The NIH Common Fund Metabolomics Consortium

Michael Conlon[1], Padma Maruvada[2], Richard A. Yost[1,3]

[1]Metabolomics Consortium Coordinating Center (M3C), University of Florida, Gainesville, Florida, USA
[2]NIH/NIDDK, Bethesda, Maryland, USA
[3]Chemistry Department, University of Florida, Gainesville, Florida, USA

This material was presented at the Metabolomics Society meeting, June 26-27, 2019, The Hague, Netherlands

Consortium Goals

1. *Establish an enduring public repository* for metabolomic data
2. *Overcome technical hurdles* in analyzing and interpreting metabolomics data, including the ability to determine metabolite identities
3. *Develop consensus* for, and promote adoption of, best practices and guidelines to promote accuracy, reproducibility, and re-analysis of metabolomics data in collaboration with the national and international communities.

Scan for additional information on consortium cores

Consortium Cores

National Metabolomics Data Repository, Metabolomics Workbench http://metabolomicsworkbench.org
Shankar Subramanian, University of California San Diego, National Metabolomics Data Repository NextGen Metabolomics Workbench

Five Compound Identification Cores
Arthur S. Edison, University of Georgia, Genetics and Quantum Chemistry as Tools for Unknown Metabolite Identification
Alexey Nesvizhskii, Charles R. Evans, University of Michigan, Michigan Compound Identification Development Cores (MCIDC)
Oliver Fiehn, University of California Davis, West Coast Metabolomics Center for Compound Identification
Dean Paul Jones, Shuzhao Li, Gary W. Miller, Edward T. Morgan, Emory University, Mega-scale Identification Tools for Xenobiotic Metabolism
Thomas O. Metz, Pacific Northwest National Laboratory, Pacific Northwest Laboratory Compound Identification Core

Seven Data and Tools Cores
John Weinstein, Rehan Akbani, Bradley M. Bloom, University of Texas MD Anderson Cancer Center, Computational Tools for Analysis and Visualization of Quality Control Issues in Metabolomic Data
Jamey Young, Doug Allen, Vanderbilt University, Tools for Leveraging High-resolution MS Detection of Stable Isotope Enrichments to Upgrade the Information Content of Metabolomics Datasets
XiuXia Du, University of North Carolina Charlotte, Cross-platform and Graphics Software Tool for Adaptive LC/MS and GS Metabolomics Data Processing
Shuzhao Li, Emory University, Gary Siuzdak, Scripps Research Institute, Mummichog 3, Aligning Mass Spectrometry Data to Biological Networks
Alla Karnovsky, University of Michigan, George Michailidis, University of Florida, Methods and Tools for Integrative Functional Enrichment Analysis of Metabolomics Data
Katerina Kechris, Debashis Ghosh, University of Colorado Denver, Addressing Sparsity in Metabolomics Data Analysis
Garry Patti, Washington University at St. Louis, A Comprehensive Platform for High-Throughput Profiling of the Human Reference Metabolome

Consortium Coordinating Center http://metabolomics.info
Richard A. Yost, Michael Conlon, University of Florida, Metabolomics Consortium Coordinating Center

Participating in the Consortium

- ❑ Consortium workgroups are open to all. Workgroups identify hurdles, develop best practices, contribute to repository development, and build consensus. Current workgroups include:
  - Software standards for Metabolomics Tool Development, Jamey Young, Vanderbilt University, chair.
  - Quantum Mechanical Computing Best Practices, Art Edison, University of Georgia, chair
  - Unknown lipids data exchange, Charles Evans, University of Michigan, chair
- ❑ The Metabolomics Workbench is open to all. Find, use, and contribute FAIR metabolomics data. See https://www.metabolomicsworkbench.org/data/
- ❑ For more info, or to participate in a workgroup, visit http://metabolomics.info, contact us at *info@metabolomics.info*, or follow us on Twitter @metabinfo

The NIH Common Fund (http://commonfund.nih.gov) addresses emerging scientific opportunities and pressing challenges in biomedical research that no single NIH Institute or Center can address on its own, but are of high priority for the NIH as a whole.

Metabolomics Stakeholder Engagement and Program Promotion

Michael Conlon, Alisha Mitchell-Roberts, Richard A. Yost
University of Florida, Gainesville, Florida, USA

Background

Stakeholder Engagement and Program Promotion is a fundamental work activity of the NIH Common Fund Metabolomics Consortium, and is facilitated by the Metabolomics Consortium Coordinating Center (M3C). At the April 2019 investigator meeting at UCSD, the investigators discussed focusing efforts on stakeholders who are experts in the field of metabolomics. Future efforts may broaden the reach of stakeholder engagement and program promotion. At the Metabolomics Society meeting at The Hague, data curation, reporting standards, and alignment of data repository efforts were discussed as potential areas of joint work.

Upcoming stakeholder engagement and program promotion events

September 25-26, 2019 – Annual Consortium Meeting, Shady Grove, MD
Investigators present progress on their scientific work. External Program consultants review consortium progress and make recommendations for improvement.

November 15-17, 2019 – First annual meeting of the Metabolomics Association of North America, Atlanta, GA
M3C representatives will meet as needed with stakeholders regarding needs that may be addressed by the consortium, plus support a Workshop on Compound ID

January 2020 – M3C Pilot and Feasibility Call for Proposals
Focus on best practices for reproducibility

January 25-27, 2020 Metabolomics short course, Gainesville, FL

March 29-31, 2020 – MSACL Short Course, Palm Springs, CA
Outreach to clinical labs on high resolution MS and metabolomics

May 4-5, 2020 – Consortium spring meeting, Shady Grove, Maryland
Using multiple breakouts, principal investigators discuss how best to advance the goals of the consortium and create recommendations for program components.

May 31 – June 4, 2020 – American Society for Mass Spectrometry meeting, Houston, TX
TBD

June 2020 – Metabolomics summer school, Gainesville, FL

July 6-10, 2020 – Metabolomics Society meeting, Shanghai, China
Follow-up to meeting at the Hague. Review progress on reporting standards and data repository coordination. Discuss opportunities and focus for the coming year.

A Cross-Institutional, FAIR VIVO for Metabolomics

Michael Conlon, Kevin S. Hanson, Taeber Rapczak,
Naomi Braun, Christopher P. Barnes, University of Florida, Gainesville, Florida, USA

This material was presented at the 10th annual VIVO Conference, September 4-6, 2019, Podgorica, Montenegro

Metabolomics

Metabolomics is the scientific study of metabolites present within an organism, cell, or tissue. Human metabolites are small molecules found in human tissue that occur naturally as a result of human metabolism, or are present as a result of drugs, food components, or exposure to environmental conditions. Along with genomics (the study of DNA), transcriptomics (RNA), proteomics (proteins), metabolomics provides information regarding biochemical compounds and processes in cells, leading to a better understanding of cellular biology. "Metabolic profiling can give an instantaneous snapshot of the physiology of a cell, and thus, metabolomics provides a direct "functional readout of the physiological state" of an organism." (Wikipedia http://bit.ly/2Pa06X3)

Such profiles can be used to distinguish tissue types, disease states, and health status of individuals. Understanding the changes in profiles over time can lead to improved understanding of diseases such as cancer and diabetes, leading to potential treatments.

Over 450,000 human metabolites have been identified to date. Metabolomics is considered an emerging field. Technical challenges in compound identification and data analysis are significant.

The NIH Common Fund Metabolomics consortium originated in 2013 to help address these and other challenges. Fourteen investigators across the US are supported to advance metabolomics. Five are engaged in compound identification, seven in software tools and data analysis. The group at UCSD supports the Metabolomics Workbench, a data repository. The group at the University of Florida provides overall coordination for the consortium.

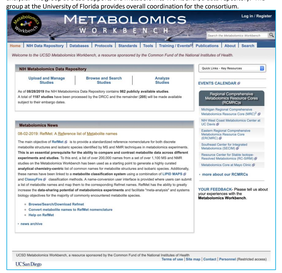


Figure 1. The Metabolomics Workbench is a data repository for data from metabolomics experiments. An API provides metadata regarding studies and investigators for all to use.

Metabolomics Data

The Metabolomics Workbench (MWB) (https://metabolomicsworkbench.org) is the National Metabolomics Data Repository of the US National Institutes of Health (NIH). Anyone can deposit data to the MWB. As of August 28, 2019, the workbench provides data from 982 publicly available studies. Another 205 studies are currently embargoed and will be available subject to their embargo dates. MWB develops and uses RefMet (http://bit.ly/2RwY5pk), a nomenclature for representing metabolites found using mass spectroscopy (MS) and nuclear magnetic resonance (NMR) techniques. Investigators upload study data to the workbench and provide metadata regarding themselves and their work. The MWB provides an API that can be used to access metadata about studies and investigators. MWB metadata has been mapped to an ontology developed by the authors to represent it as RDF and load it to VIVO.

PubMed (http://pubmed.gov) is an open access index to literature in metabolomics. Its API can be used to find and retrieve publication data regarding metabolomics investigators. Many groups use PubMed data in VIVO. PubMed data has been mapped to the VIVO Ontology (http://vivoweb.org/ontology/core)

Data regarding software used in metabolomics studies has been difficult to find. An index to such software will be created and curated by a group at the University of Colorado Anschutz Medical Campus funded by the NIH. This data will be in the form of a spreadsheet. The authors will use the Software Ontology (SWO) (http://www.ebi.ac.uk/ols/ontologies/swo.html) to represent the data as RDF and load the data into VIVO.

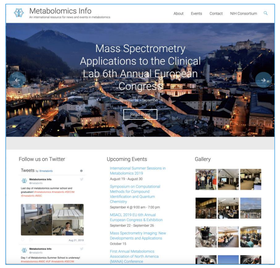


Figure 2. Metabolomics.info is a WordPress site that shares information about events, and the NIH consortium. A VIVO (not shown) provides a TPF endpoint used to display metadata regarding datasets in Metabolomics workbench, investigators, their papers and their software.

FAIR Data Principles

The FAIR Data Principles (Wilkinson, et. al https://www.nature.com/articles/sdata201618) provide a framework for creating data that can be used by groups beyond the group that created the data. The principles are difficult to achieve in practice and much has been written about implementation. VIVO supports the principles naturally, it is designed to share data.

- Findable – data should be found using search engines on the Internet
- Accessible – the data can be accessed without technical, legal or operational barriers
- Interoperable – the data is available in a common format, so that data from more than one source can be readily combined
- Reusable – the data can be reused by those finding it. This typically means that enough information is provided for the recipients of the data to determine if the data is suitable for their purposes.

The Metabolomics Workbench (https://metabolomicsworkbench.org), Metabolites (https://www.ebi.ac.uk/metabolights/index), and the MassBank (http://massbank.jp) provide data on experimental results, and spectra of compounds according to the FAIR principles. In each case, data can be found by using internal search capabilities of each of the sites. This is not ideal. Generally findable results using search engines can be created for metabolomic data sets using the techniques described here. Each dataset or study will have its own page which can be found and indexed by search engines. Our approach will result in pages for datasets, publications, investigators, and software.

Figure 3. VIVO is used to collect and represent data as RDF triples. JavaScript is used to query the VIVO TPF endpoint and display results. The profile of Dr. Rick Yost above was generated using TPF queries. The counts are for all data in the system. Future work will build out counts for each investigator, and drill down to displays of datasets, publications, and software for each investigator. Faceted search will allow users to select datasets, software, papers, and investigators by organization, metabolite, technology (NMR or MS) and other criteria. Visualizations such as word clouds and bar charts are produced using TPF as well.

Use Cases

The metabolomics community is maturing. There is interest in showcasing datasets, investigators, software, and publications. Some use cases for the web site being developed include:

1. Discover metabolomics datasets using search engines. The pages created by the approach described here will be findable by standard search engines.
2. Support discovery of datasets by web site users interested in particular metabolites, techniques, investigators, organizations and projects. Faceted search will provide this form of discovery.
3. Organize metabolomics information by investigator. Creating web pages for each investigator will emphasize the number of datasets deposited by each investigator, the number of metabolomics papers of the investigator, and software systems created by the investigator. Such displays help inform the community of investigators in the field of metabolomics.
4. Provide summaries of contributions by investigator. Summaries support the discovery process.
5. Provide contact information for investigators leading to potential support and collaboration.
6. Boost the visibility of the work products (papers, software, and datasets) of the NIH Common Fund Metabolomics Consortium, leading to increased collaboration between consortium members and others around the world.

Overall goals for the application and the consortium include:

1. Increase deposits to Metabolomics workbench
2. Increase reuse of data in Metabolomics workbench
3. Through the FAIR data principles, encourage data sharing and reuse in the metabolomics community.

Triple Pattern Fragments (TPF)

Triple Pattern Fragments (TPF), (Verborgh, et. Al doi:10.1016/j.websem.2016.03.003) is a low cost knowledge interface for the web. TPF provides a simple, fast, general mechanism for querying knowledge graphs of RDF triples. TPF answers queries of the form:

```
Subject predicate object
```

Where any of the three components can be a wildcard. A query such as

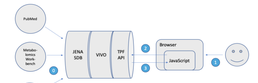
```
* rdf:type *
```

returns all the type assertions in a triple store. When executed at http://openvivo.org/tpf/core, this query returns 3.8M triples. The query

```
uri * *
```

returns all triples with the specified uri as a subject. When executed on a person's URI in VIVO, the resulting set of triples is the first order graph of the person's information in VIVO.

A TPF endpoint is included as part of the VIVO installation in version 1.10 and above based on the the TPF server available here: https://github.com/LinkedDataFragments



JavaScript

A small JavaScript library is available (http://github.com/ctsit/tpf) to make TPF queries and handle TPF replies. The code below adds the library to a page, and creates a client pointing at the OpenVIVO (http://openvivo.org) TPF endpoint. The client then requests the triples for a particular entity. The triples returned by the query are then used in a second query (`.Link`) to return all the `rdfs:label` values from the first query. The `.Single` function provides a callback function that executes on one returned value when the queries are completed. In the example, the callback displays a single label in the browser console log.

```
<script href="tpf.js"></script>
<script>
    const rdfs = "http://www.w3.org/2000/01/rdf-schema#"
    const endpoint = "http://openvivo.org/tpf/core"
    const client = new tpf.Client(endpoint)

    client
        .Entity("http://openvivo.org/a/orcid0000-0002-1304-8447")
        .Link(rdfs, "label")
        .Single(function (label) { console.log(label) })
</script>
```

Additional functions in the library include: 1) `.Results` provides a callback that can process a list of returned results, 2) `.Query` can be used to specify an arbitrary subject, predicate, object TPF query, returning a set of triples. 3) `.Type` is used to reduce a returned set of triples to entities of a specified `rdf:type`.

Technical Approach

See figure below. 1) Publication data from PubMed, dataset metadata from Metabolomics workbench, and software metadata from collaborators at the University of Colorado Anschutz Medical Campus are retrieved, transformed to triples and loaded to the JENA SDB triple store (part of VIVO) using custom scripts on a regular basis; 1) User makes request for page that will contain data from VIVO; 2) TPF Client JavaScript software makes a TPF query of the VIVO TPF endpoint. The JavaScript software is generic – any ontology and any TPF endpoint can be used; 3) the TPF API returns a set of triples to the client. No modifications of VIVO are necessary.

Ontology

A small ontology was created to represent the data for the metabolomics consortium application. The ontology simplifies the VIVO ontology and most assertions can be inferred from the VIVO ontology.

For example, in the VIVO ontology, is a person p is the pi of a grant g, the VIVO ontology would assert:

```
p beareOf r
r a PIrole
r relatedBy g
```

A new object property `isPIOF` ("is principal investigator of") is used to simplify the above assertions to the equivalent assertion below:

```
p isPIOF g
```

A TPF query of the form

```
p isPIOF *
```

returns all the grants for which p is the principal investigator. Other "shortcut" object properties include `isAuthorOf` for publications and `isCreatorOf` for datasets. Such shortcuts speed up the retrieval of metadata using TPF. The ontology is available here: https://github.com/ctsit/metabolomics_ontology

Progress and Next Steps

TPF is being used to develop web pages for metadata regarding metabolomics coming from multiple sources – Metabolomics Workbench, PubMed, and a spreadsheet describing metabolomics software tools. We have demonstrated the feasibility of using TPF to generate web pages for metabolomics. Some pages requires 500 queries. These pages load in under one second.

We are developing pages for datasets, publications and software, as well as faceted search pages for each.

The approach described here is easily extended to support:

1. **Internationalization.** The client can accept a language parameter that is used to filter triples to those containing the preferred language, and any secondary languages desired. This feature is not currently needed for the metabolomics application.
2. **Cross-site TPF.** A web page can have any number of JavaScript clients, each client associated with a specified TPF endpoint. References in one VIVO to triples in a second VIVO can result in the the creation of a second client for the second VIVO and retrieval of desired triples from the second VIVO using TPF. This capability is not currently needed for the metabolomics application, but can be demonstrated for use in future applications.

This poster is the work of the Metabolomics Consortium Coordinating Center (M3C) at the University of Florida and is funded in part by the grant 1U2CDK119889-01 of the NIH Common Fund Metabolomics Program. The content of this poster is solely the responsibility of the M3C and does not necessarily represent the official views of the NIH.

Metabolomics Info
An international resource for news and events in metabolomics

UF UNIVERSITY of FLORIDA