Journal of proteome research

Technical Note

Cite This: J. Proteome Res. 2018, 17, 4345−4357

pubs.acs.org/jpr

A Cloud-Based Metabolite and Chemical Prioritization System for the Biology/Disease-Driven Human Proteome Project

Kun-Hsing Yu,[†,‡] Tsung-Lu Michael Lee,[§] Yu-Ju Chen,[∥] Christopher Ré,[⊥] Samuel C. Kou,[‡] Jung-Hsien Chiang,[*,#] Michael Snyder,[*,∇,○] and Isaac S. Kohane[*,†,○]

[†]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, United States

[‡]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, United States

[§]Department of Information Engineering, Kun Shan University, Tainan City 710, Taiwan

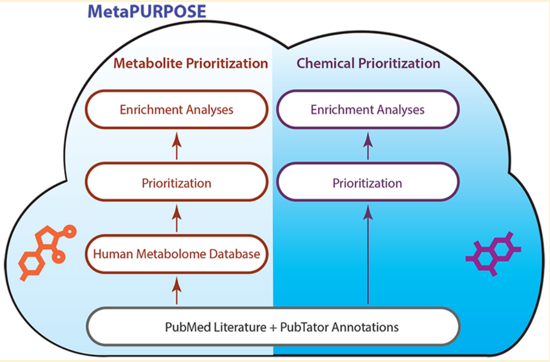
[∥]Institute of Chemistry, Academia Sinica, Taipei City 115, Taiwan

[⊥]Department of Computer Science, Stanford University, Stanford, California 94305, United States

[#]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan City 701, Taiwan

[∇]Department of Genetics, School of Medicine, Stanford University, Stanford, California 94305, United States

S *Supporting Information*



ABSTRACT: Targeted metabolomics and biochemical studies complement the ongoing investigations led by the Human Proteome Organization (HUPO) Biology/Disease-Driven Human Proteome Project (B/D-HPP). However, it is challenging to identify and prioritize metabolite and chemical targets. Literature-mining-based approaches have been proposed for target proteomics studies, but text mining methods for metabolite and chemical prioritization are hindered by a large number of synonyms and nonstandardized names of each entity. In this study, we developed a cloud-based literature mining and summarization platform that maps metabolites and chemicals in the literature to unique identifiers and summarizes the copublication trends of metabolites/chemicals and B/D-HPP topics using Protein Universal Reference Publication-Originated Search Engine (PURPOSE) scores. We successfully prioritized metabolites and chemicals associated with the B/D-HPP targeted fields and validated the results by checking against expert-curated associations and enrichment analyses. Compared with existing algorithms, our system achieved better precision and recall in retrieving chemicals related to B/D-HPP focused areas. Our cloud-based platform enables queries on all biological terms in multiple species, which will contribute to B/D-HPP and targeted metabolomics/chemical studies.

KEYWORDS: *metabolomics, chemicals, Biology/Disease-Driven Human Proteome Project, literature mining, Protein Universal Reference Publication-Originated Search Engine (PURPOSE), Finding Associated Concepts with Text Analysis (FACTA+), Biomedical Entity Search Tool (BEST)*

■ INTRODUCTION

The Human Proteome Organization (HUPO) Biology/Disease-Driven Human Proteome Project (B/D-HPP) is a coordinated comprehensive proteomics profiling effort that focuses on human biology and diseases.[1−3] Investigations of metabolites and chemicals associated with human biology and diseases can enhance and complement the ongoing studies on B/D-HPP.[1] With the advancement of targeted assays, researchers can quantify hundreds of metabolites or chemical compounds simultaneously.[4] These high-throughput approaches have the potential to characterize the chemical landscape of human biology in various organs and identify metabolomics disturbances under disease conditions,[5,6] which will contribute to a holistic understanding of biology and diseases.

Similar to proteomics studies, target prioritization is crucial for targeted metabolomics and chemical investigations.[7] There are more than tens of thousands of metabolites and hundreds of thousands of exogenous and endogenous chemicals;[8] however, many modern targeted assays can handle only hundreds to thousands of targets at a time.[9] In order to maximize the utility of the targeted approaches, it is crucial to prioritize the metabolites and chemicals relevant to the study. Previously, researchers have proposed computational approaches to prioritize proteins using literature mining

Special Issue: Human Proteome Project 2018

Received: May 28, 2018
Published: August 10, 2018

DOI: 10.1021/acs.jproteome.8b00378
J. Proteome Res. 2018, 17, 4345−4357

algorithms.[10−12] Nevertheless, because of the plethora of metabolites and chemicals, a comprehensive tool for their prioritization is lacking. In addition, many metabolites and chemicals have a great number of synonyms and nonstandardized names,[13,14] which has hindered the development of automated approaches for their identification.[15]

Recent studies have presented efficient algorithms that summarize the strength and specificity of protein-topic copublication patterns in the PubMed literature.[11,12] Such methods prioritize the associations between any topic and any protein in the PubMed abstract. With the ongoing curation efforts of the Human Metabolome Database (HMDB),[8] Chemical Entities of Biological Interest (ChEBI),[16] and updates in the Medical Subject Headings (MeSH),[17] there is an opportunity to extend the literature mining algorithms to characterize the relations between metabolites/chemicals and any search topic systematically.

In this study, we implemented a cloud-based system for prioritizing metabolites and chemicals for the B/D-HPP targeted fields and any custom search terms. Our system employs the state-of-the-art approach of bioentity tagging and PubMed literature mining,[18] searches the PubMed database in real time, compiles the results automatically, and ranks the retrieved metabolites and chemicals within a few seconds using an efficient copublication summarization algorithm.[12] Our system will enable comprehensive investigations of metabolites and chemicals in all targeted areas of B/D-HPP, complementing the ongoing efforts on proteomic profiling in these areas of interest.

■ METHODS

Data Retrieval for Literature Mining

The targeted areas of B/D-HPP are retrieved from the B/D-HPP Web site.[19] The identified B/D-HPP topics are brain, cancers, cardiovascular, diabetes, extreme conditions, EyeOme, food and nutrition, glycoproteomics, immune-peptidome, infectious diseases, kidney and urine, liver, mitochondria, model organisms, musculoskeletal, PediOme, plasma, protein aggregation, and rheumatic disorders. Table S-1 shows the PubMed search terms for the B/D-HPP targeted fields.

To systematically identify metabolites and chemicals from the PubMed literature, the chemical and species tags from PubTator were obtained for each PubMed article.[18] The retrieved tags were intersected with the MeSH subtrees[17] of known chemicals. Through obtaining the PubTator taggings and filtering them by the MeSH ontology tree, the unique identifier of each chemical was identified. This approach effectively mapped the synonyms of chemicals to unique identifiers. To ensure that the most updated metabolite, chemical, and species tags were retrieved, an automated downloader was implemented to retrieve PubTator data files from its File Transfer Protocol (FTP) site periodically. To enable metabolite prioritization, the list of human metabolites was retrieved from the HMDB.[8] The chemicals included in the HMDB list were employed in the metabolite prioritization tasks.

For each PubMed article with relevant tags, the NLM Entrez Programming Utilities (E-utilities)[20] were used to obtain the title, authors, journal, year of publication, and number of citations.

Metabolite and Chemical Prioritization through PURPOSE Scores

Protein Universal Reference Publication-Originated Search Engine (PURPOSE) scores were used to prioritize metabolites and chemicals for each of the B/D-HPP targeted areas.[12] The PURPOSE score is defined as

$$\left(1 + \log_{10} \mathrm{nTC} + \log_{10} \frac{\mathrm{Sum}\left(\frac{\mathrm{Cit}}{\mathrm{Year}}\right) + 1}{10}\right) \times \left(1 + \log_{10} \frac{\mathrm{nU}}{\mathrm{nT}} + \log_{10} \frac{\mathrm{nU}}{\mathrm{nC}}\right)$$

where nTC is the number of papers associated with both the topic and the chemical/metabolite (TC), Sum(Cit/Year) is the sum of yearly citation numbers of TC, nU is the number of PubMed publications, nT is the number of publications related to the topic, and nC is the number of publications associated with the chemical/metabolite. This scoring scheme accounts for the strength and the specificity of topic−chemical associations. In particular, the quantity in the first set of parentheses in the formula summarizes the frequency of the topic−chemical copublication, and the number of annualized citations is included in the algorithm to put higher weights on seminal papers and landmark studies.[12] The quantity in the second set of parentheses in the formula takes into account the overall popularity of the queried topic and the chemicals. This scoring formula is related to the term frequency−inverse document frequency statistic,[21] and a similar approach achieved superior performance in proteomics literature mining.[12]

Enrichment Analyses and Pathway Visualization

In order to identify the biological pathways associated with the retrieved chemicals and metabolites, the Search Tool for Interactions of Chemicals (STITCH) tool was employed to identify the known associations among chemicals, metabolites, genes, and proteins.[22] The STITCH tool conducts enrichment analysis on an open-source database containing 500 000 chemicals, 9.6 million proteins, and 1.6 billion interactions.[22] The database is maintained by the European Molecular Biology Laboratory, the Swiss Institute of Bioinformatics, and the Center for Protein Research.[22] Gene Ontology enrichment analyses, KEGG pathway analyses, and network analyses were performed by the STITCH tool.[22] Network statistics of the gene−metabolite and gene−chemical interaction networks, including centralization, Krackhardt efficiency, transitivity, and connectedness scores, were computed using the R package sna.[23] The centralization of a network was evaluated by Freeman's centrality score.[24] The Krackhardt efficiency score computed the proportion of necessary edges that could not be removed without disconnecting the nodes in the network. The transitivity score assessed the proportion of connections where transitivity holds (whether node A is directly connected to node C when node A is connected to node B and node B is connected to node C). The connectedness score identified the proportion of connected node pairs in the networks.[23] The Metscape app[25−27] in Cytoscape[28] was used to visualize the interactions among metabolites, genes, and enzymes. All of the analyses were conducted on May 20, 2018.

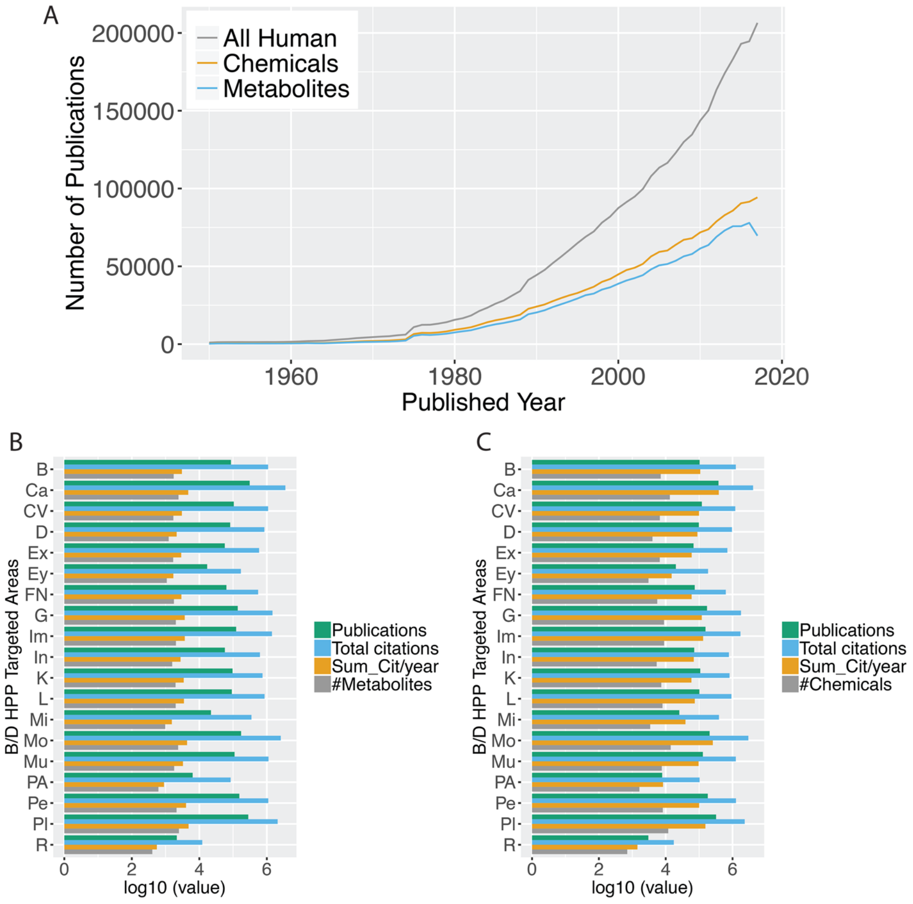


Figure 1. Summary of metabolite and chemical publication patterns in the B/D-HPP targeted areas. (A) Numbers of all PubMed publications on human, publications associated with any chemical, and publications associated with any metabolite since 1950. The numbers of PubMed publications have increased exponentially since 1975. (B) Number of publications, total citations, citations per year (Sum_Cit/year), and number of associated metabolites in each of the B/D-HPP fields. (C) Number of publications, total citations, citations per year (Sum_Cit/year), and number of associated chemicals in the B/D-HPP areas. It should be noted that in (B) and (C) the X axis is $\log_{10}$-transformed. Abbreviations: B, brain; Ca, cancers; CV, cardiovascular; D, diabetes; Ex, extreme conditions; Ey, EyeOme; FN, food and nutrition; G, glycoproteins; Im, immunepeptidome; In, infectious diseases; K, kidney and urine; L, liver; Mi, mitochondria; Mo, model organisms; Mu, musculoskeletal; PA, protein aggregation; Pe, PediOme; Pl, plasma; R, rheumatic disorders.

Metabolites/Chemicals–B/D-HPP Linkage Visualization

To summarize the linkages among the B/D-HPP and metabolites/chemicals, the correlations among B/D-HPP targeted fields and the associations between the most prominent metabolites/chemicals and the related B/D-HPP areas were visualized. For each pair of B/D-HPP targeted areas, the pairwise Spearman's correlation coefficient was computed for the associated metabolites' or chemicals' PURPOSE scores, and 1 minus the Spearman's correlation coefficient was defined as the distance between the B/D-HPP fields. Multidimensional scaling (MDS)[29] was employed to map the distances between B/D-HPP fields into a two-dimensional graph. The most prominent metabolites and chemicals were added to the resulting graph. The pairwise distances among the B/D-HPP areas reflected their correlations in the PURPOSE scores, and the connections between metabolites/chemicals and B/D-HPP areas visualized the most prominent linkages (metabolites and chemicals were shown in the graphs if their PURPOSE scores in the respective B/D-HPP areas were in the top 2.5 percentile and the scores were greater than 20). For metabolites/chemicals strongly associated with only one B/D-HPP, the distances between the metabolites/chemicals and the B/D-HPP areas were inversely proportional to their PURPOSE scores. For metabolites/chemicals strongly correlated with two or more B/D-HPP areas, the distances between the metabolites/chemicals and the associated B/D-HPP areas reflected both their PURPOSE scores in the associated B/D-HPP areas and the general correlations among the associated B/D-HPP areas. The figures were generated by R version 3.3 on the Extreme Science and Engineering Discovery Environment (XSEDE) platform.[30]

Evaluation of the Prioritization Results

Curated chemical–topic associations in the Comparative Toxicogenomics Database (CTD)[31] were employed as the ground truth for evaluating the chemical prioritization results. The precision, recall, and F1 measure (the harmonic mean of precision and recall) of the PURPOSE algorithm and those of the Finding Associated Concepts with Text Analysis (FACTA+) tool[32,33] and the Biomedical Entity Search Tool (BEST)[34] were compared. MeSH terms were used to aggregate the synonyms of a chemical. The B/D-HPP areas cancers, diabetes, rheumatic, and liver were selected as the topics for evaluation because of the availability of the curated annotations and their clean MeSH organization.

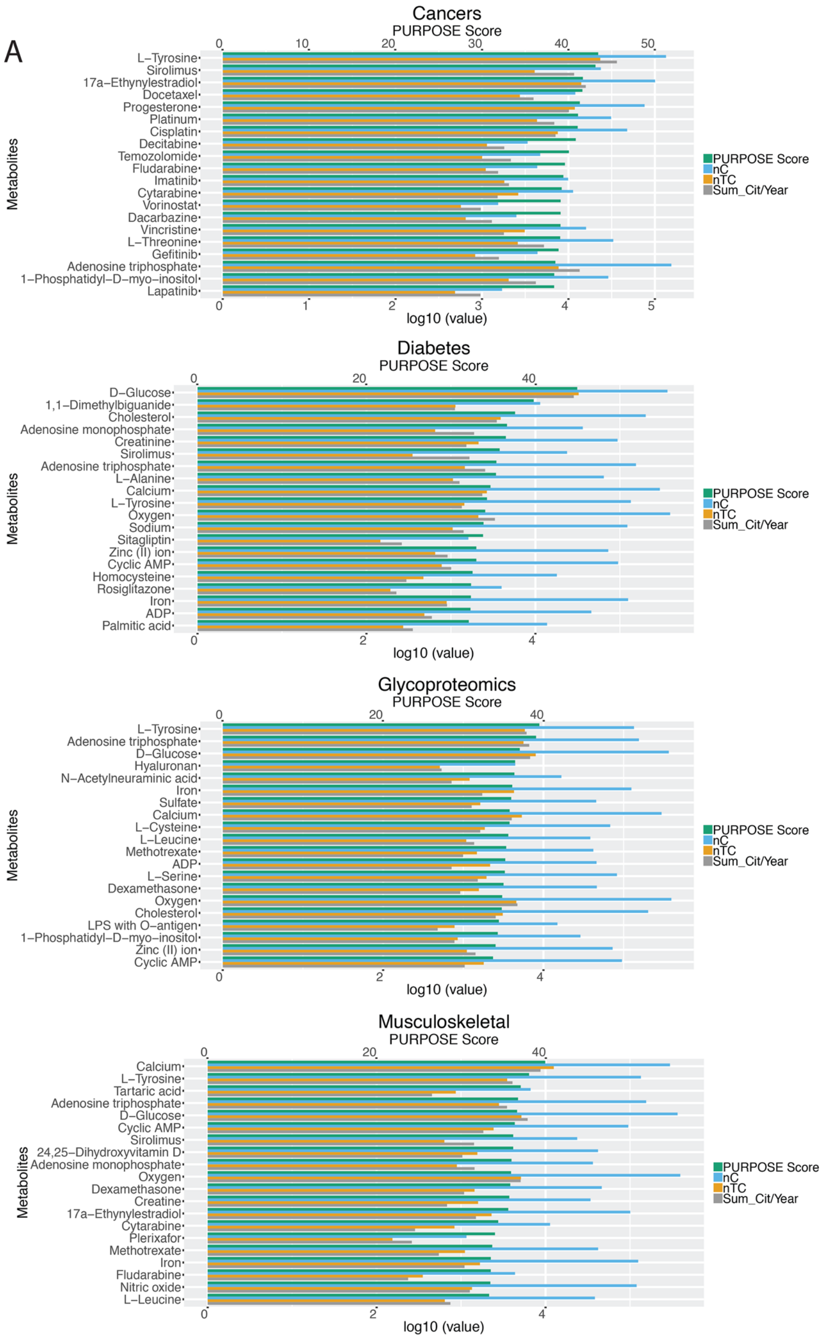


Figure 2. continued

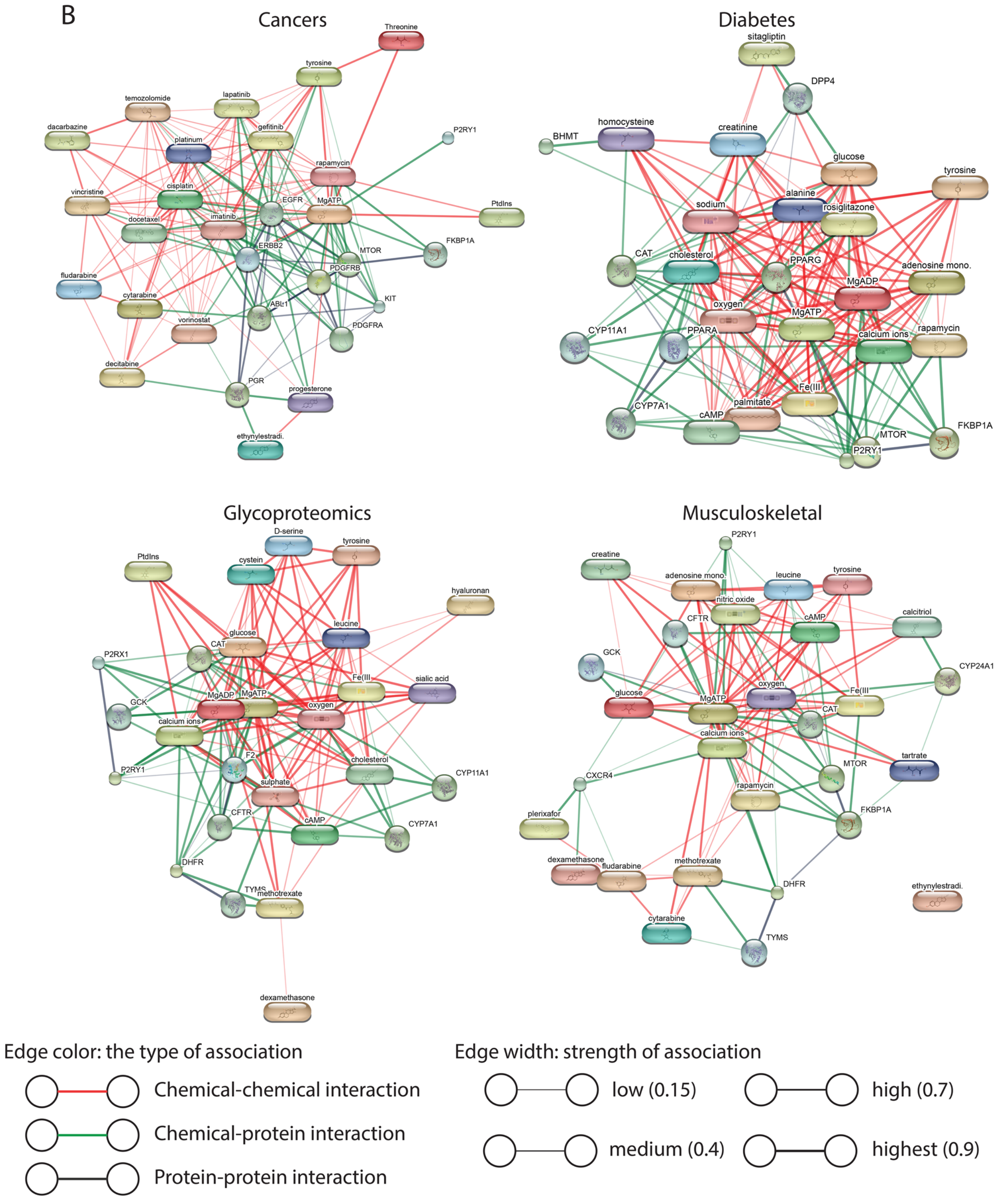


Figure 2. continued

DOI: 10.1021/acs.jproteome.8b00378
J. Proteome Res. 2018, 17, 4345−4357

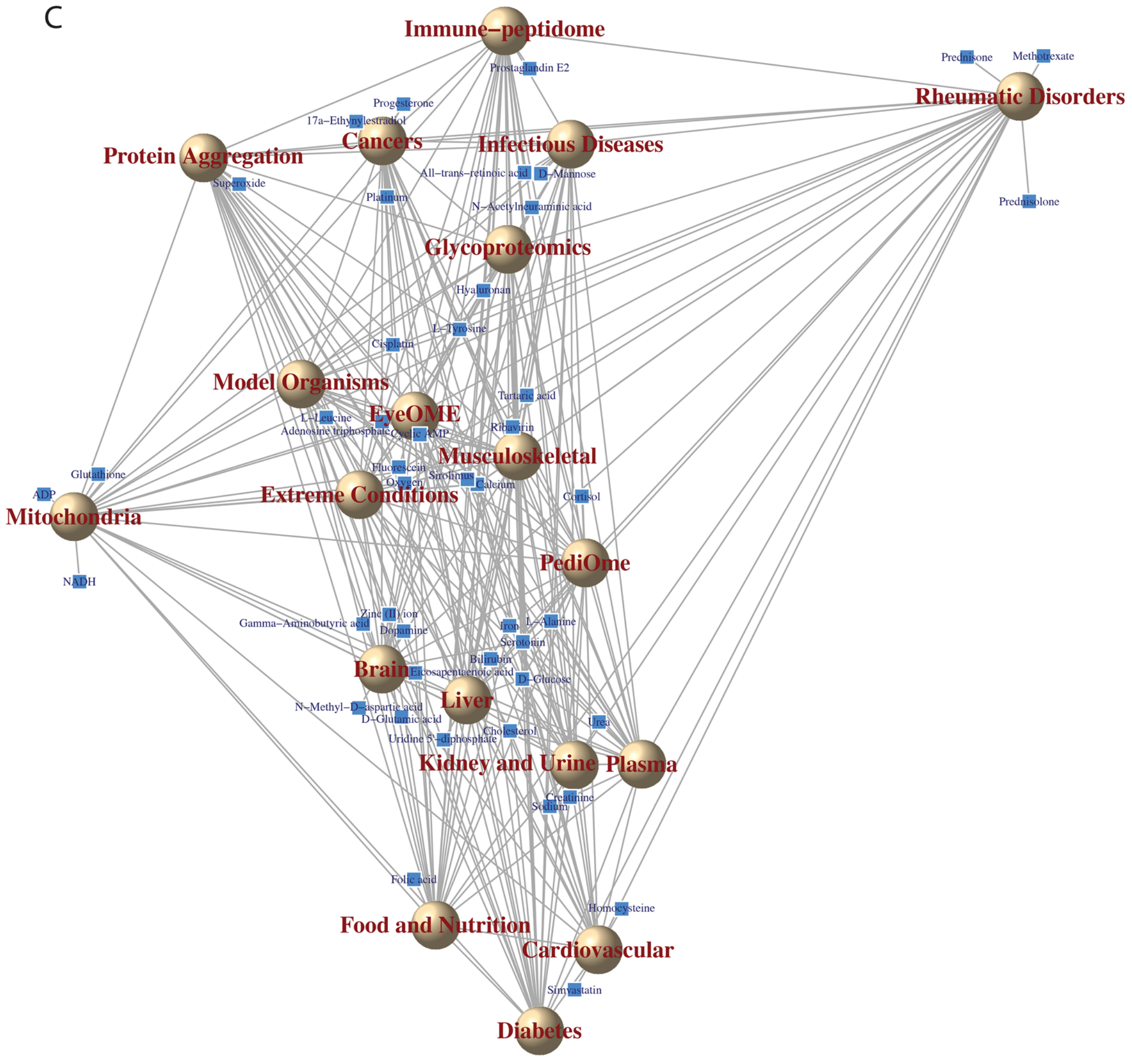


Figure 2. Metabolite prioritization in the selected B/D-HPP targeted areas. (A) Distributions of the Protein Universal Reference Publication-Originated Search Engine (PURPOSE) scores of the top metabolites associated with cancers, diabetes, glycoproteomics, and the musculoskeletal system. In each graph, the top *X* axis is the PURPOSE score and the bottom *X* axis is $\log_{10}$(value), where "value" is either nC (the number of publications associated with the metabolite), nTC (the number of papers associated with both the topic and the metabolite (TC)), or Sum_Cit/Year (citations per year of TC). (B) Network analysis results using the Search Tool for Interactions of Chemicals (STITCH) tool. Metabolites with the highest PURPOSE scores and their interacting proteins are shown. (C) Multidimensional scaling (MDS) visualization of the connections among B/D-HPP targeted fields and their associated metabolites. B/D-HPP fields with higher correlation in their associated metabolites' PURPOSE scores have shorter distances on the graph.

Cloud-Based User Interface

To facilitate user interaction, a user interface was built with the shiny package in R. The system has been deployed to a cloud server, allowing researchers to access the system with ease. All statistical analysis was conducted using R version 3.3. The source codes for the cloud-based system, the literature mining back end, and the automated updater for PubTator data files are available at http://rebrand.ly/metapurposesourcecode.

■ RESULTS

Summary of Metabolites and Chemicals Published in the PubMed Literature

At the time of evaluation, there were 27 million PubMed articles. PubTator tagged 79 948 chemicals in 9.04 million PubMed articles; 7508 chemicals (9.39%) are labeled as human metabolites by the HMDB and are mentioned in 7.29 million articles in PubMed. The publication trends of all PubMed articles on human and articles associated with at least one chemical or metabolite since 1950 are shown in Figure 1A. The numbers of publications per year on human, chemicals related to human, and human metabolites have increased steadily since 1950. The annualized number of publications on human is strongly correlated with the annualized number of papers describing human metabolites (Spearman's correlation coefficient = 0.998) and the annualized number of publications mentioning chemicals related to human (Spearman's correlation coefficient = 0.996).

Publication Patterns of Metabolites and Chemicals Related to the B/D-HPP Targeted Areas

To prioritize the metabolites and chemicals associated with the B/D-HPP targeted areas through literature mining, we implemented the PURPOSE algorithm to summarize the topic−metabolite/chemical copublication strengths in the PubMed literature. For each targeted area of the B/D-HPP,

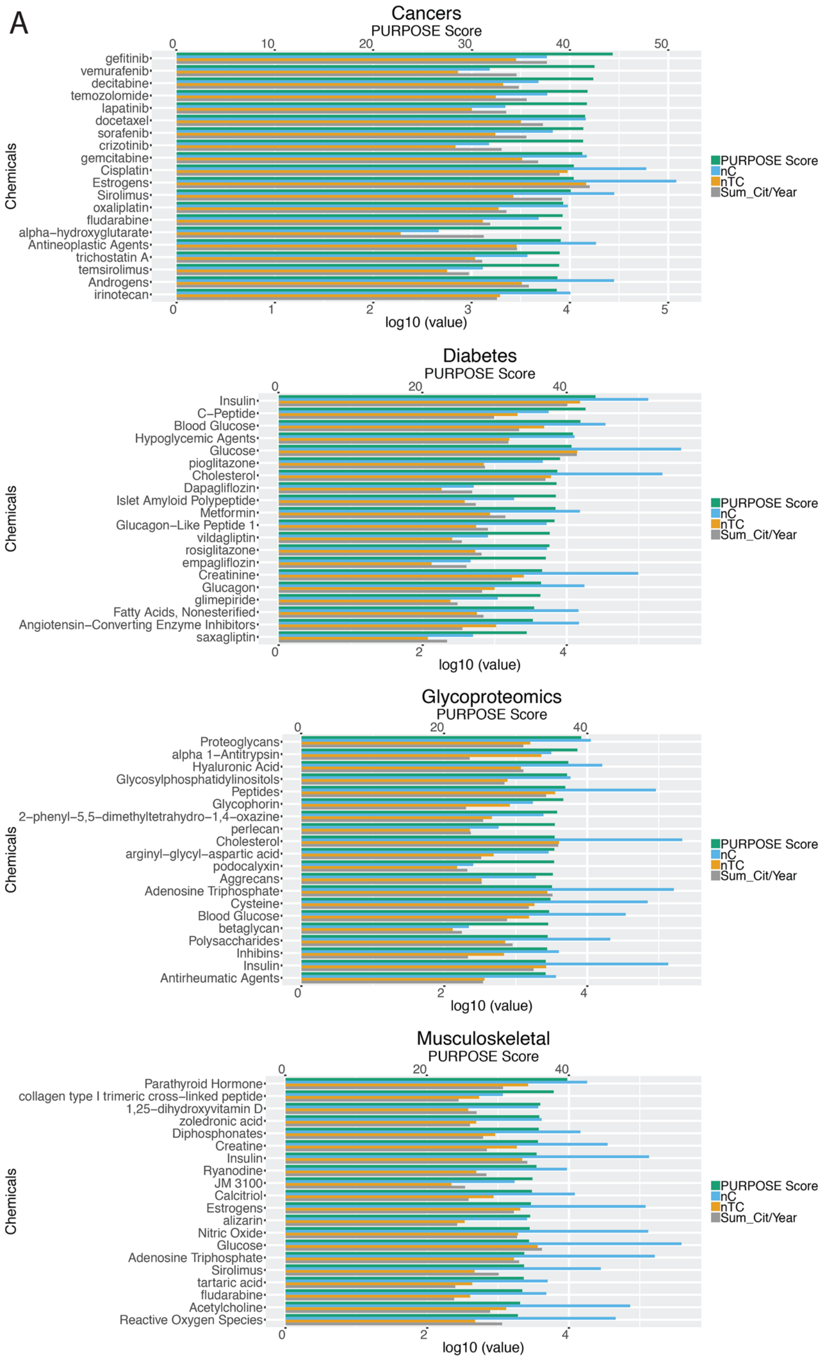


Figure 3. continued

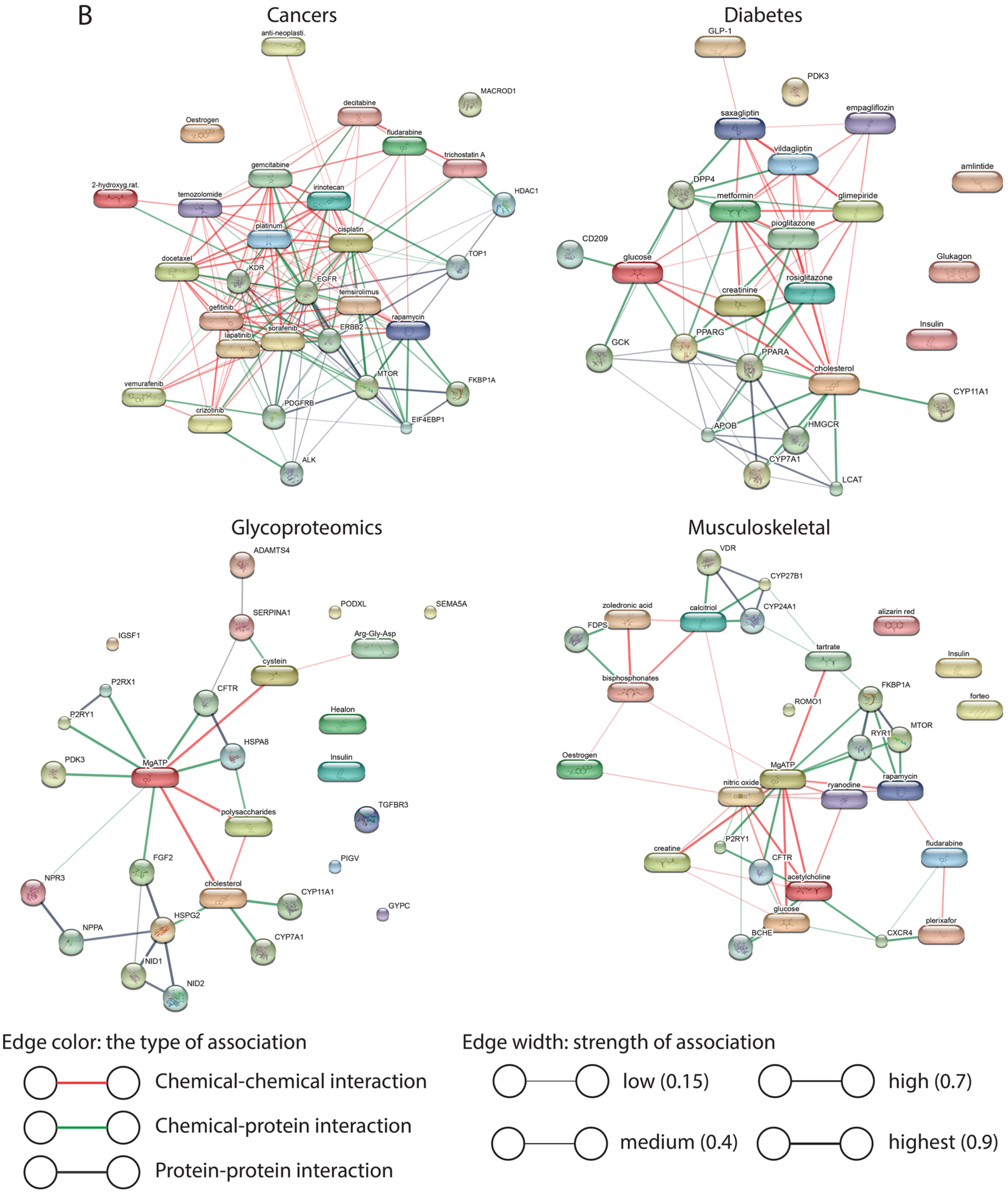


Figure 3. continued

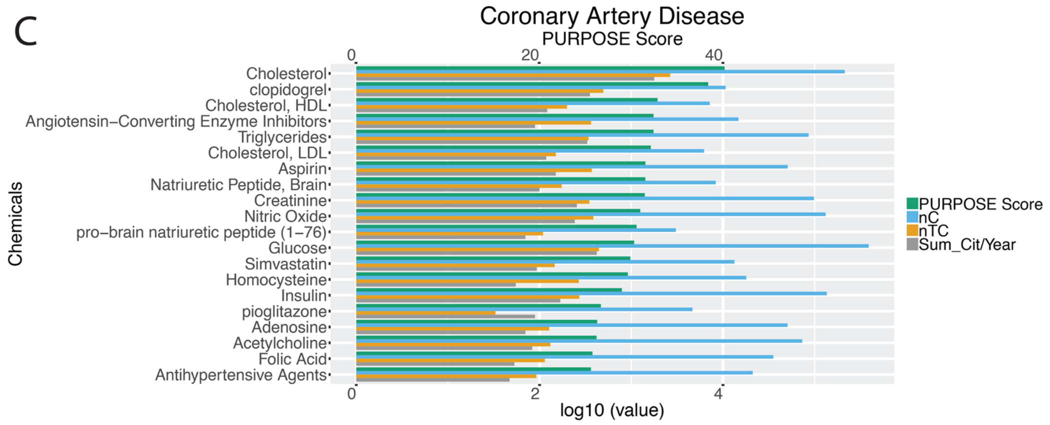


Figure 3. Chemical prioritization in the selected B/D-HPP targeted areas. (A) Distributions of the PURPOSE scores of the top chemicals associated with cancers, diabetes, glycoproteomics, and the musculoskeletal system. In each graph, the top X axis is the PURPOSE score and the bottom X axis is $\log_{10}$(value), where "value" is either nC, nTC, or Sum_Cit/Year. (B) Network analysis results using the STITCH tool. Chemicals with the highest PURPOSE scores and their interacting proteins are shown. (C) PURPOSE scores of the top chemicals associated with coronary artery disease.

the numbers of associated metabolites/chemicals, publications, total citations, and citations per year are summarized in Figure 1B,C. The total number of metabolites associated with each B/D-HPP area is between 405 (rheumatic) and 2483 (plasma), whereas that of chemicals is between 705 (rheumatic) and 14 070 (model organisms). Across the B/D-HPP topics, the Spearman's correlation coefficient between the number of identified chemicals and that of metabolites is 0.98. The targeted areas with the greatest number of metabolite-related publication are cancers (310 537 publications), plasma (281 874), model organisms (172 860), PediOme (153 053), and glycoproteomics (137 234). The areas with the most chemical-related publications are also cancers (378 929 publications), plasma (321 201), model organisms (205 747), PediOme (181 353), and glycoproteomics (172 064). For the 19 B/D-HPP topics, the Spearman's correlation coefficient between the number of publications associated with metabolites and the number of publications associated with chemicals is 0.998, and the correlation coefficient between the annualized citation numbers associated with metabolites and that of chemicals is 0.875. All of the B/D-HPP topics have at least 2150 publications associated with metabolites or chemicals, indicating the rich information in the published literature.

Prioritizing Metabolites in the B/D-HPP Targeted Fields

To prioritize metabolites related to the B/D-HPP targeted areas, a list of human metabolites were identified from the HMDB,[8] where a number of drugs, drug metabolites, and chemical compounds were annotated as metabolites. The metabolites associated with each B/D-HPP area were ranked by their PURPOSE scores, which balanced the strength (quantified by the number of copublications in PubMed and the citation numbers of the papers per year) and the specificity (quantified by the number of publications associated with the topics and that of the proteins in general) of the associations.[12] As an illustration, L-tyrosine (PURPOSE score = 43.44), sirolimus (43.13), 17a-ethynylestradiol (41.67), docetaxel (41.62), and progesterone (41.32) were the metabolites with the highest PURPOSE scores in cancers (Figure 2A). These metabolites were enriched in the epidermal growth factor receptor signaling, protein autophosphorylation, and Fc receptor signaling pathways (Figure 2B). Metscape revealed that these metabolites participated in the metabolism of phosphatidylinositol phosphate and purine (Figure S-1). For diabetes, the metabolites D-glucose (44.95), 1,1-dimethylbiguanide (39.80), cholesterol (37.59), adenosine monophosphate (36.64), and creatinine (36.43) had the highest scores (Figure 2A). These metabolites and chemicals were enriched in the PPAR signaling pathway and a number of biological processes, including regulation of the cellular ketone metabolic process (Figure 2B). Metscape showed that the prioritized metabolites were involved in glycolysis, gluconeogenesis, cholesterol biosynthesis, and de novo fatty acid biosynthesis pathways (Figure S-1). The metabolites L-tyrosine (39.49), adenosine triphosphate (39.08), D-glucose (37.05), hyaluronan (36.47), and N-acetylneuraminic acid (36.37) attained the highest scores in glycoproteomics (Figure 2A). These metabolites participated in the aminosugar metabolism, fructose and mannose metabolism, and glycerophospholipid metabolism pathways (Figures 2B and S-1). Calcium (39.97), L-tyrosine (38.05), tartaric acid (37.06), adenosine triphosphate (36.74), and D-glucose (36.63) were the metabolites most relevant to the musculoskeletal system (Figure 2A). Pathway analysis revealed that these metabolites were associated with the metabolism pathways of carbohydrates (including fructose, mannose, and galactose) and amino acids (e.g., tyrosine, arginine, proline, glutamate, aspartate, and asparagine) (Figures 2B and S-1). The results indicated that our methods successfully retrieved many known associations between metabolites and the B/D-HPP areas. Network analysis across the four areas revealed that these gene–metabolite interaction networks (Figure 2B) were highly connected (connectedness scores (proportions of connected node pairs in the network) > 0.93) and moderately centralized (centralization scores of 0.40–0.51). In addition, there were multiple interactions connecting the nodes, resulting in moderate connectivity efficiency (efficiency scores (proportions of edges that could not be removed without disconnecting the nodes) of 0.54–0.72) and transitivity (transitivity scores (probabilities that nodes A and C are directly connected in the network when node A is connected to node B and node B is connected to node C) of 0.53–0.65) (Figure S-2A). We further computed the scores of all related metabolites for each of the B/D-HPP targeted areas, and the results are summarized in Data S-1. Figure 2C shows the

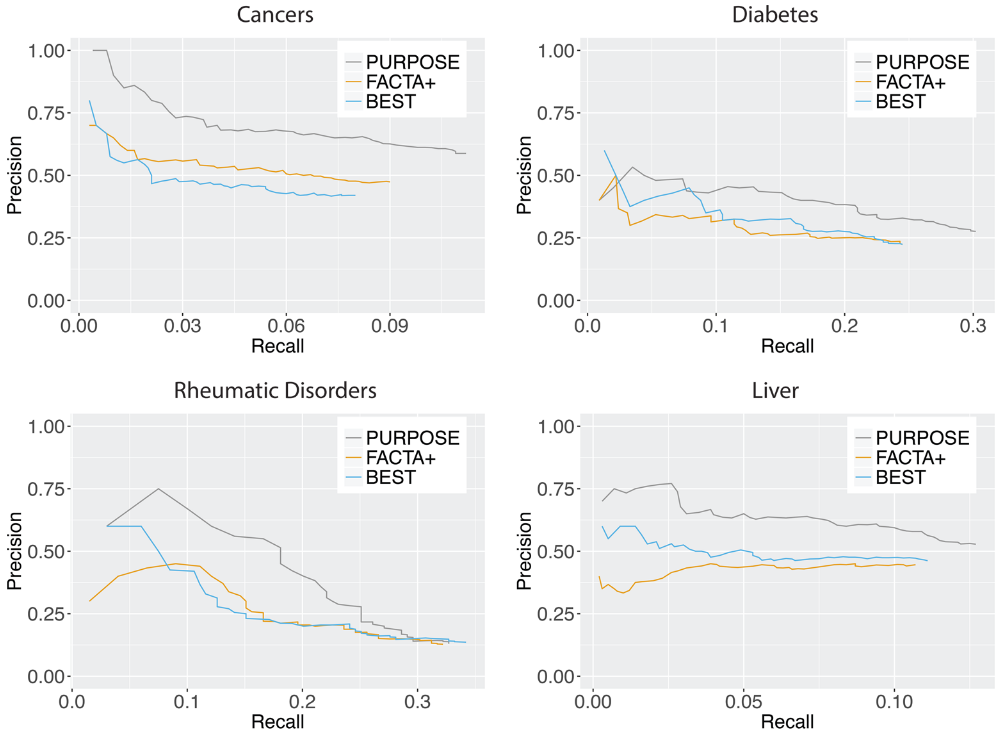


Figure 4. Performance comparison among Protein Universal Reference Publication-Originated Search Engine (PURPOSE), Finding Associated Concepts with Text Analysis (FACTA+), and Biomedical Entity Search Tool (BEST). Precision−recall curves for chemical prioritization for cancers, diabetes, rheumatic diseases, and liver are shown. Biologist-curated topic−chemical relations from the Comparative Toxicogenomics Database (CTD) were used as the ground truth. PURPOSE achieved the best precision and recall in cancers, diabetes, and liver and exhibited similar performance in rheumatic diseases compared with FACTA+ and BEST. BEST performed better than FACTA+ in liver but worse in cancers, and the two systems had similar performance in diabetes and rheumatic diseases.

correlations among the B/D-HPP targeted areas and highlights metabolites strongly associated with each B/D-HPP. Biologically related concepts, such as cardiovascular, diabetes, and food and nutrition, formed a cluster in the figure.

Prioritizing Chemicals in the B/D-HPP Targeted Fields

In addition to metabolites, our algorithm successfully prioritized chemicals associated with the B/D-HPP targeted fields (Data S-2). The identified chemicals ranged from endogenous chemicals (including hormones and neurotransmitters) to drugs, drug metabolites, ions, and environmental pollutants, as defined by the PubTator tool.[18] On the basis of the DrugBank[35] definition, 1630 chemicals tagged by PubTator are drugs. Drugs tended to have more PubMed publications (median number of publications = 1195.5) than nondrug chemicals (median number of publications = 4).

Using the PURPOSE scores, we identified chemicals implicated with each of the B/D-HPP focused areas. For instance, gefitinib (PURPOSE score = 44.35), vemurafenib (42.50), decitabine (42.38), temozolomide (41.79), and lapatinib (41.73) scored the highest among all chemicals in cancers (Figure 3A). These chemicals were enriched in the protein autophosphorylation and transmembrane receptor protein tyrosine kinase signaling pathways (Figure 3B). For diabetes, insulin (44.00), C-peptide (42.61), and blood glucose (41.90) had the highest scores (Figure 3A). These chemicals are involved in the PPAR signaling pathway and carbohydrate metabolism mechanisms (Figure 3B). The chemicals proteoglycans (39.15), alpha-1-antitrypsin (38.62), hyaluronic acid (37.35), and glycosylphosphatidylinositols (37.17) scored the highest in glycoproteomics (Figure 3A). The chemicals were highly enriched in the cholesterol metabolic process, carbohydrate derivative binding, and ATP binding functions (Figure 3B). For the musculoskeletal system, parathyroid hormone (39.77), collagen type I trimeric cross-linked peptide (37.85), 1,25-dihydroxyvitamin D (35.96), zoledronic acid (35.82), and diphosphonates (35.74) were the highest-scoring chemicals (Figure 3A). Pathway analysis revealed that these chemicals were associated with positive regulation of vitamin D 24-hydroxylase activity and the vitamin D catabolic process (Figure 3B). Quantitative analyses on the gene−chemical interaction networks (Figure 3B) showed that these networks are less well-connected than the gene−metabolite interaction networks (Figure 2B) of the same query topic (connectedness scores of 0.54−0.87) with low to moderate centralization scores (0.33−0.41). In these B/D-HPP targeted fields, many drugs had high PURPOSE scores, which is consistent with the fact that there were more publications associated with drugs than nondrugs in general. Compared with the gene−metabolite interaction networks, the gene−chemical interaction networks had relatively sparse edge connections, resulting in higher connectivity efficiency scores in general (efficiency scores of 0.58−0.90) and variable transitivity scores (0.20−0.66) (Figure S-2B). Figure S-3 visualizes the connections among the B/D-HPP targeted areas and illustrates chemicals strongly associated with each B/D-HPP.

Our algorithm can also identify the chemicals associated with specific biological or medical conditions. As an illustration, in response to the query “coronary artery disease” in human, our method retrieved many well-known chemicals associated with the disease (Figure 3C), such as cholesterol (PURPOSE score = 40.18), HDL (32.87), triglycerides

(32.42), LDL (32.13), brain natriuretic peptide (31.54), and homocysteine (29.62). In addition, many drugs related to the treatment of coronary artery disease and related comorbidities were identified by our system. For instance, clopidogrel (38.39) and aspirin (31.55) ranked among the top 10 chemicals in this query. These results suggested the extensibility of the PURPOSE algorithm to specific biomedical conditions of clinical importance.

Evaluation of the Prioritization Results

In comparison with the curated topic−chemical relations obtained from the CTD,[31] our tool successfully retrieved relevant chemicals from the literature. The precision and recall of our tool were better than those of the FACTA+[32] and BEST[34] systems in most B/D-HPP fields with CTD annotations (Figure 4). Among the top 500 retrieved chemicals associated with cancers, diabetes, or liver, our tool achieved a 5.2−11.4% improvement in precision and a 2.0−5.7% improvement in recall compared with FACTA+ and a 5.2−16.8% improvement in precision and a 1.6−3.2% improvement in recall compared with BEST. FACTA+ performed better than BEST in cancers but had worse performance in liver, and the two systems had similar performance in diabetes. For rheumatic diseases, which had the least number of PubMed publications, the first 390 chemicals retrieved by PURPOSE attained the highest precision and recall among all three tools, but the precision gradually decreased as we went further down the retrieved list to include chemicals with lower PURPOSE scores, indicating that the PURPOSE algorithm worked better in well-published fields and for well-studied chemicals. These results validated the relevance of the PURPOSE algorithm in chemical prioritization tasks.

Cloud-Based System Deployment

To facilitate real-time metabolite and chemical prioritization, a cloud-based system was deployed. In addition to the B/D-HPP targeted areas, our system allows users to input any search term of interest and retrieves the results in a few seconds. Modules for enrichment analyses, visualization of PURPOSE score distributions, and summarization of highly cited publications are available in the browser-based user interface. Our system is freely accessible at http://rebrand.ly/metapurpose.

■ DISCUSSION

We have presented a novel general-purpose tool for metabolite and chemical prioritization with direct applications to the ongoing B/D-HPP investigations.[1−3] Our cloud-based system automatically obtains the most updated PubMed literature and bioentity taggings and employs the state-of-the-art literature mining approach to prioritizing metabolites and chemicals, and the results were successfully validated in the curated Comparative Toxicogenomics Database.[31] Our approach will facilitate targeted metabolomics and chemical analyses, which is expected to expedite multiomics integration for investigations of human biology and disease states.[5,36,37]

As many metabolites and chemicals possess a number of evolving synonyms,[13] it was difficult to track their publication trends, and there was no available tool that prioritizes metabolites for targeted investigations. To address this challenge, our system employs the tagged entities from PubTator,[18] identifies tags for chemicals using the MeSH ontological structure,[17] and filters known human metabolites using the curated information from the HMDB.[8] In addition, we demonstrated the extensibility of the PURPOSE algorithm,[12] which achieved improved precision and recall compared with the previously proposed literature mining methods.[32,34] Our system allows users to input any search term of interest, queries the most updated PubMed database, retrieves and prioritizes the metabolites and chemicals in real time, and summarizes the results for the users. Our cloud-based system enables enrichment analyses of the retrieved results,[14] provides external links to curated databases,[8] and shows the landmark publications describing the relations between the queried topic and the prioritized metabolites and chemicals.

Our results demonstrate that there are a great number of publications describing metabolites and chemicals associated with each of the B/D-HPP targeted fields, indicating the feasibility of building literature mining systems for prioritizing metabolites and chemical targets. The numbers of publications on human metabolites and chemicals have increased steadily since 1950. In recent years, more than 70 000 new publications on human metabolites and chemicals (including more than 50 000 papers mentioning drugs) have been added to the literature each year. The amount of information posed a challenge to manual literature curation but a unique opportunity for text mining algorithms in retrieving and aggregating the most updated and relevant information from the literature.[38] Our system showcases a novel way of utilizing such information, and the prioritized metabolites and chemicals can guide targeted analysis as well as serve as dynamic summaries of the publication trends in the queried fields.

One limitation of our approach is that some newly synthesized chemicals may not have a MeSH term or identifier. Such new chemicals could be missed by PubTator tagging and hence not prioritized by our system. To address this challenge, we have implemented an automated updater to obtain the most recent MeSH entries and PubTator taggings regularly. In addition, like all literature mining tools, undiscovered topic−chemical associations would not receive high PURPOSE scores. The ongoing efforts on high-throughput metabolomics and chemical profiling could mitigate this issue.[39]

In summary, our system successfully identifies relevant metabolites and chemicals associated with each of the B/D-HPP focused fields. Together with the previously described protein prioritization framework,[12] our tools can compile lists of proteins, metabolites, and chemicals related to the B/D-HPP targeted areas and other human organ systems or disease states, which will facilitate the design of targeted proteomic, metabolomic, and biochemical profiling methods and expedite integrative multiomic analyses. The cloud-based metabolites and chemicals prioritization platform can accommodate any custom search term, enabling scientific investigations of any diseases or organs of interest, and contribute to the development of precision medicine.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00378.

Metscape visualized the pathways involved with the prioritized metabolites (Figure S-1); quantitative network statistics of the gene-metabolite interaction and gene-chemical interaction networks (Figure S-2); MDS visualization of the connections among B/D-HPP targeted fields and their associated chemicals (Figure S-3); PubMed search terms for the B/D-HPP targeted fields (Table S-1) (PDF)

Data S-1: metabolite prioritization results for B/D-HPP (XLSX)

Data S-2: chemical prioritization results for B/D-HPP (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*Tel: (617) 432-2144. E-mail: Isaac_Kohane@hms.harvard.edu.
*Tel: (650) 736-8099. E-mail: mpsnyder@stanford.edu.
*Tel: +886 6-2757575 ext 62534. E-mail: jchiang@mail.ncku.edu.tw.

ORCID

Kun-Hsing Yu: 0000-0001-9892-8218
Yu-Ju Chen: 0000-0002-3178-6697

Author Contributions

○M.S. and I.S.K. contributed equally to this study. K.-H.Y. conceived, designed, and performed the analysis, interpreted the results, developed the cloud-based query system and user interface, and drafted the manuscript. T.-L.M.L. implemented the backend literature mining system, interpreted the metabolite and chemical prioritization results, evaluated the system performance, and revised the manuscript. Y.-J.C., C.R., S.C.K., J.-H.C., M.S., and I.S.K interpreted the results and revised the manuscript. I.S. K., M.S., and J.-H.C. supervised the work. All of the authors approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

K.-H.Y. is a Harvard Data Science Fellow. This work was supported in part by grants from the National Human Genome Research Institute, National Institutes of Health (Grant 5P50HG007735), the National Cancer Institute, National Institutes of Health (Grant 5U24CA160036), the Office of the Director, National Institutes of Health (Grant 3OT3OD025466), the Defense Advanced Research Projects Agency (Simplifying Complexity in Scientific Discovery (SIMPLEX) Grant N66001-15-C-4043 and Data-Driven Discovery of Models Contract FA8750-17-2-0095), and the Ministry of Science and Technology of Taiwan (Research Grant MOST 103-2221-E-006-254-MY2). The authors express their appreciation to Professor Griffin Weber for his insight on citation counts, Dr. Stephen Bach, Mr. Alex Ratner, and Mr. Chen-Ruei Liu for their valuable advice on literature mining and suggestions on the manuscript, Dr. Mu-Hung Tsai for pointing out the literature mining resources, and Ms. Samantha Lemos for her administrative support. The authors thank the AWS Cloud Credits for Research, Microsoft Azure Research Award, and the NVIDIA Corporation for their support on the computational infrastructure. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges Pylon at the Pittsburgh Supercomputing Center, which is supported by National Science Foundation (Grant ACI-1548562), through allocation TG-BCS180016.

ABBREVIATIONS

HUPO, Human Proteome Organization; B/D-HPP, Biology/Disease-Driven Human Proteome Project; MeSH, Medical Subject Headings; STITCH, Search Tool for Interactions of Chemicals; PURPOSE, Protein Universal Reference Publication-Originated Search Engine; FACTA+, Finding Associated Concepts with Text Analysis; BEST, Biomedical Entity Search Tool

REFERENCES

(1) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12* (1), 23−7.

(2) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J.; Kussman, M.; Qin, J.; Omenn, G. S. Highlights of B/D-HPP and HPP Resource Pillar Workshops at 12th Annual HUPO World Congress of Proteomics: September 14−18, 2013, Yokohama, Japan. *Proteomics* **2014**, *14* (9), 975−88.

(3) Van Eyk, J. E.; Corrales, F. J.; Aebersold, R.; Cerciello, F.; Deutsch, E. W.; Roncada, P.; Sanchez, J. C.; Yamamoto, T.; Yang, P.; Zhang, H.; Omenn, G. S. Highlights of the Biology and Disease-driven Human Proteome Project, 2015−2016. *J. Proteome Res.* **2016**, *15* (11), 3979−3987.

(4) Wei, R.; Li, G.; Seymour, A. B. High-throughput and multiplexed LC/MS/MRM method for targeted metabolomics. *Anal. Chem.* **2010**, *82* (13), 5527−33.

(5) Chen, R.; Mias, G. I.; Li-Pook-Than, J.; Jiang, L.; Lam, H. Y.; Chen, R.; Miriami, E.; Karczewski, K. J.; Hariharan, M.; Dewey, F. E.; Cheng, Y.; Clark, M. J.; Im, H.; Habegger, L.; Balasubramanian, S.; O'Huallachain, M.; Dudley, J. T.; Hillenmeyer, S.; Haraksingh, R.; Sharon, D.; Euskirchen, G.; Lacroute, P.; Bettinger, K.; Boyle, A. P.; Kasowski, M.; Grubert, F.; Seki, S.; Garcia, M.; Whirl-Carrillo, M.; Gallardo, M.; Blasco, M. A.; Greenberg, P. L.; Snyder, P.; Klein, T. E.; Altman, R. B.; Butte, A. J.; Ashley, E. A.; Gerstein, M.; Nadeau, K. C.; Tang, H.; Snyder, M. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **2012**, *148* (6), 1293−307.

(6) Yu, K. H.; Snyder, M. Omics Profiling in Precision Oncology. *Mol. Cell. Proteomics* **2016**, *15* (8), 2525−36.

(7) Borràs, E.; Sabidó, E. What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry. *Proteomics* **2017**, *17* (17−18), 1700180.

(8) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vazquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D608−D617.

(9) Kusebauch, U.; Campbell, D. S.; Deutsch, E. W.; Chu, C. S.; Spicer, D. A.; Brusniak, M. Y.; Slagel, J.; Sun, Z.; Stevens, J.; Grimes, B.; Shteynberg, D.; Hoopmann, M. R.; Blattmann, P.; Ratushny, A. V.; Rinner, O.; Picotti, P.; Carapito, C.; Huang, C. Y.; Kapousouz, M.; Lam, H.; Tran, T.; Demir, E.; Aitchison, J. D.; Sander, C.; Hood, L.; Aebersold, R.; Moritz, R. L. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* **2016**, *166* (3), 766−778.

(10) Lam, M. P.; Venkatraman, V.; Cao, Q.; Wang, D.; Dincer, T. U.; Lau, E.; Su, A. I.; Xing, Y.; Ge, J.; Ping, P.; Van Eyk, J. E. Prioritizing Proteomics Assay Development for Clinical Translation. *J. Am. Coll. Cardiol.* **2015**, *66* (2), 202−4.

(11) Lam, M. P.; Venkatraman, V.; Xing, Y.; Lau, E.; Cao, Q.; Ng, D. C.; Su, A. I.; Ge, J.; Van Eyk, J. E.; Ping, P. Data-Driven Approach To

Determine Popular Proteins for Targeted Proteomics Translation of Six Organ Systems. *J. Proteome Res.* **2016**, *15* (11), 4126−4134.

(12) Yu, K. H.; Lee, T. M.; Wang, C. S.; Chen, Y. J.; Re, C.; Kou, S. C.; Chiang, J. H.; Kohane, I. S.; Snyder, M. Systematic Protein Prioritization for Targeted Proteomics Studies through Literature Mining. *J. Proteome Res.* **2018**, *17* (4), 1383−1396.

(13) Mattingly, C. J.; Rosenstein, M. C.; Davis, A. P.; Colby, G. T.; Forrest, J. N., Jr.; Boyer, J. L. The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.* **2006**, *92* (2), 587−95.

(14) Kuhn, M.; von Mering, C.; Campillos, M.; Jensen, L. J.; Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* **2008**, *36* (Databaseissue), D684−8.

(15) Shatkay, H.; Feldman, R. Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.* **2003**, *10* (6), 821−55.

(16) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **2016**, *44* (D1), D1214−9.

(17) Lipscomb, C. E. Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc.* **2000**, *88* (3), 265−6.

(18) Wei, C. H.; Kao, H. Y.; Lu, Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **2013**, *41* (W1), W518−22.

(19) The Human Proteome Organization Biology/Disease-driven HPP. https://www.hupo.org/B/D-HPP (accessed August 1, 2018).

(20) Sayers, E. Entrez programming utilities help. http://www.ncbi.nlm.nih.gov/books/NBK25499 (accessed August 1, 2018).

(21) Leskovec, J.; Rajaraman, A.; Ullman, J. D. *Mining of Massive Datasets*; Cambridge University Press: Cambridge, U.K., 2014; pp 1−15.

(22) Szklarczyk, D.; Santos, A.; von Mering, C.; Jensen, L. J.; Bork, P.; Kuhn, M. STITCH 5: augmenting protein−chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **2016**, *44* (D1), D380−4.

(23) Butts, C. T. *R Package sna: Tools for Social Network Analysis*, version 2.2-0, 2010.

(24) Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks* **1978**, *1* (3), 215−239.

(25) Basu, S.; Duren, W.; Evans, C. R.; Burant, C. F.; Michailidis, G.; Karnovsky, A. Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. *Bioinformatics* **2017**, *33* (10), 1545−53.

(26) Karnovsky, A.; Weymouth, T.; Hull, T.; Tarcea, V. G.; Scardoni, G.; Laudanna, C.; Sartor, M. A.; Stringer, K. A.; Jagadish, H. V.; Burant, C.; Athey, B.; Omenn, G. S. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **2012**, *28* (3), 373−80.

(27) Gao, J.; Tarcea, V. G.; Karnovsky, A.; Mirel, B. R.; Weymouth, T. E.; Beecher, C. W.; Cavalcoli, J. D.; Athey, B. D.; Omenn, G. S.; Burant, C. F.; Jagadish, H. V. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics* **2010**, *26* (7), 971−3.

(28) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13* (11), 2498−504.

(29) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*; Chapman and Hall/CRC: Boca Raton, FL, 2000; pp 5−8.

(30) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkens-Diehr, N. XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* **2014**, *16* (5), 62−74.

(31) Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; King, B. L.; McMorran, R.; Wiegers, J.; Wiegers, T. C.; Mattingly, C. J. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D972−D978.

(32) Tsuruoka, Y.; Tsujii, J.; Ananiadou, S. FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* **2008**, *24* (21), 2559−60.

(33) Tsuruoka, Y.; Miwa, M.; Hamamoto, K.; Tsujii, J.; Ananiadou, S. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* **2011**, *27* (13), i111−9.

(34) Lee, S.; Kim, D.; Lee, K.; Choi, J.; Kim, S.; Jeon, M.; Lim, S.; Choi, D.; Kim, S.; Tan, A. C.; Kang, J. BEST: Next-Generation Biomedical Entity Search Tool for Knowledge Discovery from Biomedical Literature. *PLoS One* **2016**, *11* (10), e0164680.

(35) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074.

(36) Yu, K. H.; Berry, G. J.; Rubin, D. L.; Re, C.; Altman, R. B.; Snyder, M. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. *Cell Syst* **2017**, *5* (6), 620−627.

(37) Yu, K. H.; Fitzpatrick, M. R.; Pappas, L.; Chan, W.; Kung, J.; Snyder, M. Omics AnalySIs System for PRecision Oncology (OASISPRO): A Web-based Omics Analysis Tool for Clinical Phenotype Prediction. *Bioinformatics* **2018**, *34* (2), 319−320.

(38) Zhu, F.; Patumcharoenpol, P.; Zhang, C.; Yang, Y.; Chan, J.; Meechai, A.; Vongsangnak, W.; Shen, B. Biomedical text mining and its applications in cancer research. *J. Biomed. Inf.* **2013**, *46* (2), 200−11.

(39) Wishart, D. S. Proteomics and the human metabolome project. *Expert Rev. Proteomics* **2007**, *4* (3), 333−5.