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

Ewy A. Mathé[1], Keith J. Kelleher[1], Tim Sheils[1], Qian Zhu[1]
[1]Division of Preclinical Innovation, National Center for Advancing Translational Sciences, Rockville, MD

Abstract
Translational research efforts increasingly rely on up-to-date and comprehensive resources and associated analytics to accelerate preclinical and clinical discoveries. While many different types of resources exist, they can be broadly categorized as those providing information on diseases, molecular/omic phenotypes, target annotations, and ingredient/drug regulatory information. Our group at NCATS is developing, maintaining, and providing access and analytics to such resources in collaboration with other domain experts in academia and other federal agencies. The goal of this instructional workshop is to provide an overview of four public resources: 1) Rare Disease Alert System (RDAS) for annotations on rare diseases from biomedical literature, grant funding, and clinical trials; 2) Relational database of Metabolomics Pathways (RaMP-DB) for chemical, biological, and ontology annotations on human analytes (metabolites, genes, proteins); 3) Pharos for annotations related to targets, including the widely used Target Development Level; and 4) Inxight/Global Substance Registration System (GSRS) for integrated and curated data supplied by the FDA and private companies on the regulatory status and drug ingredient definitions and annotations. For each of these resources, an overview of the underlying datasets, their provenance and how they could be useful will be discussed, followed by a more detailed and interactive session on how to interact with the data (through web applications, R packages, and/or APIs). Further, we will provide details on the underlying standards used within each resource, as well as examples on how we integrate these resources for generating comprehensive knowledge graphs. Finally, we plan to highlight opportunities for further development, collaborations, and promote discussion on how these resources can be further integrated with others and with other analytical tools. Overall, this workshop will thus provide hands on experience with comprehensive resources on annotations at the disease, omic, target, and drug levels, and opportunities for using these data in translational research efforts.

Introduction

This era is marked by an immense amount of publicly available data, which enhances our ability to use and process what we know to instigate new discoveries or provide innovative solutions. In translational research, the goal is to offer relief to patients suffering a wide range of human diseases and conditions more quickly. A driving component in achieving this goal is useful and usable data. With this in mind, our goal in this workshop is to provide background information and hands on experience for four up-to-date and comprehensive resources: 1) Rare Disease Alert System (RDAS); 2) Relational database of Metabolomics Pathways (RaMP-DB); 3) Pharos; and 4) Inxight/Global Substance Registration System (GSRS). These resources and their access are summarized in Table 1.

Table 1. Workshop resources to be presented during the workshop.

Resource	Brief Description	Access
RDAS	Annotations on rare disease related attributes extracted from scientific publications, NIH grant fundings and clinical trials by implementing the advanced language models.	**Web app**: http://rdas-dev.ncats.nih.gov/diseases **Neo4j DB**: http://rdas-dev.ncats.nih.gov:7474/browser/
RaMP-DB	Annotations and analytics on chemical and biological properties, ontologies, and reactions for human metabolites, genes, and proteins.	**Web app**: https://rampdb.nih.gov/ **R package**: https://github.com/ncats/RaMP-DB **API:** https://rampdb.nih.gov/api
Pharos	Annotations and analytics on drug targets and associated diseases, ligand activities, and more.	**Web app**: https://pharos.nih.gov/ **API**: https://pharos-api.ncats.io/graphql
GSRS/Inxight	Curated annotations on drug substances and associated pharmacological actions, targets, treatment modalities, indications, pharmacokinetic properties, and more.	**Web app**: https://drugs.ncats.io/ **Web app**: https://gsrs.ncats.nih.gov/#/

RDAS

Rare disease patients often experience diagnosis delays or misdiagnosis. The collection and public sharing of accurate, up-to-date, standardized data on rare diseases are key priorities in shortening the diagnostic odyssey. These data inform health care providers about the up-to-date information/findings on the rare diseases they are treating, but also educate patients and their families with critical information pertinent to their conditions. Here, we introduce Rare Disease Alert System (RDAS) as a public, comprehensive rare disease resource that supports searching, informing, educating, and alerting users on the latest information and findings pertinent to rare diseases. RDAS comprises a frontend user interface (UI) and a backend data repository. The UI allows users to search, browse, and subscribe to alert services. RDAS thus provides the latest information and findings about rare disease(s) of interest. The UI is built using the Angular front-end library, with a Node/Express server layer to connect to the underlying Neo4j database [1] that stores the backend Knowledge Graphs (KGs). The backend data repository includes three KGs built by integrating information from PubMed articles, clinical trials, and NIH grant funding related to rare diseases. Specifically, the KGs are: 1) Scientific Annotation Knowledge Graph (RDAS_SAKG) [2] contains annotations generated from PubMed abstracts by PubTator,[3] and epidemiology information extracted from PubMed abstracts using the developed deep learning models[4, 5]; 2) Clinical Trial Knowledge Graph (RDAS_CTKG) semantically represents data extracted from clinicaltrial.gov from the perspectives of our focused users, patients, health care providers and informaticians; 3) NIH Grant Funding Knowledge Graph (RDAS _GFKG) contains annotations generated from NIH funded projects related to rare diseases [6].

RaMP-DB

RaMP-DB [7] is a publicly available relational database that integrates multiple sources of biological, structural/chemical, disease, and ontology annotations for analytes (metabolites, genes, proteins) from multiple resources including HMDB[8], KEGG[9], Reactome[10], WikiPathways[11], LIPIDMAPS[12], ChEBI[13], and Rhea[14]. The resource was initially created to 1) enhance the ability to interpret metabolomic and multi-omic data using standardized annotations on analytes; 2) fill the gap in the availability of benchmark annotation database that was consistently used to compare pathway enrichment methods. Today, RaMP-DB includes 254,860 chemical structures, of which 43,338 are lipids, 15,389 genes, 53,745 pathways, 807,362 metabolic enzyme/metabolite reactions, and 699 ontologies. Notably, curation of the source data within RaMP-DB is semi-automatic to help ensure sustainability of the resource. The curation process involves a first automatic pass for errors in ID mappings for metabolites across resources (e.g. comparison of MW and chemical properties), followed by a second manual curation. The main functionalities for interacting with RaMP-DB include single and batch queries as well as chemical and biological pathway analyses. These functionalities are accessible through an R package (https://github.com/ncats/RaMP-DB) and a public, user-friendly web interface (https://rampdb.nih.gov/). Some useful utilities include the ability to input mixed IDs for analytes (e.g. LIPIDMAPS, PubChem, HMDB, etc for metabolites and UniProt, HMDB, and Ensembl IDs for genes/proteins), the ability to globally evaluate how much is known about analytes of interest, and clustering of enriched pathways for ease of interpreting enrichment results. The web interface is served by APIs that can also serve as endpoints for integrating RaMP-DB with other tools. Finally, RaMP-DB users can also directly download the database as a MySQL dump for their own mining or incorporation into other tools.

Pharos

Pharos [15] is a collaborative project from the Illuminating the Druggable Genome program, an NIH Common Fund program that includes 79 resources covering many different data types that annotate drug targets. The data types include gene and protein expression, publications, disease/pathway annotations, GO annotations, ligand/drug information, and many more. Currently, Pharos provides browsing and API access on annotations for 20,412 targets, 13,953 diseases, and 355,932 ligands. In the past two years, the team has greatly enhanced the usability of the website providing incredible flexibility in searching the data, programmatic access to data, tutorials, prediction models, etc. Basic workflows, such as searching, browsing, or downloading data, are naturally supported. Additionally, Pharos allows for creation of many illustrative visualizations, such as UpSet plots, and heatmaps, as well as supports advanced analysis capabilities such as enrichment calculation. Using these features together, a user can perform many interesting data analysis tasks, such as finding data about a disease of interest, exploring a list of targets documented to be associated with that disease, and calculating enrichment scores for annotations on targets in the list, like pathway annotations, or GO processes, etc. Pharos is a widely used resource averaging 2,000 new visitors per month.

Inxight Drugs/GSRS

Inxight Drugs [16] is a web resource that aggregates curated, reliable drug development data from the FDA and other resources. Using the web application, researchers can query prescription and over-the counter drug substances

approved or withdrawn from the US, globally marketed, or investigational. Each substance in Inxight is supplemented by information on pharmacological actions, targets, treatment modalities, indications (approved and off-label), pharmacokinetic properties, and more. Further, each substance includes information, when available, on their synonyms, molecular targets, pharmacology, diseases, and other conditions, and corresponding marketed projects. Currently, Inxight Drugs provides information on over 144,000 substances. Notably, all substances within Inxight are defined according to the ISO 11238 standard, thereby complying with existing regulatory standards for unique drug substance definition. This standard is defined by the software GSRS (Global Substance Registration System), a partnership between the FDA, and is a requirement for the identification of medicinal products initiative supported by WHO, FDA, EMA, BfArM, USP, and other regulators. Notably, the NCATS instance of GSRS also stores public substance-related information, which is also available through Inxight, along with data provided from NCATS and industry collaborators.

Workshop Plan

Workshop Type

This proposal is for a 3 hour instructional workshop.

Timeline and Topics Covered

Forty-five minute segments will be attributed to each resource, for a total of 3 hours. The 45 min segment will be split into the following components:

1) 15 min introduction on the data sources and examples on how they can be used
2) 10 min 'live tutorial' on how to access the resource, and a brief overview of the functionalities
3) 20 min guided discussion, QA, and self-exploration for the resource

Specific Objectives

Our main objective for this workshop is two-fold. First, we aim to provide details on the data provenance for our resources, so that our users understand how the data could be useful for their work. Second, we aim to provide hands on experience interacting with our resources in various ways (e.g. web applications, APIs, or other packages). As a secondary objective, we wish to provide ample time for interactions and discussions on how the data can be used, augmented, and further incorporated with other efforts. We anticipate new collaborative opportunities through the workshop.

Instructional Level and Intended Audience

The anticipated level is introductory to intermediate. The intended audience for the workshop is, broadly speaking, anyone who is interested in using comprehensive and up-to-date data resources that incorporate annotations on rare diseases, metabolites/genes/proteins, drug targets, and drugs/drug substances. Because our public resources are accessible via web-friendly applications and through programmatic access, there is no set of computational expertise required. Thus, any participant interested in learning more about our resources and how to interact with them would benefit from the workshop. Workshop attendees will benefit most from this workshop if they fully read the proposal and explore the resources on their own prior to workshop day.

Workshop presenters

Ewy Mathé – Dr. Mathé has planned many interactive workshops on bridging the gap between computational and non-computational users (GL-BIO, ASMS) and on exploring the needs and uses of benchmarking data to support method and software development (MANA, Dagstuhl meeting on Computational Metabolomics). She has led the development of RaMP-DB since its inception and is involved in collaborative projects using all the resources planned to be discussed in this workshop.

Keith Kelleher – Dr. Kelleher is a developer for the Pharos project, developing new features, and solving problems at the backend database and web frontend levels. To date, he has added more than 245,000 lines of code to the front-end and back-end Pharos repositories. He is also involved in maintaining the Inxight web application.

Tim Sheils – Mr. Sheils has 9 years of software development experience at NCATS. During this time, he has developed the Pharos UI, the RaMP-DB web portal and API, and the RDAS web app. He has also presented extensively on Pharos, including hosting hack-a-thons, and leading a seminar workshop.

Qian Zhu – Dr. Zhu is a team lead of rare disease translational research in Informatics Core within Division of Preclinical Innovation at NCATS. She has extensive experience in medical informatics and started to work in the field of rare disease informatics since 2011. She supported Genetic and Rare Disease Information Center (GARD) program since she joined NCATS in 2018. Currently she is leading the development of Rare Disease Alert System (RDAS).

This workshop has not been presented at any conferences prior, although these resources have been presented at other venues at a regular basis. Specifically, RDAS was presented at Rare Disease Day 2023, RaMP-DB was previously presented at the 2022 AMIA Informatics Summit, MANA, Metabolomics Society, and ASMS,. Pharos has been presented at ACS, NAVBO and IDG Program meetings.

References

[1] J. J. Miller, "Graph database applications and concepts with Neo4j," in *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, 2013, vol. 2324, no. 36.

[2] Q. Zhu *et al.*, "Rare disease-based scientific annotation knowledge graph," *Frontiers in Artificial Intelligence,* vol. 5, 2022.

[3] C.-H. Wei, H.-Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," *Nucleic acids research,* vol. 41, no. W1, pp. W518-W522, 2013.

[4] J. N. John, E. Sid, and Q. Zhu, "Recurrent Neural Networks to Automatically Identify Rare Disease Epidemiologic Studies from PubMed," *AMIA Summits on Translational Science Proceedings,* vol. 2021, p. 325, 2021.

[5] W. Z. Kariampuzha *et al.*, "Precision information extraction for rare disease epidemiology at scale," *Journal of Translational Medicine,* vol. 21, no. 1, p. 157, 2023.

[6] Q. Zhu *et al.*, "Scientific evidence based rare disease research discovery with research funding data in knowledge graph," *Orphanet journal of rare diseases,* vol. 16, no. 1, pp. 1-12, 2021.

[7] J. Braisted *et al.*, "RaMP-DB 2.0: a renovated knowledgebase for deriving biological and chemical insight from metabolites, proteins, and genes," *Bioinformatics,* vol. 39, no. 1, p. btac726, 2023.

[8] D. S. Wishart *et al.*, "HMDB 5.0: the human metabolome database for 2022," *Nucleic Acids Research,* vol. 50, no. D1, pp. D622-D631, 2022.

[9] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research,* vol. 28, no. 1, pp. 27-30, 2000.

[10] B. Jassal *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research,* vol. 48, no. D1, pp. D498-D503, 2020.

[11] M. Martens *et al.*, "WikiPathways: connecting communities," *Nucleic acids research,* vol. 49, no. D1, pp. D613-D621, 2021.

[12] E. Fahy, M. Sud, D. Cotter, and S. Subramaniam, "LIPID MAPS online tools for lipid research," *Nucleic acids research,* vol. 35, no. suppl_2, pp. W606-W612, 2007.

[13] K. Degtyarenko *et al.*, "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic acids research,* vol. 36, no. suppl_1, pp. D344-D350, 2007.

[14] P. Bansal *et al.*, "Rhea, the reaction knowledgebase in 2022," *Nucleic acids research,* vol. 50, no. D1, pp. D693-D700, 2022.

[15] K. J. Kelleher *et al.*, "Pharos 2023: an integrated resource for the understudied human proteome," *Nucleic Acids Research,* vol. 51, no. D1, pp. D1405-D1416, 2023.

[16] V. B. Siramshetty *et al.*, "NCATS Inxight Drugs: a comprehensive and curated portal for translational research," *Nucleic Acids Research,* vol. 50, no. D1, pp. D1307-D1316, 2022.