Jay *et al. BMC Res Notes* (2018) 11:883
https://doi.org/10.1186/s13104-018-3997-y

BMC Research Notes

DATA NOTE

Open Access

# Connecting nutrition composition measures to biomedical research

Jeremy J. Jay*, Alexa Sanders, Robert W. Reid and Cory R. Brouwer

**Abstract**

**Objectives:** Biomedical research is gaining ground on human disease through many types of "omics", which is leading to increasingly effective treatments and broad applications for precision medicine. The majority of disease treatments still revolve around drugs and biologics. Although food is consumed in much higher quantities, we understand very little about how the human body metabolizes and uses the full range of nutrients, or how these processes affect human health and disease risk. Nutrient composition databases are used by dietitians to describe common consumer food products, but these fail to identify chemicals with the same nomenclature as metabolic pathways in basic life sciences research and with far less precision. Consumer-oriented nutrient compositions often describe generic substances (e.g. Sugars) while scientific reporting is often much more specific (e.g. Dextrose, Fructose, etc.). Integrating these two fields of research presents a difficult challenge for novel applications of precision nutrition.

**Data description:** This data set provides a manually curated collection of nutrient identifiers from the USDA's Nutrition Data Bases and maps them to PubChem (a resource for cheminformatics and drug discovery research), biomedical literature records in PubMed using Medical Subject Headings, biological pathways using the Chemical Entities of Biological Interest ontology.

**Keywords:** Nutrition, Nutrient, Diet, Food crops, Agriculture, Nutrition informatics

## Objective

Biomedical research is gaining ground on human disease through many types of 'omics, leading to increasingly effective treatments and broad applications for precision medicine. However, the majority of disease treatments still revolve around drugs and biologics. Food is consumed in much higher quantities and yet we understand very little of the specifics about how the human body metabolizes and uses nutrients, and how these processes affect human health and disease. Understanding this better is one of the primary goals of the Plant Pathways Elucidation Project (P2EP) [1].

The P2EP Knowledgebase aims to collect and integrate data from food crops, nutrients, biochemical pathways and reactions, and human health and disease. This is a large undertaking that will require years of effort and data collection starting with genome assembly for many crop plants, and gene mapping and annotation to determine pathway representation. In order to jumpstart the informatics component of this project, we have integrated existing consumer-oriented nutrient composition databases with life sciences and biomedical resources.

Nutrient composition databases [2, 3] often describe food nutrients at a high level of generality (e.g. "Sugars") while scientific reporting is often much more specific (e.g. "Dextrose", "Fructose", etc.). In addition, consumer reporting lists broad classifications that are not immediately useful in a scientific context (e.g. the broad spectrum of chemicals denoted by "trans fats"). Translating between these two worlds is a laborious process, but immensely valuable as more work is done towards precision nutrition.

Nutrition intersects various biomedical domains: basic science research, biomedical literature, and drug discovery and cheminformatics. There are many biological pathways resources such as the Gene Ontology [4] and Reactome Knowledgebase [5], which can be associated

*Correspondence: Jeremy.Jay@uncc.edu
Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, NC, USA

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

with the Chemical Entities of Biological Interest (ChEBI) ontology [6]. The Medical Subject Headings (MeSH) are a comprehensive terminology for the biomedical literature applied to the over 28 million articles in PubMed [7]. PubChem is a large resource providing further links to various cheminformatics and drug discovery data sets, tools, and applications [8]. Including these three public resources allows our mappings to be used with a wide variety of scientific resources.

## Data description

Our baseline sample of nutrient identifiers and names were extracted from multiple USDA Nutrient Databases. The primary resource was the Standard Reference 28 [2], which was supplemented with the contents of the Special Interest Databases on Flavonoids [3]. Together these records contained 188 unique nutrient identifiers related to 7793 foods in the standard reference database. Table 1 contains a listing of the resources produced.

The list of Nutrient Identifiers and Names were given to two student interns of the Plant Pathways Elucidation Project (P2EP) [1]. The students were instructed to independently search the ChEBI web portal [6] and record the appropriate matching identifiers, making sure to select the appropriate biologically relevant enantiomers and avoid erroneous ions. As expected for some fatty acids, naming conventions varied between USDA documentation and the ChEBI nomenclature, for which manual research and annotation was performed to confirm. Student results were then cross-compared to verify and investigate differences, then confirmed by a domain scientist. After an initial technical validation against pathway databases, we modified the ChEBI annotations to use more general ancestor classifications for many nutrients, which allowed for variation in ambiguity, bioavailability, and in turn greater access to other linked resources. Approximately 22% of NDB Nutrients have no ChEBI equivalent.

The prior techniques were then repeated for assignment to Medical Subject Headings (MeSH) descriptors [7]. When multiple terms were available, we preferred the term with higher occurrence in PubMed metadata records (e.g. elemental "Phosphorus" has over 10 × more annotations than "Phosphorus, Dietary"). These counts were accessed from the MeSH browser at the web portal using the "Related Information" sidebar link to PubMed. Approximately 57% of NDB Nutrients have no equivalent MeSH identifier, as expected since MeSH is used for generic document annotation and not pathway mapping.

Finally, links to the PubChem Compound database [8] were collected using the prior mappings to MeSH and ChEBI as starting points. When multiple PubChem Compound IDs were available, the CID with the most appropriate IUPAC name was selected. PubChem records were also cross-referenced with AOAC International references as specified in the NDB documentation and Chemical Abstracts Service (CAS) identifiers [9]. Approximately 33% of NDB Nutrients have no PubChem annotations.

Interestingly, none of the resources annotated contains a complete matching to all of the USDA Nutrients listed. Twenty-nine of the NDB nutrients have no annotations to ChEBI, MeSH, or PubChem. While in some cases this can be attributed to the generic consumer-oriented nutrient classifications (8/29 NDB identifiers are terms such as Ash, Energy, Sugars, etc.), the remaining instances may be classifications that warrant further consideration into the scientific databases. Now that this resource exists, we can do further research into the properties and connections of these nutrients and find more information about the effects of both known and uncharacterized phytochemicals on human health.

## Limitations

The high specificity of scientific terminology does not map cohesively to the current level of ambiguity in the Nutrition Composition databases. While it is not immediately obvious that adding new generic terminology to existing resources like ChEBI or MeSH is beneficial, adding more comprehensive lists of chemical variations with similar bioavailability would be desirable for increasing utility and accessibility of more diverse data sets.

**Table 1 Overview of data files/data sets**

| Label | Name of data set | File types (file extension) | Data repository and identifier (DOI) |
|---|---|---|---|
| Data set 1 | MeSH to USDA NDB mappings | .csv | https://doi.org/10.6084/m9.figshare.7033190 [11] |
| Data set 2 | ChEBI to USDA NDB mappings | .csv | https://doi.org/10.6084/m9.figshare.7033208 [12] |
| Data set 3 | PubChem CID to USDA NDB mappings | .csv | https://doi.org/10.6084/m9.figshare.7033217 [13] |

### Abbreviations

USDA: United States Department of Agriculture; NDB: Nutrient Database (aka USDA National Nutrient Database for Standard Reference); MeSH: Medical Subject Headings; ChEBI: Chemical Entities of Biological Interest; CAS: Chemical Abstracts Service; P2EP: Plant Pathways Elucidation Project; IUPAC: International Union of Pure and Applied Chemistry.

### Authors' contributions

JJ developed the overall project design, directed student interns, verified results, and wrote the manuscript. AS and RR provided feedback on the results and additional insights into problem areas. CB recruited interns and directs the bioinformatics components of the P2EP. All authors read and approved the final manuscript.

### Acknowledgements

We wish to thank the P2EP interns Samantha Kaiser and Tyler Robbins for their valuable contributions to this project.

### Competing interests

The authors declare that they have no competing interests.

### Availability of data and materials

The data described in this Data note can be freely and openly accessed on FigShare using Collection https://doi.org/10.6084/m9.figshare.c.4322714.v1 [10]. Individual data sets listed in Table 1 can be found at DOIs https://doi.org/10.6084/m9.figshare.7033190.v1, https://doi.org/10.6084/m9.figshare.7033208.v1, and https://doi.org/10.6084/m9.figshare.7033217.v1 [11–13].

### Consent for publication

Not applicable.

### Ethics approval and consent to participate

Not applicable.

### Funding

This research was supported by the University of North Carolina at Charlotte and is part of the Plant Pathways Elucidation Project's (p2ep.org) ongoing effort to improve nutrition through plant genomics-based strategies. This support included direction and feedback on the project goals and design, and student support, but did not influence the results.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 1 November 2018 Accepted: 6 December 2018
Published online: 12 December 2018

### References

1. Reid RW, Brouwer CR, Jackson EW, Lila MA. A need for a transdisciplinary environment: the Plant Pathways Elucidation Project. Trends Plant Sci. 2014;19(8):485–7. http://www.sciencedirect.com/science/article/pii/S136013851400171X.
2. US Department of Agriculture (USDA), Agricultural Research Service, Nutrient Data Laboratory. USDA National Nutrient Database for Standard Reference, Legacy. http://www.ars.usda.gov/nutrientdata. Accessed Oct 8 2018.
3. US Department of Agriculture (USDA), Agricultural Research Service, Nutrient Data Laboratory. USDA Database for the Flavonoid Content of Selected Foods, Release 3.1. 2013. https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/nutrient-data-laboratory/docs/usda-database-for-the-flavonoid-content-of-selected-foods-release-31-december-2013/. Accessed Oct 8 2018.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
5. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55.
6. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016;44(D1):D1214–9. https://academic.oup.com/nar/article/44/D1/D1214/2502583. Accessed Oct 8 2018.
7. Rogers FB. Communications to the editor. Bull Med Libr Assoc. 1963;51(1):114–6.
8. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res. 2016;44(Database issue):D1202–13.
9. Huffenberger MA, Wigington RL. Chemical Abstracts Service approach to management of large data bases. J Chem Inf Comput Sci. 1975;15(1):43–7.
10. Jay J. Nutrient Composition to Chemical resources. 2018. https://doi.org/10.6084/m9.figshare.c.4322714.v1. Accessed Oct 8 2018.
11. Jay J. MeSH to USDA NDB mappings [Internet]. 2018. https://doi.org/10.6084/m9.figshare.7033190.v1. Accessed cited 2018 Oct 8.
12. Jay J. ChEBI to USDA NDB mappings. 2018. https://doi.org/10.6084/m9.figshare.7033208.v1. Accessed Oct 8 2018.
13. Jay J. PubChem CID to USDA NDB mappings. 2018. https://doi.org/10.6084/m9.figshare.7033217.v1. Accessed Oct 8 2018.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

**Learn more** biomedcentral.com/submissions

BMC