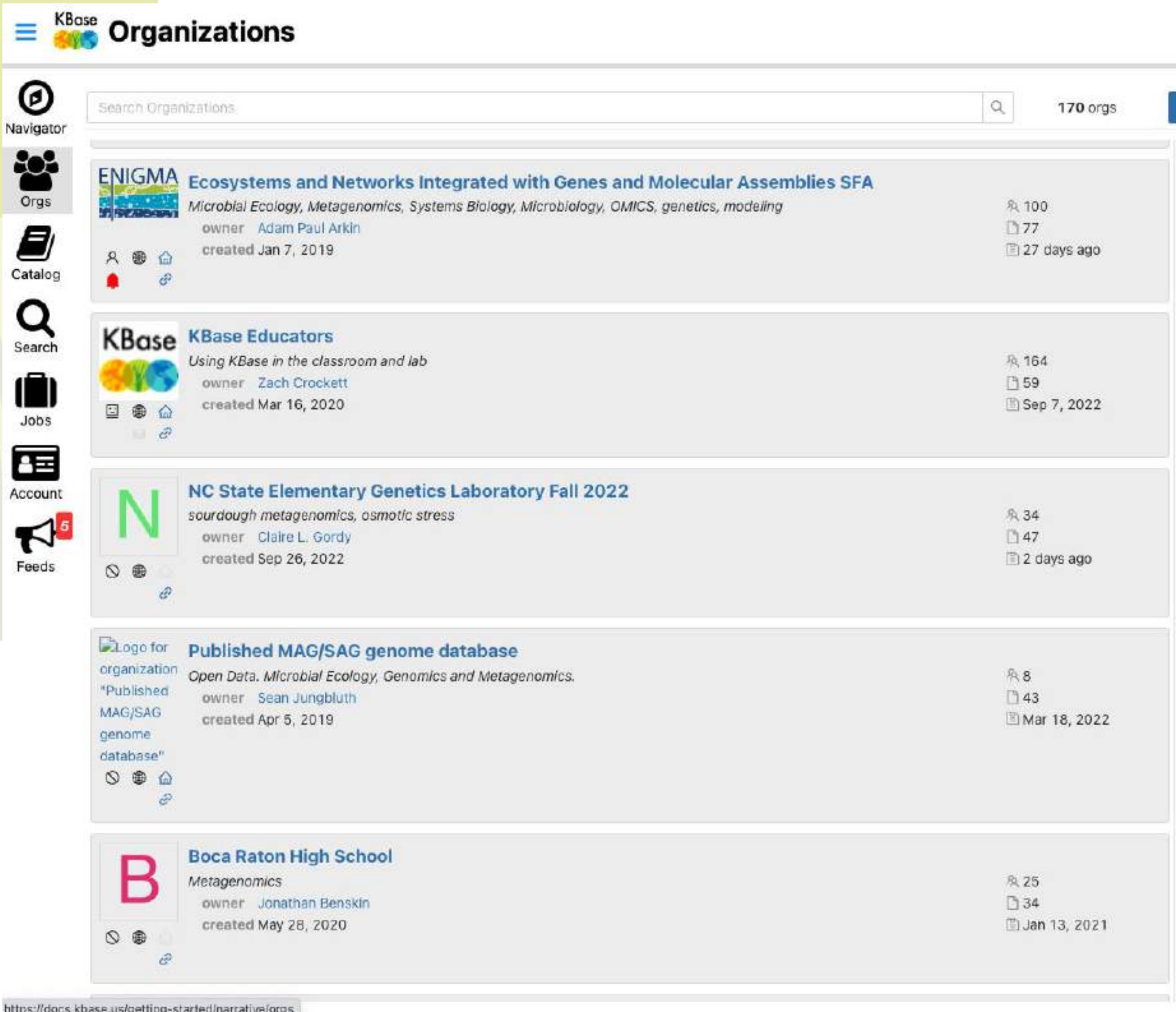
- App Catalog
  - ~280 apps
  - Read processing
  - Genome assembly
  - Genome annotation
  - Sequence alignment
  - Comparative genomics
  - Metabolic modeling
  - Expression
  - Microbial communities



What is KBase? Collaboration

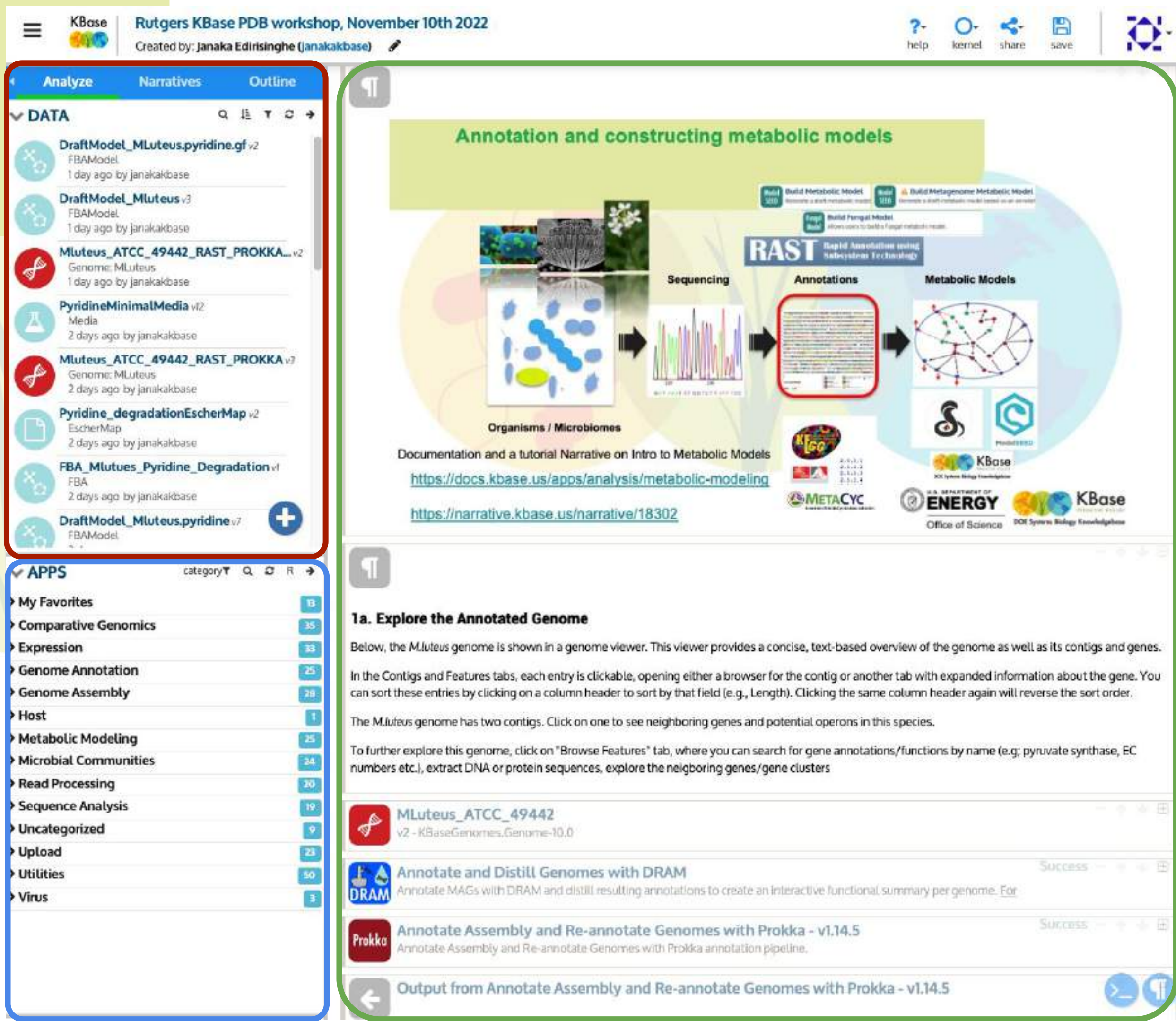
- As users conduct their work within notebooks called “narratives”, they can share those narratives with other users

- 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

EMSL Summer School 2021

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

Course Page →

Searching For Features

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

View Tutorial →

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 shows how to draft isolate genomes in...

View Tutorial →

Multiscale Microbial Dynamics Modeling

EMSL Summer School 2020

Learn how to incorporate microbial metagenomic and environmental metabolite data from watershed ecosystems into metabolic and community modeling using computational frameworks in this virtual course.

KBase For Educators

Made by Ellen Dow

Join the community of educators using KBase to teach bioinformatics and computational biology!

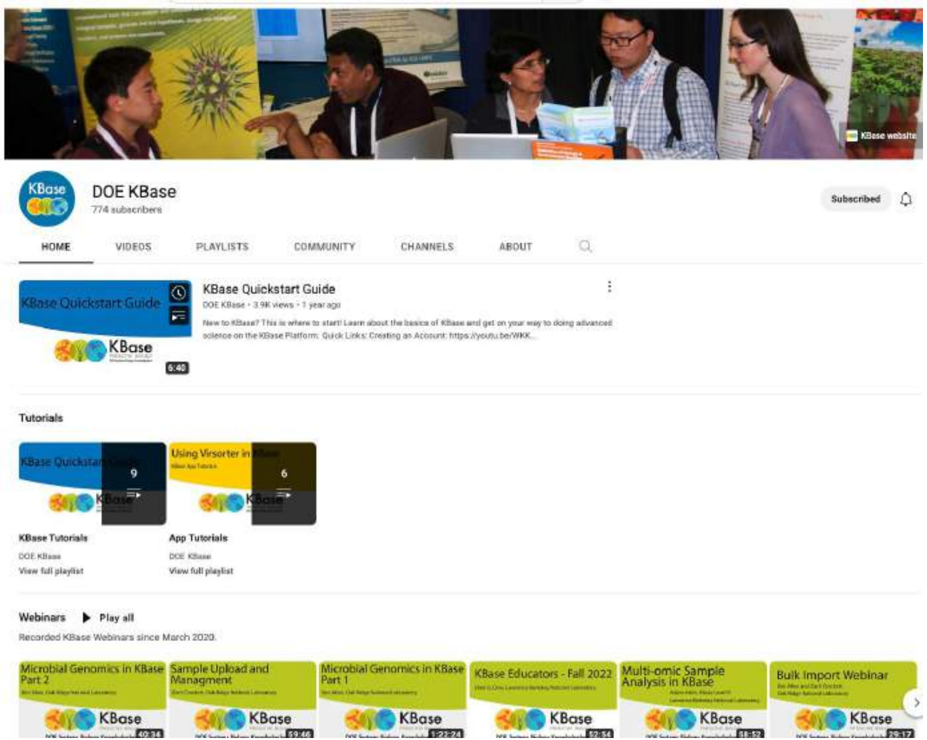
Why KBase?

While KBase is routinely used for advanced analysis by scientists around

Assembly And Annotate Prokaryotic Genome Tutorial

Made 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

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Why KBase and PDB?

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

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Case study in Arabidopsis…

Acknowledgements


Stephen Burley


John Westbrook


Ed O’loughlin


Qizhi Zhang


Kelly Skinner


Jack Gilbert


Sam Seaver


Claudia Lerma-Ortiz


Nidhi Gupta

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!

KBase

PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Protein Candidates from Function Queries in KBase

Janaka N. Edirisinghe, Ph.D

November 10th, 2022

INTEGRATION and MODELING *for* PREDICTIVE BIOLOGY

RUTGERS

Argonne NATIONAL LABORATORY

RCSB PDB PROTEIN DATA BANK

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

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?

Science example: Discover potential novel degradation pathway genes for heteroaromatic compound Pyridine in *Micrococcus letus* using KBase Functional Discovery Pipeline

- 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*

U.S. DEPARTMENT OF ENERGY Office of Science

KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase

MaSuRCA Assembler - v3.2.9
kb_MaSuRCA v1.1.3
4719 genes

Annotate Genome/Assembly with RASTtk - v1.073
RAST_SDK v1.9.4
4719 genes

Build Metabolic Model
fba_tools v2.0.0
1127 genes
1135 reactions

PickAxe - Generate novel compounds from reaction rules
kb_pickaxe v1.3.1
7320 novel reactions

Gapfill Metabolic Model
fba_tools v2.0.0
Fill gaps with novel and known compounds

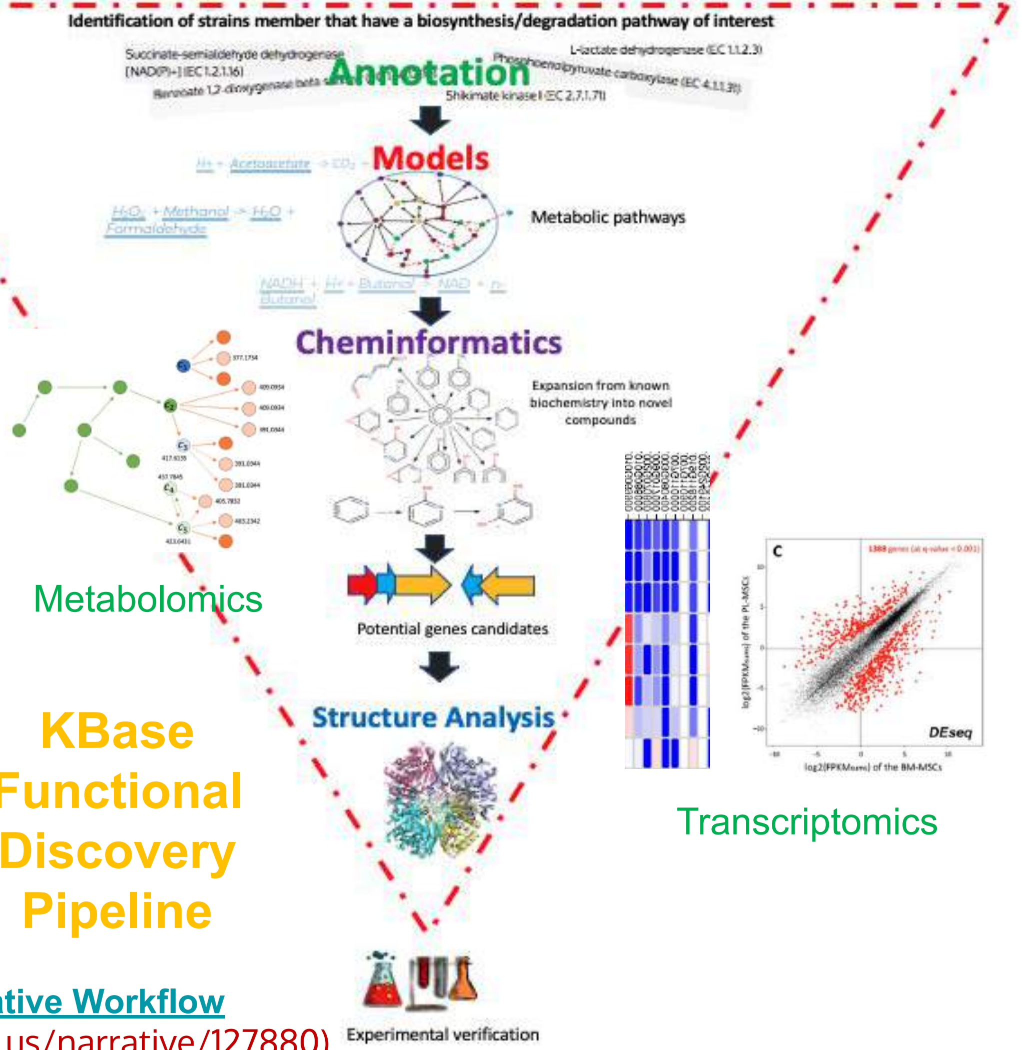
Escher Pathway Viewer
kb_escher v0.0.1
View key reactions and pathways

Create Differential Expression Matrix using DESeq2 - v1.20.0
kb_deseq v1.1.2
> 20 gene candidates (~15)

PDB - Import PDB Metadata into KBase Genome
KBAnnotationApps v0.0.1

Query RCSB databases for protein structures
ProteinStructureUtils v0.0.2
Query candidate proteins from RCSB
few genes

KBase Functional Discovery Pipeline



Link: [Base Narrative Workflow](https://narrative.kbase.us/narrative/127880)
(https://narrative.kbase.us/narrative/127880)

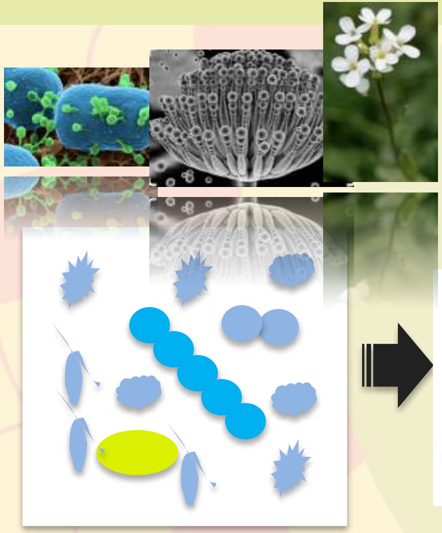
Annotation of genomes and constructing metabolic models

Build Metabolic Model
Generate a draft metabolic model

Build Metagenome Metabolic Model
Generate a draft metabolic model based on an annotat

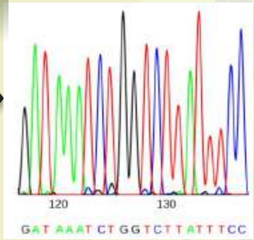
Build Fungal Model
Allows users to build a Fungal metabolic model.

RAST Rapid Annotation using Subsystem Technology



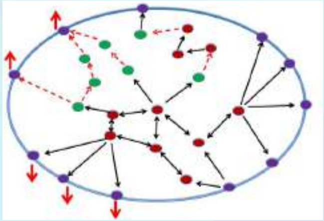
Organisms / Microbiomes

Sequencing



Annotations

Metabolic Models



2.5.1.1
2.5.1.2
2.5.1.3
2.5.1.4

METACYC A member of the BioCyc database collection

KBase DOE Systems Biology Knowledgebase

U.S. DEPARTMENT OF ENERGY Office of Science

KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase

Documentation and a tutorial Narrative on Intro to Metabolic Models

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

https://narrative.kbase.us/narrative/18302

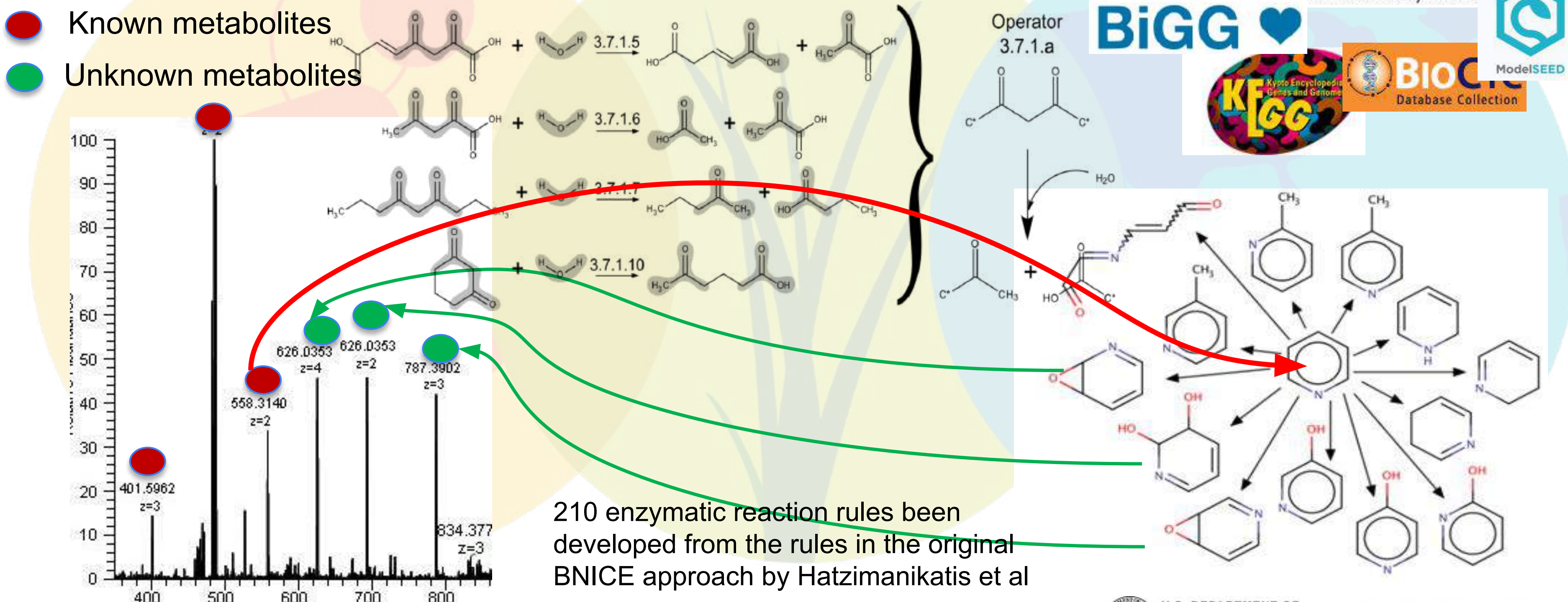
Introducing new biochemistry in filling metabolic gaps

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

Biochemistry Databases

Known metabolites

Unknown metabolites



210 enzymatic reaction rules been developed from the rules in the original BNICE approach by Hatzimanikatis et al

PickAxe - Generate novel compounds from reaction rules

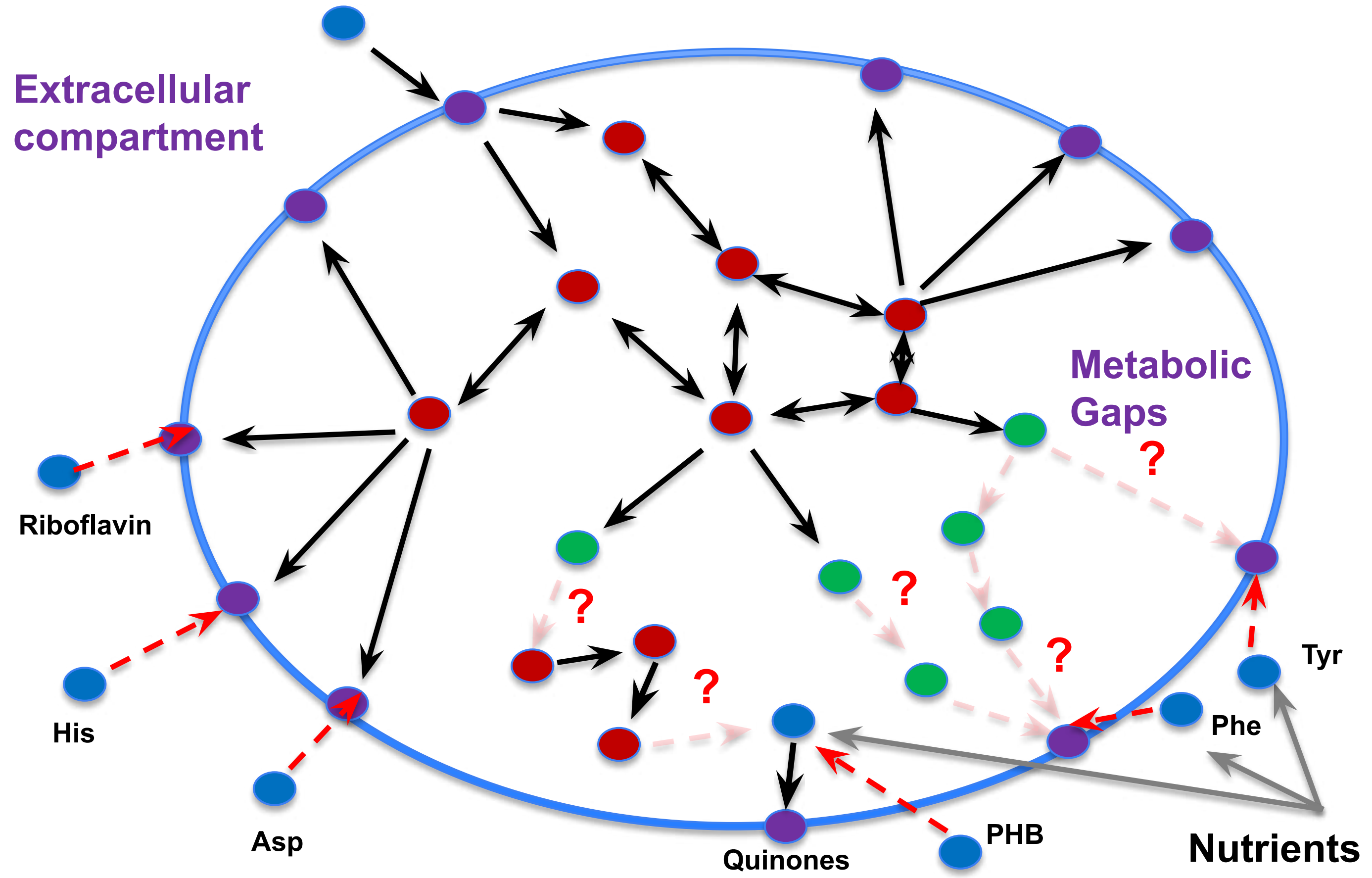
Generate novel compounds based enzymatic and spontanios reaction rules

Filling gaps in a metabolic network

Model SEED

Gapfill Metabolic Model

Identify the minimal set of biochen



Genome

Draft Model

Gapfill on Pyridine media (No growth)

Cheminformatics (Introduce Pyridine degradation reactions)

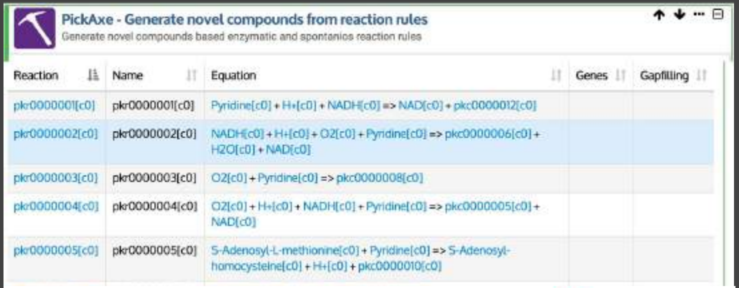
Gapfill on Pyridine media (able to grow)

Model Simulations - FBA

Discovering genes involved in pyridine degradation

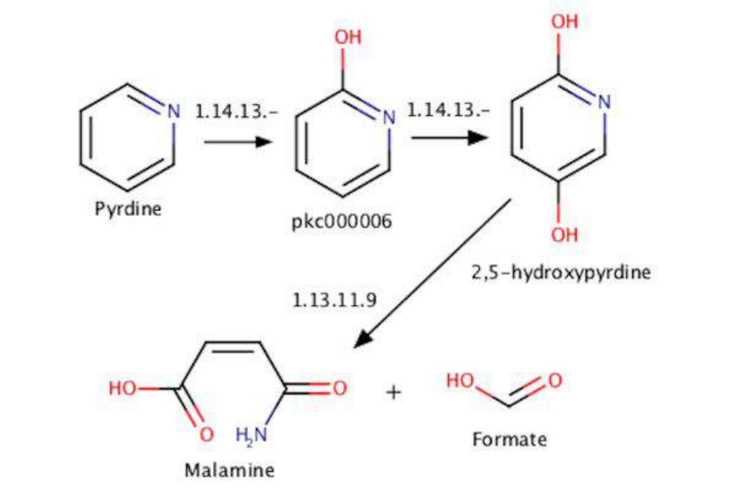
Using cheminformatics to predict a novel pathway

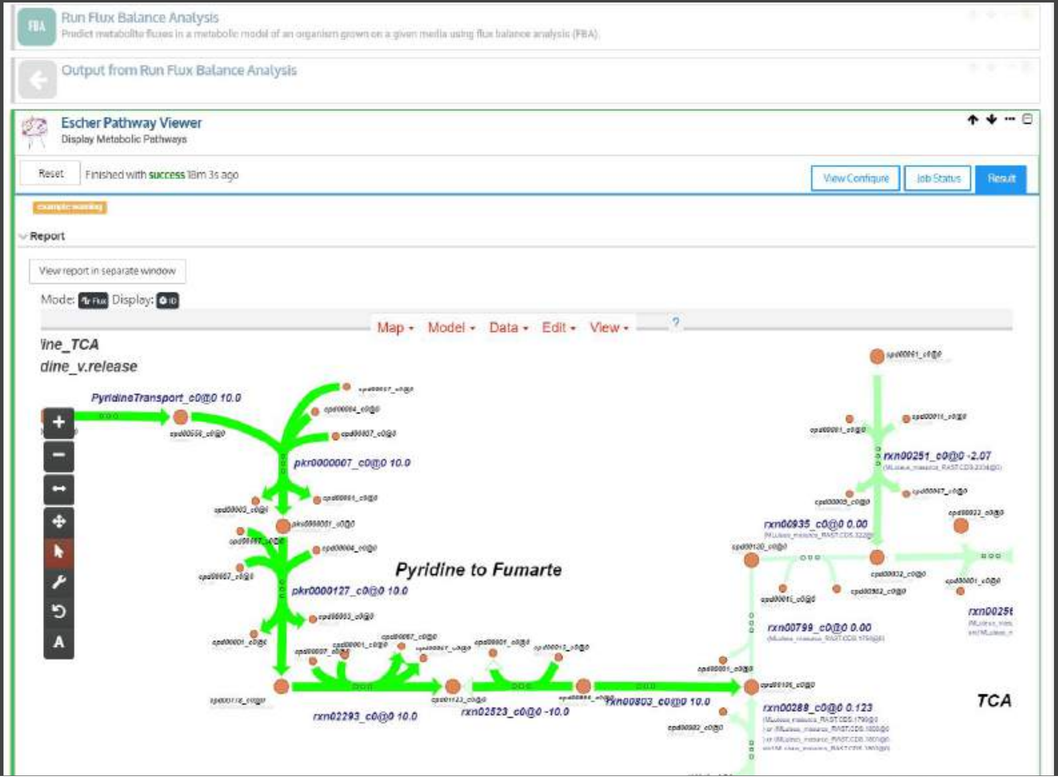
Metabolic modeling workflow in KBase is used to predict the most likely pathway, then Escher is used to visualize with flux



PickAxe - Generate novel compounds from reaction rules
Generate novel compounds based enzymatic and spontanios reaction rules

Reaction	Name	Equation	Genes	Gapfilling
pkr0000001[c0]	pkr0000001[c0]	Pyridine[c0] + H+[c0] + NADH[c0] => NAD[c0] + pkc0000012[c0]		
pkr0000002[c0]	pkr0000002[c0]	NADH[c0] + H+[c0] + O2[c0] + Pyridine[c0] => pkc0000006[c0] + H2O[c0] + NAD[c0]		
pkr0000003[c0]	pkr0000003[c0]	O2[c0] + Pyridine[c0] => pkc0000008[c0]		
pkr0000004[c0]	pkr0000004[c0]	O2[c0] + H+[c0] + NADH[c0] + Pyridine[c0] => pkc0000005[c0] + NAD[c0]		
pkr0000005[c0]	pkr0000005[c0]	S-Adenosyl-L-methionine[c0] + Pyridine[c0] => S-Adenosyl-homocysteine[c0] + H+[c0] + pkc0000010[c0]		

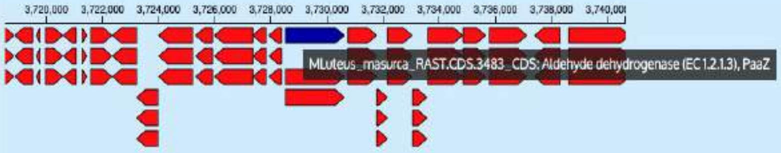




Linking cheminformatics to genes in the genome

Feature ID	Type	Function	Ontology
MLuteus_masurca_RAST.CDS.3484	gene	1,2-phenylacetyl-CoA epoxidase, subunit A	ec:1.14.13.149- Phenylacetyl-CoA 1,2-epoxidase.

Product Function	Aldehyde dehydrogenase (EC 1.2.1.3), PaaZ
Function Descriptions	
Ontology Terms	None
Location	Contig: scf7180000000004 3,728,527 - 3,730,639 (+ Strand)
Feature Context	



PickAxe - Generate novel compounds from reaction rules

Generate novel compounds based enzymatic and spontanios reaction rules

PyridineNovelReactions

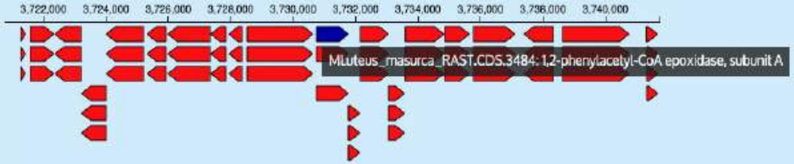
v2 - KBaseFBA.FBAModel-14.0

Overview | Reactions | Compounds | Genes | Compartments | Biomass | Gapfilling | Pathways | enzc274_c0

Show 10 entries

Reaction	Name	Equation
enzr106[c0]	1.14.13_10_14	2-Pyridone[c0] + O2[c0] + NADH[c0] + H+[c0] => dihydroxypyridine[c0] + H2O[c0] + NAD[c0]
enzr10[c0]	2.1.1_02_3	Pyridine[c0] + S-Adenosyl-L-methionine[c0] => e S-Adenosyl-homocysteine[c0] + H+[c0]
enzr1148[c0]	1.14.13_10_136	2,5-Dihydroxypyridine[c0] + O2[c0] + NADH[c0] => enzc15800[c0] + H2O[c0] + NAD[c0]
enzr118[c0]	1.3.1_03;1.3.1_05_1	enzc3[c0] + NAD[c0] => 2,3-dihydroxypyridine[c NADH[c0] + H+[c0]
enzr11[c0]	1.14.13_08_1	Pyridine[c0] + O2[c0] + NADH[c0] + H+[c0] => er H2O[c0] + NAD[c0]
enzr12[c0]	1.14.13_08_2	Pyridine[c0] + O2[c0] + NADH[c0] + H+[c0] => er H2O[c0] + NAD[c0]
enzr13[c0]	1.14.13_10_3	Pyridine[c0] + O2[c0] + NADH[c0] + H+[c0] => enzc7[c0] + H2O[c0] + NAD[c0]
enzr14[c0]	1.3.1_05.rev_3	Pyridine[c0] + NADH[c0] + H+[c0] <=> enzc10[c0] + NAD[c0]

Ontology Terms	ec:1.14.13.149- Phenylacetyl-CoA 1,2-epoxidase.
Location	Contig: scf7180000000004 3,730,746 - 3,731,772 (+ Strand)
Feature Context	

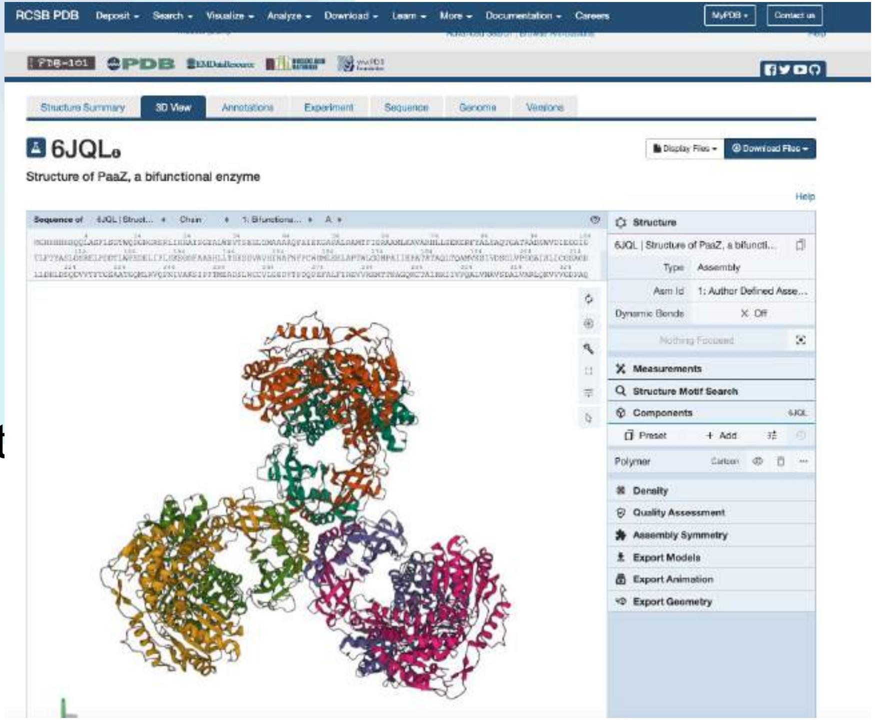


U.S. DEPARTMENT OF ENERGY Office of Science

KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase

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 substrat
- Align experimental and computational structures to aid binding site identification and functional characterization



U.S. DEPARTMENT OF ENERGY
Office of Science

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Acknowledgements


Stephen Burley


John Westbrook


Ed O'loughlin


Qizhi Zhang


Kelly Skinner


Jack Gilbert


Sam Seaver


Claudia Lerma-Ortiz

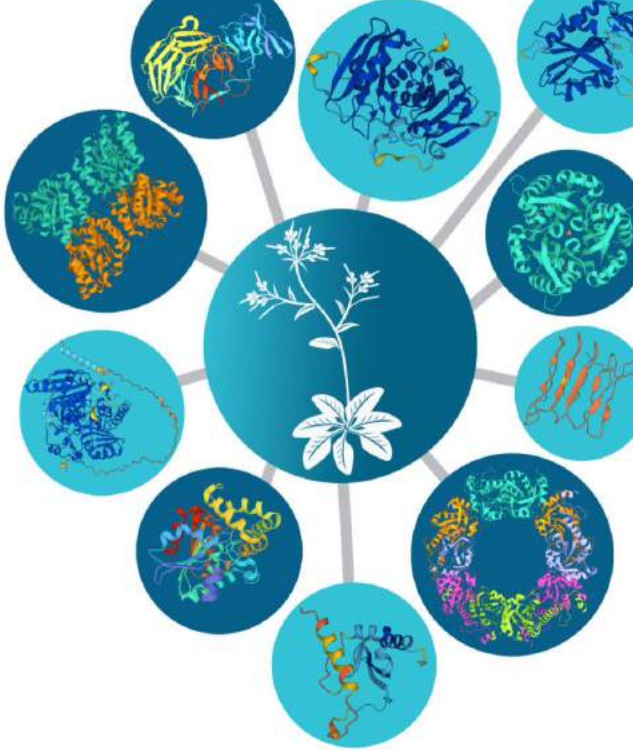

Nidhi Gupta

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!

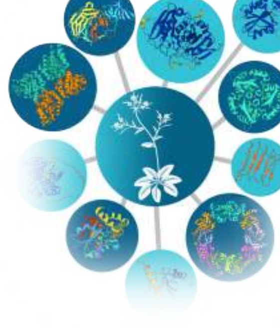
rcsb.org



Accessing Experimental Structures from the PDB

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

Outline

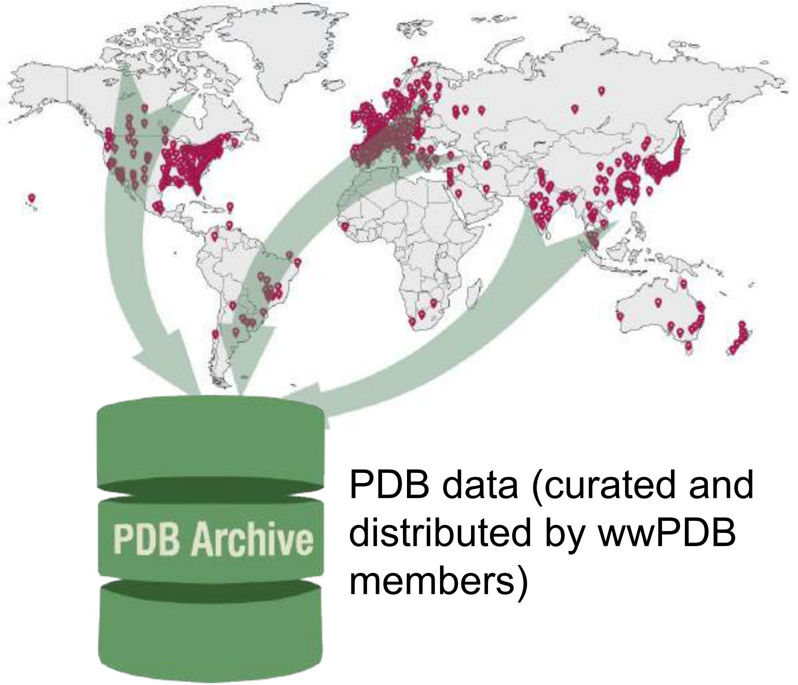


- Introduction to the PDB Archive and RCSB PDB
- Overview of search tools
- Overview of download services and structure data format
- Tutorial: Accessing data from RCSB.org
    - Case study: *Arabidopsis thaliana* resistosome
    - Searching and browsing
    - Structure summary pages
    - Downloading 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

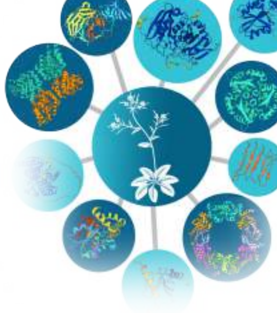
3D structural data from around the world



PDB data (curated and distributed by wwPDB members)

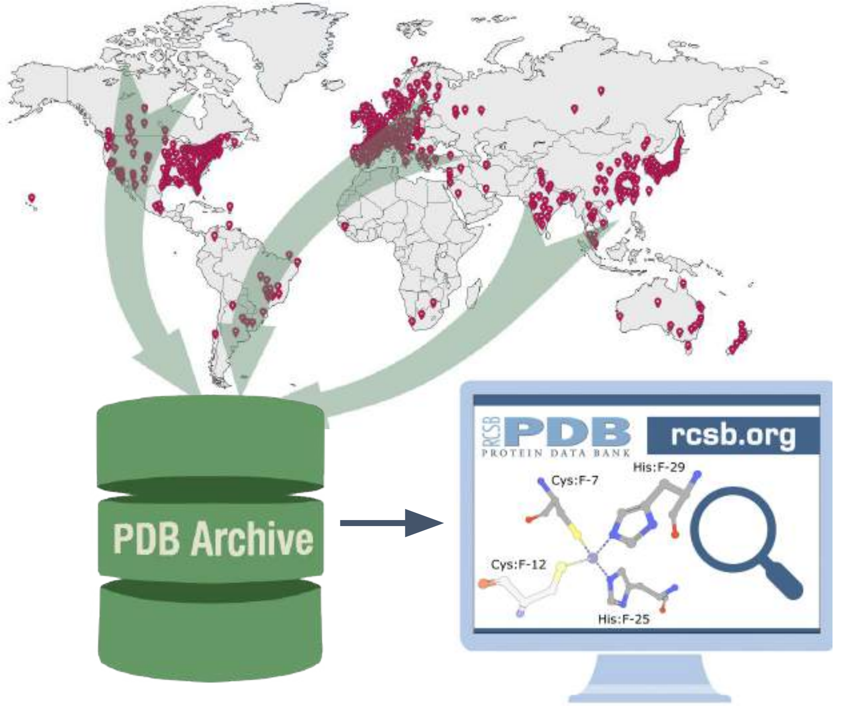
Protein Data Bank (1971) *Nature New Biology* **233**, 223.
Worldwide Protein Data Bank (2019) *Nucleic Acids Research* **47**, D520–D528.

The RCSB PDB Web Portal (RCSB.org)



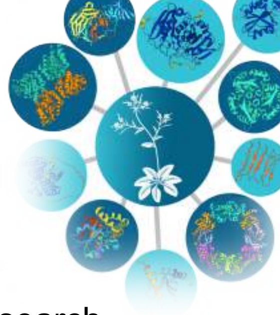
- **RCSB.org:** Tools for searching, visualizing, analyzing and downloading the contents of the PDB Archive
- Integration of PDB Archive data with annotations from ~50 external data resources (UniProt, SCOPe, CATH, …)
- Recently introduced tools and features:
  - Protein 1D-3D Feature View
  - Groups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

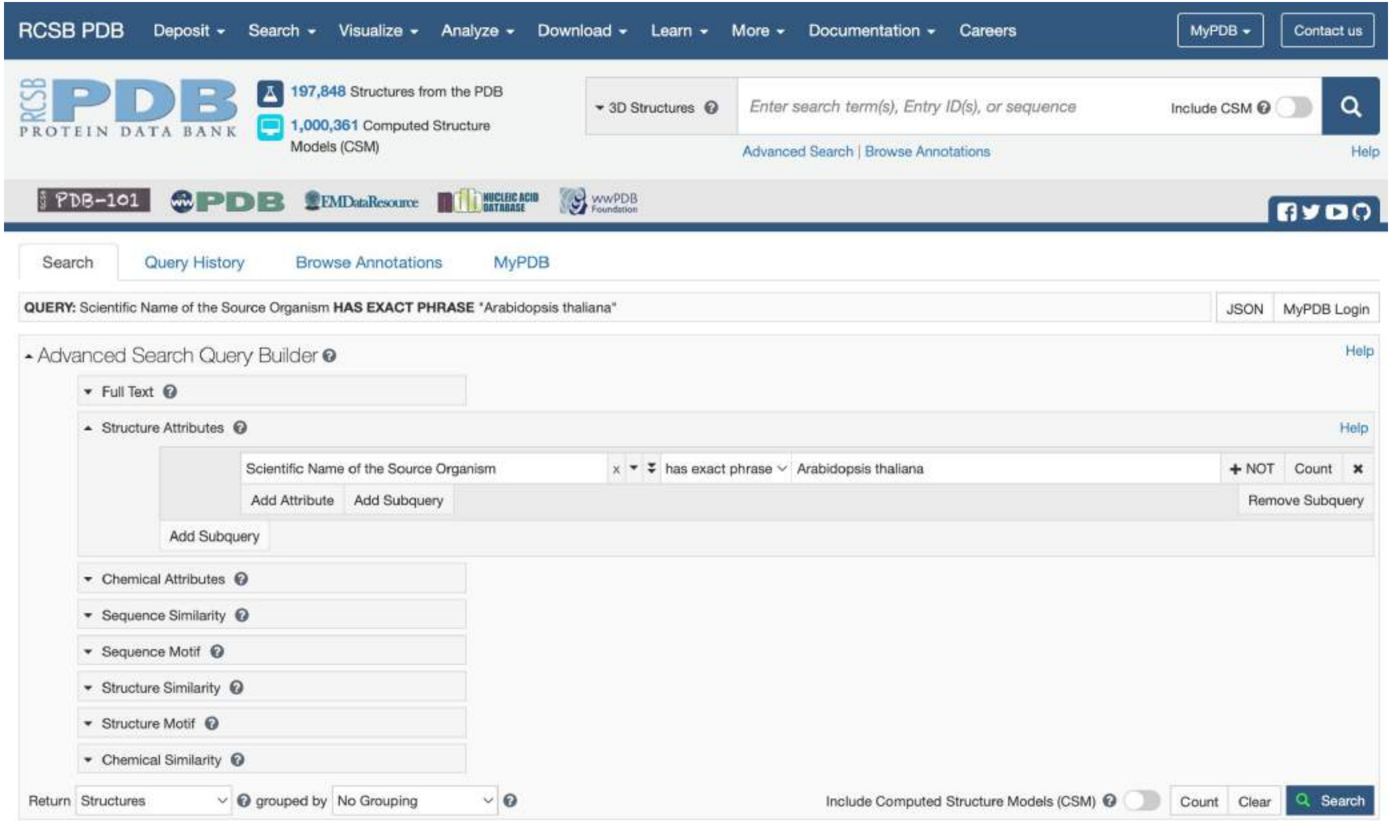
3D structural data from around the world



RCSB Protein Data Bank (2022) *Protein Science* **31**, 187–208.

Search Tools at RCSB.org

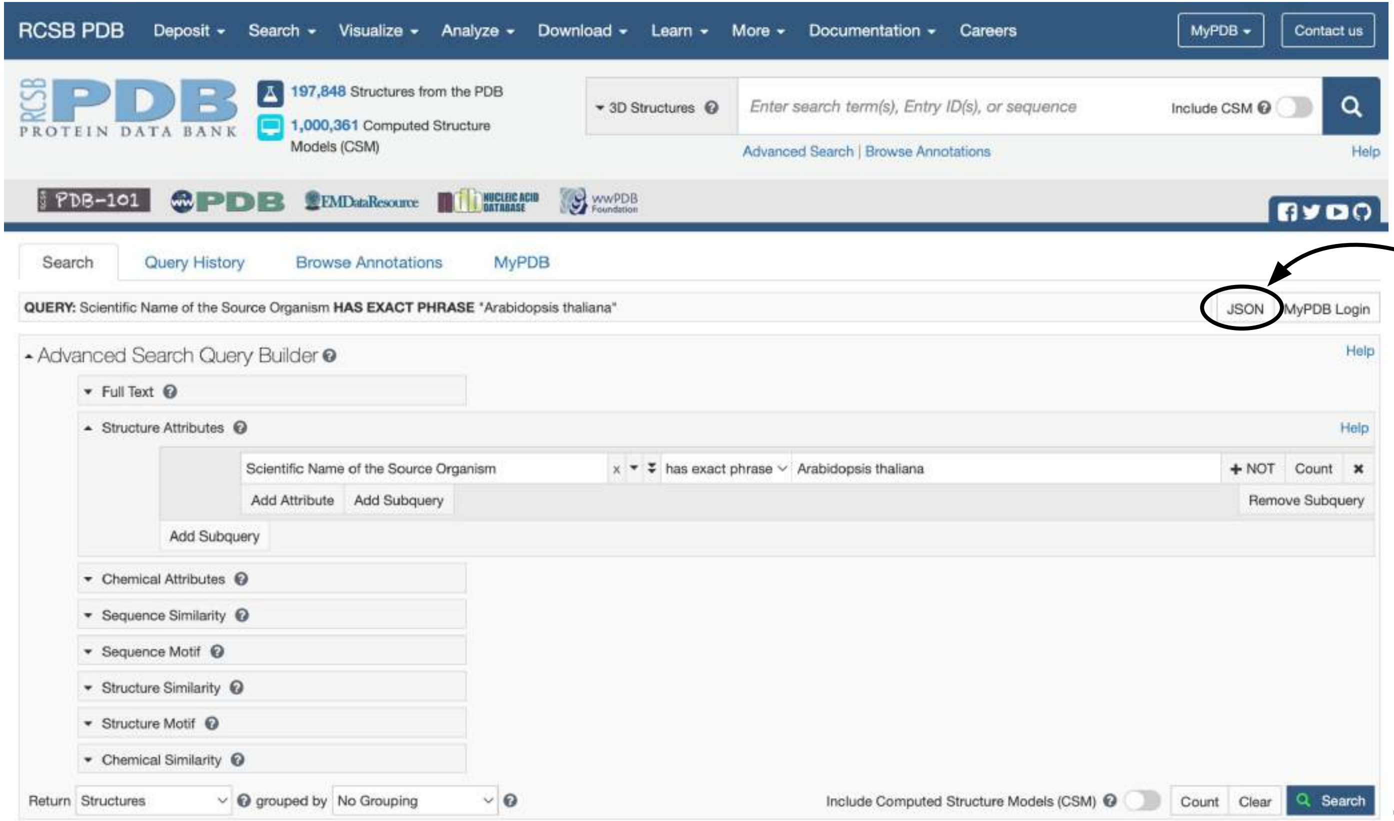




Basic search

Advanced search

Search Tools at RCSB.org

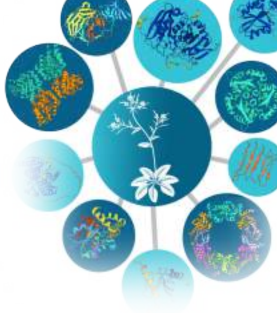


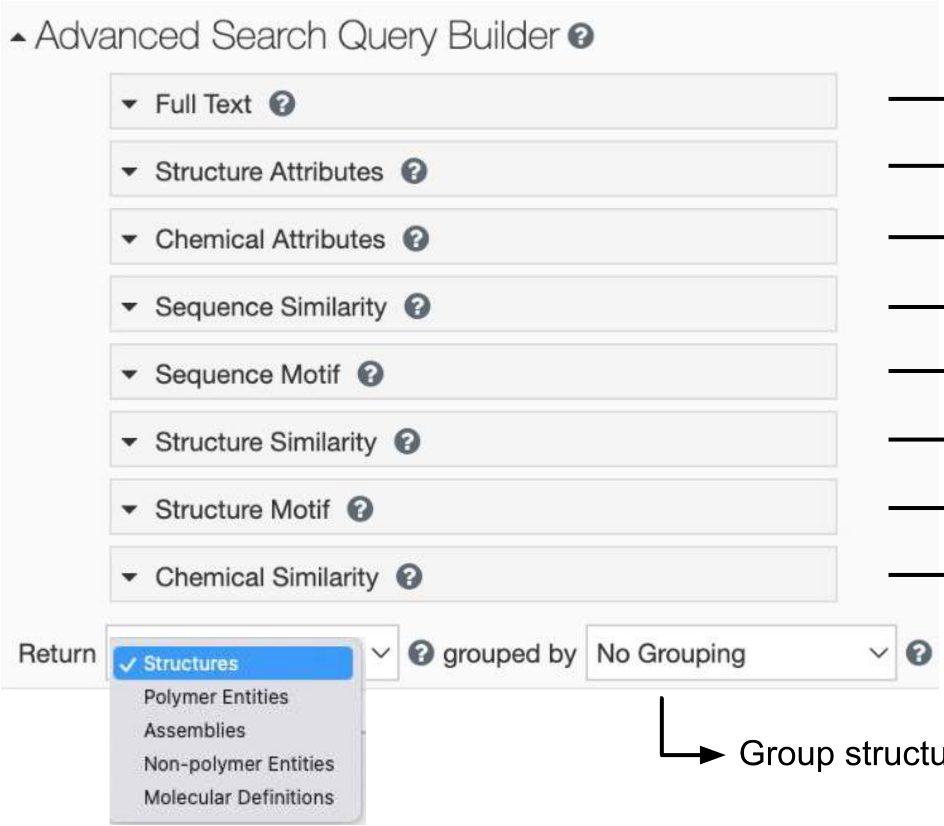
Basic search

Search API (search.rcsb.org)

Advanced search

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)

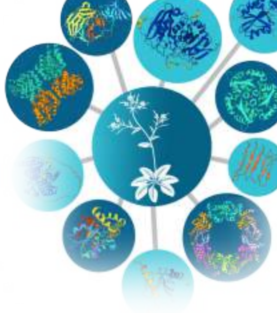
→ Global structure similarity (in-house algorithm)

→ 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

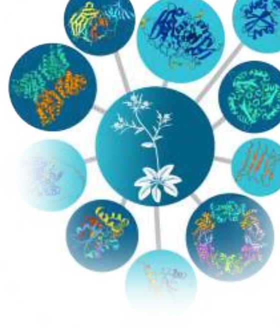


search.rcsb.org/query-editor.html?json=%7B"query"%3A%7B"type"%3A"terminal"%2C"label"%3A"text"%2C"service"%3A"text"%2C"parameters"...

RCSB PDB: Search API Query Editor


```
{
  "query": {
    "type": "terminal",
    "label": "text",
    "service": "text",
    "parameters": {
      "attribute": "rcsb_entity_source_organism.ncbi_scientific_name",
      "operator": "contains_phrase",
      "negation": false,
      "value": "Arabidopsis thaliana"
    }
  },
  "return_type": "entry",
  "request_options": {
    "paginate": {
      "start": 0,
      "rows": 25
    },
    "results_content_type": [
      "experimental"
    ],
    "sort": [
      {
        "sort_by": "score",
        "direction": "desc"
      }
    ],
    "scoring_strategy": "combined"
  }
}
```

```
{
  "query_id": "330fcd0f-62c0-4b9e-a710-fbe40c9c93d7",
  "result_type": "entry",
  "total_count": 1879,
  "result_set": [
    {
      "identifier": "7AR8",
      "score": 1
    },
    {
      "identifier": "7ARB",
      "score": 1
    },
    {
      "identifier": "6XYW",
      "score": 0.93261934140986
    },
    {
      "identifier": "7AR7",
      "score": 0.843479411762288
    },
    {
      "identifier": "7WG5",
      "score": 0.7614635080689911
    },
    {
      "identifier": "7OUI",
      "score": 0.7074923938322015
    },
    {
      "identifier": "7AQQ",
```

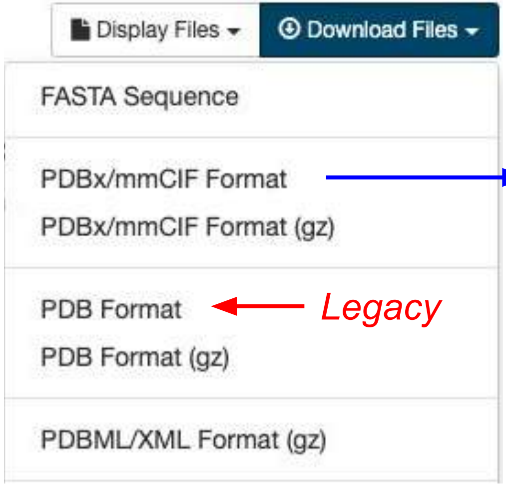


Download Services

- 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
  - Using FTP: ftp://ftp.wwpdb.org

Structure Data Format: PDBx/mmCIF

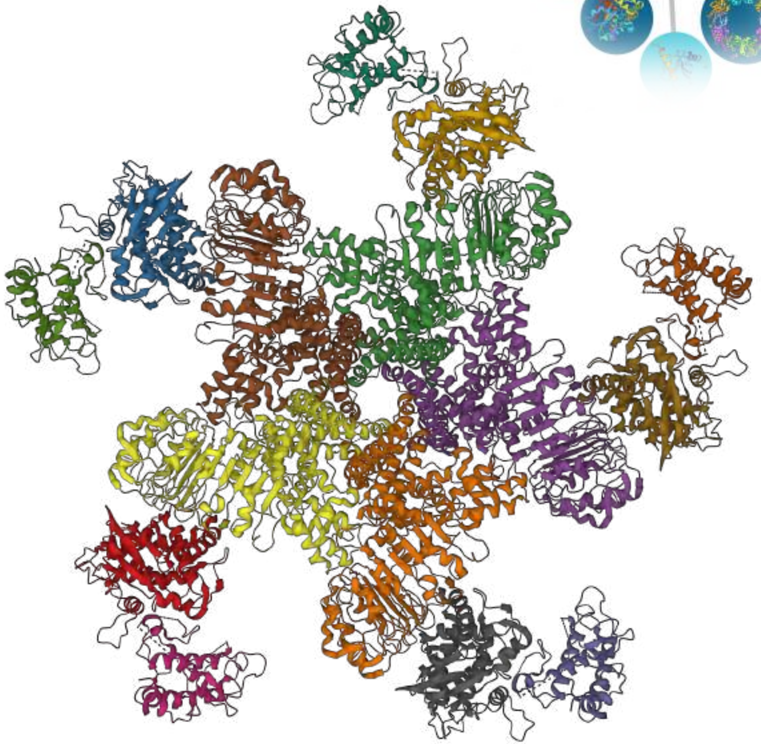
- PDBx/mmCIF Data Standard (https://mmcif.wwpdb.org/)
  - Became the standard file format for data in the PDB Archive in 2014
  - Files are given the extension ".cif" (or ".cif.gz")
  - Supersedes the now legacy "PDB" (".pdb") file format, which is limited by fixed-column width restrictions and is no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




```
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
WWPDB  D_1300010525
EMDB   EMD-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          EMD-0680
_pdbx_database_related.content_type   'associated EM volume'
#
...
```

Case study: *Arabidopsis thaliana* resistosome

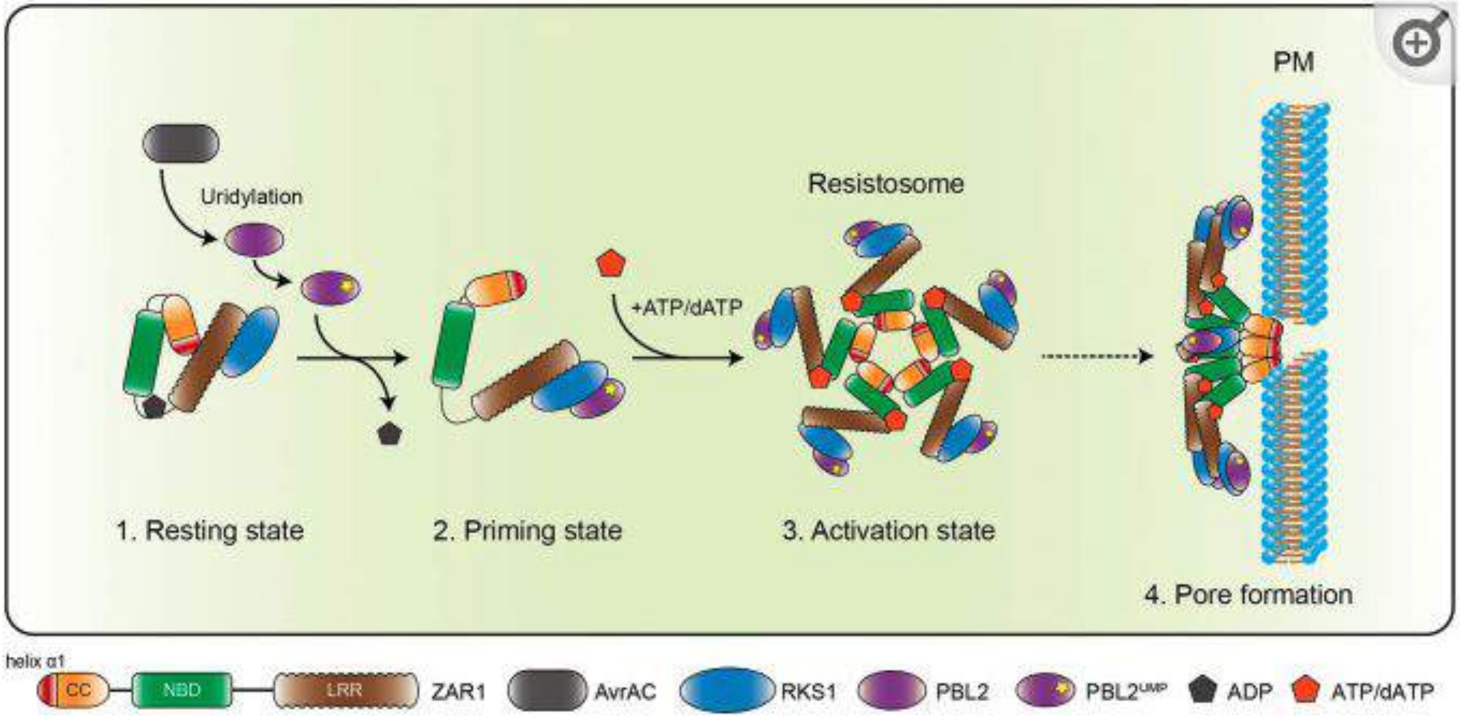
- *Arabidopsis thaliana*: the gold standard model organism in plant molecular biology
- In plants, the innate immune response is triggered via the sensing of pathogens by nucleotide-binding, leucine-rich repeat receptors (NLRs)
- Upon pathogen sensing, NLRs oligomerize to form the activated “resistosome”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



***A. thaliana* resistosome**
PDB 6J5T (Wang 2019)

Wang, J., et al. (2019) *Science* **364**(6435), eaav5870.

Case study: *Arabidopsis thaliana* resistosome

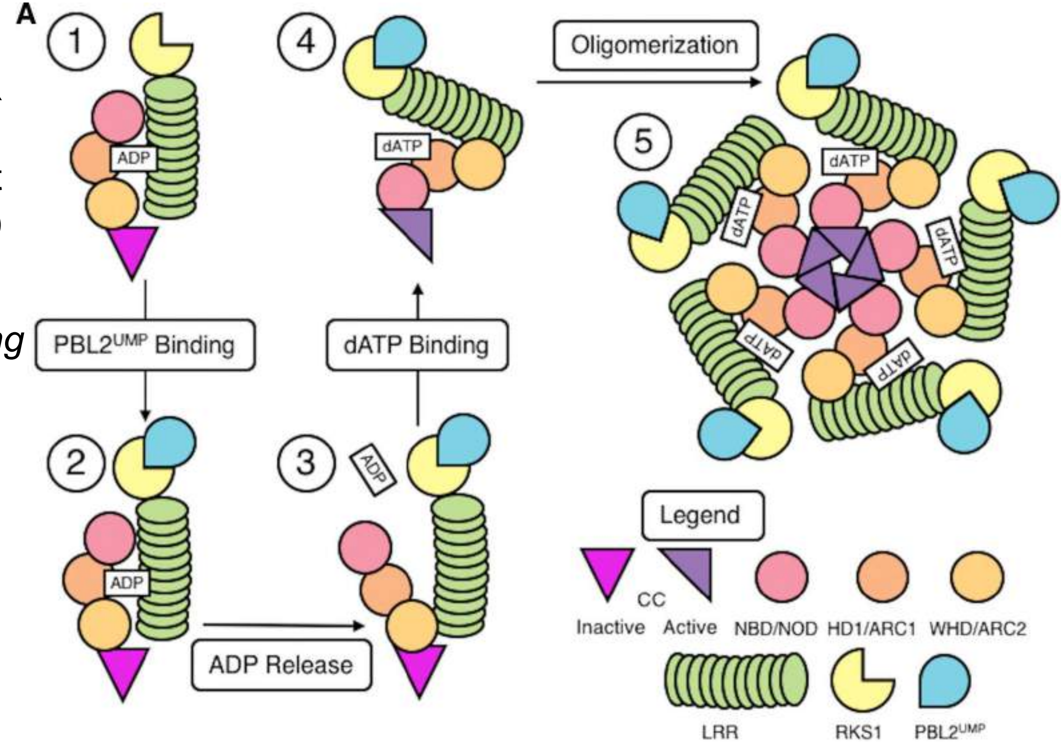


Shi, X., et al. (2019) *Abiotech* **1**(2), 147–150.

Case study: *Arabidopsis thaliana* resistosome

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

Indirect pathogen sensing



Resistosome

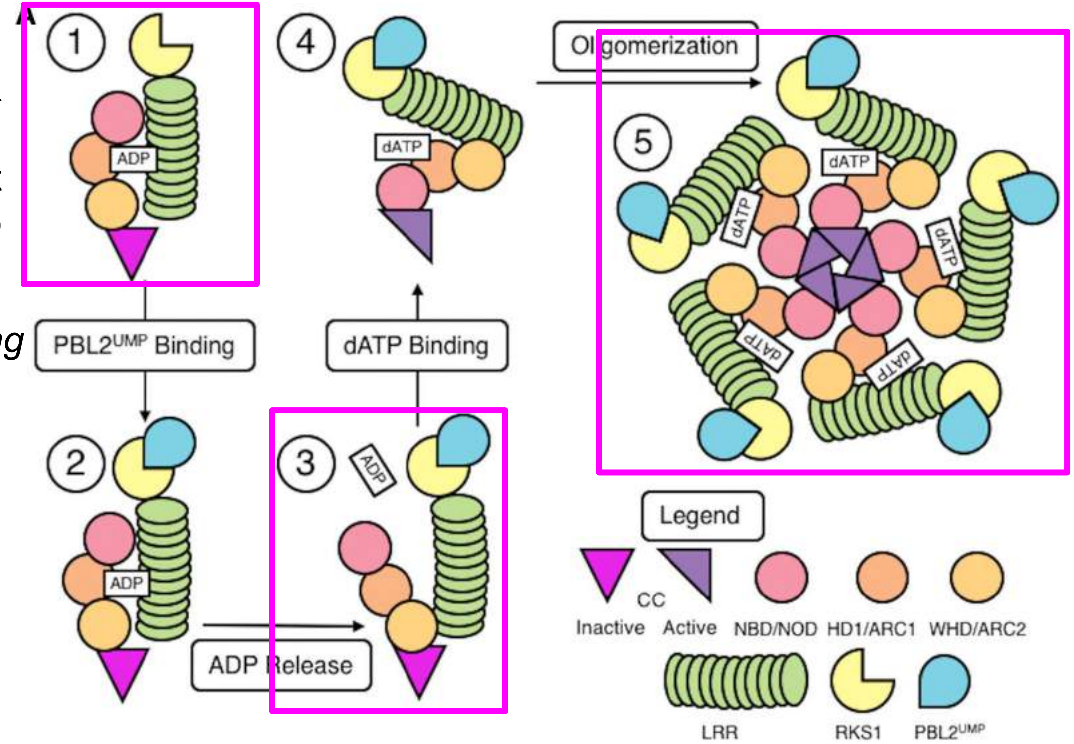
Burdett, H., et al. (2019) *Cell Host & Microbe* **26**(2), 193–201.
Wang, J., et al. (2019) *Science* **364**(6435), eaav5870.

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

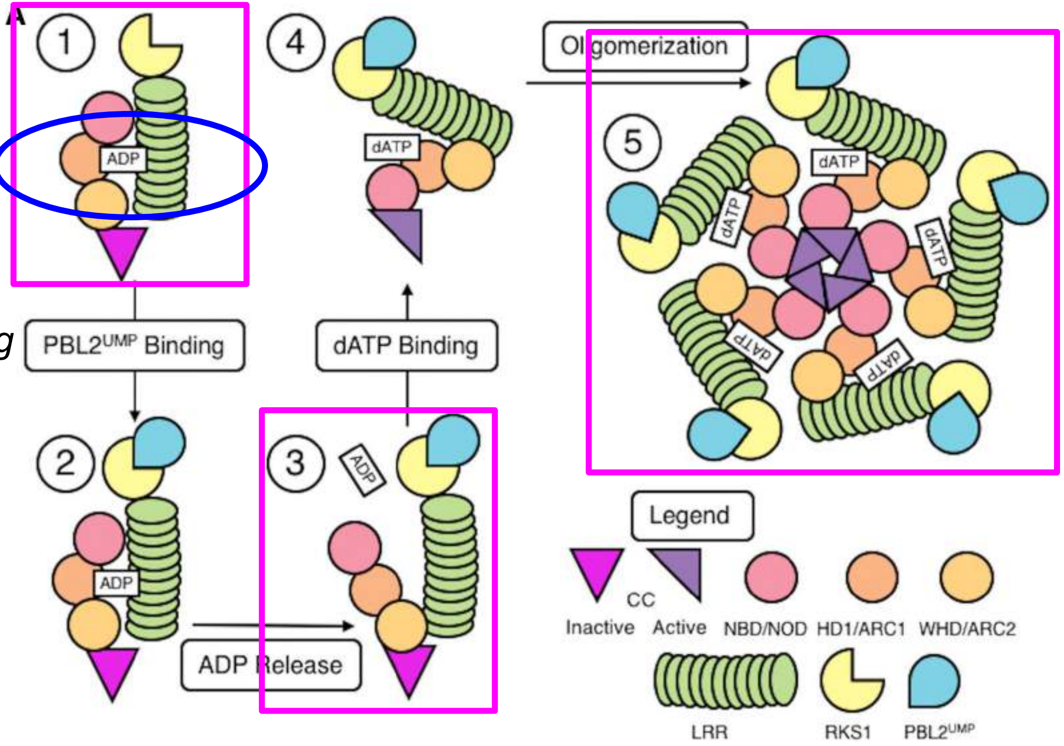
Burdett, H., et al. (2019) *Cell Host & Microbe* **26**(2), 193–201.
Wang, J., et al. (2019) *Science* **364**(6435), eaav5870.

Tutorial: Accessing structure data in the PDB

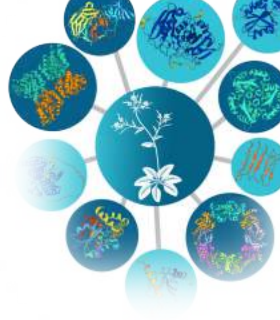
NLR
(nucleotide-binding, leucine-rich repeat receptor)

Indirect pathogen sensing

Let's try to find these structures



Resistosome



Burdett, H., et al. (2019) *Cell Host & Microbe* **26**(2), 193–201.
Wang, J., et al. (2019) *Science* **364**(6435), eaav5870.

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
    - Using FTP: ftp://ftp.wwpdb.org

Resources and Documentation

- Basic and advanced searching:
  - Infographic: https://cdn.rcsb.org/rcsb-pdb/search/SearchnBrowse2go.pdf
  - https://www.rcsb.org/docs/search-and-browse/overview-search-and-browse
- Programmatic access:
  - Search API: https://search.rcsb.org/index.html#search-api
  - Downloading structures:
    - https://www.rcsb.org/docs/programmatic-access/file-download-services
    - https://www.wwpdb.org/ftp/pdb-ftp-sites
    - AWS S3: https://www.rcsb.org/news/6266e0379c3931864b072861
    - 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

RUTGERS

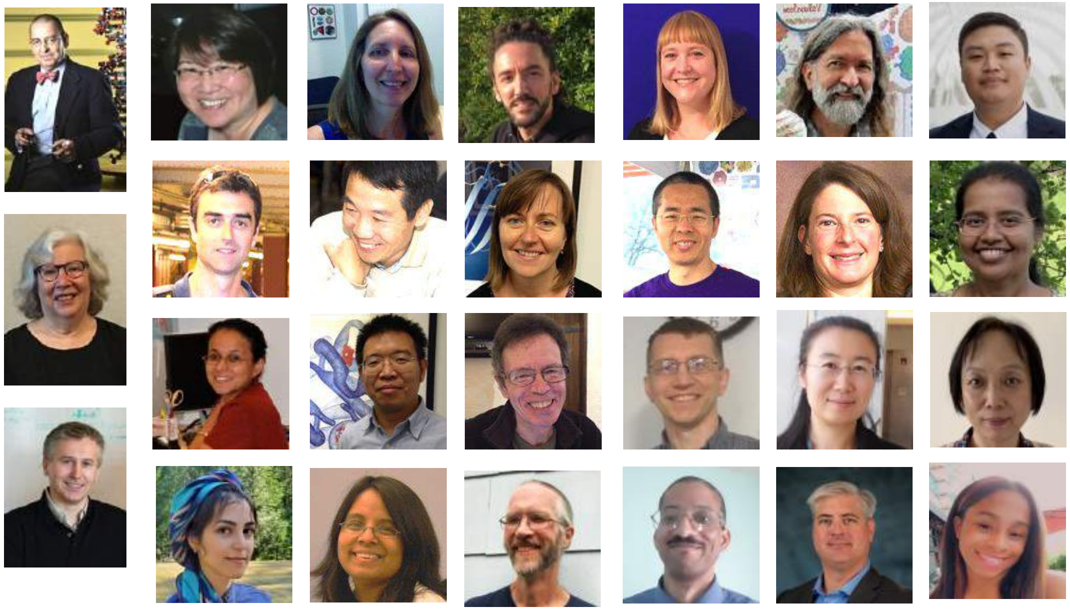
UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UCSF University of California San Francisco

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

Follow us



John D. Westbrook
In memoriam
1957-2021

RCSB PDB PROTEIN DATA BANK

rcsb.org

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

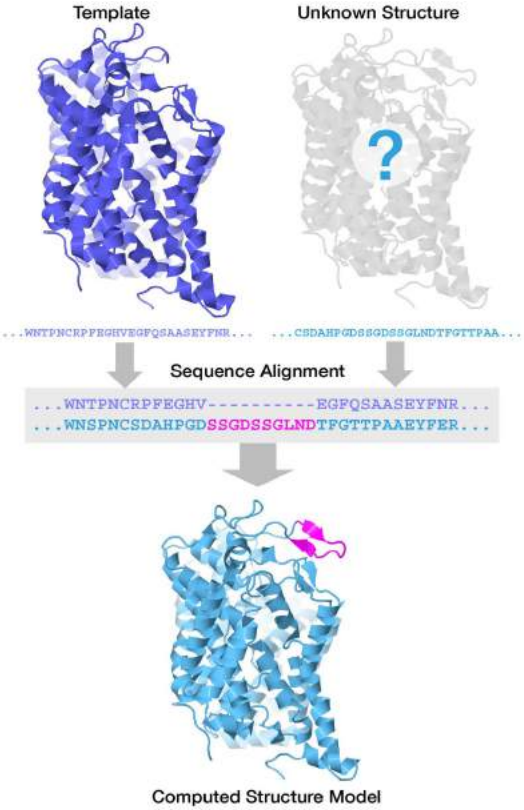
- CSMs in RCSB.org

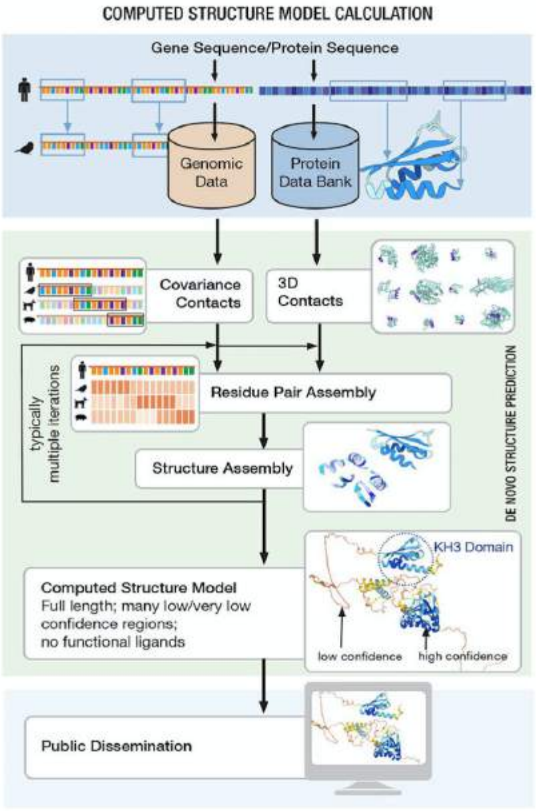
- Accessing CSMs on RCSB.org
  - Live demo: RCSB.org tools and functionalities
  - Case studies: Disease resistance RPP13-like protein 4 from Arabidopsis thaliana resistosome and Micrococcus luteus aldehyde dehydrogenase

What are computed structure models?

Protein structure prediction

Template-based modeling





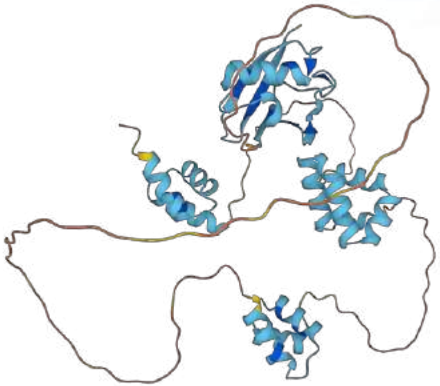
Template-free modeling

Baker, D., Sali, A. Science. 2001. 294, 93-6; Abriata, L.A., et al., Proteins. 2018. 86 Suppl 1, 97-112
https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/computed-structure-models

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**



Jumper, J., et al. Nature. 2021. 596, 583–589
Baek, M., et al. Science. 2021. 373, 871-876
Humpheys, I.R., et al. Science. 2021. 374, eabm4805
Shao, C. et al. Structure. 2022. 30, 1385-1394

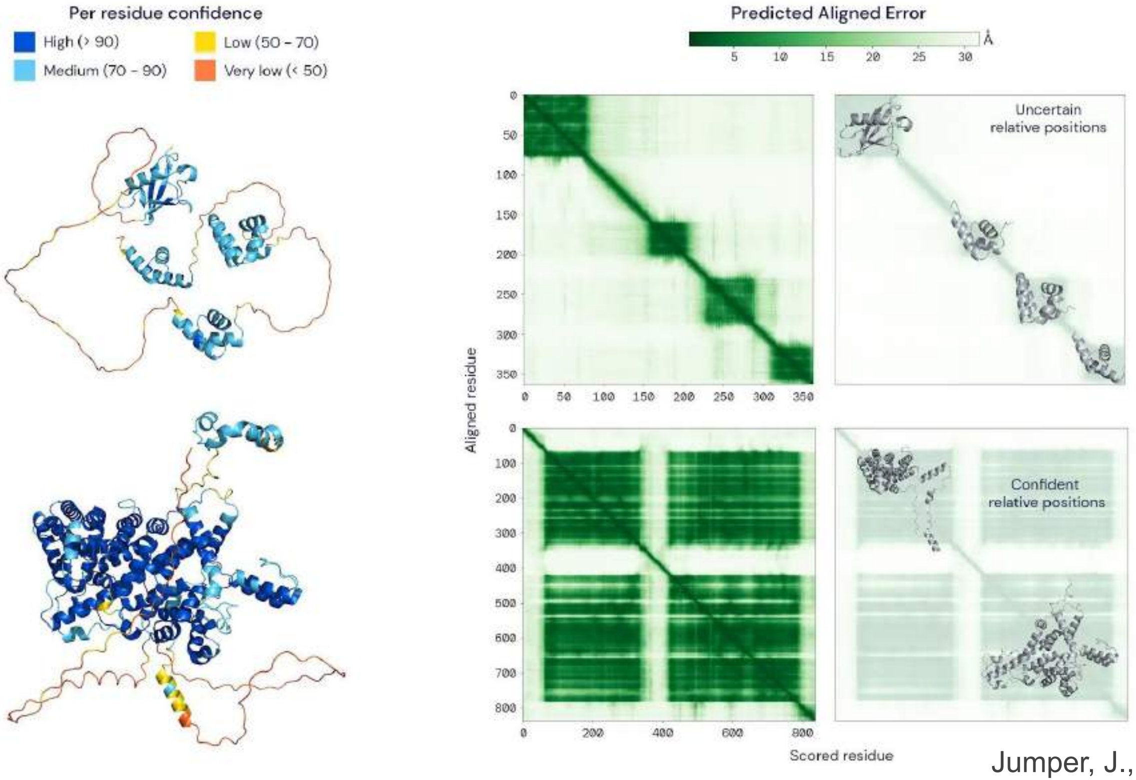
AlphaFold Protein Structure Database and ModelArchive

- 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

Varadi, M., et al. Nucleic Acids Res. 2022. 50, D439–D444
UniProt Consortium. Nucleic Acids Res. 2021. 49, D480-D489
Humpheys, I.R., et al. Science. 2021. 374, eabm4805

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

Jumper, J., et al. Nature. 2021. 596, 583–589
Tunyasuvunakool, K., et al. Nature. 2021. 596, 590-6
https://www.deepmind.com/publications/enabling-high-accuracy-protein-structure-prediction-at-the-proteome-scale

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
https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/computed-structure-models

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

Experimental Structures

CSMs

https://www.rcsb.org/docs/general-help/computed-structure-models-and-rcsborg
https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/computed-structure-models

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)

Humpheys, I.R., et al. Science. 2021. 374, eabm4805
Baek, M., et al. Science. 2021. 373, 871-876
https://www.modelarchive.org/

Jumper, J., et al. Nature. 2021. 596, 583–589
Varadi, M., et al. Nucleic Acids Research. 2022. 50, D439–D444

Accessing CSMs on RCSB.org
Live Demo

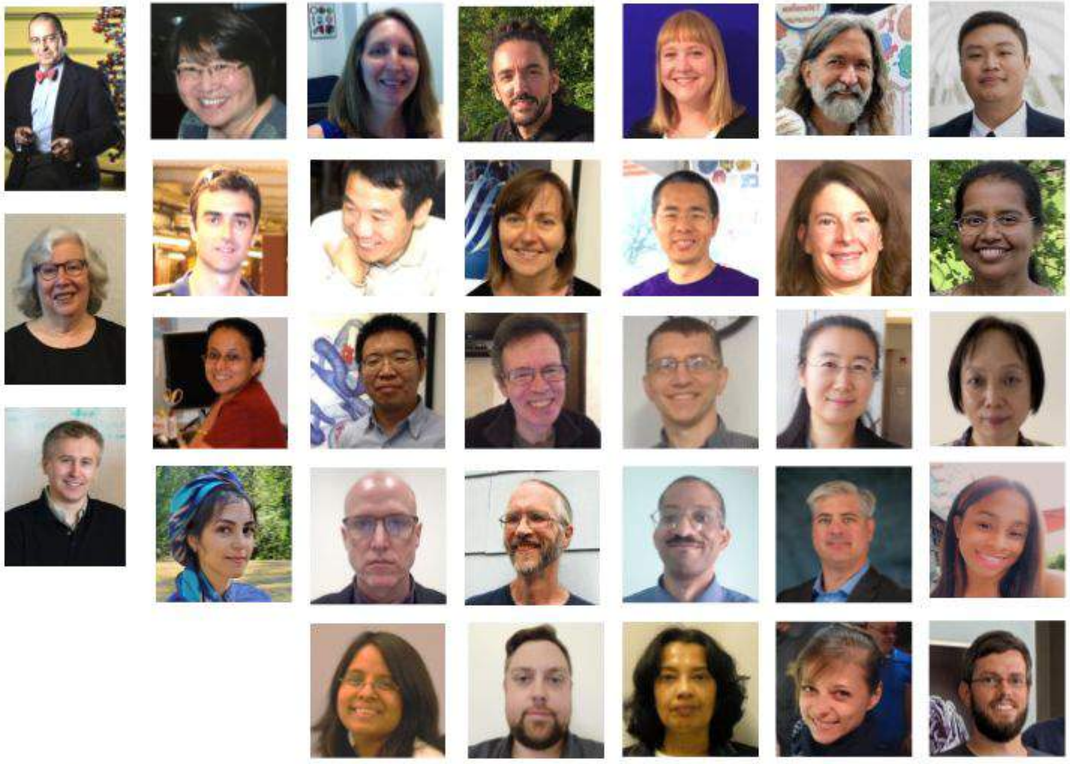
Summary

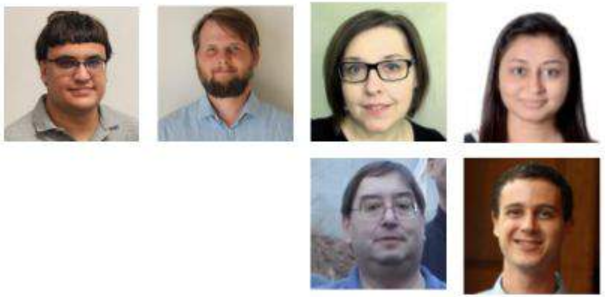
- RCSB.org now provides access to >1,000,000 CSMs from AlphaFold2 and RoseTTAFold alongside ~200,000 experimental structures of macromolecules

- 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 PDB PROTEIN DATA BANK

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

RUTGERS

UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UCSF University of California San Francisco

WORLDWIDE PDB PROTEIN DATA BANK

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

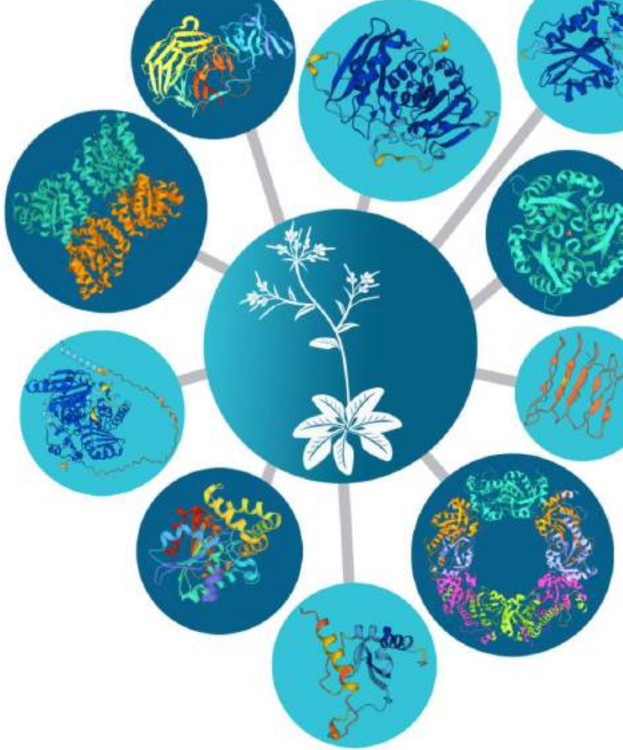
Follow us

John D. Westbrook
In memoriam
1957-2021

Thank you!

RCSB PDB
PROTEIN DATA BANK

rcsb.org

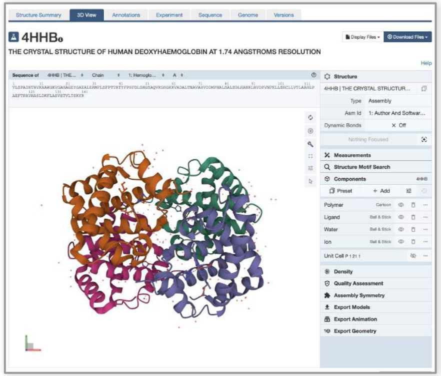


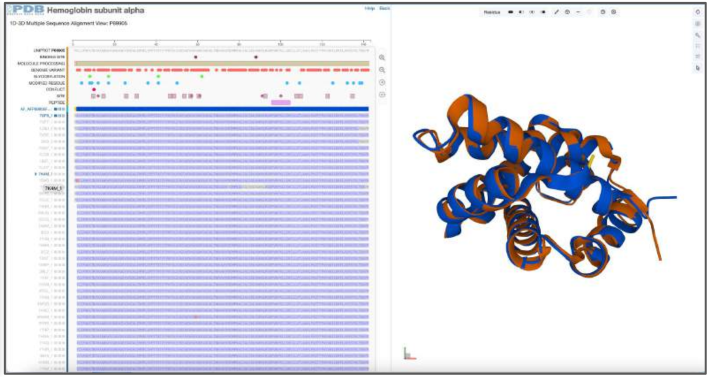
Introduction to the Mol* Molecular Graphics System

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

Overview

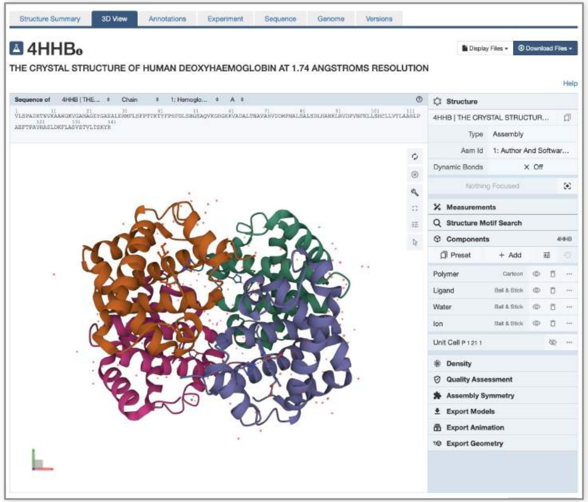
- Exploring a 3D structure from RCSB.org using Mol*
- Standalone Mol*
- Exploring and Comparing multiple 3D structures using Mol*



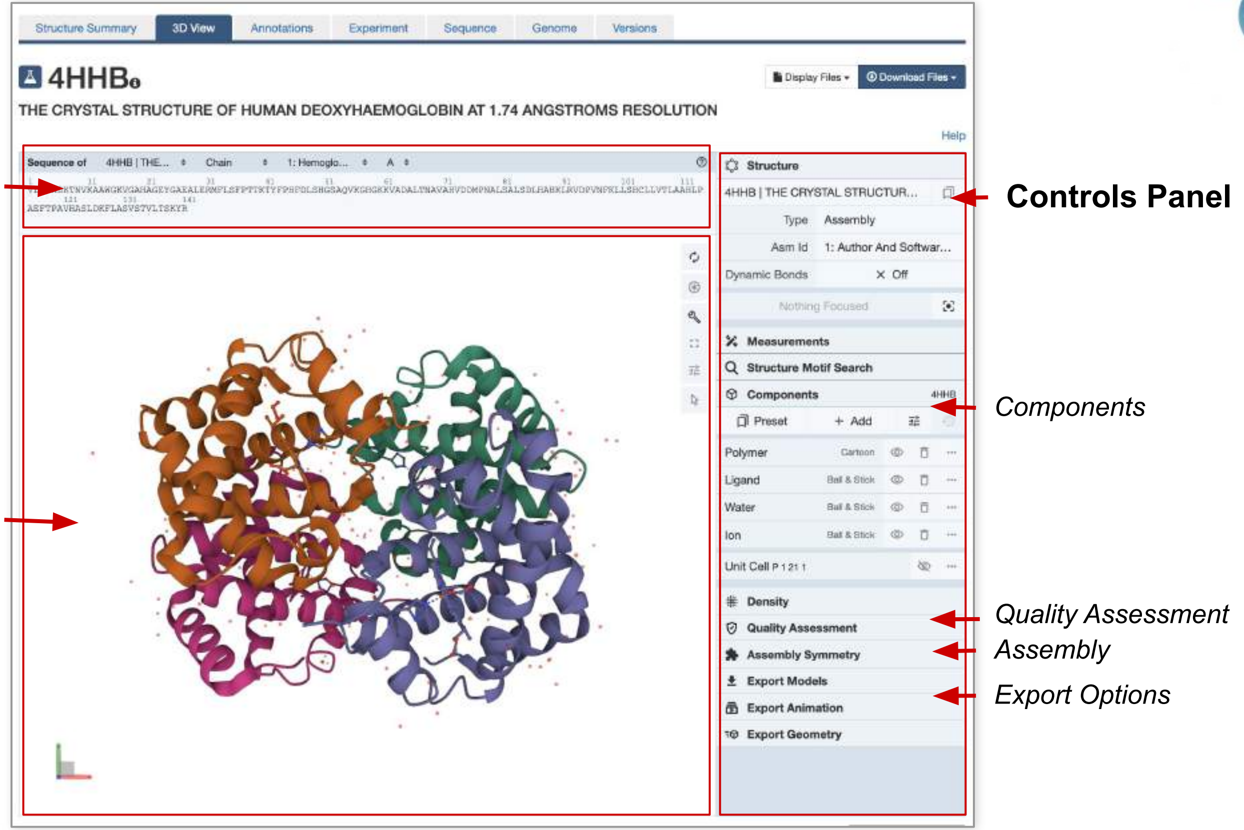


Overview

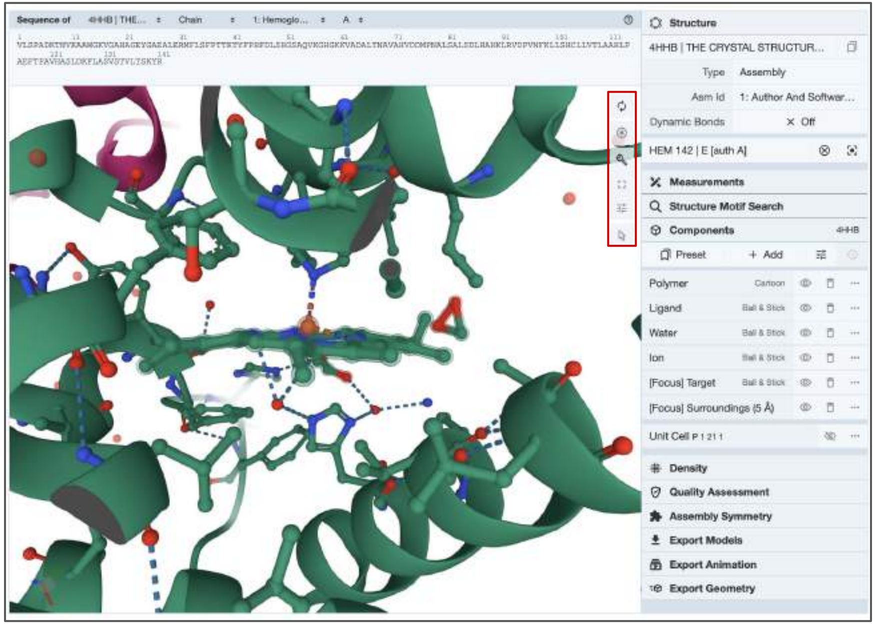
- Exploring a 3D structure from RCSB.org using Mol*
  - Overall structure
  - Focus vs Select
  - Display and Color
  - Measure
  - Explore Annotations (1D-3D viewer)
- Standalone Mol*
- Exploring and Comparing multiple 3D structures using Mol*

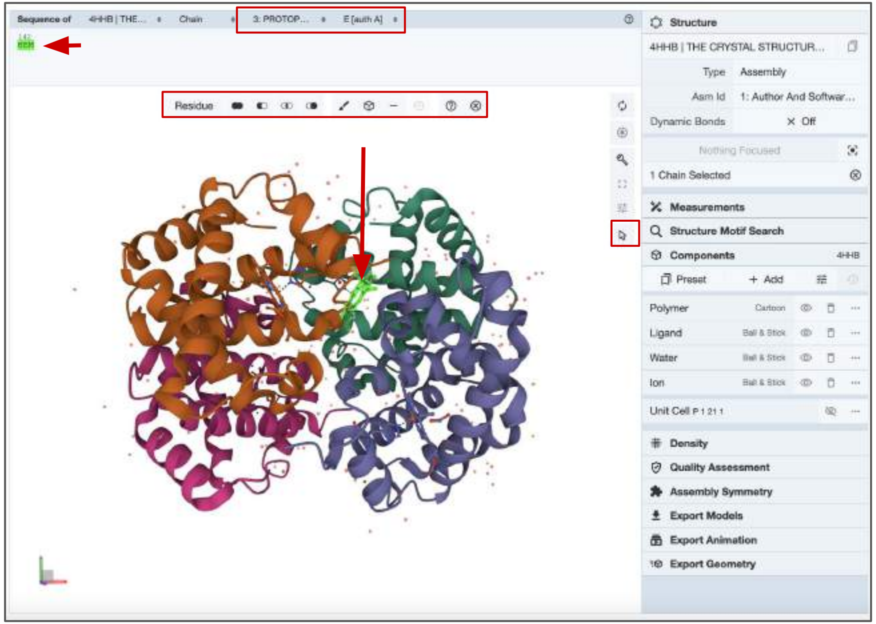


Overall Structure (Access from Structure Summary Page → 3D View)

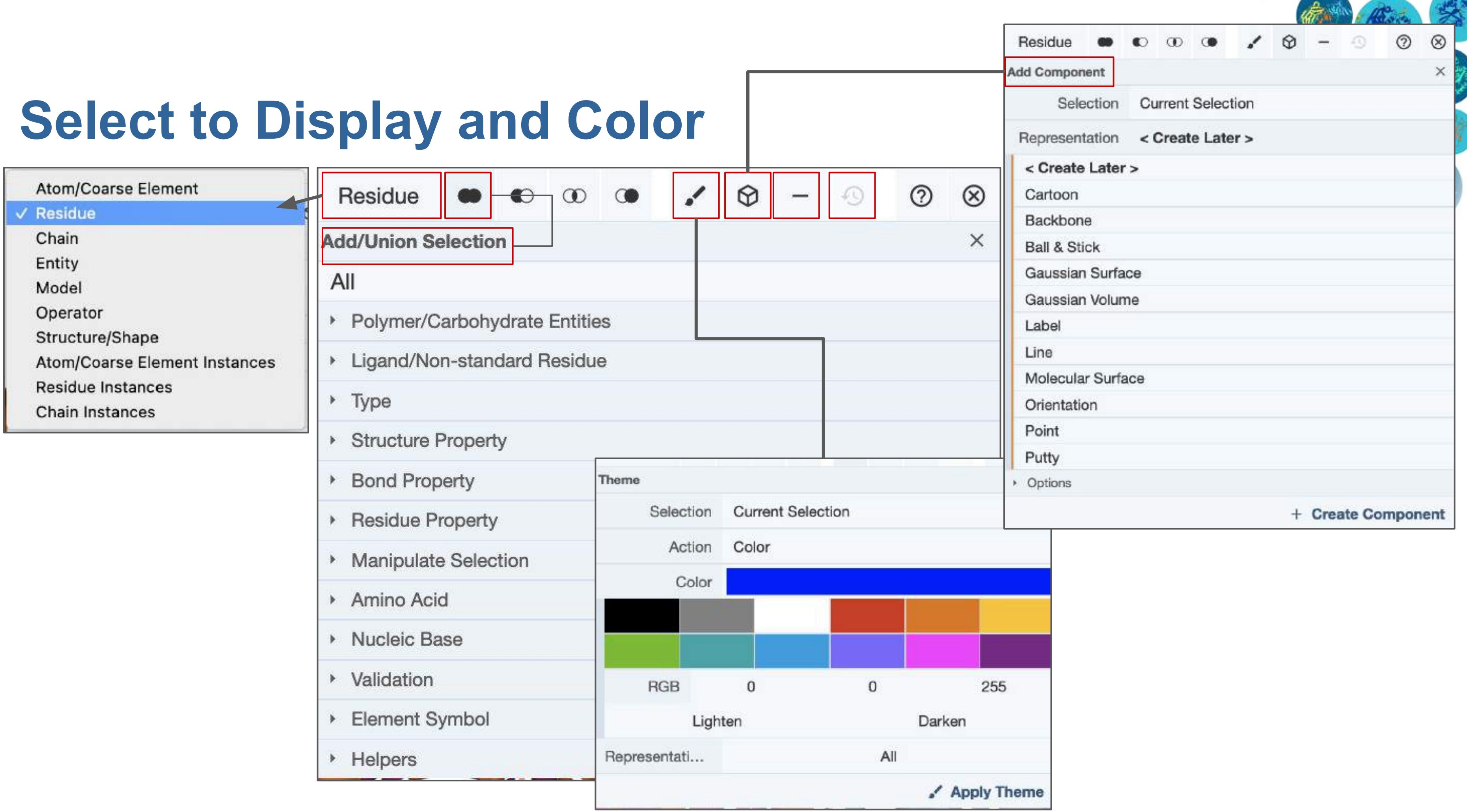


Focus vs Select



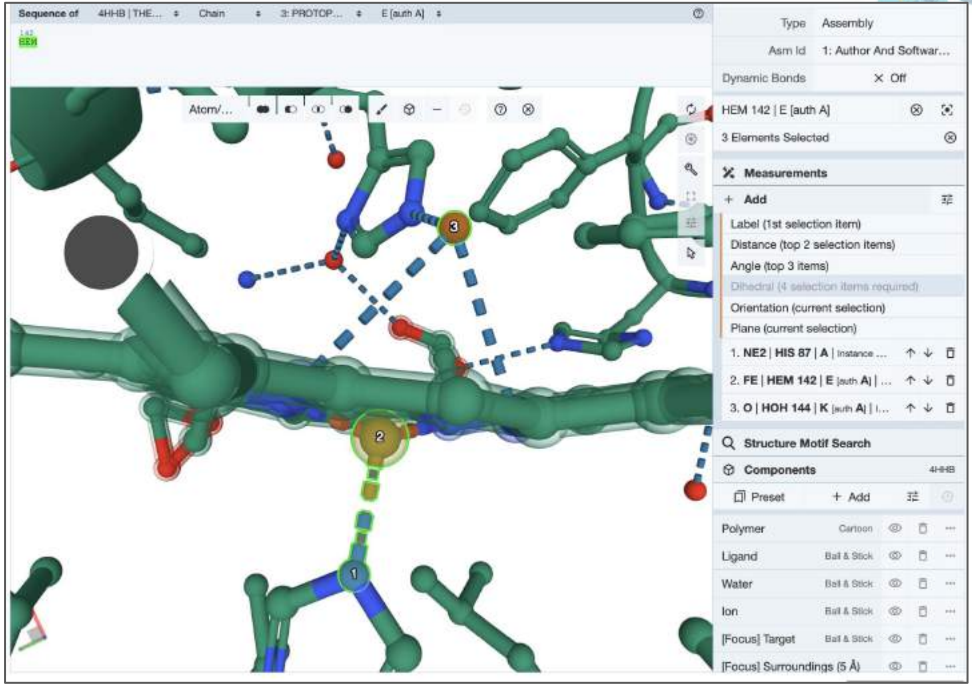


Select to Display and Color

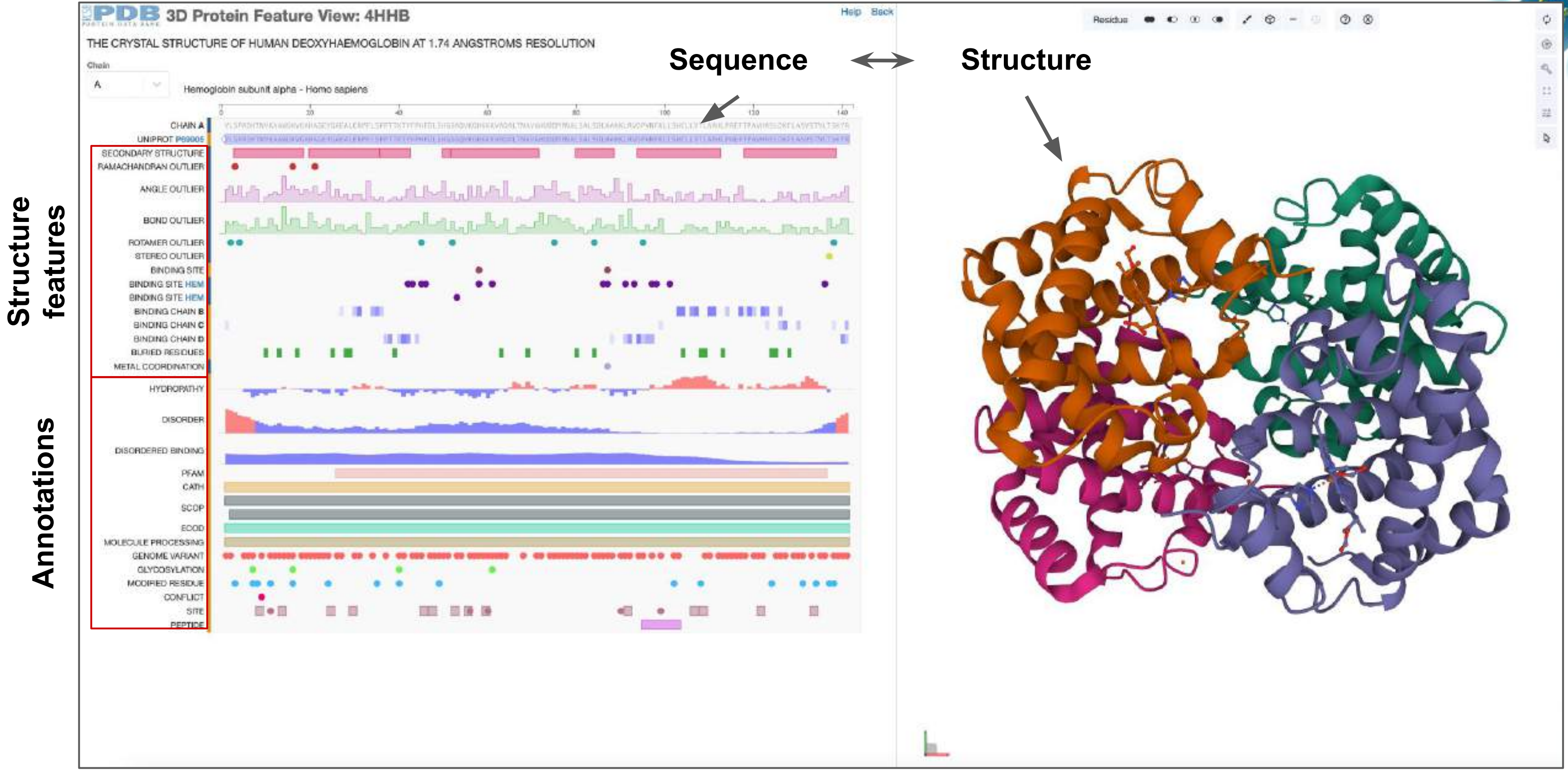


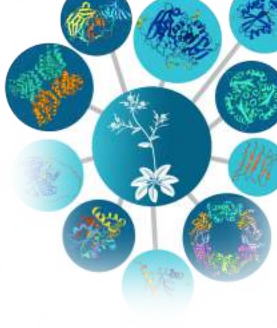
Measure

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



Explore Annotations (1D-3D View)

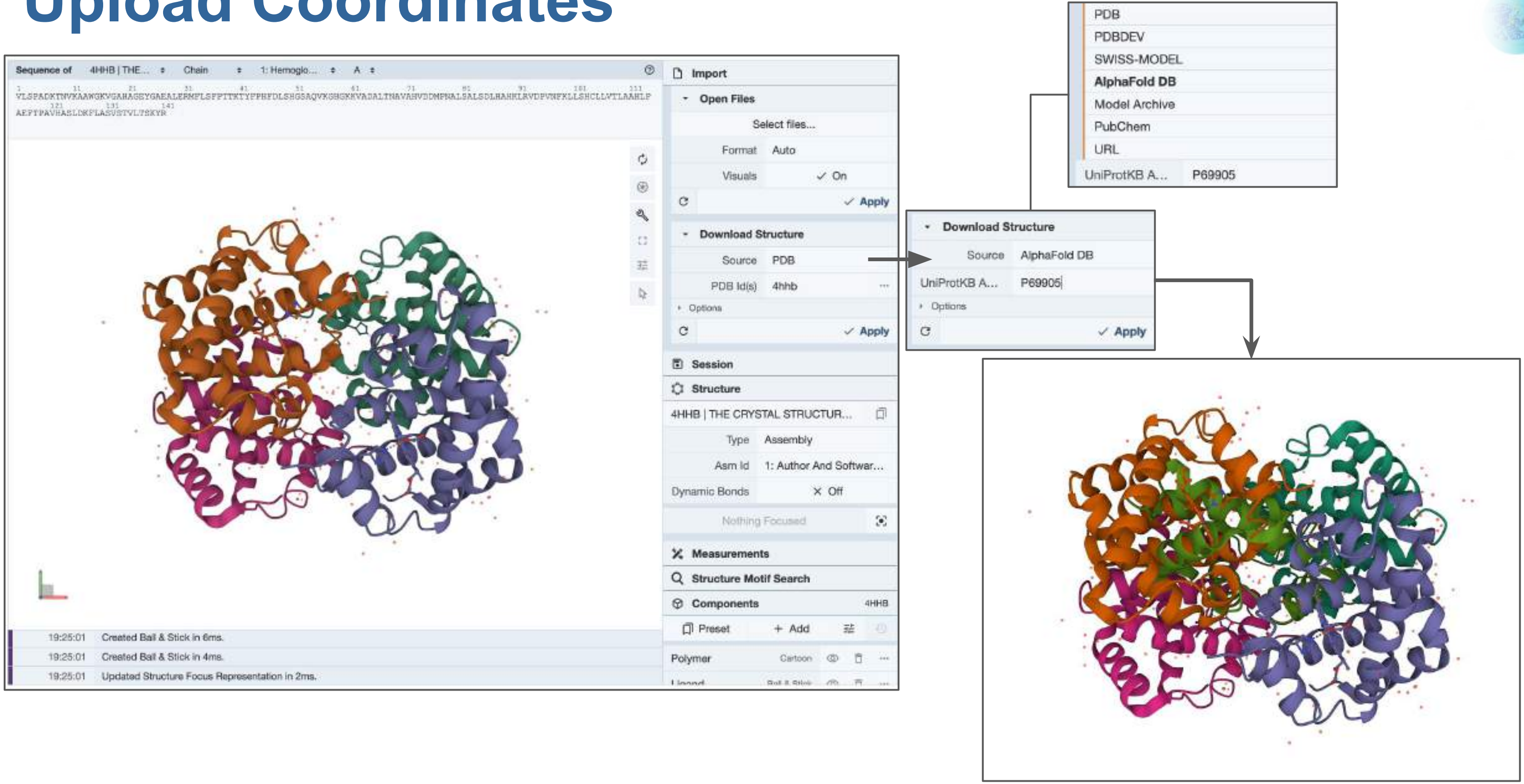




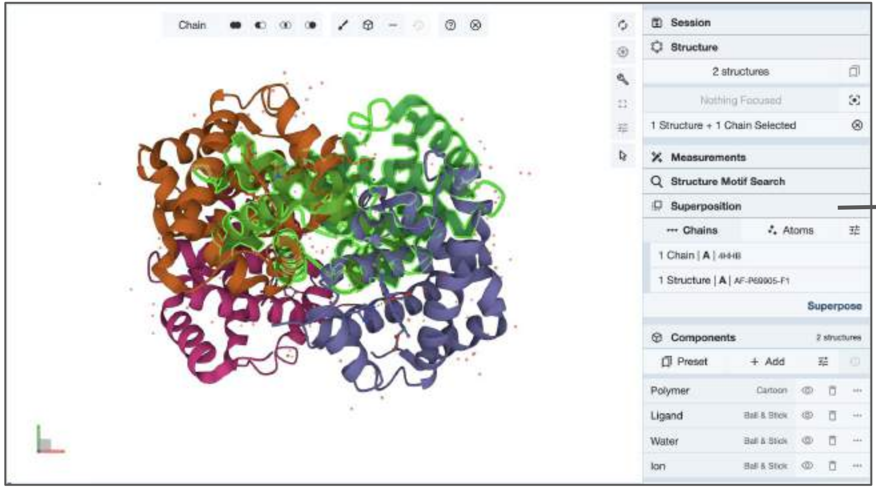
Overview

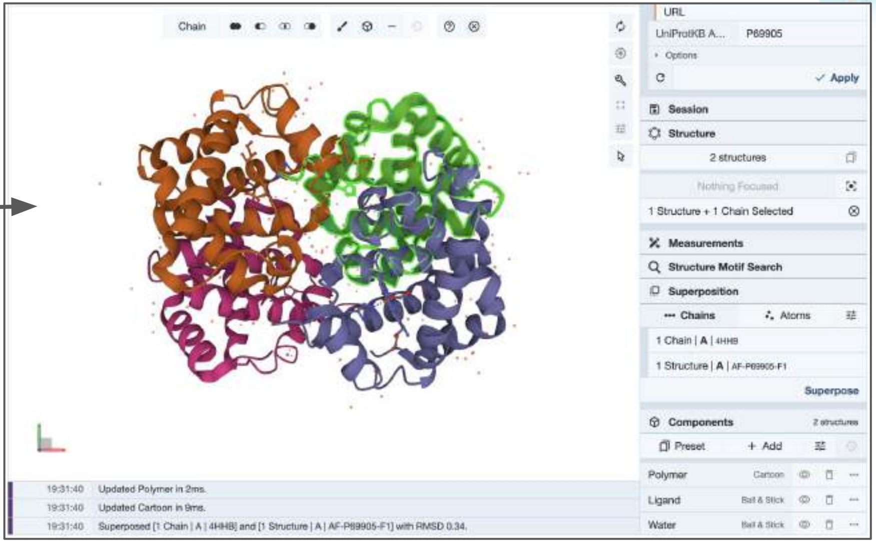
- Exploring a 3D structure from RCSB.org using Mol*
- Standalone Mol* (https://www.rcsb.org/3d-view)
    - Upload your coordinates
    - Compare structures
- Exploring and Comparing multiple 3D structures using Mol*

Upload Coordinates



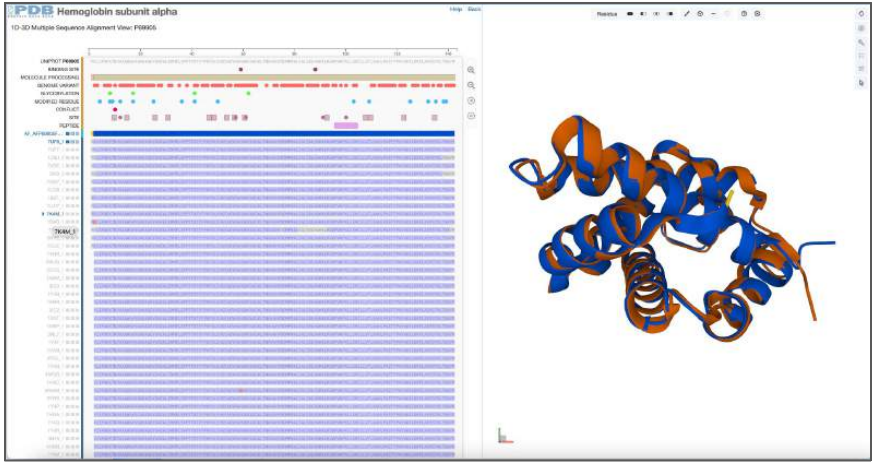
Compare Structures



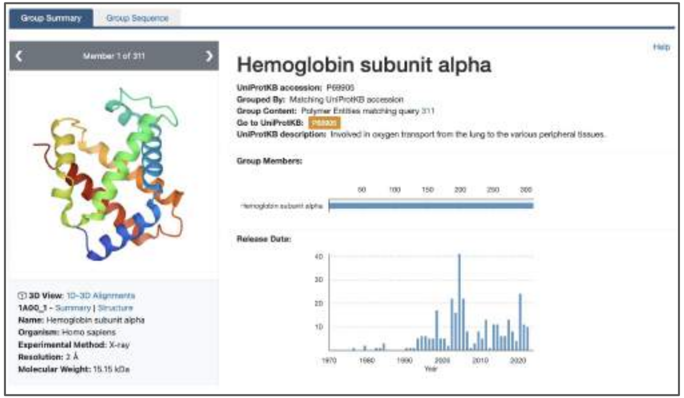


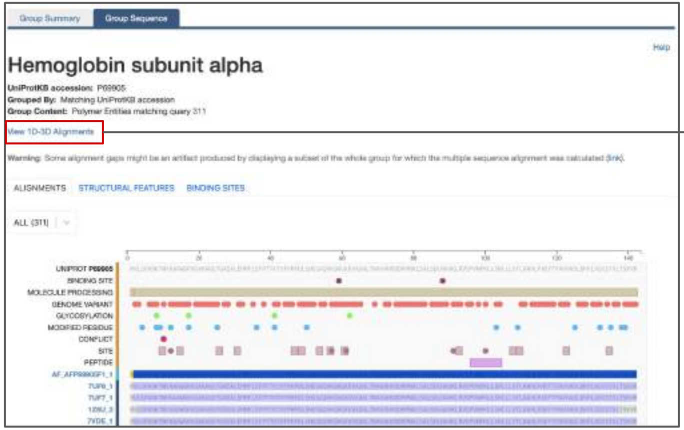
Overview

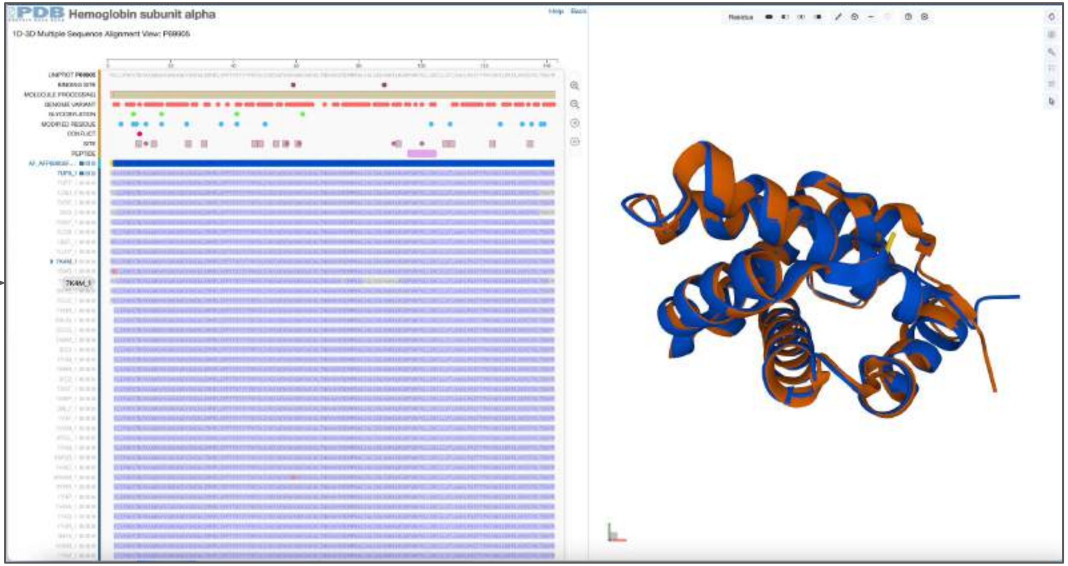
- Exploring a 3D structure from RCSB.org using Mol*
- Standalone Mol*
- Exploring and Comparing multiple 3D structures using Mol*
    - View 1D-3D Alignments



View 1D-3D Alignments

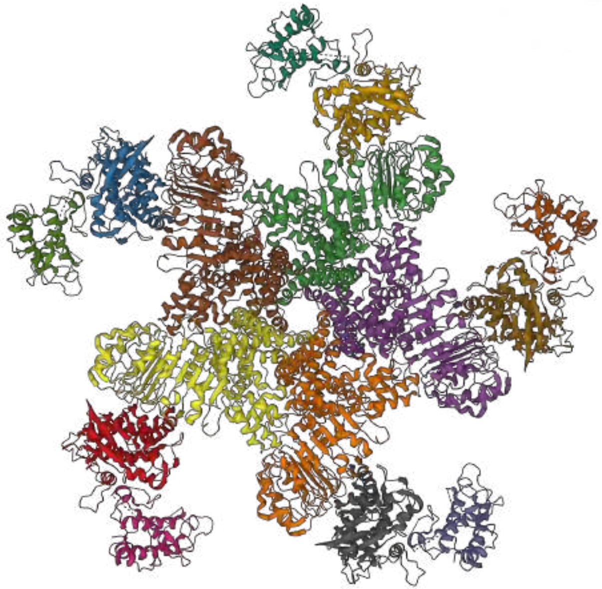






Case Study: *Arabidopsis thaliana* Resistosome

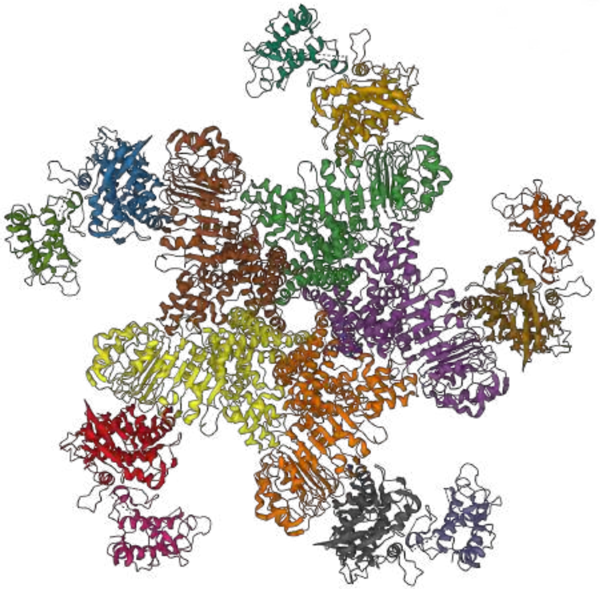
- Which is the Disease resistance RPP13-like protein 4?
- What proteins does it interact with?
- What ligands does it interact with?
- Mutations @ 297 vs 359?
- How do other (experimental/CSM) structures of this protein compare with this structure



PDB 6J5T (Wang 2019)

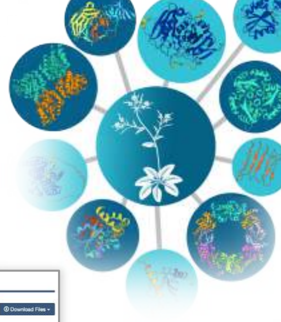
Exploring: *Arabidopsis thaliana* Resistosome

- 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

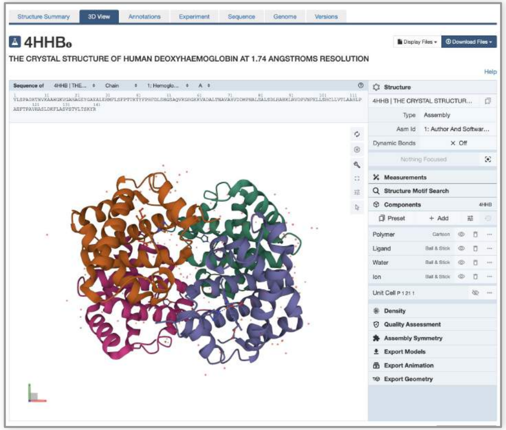


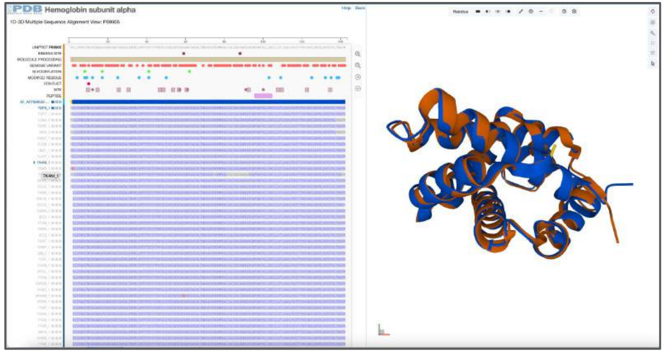
PDB 6J5T (Wang 2019)

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

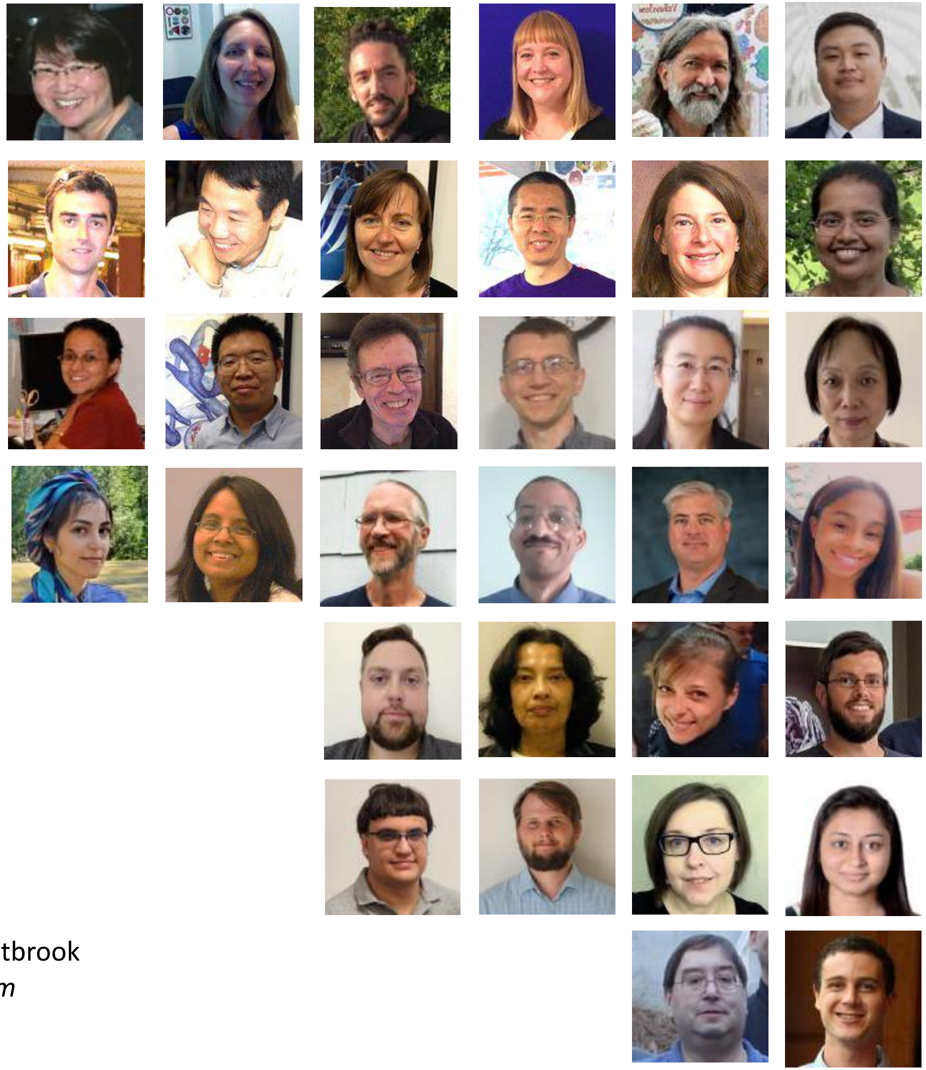
National Science Foundation (DBI-1832184), National Institute of General Medical Sciences, National Institute of Allergy and Infectious Disease, and National Cancer Institute (NIH R01GM133198), and the US Department of Energy (DE-SC0019749)

Management

Rutgers
UC San Diego
SDSC San Diego Supercomputer Center
UCSF University of California San Francisco

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

Follow us



John D. Westbrook
In memoriam
1957-2021

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

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

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

INTEGRATION and MODELING *for* PREDICTIVE BIOLOGY

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

Import ProteinStructures from a Metadata File in Staging Area

Narrative ID	Object name (Genome AMA feature set)	Feature ID	PDB filename	Is model
107138	MLuteus_ATCC_49442	MLuteus_masurca_RAST.CDS.3483	MLuteus_AlphaFold_3483.pdb	yes
107138	MLuteus_ATCC_49442	MLuteus_masurca_RAST.CDS.3484	MLuteus_Alphafold_3484.pdb	yes
107138	MLuteus_ATCC_49442	MLuteus_masurca_RAST.CDS.133	pdb_extract_MLuteus_133.cif	yes
107138	MLuteus_ATCC_49442	MLuteus_masurca_RAST.CDS.133	ma-bak-cepc-0245.cif	yes
107138	MLuteus_ATCC_49442	MLuteus_masurca_RAST.CDS.133	vinaoutNtarget_133.pdb	yes

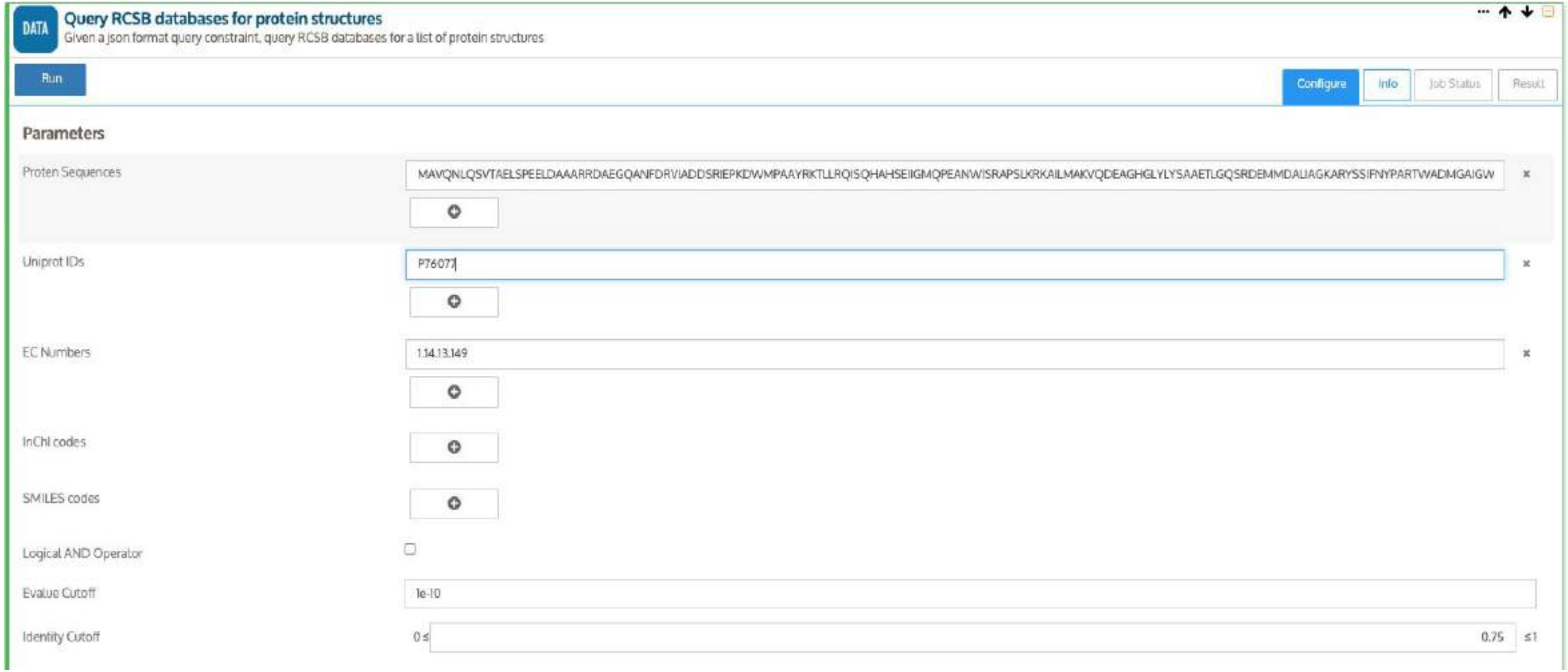
Purpose: With your own protein structure files, you can link them with KBase genome data that is in your narratives or shared with you. Given certain sequence similarity threshold, annotation data of structures and genes that satisfy the threshold will be uploaded (saved) to KBase as an object of the type of **ProteinStructures**.

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Query RCSB databases for protein structures

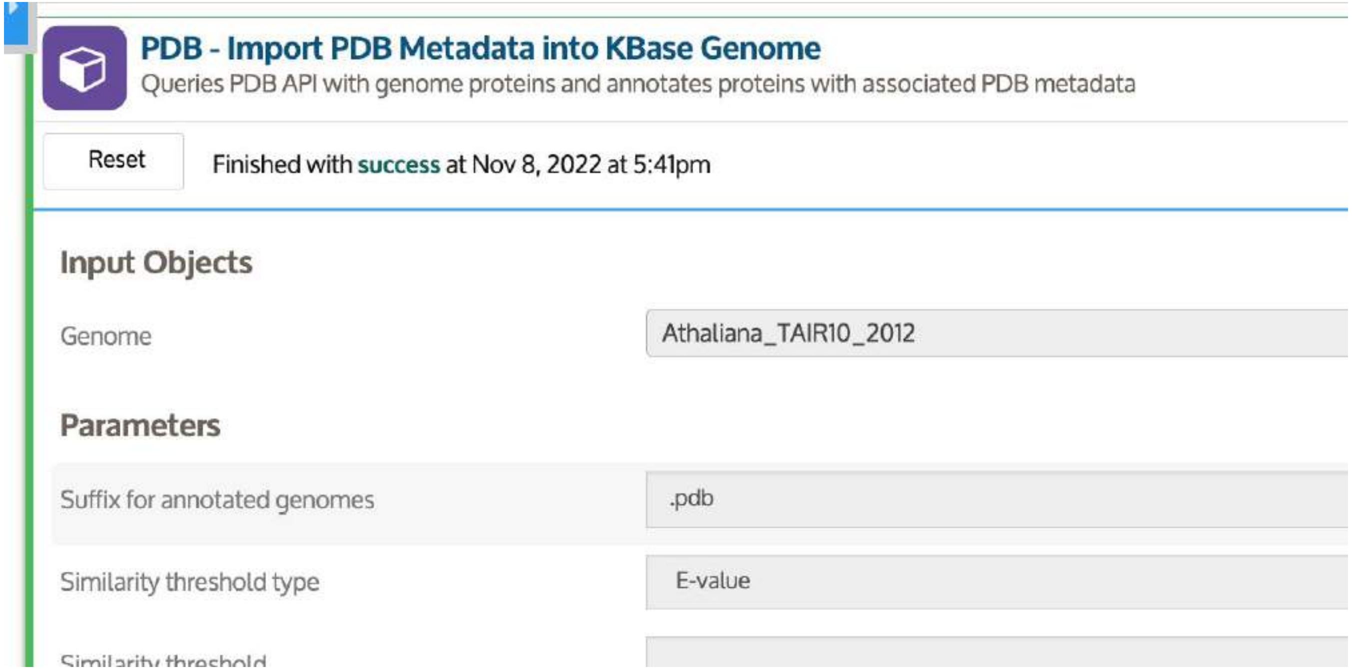
Purpose: By specifying query filters, you can search the RCSB database to find structure hits that meet the criteria you give.



U.S. DEPARTMENT OF ENERGY
Office of Science

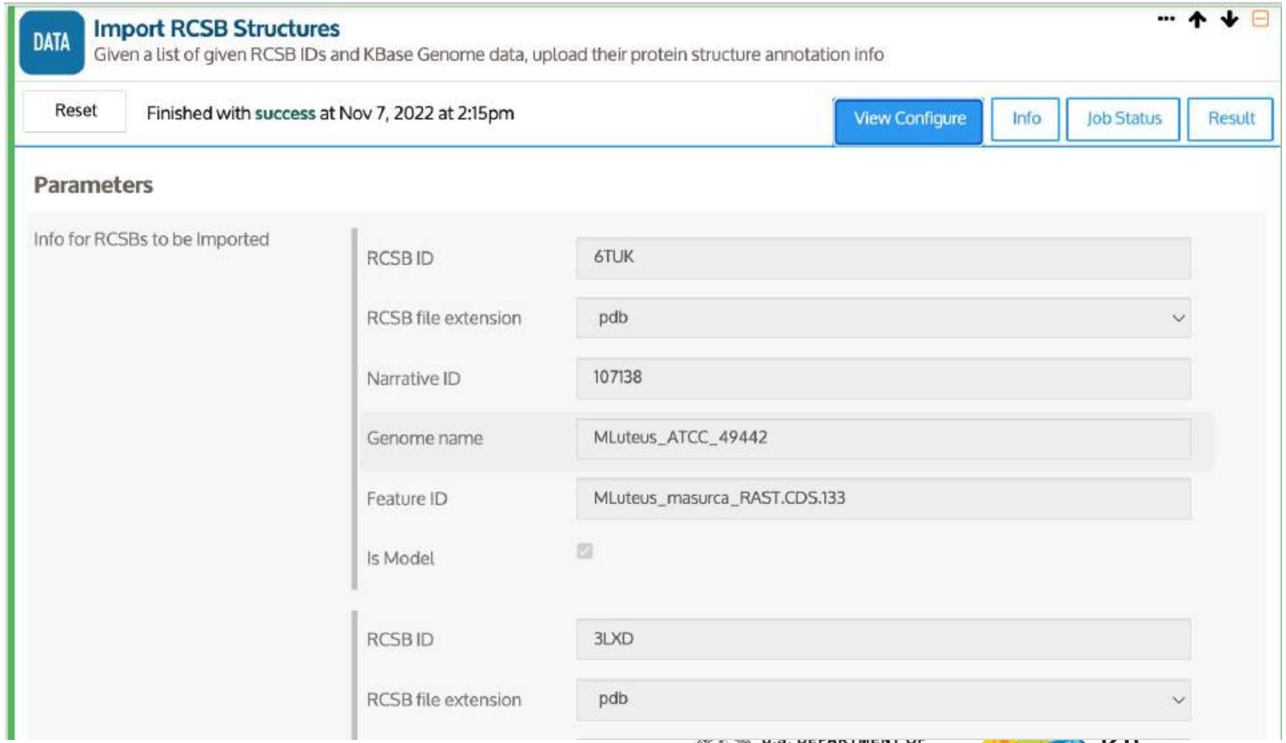
KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

PDB - Import PDB Metadata into KBase Genome



Import RCSB Structures

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.



To the demo…

Demo in a KBase Narrative:

https://narrative.kbase.us/narrative/130799

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

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?

U.S. DEPARTMENT OF ENERGY
Office of Science

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

KBase
PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Making the Best use of Protein Structure Data in KBase and PDB

Christopher S. Henry

November 10th, 2022

INTEGRATION and
MODELING *for*
PREDICTIVE BIOLOGY

RUTGERS

Argonne NATIONAL LABORATORY

RCSB PDB PROTEIN DATA BANK

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

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

Stephen Burley

John Westbrook

Ed O'loughlin

Qizhi Zhang

Kelly Skinner

Jack Gilbert

Sam Seaver

Claudia Lerma-Ortiz

Nidhi Gupta

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!