Constructing a Data Validation Tool for Identifying Structure-Function Matches in Protein & Assay Databases

CS520: Data Curation Project | Group 13

BACKGROUND

Starting With Some Important Definitions

- What are we doing?
  - Working with Dr. David Minh's lab to find mapping between an assay database and a 3-D Function database
- What is an assay?
  - An assay is a test conducted in a lab (use uniProt + InChI)
- What is a protein?
  - Proteins are building blocks for the human body (GID: uniProt)
- Why is the structure-function relationship important?
  - It helps us identify *how* biological functions are facilitated

BACKGROUND

A bit more on the scientific background

- Proteins
  - Large complex molecules responsible for catalyzing biochemical reactions in the body.
  - Can be identified globally through their **UniProt ID**
- Ligands
  - Drugs/ chemicals that bind to the proteins to elicit a desired cellular response.
  - Can be identified through their **InChi Key**

UNPACKING THE DATABASES

PubChem BioAssay

- Open chemistry database at the National Institutes of Health
- Biological activity data and descriptions of assay (test) procedures and experiments on protein targets
- Data used from PubChem DB in this project:
  - Assay information (Assay ID)
- API:
  - REST-based API

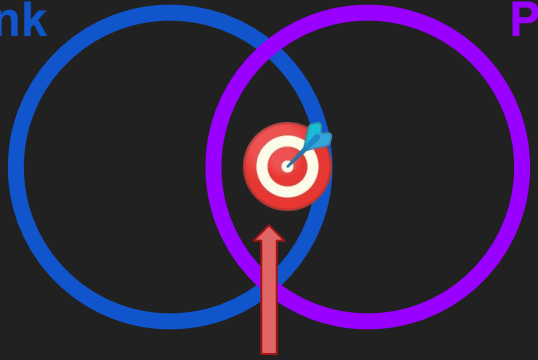
UNPACKING THE DATABASES

RCSB Protein Data Bank (PDB)

- 3D models of proteins, ligands and other important biological molecules (nucleic acids, and complex assemblies).
- Both proteins and ligands are identified by PDB ID
- Data used from RCSB in this project:
  - Protein PDB IDs
  - InChIKeys for non-polymer entities (small molecules) associated with a PDB ID
- API:
  - REST-based API
  - GraphQL-based API

Protein Data Bank

PubChem BioAssay



Target: Pairs of Ligands and Proteins Where The 3D Structure AND Assay Are Given

UniProt ID

Represents a protein.
There are many assays per UniProt ID.

PDB ID

Local key for PDB.
Each UniProt ID has 1 PDB ID.

Assay ID

Represents 1 assay.
There may be multiple InChi keys per Assay.
There is one UniProt ID per Assay.

InChi Key

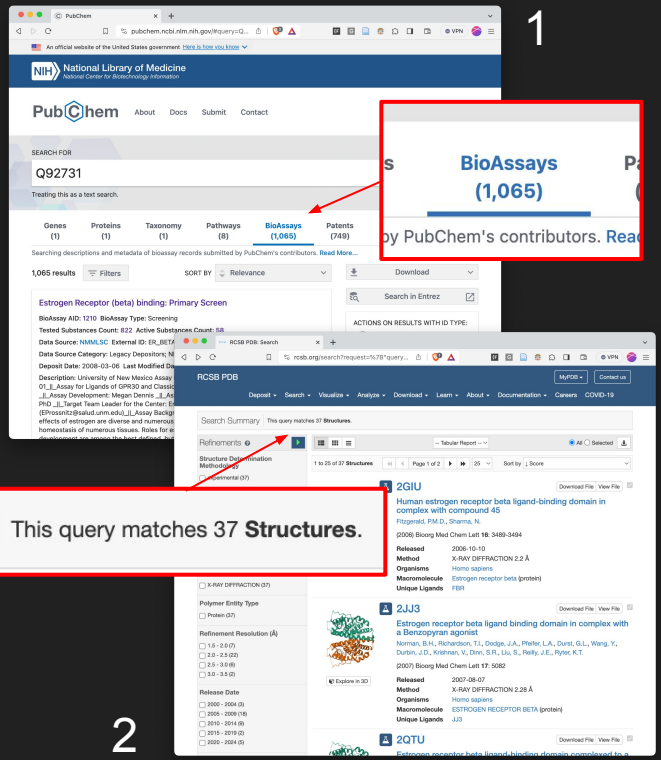
Each InChi Key maps to a PDB ID and Assay ID.

Associated Data

Any relevant data from PDB or PubChem.

IDENTIFYING THE CHALLENGES

Data Validation Challenges



- No 1-to-1 relationship between protein structures and assays through simple query
- There are too many bioAssay results (1) and too few PDB structure (2) results by UniProt query
- The structure-function relationship isn't elucidated by either individual database

ADDRESSING THE CHALLENGES

Synthesizing Multiple Databases

UniProt API used for validation

3D structures are organized by PDB ID and queryable by InChI key + UniProt ID

Data is currently limited without placing sequence similarity requirement in

Assays are organized by BioAssay AID. Tested substances are not directly queryable.

AID is the data of value

ADDRESSING THE CHALLENGES

Our Solution Algorithm

- Given a UniProt ID (user input), check if it is a valid UniProt ID using uniprot.org
- Use UniProt ID to retrieve a FASTA*
- Perform a sequence similarity search against the RCSB Protein Data Bank (PDB) using the FASTA sequence to retrieve PDB IDs
- Fetch InChIKeys for non-polymer entities (small molecules) associated with a PDB ID.
- Fetch Assay IDs given either InChIKeys or UniProt ID

* FASTA is simple way of representing the basic building blocks of proteins or genetic material (like DNA or RNA) in a text format.

ADDRESSING THE CHALLENGES

Our Solution Architecture

Phase I: Find Sequence Similar Proteins

Query Conditions
- UniProt ID: User Input

Query Source
- UniProt for validation
- PDB for sequence similar proteins

Desired Output
- Sequence Similar Proteins

Phase II: Query Structures in PDB

Query Conditions
- UniProt ID: User Input (translated to FASTA)
- e_val cutoff = 0.0001, identity_cutoff = 0.6

Query Source
- RCSB PDB

Desired Output
- An array of PDB identifiers

Phase II: Find Substances From Structures

2a. Get a List of Substance IDs (SIDs)
Output a dictionary of SIDs that map to each UniProt ID. {(UniProt, [SID1, ..., SIDN]), ...}

2b. Find corresponding InChI Key for Each SID
From the PDB ID list, find the corresponding InChI for each substance if it exists. Otherwise remove SID.
{(UniProt, [{SID1, InChI1}, ..., {SIDN, InChIN}]), ...}

Phase III: InChi + UniProt in PDB

Collect all pairs of InChi and UniProt. For pairs yielding a result, keep in the database. If there is no result from PDB, remove the UniProt from the dictionary.

- Tested Substance Count: > 10
- BioAssay Type: Confirmatory

Phase IV: Outputting the Data

Output the remaining pairs to the user. These pairs have corresponding 3D structures and assays.

Key Output Data Format

AssayID	UniProt ID	InChi Key	PDB ID

ADDRESSING THE CHALLENGES

Output

- Association between Protein (UniProt ID) and Assays (AIDs)

	UniProt	InChI	PubChem CID	AIDs
0	Q92731	YHEHVRSGKUYDON-UHFFFAOYSA-N	6102690	[977608, 243121, 1811, 242135, 250911]
1	Q92731	NSSOSHDCWCMNDM-UHFFFAOYSA-N	6102691	[977608, 250895, 243121, 1811, 242135]
2	Q92731	TZBJGXHYKVUXJN-UHFFFAOYSA-N	5280961	[1800065, 70660, 70021, 293384, 102409, 102410, 231689, 231690, 68744, 255369, 336145, 1410193, 1410194, 257300, 257301, 1811, 231319, 70170, 1186205, 70174, 1186207, 300319, 241825, 244129, 300322, 262948, 1797922, 231463, 262952, 262953, 240680, 262955, 262957, 250672, 242353, 243121, 566838, 254777, 244155, 1855804, 977608, 274384, 244177, 274385, 242135, 1077852, 297565, 1077853, 1797856, 70504, 70505, 70506, 625259, 70507, 1127149, 292715, 70511, 1797995, 1797996, 70514, 70515, 361463, 70520, 235514, 70654]
3	Q92731	MASYAWHPJCQLSW-ZIAGYGMSSA-N	446849	[70665, 232588]
4	Q92731	XIESSJVMWNJCGZ-VKJFTORMSA-N	10286159	[301349, 1797927, 1797928, 1797929, 977610, 1797930, 1797931, 273104, 273105, 300595, 273108, 273109, 273110, 1811, 300826]
5	Q92731	RHQLNMNKTIOREN-AOIWGVFYSA-N	9927355	[301349, 977610, 1797931, 1797930, 1811, 300826]
6	Q92731	GPFRMIHXGMVMGF-BZSNNMDCSA-N	16758226	[1811, 977610, 1797931, 301349]
7	Q92731	QJSMFUTULGSHNQ-ZOBUZTSGSA-N	11197931	[301349, 1797929, 977610, 1797931, 300595, 1811]

Thanks, any questions?