Linking Proteins, Mutations, and Drugs

Applications of Data Integration using RDF

Emily Doughty
Collaborator: Brian Kirk
Mentor: Olivier Bodenreider
August 11, 2010

Outline

- Background on pharmacogenomics and our contribution to the field
- Brief overview of the framework and data integration of our system
- Validation applications and system traversals
- System limitations and future directions

Pharmacogenomics

Definition:

"The study of how variations in the human genome affect the response to medications"[1]

- Mutations are one of the variations existing between individual human genomes
- How does the existence of a mutation impact drugs or vice versa?
- How does the association with a mutation impact drugs on proteins?

1. http://www.medterms.com/script/main/art.asp?articlekey=15313

Relating Drugs, Mutations, and Proteins

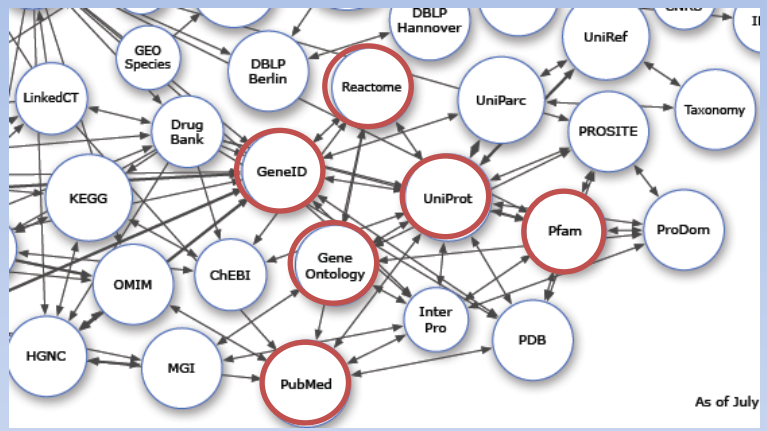
- Want to relate clinical drugs and mutations to protein structure and function

- Specifically, want to look at protein domains and pathways and relate to drug classes and properties

- Formulate new hypotheses for validation and future research

Framework

From Brian Kirk:

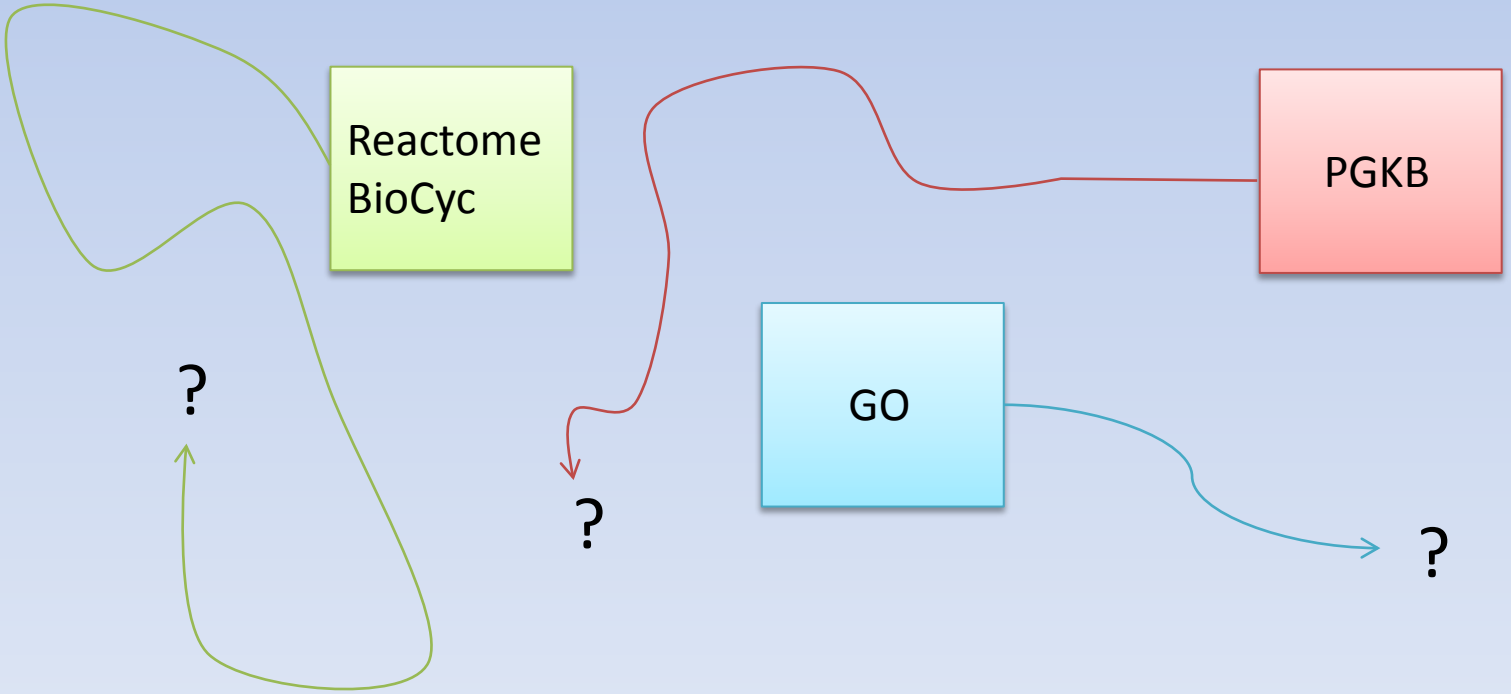


- Database Integration:
  - UniProt
  - NDF-RT
  - PharmGKB
  - Reactome and BioCyc
  - Gene Ontology
  - Protein to Domain table provided by Dr. Maricel Kann's lab at UMBC

- High-throughput method for text mining for drug-related mutations
  - Extractor of MUtations (EMU) developed at Dr. Kann's lab

Making Connections: Data Integration

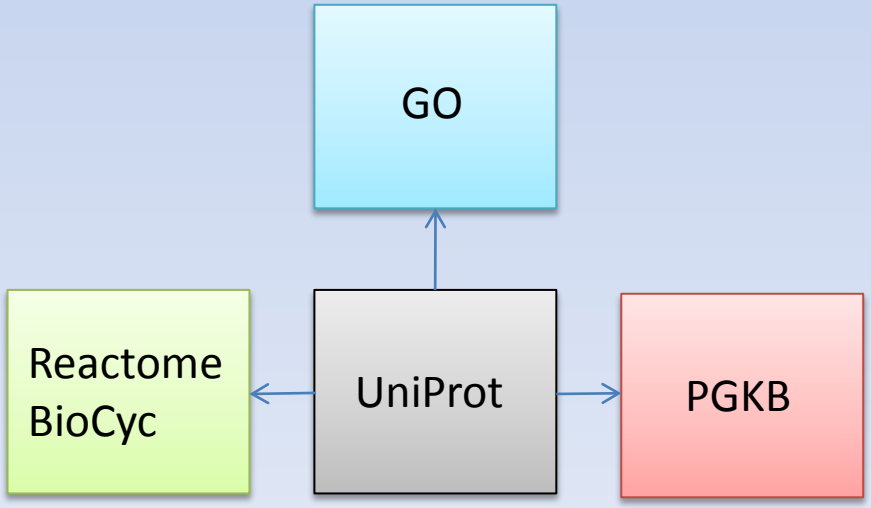
- Need to be able to connect database "pieces" together
- Solution: database identifiers and UniProt



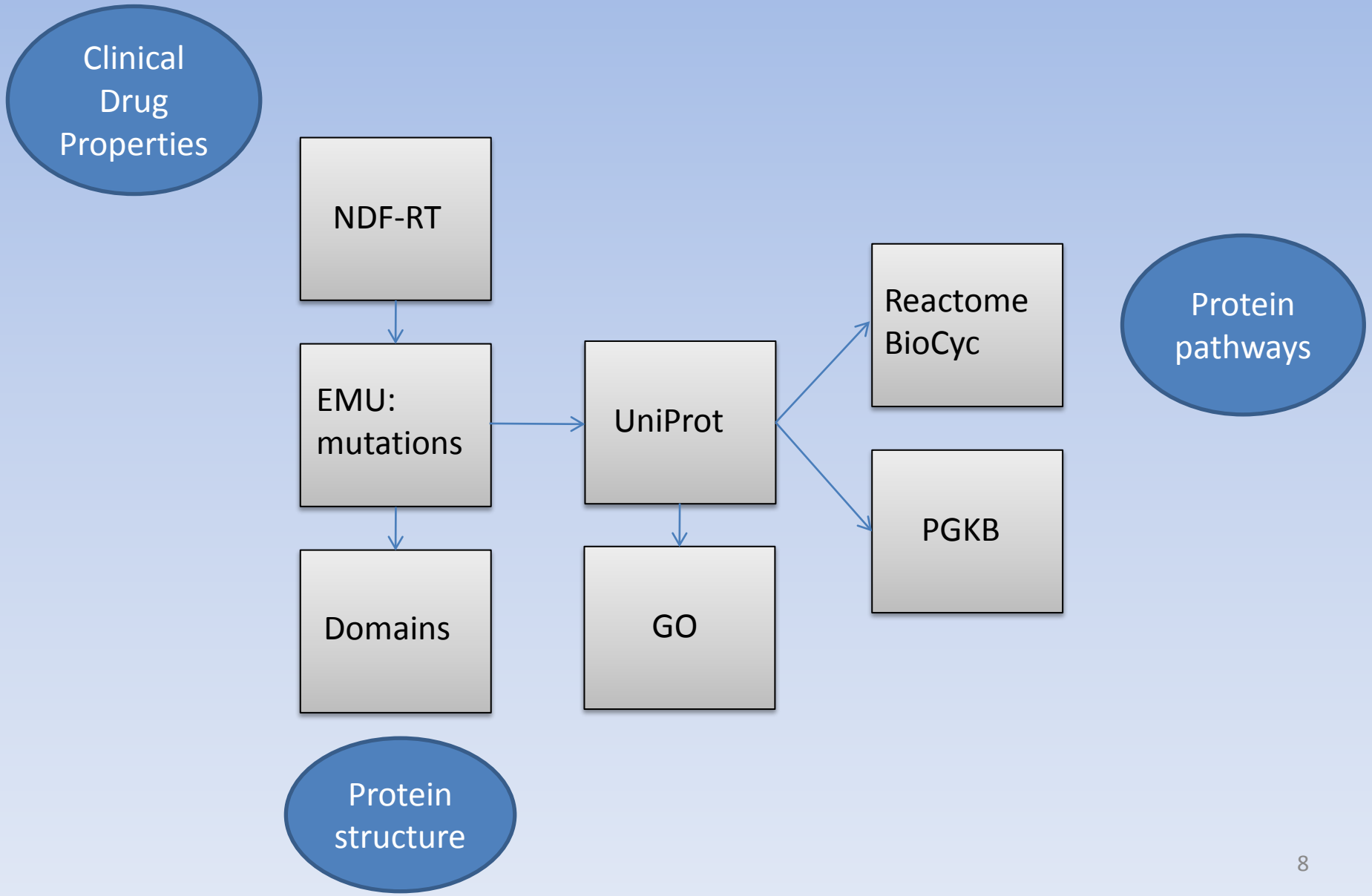
Making Connections: Data Integration

UniProt

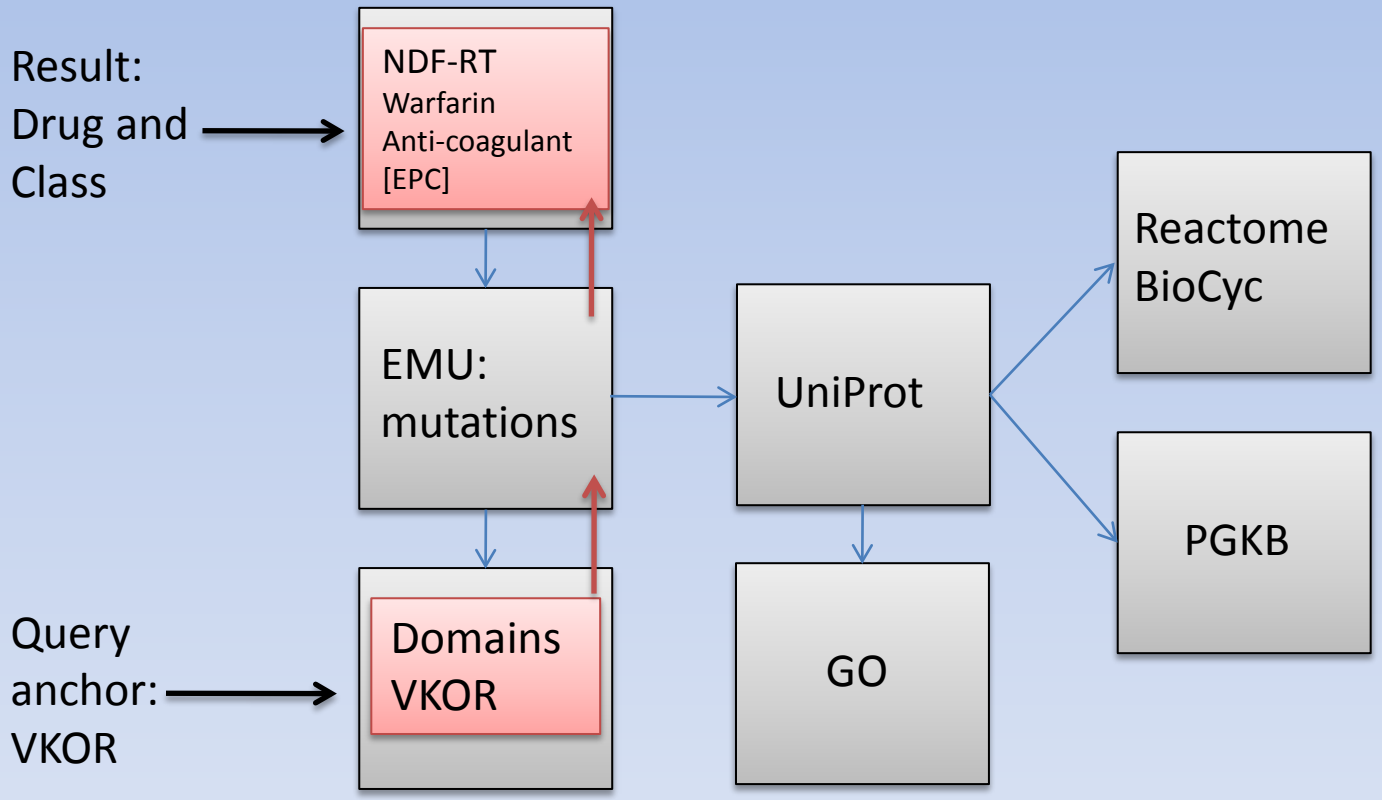
- For proteins and genes, contains database identifiers to other databases, including PharmGKB, GO, Reactome, BioCyc
- Connect to the mutational information through genes (or UniProt identifiers if available)



Schema


Clinical Drug Properties
NDF-RT
EMU: mutations
Domains
Protein structure
UniProt
GO
Reactome BioCyc
PGKB
Protein pathways

Query Strategy: an Example



What Can The System Accomplish?

- Able to link drugs to mutations to protein domains, pathways, and drug properties

- Can help with drug validation (ex. PharmGKB)

- Dynamic structure: can integrate new datasets easily for other drug-mutation questions (Ex. Drug adverse events)

Analysis: Warfarin

PharmGKB
Pharmacogenomics Knowledge Base

Search PharmGKB | Sign In | Feedback

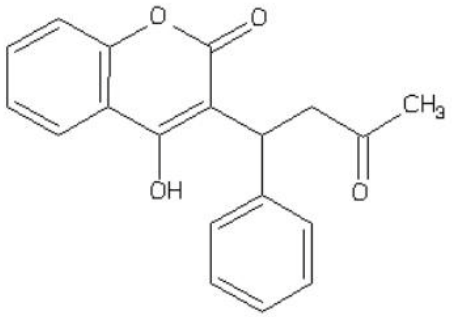
Home | Search | Submit | Download | Help | PGRN | Contributors | Clinical PGx

DRUG/SMALL MOLECULE:
warfarin

Overview | Properties | Genetics | Related Genes | Related Drugs | Related Diseases | Datasets | Downloads/LinkOuts

Overview

Generic Names:	Warfarin sodium
Trade Names:	Athrombin; Athrombin-K; Athrombine-K; Brumolin; Co-Rax; Coumadin; Coumafen; Coumafene; Coumaphen; Coumaphene; Coumarins; Coumefene; D-Con; Dethmor; Dethnel; Dicusat E; Frass-Ratron; Jantoven; Kumader; Kumadu; Kumatox; Kypfarin; Latka 42; Mar-Frin; Marevan; Maveran; Panwarfin; Place-Pax; Prothromadin; RAX; Rosex; Sofarin; Solfarin; Sorexa Plus; Temus W; Tintorane; Tox-Hid; Vampirinip II; Vampirinip III; Waran; Warf 42; Warfarat; Warfarin Plus; Warfarin Q; Warfarine; Warficide; Warfilone; Zoocoumarin
PharmGKB Accession Id:	PA451906



Description

(source: Drug Bank)

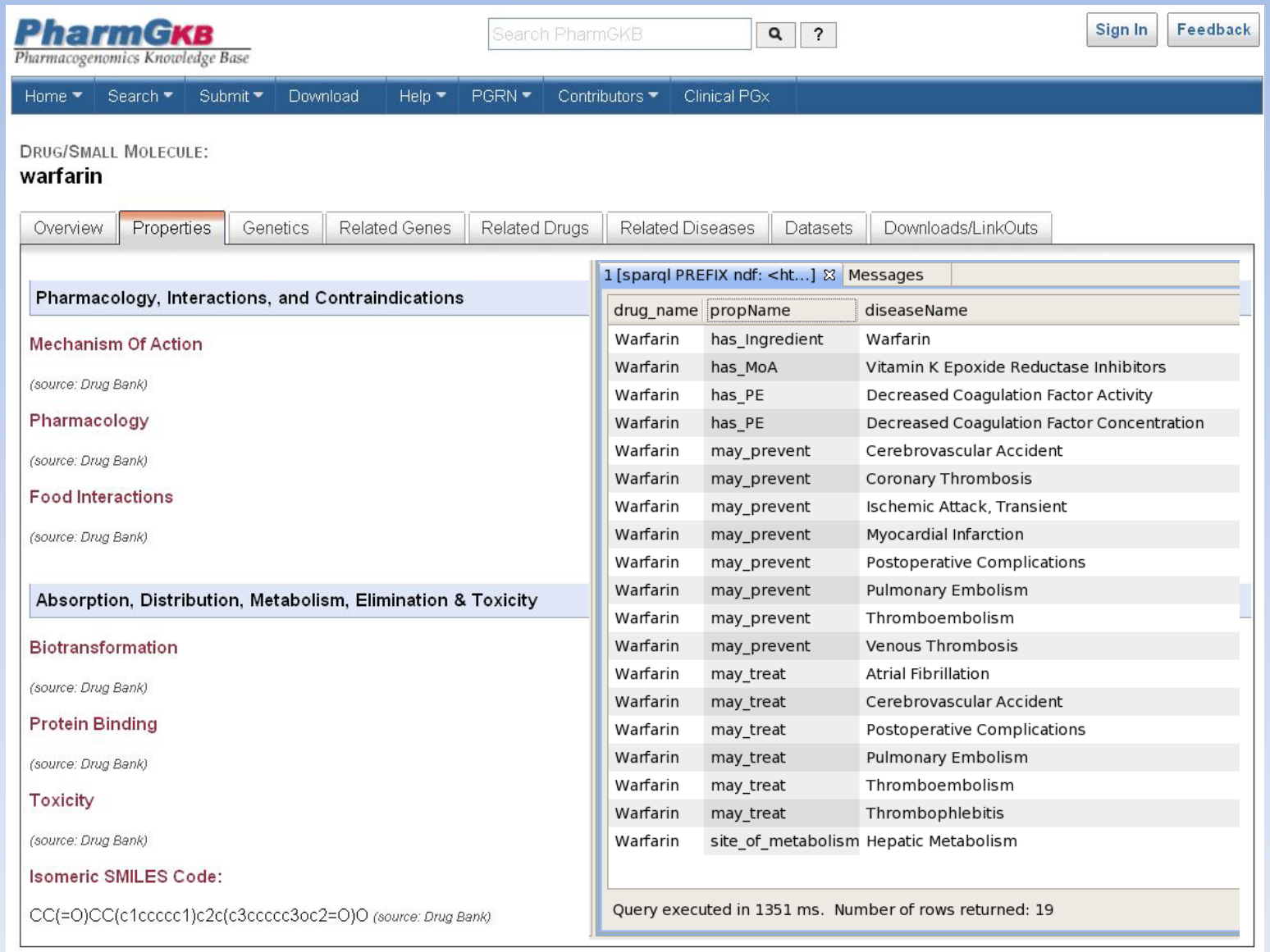
Indication

(source: Drug Bank)

ATC Therapeutic Category

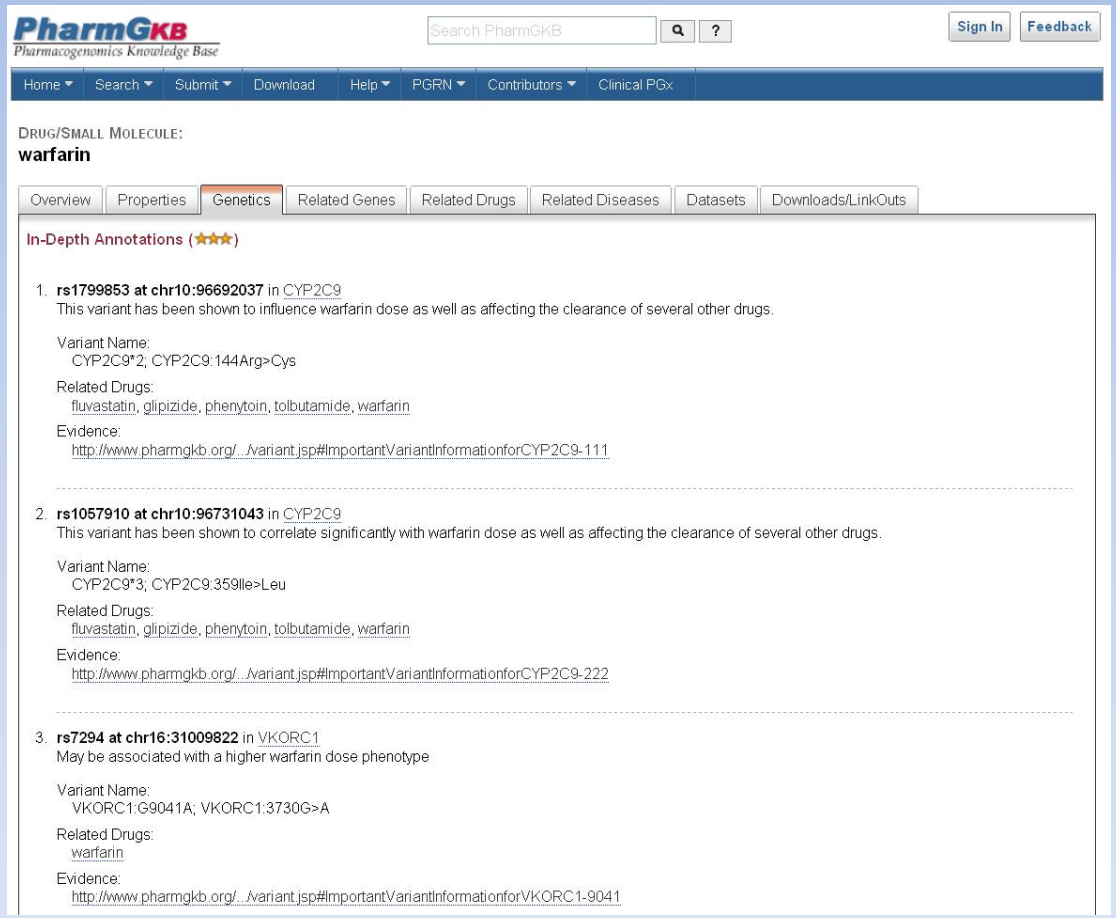
- B01AA:Vitamin K antagonists

Analysis: Warfarin

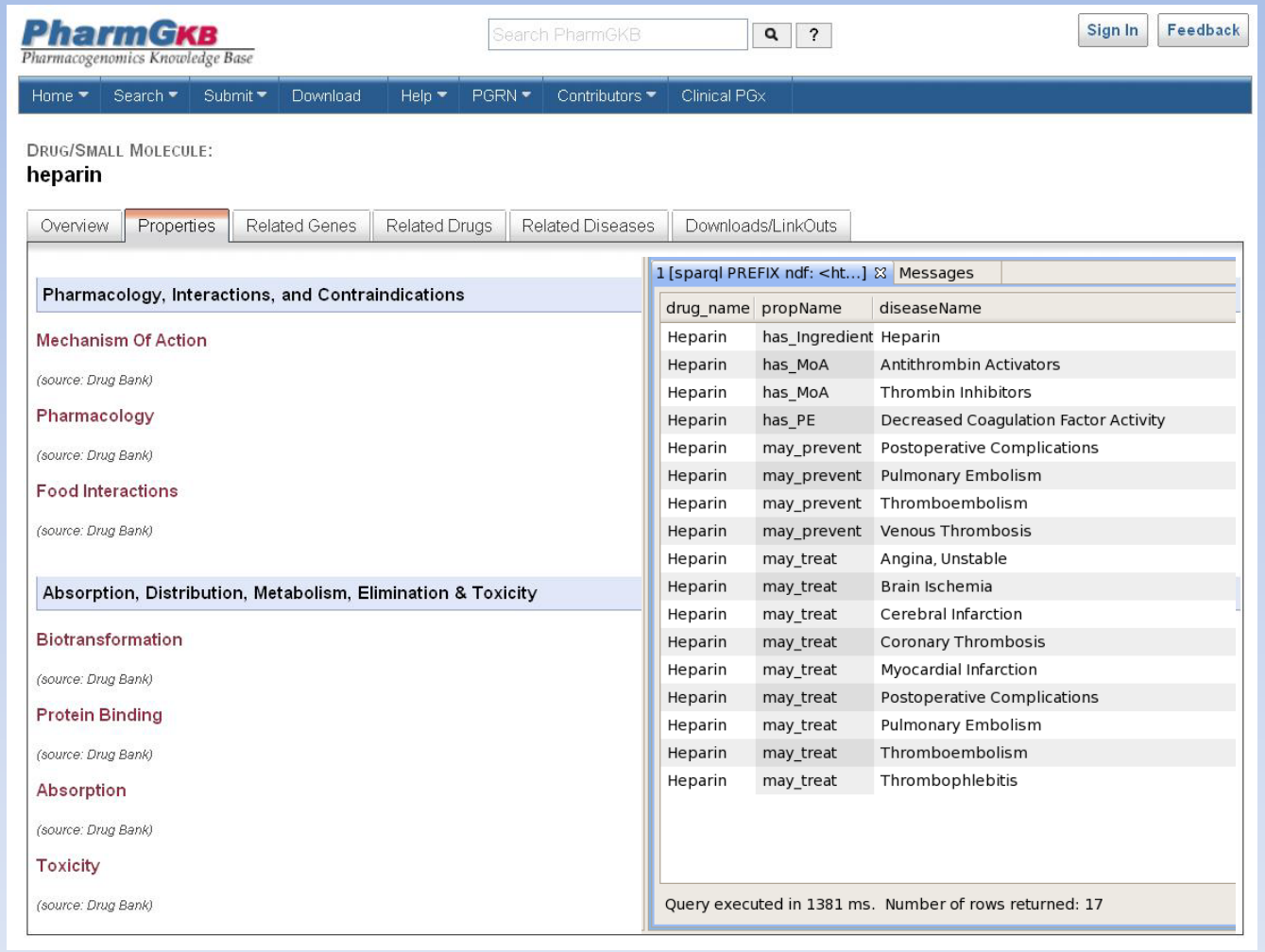


Analysis: Warfarin

- Found two genes associated with mutations: CYP2C9 and VKORC1
- Both genes annotated with mutations in PharmGKB
- Reconstruction of known information



Analysis: Heparin



Analysis: Heparin

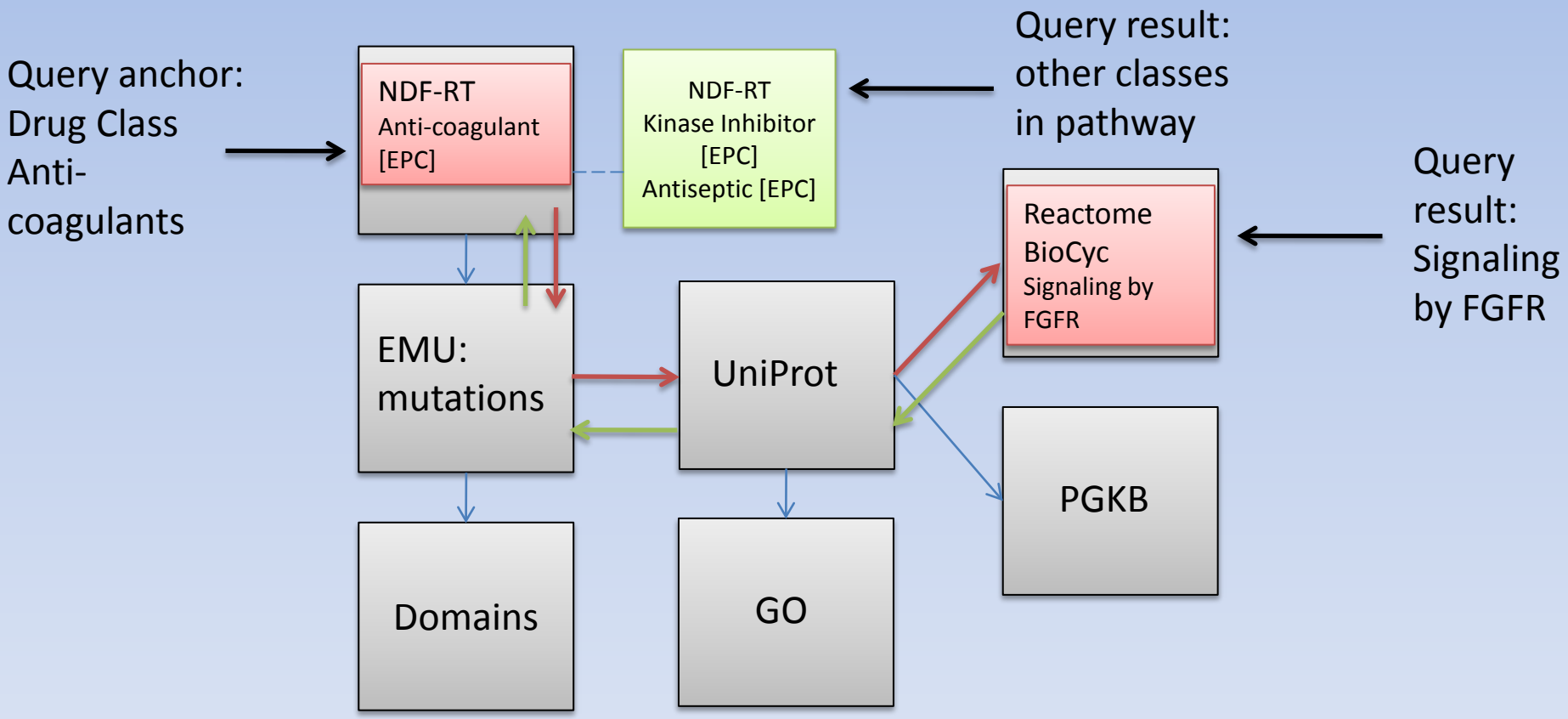
- Uncovered 239 domains from two sources (name variations possible)
- 29 citations found with associations to mutations
- PharmGKB has no information on genetics and thus no mutational links easily available
- Recommendation: validation of Heparin towards genetics using our list of citations as a starting point

Further System Applications

For a given class (ex. anti-coagulants):

- What are the pathways associated with that class through mutations
- What are the other classes associated with that pathway

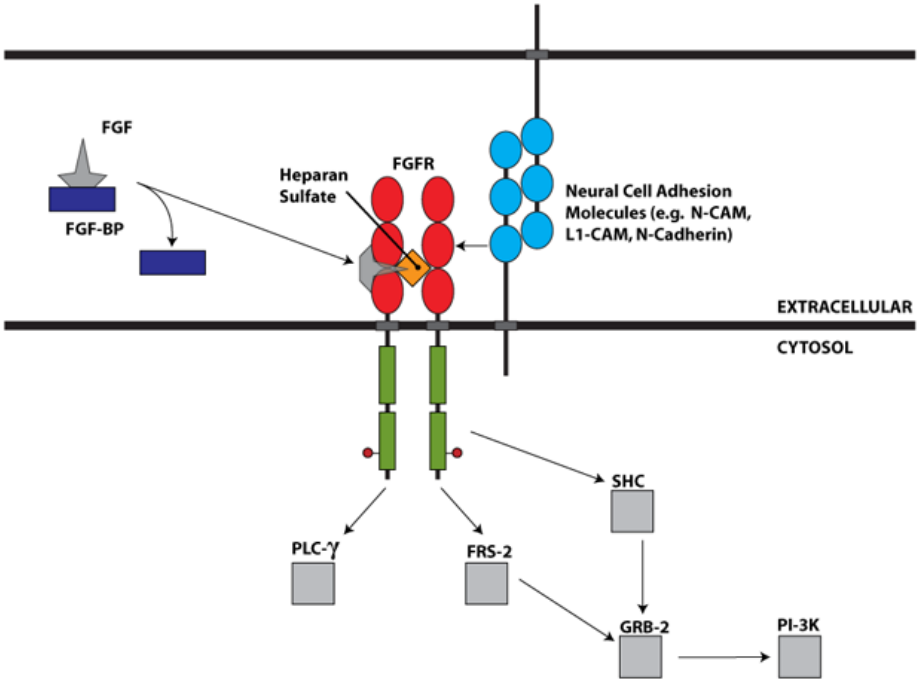
Query Classes to Pathways



Demonstrating the System

Signaling by FGFR	
Stable identifier	REACT_9470.2
Authored	de Bono, B, 2007-01-10
Reviewed	Mohammadi, M, 2007-02-06
Your feedback	Let us know what you think of this article (click here)

The 22 members of the fibroblast growth factor (FGF) family of growth factors mediate their cellular responses by binding to and activating the different isoforms encoded by the four receptor tyrosine kinases (RTKs) designated FGFR1, FGFR2, FGFR3 and FGFR4. These receptors are key regulators of several developmental processes in which cell fate and differentiation to various tissue lineages are determined. Unlike other growth factors, FGFs act in concert with heparin or heparan sulfate proteoglycan (HSPG) to activate FGFRs and to induce the pleiotropic responses that lead to the variety of cellular responses induced by this large family of growth factors. An alternative, FGF-independent, source of FGFR activation originates from the interaction with cell adhesion molecules, typically in the context of interactions on neural cell membranes and is crucial for neuronal survival and development. Upon ligand binding, receptor dimers are formed and their intrinsic tyrosine kinase is activated causing phosphorylation of multiple tyrosine residues on the receptors. These then serve as docking sites for the recruitment of SH2 (src homology-2) or PTB (phosphotyrosine binding) domains of adaptors, docking proteins or signaling enzymes. Signaling complexes are assembled and recruited to the active receptors resulting in a cascade of phosphorylation events. This leads to stimulation of intracellular signaling pathways that control cell proliferation, cell differentiation, cell migration, cell survival and cell shape, depending on the cell type or stage of maturation. [Eswarakumar *et al* 2005, Dailey *et al* 2005, Schlessinger 2000]

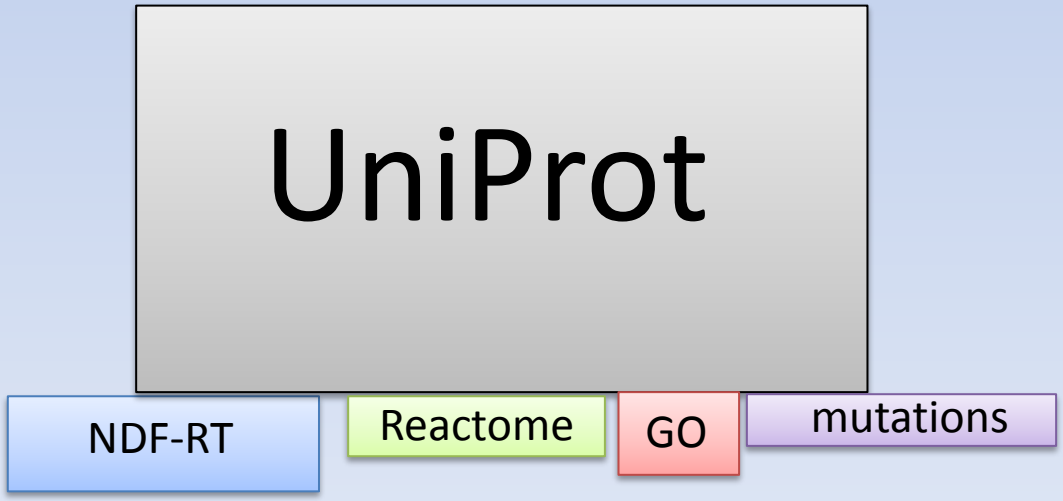


Preliminary Quantitative Results

Protein Characteristics	Drug Characteristics: Class (EPC)	Drug Characteristics: Properties
Pathway	5,724 (pathway, gene, drug, class)	In Progress…
Domain	55,930 (domain, drug, class)	In Progress…

Technical Challenges

- UniProt
  - massive database with ~1.5 billion triples (majority of data)
  - needs a robust server and efficient queries in order to be used effectively
  - however, contains useful information, not overwhelming and still query-able



Technical Challenges

- Interoperability
  - No way to link Reactome and BioCyc together
  - Solution: go through UniProt

- Smart and efficient query strategies
  - Due to size of the entire system and complexity, need efficient query strategies to get results
  - Need an entry point for quick querying (starting from specific drug, domain, etc)
  - Reducing UniProt will help with this problem, though not completely

System Limitations

- With current level of knowledge, difficult to make generalized hypotheses

- Difficult to makes direct connections between properties of drugs and protein levels due to interaction complexity

- System is weakened overall by its weakest links, but difficult to fully exploit in 2.5 months

- Post processing difficult with the high volume of data

Additional Limitations

- Mutations
  - EMU mutation extractions were not curated for drug-mutation relationships
  - All output automatically generated
  - EMU was evaluated on mutation extraction on a corpus of disease-related abstracts, and not evaluated on a mutation-drug corpus

- Mutation-Drug Association
  - Due to lack of curation, only say that drugs are associated with mutations, and cannot mention drugs specifically causing or affecting a mutation

Future Work

- Add additional functionality such as drug adverse events
  - ➢ easy to do if identifiers to other databases are in UniProt or if additional databases use Entrez geneID
- Curate a random portion of EMU output to create a gold standard
- Use only a subset of UniProt to improve query runtime
- Submit our findings for the Pacific Symposium on Biocomputing workshop on Mining the Pharmacogenomics Literature

Acknowledgements

- Brian Kirk, collaborator
- Jonathan Mortensen
- Olivier Bodenreider, mentor
- Dr. Maricel Kann’s lab, UMBC
- May Cheh
- Dr. McDonald