INFORMATICS PROFESSIONALS. LEADING THE WAY.

Pharos - Illuminating the Druggable Genome

An Introduction to NCATS Public Resources and Analytics for Rare Diseases, Targets, Drug Substances, and Analytes

Session Number: W20

Keith Kelleher PhD

National Center for Advancing Translational Sciences

#AMIA2023

Pharos - Illuminating the Druggable Genome

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

- Introduction
  - Background
  - Tools for your toolbox
  - Using the tools together - use cases
- Retrieving data
  - database download
  - csv download
  - Programmatic Access
- Getting data into Pharos

The need for the IDG

75% of protein research still focused on 10% genes known before the human genome was mapped

AM Edwards et al., Nature, 2011

This prompted the NIH to start the Illuminating the Druggable Genome Initiative

COMMENT

ANTHROPOLOGY Call for unity in the science of human beings p.166

GENETICS Reviewed: two primers on personal genomics p.169

POLICY Sanitation, not vaccination, is most important in Haiti p.175

OBITUARY Jack Oliver, key player in proof of plate tectonics, remembered p.176

ILLUSTRATION BY JONATHAN BURTON

Too many roads not taken

Most protein research focuses on those known before the human genome was mapped. Work on the slew discovered since, urge **Aled M. Edwards** and his colleagues.

When a draft of the human genome was announced in 2000, funders, governments, industry and researchers made grand promises about how genome-based discoveries would revolutionize science. They promised that it would transform our understanding of human biology and disease, and provide new targets for drug discovery. Yet more than 75% of protein research still focuses on the 10% of proteins that were known before the genome was mapped — even though many more have been genetically linked to disease.

We performed a bibliometric analysis to assess how research activity has altered over time for three protein families that are central in disease and drug discovery: kinases, ion channels and nuclear receptors. For all three, we found very little change in the pattern of research activity — which proteins are associated with the highest number of publications — over the past 20 years[1]. Even those proteins that have been directly associated with disease remain 'hidden in plain sight', with scientists proving very reluctant to study them.

Where there has been a shift in research activity, it was often spurred by the emergence of tools to study a particular protein, not by a change in the protein's perceived importance. We believe that ensuring high-quality tools are developed for all the proteins discovered may be all that is needed to drive research into the unstudied parts of the human genome — even within funding and peer-review systems that are inherently conservative.

We searched for mention of every human ▸

NATURE.COM
Protein mapping gains a human focus:
go.nature.com/vbqvff

Pharos - Introduction

Pharos began as the frontend for the Target Central Resource Database (TCRD)

TCRD integrates 79 data sources

- Uniprot
- DISEASES
- GTeX
- ChEMBL
- DrugCentral
- MONDO
- etc.

Pharos is the web frontend that allows you to:

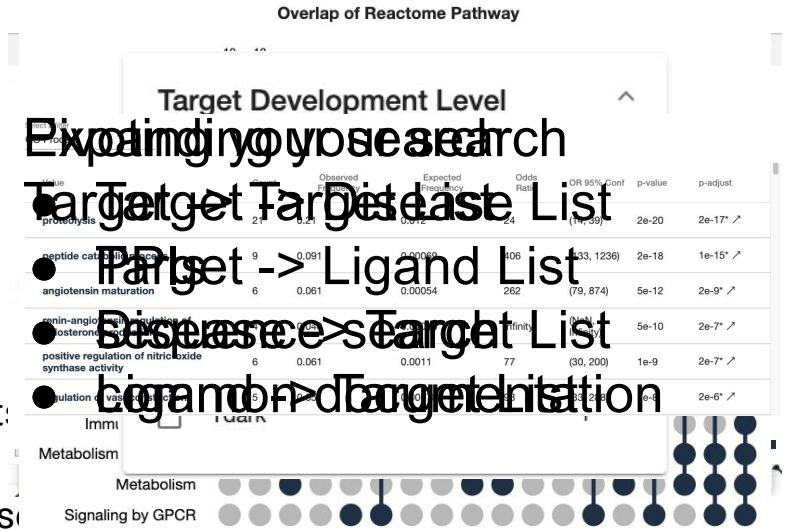
- search
- browse
- visualize
- analyze
- download

Basic Tools for Pharos users

Search for targets, diseases, ligands
Download data
Interactive components
List Analysis tools
- sort, filter
- facet counts
- visualizations (UpSet plots, Heatmaps)
- calculate enrichment

Expanding your search
- change from thinking of one target, to a list of targets

Pivoting your search
- change from thinking of one target, to a list of diseas



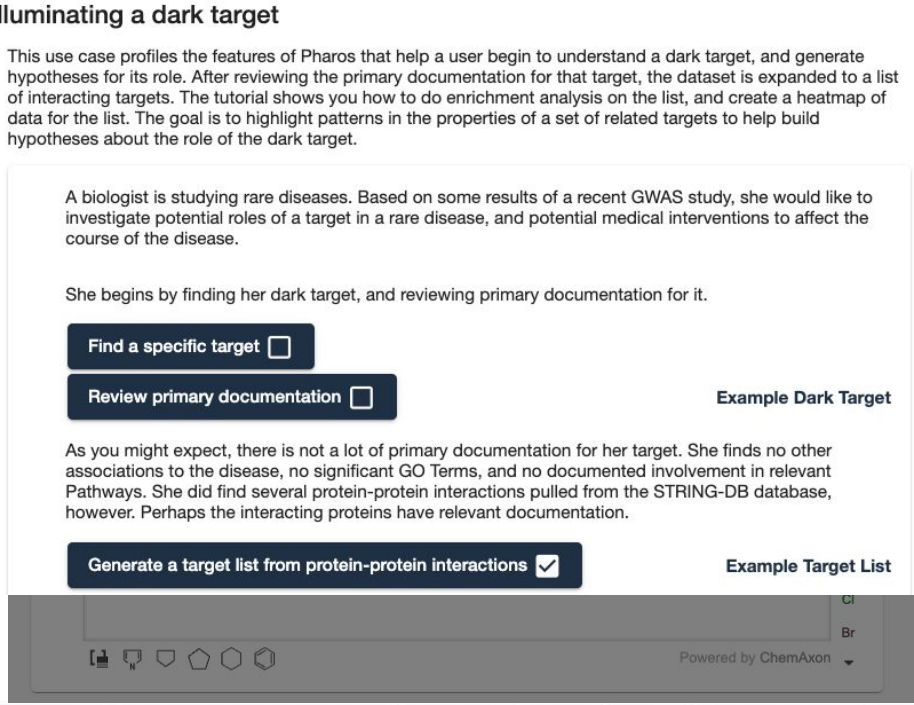
Expanding your search
- Target -> Disease List
- Target -> Ligand List
- Disease -> Target List
- Ligand -> Target List

Using the tools together

https://pharos.nih.gov/usecases

Use cases

- Finding primary documentation for targets a
- Finding chemical compounds
- Illuminating a dark target
- Finding an appropriate dark target
- Exploring effects of a novel chemical comp
- Identify commonalities between ligands ide



Fetching data from Pharos

Three use cases

Fetching lots and lots of data - download the database

- http://juniper.health.unm.edu/tcrd/download/ - Base TCRD
- https://opendata.ncats.nih.gov/public/pharos/ - Pharos Version

For a focused dataset

- download CSVs from Pharos UI
- every list page and details page has a download link

Programmatically

- PROD : https://pharos-api.ncats.io/graphql
- DEV : https://ncatsidg-dev.appspot.com/graphql
- Example Queries : https://pharos.nih.gov/api

Pharos GraphQL overview

Main useful query types:

`batch`: useful for loading disparate id types (target/ligand/disease)

`target`: details about a **single target**

`targets`: details and facets for a **list of targets**

`ligand`: details about a **single ligand**

`ligands`: details and facets for a **list of ligands**

`disease`: details about a **single disease**

`diseases`: details and facets for a **list of diseases**

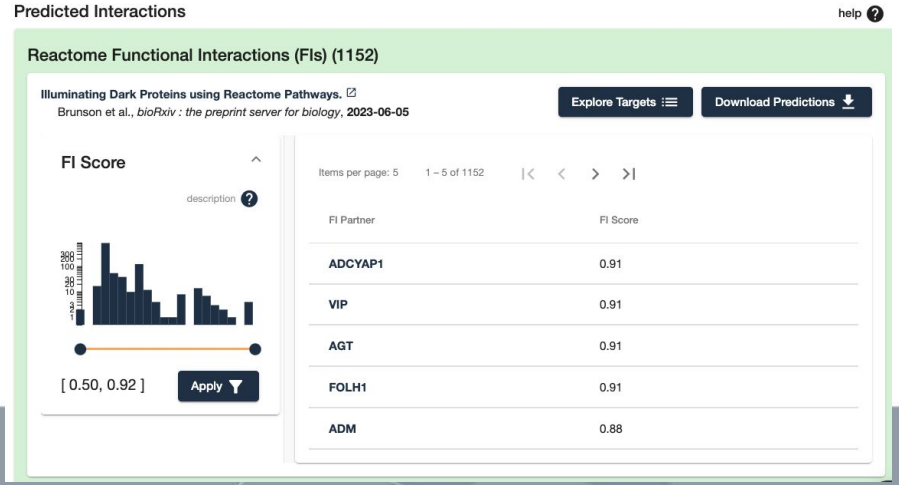
Getting Data back into Pharos

Community Data API

Schema.org structured data is shown in table or card layouts
- More formats can be created. Work with us to let us know what you need

Interactive facets are shown for your data fields

Tutorial on the repo
- https://github.com/ncats/pharos-community-data-api



Why share your data in Pharos?

Pharos sees 1k-2k users per week
Show your data in context of other target, disease, compound knowledge
Take advantage of the list analysis features and visualizations that Pharos has to offer
Bite size project for students, etc.

Criteria for inclusion in Pharos

https://pharos.ncats.nih.gov/faq

Pharos and IDG

What are the criteria for including predictive models in Pharos?

Utility

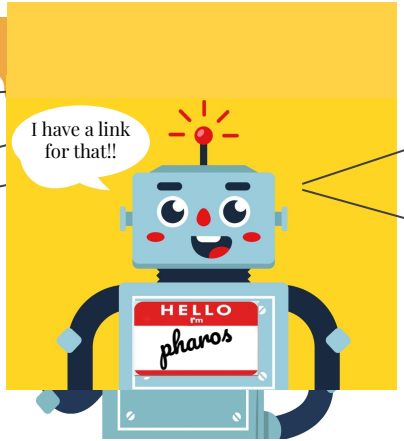
- Adds value to Pharos' users. Examples include:
  - Predictions that fill in gaps for which experimental evidence is not available
  - Confidence metrics or rankings for experimental data
  - Aggregation of knowledge across sources to generate insight into a target's functional role
- Predictions / calculations are well defined and described
  - Standardized confidence metrics that can be compared across targets
  - Details on how metrics are calculated
- Predictions apply to a significant number of targets or diseases
  - i.e. at least 100 targets / diseases
- Ideally, predictions for targets include dark targets

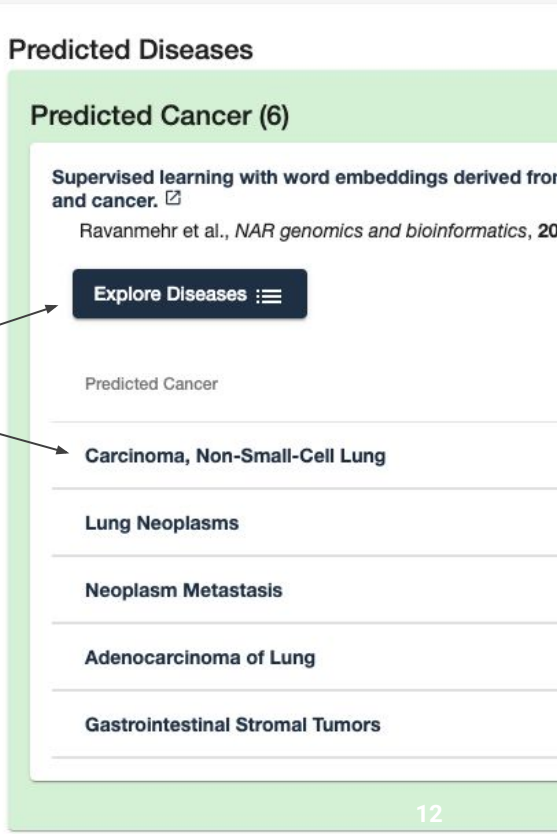
Quality

- High performing predictions for its domain
  - As shown in publication

Benefits of Structured Data

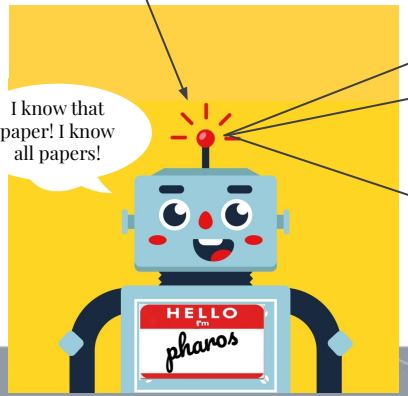
```
predictions: [{
  @type: 'Prediction',
   name: 'Predicted Cancer',
   value: {
      '@type': 'MedicalCondition',
      Name: 'Carcinoma, Non-Small-Cell Lung'
      alternateName: 'MESH:D002289'},
     Confidence: {
            @type: 'QuantitativeValue',
            alternateName: 'probability',
            Value: 0.87}
Citation: {...},
Style: 'table' (or 'card')
```





Benefits of Structured Data

```
Citation: {
      @type: 'ScholarlyArticle',
       Identifier: {
            Name: 'PMID',
            Value: 34888523
       }
}
```



Predicted Diseases

Descriptions and Definitions

Disease Associations:

Predicted disease associations retrieved from external APIs.

Citation for Predicted Data

Supervised learning with word embeddings derived from PubMed captures latent knowledge about protein kinases and canc

Vida Ravanmehr, Hannah Blau, Luca Cappelletti, Tommaso Fontana, Leigh Carmody, Ben Coleman, Joshy George, Justin Reese, Joachimiak, Giovanni Bocci, Peter Hansen, Carol Bult, Jens Rueter, Elena Casiraghi, Giorgio Valentini, Christopher Mungall, Tudor Peter N Robinson

PMID: 34888523

Abstract:

Inhibiting protein kinases (PKs) that cause cancers has been an important topic in cancer therapy for years. So far, almost 8% of > have been targeted by FDA-approved medications, and around 150 protein kinase inhibitors (PKIs) have been tested in clinical tria present an approach based on natural language processing and machine learning to investigate the relations between PKs and ca predicting PKs whose inhibition would be efficacious to treat a certain cancer. Our approach represents PKs and cancers as sema meaningful 100-dimensional vectors based on word and concept neighborhoods in PubMed abstracts. We use information about trials in ClinicalTrials.gov to construct a training set for random forest classification. Our results with historical data show that asso between PKs and specific cancers can be predicted years in advance with good accuracy. Our tool can be used to predict the rel inhibiting PKs for specific cancers and to support the design of well-focused clinical trials to discover novel PKIs for cancer therap

Demo

use cases

Illuminating a dark target

- use case
  - given a relatively unknown target - MAPK11
  - explore primary documentation
  - expand your search
  - calculate enrichment

characterizing a novel compound

- use case
  - given a new chemical compound
    - CC1CC(O)(CCN1CCCC(=O)C1=CC=C(F)C=C1)C1=CC=C(Cl)C=C1
  - explore similar structures
    - target enrichment
    - pathway enrichment
  - explore predicted targets
- variations
  - active ligands for a target
    - https://pharos.nih.gov/ligands?associatedTarget=CAMK2A
      - find potent compounds
      - find selective compounds
  - ligands from a screen
    - https://pharos.nih.gov/analyze/ligands?collection=0xL5foFCpKfCVYb71K4F
      - find patterns in noise - panther class

Let's make this interactive

All workshop materials can be found here:

https://shorturl.at/kvzIZ

Some installation prerequisites for RaMP-DB:

https://shorturl.at/bgqrQ

Provide your comments/thoughts here:

https://shorturl.at/sS138

Thank you!

Email me at:
keith.kelleher@nih.gov