Alex Perez, Dan Krasnonosenkikh, Daniel Margolis, Sidney LaFontaine, Hari Muralikrishnan

BioLink/PubChem Group - Final Report:

Motivation:

BioLink is a biological database that leverages Neo4J's functionality to create a graph of biological data, complete with relationships between different types of biological information. BioLink currently features genes and diseases, as well as the relationships between them, and is working on expanding other biological entities into the graph. In its completed state, BioLink will serve as an information exploration/retrieval tool that will allow researchers to discover interesting relationships between different biological entities. The final product will also feature a UI that will have completely abstracted away Neo4J, allowing for querying without knowledge of Cypher.

PubChem is the world's largest chemical database. Operated by the United States National Institute of Health, PubChem offers API access to millions of chemical compounds and bioassays. While there is plenty of data in PubChem, it is hard to traverse, and the relational components of compounds and bioassays are abstracted out by a middle-man class of data called a substance. A substance is essentially a compound that was tested on by a specific bioassay, and the existence of substances in PubChem slows down a user's ability to understand the relationships between bioassays and compounds. If this data were to be imported into a graph database, the connectivity of compounds and bioassays to each other (and even genes and diseases) would be much easier to visualize and draw conclusions from.

Project Goals:

In this project, we aimed to add node types and associations from the data found in PubChem. This database was primarily focused around three major biological entities: the substance, the compound, and the bioassay. Contributors to the PubChem database can add information about a particular chemical substance, which are the Substance entities. But different contributors could upload slightly different information for one specific chemical structure, so there ends up being multiple substance records for one actual substance. This is where the compound entity comes in, by merging the most important content from all of these substance records into one normalized representation of that chemical structure. Knowing all this, we decided to focus on the compound entity over each specific substance entity. An additional goal was to extract from PubChem details about bioassays, which represent tests of the compound entities. Once we have these two entities from PubChem, we planned to make sure to identify the specific associations between them. Finally, a major goal for our project was to make sure that all these new node types and links were not only added into the BioLink database but also integrated by additionally including links from all this new data back to the existing data. We aimed to do this by finding connections between bioassays and genes, the latter being an already prevalent feature in the database.

Because PubChem houses data for millions of compounds/bioassays, we chose to import only the first 3000 compounds and their associated bioassays. Because we were not importing all the data, it was important to make sure that our work was easily extensible, so that, in the future, more PubChem data could be easily imported using our extensions to the existing BioLink import tool, provided the data was formatted as a CSV, as we made sure our data was.

Significance:

The reason this project is significant is because it lets us leverage graph databases in order to derive new associations that are otherwise not easily accessible. In the context of the project, we can easily understand far-away relationships between compounds and the respective genes they interact with by using their common factor in bioassays. By tying in more information with what BioLink already has in the form of proteins and diseases, we can help the medical field more quickly understand various causes of diseases.

Additionally, this platform helps serve as a multi-modal database that can support various types of queries within its domain. As expected, by adding support for an even greater domain, we can help expand not only the number of features we can query about but also the types of complexity of the queries themselves.

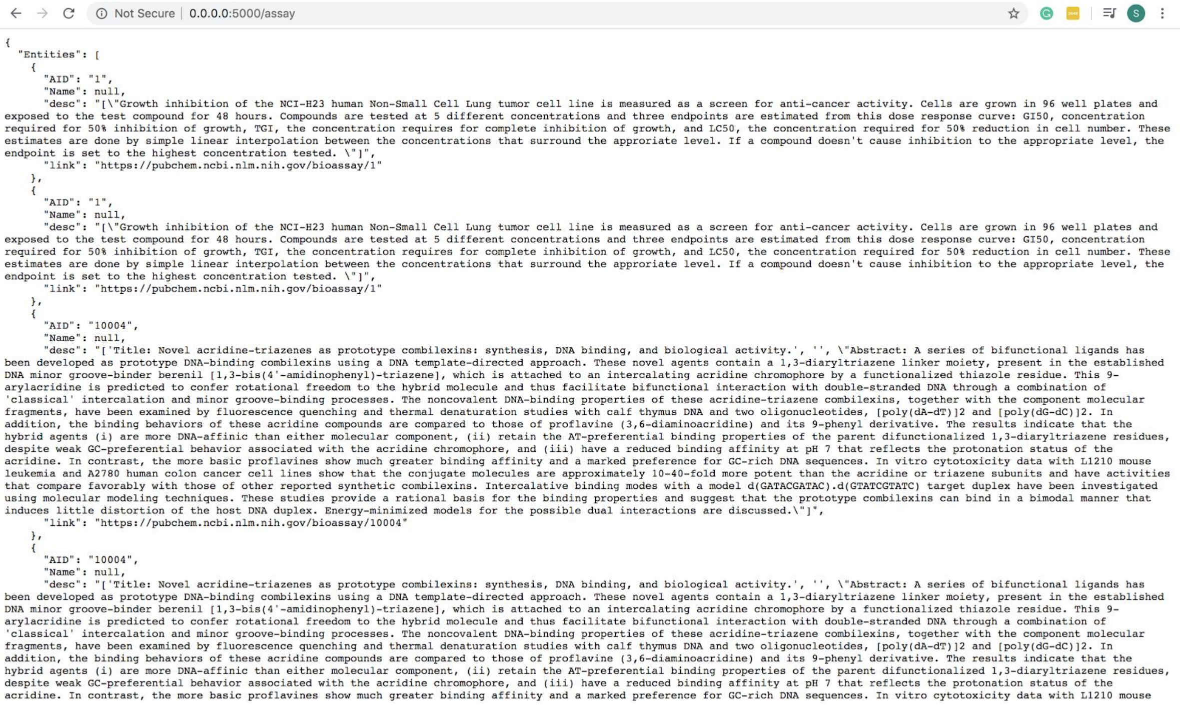
As a final result, now that we have added compounds and their associated bioassays, we can understand the effects of compounds on various diseases - since compounds are tested by specific bioassays, which interact with genes, which are associated with diseases.

Methodology:

We used the PubChemPy API and the PubChem REST API to load our data. We ran a python script to load 3000 compounds, find and load each compound's associated bioassays, and locate associated genes to those bioassays. We then exported the data on compounds, assays, genes, and the interconnections between them as CSVs to upload via a slightly modified version of the existing BioLink data import tool (specifically the config.json and db.py files).

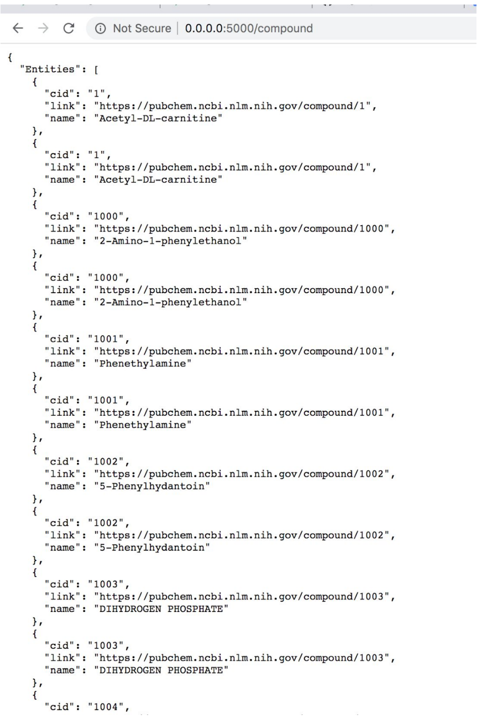
Results

After we wrote code to extract the information present in the PubChem and added the data to the BioLink database, here are the results of our successful work. We specifically show in the screenshots below how we added bioassays and compounds to the backend of the database, how we added different edge types (InteractsWith and TestsFor) to the biolink graph database, and how these changes translate to the front end querying:



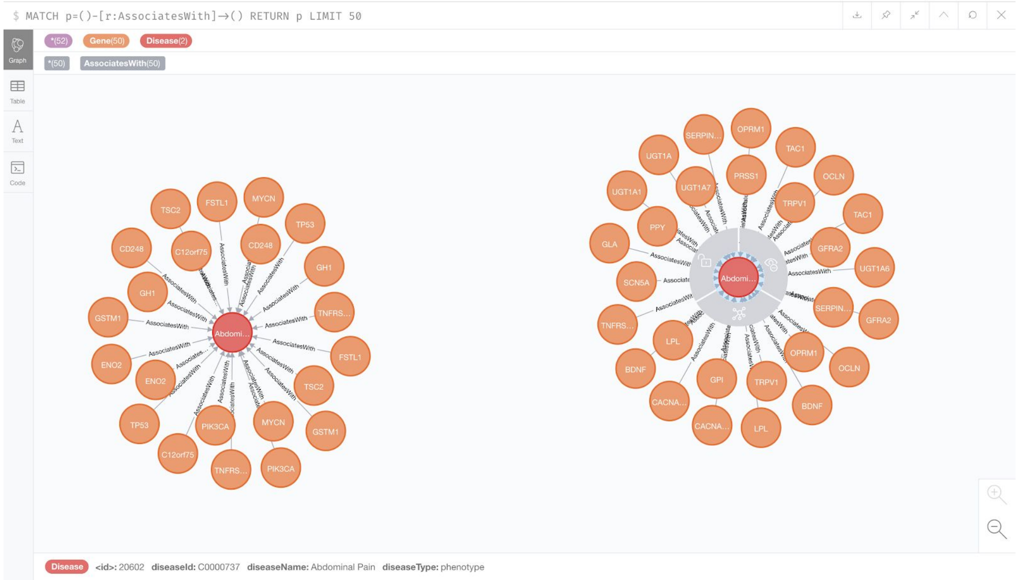
```
{
  "Entities": [
    {
      "AID": "1",
      "Name": null,
      "desc": "[\"Growth inhibition of the NCI-H23 human Non-Small Cell Lung tumor cell line is measured as a screen for anti-cancer activity. Cells are grown in 96 well plates and
exposed to the test compound for 48 hours. Compounds are tested at 5 different concentrations and three endpoints are estimated from this dose response curve: GI50, concentration
required for 50% inhibition of growth, TGI, the concentration requires for complete inhibition of growth, and LC50, the concentration required for 50% reduction in cell number. These
estimates are done by simple linear interpolation between the concentrations that surround the approriate level. If a compound doesn't cause inhibition to the appropriate level, the
endpoint is set to the highest concentration tested. \"]",
      "link": "https://pubchem.ncbi.nlm.nih.gov/bioassay/1"
    },
    {
      "AID": "1",
      "Name": null,
      "desc": "[\"Growth inhibition of the NCI-H23 human Non-Small Cell Lung tumor cell line is measured as a screen for anti-cancer activity. Cells are grown in 96 well plates and
exposed to the test compound for 48 hours. Compounds are tested at 5 different concentrations and three endpoints are estimated from this dose response curve: GI50, concentration
required for 50% inhibition of growth, TGI, the concentration requires for complete inhibition of growth, and LC50, the concentration required for 50% reduction in cell number. These
estimates are done by simple linear interpolation between the concentrations that surround the approriate level. If a compound doesn't cause inhibition to the appropriate level, the
endpoint is set to the highest concentration tested. \"]",
      "link": "https://pubchem.ncbi.nlm.nih.gov/bioassay/1"
    },
    {
      "AID": "10004",
      "Name": null,
      "desc": "['Title: Novel acridine-triazenes as prototype combilexins: synthesis, DNA binding, and biological activity.', '', \"Abstract: A series of bifunctional ligands has
been developed as prototype DNA-binding combilexins using a DNA template-directed approach. These novel agents contain a 1,3-diaryltriazene linker moiety, present in the established
DNA minor groove-binder berenil [1,3-bis(4'-amidinophenyl)-triazene], which is attached to an intercalating acridine chromophore by a functionalized thiazole residue. This 9-
arylacridine is predicted to confer rotational freedom to the hybrid molecule and thus facilitate bifunctional interaction with double-stranded DNA through a combination of
'classical' intercalation and minor groove-binding processes. The noncovalent DNA-binding properties of these acridine-triazene combilexins, together with the component molecular
fragments, have been examined by fluorescence quenching and thermal denaturation studies with calf thymus DNA and two oligonucleotides, [poly(dA-dT)]2 and [poly(dG-dC)]2. In
addition, the binding behaviors of these acridine compounds are compared to those of proflavine (3,6-diaminoacridine) and its 9-phenyl derivative. The results indicate that the
hybrid agents (i) are more DNA-affinic than either molecular component, (ii) retain the AT-preferential binding properties of the parent difunctionalized 1,3-diaryltriazene residues,
despite weak GC-preferential behavior associated with the acridine chromophore, and (iii) have a reduced binding affinity at pH 7 that reflects the protonation status of the
acridine. In contrast, the more basic proflavines show much greater binding affinity and a marked preference for GC-rich DNA sequences. In vitro cytotoxicity data with L1210 mouse
leukemia and A2780 human colon cancer cell lines show that the conjugate molecules are approximately 10-40-fold more potent than the acridine or triazene subunits and have activities
that compare favorably with those of other reported synthetic combilexins. Intercalative binding modes with a model d(GATACGATAC).d(GTATCGTATC) target duplex have been investigated
using molecular modeling techniques. These studies provide a rational basis for the binding properties and suggest that the prototype combilexins can bind in a bimodal manner that
induces little distortion of the host DNA duplex. Energy-minimized models for the possible dual interactions are discussed.\"]",
      "link": "https://pubchem.ncbi.nlm.nih.gov/bioassay/10004"
    },
    {
      "AID": "10004",
      "Name": null,
      "desc": "['Title: Novel acridine-triazenes as prototype combilexins: synthesis, DNA binding, and biological activity.', '', \"Abstract: A series of bifunctional ligands has
been developed as prototype DNA-binding combilexins using a DNA template-directed approach. These novel agents contain a 1,3-diaryltriazene linker moiety, present in the established
DNA minor groove-binder berenil [1,3-bis(4'-amidinophenyl)-triazene], which is attached to an intercalating acridine chromophore by a functionalized thiazole residue. This 9-
arylacridine is predicted to confer rotational freedom to the hybrid molecule and thus facilitate bifunctional interaction with double-stranded DNA through a combination of
'classical' intercalation and minor groove-binding processes. The noncovalent DNA-binding properties of these acridine-triazene combilexins, together with the component molecular
fragments, have been examined by fluorescence quenching and thermal denaturation studies with calf thymus DNA and two oligonucleotides, [poly(dA-dT)]2 and [poly(dG-dC)]2. In
addition, the binding behaviors of these acridine compounds are compared to those of proflavine (3,6-diaminoacridine) and its 9-phenyl derivative. The results indicate that the
hybrid agents (i) are more DNA-affinic than either molecular component, (ii) retain the AT-preferential binding properties of the parent difunctionalized 1,3-diaryltriazene residues,
despite weak GC-preferential behavior associated with the acridine chromophore, and (iii) have a reduced binding affinity at pH 7 that reflects the protonation status of the
acridine. In contrast, the more basic proflavines show much greater binding affinity and a marked preference for GC-rich DNA sequences. In vitro cytotoxicity data with L1210 mouse
```

This is a picture of how bioassays, which we added to the biolink database, are stored in the backend of the biolink application. A bioassays here is made up of a bioassay ID (AID), name, and in-depth description (desc).



```
{
  "Entities": [
    {
      "cid": "1",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1",
      "name": "Acetyl-DL-carnitine"
    },
    {
      "cid": "1",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1",
      "name": "Acetyl-DL-carnitine"
    },
    {
      "cid": "1000",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1000",
      "name": "2-Amino-1-phenylethanol"
    },
    {
      "cid": "1000",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1000",
      "name": "2-Amino-1-phenylethanol"
    },
    {
      "cid": "1001",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1001",
      "name": "Phenethylamine"
    },
    {
      "cid": "1001",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1001",
      "name": "Phenethylamine"
    },
    {
      "cid": "1002",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1002",
      "name": "5-Phenylhydantoin"
    },
    {
      "cid": "1002",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1002",
      "name": "5-Phenylhydantoin"
    },
    {
      "cid": "1003",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1003",
      "name": "DIHYDROGEN PHOSPHATE"
    },
    {
      "cid": "1003",
      "link": "https://pubchem.ncbi.nlm.nih.gov/compound/1003",
      "name": "DIHYDROGEN PHOSPHATE"
    },
    {
      "cid": "1004",
```

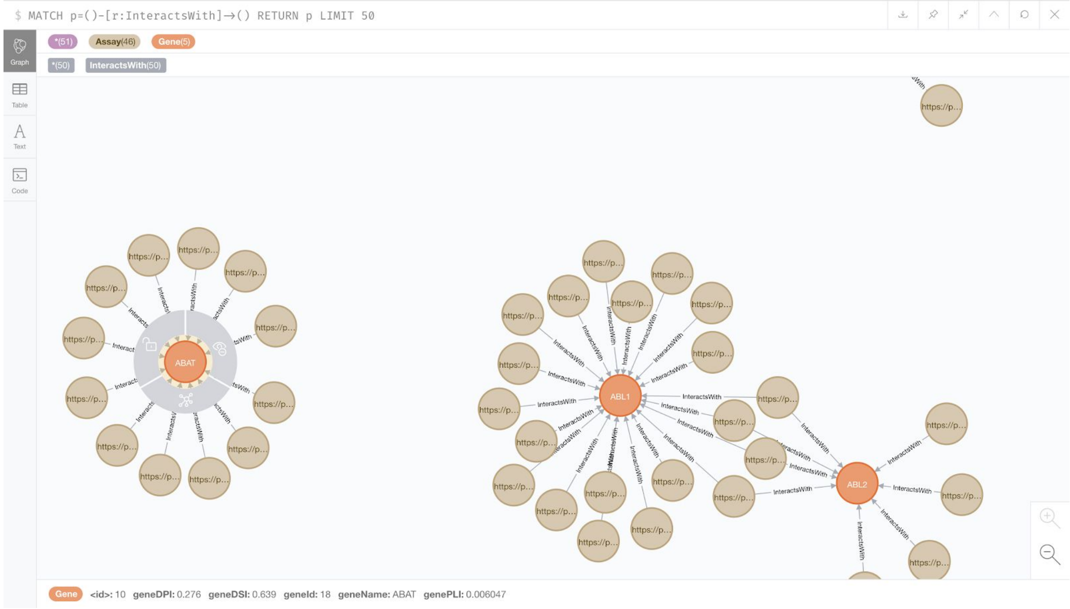
This is a picture of how compounds, which we added to the biolink database, are stored in the backend of the biolink applications. A compound here is made up of a compound ID (cid), name, and link to more information about the compound (link).



```
$ MATCH p=()-[r:AssociatesWith]→() RETURN p LIMIT 50
```

Disease <id>: 20602 **diseaseId:** C0000737 **diseaseName:** Abdominal Pain **diseaseType:** phenotype

This is a picture of how genes are linked to diseases, the AssociatesWith edge. These edge and node types already existed in biolink before our project, but since we connect bioassays and compounds to genes we wanted to show how those gene nodes are connected to diseases to start off.



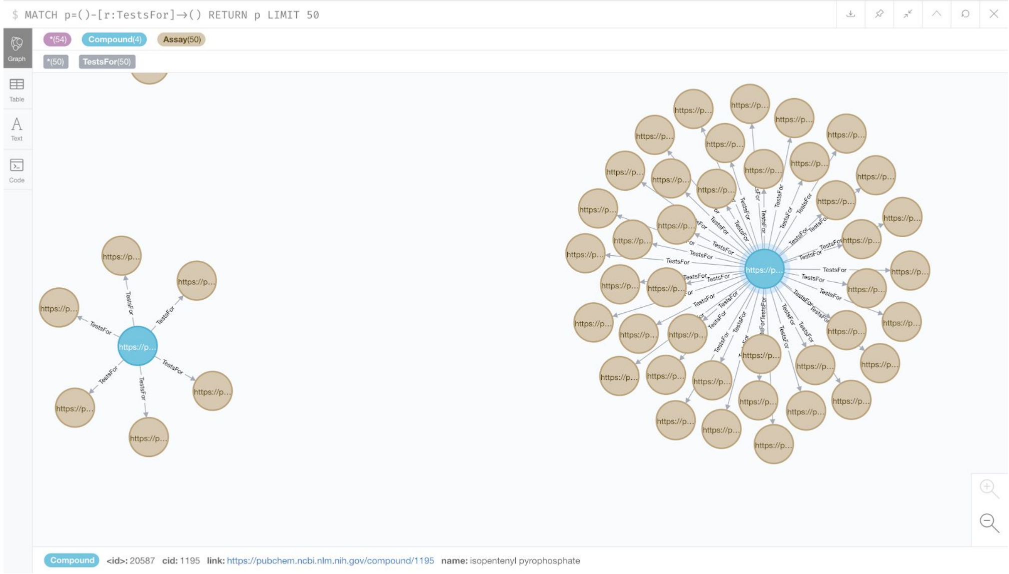
```
$ MATCH p=()-[r:InteractsWith]→() RETURN p LIMIT 50
```

*(51) Assay(46) Gene(5)

*(50) InteractsWith(50)

Gene <id>: 10 geneDPI: 0.276 geneDSI: 0.639 geneId: 18 geneName: ABAT genePLI: 0.006047

This is a picture of how bioassays are linked to genes, the InteractsWith edge. We added the bioassay node types to the biolink database and were able to create this edge type through the data pubchem provides on bioassay to gene interactions.



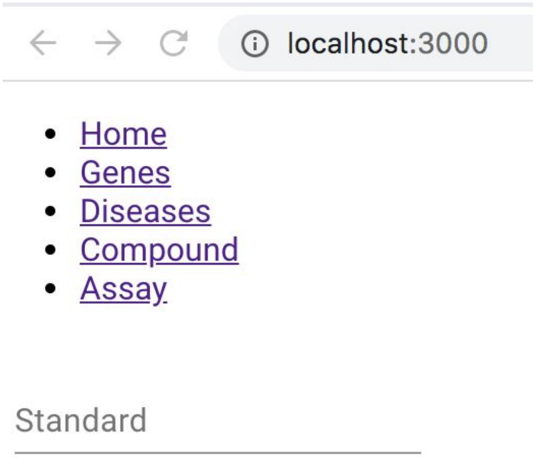
```
$ MATCH p=()-[r:TestsFor]→() RETURN p LIMIT 50
```

*(54) Compound(4) Assay(50)

*(50) TestsFor(50)

Compound <id>: 20587 cid: 1195 link: https://pubchem.ncbi.nlm.nih.gov/compound/1195 name: isopentenyl pyrophosphate

This is a picture of how compounds are linked to bioassays, the TestsFor edge. We added the bioassay and compound node types to the biolink database and were able to create this edge type through the data pubchem provides on compound to bioassay interactions.



This is a picture of the front-end of the biolink application, done in React. We added the compound and assay links.

Conclusion

Ultimately, our project goals were successfully met in this final project. The goals we accomplished were understanding the complexities of PubChem and gaining some primitive domain knowledge on the biological data it stored, then formulating a procedure for how to extract the data we felt was important to add to the BioLink database through the use of the API's, and then we updated the files that import data directly into the database in a scalable way. In the future, contributors to BioLink can leverage our work by adding in more data from PubChem relating to bioassays and compounds, as well as their interactions, to gain a more complete understanding of how bioassays and compounds interact with different genes and diseases.