Linking Proteins to Drugs

Functional RDF Utilization

Brian Kirk
Rice University
07/29/10

Introduction

- Scientific methodology
- Preliminary results
- Implementation
- Challenges
- Conclusions

Working Collaboration

- Emily Doughty and Jonathan Mortensen (who contributed the version of NDF-RT that we are using)
- The subsequent research was produced as a joint exercise between Emily and myself in all aspects including planning, creation of system, and application of the research
- In the interest of creating this presentation, I will be presenting on the building aspects of the project, while later Emily will present on some of the findings

Project Goals

- To find relationships between Drug Properties and Protein Features
  - Some research questions
    - How closely do drugs interact with proteins containing the same domain classes as with other known drug targets?
    - How closely do drugs that treat the same disease relate to proteins that are known to cause said disease?
    - How often do known mutations correspond to domains within the literature?
    - How well does the presence of a drug target within a given molecular pathway correspond to noted methods of action of other drugs that affect the same pathway?

Project Goals

- To find relationships between Drug Properties and Protein Features
  - Some research questions
    - How closely do drugs interact with proteins containing the same domain classes as with other known drug targets?
    - How closely do drugs that treat the same disease relate to proteins that are known to cause said disease?
    - How often do known mutations correspond to domains within the literature?
    - How well does the presence of a drug target within a given molecular pathway correspond to noted methods of action of other drugs that affect the same pathway?

Pharmacogenetics

- Relationship between Genes (*Genetics)* and Drugs, (*Pharmacology)*

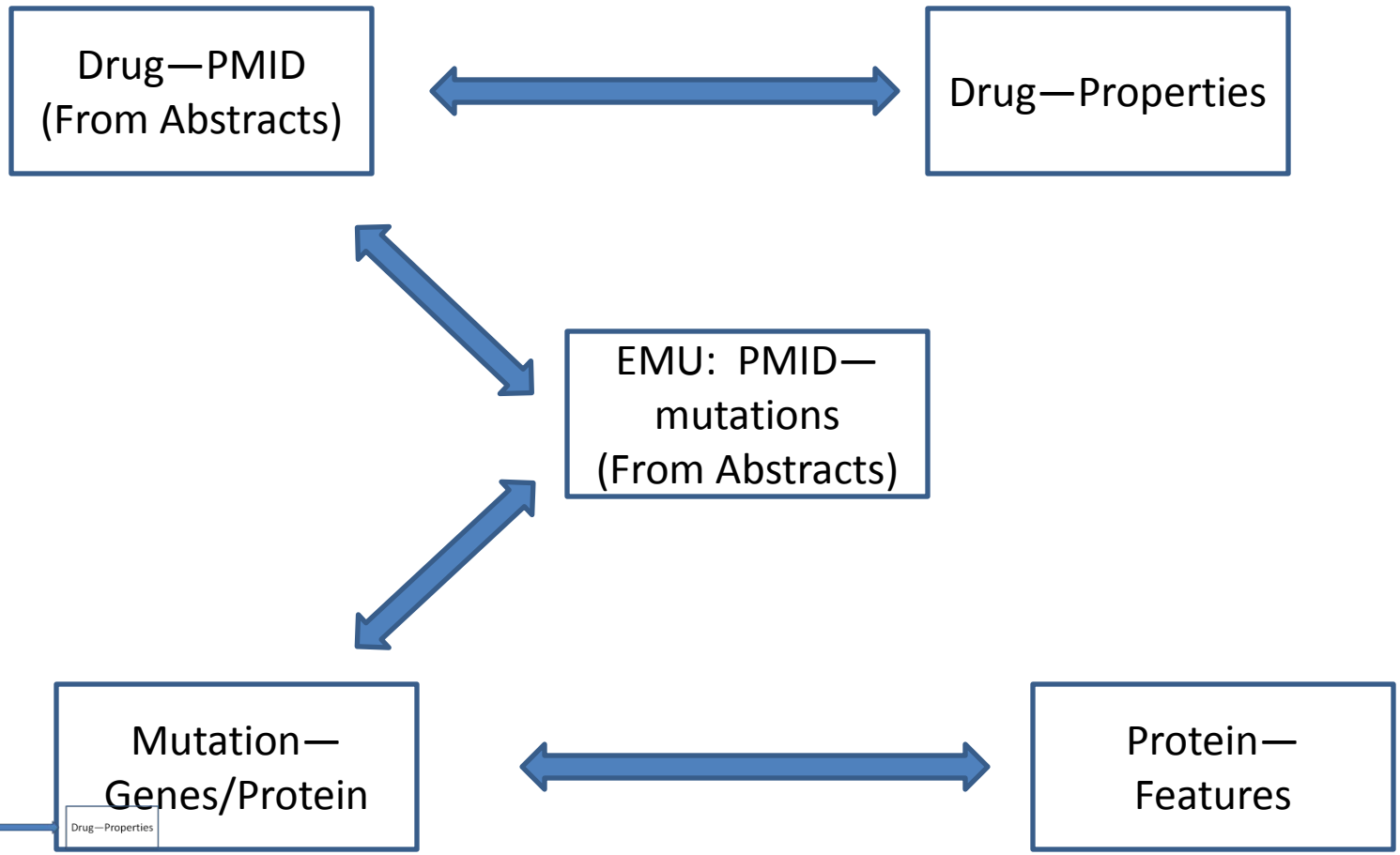
- Current state of the art reflects research conducted in:
  - Elucidated physiological effects
  - Measuring gene expression changes
  - Genotyping specific drug responses
    - Building Individualized Medicine: Prevention of Adverse Reactions to Warfarin Therapy, Krynetskiy, E., McDonnell P., JPET 322:427-434, 2007
    - Mutations on Cytochrome P450 2C9 affect ability to metabolize the drug; leads to needing lower dose

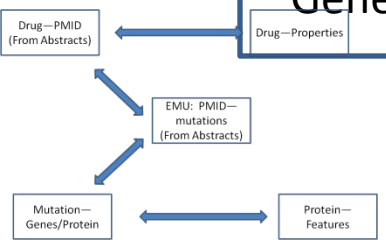
Previous Work: Text Mining

- EMU - (Developed by Emily and her lab)
  - Takes PubMed abstracts that have been pre-filtered by MeSH terms and identifies mentions of nucleotide or amino acid mutations and then verifies that that mutation is possible with the genes or proteins mentioned in the abstract

- Drug names from the same PubMed abstracts extracted by MetaMap and converted into RxNorm concepts

Design

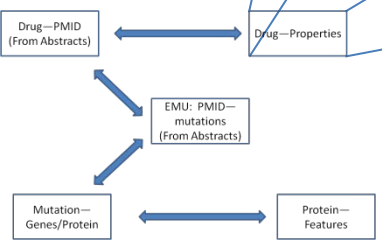

Drug—PMID
(From Abstracts)
Drug—Properties
EMU: PMID—
mutations
(From Abstracts)
Mutation—
Genes/Protein
Protein—
Features


Drug—PMID
(From Abstracts)
Drug—Properties
EMU: PMID—
mutations
(From Abstracts)
Mutation—
Genes/Protein
Protein—
Features

Drug Properties

Example: Drug "Warfarin"

drug_class:	Anticoagulants and Blood Products/modifiers/volume expanders
may_treat:	Thrombophlebitis, Thromboembolism, Atrial Fibrillation…
may_prevent:	Venous Thrombosis, Myocardial Infarction, Thromboembolism…
CI_with:	Eclampsia, Cerebral Hemorrhage, Ascorbic Acid Deficiency…
has_Ingredient:	Warfarin
has_MoA:	Vitamin K Epoxide Reductase Inhibitors
has_PE	Decreased Coagulation Factor Concentration
site_of_metabolism:	Hepatic Metabolism



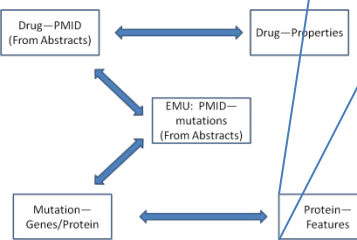
Protein Features

Pathways—Reactome and BioCyc describe ~1000 genomic and metabolic pathways

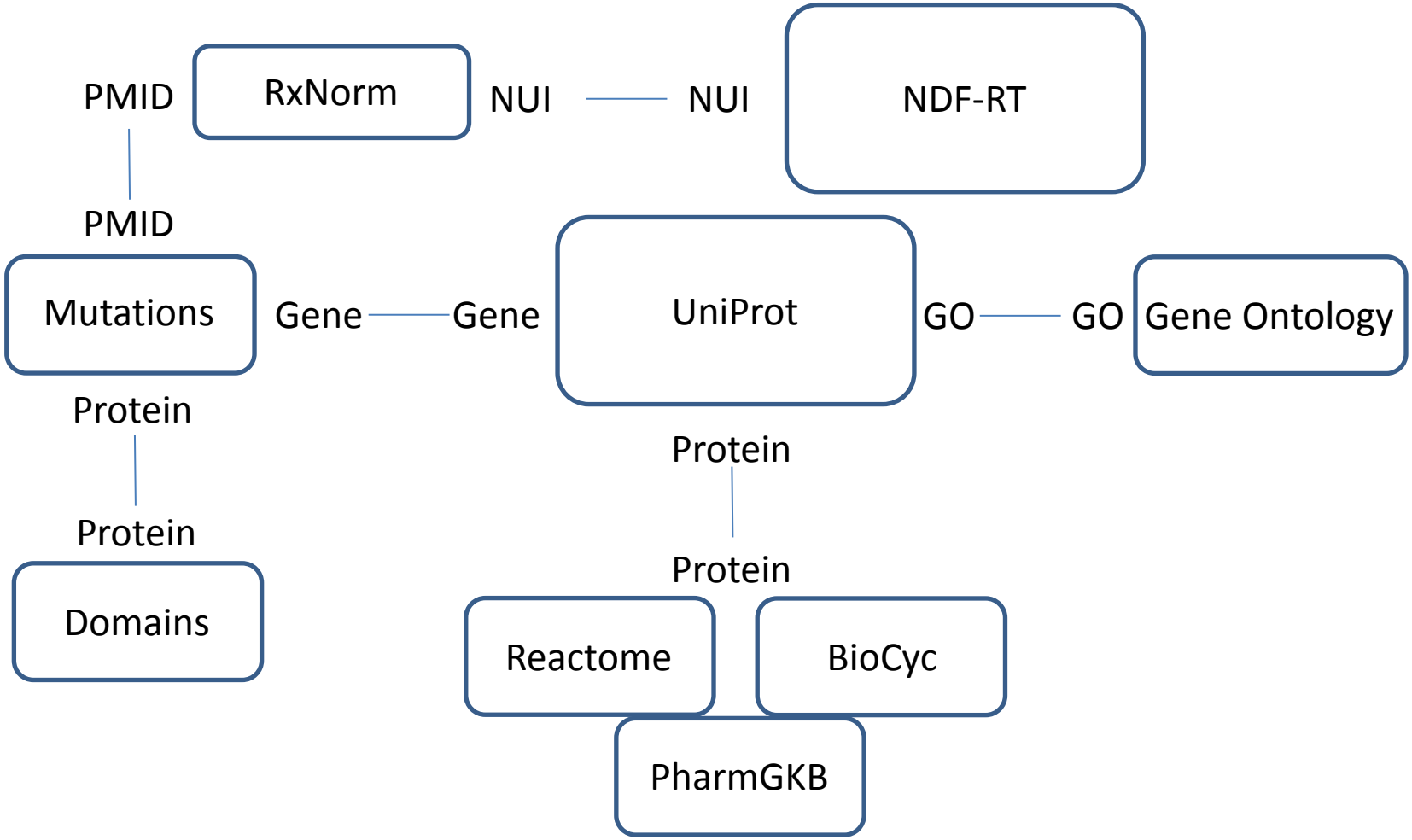
Functional properties—Gene Ontology describes the Biological Process, Cellular Component, Molecular Function

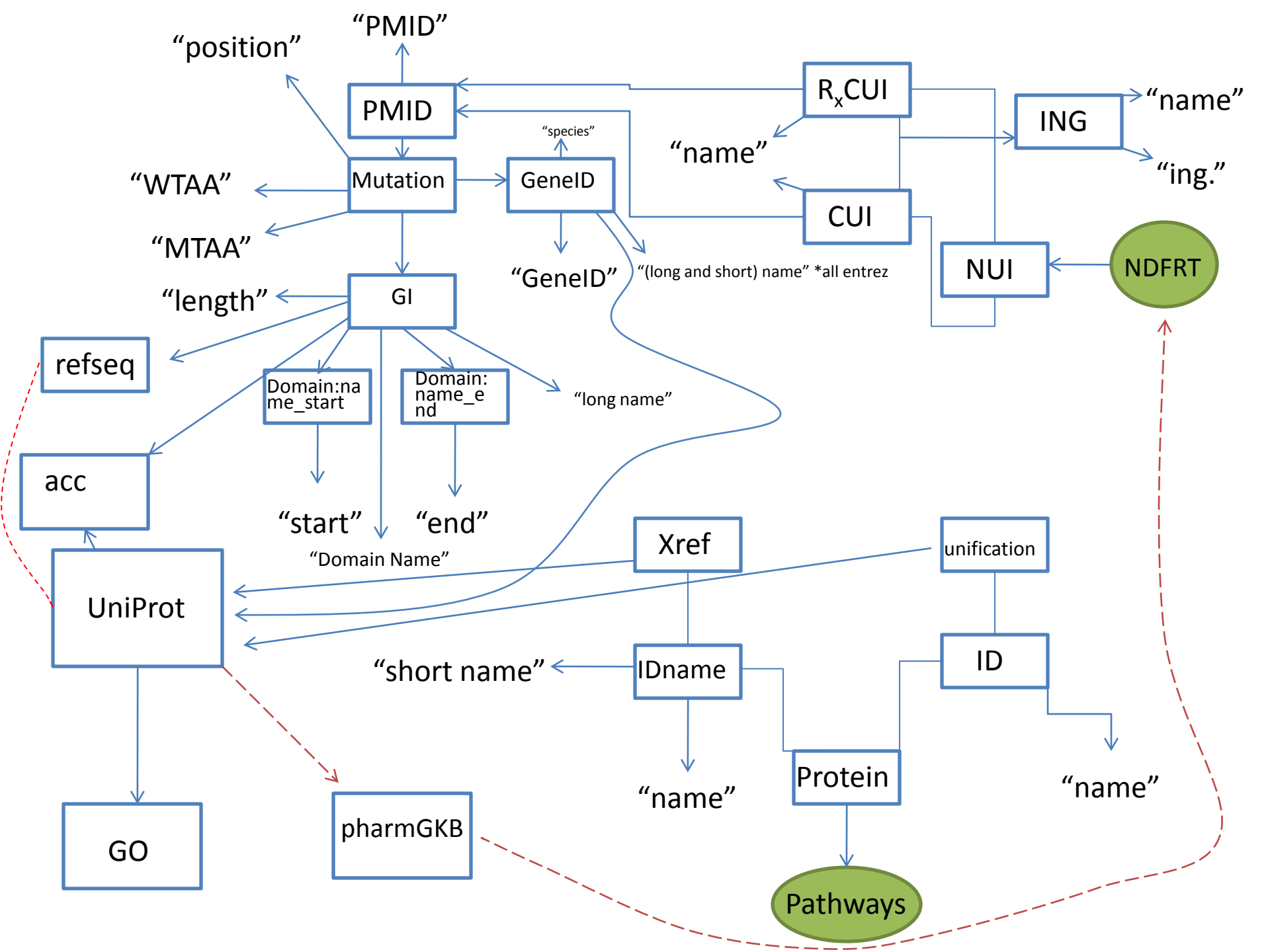
Disease—PharmGKB relates genes to drugs to diseases across ~6000 relationships

Domains—Pfam and NCBI Structure describe the structural and functional domains for proteins



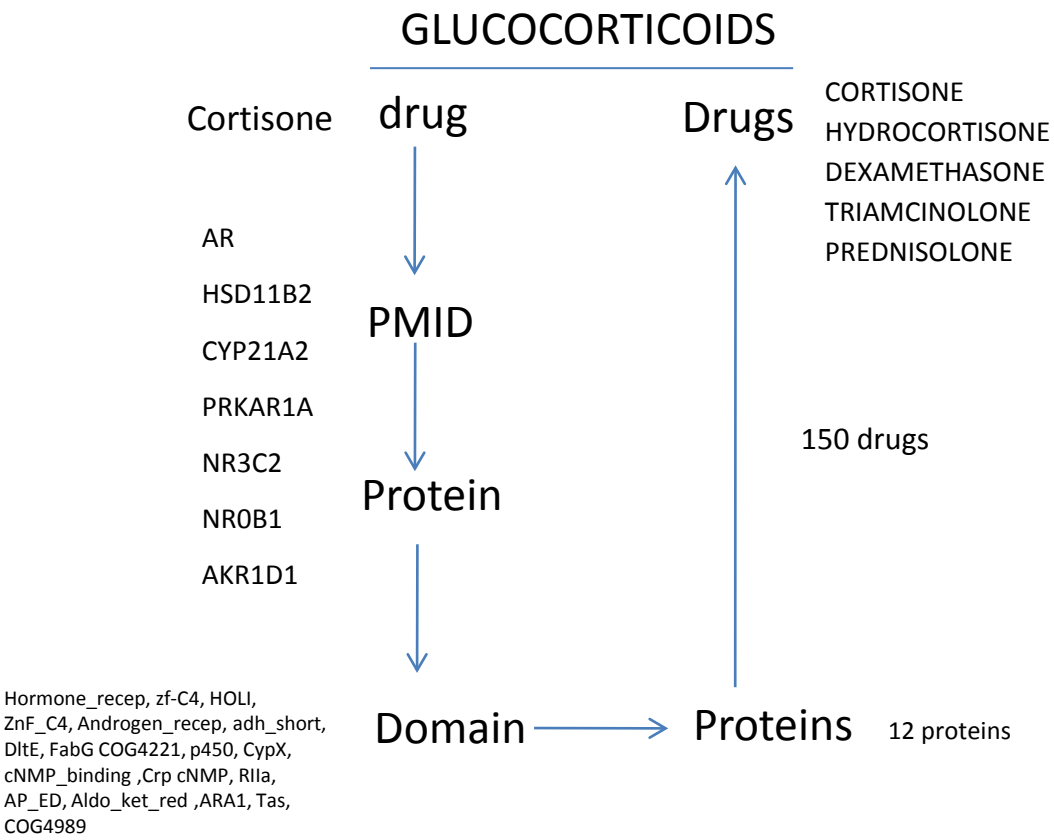
Traversing the graph




"PMID"
"position"
PMID
Mutation
"WTAA"
"MTAA"
"species"
GeneID
"GeneID"
"(long and short) name" *all entrez
GI
"length"
refseq
Domain:name_start
Domain:name_end
"long name"
acc
"start"
"end"
"Domain Name"
UniProt
GO
pharmGKB
Xref
unification
"short name"
IDname
ID
"name"
Protein
"name"
Pathways
RxCUI
"name"
CUI
ING
"name"
"ing."
NUI
NDFRT

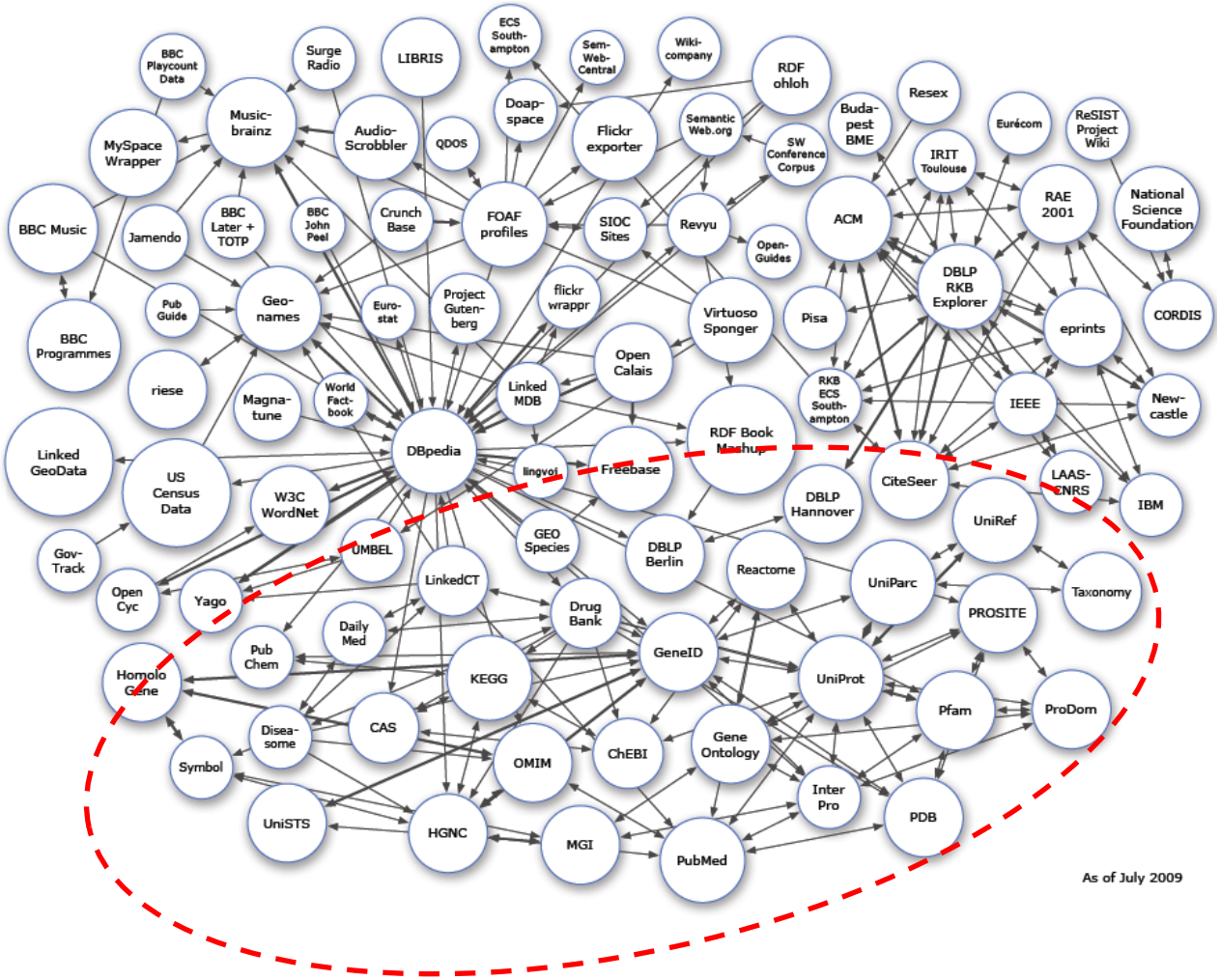
Preliminary Result

- Looking at the drug cortisone we find that proteins that are related to this drug and their domains

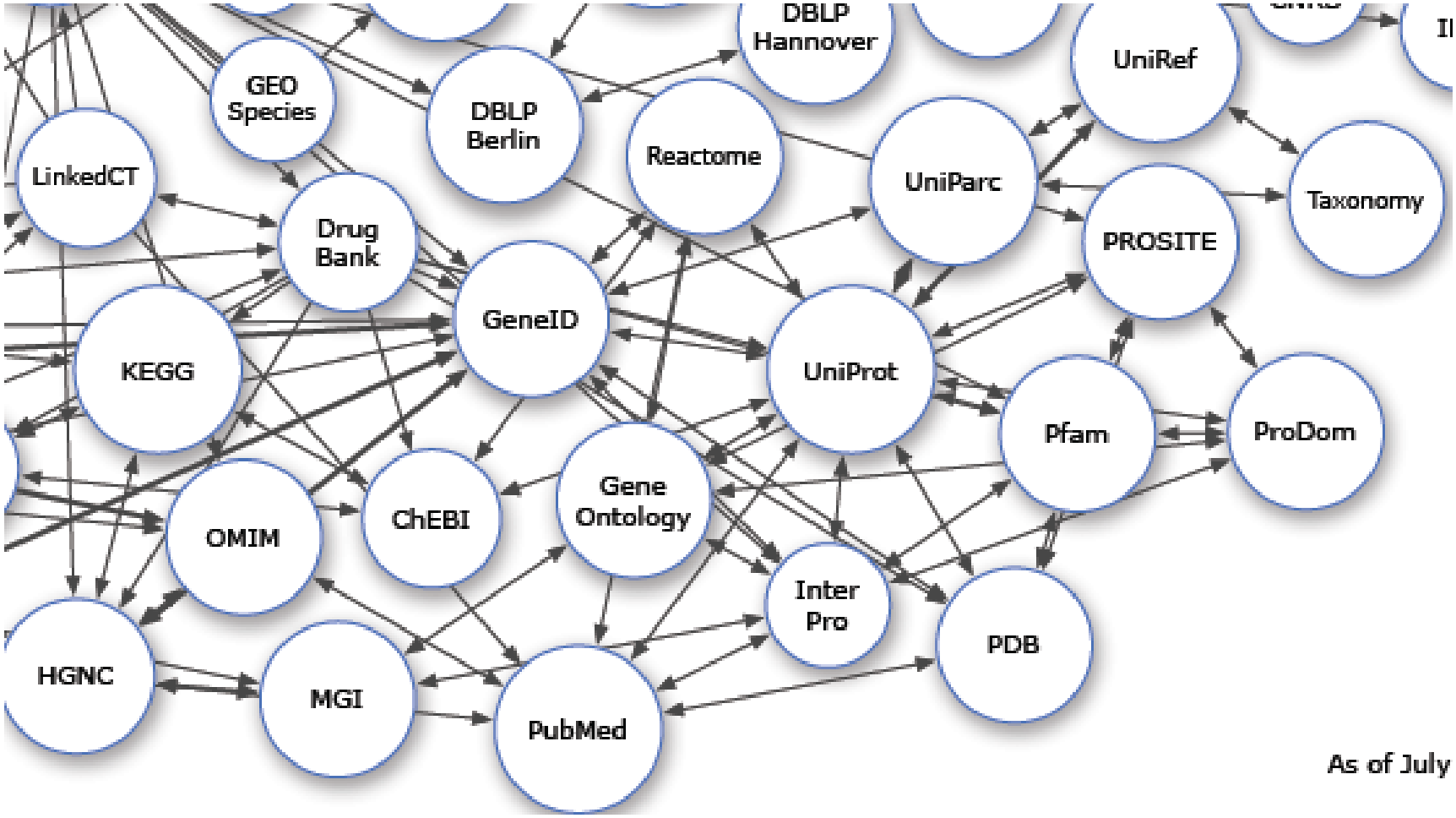


IMPLEMENTATION

Linked Data on the Web



Biological Resources



Resource Description Framework (RDF)

- **Semantic Web (Most commonly seen as RSS)**
  - **Subject – Predicate – Object Triples**

- **For Example: A table such as:**

Entrez Id	Ensembl Id	UniProt Id	Name	Symbol
1017	ENSG00000123374	P24941	cyclin-dependent kinase 2	CDK2
1019	ENSG00000135446	P11802	cyclin-dependent kinase 4	CDK4
1021	ENSG00000105810	Q00534	cyclin-dependent kinase 6	CDK6

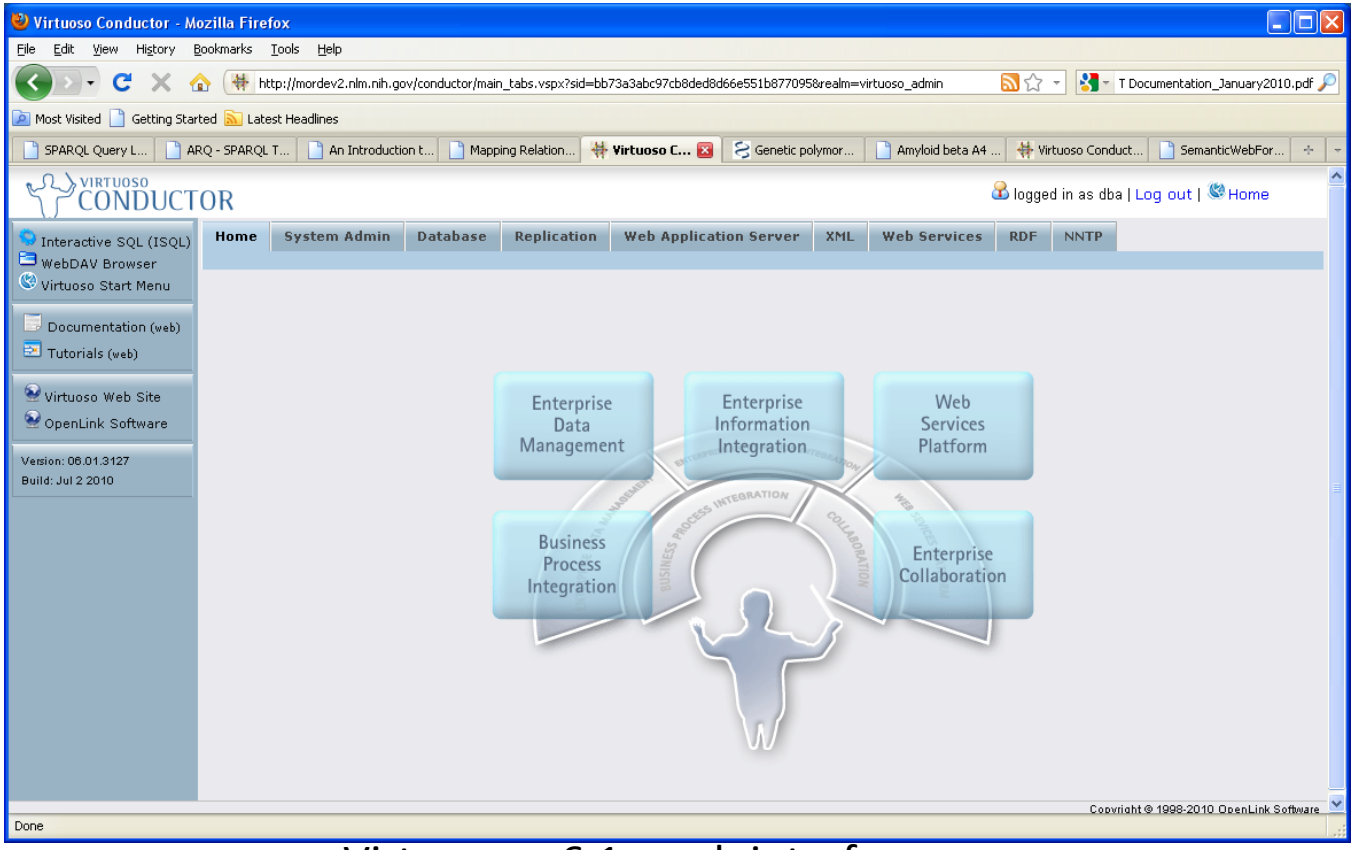
  - ***<URI_Entrez#1017> <URI#has_name> ("***cyclin-dependent kinase 2***" or <URI_uniprot#*P24941*>)***

    Subject Predicate Object
  - **URI: Uniform Resource Identifier**
    - Similar to URL, only identifies resource location
      - Example: http://purl.uniprot.org/uniprot

Triple Store

A Database manager that has been specifically designed to deal with indexing and storing RDF triples, can functionally deal with billions of triples (1.5 B)

Virtuoso allows web access, has good scalability, and is open source



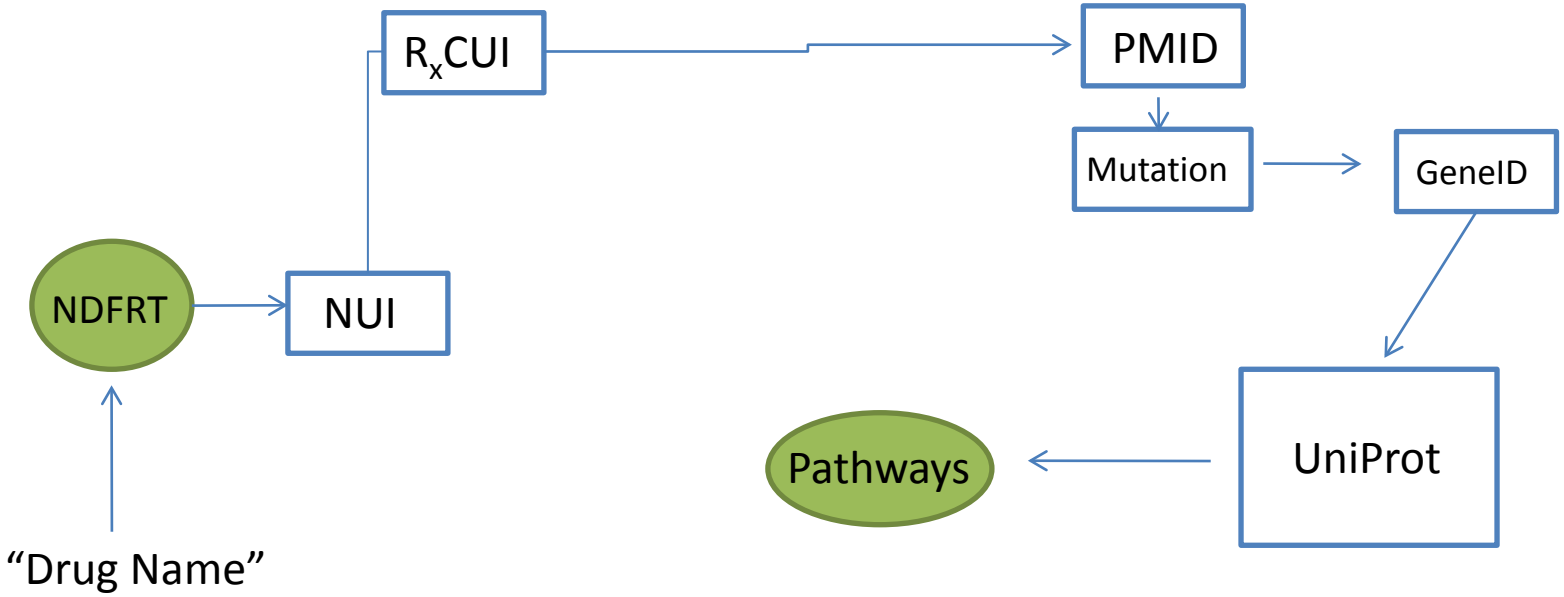
Virtuoso v. 6.1 –web interface

SPARQL Queries

```
select ?pmid ?mutation ?gene ?uniprot ?name ?path
from <http://localhost:8890/DAV/prot_domain>
from <http://localhost:8890/DAV/mutations>
from <http://localhost:8890/DAV/PMID_drug>
from <http://mor.nlm.nih.gov/ndfrt>
from <http://mor.nlm.nih.gov/uniprot>
from <http://localhost:8890/DAV/biopax>
from <http://mor.nlm.nih.gov/GO>
where{
        ?nui  <http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#RxNorm_Name>
                "Warfarin"^^<http://www.w3.org/2001/XMLSchema#string> .
        ?nui <http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#URI_RxNorm_CUI> ?rxcui.
        ?rxcui <http://www.biopax.org/release/biopax-level2.owl#XREF> ?pmid.
        ?pmid <http://mor.nlm.nih.gov#containsMutation> ?mutation .
        ?mutation <http://mor.nlm.nih.gov#hasGeneID> ?gene.
        ?uniprot <http://www.w3.org/2000/01/rdf-schema#seeAlso> ?gene.
        ?uniprot <http://purl.uniprot.org/core/reviewed> ?y.
        FILTER (?y >= 1).
        ?uniprot <http://purl.uniprot.org/core/recommendedName> ?e .
        ?e <http://purl.uniprot.org/core/fullName> ?name .
        ?uniprot <http://www.w3.org/2000/01/rdf-schema#seeAlso> ?reac.
        ?react <http://mor.nlm.nih.gov/Reactome> ?reac.
        ?path <http://www.biopax.org/release/biopax-level2.owl#XREF> ?react
}
```

SPARQL Query

Find the NDF-RT drug “Warfarin” and get all of the pathways that the genes it affects are in



pmid	mutation	gene	uniprot	name	pathway
11588061	1559.ARG144CYS	1559	P11712	Cytochrome P450 2C9	Biological_oxidations

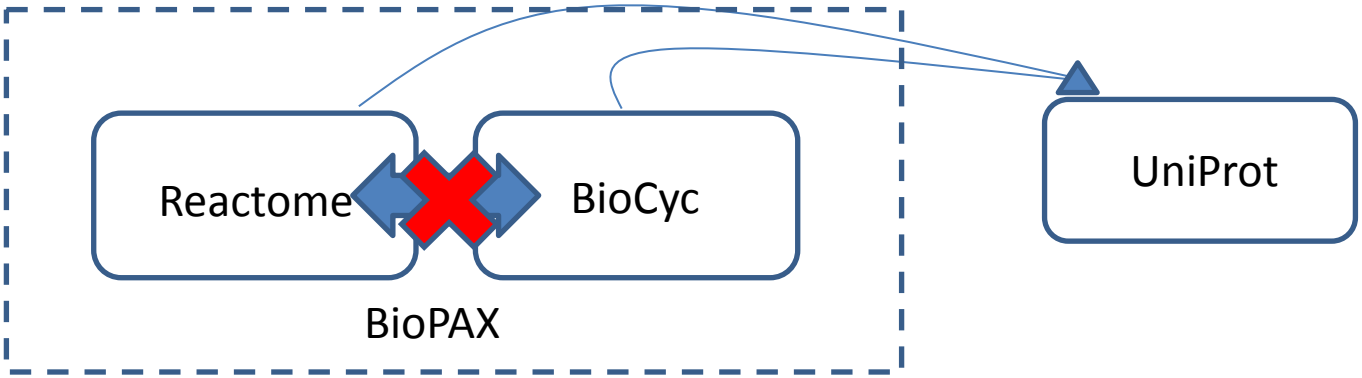
CHALLENGES

Challenges: Scalability

- Uniprot is a huge database
  - Has ~1.5 billion triples and exists as a file that is over 100 Gb
  - Loading it into Virtuoso, crashed our server, several times
  - Required upgrading to a more dedicated server with larger RAM
  - Optimization for speed is still required to accommodate large size

Challenges: Interoperability

- BioPAX's ontology schema includes Reactome and BioCyc
  - Each DB interprets the common ontology differently
    - The same architecture is used to describe the same concept differently, which requires leaving the ontology to find a unifying descriptor and then coming back
- The solution to this is that both DB's call all of their proteins in UniProt accessions



Challenges: Interoperability

- A protein can be identified in three ways
  - UniProt Accession
  - NCBI Protein GI
  - NCBI RefSeq (comes from sequence)
- All of the output from EMU is created into RefSeq accession format, while ~half of protein description data started as UniProt and half as RefSeq as a result of an attempt reaching completeness of genome

Solution

- All of the UniProt Accessions are converted into RefSeq NCBI Protein GI's to facilitate bridging mutations at Protein level
- For general traversing purposes, specific isoform information is lost as Entrez Gene ID's are converted into UniProt Accession
  - Isoforms can still be retrieved if purpose requires it
- RefSeq Accessions that have no UniProt equivalent are dead ends, but are retained in graph

Conclusions

- Our triple store can handle ~1.5 billion triples
- All nodes and edges link and traverse correctly
- Preliminary results
  - More results to be presented by Emily on Aug. 11
  - Results show that we can find the same class of drug interacting with the same types of domains, which is what you would expect

Future Directions

- To add additional functionality by including new resources, such as drug adverse events and a more general gene-to-disease knowledge base
- To submit our findings for the Pacific Symposium on Biocomputing workshop on Mining the Pharmacogenomics Literature

Acknowledgements

- I would like to thank the NLM its rich support while here
- Olivier Bodenreider, my mentor while here
- Emily Doughty and Jonathan Mortensen, who I have collaborated with
- Dr. Jianpeng Ma, my mentor at Rice
- May Cheh, for putting all of this together
- Dr. McDonald

Thank you

Things That I Have Learned

- When dealing with large quantities of data from disparate systems,

Project Rational

Existing web resources exist as Islands

While many resources are quite large, both in their scope and interlink ability with other databases, for example an Entrez Gene search result will let you see the presence of that gene within all of the other tools featured by NCBI in addition to multiple pathway and gene expression databases generated by others.

This allows the user the jump between to Islands and to search many Islands at once.

However, this process is not suited well for batch queries across multiple databases, or for taking the result from one database and seeing how relates to a number of different things within a second database.

Or for propagating a query across multiple DB's where the results from one DB become input for the next

Drug Properties

National Drug File – Reference Terminology (NDF-RT)

Created by the Veterans Health Administration, it is:

- A concept-oriented terminology
- Organized into hierarchies of concepts based on generalization
- A listing of drugs by their ingredients, method of action, diseases that are treated, physiologic effect, drug interactions, and drug class