

RESEARCH NOTE

Open Access



The Aliment to Bodily Condition knowledgebase (ABCkb): a database connecting plants and human health

Aaron Trautman^{1,2} , Richard Linchangco^{1,2}, Rachel Walstead², Jeremy J. Jay^{1,2} and Cory Brouwer^{1,2*}

Abstract

Objective: Overconsumption of processed foods has led to an increase in chronic diet-related diseases such as obesity and type 2 diabetes. Although diets high in fresh fruits and vegetables are linked with healthier outcomes, the specific mechanisms for these relationships are poorly understood. Experiments examining plant phytochemical production and breeding programs, or separately on the health effects of nutritional supplements have yielded results that are sparse, siloed, and difficult to integrate between the domains of human health and agriculture. To connect plant products to health outcomes through their molecular mechanism an integrated computational resource is necessary.

Results: We created the Aliment to Bodily Condition Knowledgebase (ABCkb) to connect plants to human health by creating a stepwise path from plant → plant product → human gene → pathways → indication. ABCkb integrates 11 curated sources as well as relationships mined from Medline abstracts by loading into a graph database which is deployed via a Docker container. This new resource, provided in a queryable container with a user-friendly interface connects plant products with human health outcomes for generating nutritive hypotheses. All scripts used are available on github (<https://github.com/atrautm1/ABCkb>) along with basic directions for building the knowledgebase and a browsable interface is available (<https://abckb.charlotte.edu>).

Keywords: Knowledgebase, Precision nutrition, Graph database, Neo4j, Text mining, Knowledge discovery

Introduction

The growth of obesity worldwide correlates strongly with overconsumption of processed foods [1]. This has contributed to an increase in chronic diet-related diseases like type 2 diabetes (T2D), heart disease, and some cancers [2]. Exercise and diets high in fruit, vegetables, whole grains, and nuts have been linked with healthier outcomes and reduce the risk of developing these diseases [3]. Unfortunately, the specific mechanisms driving these associations are poorly understood. The Plant Pathways Elucidation Project (P2EP) was a collaboration

started to uncover the mechanisms between plant-pathway products and human health [4]. Three questions drove this collaboration: “What do plants make,” “How do they make them,” and “What is their effect on human health?” The ABCkb was developed to capture the information required to answer these questions and provide researchers with a tool to build informed, nutritive hypotheses with molecular mechanisms as the linking factor between dietary plants and human health.

These questions closely align to the recently released “2020–2030 Strategic Plan for NIH Nutrition Research.” This plan contains 4 strategic goals for further study to move closer to a precision nutrition approach including foundational research into “What do we eat and how does it affect us?” as well as understanding “How can we improve the use of food as medicine?” A cornerstone for

*Correspondence: cory.brouwer@uncc.edu

² Department of Bioinformatics and Genomics, UNC Charlotte, Charlotte, NC, USA

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. 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 in a credit line to the data.

answering these questions and the questions of the P2EP collaboration is an understanding of the mechanism of action of how our diet affects our health.

However, manually capturing this information is a difficult, time-consuming task due to scattered bodies of scientific knowledge. Currently available resources contain partial information to answer these questions, but they do not address mechanism of action. For example, the Comparative Toxicogenomics Database (CTD) connects chemicals to human health through human genes by manually curating associations between chemicals, genes, pathways and phenotypes but excludes nutritional data [5]. Specialized nutritional databases like FooDB (<https://foodb.ca>) and Phenol-Explorer aid researchers in estimating quantity of phytochemical content, but lack human phenotypic information [6]. NutriChem was developed to bridge the gap between plant-based nutrition and human disease through the chemicals contained in those plants, but does not contain gene-chemical associations, a key part of the driving molecular mechanisms between diet and human health [7]. While a small proportion of assertions are in available databases, others are hidden in published research and can only be extracted through extensive reading or by natural language processing (NLP) the literature. Given the rise in diet-related diseases, and the pursuit of personalized nutrition, an integrated resource to develop nutritive hypotheses is necessary.

Main text

We have developed the Aliment to Bodily Condition Knowledgebase (ABCkb) to address the gap of connecting plant compounds to human indications through their mechanism of action. The ABCkb integrates multiple resources for building informed hypotheses with molecular mechanisms as the linking factor between dietary plants and human health. To accomplish this, the ABCkb uses both structured and unstructured data sources (Fig. 1). The structured resources are publicly accessible, curated databases and the unstructured data is in the form of Medline abstracts. Since this data, composed of entities and relationships or nodes and edges, composes a graphical network, we extracted, transformed, and then loaded into a Neo4j graph database. To help users begin discovering these nutritive connections, the knowledgebase is available on GitHub and a simplified online web interface.

Structured resource collection

Structured data from 11 resources (Additional file 1: Fig. S1) produce five major node types (Plant, Chemical, Gene, Pathway, Phenotype) in a Neo4j graph database. Connections, or edges between these nodes are provided by both structured data, and unstructured MEDLINE Abstracts through NLP. The ABCkb utilizes three types of structured data sources: ontologies, structured vocabularies, and databases.

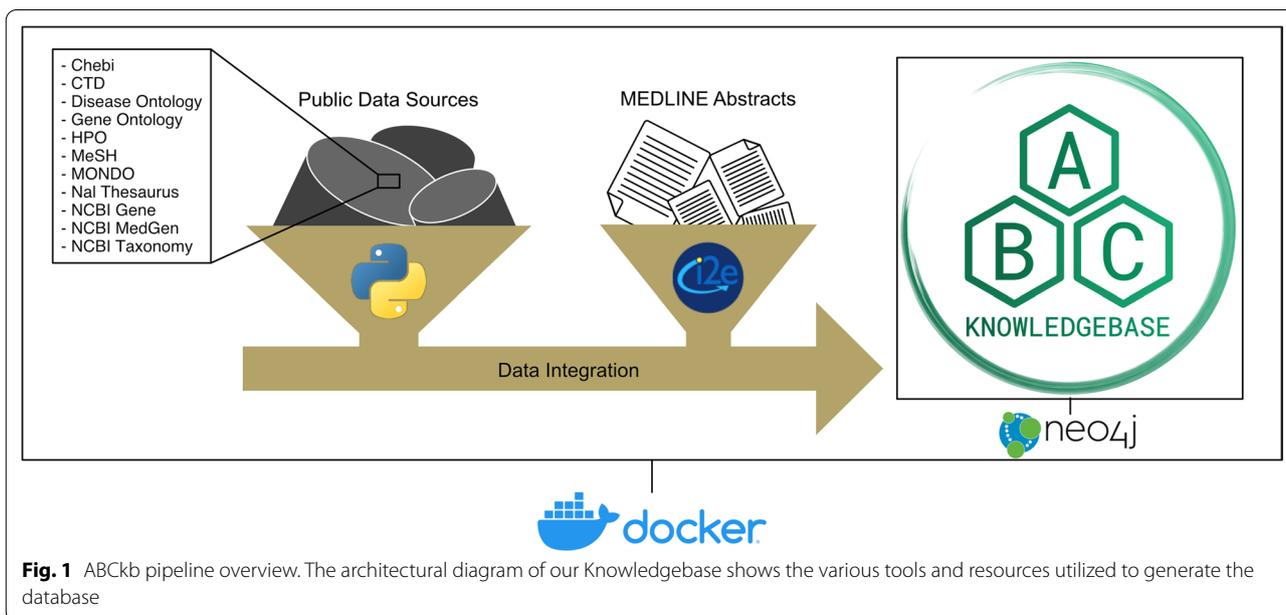


Fig. 1 ABCkb pipeline overview. The architectural diagram of our Knowledgebase shows the various tools and resources utilized to generate the database

Ontologies and structured vocabularies

The ontologies and structured vocabularies create well-controlled edges between chemicals, pathways, and phenotypes. The Chemical Entities of Biological Interest provide chemical nodes and semantic connections (edges) between chemicals [8]. Genes are grouped into pathways from the Gene Ontology resource [9, 10]. Human phenotypes are represented from three sources. The Disease Ontology categorizes human diseases with phenotypic characteristics [11]. The Human Phenotype Ontology provides phenotypic abnormalities not found within the Disease Ontology which allows researchers to focus on specific phenotypic symptoms and the associated molecular mechanisms [12]. Finally, the MONDO Disease Ontology was used to collapse similar phenotype nodes from multiple sources using their source identifiers [13]. The Medical Subject Headings resource provided nodes and connections for all major labels with the exception of Genes [14]. Additional plant, chemical, and phenotype nodes were extracted from the National Agricultural Library Thesaurus [15]. Terms from different ontologies or vocabularies with the same identifiers are collapsed into the same node. All other nodes are left separate to retain their hierarchical relationships.

Databases

Several databases were utilized to increase molecular mechanisms from plant to human disease in the ABCkb. The Comparative Toxicogenomics Database added over 7.4 million manually curated edges between chemicals, genes, pathways, and phenotype nodes [5]. We utilized three public databases from The National Center for Biotechnology Information. All plants under the Embryophyta clade from the NCBI Taxonomy database produced plant nodes and phylogenetic relationships between plants [16, 17]. The Gene database provided gene names, types, and synonyms [18]. Finally, additional edges were added utilizing NCBI gene nodes and MONDO phenotypes were extracted from the NCBI MedGen database [19]. The compendium of structured data sources provide many of the node and edges connecting plants to disease. However, unstructured literature contains informative relationships not contained within these sources, leaving many gaps in our understanding.

Unstructured NLP collection

To uncover relationships in literature, elucidate molecular mechanisms, and answer the three questions of the P2EP, we mined the literature using Linguamatics' I2E NLP text mining platform (<https://www.linguamatics.com/products/i2e>). This platform utilizes ontologies and structured vocabularies to transform unstructured text into structured assertions (nodes and edges).

Natural Language Processing of MEDLINE Abstracts

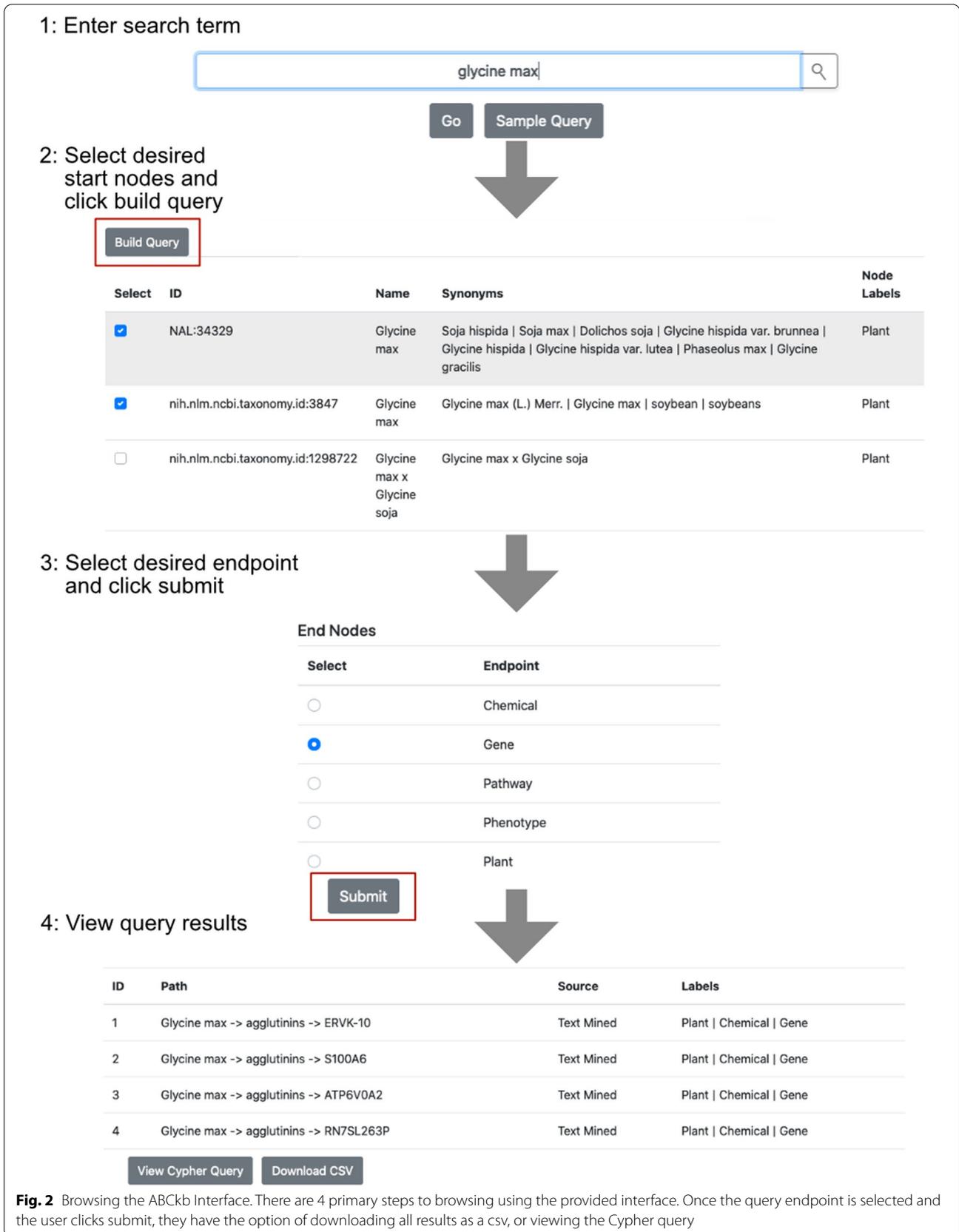
The I2E platform employs a graphical user interface for NLP query development, where each query extracts a set of subjects, objects, and predicates, or relationships from user-specified ontologies and structured vocabularies. From published abstracts, and titles extracted from MEDLINE in May, 2019, NLP queries were developed with I2E for each of the 4 steps (plant to chemical, chemical to gene, gene to pathway, pathway to phenotype) with an additional query from genes to phenotypes. All I2E assertions generated are provided to users of the ABCkb as source files and are parsed when the graph database is built.

Statistics and application

Extracted public data sources generated over 957,000 nodes with over 11 million edge relationships. NLP results from I2E queries make up 1.26 million of the overall relationship count, of which 1.25 million relationships were novel, not from structured public data sources. Additional file 2: Fig. S2 gives a visual presentation of (a) the relative number of each node type and their source, (b) the edge relationships from each source and (c) the relative comparison of edge relationship types between each type of node.

This collection of nodes and edge relationships forming semantic triples, naturally forms a biological network of knowledge that is best stored in a graph database like Neo4j. Chaining these triples together in the ABCkb highlights connections between dietary plants and human phenotypes that would otherwise go unseen if left in their original sources, particularly unstructured literature sources. The intention of the knowledgebase is for information in the network to flow from plants to phenotypes/disease indications, however, assertions are maintained in both directions, which allows for query flexibility of relationships between any nodes. Start and end node types are not enforced which allows queries from any point, to any point. All associations are kept along with references to the original source allowing the user to evaluate potential inconsistencies using the original evidence. To explore the database and discover connections, users have two choices. One, use the online interface (available at <https://abckb.charlotte.edu>). Otherwise, download from GitHub and build the database on a local machine which can then be queried in the Neo4j interface, or on the command line. A prebuilt data folder with the neo4j database is also available [20].

The provided user-friendly interface aids users unfamiliar with Neo4j query language (Cypher) to browse the contents within and examine nutritive connections (Fig. 2). On the home page, users are provided a search box to enter in a search term. This scans the nodes in the



knowledgebase and returns results ranked by similarity to search term. Users can select nodes and continue to build a query to any end point within the knowledgebase (plant, chemical, gene, pathway, or phenotype). Running the query scans the database for all paths to the selected end point and returns them to the user, which are available to download. Additionally, a Cypher query is available to users that can be used in the built in Neo4j interface or the terminal for further exploration.

Oat and T2D

To demonstrate how the ABCkb connects dietary plants to separate human indications through molecular mechanisms, a graph was created in the ABCkb, through the Neo4j browser, depicting the diet-disease network between *Avena sativa*, T2D, and heart failure (Fig. 3). The detailed associations are in the attached supplementary file (Additional file 3: File S3). Connections from the CTD indicate genes commonly associated with cholesterol and heart failure. However, text-mining indicates that consumption of oats affects cholesterol levels in the body, which is associated with the gene HSD11B1 that affects lipid metabolic processes with both positive and negative impacts on the incidence of T2D. These relationships are due to the presence of beta-glucan in oat grains. Consumption of beta-glucan-containing oat can help lower LDL cholesterol [21]. The cholesterol lowering effects of oat can also be attributed to the presence of

certain lipids and proteins [22]. The proteins in oat with low lysine-arginine and methionine-glycine ratios contribute to lower total cholesterol and LDL cholesterol levels. Hypocholesterolemic properties of oat cannot simply be attributed to one factor, but a combination of many, including oleic acid, vitamin E, and plant sterols [22].

T2D patients frequently have abnormal levels of many different lipids, as well as abnormal qualities to these lipids, for example, T2D patients experience normal or slightly elevated LDL cholesterol with increased LDL oxidation and glycation [23]. Dyslipidemia in T2D patients is associated with cardiovascular disease [24, 25]. This creates an elevated risk for cardiovascular diseases including atherosclerosis, and dislipidemia may play a role in these risks [25]. In the graph, HSD11B1 is the human gene connecting this relationship. HSD11B1 expression is increased in adipose tissues of obese individuals [26]. Dysregulation of HSD11B1 is associated with an imbalance of glucocorticoid in adipose tissues, glucose imbalance, and visceral fat accumulation [27]. These factors contribute to metabolic syndrome, which puts patients at a higher risk for cardiac diseases [28]. Various SNPs in HSD11B1 have associations with T2D, metabolic syndrome, and hypertension [29–32].

Due to the established relationship between oat beta-glucans, cholesterol, and weight, the connection to T2D is logical [21, 27]. Decreased weight, specifically visceral fat in the abdomen, would result in reduced expression

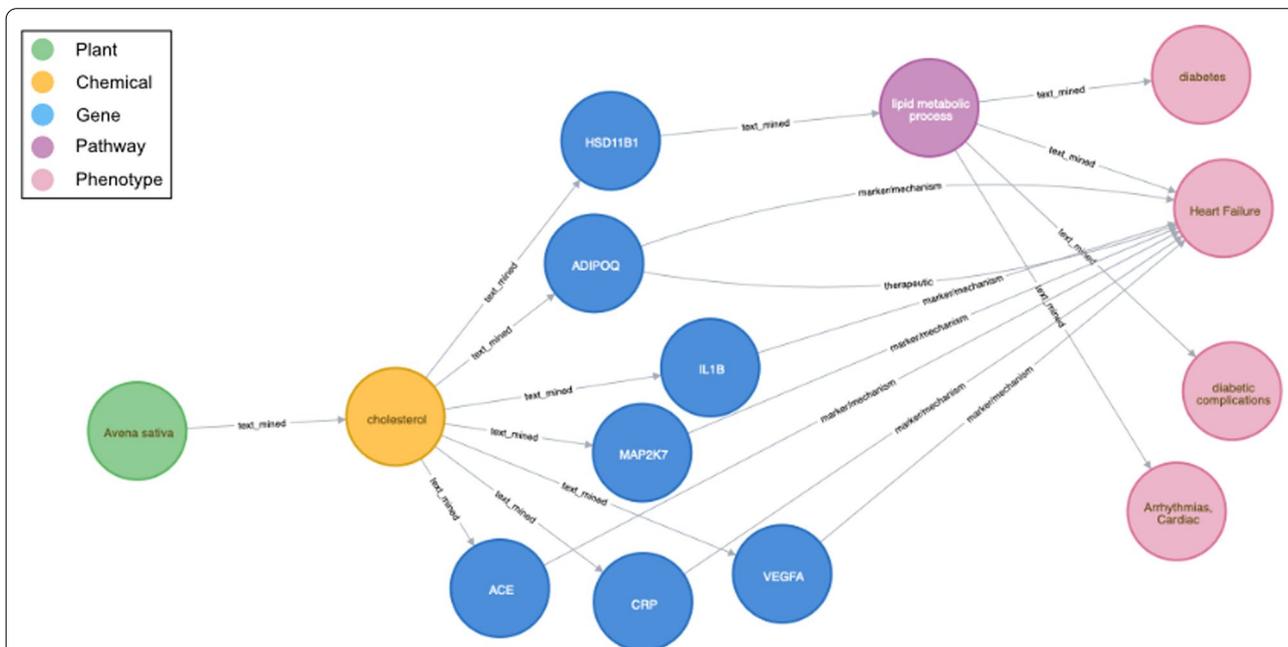


Fig. 3 Visualizing the results of Avena sativa to diabetes and heart failure via Hydroxysteroid 11-Beta Dehydrogenase 1. This meta-path highlights the connectivity between oats, diabetes, and heart failure through the gene HSD11B1 from the ABCkb

of HSD11B1, which would improve regulation of cortisol. Further examination of the oat–cholesterol–HSD11B1 relationship could be very informative to both patients and doctors in making more informed dietary choices and reducing the risk of developing T2D. This example demonstrates the ABCkb ability to connect seemingly separate conditions through the molecular mechanistic links within.

Discussion

The ABCkb integrates structured and unstructured resources in a network that connects plants to human disease through molecular mechanisms. This reduces the time required to manually connect these links through each individual resource. Additionally, knowledge discovery is aided by the development of a user-friendly interface. All of these components provide precision nutrition a path to better understand the mechanisms behind diet-related conditions. The ABCkb is available from the interface (<https://abckb.charlotte.edu>).

Limitations

- Microbiota contributions to diet and human disease. Bacteria within the gut are known to affect disease both through the production of metabolites and the conversion of plant phytochemicals. In addition, gut bacteria are affected by diet. Future implementations of the ABCkb will contain microbiota associations to enhance precision nutrition hypotheses.
- Mining abstracts versus full text. Abstracts contain valuable associations, however associations full text articles would provide a greater number of associations.
- Incorporating genomic data. Precision nutrition hypotheses and treatment plans will depend on patient genomic data, to provide optimal dietary solutions for each individual. Future versions of the ABCkb should incorporate human genomic data.

Abbreviations

ABCkb: Aliment to Bodily Condition knowledgebase; P2EP: Plant Pathways Elucidation Project; CTD: Comparative Toxicogenomics Database; NLP: Natural language processing; LDL: Low-Density Lipoprotein; T2D: Type 2 diabetes.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13104-021-05835-x>.

Additional file 1: Figure S1. ABCkb data sources. Data from each source is transformed into one of the 5 labels and may provide external and internal references to nodes within the knowledgebase. The CTD provides manually curated references between labels with no original node labels.

Additional file 2: Figure S2. ABCkb node and relationship statistics.

a) The pie chart shows primary labels indicated by color with named secondary (source) labels, shaded and sized by proportion of total nodes in the knowledgebase. b) The sum of relationship counts for each source is indicated by the bar chart. c) Relative relationship counts indicated from node-node in rows, columns in a bar chart in order by type (Internal Descriptor, External Connector, Cross Reference, and Text Mined).

Additional file 3: File S3. Query node and relationship information. The query from the application portion from *Avena sativa* to Heart Disease and Diabetes resulted in the nodes and relationships as previously discussed. This file contains the more detailed information contained in the Neo4j database about the nodes and the relationship connections.

Acknowledgements

Steven Blanchard, the P2EP program, and the many interns that aided the development of this resource.

Authors' contributions

AT and RL designed and built the ABCkb, AT, RW, and CB wrote the manuscript, JJ and CB provided guidance, all authors provided corrections and feedback. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The browsable interface is available at: <https://abckb.charlotte.edu> All source code is available on the project github (<https://github.com/atratm1/ABCkb>) along with basic instructions for building and running the knowledgebase. The prebuilt data folder [20] with the corresponding neo4j database is available from: <https://doi.org/10.6084/m9.figshare.13653539.v1> Operating system(s): Platform independent Database version: Neo4j 3.5 License: GNU GPL v3.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Bioinformatics Services Division, UNC Charlotte, Charlotte, NC, USA. ²Department of Bioinformatics and Genomics, UNC Charlotte, Charlotte, NC, USA.

Received: 23 March 2021 Accepted: 3 November 2021

Published online: 27 November 2021

References

1. Laster J, Frame LA. Beyond the calories-is the problem in the processing? *Curr Treat Options Gastroenterol.* 2019;17(4):577–86. <https://doi.org/10.1007/s11938-019-00246-1>.
2. Popkin BM. Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases. *Am J Clin Nutr.* 2006;84(2):289–98.

3. Schulze MB, Martínez-González MA, Fung TT, Lichtenstein AH, Forouhi NG. Food based dietary patterns and chronic disease prevention. *BMJ*. 2018;361:k2396.
4. 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.
5. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wieggers J, et al. The comparative toxicogenomics database: update 2019. *Nucleic Acids Res*. 2019;47(D1):D948–54.
6. Rothwell JA, Perez-Jimenez J, Neveu V, Medina-Remón A, M'Hiri N, García-Lobato P, et al. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. Database. 2013;<https://doi.org/10.1093/database/bat070/342410>.
7. Jensen K, Panagiotou G, Kouskoumvekaki I. NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Res*. 2015;43(D1):D940–5.
8. 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.
9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25–9.
10. The Gene Ontology Consortium. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330–8.
11. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 2019;47(D1):D955–62.
12. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gouridine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res*. 2019;47(D1):D1018–27.
13. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. 2017;45(D1):D712–22.
14. Medical Subject Headings - Home Page [Product, Program, and Project Descriptions]; Library Catalog: www.nlm.nih.gov Publisher: U.S. National Library of Medicine. <https://www.nlm.nih.gov/mesh/meshhome.html>.
15. Agricultural Thesaurus and Glossary Home Page; <https://agclass.nal.usda.gov/agt.shtml>.
16. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res*. 2012;40:D136–143.
17. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. *Nucleic Acids Res*. 2019;47(D1):D94–9.
18. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*. 2015;43:D36–42.
19. Halavi M, Maglott D, Gorenkov V, Rubinstein W. MedGen. National Center for Biotechnology Information (US); 2018. Publication Title: The NCBI Handbook [Internet]. 2nd edition. <https://www.ncbi.nlm.nih.gov/books/NBK159970/>.
20. Trautman A. Aliment to bodily condition knowledgebase. figshare. 2021. <https://doi.org/10.6084/m9.figshare.13653539.v1>.
21. Wolever TMS, Duss DRaR. Oat beta-glucan reduces serum LDL cholesterol in humans with serum LDL cholesterol $\leq 160\text{mg/dL}$; 2016. <http://www.eurekaselect.com/142731/article>.
22. Guo L, Tong LT, Liu L, Zhong K, Qiu J, Zhou S. The cholesterol-lowering effects of oat varieties based on their difference in the composition of proteins and lipids. *Lipids Health Dis*. 2014;13:182.
23. Vergès B. New insight into the pathophysiology of lipid abnormalities in type 2 diabetes. *Diabetes Metab*. 2005;31(5):429–39.
24. Pokharel DR, Khadka D, Sigdel M, Yadav NK, Acharya S, Kafle R, et al. Prevalence and pattern of dyslipidemia in Nepalese individuals with type 2 diabetes. *BMC Res Notes*. 2017;10(1):146.
25. Shahwan MJ, Jairoun AA, Farajallah A, Shanabli S. Prevalence of dyslipidemia and factors affecting lipid profile in patients with type 2 diabetes. *Diabetes Metab Syndr*. 2019;13(4):2387–92.
26. Paulsen SK, Pedersen SB, Fisker S, Richelsen B. 11Beta-HSD type 1 expression in human adipose tissue: impact of gender, obesity, and fat localization. *Obesity*. 2007;15(8):1954–60.
27. Dammann C, Stapelfeld C, Maser E. Expression and activity of the cortisol-activating enzyme 11beta-hydroxysteroid dehydrogenase type 1 is tissue and species-specific. *Chem Biol Interact*. 2019;303:57–61.
28. Turek LV, Leite N, Rodrigues Souza RL, Lima JK, Milano GE, Timossi LS, et al. Gender-dependent association of HSD11B1 single nucleotide polymorphisms with glucose and HDL-C levels. *Genet Mol Biol*. 2014;37(3):490–5.
29. Nair S, Lee YH, Lindsay RS, Walker BR, Tataranni PA, Bogardus C, et al. 11beta-Hydroxysteroid dehydrogenase Type 1: genetic polymorphisms are associated with type 2 diabetes in Pima Indians independently of obesity and expression in adipocyte and muscle. *Diabetologia*. 2004;47(6):1088–95.
30. Gambineri A, Tomassoni F, Munarini A, Stimson RH, Mioni R, Pagotto U, et al. A combination of polymorphisms in HSD11B1 associates with in vivo 11beta-HSD1 activity and metabolic syndrome in women with and without polycystic ovary syndrome. *Eur J Endocrinol*. 2011;165(2):283–92.
31. Freedman DS, Bowman BA, Srinivasan SR, Berenson GS, Otvos JD. Distribution and correlates of high-density lipoprotein subclasses among children and adolescents. *Metab Clin Exp*. 2001;50(3):370–6.
32. Goff DC, D'Agostino RB, Haffner SM, Otvos JD. Insulin resistance and adiposity influence lipoprotein size and subclass concentrations. Results from the Insulin Resistance Atherosclerosis Study. *Metab Clin Exp*. 2005;54(2):264–70.

Publisher's Note

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

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

