Nutritional Systems Biology

Jensen, Kasper

Publication date:
2014

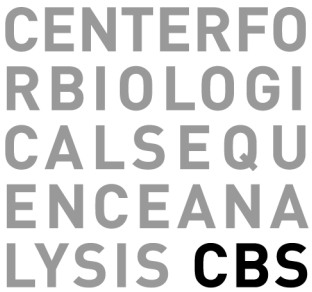
Document Version
Publisher's PDF, also known as Version of record

NUTRITIONAL SYSTEMS BIOLOGY

Kasper Jensen

Supervisors:
Irene Kouskoumvekaki and
Gianni Panagiotou

Funded by DTU's PhD scholarships



Center for Biological Sequence Analysis
Department of Systems Biology
Technical University of Denmark

09/06/14

Preface

The thesis was prepared at the Center for Biological Sequence Analysis, Department of Systems Biology at the Technical University of Denmark. The work was funded by DTU's PhD scholarship program.

Contents

Summary

"Prevention is better than cure" and when it comes to human health, this strategy translates into many socioeconomic benefits. Practically all the cellular processes, including every step in the flow of genetic information from gene expression to protein synthesis and degradation, can be affected by diet and lifestyle. Similar to the role of pharmaceuticals, nutrients contain a number of different compounds that act as modifiers of network function and stability. However, the level of complexity in nutrition studies is further increased by the simultaneous presence of a variety of nutrients, with diverse chemical structures that can have numerous targets with different affinities and specificities. Obviously, this differentiates the nutritional from the pharmacological studies, where single elements are used at low concentrations and with a relatively high affinity and specificity in a small number of thoroughly selected targets. Our need for fundamental understanding of the building blocks of the complex biological systems had been the main reason for the reductionist approach that was mainly applied in the past to elucidate these systems. Nowadays, it is widely recognized that systems and network biology has the potential to increase our understanding of how small molecules affect metabolic pathways and homeostasis, how this perturbation changes at the disease state, and to what extent individual genotypes contribute to this. A fruitful strategy in approaching and exploring the field of nutritional research is, therefore, to borrow methods that are well established in medical and pharmacological research.

In this thesis, we use advanced data-mining tools for the construction of a database with available, state-of-the-art information concerning the interaction of food and its molecular components with biological systems and their connection to health and disease. The database will be enriched with predicted interactions between food components and protein targets, based on their structural and pharmacophore similarity with known small molecule ligands. Further to this, the associations of bioactive food components with metabolic pathways will be investigated from a chemical-protein network perspective, while their effects in network robustness will be further confirmed by proteome analyses and high-throughput genotype-phenotype characterization.

The first chapter of the thesis is about the development of our data resource. In this work, we applied text mining and Naïve Bayes classification to assemble the knowledge space of food-phytochemical and food-disease associations, where we distinguish between disease prevention/amelioration and disease progression. We subsequently searched for frequently occurring phytochemical-disease pairs and we identified 20,654 phytochemicals from 16,102 plants associated to 1,592 human disease phenotypes. We selected colon cancer as a case study and analyzed our results in three directions; i) one stop legacy knowledge-shop for the effect of food on disease, ii) discovery of novel bioactive compounds with drug-like properties, and iii) discovery of novel health benefits from foods.

This works represents a systematized approach to the association of food with health effect, and provides the phytochemical layer of information for nutritional systems biology research. The paper also shows as a proof-of-concept that a systems biology approach to diet is meaningful and demonstrates some basic principles on how to work with diet systematic.

The second chapter of this thesis we developed the resource NutriChem v1.0. A food-chemical database linking the chemical space of plant-based foods with human disease phenotypes and provides a fundamental foundation for understanding mechanistically the consequences of eating behaviors on health. Dietary components may act directly or indirectly on the human genome and modulate multiple processes involved in disease risk and disease progression. The database has been created from text mining. The database and its content have been made available to the public from our webserver NutriChem: http://cbs.dtu.dk/services/NutriChem-1.0

The third chapter of the thesis is on developing a molecular roadmap of drug-food interactions. Our main hypothesis in the current work is that the complex interference of food on drug pharmacokinetic or pharmacodynamics processes is mainly exerted at the molecular level via natural compounds in food that are biologically active towards a wide range of proteins involved in drug ADME and drug action. Hence, the more information we gather about these natural compounds, such as molecular structure, experimental and predicted bioactivity profile, the greater insight we will gain about the molecular mechanisms dictating drug-food interactions, which will help us identifying, predicting and preventing potential unwanted interactions between foods and marketed- or novel drugs. Unlike drug bioactivity information that has already been made available for system-level analyses, biological activity data and source origin information of natural compounds present in food are scarce and unstructured. To this end, we integrate protein-chemical interaction networks, gene expression signatures and molecular docking to provide the foundation for understanding mechanistically the effect of eating behaviors on therapeutic intervention strategies.

The fourth chapter of the thesis is a case study on diet-colon cancer through candidate molecular interaction networks. The study shows a holistic examination of the dietary components for exploring the mechanisms of action and understanding the nutrient-nutrient interactions. In this paper we used colon cancer as a proof-of-concept for understanding key regulatory sites of diet on the disease pathway. We propose a framework for interrogating the critical targets in colon cancer process and identifying plant-based dietary interventions as important modifiers using a systems chemical biology approach.

The fifth chapter of the thesis is on discovering of novel anti-ovarian cancer compounds from our diet. Ovarian cancer is the leading cause of death from gynecological disorders with an increasingly high incidence, especially in the western world. Epidemiological studies suggest that some dietary factors may play a role in the development of ovarian cancer; so far most studies have shown up inconclusive. In the present study we disclose novel anti-ovarian cancer compounds from our diet with activity against ovarian cancer, through text mining and a system-wide association of phytochemicals, foods and health benefits on human ovarian cancer. We selected several compounds that where predicted to have anti-ovarian cancer activities, using chemoinformatics approaches and evaluated and confirm their activities in vitro.

Resumé (Dansk)

"Forebyggelse er bedre end helbredelse", og når det kommer til menneskers sundhed, udmønter dette sig i mange samfundsøkonomiske fordele. Stort set alle de cellulære processer, herunder alle trin i strømmen af genetisk information fra gen udtryk til proteinsyntese og nedbrydning, kan påvirkes af kost og livsstil. Svarende til lægemidler indeholder næringsstoffer en række forskellige forbindelser, der virker som modifikatorer for funktion og stabilitet. Imidlertid er graden af kompleksitet i ernæring studier yderligere påvirket af den samtidige tilstedeværelse af en række næringsstoffer, med forskellige kemiske strukturer, der kan have mange størrelse med forskellige tilhørsforhold. Det er klart, dette adskiller den ernæringsmæssige studier fra farmakologiske studier, hvor enkelte elementer anvendes ved lave koncentrationer og med en relativ høj affinitet og specificitet i et lille antal omhyggeligt udvalgte mål. Vores behov for grundlæggende forståelse af byggestenene i de komplekse biologiske systemer havde været den vigtigste årsag til den reduktionistiske tilgang, der blev primært anvendt i førhen for at belyse disse systemer. I dag er det almindeligt anerkendt, at systemer og netværks biologi har potentiale til at øge vores forståelse af, hvordan små molekyler påvirker metaboliske veje og homeostase, hvordan disse perturbationsmetoder påvirker sygdomstilstand, og i hvilket omfang de enkelte genotyper bidrage til dette. En frugtbart strategi til ernæringsmæssige forskning er derfor at låne metoder, der er godt etableret i medicinsk og farmakologisk forskning.

I denne afhandling, bruger vi avancerede data mining-værktøjer til konstruktion af en database med tilgængelige, state-of-the-art oplysninger om samspillet mellem mad og dets molekylære komponenter med biologiske systemer og deres forbindelse til sundhed og sygdom. Databasen vil blive beriget med forudsete vekselvirkninger mellem fødevare-komponenter og proteinerne i vores krop, baseret på deres strukturelle og pharmacophore lighed med kendte ligander. I forlængelse af dette, vil de sammenslutninger af bioaktive fødevarer komponenter med metaboliske veje undersøges ud fra et kemisk-protein-netværk perspektiv, mens deres virkninger i netværk robusthed vil blive yderligere bekræftet af proteom analyser og high-throughput genotype-fænotype karakterisering.

Det første kapitel i denne afhandling handler om udviklingen af vores datagrundlag. I dette arbejde anvender vi tekst mining og Naïve Bayes klassifikation til at samle viden om af fytokemikalier i plantefødevarer og deres fødevare-sygdom sammenhænge, hvor vi skelner mellem sygdomsforebyggelse/forbedring og sygdomsprogression. Vi søger efterfølgende forekomsten af fytokemikalie-sygdoms par og vi identificerede 20.654 fytokemikalier fra 16.102 planter, der er forbundet til 1.592 humane sygdomsfænotyper.

Vi valgte tyktarmskræft som case og har analyseret vores resultater i tre retninger; i) et one-stop videns punkt for effekten af ernæring på sygdom, ii) opdagelsen af helt nye bioaktive forbindelser med stof-lignende egenskaber, og iii) opdagelse af nye sundhedsmæssige fordele fra fødevarer. Dette arbejde repræsenterer en systematiseret tilgang til forståelsen af fødevarer og deres sundhedseffekt, og giver et fytokemikalie lag af oplysninger til brug inden for ernæringsmæssig systembiologisk forskning. Dette arbejde virker også som et proof-of-concept, at en systembiologisk tilgang til forskning inden for ernæring er meningsfuldt og demonstrerer nogle grundlæggende principper om, hvordan man arbejder med ernæring systematisk.

Det andet kapitel i denne afhandling er en database publikation. Vi udviklede ressourcen NutriChem v1.0. En fødevare-kemisk database der forbinder kemi fra plante-baserede fødevarer med humane sygdomsfænotyper og giver et grundlæggende fundament for at forstå konsekvenserne af spiseadfærds virkning på sundhed mekanistisk. Kostbestanddele kan fungere direkte eller indirekte på det menneskelige genom og modulere processer involveret i risiko for sygdom og sygdomsprogression. Databasen er udarbejdet ved brug af tekst mining. Databasen og dens indhold er blevet gjort tilgængelige for offentligheden via vores webserver NutriChem: http://cbs.dtu.dk/services/NutriChem-1.0

Den tredje kapitel i denne afhandling handler om udvikling af en molekylær køreplan for lægemiddel-fødevare interaktioner. Vores vigtigste hypotese i det aktuelle arbejde er, at den komplekse indblanding af fødevarer på lægemidlers farmakokinetiske eller farmakodynamiske natur hovedsageligt udøves på det molekylære plan via de naturlige stoffer der findes i fødevarer. Disse stoffer er biologisk aktive mod en bred vifte af proteiner involveret i ADME og lægemiddelvirkning. Derfor, jo flere oplysninger, vi indsamler om disse naturlige forbindelser, såsom molekylær struktur og bioaktivitets profil, jo større indsigt, får vi om de molekylære mekanismer dikteret af lægemiddel-fødevarer interaktioner, som vil hjælpe os med at identificere, forudsige og forebygge potentielt uønsket samspil mellem mad og markedsførte- eller nye lægemidler. I modsætning til lægemidlers bioaktivitets oplysninger, der allerede er gjort tilgængeligt for system-niveau analyser, biologiske aktivitetsdata og kilde oprindelsesoplysninger af naturlige forbindelser til stede i fødevarer er begrænsede og ustruktureret. Til dette formål har vi integreret protein-kemiske interaktions netværk, gen-ekspression signaturer og anvendt molekylær docking for at give et grundlag for at forstå effekten af spise adfærd på terapeutisk interventions strategier.

Det fjerde kapitel i denne afhandlingen er et casestudie om diæt og tyktarmskræft gennem kandidat molekylærer interaktion netværk. Undersøgelsen viser en holistisk gennemgang af kost komponenter for at udforske de virkningsmekanismer og forstå samspillet af næringsstoffer. I denne artikel har vi brugt tyktarmskræft som et proof-of-concept for at forstå vigtige regulatoriske komponenter i kosten.

Vi foreslår en ramme for undersøgelse af kritiske mål i tyktarmskræfts proces og identificere plantebaserede diætinterventioner som vigtige modifikatorer ved hjælp af en system-kemisk biologi tilgang.

Det femte kapitel i denne afhandling handler om opdagelsen af nye anti- kræft i æggestokkene stoffer fra vores kost. Kræft i æggestokkene er den førende dødsårsag fra gynækologiske lidelser med en stadig høj forekomst, især i den vestlige verden. Epidemiologiske undersøgelser tyder på, at nogle kost faktorer kan spille en rolle i udviklingen af kræft i æggestokkene. I den foreliggende undersøgelse afslører vi nye potentielle anti- kræft i æggestokkene stoffer fra vores kost, via tekst mining og et system af fytokemikalier, fødevarer og sundhedsmæssige fordele på menneskers kræft i æggestokkene. Vi valgte flere stoffer, der blev forudsagt til at være aktive fra vores analyse, afprøvede deres aktivitet i celle linje studier og kunne heraf bekræfte den forudsagte aktivitet for flere af stofferne.

Acknowledgements

No thesis is a one-man show, therefore I would like to acknowledge a number of people that have helped and supported me throughout my work on this thesis.

I have had the great pleasure to be staying at the Center for Biological Sequence Analysis, with its helpful staff and fantastic atmosphere. The center is led by Professor Søren Brunak who undoubtedly provided important feedback during the development of the thesis and helped me pointing the research the right direction. Without your feedback the thesis wouldn't have been possible. I have also had the great pleasure of working with a truly brilliant mind, Professor Lars Juhl Jensen at the University of Copenhagen. You expertise and feedback on the development of our text-mining pipeline has been crucial for our work to reach its high quality. I would also like to thank Sune Pletscher-Frankild who was working at the NNF Centre for Protein Research. You introduced me to the field of text-mining and help me structure my coding. Your feedback truly saved me a lot of work and time.

I would also like to give a special thanks to Sonny Kim Kjærulff who introduced me to the computational methods in Chemoinformatics. The methods used widely throughout this thesis. You built the foundation for both me and a lot of other students in our group, to rapidly pick up on the computational methods. I would also like to send a special thanks to Melanie Khodaie who introduced me to the world of Adobe Illustrator and Adobe Photoshop. Your helped me improve my skills on presentations, illustrations and figures.

I would also like to thank Peter Wad Sackett and John Damm Sørensen for providing incredible technical assistance on the department servers. Thanks to Lone Boesen and Dorthe Kjærsgaard for helping out on all the non-scientific issues. You truly made my studies a smooth.

I would also like to thank my officemates, Ulrik Plesner Jacobsen, Karin Marie Brandt Wolffhechel and Juliet Wairimu Frederiksen for always helping me out whenever needed.

In the end, I would like to give a special thanks to my amazing supervisors' associate professor Gianni Panagiotou and associate professor Irene Kouskoumvekaki. Thank you for believing in me and for always being there for me, guiding me safely through the transition process of my PhD. I am confident that your training will help me in all future career. The four years working together has been a truly inspiring and fascinating. Also thanks to Bernard Ni from the University of Hong Kong, for helping out with the curation of our database.

Publications

Included in thesis

Chapter I — **Jensen, K.**, Panagiotou, G., and Kouskoumvekaki, I. (2014). Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level. PLOS Computational Biology.

Chapter II — **Jensen, K.**, Panagiotou, G. and Kouskoumvekaki, I. (2014). NutriChem: a systems chemical biology resource to explore the medicinal value of diet. Nucleic Acids Research, Database issue. (Submitted)

Chapter III — **Jensen, K.**, Ni, B., Panagiotou, G., Kouskoumvekaki, I. (2014). Developing a molecular roadmap of drug-food interactions. PLOS Computational Biology. (Submitted)

Chapter IV — Westergaard, D., Li, J., **Jensen, K.**, Kouskoumvekaki, I., Panagiotou, G. (2014). Exploring Mechanisms of Diet-Colon Cancer Associations through Candidate Molecular Interaction Networks, BMC Genomics.

Chapter V — **Jensen, K.**, Panagiotou, G., and Kouskoumvekaki, I. Discovering novel anti-ovarian cancer compounds from our diet. (In preparation)

Not included in thesis

Jensen, K., Plichta, D., Panagiotou, G., and Kouskoumvekaki, I. (2012). Mapping the genome of Plasmodium falciparum on the drug-like chemical space reveals novel anti-malarial targets and potential drug leads. Mol Biosyst 8, 1678–1685.

Introduction

The structure of the thesis is illustrated in Figure 1. In the following we will walk through the basic principles of the thesis, and provide an overview of the theoretical background of its content. First we define the basic principles of nutritional systems biology that govern our research. Following this, we introduce our data resource, PubMed/MEDLINE, and the main concepts behind data extraction. Finally we present publicly available chemical and biological data sources that we have integrated in our work.

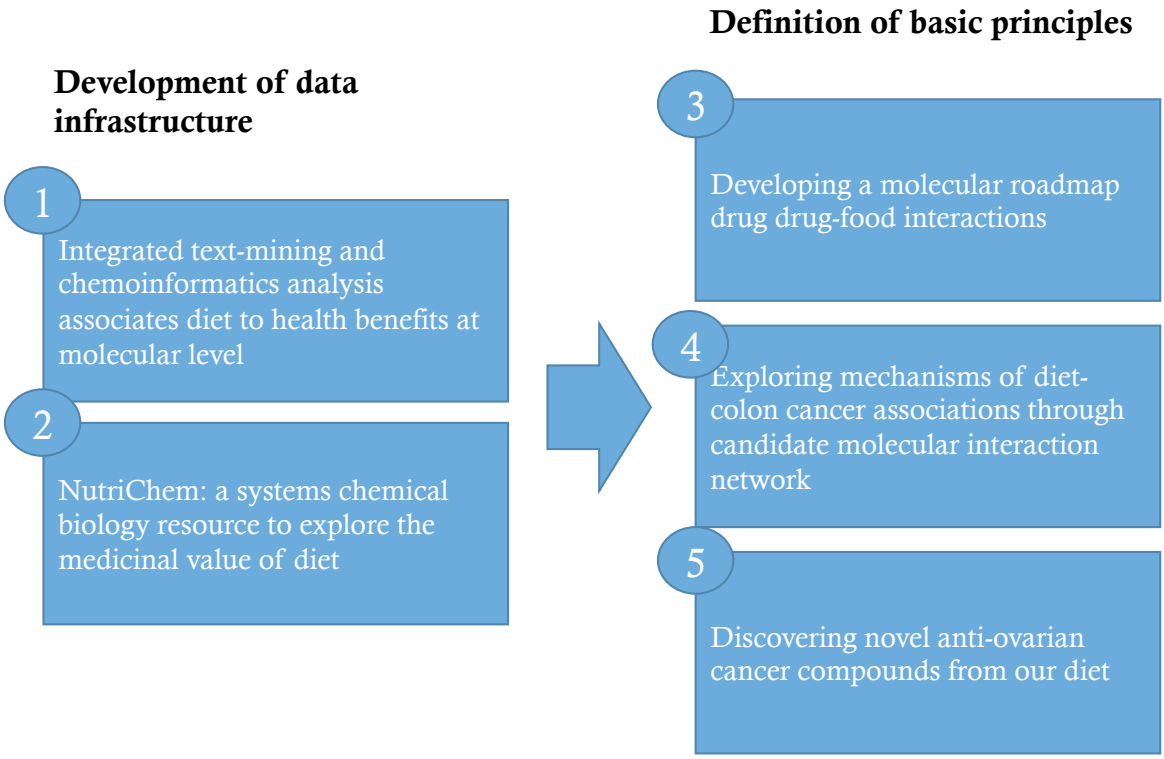


Figure 1: In the first and second chapter we develop the data infrastructure and define the basic principles of our research. In the following three chapters we explore our data warehouse in respective case studies on drug-food interactions and effects of food and its components on colon and ovarian cancer.

Nutritional systems biology

Similar to the role of pharmaceuticals, nutrients contain a number of different compounds that act as modifiers of network function and stability. However, the level of complexity in nutrition studies is further increased by the simultaneous presence of a variety of nutrients, with diverse chemical structures that can have numerous targets with different affinities and specificities. Obviously, this differentiates the nutritional from the pharmacological studies, where single elements are used at low concentrations and with a relatively high affinity and specificity in a small number of thoroughly selected targets.

Our need for fundamental understanding of the building blocks of the complex biological systems had been the main reason for the reductionist approach that was mainly applied in the past to elucidate these systems. Nowadays, it is widely recognized that systems and network biology has the potential to increase our understanding of how small molecules affect metabolic pathways and homeostasis, how this perturbation changes at the disease state, and to what extent individual genotypes contribute to this. A fruitful strategy in approaching and exploring the field of nutritional research is, therefore, to borrow methods that are well established in medical and pharmacological research. Molecular interaction networks could provide a convenient and practical scaffold to bridge the gap between nutritional research and systems biology and to make possible the designing of optimal diets that would allow health maintenance and disease preventions for individuals.

PubMed

MEDLINE (Medical Literature Analysis and Retrieval System Online), the database behind PubMed, is a literature database of life sciences and biomedical information. It is a database developed by the U.S. National Library of Medicine database (NLM)(1982). Currently, there are 5080 journals indexed as ‘Index Medicus’, which is the core content of the database. There are additional 575 journals not indexed in ‘Medicus’ within the following areas: 90 journals within dentistry, 18 within AIDS/HIV, 15 Consumer Health, 183 Nursing, 101 Health care administration, 85 health care technology, 80 history of medicine (http://www.nlm.nih.gov/bsd/num_titles.html, Accessed Feb. 28 2014).

The first version of the database can be traced back to a collection of books in the US Surgeon General’s office and was in time expanded with the aim to become more complete within health science as the ‘Index Medicus’. It was later developed in an electronic version, MEDLARS, and became online in 1996 as MEDLINE (Pritchard and Weightman, 2005).

For many years, the U.S. National Library of Medicine (NLM) has considered MEDLINE to be the definitive version of its indexing data (http://www.nlm.nih.gov/pubs/techbull/mj04/mj04_im.html, Accessed Feb. 28 2014). Today, the scope of journals indexed in MEDLINE ranges from life sciences to biomedicine, biochemistry, behavioral sciences, chemical sciences, and bioengineering. The objective of the choice of journals is to provide information on relevant literature to health professionals and others engaged in research or education. January 2014, the database had more than 22 million references to journal articles within life science (2014). The development of PubMed has been illustrated in Figure 2A.

The NLM, National Center for Biotechnology Information has invested huge efforts on indexing the literature and this has led to an extensive database with literature meta-data. The meta-data includes among others information on 'genes', 'proteins', 'chemicals' or 'diseases' related to the indexed literature. While biological sciences are moving towards a systems approach, meta-data has become more and more important for literature search for scientist and researchers. This has led to the development of the National Center for Biotechnology information (NCBI) E-utilities, which are a set of eight server-side programs that provide a stable interface into the Entrez query and database system at the NCBI (2010a).

MEDLINE is provided in XML file-format (http://www.nlm.nih.gov/bsd/mms/medlineelements.html, http://www.w3schools.com/xml/xml_whatis.asp, Accessed Feb. 28 2014). In this file each journal article is listed as an xml-object with a set of associated (meta) data fields. Examples of such fields are 'title', 'author', 'abstract´ and 'journal'. However, there is an additional vast amount of data-fields available for each article record. MEDLINE includes some special data-fields, which function as controlled vocabularies to ease the search within certain scientific areas. The MeSH term is one such term that is a controlled vocabulary of medical subject headings (http://www.nlm.nih.gov/bsd/mms/medlineelements.html, Accessed Mar. 2 2014). The data-structure of MEDLINE is shown in Figure 2B.

The MeSH term is used to provide a 'medical' characterization of the content of the articles. The substance name term is another type of meta-data that contains any of the 3 types of supplementary concept record data 1) MeSH SCR chemical and drug terms, 2) protocol terms and 3) non-MeSH rare disease terms. The gene symbol field is another meta-data that contains the "symbol" or abbreviated form of gene names as reported in the literature.

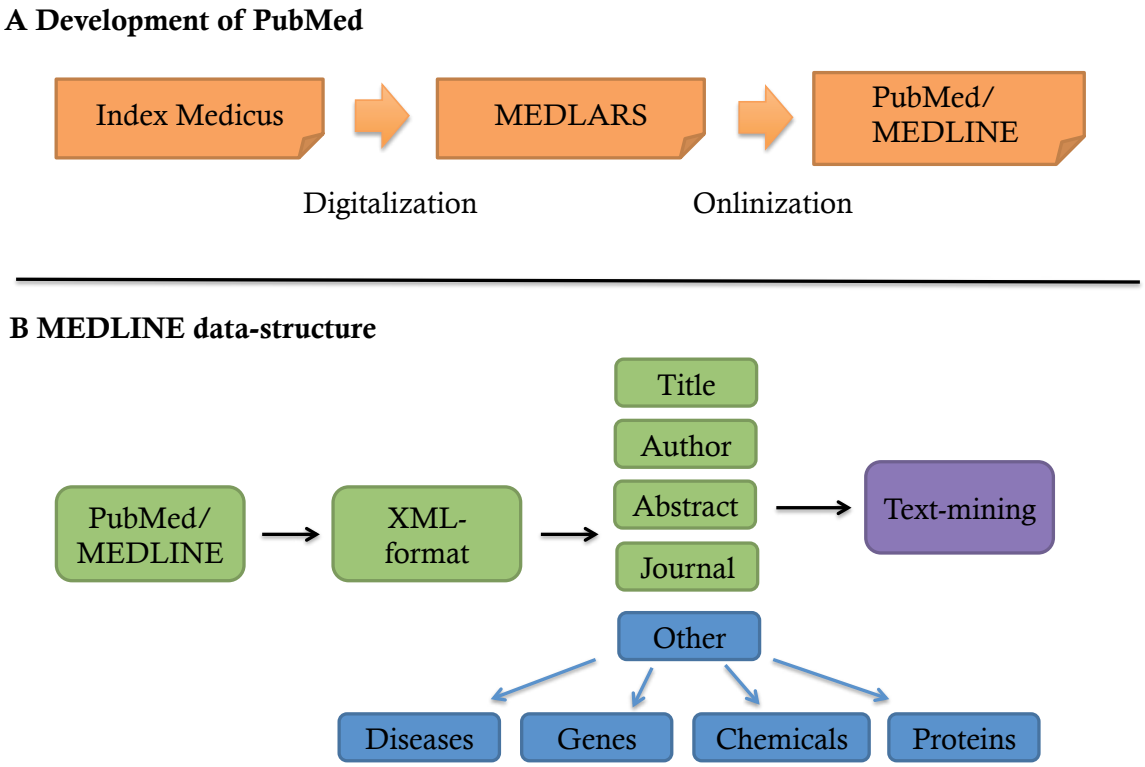


Figure 2: A) The development of PubMed/MEDLINE through time. The first version of the database can be traced back to an expanded version of the US Surgeon General's Office named 'Index Medicus'. The database was later developed into an electronic version 'MEDLARS' and become online in 1996 as the MEDLINE database. B) The MEDLINE data-structure. MEDLINE is provided in XML-file format that stores the articles as xml-objects. These objects have a set of properties such as 'title', 'author', 'abstract', 'journal' and a set of biochemical meta-data with disease, genes, chemicals and protein annotations.

PubMed's secondary uses

The secondary use of health information includes uses outside of direct health care delivery; including activities such as research, analysis and quality measurement. The secondary uses of health information typically aims to advance our health care knowledge or expertise (Safran et al., 2007). Searching and identifying literature in PubMed/MEDLINE effectively is a learned skill. For some searches the large numbers of retrieved articles causes frustration because of the lack of overview. A search that returns thousands of articles is not comprehensive, because it's hard to get an overview of the content (Jain and Raut, 2011). Therefore, scientists have begun exploring other means of utilizing and navigating vast amount of relevant literature. This has led to an increasing interest in text mining and automated information extraction methods driven by an increasing number of electronically available publications, stored in databases such as PubMed.

Biomedical text mining refers to text mining applied to text and literature at the edge of the field of natural language processing, bioinformatics, medical information and computational linguistics. It is related to identification of biological entities, protein and gene names in free text. However, automated extraction of information has also been used to extract and compile databases on protein-protein interactions and to develop functional concepts of genes and gene ontologies (http://en.wikipedia.org/wiki/Biomedical_text_mining, Accessed Mar. 2 2014).

Protein-Protein Interaction (PPI) extraction has become an important area in biomedical text mining, because the PPI information has become critical in the understanding of biological processes (Kim et al., 2008). Today, there is a vast amount of a variety of web-servers that provides PPI information.

The web-server 'PIE' (Protein Interaction information Extraction) is such a web-service that is created to extract relevant PPI articles from MEDLINE (Kim and Wilbur, 2011; Kim et al., 2008, 2012). Another, biomedical text-mining web-server is 'KLEIO'. KLEIO is an advanced information retrieval system that provides an enriched searching for relevant literature within biomedicine. The web service provides searches in categories such as protein, gene, metabolite, disease, symptom, organ etc. and scores them based on their relevance (Nobata et al., 2008). A common feature for most biomedical text-mining services is that they provide 'improved' searches or structured searches. Instead of just providing another searching option, there have been some advances such as recognizing named entities (Rocktäschel et al., 2012). However, the more interesting directions biomedical text mining is when it's used to compile databases that can be used for system-wide modeling of biological systems. Such as ChEMBL, STRING or STITCH (Kuhn et al., 2010; Overington, 2009; Szklarczyk et al., 2011).

Text-mining

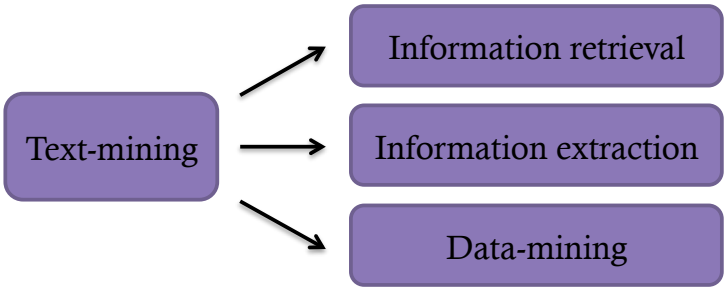
The increasing amount of published literature on biomedicine represents an extremely large source of information (Bundschus et al., 2008). With an overwhelming amount of textual information on molecular biology and biomedicine, there is a need for effective and efficient literature mining and knowledge discovery that can help biologist to gather and make use of the knowledge encoded in text documents (Zhou et al., 2004a). Automatic extraction of names plays an important role in the increasing challenge of extracting information from free text literature.

Traditionally, text mining is defined as the automatic discovery of previously unknown information by extracting information from text. However, in the community the text mining is often reduced to the process of highlighting small pieces of relevant information from large collections of raw text data (Spasic et al., 2005).

Text-mining can be divided into the following three categories: 1) Information retrieval, which gathers and filters relevant documents. 2) Information extraction, which extracts specific facts about predefined names or relationships. 3) Data mining, which is used to discover unsuspected associations between known facts, such as linking a plant to a disease (Spasic et al., 2005). The categories are shown in Figure 3A. The most recent development of text mining applications aim to assist researchers in obtaining and managing additional information by incorporating text-mining and natural-language processing tools for the extraction and compilation of functional characteristics (Krallinger et al., 2005). Most natural language processing and text-mining applications take advantage of a range of domain-independent methods such as part-of-speech (POS) taggers, which label each word with its corresponding part of speech or stemmers, that return the morphological root of a word form (Krallinger et al., 2005). Although extraction of relations between the recognized entities is also possible, most work to date is focused on the mere detection of entities (Bundschus et al., 2008). A flow diagram showing the relationship between natural language processing, classification and its use is shown in Figure 3B.

Text-mining applications are very powerful in their ability to integrate a broad spectrum of heterogeneous data resources, by transforming unlinked data into usable information and knowledge (Krallinger et al., 2005). All text-mining algorithms make errors when extracting facts from natural-language texts (Rodriguez-Esteban et al., 2006). Unfortunately, the current tools of information extraction produce imperfect, noisy results (Rodriguez-Esteban et al., 2006). Therefore, in biomedical applications it is crucial to assess the quality of individual facts – to resolve data conflicts and inconsistencies (Rodriguez-Esteban et al., 2006). Biomedical language and vocabularies are highly complex and rapidly evolving, making the identification of entities a cumbersome task (Krallinger et al., 2005). Among the strategies adopted to tag entities are methods such as ad hoc rule-based approaches, approaches using dictionaries with subsequent exact or inexact pattern matching, various machine-learning techniques and hybrid approaches that take advantage of different techniques (Krallinger et al., 2005).

A Text-mining categories



B Text-mining flow-diagram

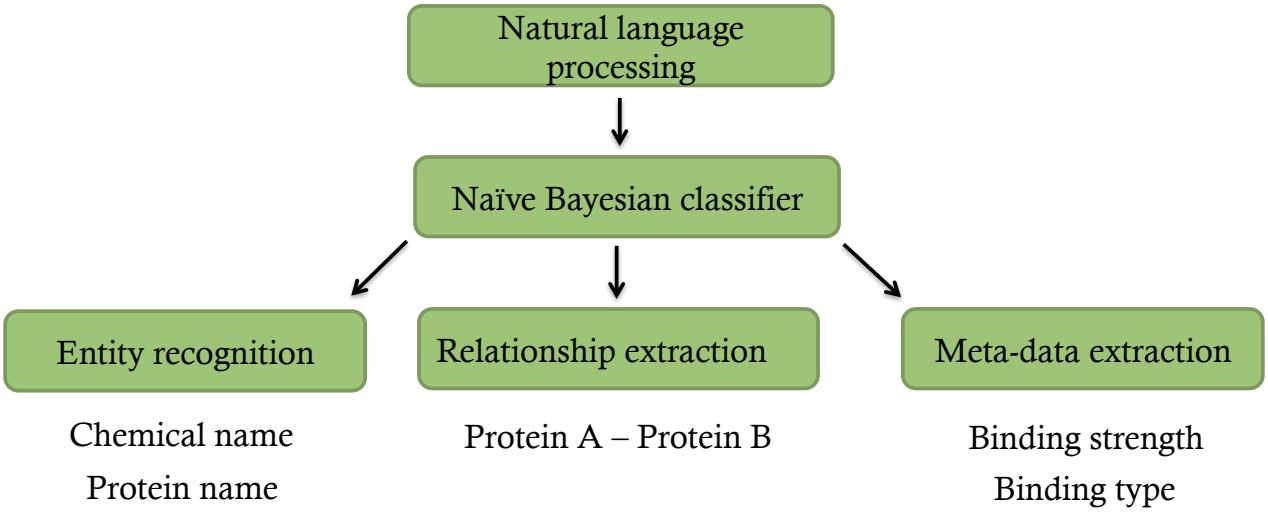


Figure 3: A) The three categories of text mining. The first category is information retrieval that gathers and filters relevant documents. The second category is information extraction that extracts specific facts about predefined names or relationships. The third category is data-mining that is used to discover unsuspected associations between known facts, such as linking a plant to a disease.
B) Flow-diagram showing the relationship between natural language processing, classification and its use. A classifier like the Naïve Bayesian classifier can be used as a classifier for natural language processing in order to recognize entities in textual data or to extraction relationships or meta-data.

Semantic browsing and automatic annotation (information retrieval)

Browsing biomedical literature involves two basic tasks: 1) Finding the right literature and 2) Making sense of its content. A lot of research has gone into supporting the task of finding the right literature either by means of standard literature searching or by means of semantically enhanced search (Guarino et al., 1999; McGuinness, 1998). Anyone who regularly reads life science literature often comes across names of genes, proteins or small molecules that one would like to know more about. Searching and retrieving information about these name entities is only possible if they have been annotated from the textual data. Annotation technologies allow users to associate meta-information with textual data, which can then be used to facilitate their interpretation (Kahan et al., 2001; Ovsiannikov et al., 1999; Vargas-Vera et al., 2002). While such technologies provide a useful way to support group-based and shared interpretation, they are nonetheless very limited, because the annotation is carried out manually. The quality of the sense making abilities of the annotations depends on the willingness of stakeholders to provide annotation, and their ability to provide valuable information. Annotation of literature meta-data is time consuming and expensive; most literature search databases such as PubMed does only provide annotation of named entities to some extent. This makes sematic browsing or exploration of the textual data complicated.

Because of the extensive manual annotation, semantic browsing does not work well without a system to automatically annotate the textual data. This has brought some focus to the development of automatic annotations systems and lead to the pioneering of the system Cohse (http://cohse.cs.manchester.ac.uk/, Accessed March 14 2014) that is a system for automatic annotation of literature using ontologies (Goble et al.). The system enables users to choose from different ontologies, including those outside life science. However, the public available version of Cohse has only very little functionality(Pafilis et al., 2009). Another tool has recently been developed, Reflect (http://reflect.ws/, Accessed March 14 2014), which is an extendable platform that allows users to annotate documents and webpages with scientific entities. Reflect servers as an augmented browsing tool, broadly useful to life science (Pafilis et al., 2009).

The idea about semantic browsing is that when browsing literature the system is capable of understanding the content of web pages. This is important when we are trying to retrieve the relevant literature within a specific field of study. The idea of the semantic browsing is to find a way for computers to understand the content and not just the structure. The power of semantic browsing becomes clear when, for example, one wishes to find the protein interactions related to the development of a particular disease.

The likelihood of knowing which keywords to search with to find all likely candidates is extremely low. However, with semantic browsing, it is unnecessary to know which keywords to search on, because the search engine automatically finds the words needed, making the search itself a subject area expert on every search performed. This is the fundamental principle of text mining and automatic information retrieval.

ChemTagger

Most of the literature has no or little annotation. Therefore, to extract useful information from textual data, the textual data needs first to be annotated. The automatic annotation of the textual data is achievable using semantic browsers such as Cohse (http://cohse.cs.manchester.ac.uk/, Accessed March 14 2014) or Reflect (http://reflect.ws/, Accessed March 14 2014). These web-services require the documents or textual data to be sent to a remote server through the Internet. Thus, the amount of text that can be annotated in a realistic time frame is limited.

ChemTagger that has been developed within the present project (https://pypi.python.org/pypi/ChemTagger, Accessed March 15 2014) is a module that can be loaded into the python3 interpreter and incorporated into other python3 programs. The words from the ontologies loaded into the module and the words in the textual data are linked with a rule-based algorithm designed to match names of chemical compounds, therefore the name 'ChemTagger'. However, the module is perfectly capable of matching names and terms of any other types as well. Through the python-shelve technology the module allows users to upload 'huge' ontologies of names and terms into virtual memory without putting pressure on the machine memory. This allows the user to annotate documents using ontologies at a size far exceeding the memory available on the machine. The technology also enables the user to do the annotation using several processors sharing the same virtual memory.

Annotation of textual data has so far only been possible using huge memory machines with a lot of processors, where the full ontologies to be used for the annotation has to be loaded into the machine memory before annotation can be performed. The makes the annotation processor expensive and resource consuming. The ChemTagger module allows users to do the annotation of a resource such as PubMed/Medline in a reasonable amount of time using only few system resources.

Natural Language processing

Initially, natural language processing (NLP) systems were based on complex sets of hand-written rules. However, with increasing computational power and the gradual lessening of the dominance of Chomskyan theories of linguistics (such as transformational grammar), (http://en.wikipedia.org/wiki/Natural_language_processing, Accessed March 17 2014), the machine-learning approach to language processing has been gaining popularity. There are several advantages to machine-learning algorithms compared to hand-produced rules: The learning procedures used during machine learning automatically snap-in on the common cases, whereas the hand-written rules are often not at all obvious in relation to where the effort should be put (http://en.wikipedia.org/wiki/Natural_language_processing, Accessed March 17 2014). Another important advantage is that automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar or erroneous input (http://en.wikipedia.org/wiki/Natural_language_processing, Accessed March 17 2014).

Naive Bayes classifier

When extracting information on the relationship between entities, the relationship of the entities depends on the context the entities. In order to make sense of textual data, the complex nature of natural language has to be processed (Rodriguez-Esteban et al., 2006). A means to process the complex nature of natural language and to extract information about the relationships of entities is to create a vector with text fragments for each entity relationship. These vectors can then be assigned a label, and a machine-learning algorithm can be trained to assign labels to new textual vectors (Rodriguez-Esteban et al., 2006). In text mining there can be differences in how the textual vectors are created as well as how they are assigned labels. Unfortunately, so far the current tools of information extraction still produce imperfect, noisy results (Rodriguez-Esteban et al., 2006).

A popular method for assigning new labels to textual vectors is the Naïve Bayes Classifier, which is a machine-learning method. The Naïve Bayes classifier in based on Bayes' theorem with independence assumptions between the predictors (http://www.saedsayad.com/naive_bayesian.htm, Accessed 15 March 2014). The model is also known as an independent feature model. For some types of probability models, the naïve Bayes classifier is easily trained and can be very efficient despite its relative simplicity. (http://en.wikipedia.org/wiki/Naive_Bayes_classifier, Accessed 15 March 2014). The Bayesian classifier assigns objects to classes if the posterior probability is greater for the class than for any alternative.

This posterior probability is computed in the following way (Rodriguez-Esteban et al., 2006):

$$P(C = c_k|F = F_i) = P(C = c_k) \times \frac{P(F = F_i|C = c_k)}{P(F = F_i)}$$

When training, the probability $P(F = F_i|C = c_k)$ is estimated from the training data as a ratio of the number of objects that belong to the class c_k and have the same set of feature values as specified by the vector F_i to the total number of objects in class c_k (http://en.wikipedia.org/wiki/Naive_Bayes_classifier, Accessed March 16 2014). The method handles a binary or a real value feature vector. While binary features are directly used in a frequency count for the classifier, the real value features are transformed, typically using a normal distribution (http://en.wikipedia.org/wiki/Naive_Bayes_classifier, Accessed March 16 2014).

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification. Independent variables are assumed and only the variance of the variables for each class needs to be determined (http://en.wikipedia.org/wiki/Naive_Bayes_classifier, Accessed March 16 2014). Even through the naive Bayes classifier represents some oversimplified assumptions, it works quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are some theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers (Zhang, 2004).

The main reason for choosing the naïve Bayesian classifier for natural language processing is that the classifier is relatively easy to train without the need of huge training sets. The theoretical principle behind the method is relatively simple, which makes it easy to implement and to debug. Due to the simplicity of the classifier results are easily reproduced. A python3 implementation of the naïve Bayesian classifier has been made available for download at the Python Package Index (PyPI) (https://pypi.python.org/pypi/NaiveBayes, Accessed March 16 2014).

Training and evaluating machine learning methods

Machine learning methods, as with all statistical models, need to be trained on an input data set. Training can take place either using a supervised approach, where the model fits training data to a specific parameter value, or unsupervised approach, where the model fits hidden structures in unlabeled training data (http://en.wikipedia.org/wiki/Unsupervised_learning, Accessed March 17 2014). In both cases, evaluation of the quality model is required to ensure that there has been no ‘overfitting’ of data and that the parameters of the model are meaningful. Overfitting can be avoided by using cross-validation during training, which predicts the ‘fit’ of the model to a hypothetical validation set (http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29,

Accessed March 17 2014). The most common form of cross-validation is the 10-fold cross validation, where the training set is divided into 10 parts. A model is then estimated (trained) on nine parts and evaluated on one part. The performance is measured and this is repeated 10 times until the model has been trained and validated on the whole set. The final performance measure is then calculated as the average from all 10 repeats (http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29, Accessed March 17 2014).

The confusion matrix is a table that allows for the visualization of the performance of a statistical model. In this table each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class: *True positives (tp)*, *false negatives (fn)*, *false positives (fp)* and *true negatives (tn)* (http://en.wikipedia.org/wiki/Confusion_matrix, Accessed March 17 2014). The confusion matrix is shown in Table 1.

Table 1: The table illustrates the confusion matrix. The table has the 'actual' labels in rows and the 'predicted' labels in columns.

	P' (Predicted)	**N' (Predicted)**
P (Actual)	True positives (tp)	False negatives (fn)
N (Actual)	False positives (fp)	True negatives (tn)

Model performance can be measured in a variety of ways and different fields of science tend towards using specific performance measures. *Accuracy* is the most common performance measure and is calculated from the number of true positives (tp) and true negatives (tn) divided by the total number of instances (http://en.wikipedia.org/wiki/Accuracy_and_precision, Accessed March 17 2014).

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}}$$

For text-mining a common performance measure is the *F1-score*, which considers both precision and recall of the model (http://en.wikipedia.org/wiki/F1_score, Accessed March 17 2014).

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The precision is the fraction of correct instances retrieved compared to the total amount of instances, while the recall is the fraction of correct instances that are retrieved correctly (http://en.wikipedia.org/wiki/Precision_(information_retrieval), Accessed March 17 2014).

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

Black listing of words

There are different types of biomedical text mining and depending on the type of text mining one is interested in, some specific issues is relevant and should be taken into consideration. Text-mining were information is extracted about named entities or terms such as protein-protein interactions or plant-compound associations has the limitation that the 'quality' words and their meaning depends on context (Spasic et al., 2005). The words in the dictionaries used for this type of text mining have a 'quality' because the information that can be extracted depends on the names and terms defined in the ontology and dictionary. Even when a single standardized ontology is used, it is not always straightforward to link textual information with the ontology. The two major obstacles are 1) inconsistent and imprecise practice in the naming of biomedical concepts (terminology), and 2) incomplete ontologies as a result of rapid knowledge expansion (Spasic et al., 2005).

Text mining of the type were information is extracted about name entities is limited to only recognize those names and terms that are included in the dictionary. The 'quality' issue of names and terms is not commonly addressed which we believe may be due to the fact that most of the dictionaries and ontologies used are curated manually (International Conference on Information and Knowledge Engineering and IKE '04, 2004). However, despite of whether the dictionaries and ontologies are compiled or curated manually a 'quality' of names and terms still applies. In order to control the quality of our dictionaries we need to use 'black listing' of certain words and terms to ensure the outcome from our extraction. For example, the word 'syndrome' is a valid disease term. However, the term is not very specific and does not specify anything biological meaningful. Therefore, this word should be put on the blacklist. The same applies to the word 'isolated', which is part of a synonym for the inflammatory disease of the myocardium: 'isolated fielder's'. However this term is too common to be of interest and its inclusion will yield a high number of false positive associations. It should therefore be either excluded our controlled during the text-mining process.

Dictionaries and ontologies

For text mining the dictionaries define the words or named entities that we are interested in extracting information about. The kind of information we could be interested in extraction could be the relationship between two named entities. The word ontology has long been used to describe the branch of philosophy that deals with the study of being. In the context of text mining the ontology (or dictionary) refers to a formal specification of a conceptualization of words (Cimino and Zhu, 2006). The ontologies can range from a verity of methods; from terminologies that are little more than manually created hierarchical arrangements of terms, whose developers nevertheless consider them to be ontologies; to ontologies which are compiled semi-automatic (Cimino and Zhu, 2006). A concept for ontologies which have had a great impact on how we make dictionaries and ontologies today, is the Unified Medical Language System (UMLS) (Cimino and Zhu, 2006). Unified Medical Language System (UMLS) was created by US National Library of Medicine and identifies terminological entities at three levels: The string (any name for a term in a terminology), the lexical group (to which strings of identical or near-identical lexical structure can be mapped) and the concept (to which strings of identical meaning can be mapped) (Cimino and Zhu, 2006).

Ontologies have been widely accepted as the most suitable representation model for conceptual information and an important building block of semantic web (International Conference on Information and Knowledge Engineering and IKE '04, 2004). Ontologies have been developed to capture the knowledge of the real world domain as a formal and explicit specification of a shared conceptualization (International Conference on Information and Knowledge Engineering and IKE '04, 2004). The benefits from ontologies come from the ontologies usage from knowledge sharing and reusability of domain knowledge (International Conference on Information and Knowledge Engineering and IKE '04, 2004). As controlled medical terminologies develop from simple code-name-hierarchy arrangements, into rich, knowledge-based ontologies of medical concepts. There has been an increasing demand on both the developers and users of the both ontologies and terminologies (Cimino, 2001).

NCBI Taxonomy

An important source of names of organisms, plants and foods for text mining is the NCBI taxonomy (McEntyre and Ostell, 2003). The NCBI taxonomy is a curated database of names and classifications for all organisms that are represented in GenBank (McEntyre and Ostell, 2003). The names are derived when new sequences are submitted to GenBank, the submission is checked for new organism names, which are then classified and added to the taxonomy database (McEntyre and Ostell, 2003). In a phylogenetic classification used for the NCBI taxonomy

classification scheme, the structure of the taxonomic tree approximates the evolutionary relationships among the organisms included in the classification (McEntyre and Ostell, 2003). April 2003, the database had 176,890 registered taxa's (McEntyre and Ostell, 2003). The NCBI taxonomy is available for download from the NCBI taxonomy website (http://www.ncbi.nlm.nih.gov/taxonomy, Accessed March 5 2014). The ontology is available in 'dmp' format and for text-mining purpose one can use the files: 'names.dmp' that has the names of the taxa's and 'nodes.dmp' which contains the taxonomic hierarchy. The file 'names.dmp' lists the scientific name and synonyms for organisms in 'nodes.dmp' (http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NCBI/metarepresentation.html, Accessed March 5 2014).

Human Disease Ontology

An important source of human disease names for text mining is the human disease ontology developed by the Center for Genetic Medicine of Northwestern University. The human disease ontology project is driven to large extent by the data aggregation and analysis needs of the NUgene Project at the Center for Genetic Medicine (http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page, Accessed March 5 2014).

The human disease ontology is designed to link disparate datasets through disease concepts and provides a computable structure of inheritable, environmental and infectious origins of human disease to facilitate the connection of genetic data, clinical data, and symptoms through the lens of human disease (http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page, Accessed March 5 2014). The human disease ontology is available for download in 'obo' format. The OBO flat file format is an ontology representation language. The format is similar to the tag-value format of the GO definitions file. 'obo' format is the text file format used by OBO-Edit, the open source, platform-independent application for viewing and editing ontologies (http://www.geneontology.org/GO.format.obo-1_4.shtml, Accessed March 5 2014).

The required tags for a human disease term on the ontology are: 'id' tag that is a unique id of the term and 'term name'. Each term has only one name defined. The terms have several optional tags that can be useful for text-mining purposes. The 'is_a' tag describes a sub-class relationship between one term and another. A term may have any number of 'is_a' relationships (http://www.geneontology.org/GO.format.obo-1_4.shtml, Accessed March 5 2014).

The PubChem Repository

PubChem (https://pubchem.ncbi.nlm.nih.gov, Accessed March 5 2014) is a database where users deposit chemical structures. The chemical structures are then in time validated and standardized to comprise a non-redundant set of chemical structures. The chemical names shown in the PubChem compound records are a composite derived from all linked substances where the names are ranked by frequency of use (http://pubchem.ncbi.nlm.nih.gov/help.html, Accessed March 5 2014). Users deposit their own compound records in PubChem. This has the advantage that PubChem has an incredible high coverage of compound names and structures. However, the compound names and structures deposited are not always accurate and curation of PubChem is slow. Therefore, the compound names and structures are not always meaningful but the database has the advantage of high coverage.

The PubChem database is available for download in ASN (ASN.1 formatted data), SDF (SDF formatted data) and XML (XML formatted data). The ANS and XML data formats contain the information related to the PubChem records while the SDF data format contains the chemical structure of the compounds ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/README-Compound, Accessed March 5 2014).

ChEBI

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds. The dictionary has been compiled from a number of different sources were incorporated and then merged (Degtyarenko et al., 2008). For the first release of the database, data was drawn from three main sources: 1) The IntEnz database an integrated relational enzyme database of EBI. The IntEnz database contains the enzyme nomenclature, the recommendations of the NC-IUBMB on the nomenclature and classification of enzyme catalyzed reactions. 2) The compounds of the KEGG Ligand Database, the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/ligand.html, Accessed March 6 2014). 3) A chemical ontology developed by Michael Ashburner and Pankaj Jaiswal, the initial alpha release was merged into ChEBI (Degtyarenko et al., 2008). One advantage of ChEBI is that the terminology used is explicitly endorsed, where applicable, by international bodies such as IUPAC (http://www.iupac.org) (general chemical nomenclature) and NC-IUBMB (http://www.iubmb.org) (biochemical nomenclature). This provides a dictionary with compound terms and names of a reasonable quality. The content of the database is relative controlled, through curation and merging of other database (Degtyarenko et al., 2008).

The ChEBI dictionary is available for download in SDF and OBO data formats. The SDF data format contains the chemical structure of the compounds, while the OBO data format contains the compound names and meta-data. The OBO flat file format is an ontology representation language. The format is similar to the tag-value format of the GO definitions file. OBO format is the text file format used by OBO-Edit, the open source, platform-independent application for viewing and editing ontologies (http://www.geneontology.org/GO.format.obo-1_4.shtml, Accessed March 6 2014). The required tags for the ChEBI dictionary are: 'id' tag that is a unique id of the compound term and 'term name'. The terms have several optional tags that can be useful for text-mining purposes. The 'is_a' tag describes a sub-class relationship between one term and another. A term may have any number of 'is_a' relationships (http://www.geneontology.org/GO.format.obo-1_4.shtml, Accessed March 6 2014).

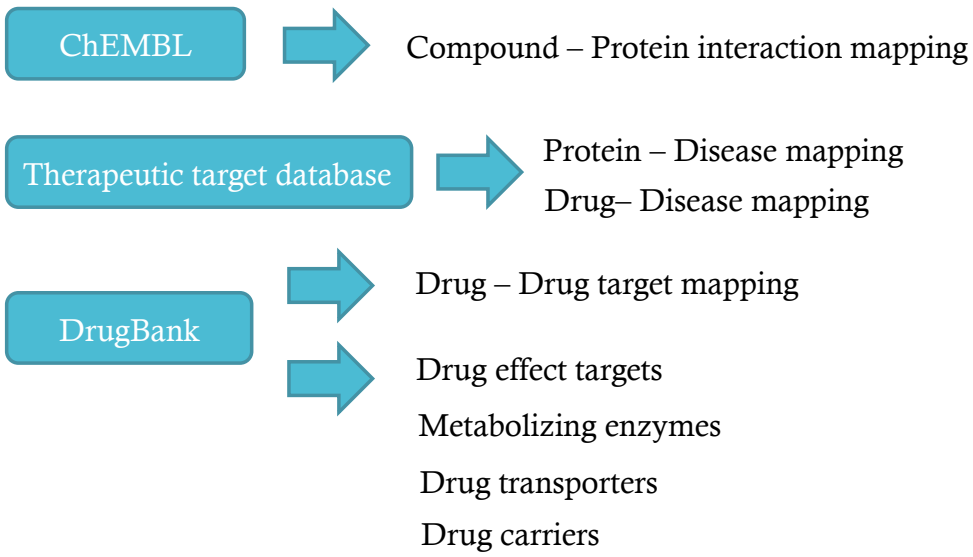


Figure 4: Diagram showing the three databases; ChEMBL, The Therapeutic Target database and DrugBank. The ChEMBL database contains compound – protein interactions and allows us to map external chemical structures to its compound – protein interaction data. The Therapeutic Target Database has information about proteins and their relation to diseases, including drugs and their relation to diseases. The database allows us to map external data to disease and drug information. The DrugBank database contains detailed information about drugs and similarly allows external data to be mapped. The database also contains information on drug effect targets, metabolizing enzymes, drug transporters and carriers.

Drug effect targets and human disease associations

ChEMBL

The effect of small molecule compounds can be linked to our proteome by mapping the small molecules to a chemical-protein interaction database. An important database with chemical-protein interactions is ChEMBL, which is a database developed by the European Molecular Biology Laboratory (Overington, 2009). At the time of writing the ChEMBL database is in version 17 with 1.520.172 compound records covering 12.077.491 biological activities. The database contains two relevant files, a file with the chemical structures in Chemical table file SDF format (Dalby et al., 1992) and a file with the mapping of proteins listed in ChEMBL to UniProt ID's (UniProt Consortium, 2014). The file with the chemical table file is convertible to a Simplified molecular-input line-entry system (SMILES) string (Weininger, 1988), which is a specification in form of a line notation for describing the structure of chemical molecules using short ASCII strings. There has since the introduction of the SMILES string been an extensive focus on the need of a simplified representation of chemical structures. For long the SMILES string has been the most dominant method of representing chemical structures in a simplified form. The SMILES is a trademark of Daylight Chemical Information Systems Inc. (Daylight). For this the development of the SMILES representation is controlled by Daylight.

In order to provide an open alternative to the SMILES string. The International Union of Pure and Applied Chemistry (IUPAC) introduced the International Chemical Identifier (InChI) as a standard for formula representation (Heller et al., 2013). The methods are similar and solve the same task. The advantage of InChI is that it's available under GNU Lesser General Public License and the development of the InChI is community driven. However, at the time of writing the SMILES still considered to have the advantage of being slightly more human-readable than InChI; it also has a wide base of software support with more robust implementations and extensive theoretical (e.g., graph theory).

ChEMBL has four main tables which allows us to link food compounds to interactions with the human proteins. The first table is a 'molecule dictionary' where the chembl_id's are mapped to molecule registration numbers (molregno). From molregno we can retrieve the activities of the compound using the 'activities' table. From the 'activities' table the 'assay_id' the target of the assay using the target id 'tid'. This target id 'tid' is then linked to a ChEMBL id specifying the protein targeted. Small molecules can then be mapped to compounds in ChEMBL, using the SMILES representation and the outcome proteins with ChEMBL id's can be linked to UniProt proteins. A diagram showing the use of the three databases ChEMBL, The Therapeutic target database and DrugBank is shown in Figure 4.

DrugBank

DrugBank is a comprehensive database that contains information on a wide range of approved drug and potential bioactive molecules. The database provides an important insight into the use of drugs and bioactive molecules by grouping them into categories; FDA-Approved, Small molecules; Experimental; Nutriceutical; Illicit Drugs and Withdrawn Drugs. For each of these categories DrugBank contains information on which proteins has are drug effect targets; drug enzymes; transporters and carriers (Knox et al., 2011; Wishart et al., 2006, 2008). Integration of the information of known drugs is a highly valuable asset to improve our understanding of less studied food-compounds. The database contains 6825 drug entries including 1541 FDA-approved small molecule drugs, 150 FDA-approved biotech (protein/peptide) drugs, 86 nutraceuticals and 5082 experimental drugs. In addition, the database contains 4323 non-redundant proteins. The advantage of this database is that the compounds can be linked through SMILES representation and the drug effect targets; absorption, distribution, metabolism, and excretion (ADME) proteins identified.

TTD: Therapeutic Target Database

The therapeutic target database (TTD) is a database that provides information on known and explored therapeutic proteins and targeted diseases. In addition, the database contains systematic information on drug to disease mapping which is less accessible in DrugBank (Chen et al., 2002; Zhu et al., 2010, 2012). Food-compounds and their interactions with human proteins, mapped through ChEMBL, becomes particular interesting when the interaction be linked with the therapeutic effects in TTD. At the time of writing, the database currently contains information on 2,025 targets, including 364 successful, 286 clinical trial, 44 discontinued and 1.331 research targets and 17.816 drugs, including 1.540 approved (Zhu et al., 2012). The therapeutic target database contains two important files. The first file is the main database with therapeutic target id’s for proteins with associated human disease names, which targeting the protein will have a therapeutic potential. The other file is the target information table, which maps the therapeutic target ids with UniProt.

Chapter 1: Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level

Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level

K. Jensen[1], G. Panagiotou[2,*], I. Kouskoumvekaki[1,*]

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark

2 School of Biological Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong

*** E-mail: Correspondence gipa@hku.hk, irene@cbs.dtu.dk**

Abstract

Awareness that disease susceptibility is not only dependent on genetic make-up, but can be affected by lifestyle decisions, has brought more attention to the role of diet. However, food is often treated as a black box, or the focus is limited to few, well-studied compounds, such as polyphenols, lipids and nutrients. In this work, we applied text mining and Naïve Bayes classification to assemble the knowledge space of food-phytochemical and food-disease associations, where we distinguish between disease prevention/amelioration and disease progression. We subsequently searched for frequently occurring phytochemical-disease pairs and we identified 20,654 phytochemicals from 16,102 plants associated to 1,592 human disease phenotypes. We selected colon cancer as a case study and analyzed our results in three directions; i) one stop legacy knowledge-shop for the effect of food on disease, ii) discovery of novel bioactive compounds with drug-like properties, and iii) discovery of novel health benefits from foods. This works represents a systematized approach to the association of food with health effect, and provides the phytochemical layer of information for nutritional systems biology research.

Author summary

Until recently diet was considered a supplier of energy and building blocks for growth and development. However, current research in the field suggests that the complex mixture of natural compounds present in our food has a variety of biological activities and plays an important role

for health maintenance and disease prevention. The mixture of bioactive components of our diet interacts with the human body through complex processes that modify network function and stability. In order to increase our limited understanding on how components of food affect human health, we borrow methods that are well established in medical and pharmacological research. By using text mining in PubMed abstracts we collected more than 20,000 diverse chemical structures present in our diet, while by applying chemoinformatics methods we could systematically explore their numerous targets. Integrating the above datasets with food-disease associations allowed us to use a statistical framework for identifying specific phytochemicals as perturbators of drug targets and disease related pathways.

Introduction

The increasing awareness of health and lifestyle in the last decade has brought significant attention from the public media to the role of diet. Typically, specific diets or single foods are associated with health and disease states through *in vivo* studies on humans or animal models, where the response of selected phenotypes, e.g. up-regulation or down- regulation of certain genes, is being monitored (Knekt et al., 2002; Wedick et al., 2012). Observational studies on populations with specific food preferences may also provide statistical evidence for the absence or prevalence of certain diseases in connection to certain dietary habits (Ferguson and Schlothauer, 2012). Even though these approaches have offered some useful insights for specific food types, they are frequently inconclusive due to small cohorts or limited focus both on the diet and the disease space. Most importantly, observations remain on the phenotypic layer, since diet is treated as a black box, when it comes to its molecular content. In the emerging field of systems chemical biology (Oprea et al., 2007) research is moving towards the network-based study of environmental exposures, (e.g. medicine, diet, environmental chemicals) and their effect on human health (Schadt et al., 2012). We believe that this shift in paradigm, where one considers the system of the molecular components of diet and their interplay with the human body, will build the basis for understanding the benefits and impact of diet on our health that will enable the rational design of strategies to manipulate cell functions through what we eat (Herrero, 2012; Panagiotou and Nielsen, 2009). However, to interpret the biological responses to diet, as well as contribute to the evidence in assigning causality to a diet-disease association, we need first to overcome the major barrier of defining the small molecule space of our diet. By assembling all available information on the complex chemical background of our diet, we can systematically study the dietary factors that have the greatest influence, reveal their synergistic interactions, and uncover their mechanisms of action.

In the present work we carried out text mining to collect in a systematic and high-throughput way all available information that links plant-based diet (fruits, vegetables, and plant-based beverages such as tea, coffee, cocoa and wine) with phytochemical content, i.e. primary and secondary metabolites, and human disease phenotypes. There are two reasons for focusing on the plant-based diet: *(1)* there is well established knowledge on the importance of fruit- and vegetable-rich diet in relation to human health e.g. nutraceuticals, antibiotics, anti-inflammatory, anti-cancer, just to name a few (Bravo, 1998; Colombo and Bosisio, 1996; Cowan, 1999; Gershenzon and Dudareva, 2007; Pandi-Perumal et al., 2006; Scalbert et al., 2005); *(2)* the huge diversity of the phytochemical space offers a fertile ground for integrating chemoinformatics with statistical analysis to go beyond the existing knowledge in the literature and suggest new associations between food and diseases.

Our text-mining strategy, based on dictionaries from the argument browser Reflect (Pafilis et al., 2009), Natural Language Processing (NLP) and Naive Bayes text classification (Berry and Kogan, 2010; Perkins, 2010), goes beyond mere retrieval of diet - disease associations, as it further assigns a positive or negative impact of the diet on the disease. With this work we aim to demonstrate how data from nutritional studies can be integrated in systems biology to boost our understanding of how plant-based diet supports health and disease prevention or amelioration. This wealth of knowledge combined with chemical and biological information related to food could pave the way for the discovery of the underlying molecular level mechanisms of the effect of diet on human health that could be translated into public health recommendations.

Results

Mining the phytochemical space

We extracted by text mining plant - phytochemical associations from 21 million abstracts in PubMed/MEDLINE, covering the period 1908-2012. We used relation keyword co-occurrences between plant names (both common names and scientific names) and small compound names and synonyms. First, the chemical name entities and plant name entities were recognized using a set of simple recognition rules. Then, a training set was manually compiled with abstracts mentioning plant - phytochemical pairs. Finally, a Naïve Bayes classifier was trained to correctly recognize and extract pairs of phytochemicals and plants that contain them. The performance of the classifier was quantitatively estimated to 88.4% accuracy and 87.5% F1-measure on an external test set of 250 abstracts. When the classifier was applied to the raw text of PubMed/MEDLINE, it associated 23,137 compounds to 15,722 plants – of which, approximately 2,768 are edible – through 369,549 edges. Since the total number of natural compounds discovered so far from all living species is estimated to be approximately 50,000

(Afendi et al., 2012), the retrieval of 23,137 phytochemicals solely by extraction of information from raw text of titles and abstracts in the PubMed domain provides a unique platform for obtaining a holistic view of the effects of our diet on health homeostasis. In order to collect all relevant available information for subsequent analyses, we integrated the data we collected via text mining with the Chinese Natural Product Database (Shen et al., 2003) (CNPD) and an Ayurveda (Polur et al., 2011) data set that we have previously curated in house. CNPD, which is a commercial, manually curated database, contains information on 16,876 unique compounds from 5,182 plant species associated through 21,172 edges. The Ayurveda data set includes information on 1,324 phytochemicals and 189 plants. After merging these two sources with the text-mined data and removing redundant information, we ended up with 36,932 phytochemicals and 16,102 plants. What further adds value to this pool of data is that all 36,932 compounds are encoded in Canonical SMILES and linked to a unique chemical structure, which allows the application of chemoinformatics tools for interrogating the human protein and disease space that these compounds may have an effect on.

Figure 5A shows the most well studied edible plants and the number of phytochemicals identified in each of them. Rice has the highest number of recorded phytochemicals (4,155 compounds), followed by soybean (4,064 compounds), maize (3,361 compounds) and potato (2,988 compounds). Figure 5B shows representative phytochemicals from our retrieved data that have made it all the way to the pharmacy shelves or have served as lead structures for drug development. Camptothecin is a natural compound that has led to the semi synthesis of the analogues irinotecan and topotecan, two antineoplastic enzyme inhibitors that are currently used in the treatment of colorectal and ovarian cancer, respectively. As camptothecin is highly cytotoxic, we have not encountered any common foods within the list of plants that contain it. Ergocalciferol (vitamin D2), on the other hand, has been traced in numerous plant sources, many of which are common foods, such as tomato, cacao and alfalfa. Ergocalciferol is an approved nutraceutical compound found in the market under various brand names that is used in the treatment of diseases related to vitamin D deficiency, such as hypocalcemia, rickets and osteomalacia.

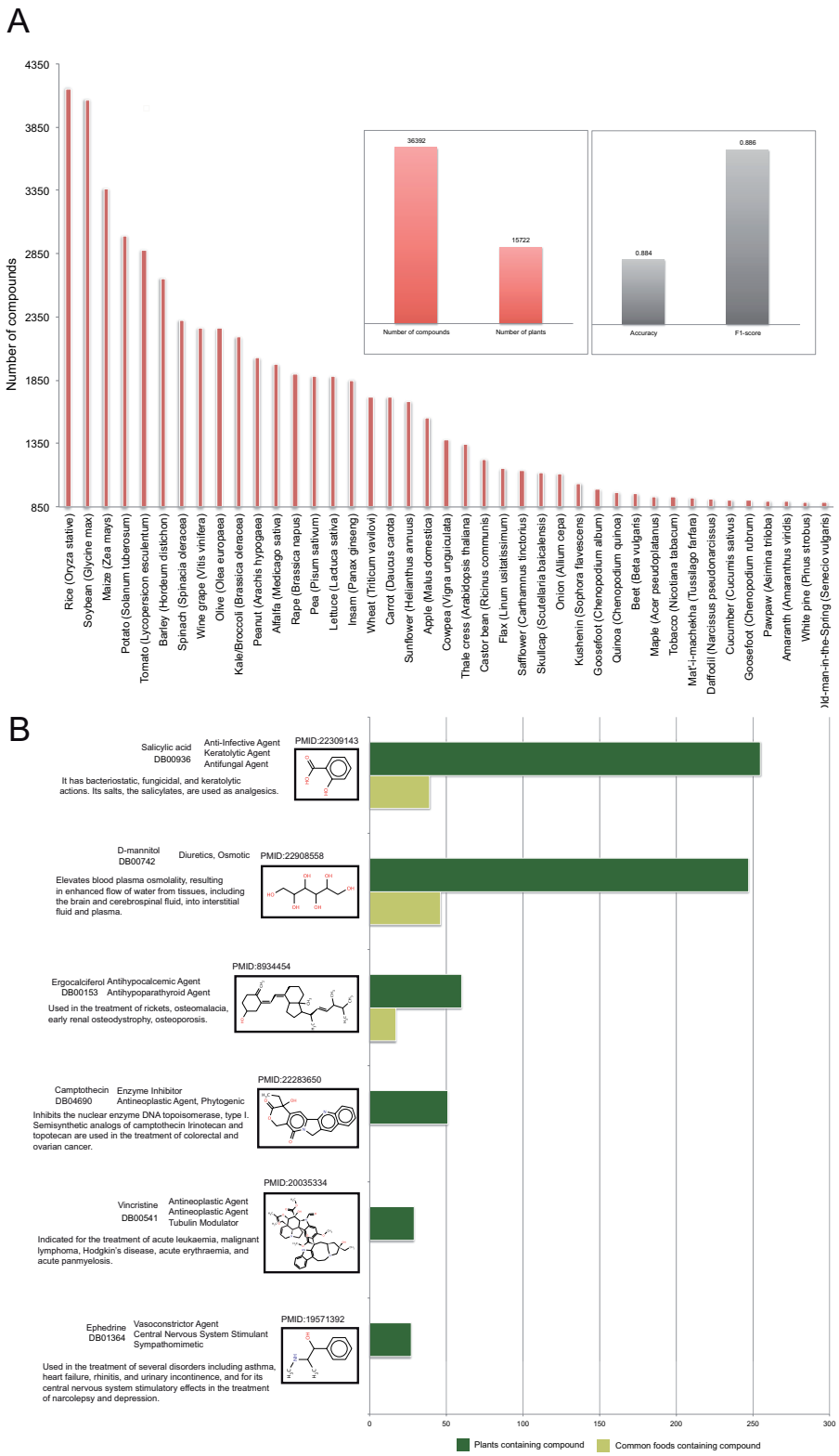


Figure 5: A) Distribution of phytochemicals on the plant space. Rice, soybean, maize and potato are the plants with the most recorded phytochemicals: 4,155, 4,064, 3,361 and 2,988 compounds respectively. B) Structures of representative phytochemicals. Structures that have made the way to the pharmacy shelves and their occurrence in respective edible sources.

Figure 5B brings also to light that natural compounds are commonly encountered in more than one plant, or family of plants. Previous studies have indicated that there are no consistent trends as to whether phytochemicals can be used as taxonomic markers or may occur in several unrelated plant families (Bravo, 1998; Wink, 2003). With this question in mind, we decided to examine how the 36,932 phytochemicals are distributed among neighboring and ancestral taxa and whether there are clusters of certain phytochemicals at specific parts of the taxonomy. Overrepresentation of phytochemicals on the taxonomy was calculated by using Fishers exact test, following the Benjamini-Hochberg procedure with a 5% False Discovery Rate (Yoav and Hochberg, 1995). Our analysis showed that only 8% of all phytochemicals are localized on certain parts of the taxonomy (Figure S1 and Table S1). For example the family of *Fabales – Fabaceae – Lens*, which includes lentils, and the *Sapindales – Rutaceae – Citrus* linkage, which includes orange, contain 60 out of 562 compounds and 42 out of 214 compounds, respectively (p-value $< 10^{-4}$) that are not found anywhere else on the taxonomy. On the other hand, compounds such as β-sitosterol, palmitic acid and catechin are spread all over the taxonomy (p-value $< 10^{-4}$). A possible interpretation of this finding is that the synthesis of small compounds in plants is mainly defined by short-term regulatory than long-term evolutionary adaptation to the environment.

Association of food with disease prevention or progression

To systematically associate plant-based diet with health effect we extracted by text mining plant - disease associations from 21 million abstracts in PubMed/MEDLINE, covering the period 1908-2012. In this manner we associated 7,106 plant species, 2,768 of which edible, with 1,613 human disease phenotypes. The performance of the classifier was quantitatively estimated to 84.5% accuracy and 84.4% F1-measure on an external test set of 250 abstracts. Natural Language Processing allowed us to add directionality to these associations, an extremely valuable feature for dietary recommendations. This enabled us not only to link a certain food to a disease, but also to characterize the association as being positive (food associated with disease prevention or amelioration) or negative (food associated with disease progress). Together with the temporal parameter that is included in the text-mined data (date of publication of articles that associate food to disease), one can make interesting observations as to when scientists began showing interest in the health effect of food and how opinion regarding a certain food has been varying throughout time.

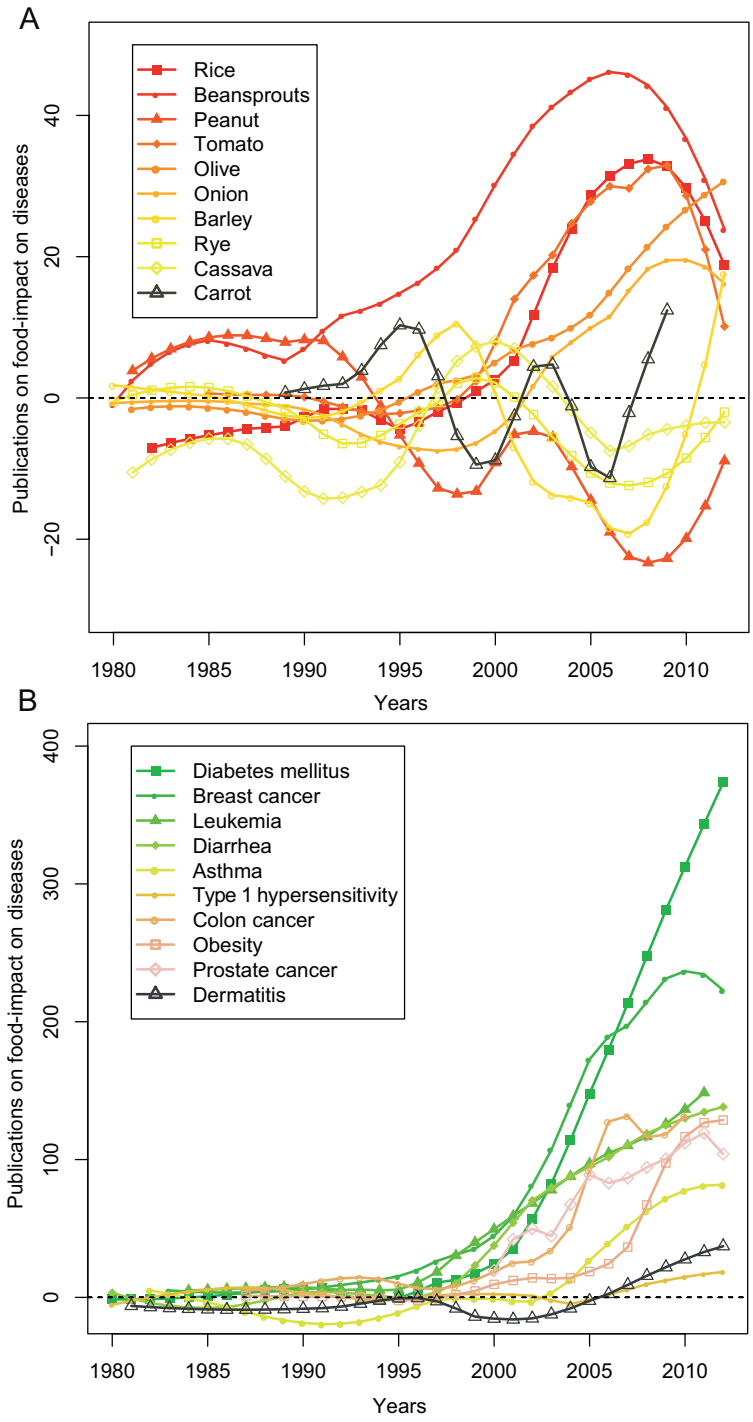


Figure 6: A) Examples of well-studied foods in relation to positive (disease prevention /amelioration) and negative (disease progression) effect on health. The focus varies from negative effects (below 0) to positive effects (above 0) over the years. The value on the y-axis denotes the number of negative publications subtracted from the number of positive publications in a given year. B) Examples of well-studied disease phenotypes in relation to food consumption. Likewise, the figure illustrates a change in focus from negative effects (below 0) to positives effects (above 0).

As shown in Figure 6A, research on the health effect of food effectively began in the early 80´s and until middle 90´s there was more research activity in relation to the negative effects of foods, such as their involvement in the development and progression of allergic reactions and asthma. However, the change of public opinion towards lifestyle and preventive strategies related to health in the last 15 years, resulted to an exponential growth of research papers reporting beneficial effects of plant-based foods against diabetes mellitus and different types of cancers (e.g. breast cancer, carcinoma and leukemia), not surprising since these diseases are the scourge of our time. Also of interest are the contradicting opinions over time on the health benefit of foods (Figure 6B). Until the beginning of the 21st century there were only sparse reports on the health benefits associated with rice consumption, while the last 10 years there are numerous reports describing the positive impact of a rice-based diet. The opposite trend is observed for peanuts, which was mainly studied for its beneficial role in cancer before a number of studies begun correlating its consumption with health problems, such as allergy and hypersensitivity.

The network of Figure 7A presents the most strongly supported associations of common foods and health benefits in the public literature. There are only a handful of common foods that have been associated either only positively or negatively with disease phenotypes. Consumption of broccoli, blueberry and camellia-tea for example, is consistently linked positively with a variety of disease phenotypes including diabetes mellitus, atherosclerosis and different types of cancers (Figure 7B). Cassava, a good source of carbohydrates but poor in protein, which constitutes the basic diet for many people in the developing world, has only negative associations with malnutrition, and malnutrition-related phenotypes (Figure 7C).

For the majority of cases however, a particular food is positively correlated with specific disease phenotypes and negatively with others, highlighting the importance of personalized dietary interventions; rice is one characteristic example, associated positively to hypertension, diabetes, colon- and breast cancer and negatively to dermatitis and hypersensitivity reactions. There are also several foods, including peanut, chestnut and avocado, consistently associated negatively with type-1 hypersensitivity and similar disease phenotypes, such as dermatitis, rhinitis and urticarial. Not surprisingly, a high number of publications exist for the negative effects of common foods such as wheat, barley and rye to celiac disease (also known as gluten intolerance). Figure 7 makes also evident that considerable research investments have been made in the past decades for enhancing our understanding of the association between diet and cancer; breast, prostate and colon cancers constitute the thickest edges on the network.

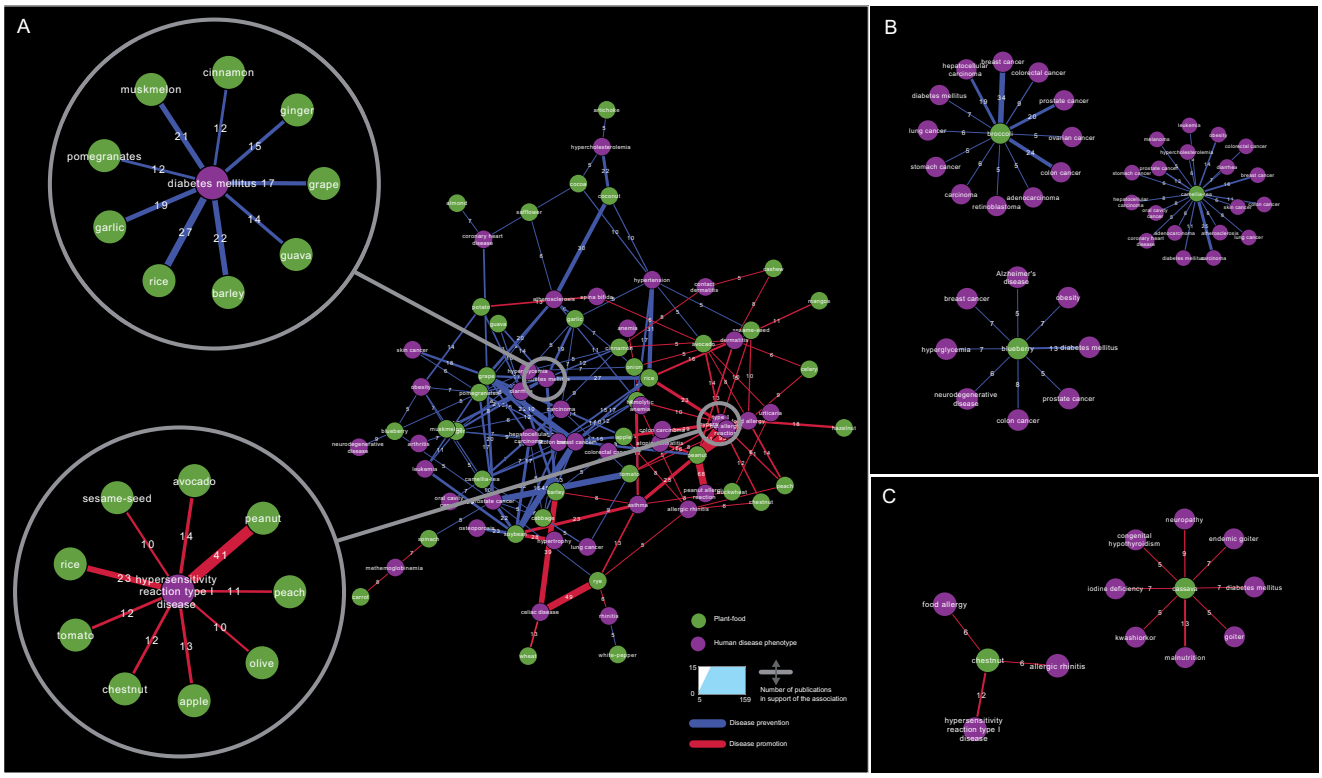


Figure 7: A) Disease phenotypes associated with common vegetables, fruits and plants of our diet. Foods are shown as green nodes and human disease phenotypes as purple nodes. Disease prevention/amelioration is depicted as a blue edge and disease promotion as a red edge. The size of the edge indicates the number of publications in support of the association. An edge is drawn between a food node and a disease node when there are at least five publications in support of this association. When a disease node has more than five edges, only the five strongest (with the most publication support) are shown on the network for the sake of clarity. Top left: zoom in the network formed between diabetes mellitus and foods that prevent/ameliorate the disease. Bottom left: zoom in the network formed between Type 1 hypersensitivity and foods that promote it. B) Examples of a vegetable (broccoli), a fruit (blueberry) and a plant-based beverage (camellia-tea) that are only positively associated with disease phenotypes. C) Two examples of foods that are only negatively associated with disease phenotypes.

Molecular level association of food to human disease phenotypes

Our main hypothesis for the molecular level association of a plant-based diet to human disease phenotypes is that the positive or negative effect of a certain food on human health is due to the presence of one or more bioactive molecules in it. Towards this end, we used Fisher´s exact test to systematically detect frequently occurring phytochemical - disease pairs through the phytochemical - food and food - disease relations that we extracted by text mining. At a 5% FDR we identified 20,654 phytochemicals connected to 1,592 human disease phenotypes, with approximately half of the disease associations being positive (Figure 8A). Some of these phytochemicals have been previously studied *in vitro* for potential biological activity. By integrating information from ChEMBL we find that, from the 20,654 phytochemicals that the above analysis suggests as bioactive, approximately 5,709 have been tested experimentally on a biological target. From the remaining phytochemicals, for which no experimental bioactivity data are available, 8,113 compounds are structurally similar to compounds with known protein targets (estimated with a Tanimoto coefficient > 0.85), indicating similar bioactivity, while the rest belong to a hitherto unexplored phytochemical space Figure 8B).

In order to get an estimate of the performance of our approach to associate phytochemicals to diseases, we used the Therapeutic Targets Database to annotate the protein targets from ChEMBL to diseases. From the 5,709 phytochemicals that are included in ChEMBL, almost half are active against a biological target that is relevant for the same disease as the one we have predicted (Figure 8C). Adding molecular-level information to food - disease associations allows us to zoom in the network of Figure 7 and generate lists of phytochemicals as promising drug-like candidates for subsequent target-based or cell line-based assay experiments, as we demonstrate in Table 2 with focus on a number of common cancer types. For example, 103 phytochemicals from 83 common foods (Kusari et al., 2011; Miller et al., 1977; Santos et al., 2011) that through our analysis are associated with lung cancer, are structurally similar with 23 drugs from DrugBank that are approved for use in lung cancer treatment. In addition, by integrating information from ChEMBL and TTD, we identify 1,070 phytochemicals from 119 common foods with experimental activity against a lung cancer drug target. For cancer types, such as endometrial cancer and adenocarcinoma, where the drugs currently available in the market are scarce, this approach could be of particular interest, as it provides new opportunities for the identification of new drug candidates.

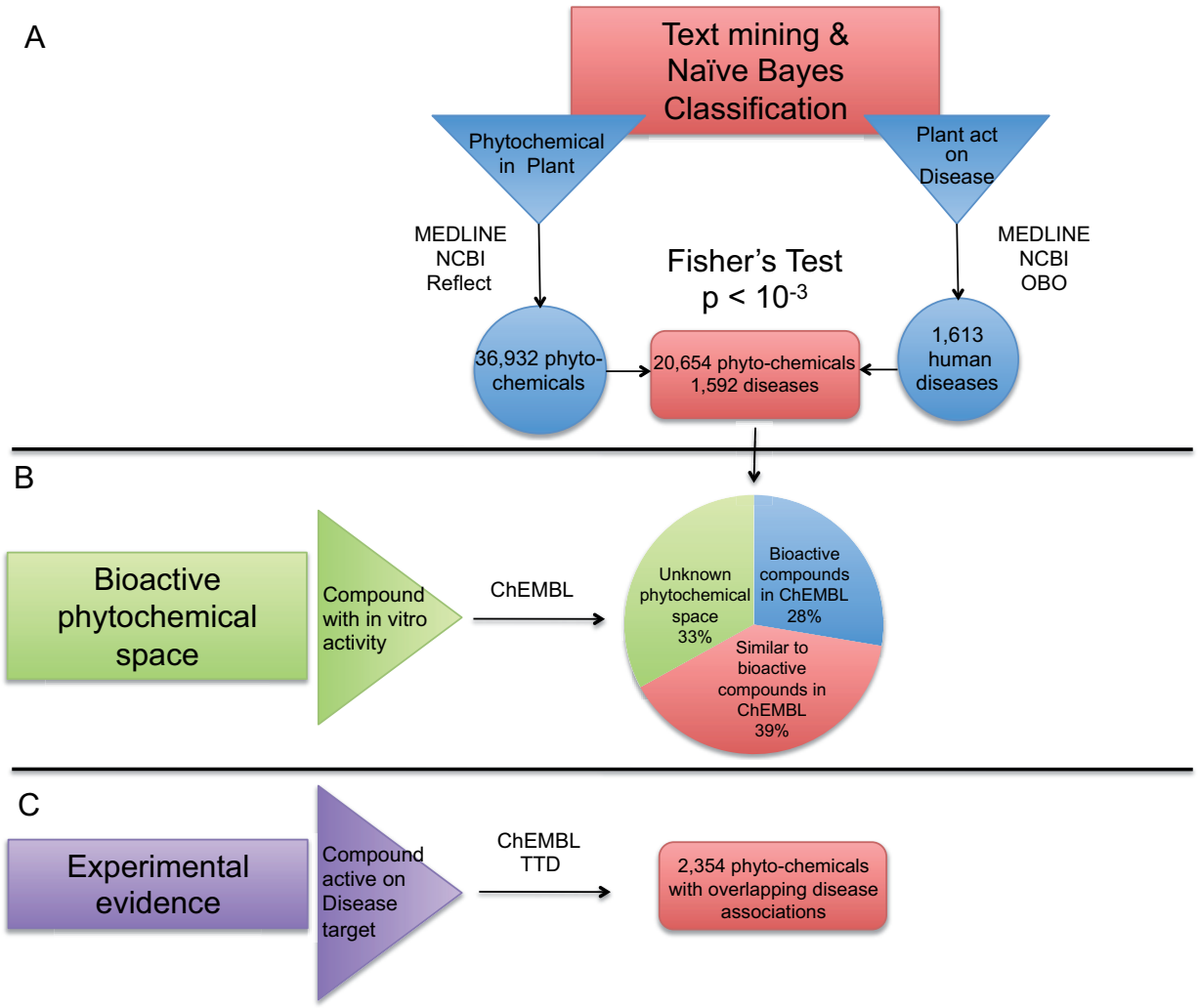


Figure 8: A) In the phytochemical - food and food - disease relations that we extracted by text mining, there are 7,077 plants with both phytochemical and human disease annotation. We used Fisher's exact test to identify statistically significant correlations between phytochemical and human disease phenotypes. At a 5% false discovery rate we identified 20,654 phytochemicals associated to 1,592 human disease phenotypes. B) 5,709 of the text-mined phytochemicals have been tested experimentally on a biological target and the activity data have been deposited in ChEMBL. For the remaining two thirds of the compounds, 8,113 phytochemicals are structurally similar to compounds with known protein targets (estimated with a Tanimoto coefficient > 0.85), indicating similar bioactivity. The rest of the compounds, 6,832 phytochemicals, are not similar to any known bioactive compound and belong to a hitherto unexplored phytochemical space. C) We used the Therapeutic Targets Database to annotate the protein targets from ChEMBL to diseases. From the 5,709 phytochemicals that are included in ChEMBL, 2,354 are active against a biological target that is relevant for the same disease as the one we have predicted.

Table 2: Phytochemicals are associated with diseases via the approach illustrated in Figure 8. For exemplary cancer types, we list the number of phytochemicals that are similar to small compound drugs that are approved for treatment of the disease (column 3), the number of phytochemicals that have experimental activity against a target implicated in this cancer type (column 4) and the corresponding number of common foods that contain these phytochemicals (column 5).

Cancer type (DOID)	# drugs[1]	# associated phytochemicals similar to a drug	# associated phytochemicals with experimental disease-related target[2]	# common foods with disease-associated phytochemicals[3]
Breast cancer (1612)	44	344	1,840	94 (120)
Leukemia (162)	36	302	1,067	95 (118)
Lung cancer (1324)	23	103	1,070	83 (119)
Prostate cancer (10283)	20	170	2,105	82 (120)
Lymphoma (0060058)	20	146	527	80 (115)
Urinary system carcinoma (3996)	11	28	1,623	58 (121)
Ovarian cancer (2394)	11	15	1,219	49 (117)
Sarcoma (1115)	8	38	45	26 (82)
Intestinal cancer (10155)	8	52	1,530	86 (120)
Testicular cancer (2998)	7	41	0	51 (0)
Kidney cancer (263)	6	12	1,605	39 (121)
Melanoma (1909)	5	4	275	11 (114)
Renal cell carcinoma (4450)	5	8	1,271	30 (120)
Pancreatic cancer (1793)	4	28	1,331	53 (119)
Liver cancer (3571)	4	24	781	53 (118)
Skin carcinoma (3451)	2	8	11	16 (58)
Adenocarcinoma (299)	2	28	7	44 (19)
Endometrial cancer (1380)	2	97	20	58 (88)

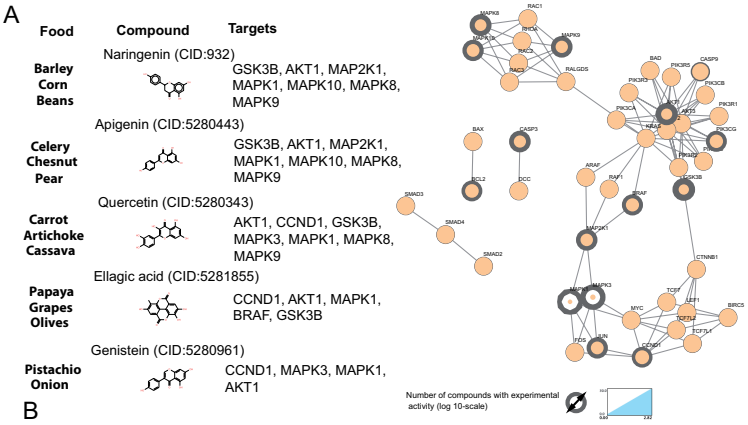
DOID: Human Disease Ontology Identifier, [1] from DRUGBANK, [2] from ChEMBL and TTD, [3] similar to a drug (with exp. disease-related target)

Case study on colon cancer

To demonstrate the full potential of our approach we selected colon (colorectal) cancer as a case study and analyzed our results in the three directions shown below. Colon cancer is the second largest cause of cancer-related deaths in western countries and various diet intervention and epidemiological studies suggest that diet is a vital tool for both prevention and treatment of the disease (Ferguson and Schlothauer, 2012; Terry et al., 2001).

(1) *One stop legacy knowledge-shop:* When one embarks into studying the effect of food on colon cancer, it is useful first to get a systems view of the existing knowledge. This includes information about what types of foods and phytochemicals have already been tested in relation to colon cancer, which are their biological targets and how these activities affect the biological networks that consist the disease pathway. Such a systems view of the influence of dietary molecules associated to colon cancer is sketched in Figure 9A, based on the knowledge derived from our text mining approach that has been projected on the colon cancer pathway from the KEGG PATHWAY Database (http://www.genome.jp/kegg-bin/show_pathway?hsadd05210, Accessed May 31 2014). By surveying our data resource we found 519 plants associated with a health benefit towards colon cancer. Statistical analysis of the data for frequently occurring phytochemical - disease pairs, reveals significant associations between 6,418 phytochemicals and colon cancer. Among the molecules associated with a health benefit for colon cancer, 623 of them have experimentally verified activity against proteins involved in the colon cancer pathway (nodes with a grey ring in Figure 9A). Naringenin, apigenin, quercetin, ellagic acid and genistein are examples of such compounds. Naringenin is commonly found in barley, beans and corn and apigenin is found in chestnuts, celery and pear. These foods have been associated with colon cancer prevention in a number of studies (Chavez-Santoscoy et al., 2009; Frédérich et al., 2009; Madhujith and Shahidi, 2007). When tested *in vivo,* both compounds have been found able to suppress colon carcinogenesis (Leonardi et al., 2010). In addition, in *in vitro* experiments naringenin and apigenin have seven targets on the KEGG colon cancer pathway. Quercetin, found in artichoke, carrot and cassava, and ellagic acid, present in grapes, papaya and olives have seven and five targets, respectively, on the KEGG colon cancer disease pathway, while genistein, found in pistachio-nuts and onions, has four. In most, if not all, of these cases, interest on the biological activity of the phytochemicals emerged after observations that the foods that contain them have some health benefit in relation to colon cancer prevention and treatment (Dinicola et al., 2012; Al-Fayez et al., 2006; Juan et al., 2006).

A



B

Target name	Uniprot ID	Compounds identical to ChEMBL compounds	Compounds similar to ChEMBL compounds
Thymidylate Synthase	P04818	16	84
Vascular Endothelial Growth Factor	P15692	0	20
DNA Topoisomerase I	P11387	24	88
Epidermal Growth Factor Receptor	P00533	383	527

C

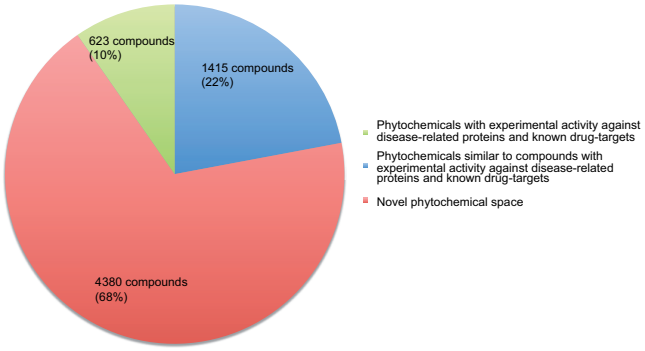


Figure 9: A) The KEGG colon cancer disease pathway map is illustrated on the right, where the number of phytochemicals with experimentally measured bioactivity data is depicted as grey ring of varying width. Examples of bioactive phytochemicals are listed on the left, along with typical food source and biological target. B) Protein targets of typical colon cancer drugs and number of phytochemicals with experimental and predicted activities against them. C) From the 6,418 molecules associated with a health benefit for colon cancer, 623 have measured experimental activity against proteins from the colon cancer pathway or targets of colon cancer drugs. On the remaining phytochemical space linked to colon cancer, we can use chemoinformatics to predict activity based on compound structure and select the most promising candidates for in vitro or in vivo experimental validation. Accordingly, we have identified 1,415 phytochemicals with potential activity against colon cancer. For reasons of consistency with the disease pathway map, protein targets are given with their corresponding gene names.

Typical drugs in the market against colon cancer are listed in Figure 9B, along with their main protein targets. By surveying our data resource we identified a number of phytochemicals that have measured experimental activity against the same proteins. Riboflavin monophosphate, for example, which is found in many common foods such as almond, broccoli and tomato, is one among the 16 phytochemicals we have identified with biological activity against thymidylate synthase, the main target of drugs 5-fluorouracil and capecitabine (Martucci et al., 2009). Similarly, reserpine, a natural compound that has found applications as antihypertensive and antipsychotic, exhibits activity against DNA topoisomerase I (Itoh et al., 2005) - the target of the colon cancer drug irinotecan - which could be interesting to investigate further in the light of drug repurposing.

(2) *Discovery of novel bioactive compounds with drug-like properties:* As we saw above, from the 6,418 molecules associated with a health benefit for colon cancer, only 623 have experimentally verified activity against colon cancer protein targets (Figure 9C). On the remaining phytochemical space linked to colon cancer, we can use chemoinformatics approaches to predict activity based on compound structure and select the most promising candidates for *in vitro* testing. By encoding the structure in 2D fingerprints and setting a Tanimoto coefficient of 0.85 as the similarity threshold, 1,415 molecules turn up as structurally similar to a phytochemical or a synthetic compound from ChEMBL with activity against a protein from the colon cancer pathway or a colon cancer drug target (Figure 9B). The compounds listed in Table 3 are such examples, for which we can infer their bioactivity from experiments performed on structurally similar compounds.

In regards to the remaining phytochemicals that our approach has associated to colon cancer, for which there exists no experimental protein target information and are not structurally similar with molecules that interact with colon cancer proteins, more advanced chemoinformatics techniques could be applied, such as pharmacophore-based similarity and docking. Alternatively, *in vivo* assays in model animals or *in vitro* experiments on disease cell lines could assist in elucidating their bioactivity. Such compounds with strong statistical support are beta-caryophyllene (Ali et al., 2012), guaiacol (Formisano et al., 2011) and alloisoleucine (Sánchez-Hernández et al., 2012) (p-value < 10^{-23}). Guaiacol, for example, has been identified in 93 plants in total, 32 of which are associated in the literature with colon cancer.

(3) Discovery of novel health benefits from foods: One of the key observations from our analysis is that the majority of phytochemicals is found in a variety of foods, even in foods that are distant taxonomically. Thus, information about the bioactive phytochemical content of one food that has been characterized as beneficial towards colon cancer could help us identify other foods, which contain the same bioactive phytochemicals that may have similar health benefits.

For example, cauliflower has been associated with a preventive effect on colon cancer (Mas et al., 2007; Temple and el-Khatib, 1987). The adzuki bean shares 800 phytochemicals with it and could potentially have a similar effect on colon cancer as well; there exists, however, no such evidence in the literature. Such comparisons of phytochemical profiles could also find applications in the design of nutrigenomics studies, with the purpose to confirm that the study group follows a reference diet as different as possible from that of the control group, i.e. the two diets do not contain foods with similar phytochemical profiles.

Table 3: Phytochemicals (column 1) from common foods (column 2) with inferred activity to a colon cancer protein (column 3), based on structural similarity with an active compound from the ChEMBL library (column 4). Listed compounds are examples of compounds predicted by our approach to have a positive effect against colon cancer, where p-values are included in column 5.

Compound name	# common foods	Predicted colon cancer target	Similar bioactive compound (Tc)	p-value for colon cancer
Vanillin	18	CCND1	CHEMBL53781 (0.86)	10^{-23}
Folic acid	19	MAPK1, MAPK3, ERBB2	CHEMBL1679 (0.85)	10^{-23}
Spermidine	16	CASP3	CHEMBL23194 (1.00)	10^{-12}
Vanillic acid	27	MAPK1, MAPK3, ERBB2	CHEMBL32749 (0.88)	10^{-12}
Chalconaringenin	1	JUN	CHEMBL129795 (0.86)	10^{-11}
Protocatechuic acid	30	EGFR	CHEMBL145 (0.86)	10^{-7}
Quercetin-3-glucoside	34	EGFR	CHEMBL486625 (0.85)	10^{-4}
Folinic acid	1	TYMS	CHEMBL439741 (0.88)	10^{-11}
Protopanaxatriol	4	TOP1	CHEMBL1096728 (0.85)	10^{-3}

Discussion

Food is a complex system that has an equally complex pattern of interactions with the human organism. As such, it consists the ideal platform for applying a systems biology approach, where different heterogeneous data sources are integrated and analyzed in a holistic way. Ferguson and Schlothauer in a review article that was published in 2012 (Ferguson and Schlothauer, 2012) illustrated how information on the beneficial effect of broccoli against cancer is enriched by the integration of genomics, proteomics and metabolomics data. For a well-studied food such as broccoli there is a rich body of evidence regarding its bioactive phytochemicals. Nevertheless, gathering and visualizing all evidence at once offered novel insights into the mechanisms by which broccoli may prevent cancer or retard cancer growth and progression.

An enormous scientific literature focusing on bioactive plant extracts and their phytochemicals, encompassing thousands of scientific papers, has emerged over the years. However, in order to utilize this wealth of information and integrate it with other types of data within systems biology studies, it is essential to first locate and then retrieve it in a high-throughput manner. The approach we have demonstrated here, which relies on the text mining of abstracts in PubMed/MEDLINE, has associated 23,137 phytochemicals with 15,722 plants, including approximately 2,768 edible fruits, vegetables and plant-based beverages. Even though there are several ongoing efforts that aim to collect information on molecular composition of food in a single resource, i.e. the Danish Food Composition Database (http://www.foodcomp.dk) centered on well-known organic nutrients, such as vitamins, amino acids, carbohydrates and fatty acids; the Phenol-Explorer (Neveu et al., 2010) with information in text format for 500 polyphenols in over 400 foods and the KNApSAcK Family Database (Afendi et al., 2012), these are rather limited in focus and size. For a molecular systems chemical biology approach of diet, the lack of chemical structures in the above databases is another significant bottleneck, as linking chemical names to a chemical structure in a high-throughput manner is not yet a straightforward process (Williams et al., 2012). The most important contribution of our study is that it uses all the evidence generated during the last 100 years supporting health benefits of vegetables, fruits and other plants for establishing associations between foods, phytochemicals and human diseases, where entities from all three classes are annotated with unique, standard identifiers, so that they can be traceable in other databases. Moreover, chemical names and synonyms of all phytochemicals are linked to a unique chemical structure, which, besides traceability in other resources, allows for the application of chemoinformatics tools and their integration in systems chemical biology analyses. Last but not least, food associations to disease are annotated with directionality, which differentiates between causative and preventive effects of the food in relation

to the specific disease. Nevertheless, and despite the enormous amount of information collected here, we should also point out that inherent bias of meta-analysis allows for further improvements in our text mining pipeline. For example, while PubMed/MEDLINE is the most appropriate database for associating dietary interventions with disease phenotypes, it is certainly lacking scientific journals focused on the chemical composition of plants (for example, the Springer journal of Metabolomics; www.springer.com/lifesciences/biochemistry&biophysics/journal/11306), Accessed May 31 2014). In order to overcome other common pitfalls of meta-analysis, such as data quality and data independence, it is our intention in the future to investigate the use of weighting parameters on the retrieved associations, so that, for example, associations generated from different labs constitute stronger evidence than associations from the same research team.

As we show in the case study on colon cancer, associating food, phytochemical content and diseases can build the basis for discovering novel bioactive compounds with drug-like properties. Furthermore, our analysis brought to the surface an undiscovered dietary component space of 8,113 phytochemicals that has not been previously linked to a health benefit and bears no structural similarities to other bioactive phytochemicals with established molecular targets. This represents a forthright opportunity for biochemists and nutritionists and offers a good basis for an attractive drug discovery platform. At the same time, food safety authorities are concerned about the presence of compounds in herbal products and dietary supplements that could exert toxicity to humans (Singh et al., 2012). For example, myristicin, a known component of nutmeg (Demetriades et al., 2005) and glycoalkaloids that are present in potatoes (Mensinga et al., 2005) can be extremely dangerous when taken in large doses. It is thus of great value to have *in silico* tools that are able to quickly list all phytochemicals associated to a given food in the public literature, and subsequently interrogate databases (e.g. the Comparative Toxicogenomics Database, http://ctdbase.org) for experimental evidence that associates the compounds in question or structurally similar compounds with a toxic effect. Similar to research in the field of nutrition, scientists in ethnomedicine are seeking for evidence that can explain at the molecular level the health effect of traditional medicine. Ethnomedicine, such as Traditional Chinese Medicine and Ayurveda has existed and supported human health for thousands of years. A major barrier for developing an ethnomedicine evidence-based knowledgebase is that the current information related to plant substances for medicinal purposes is scattered and unstructured (Sharma and Sarkar, 2013).

We provide a solution to this problem by extracting in a structured and standardized format phytochemicals that are associated with a medicinal plant, either in the open literature of the last 100 years or in the ethnomedicinal databases that we have *in-house*. Our approach facilitates the identification of novel bioactive compounds from natural sources and the repurposing of medicinal plants to other diseases than the ones traditionally used for, and builds a step towards elucidating their mechanism of action.

Conclusion

Food is a factor that exerts influence on human health on a daily basis. Modulating the expression and the activity of enzymes, transcription factors, hormones and nuclear receptors is how food and its bioactive constituents modulate metabolic and signaling processes. The aim of our study is to provide the molecular basis of the effect of food on health in the complete spectrum of human diseases and to suggest why and how diet and dietary molecules may represent a valuable tool to reinforce the effect of therapies and protect from relapse. Our systematized approach for connecting foods and their molecular components to diseases makes possible similar analyses as the one illustrated for colon cancer for approximately 2,300 disease phenotypes. In addition, it provides the phytochemical layer of information for nutritional systems biology studies with the aim to assess the systemic impact of food on health and make personalized nutritional recommendations.

Methods

Mining the literature for plant - phytochemical pairs

We retrieved the names of land plant species (embryophyta) and their synonyms from NCBI (http://www.ncbi.nlm.nih.gov/taxonomy). Chemical compound names and synonyms were taken from the argument browser Reflect (Pafilis et al., 2009). With these two dictionaries the mining of 21 million titles and abstracts of PubMed/MEDLINE (http://www.nlm.nih.gov) was carried out using ChemTagger (https://pypi.python.org/pypi/ChemTagger). A Naive Bayes Classifier (https://pypi.python.org/pypi/NaiveBayes) was trained to recognize pairs of plants and phytochemicals.

A set of 200 tags, – plant and compound name entities – from 200 abstracts was compiled for training. As positive training set (PTS) we manually compiled a set of 75 abstracts mentioning plants and their phytochemical content. As negative training set (NTS) we manually compiled a set of 125 abstracts mentioning plants and chemical compounds, which we judged that did not

refer to an actual plant - phytochemical content relationship. This includes, for example, abstracts that associate plants with synthetic small compounds in the context of chemical extraction and purification of plant extracts (e.g. ecdysonoic acid, 3-acetylecdysone 2-phosphate (Isaac and Rees, 1984). A feature vector was compiled consisting of words within the abstract that were in proximity of each nametag.

The lexical features were chosen based on the term frequency–inverse document frequency (Wu et al., 2008) and were sorted with the most frequent feature on the top and the least frequent at the bottom of the list. The training of the classifier commenced with only the highest score feature, while features with the next higher scores were added one by one, until the accuracy of the classifier stabilized at 31 features. Words such as "compound", "isolated", "extract" and "concentrated" were the features with the highest tf-idf score. Training was carried out using leave-one-out cross validation on the shuffled training data set. The performance of the classifier was subsequently evaluated on an external, balanced test set of 250 positive and negative abstracts, and resulted to 88.4% accuracy and 87.5% F1-measure. When the classifier was applied to the raw text of PubMed/MEDLINE, it retrieved 23,137 phytochemicals from 15,722 land-plant species (embryophyta) associated through 369,549 edges.

Chemical structures of the text-mined phytochemicals were retrieved from PubChem (Bolton et al.), ChEBI (Hastings et al., 2013), CHEMLIST (Hettne et al., 2009), the Chinese Natural Product Database (Shen et al., 2003) (CNPD) and the Ayurveda (Polur et al., 2011) that we have previously curated in-house. Canonical SMILES were calculated with OpenBabel (http://openbabel.org/wiki/Canonical_SMILES) with no salts, isotopic or chiral center information. Edible plant names were retrieved from Plant For A Future (PFAF) (http://www.pfaf.org) and were mapped to NCBI IDs.

The taxonomy of plant species was retrieved from NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy). Overrepresentation of phytochemicals on the taxonomy was calculated by using Fishers exact test, following the Benjamini-Hochberg procedure with a 5% false discovery rate (Yoav and Hochberg, 1995). A phytochemical that is significantly overrepresented on a specific class, order, family or genus of the taxonomy denotes that it is not randomly distributed over the whole tree. Since the association of plants to their phytochemicals was performed on the genus level, for this analysis we projected the phytochemical content of a child node to the parent node.

Mining the literature for plant - disease associations

The names of land plant species (embryophyta) and their synonyms were taken from NCBI (http://www.ncbi.nlm.nih.gov/taxonomy). We retrieved 70,005 human disease terms and synonyms from the Open Biological and Biomedical Ontologies (OBO) Foundry (Smith et al., 2007). The list of 143 common, non-processed foods was retrieved from the Danish Food Composition Database (http://www.foodcomp.dk). Names were mapped to NCBI land plant species and whenever *var.* IDs were available, they were subsequently collapsed to the corresponding species ID (e.g. broccoli and kale are varieties of the same *Brassica olerasea* species).

With these two dictionaries, text mining of 21 million titles and abstracts of PubMed/ MEDLINE (http://www.nlm.nih.gov) was carried out using ChemTagger (https://pypi.python.org/pypi/ChemTagger). A Naive Bayes Classifier (https://pypi.python.org/pypi/NaiveBayes) was trained to recognize pairs of plants and the associated human disease phenotypes. A set of 2,074 nametags, plants and human disease phenotype name entities from 333 abstracts was compiled for training. Plants and human diseases with a 'preventive' association were used as the positive training set (PTS) and plants and human diseases with a 'promoting' association as the negative training set (NTS). Name entities of plants and human diseases mentioned in other contextual associations were used as the 'noise' training set (OTS).

For the training of the Naive Bayes Classifier, the lexical features were chosen based (Wu et al., 2008) and were sorted with the most frequent feature on the top and the least frequent at the bottom of the list. The training of the classifier commenced with only the highest score feature, while features with the next higher scores were added one by one, until the accuracy of the classifier stabilized at 71 features. Words such as "treatment", "effect", "patient", "disease" and "plant" were the features with the highest tf-idf score.

Training was carried out set using leave-one-out cross validation on the shuffled training data set. The performance of the classifier was subsequently evaluated on an external, balanced test set of 250 positive and negative abstracts, and resulted to 84.5% and an F1-measure of 84.4%. When the classifier was applied to the raw text of PubMed/MEDLINE, it retrieved 7,178 land-plant species associated with 1,613 human disease phenotypes through 38,090 edges. Plant - disease networks were constructed in Cytoscape v.2.8.1.

Molecular level association of plant consumption to human disease phenotypes

We performed a categorical Fisher's exact test with the Benjamini-Hochberg procedure and a 5% false discovery rate (Yoav and Hochberg, 1995) to associate particular phytochemicals with human disease phenotypes. Our alternative hypothesis was that the proportion of plants associated with a particular phytochemical is higher among the plants with a specific human disease phenotype than among those without. Our null hypothesis was that there is no relationship between plants associated with a particular phytochemical and a specific human disease phenotype. Phytochemicals were associated to protein targets though experimental chemical-protein association data from ChEMBL, version 15 (Overington, 2009). Canonical SMILES with no salts, isotopic or chiral center information (http://openbabel.org/wiki/Canonical_SMILES) were used as the unique molecular identifier for searching for common small compound entities between the phytochemical and ChEMBL lists. Human proteins were associated to diseases through the Therapeutic Targets Database (Zhu et al., 2012) (TTD Version 4.3.02). Disease names were mapped to the OBO Foundry human disease ontology and ordered in disease categories. Disease pathway networks were constructed in Cytoscape v.2.8.1.

Case study on colon cancer

The colon cancer disease pathway was obtained from KEGG PATHWAY Database (http://www.genome.jp/kegg-bin/show_pathway?hsadd05210). The network was constructed in Cytoscape v.2.8.1. Phytochemicals were associated to the proteins from the disease pathway though experimental chemical-protein association data from ChEMBL, version 15 (Overington, 2009). Canonical SMILES with no salts, isotopic or chiral center information (http://openbabel.org/wiki/Canonical_SMILES) were used as the unique molecular identifier for searching for common small compound entities between the phytochemical and ChEMBL lists. Colon cancer drugs were obtained from KEGG Disease Entry: H00020 (http://www.genome.jp/dbget-bin/www_bget?ds:H00020) and their respective protein targets from the Therapeutic Targets (Zhu et al., 2012) (TTD Version 4.3.02).

Supplementary material

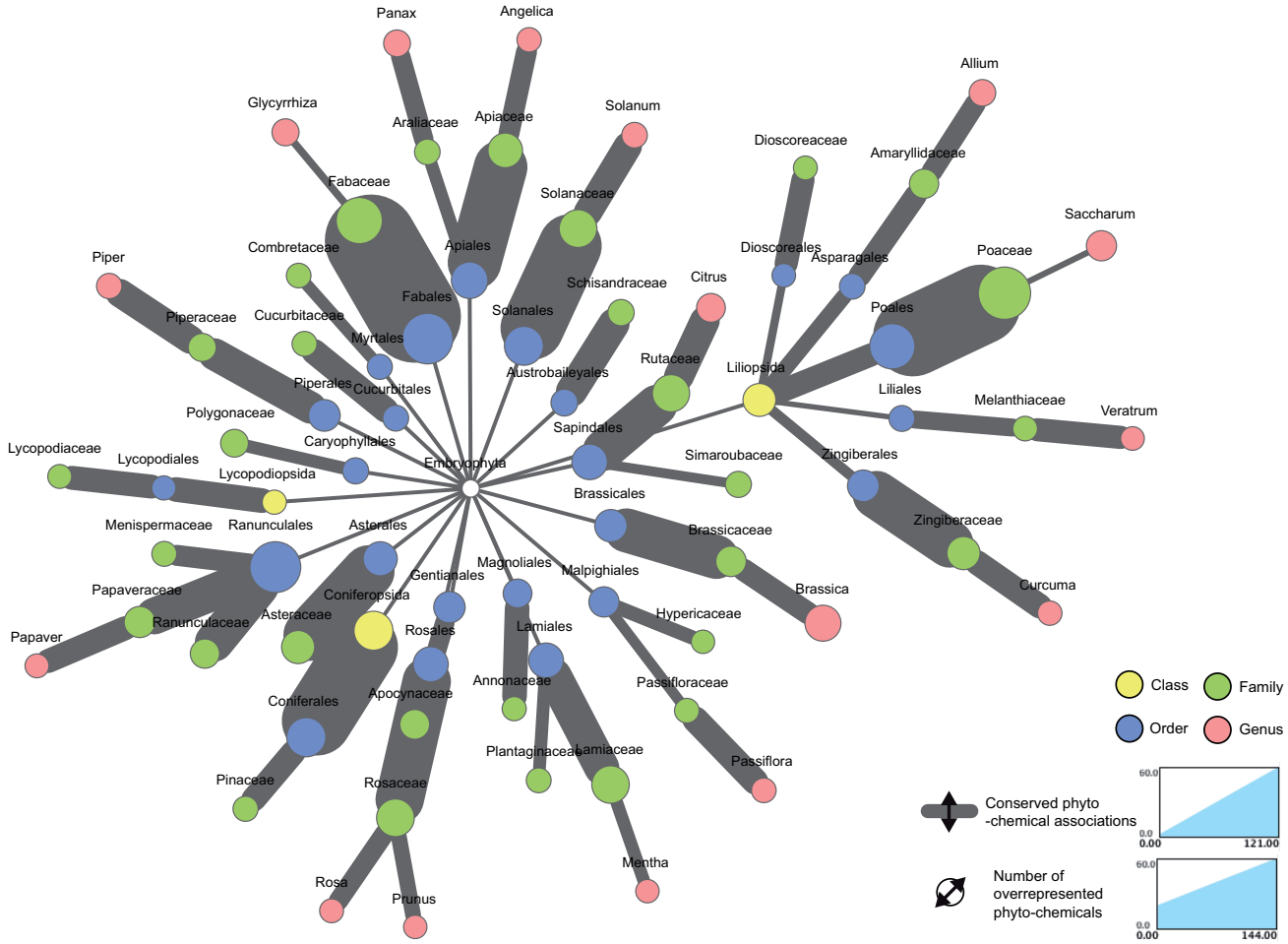


Figure S1: Mapping the phytochemical space on the plant taxonomy. 37,351 phytochemicals were mapped on the plant taxonomy. Only 8% of the recorded phytochemicals show localized enrichment (p-value < 10-4). The taxonomy of land-plant species (embryophyta) was retrieved from NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy). Nodes represent Classes (yellow), Orders (blue), Families (green) and Genera (pink) of the taxonomy tree. Links are placed between a parent and a child node, if they share conserved phytochemicals. A phytochemical is conserved, when it is overrepresented on both the parent and the child nodes. The width of the link corresponds to the number of conserved phytochemicals between parent and child nodes. The size of the node corresponds to the number of overrepresented phytochemicals on a given class, order, family or genus.

Table S1: List of phytochemicals described as SMILES that are localized on a taxonomy class, order, family or genus.

Ref: http://www.ploscompbiol.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pcbi.1003432.s002, Accessed May 14 2014.

Chapter II: NutriChem: a systems chemical biology resource to explore the medicinal value of diet

NutriChem: a systems chemical biology resource to explore the medicinal value of diet

Kasper Jensen[1], Gianni Panagiotou[2], Irene Kouskoumvekaki[1,*]

[1] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark

[2] School of Biological Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong

* To whom correspondence should be addressed. Tel: +45 4525 6162; Fax: +45 4593 15 85; Email: irene@cbs.dtu.dk

Abstract

There is rising evidence of an inverse association between chronic diseases and diets characterized by rich fruit and vegetable consumption. Dietary components may act directly or indirectly on the human genome and modulate multiple processes involved in disease risk and disease progression. However, there is currently no exhaustive resource on the health benefits associated to specific dietary interventions and there is certainly no resource covering the broad molecular content of our food. Here we present the first release of NutriChem, a database generated by text mining of 21 million MEDLINE abstracts for collecting all available information that link plant-based foods with their small molecule components and human disease phenotypes. NutriChem contains text-mined data for 18,478 pairs of 1,772 plant-based foods and 7,898 phytochemicals, and 6,242 pairs of 1,066 plant-based foods and 751 diseases. In addition, it includes predicted associations for 548 phytochemicals and 252 diseases. To the best of our knowledge this database is the only resource linking the chemical space of plant-based foods with human disease phenotypes and provides a fundamental foundation for understanding mechanistically the consequences of eating behaviors on health. NutriChem is available at http://cbs.dtu.dk/services/NutriChem-1.0

Introduction

Although both genetic and environmental factors contribute to the risks of developing chronic diseases, differences in environment are probably responsible for 70-90% of disease risks (Rakyan et al., 2011; Rappaport and Smith, 2010; Vineis et al., 2009). The numerous direct and indirect effects that environmental exposures display induce both immediate and long-term health

responses. Exposure in high doses of chemicals or vulnerability of individuals to particular chemicals, can lead to immediate health effects. On the other hand, repeated and prolonged exposures, like dietary habits, contribute to more general pathophysiological mechanisms and increase the potential health effects by influencing disease development over a lifetime. The term ´´exposome´´, which is used to describe the totality of all environmental exposures (e.g. diet, air pollutants, lifestyle factors) over the life course of an individual, has been proposed to be a critical entity for disease etiology and to complement the genetic information (Brook et al., 2010; Heinrich, 2011; Wild, 2011). In order to avoid biased inferences regarding gene-environment interactions and to discover the major causes of chronic diseases a more comprehensive and quantitative view of the exposome is required (Paoloni-Giacobino, 2011; Rappaport, 2012). Since a full characterization of the human exposome is a daunting task, cutting the pie to smaller pieces could offer critical portions of disease associations of certain exposures.

Diet is certainly one of the most dynamic expressions of the exposome and one of the most challenging to assess its effects in health homeostasis and disease development, as a consequence of its myriad components and their temporal variation. Recognize, understand and interpret the interplay between diet and biological responses to this exposure may contribute to the weight of evidence in assigning causality to a diet-disease association. Therefore, in order to open up new avenues to disease prevention through diet interventions is crucial to provide insights into the mechanisms by which an exposure to the chemical space of a food might be exerting its effects. Towards this direction we have developed a state-of-the art database with information on plant-based food (referred to simply as “food” throughout the article), its small compound constituents (also known as phytochemicals) and human disease phenotypes associated with it. This database offers a unique platform for exploring the medicinal value of diet and elucidating the synergistic effects of natural bioactive compounds on disease phenotypes.

Implementation

Data ontologies

The taxonomy of the plant species was retrieved from NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy). Food names were retrieved from the Plant For A Future (PFAF) (http://www.pfaf.org) and the Food composition database (http://www.foodcomp.dk) and were mapped to NCBI IDs (TAXIDs). Human proteins were associated to diseases through the Therapeutic Targets Database (Zhu et al., 2012) (TTD Version 4.3.02). Disease names were mapped to the OBO Foundry Human Disease Ontology (DOID) (Smith et al., 2007) and ordered in disease categories.

Chemical structures and corresponding IDs of small compounds were retrieved from PubChem (Bolton et al.) ChEBI (Hastings et al., 2013), CHEMLIST (Hettne et al., 2009), the Chinese Natural Product Database (Shen et al., 2003) (CNPD) and Ayurveda (Polur et al., 2011) ressources. Marvin 6.1.0 (http://www.chemaxon.com) was used for encoding the chemical structures in unique SMILES.

Food-compound and food-disease associations

We extracted by text-mining plant – phytochemical and plant - disease associations from 21 million abstracts in PubMed/MEDLINE, covering the period 1908-2012, as described previously (Jensen et al., 2014). We subsequently filtered for pairs involving plant-based foods using Plant For A Future (PFAF) (http://www.pfaf.org) and the food composition database (http://www.foodcomp.dk). In total, NutriChem contains 18,478 pairs of 1,772 plant-based foods and 7,828 phytochemicals, and 6,242 pairs of 1,066 plant-based foods and 751 diseases.

Association of diet to health benefit at molecular level

Fisher's exact test was used to systematically associate frequently occurring phytochemical-disease pairs through the phytochemical-food and food-disease relations extracted by text mining, with the Benjamini-Hochberg procedure and a 5% false discovery rate (the method is described in detail previously (Jensen et al., 2014)). Chemical-protein interactions data were gathered in September 2013 from the open-source database ChEMBL (version 16). We associated phytochemicals that have no direct experimental bioactivity data with structurally similar compounds from ChEMBL (Tanimoto coefficient > 0.85) and their protein targets, when such data were available. For the calculation of the Tanimoto coefficient, chemical structures were encoded in 166 MACCS keys (Durant et al., 2002) using OpenBabel (O'Boyle et al., 2011). In total, NutriChem contains 1,549 predicted associations between 548 phytochemicals and 252 diseases.

Visual interface

We implemented a visual interface in NutriChem to facilitate the visualization of the results using CytoscapeWeb (http://cytoscapeweb.cytoscape.org). At the left part of the screen a network is depicted, with the query input as the central node and the retrieved results connected to it through edges. The thickness of an edge indicates the number of references in support of the associations. By clicking on the icon "Apply layout" the user can apply the Force-directed layout on the network. At the right-hand side the results are shown as a list. The list items are

expandable upon click and detailed information about the association is shown. The list items are also expandable by clicking on the edges of the network at the left-hand side.

By clicking on the icon "Export network" the user can download the network in Cytoscape/xml format and continue working offline in Cytoscape or download the list results as a "tab-separated file" by clicking on the "Export table" icon. The user can search by: i) food name or TAXID, ii) human disease name or DOID and iii) compound (i.e. phytochemical) name, ID (ChEMBL ID, CID, etc.) or SMILES string. The different functionalities of NutriChem. The user can query NutriChem by food, disease or compound name/ID are shown in Figure 10.

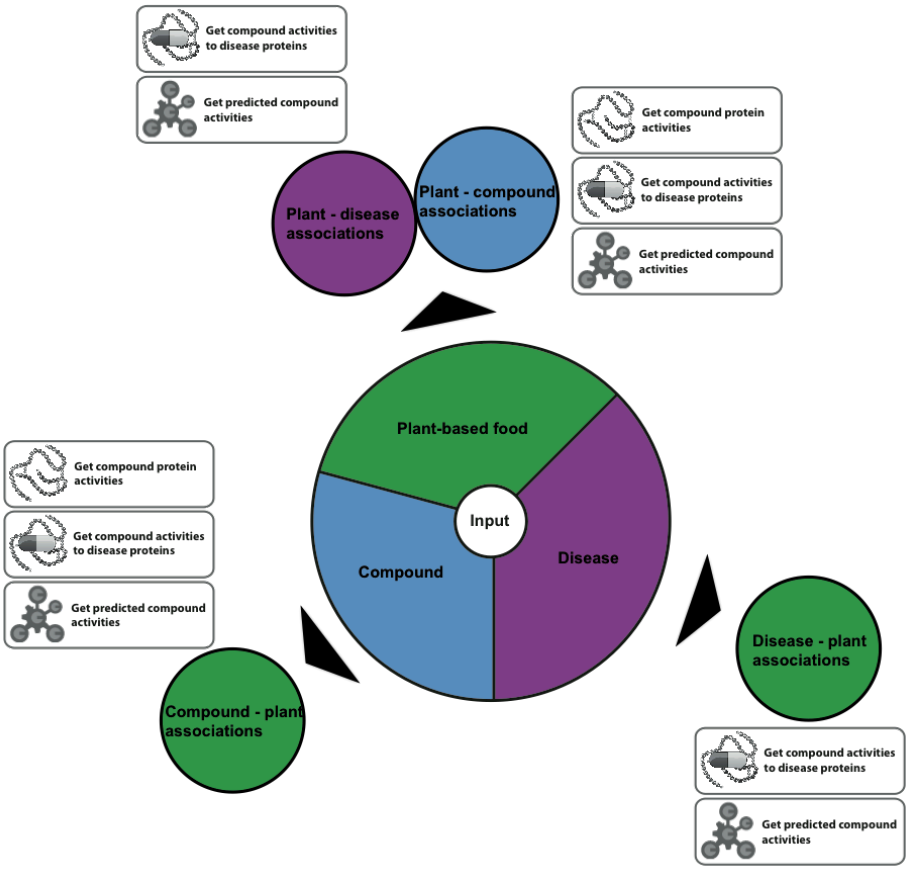


Figure 10: The different functionalities of NutriChem. The user can query NutriChem by food, disease or compound name/ID. Outcomes from each query type are in cycles pointed by arrows. The available actions that can be subsequently performed are depicted next to each cycle.

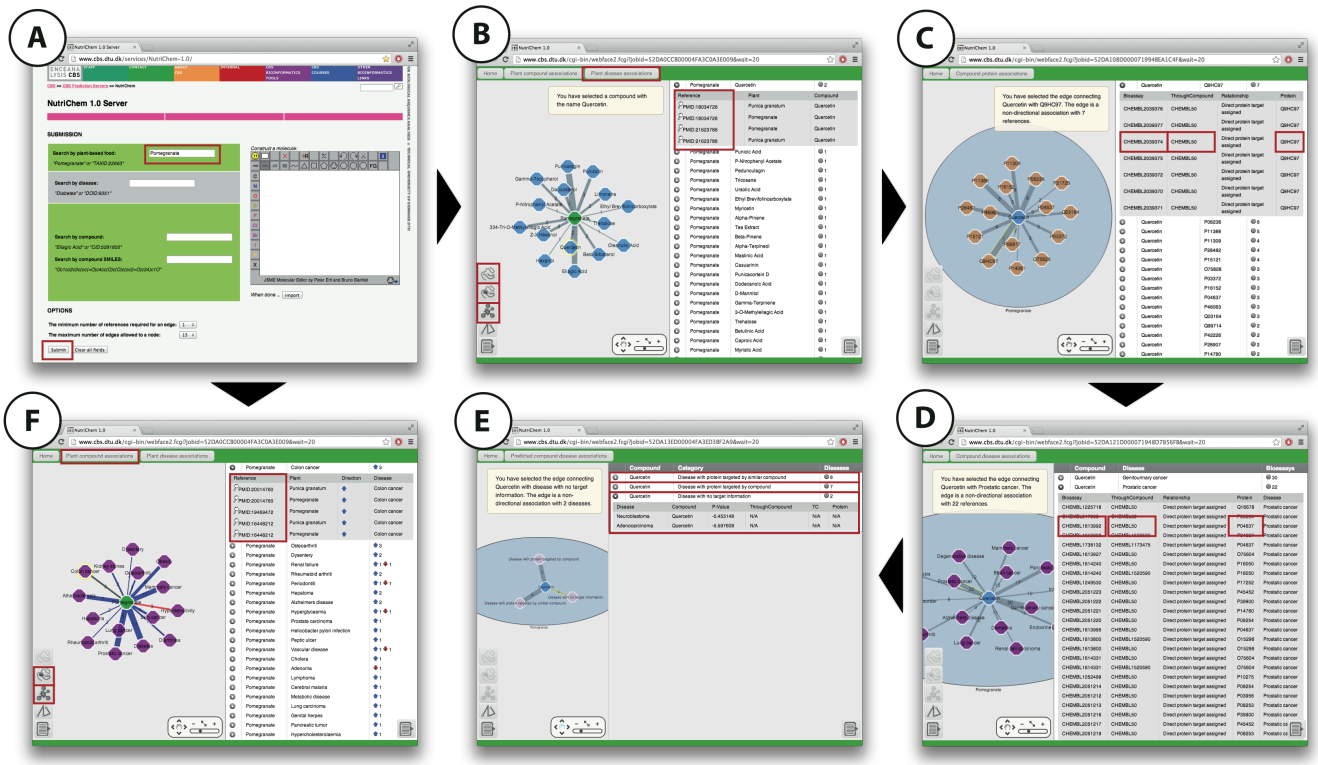


Figure 11: (A) A search with "pomegranate" as the query, (B) The plant-compound network is shown by default and by using the top buttons the user can switch to the plant-disease network or back to Home for submitting a new query. On the right-hand side the user can directly access the relevant references in PubMed in support of the pomegranate-compound associations, by clicking on the respective PMID. (C) If we select the compound "quercetin", "Get compound protein activities" lists 15 proteins with measured experimental activity. The user has the option to click either on the bioassay, the compound or the protein IDs on the right side of the panel, which will open in a new tab the respective ChEMBL and UniProt pages. (D) "Get compound activities to disease proteins" filters the above results for therapeutic proteins only (proteins annotated to a disease in TTD) and displays the disease-associated network for quercetin. (E) "Get predicted compound activities" shows predicted disease associations for quercetin as derived from Fisher's test. The results are grouped in three categories, depending on the experimental evidence that supports each prediction. (F) The disease network of the query "pomegranate" returns 15 diseases in which pomegranate is known to have a preventive/beneficial effect. Blue edge: "preventive" association, red edge: "promoting". On the right-hand side the user can directly access the relevant references in PubMed in support of the pomegranate-disease associations, by clicking on the respective PMID.

Applications

a) Food as query

When a food query is submitted to the server the user can specify the minimum number of references required for an edge and the maximum number of edges allowed to a query node. By default we use a limit of minimum 1 reference for an edge (can range from 1 to 10) and maximum 15 edges for a node (can range from 1 to 20). However, regardless of the settings, all results are listed at the right-hand side. Figure 11 illustrates a search with "pomegranate" as the query. In the top of the interface two buttons are shown. The buttons allow the user to switch between the different result sections. The plant-compound network is shown by default and by using the top buttons the user can switch to the plant-disease network or back to Home for submitting a new query. The network at the left-hand side shows 15 compound associations. On the right-hand side the user can directly access the relevant references in PubMed in support of the pomegranate-compound associations, by clicking on the respective PMID.

When a compound node is selected on the food-compound network, the user has three special action buttons. The first action button is "Get compound protein activities", the second "Get compound activities to disease proteins", and the third "Get predicted compound activities". For example, if we select the compound "quercetin", "Get compound protein activities" lists 15 proteins in UniProtID (http://www.uniprot.org) with measured experimental activity (data from ChEMBL). The user has the option to click either on the bioassay, the compound or the protein IDs on the right side of the panel, which will open in a new tab the respective ChEMBL and UniProt pages. "Get compound activities to disease proteins" filters the above results for therapeutic proteins only (proteins annotated to a disease in TTD) and displays the disease-associated network for quercetin. "Get predicted compound activities" shows predicted compound-disease associations as derived from Fisher's test. The results are grouped in three categories, depending on the experimental evidence that supports each prediction. The first category "Disease with protein targeted by compound" includes predicted disease associations, where there exist experimental activity data for quercetin and a disease-related protein target. The second category "Disease with protein targeted by similar compound" includes predicted disease associations, where there exist experimental activity data for a compound similar to quercetin (Tc > 0.85) and a disease-related protein. The user has the option to click on one of the results on the right side of the panel, to open in a new tab the respective ChEMBL compound activity data or the protein information in UniProt. The third category "Disease with no target information" includes predicted disease associations with no experimental activity data for quercetin or a structurally similar compound and a disease-related protein.

By clicking on the top right button, "Plant disease associations" the disease network of the query "pomegranate" is displayed, with 15 diseases in which pomegranate is known to have a preventive/beneficial effect. When the edge is blue, it indicates a "preventive" association (food associated with disease prevention or amelioration) and a red edge indicates a "promoting" association (food associated with disease progress). On the right-hand side the user can directly access the relevant references in PubMed in support of the pomegranate-disease associations, by clicking on the respective PMID.

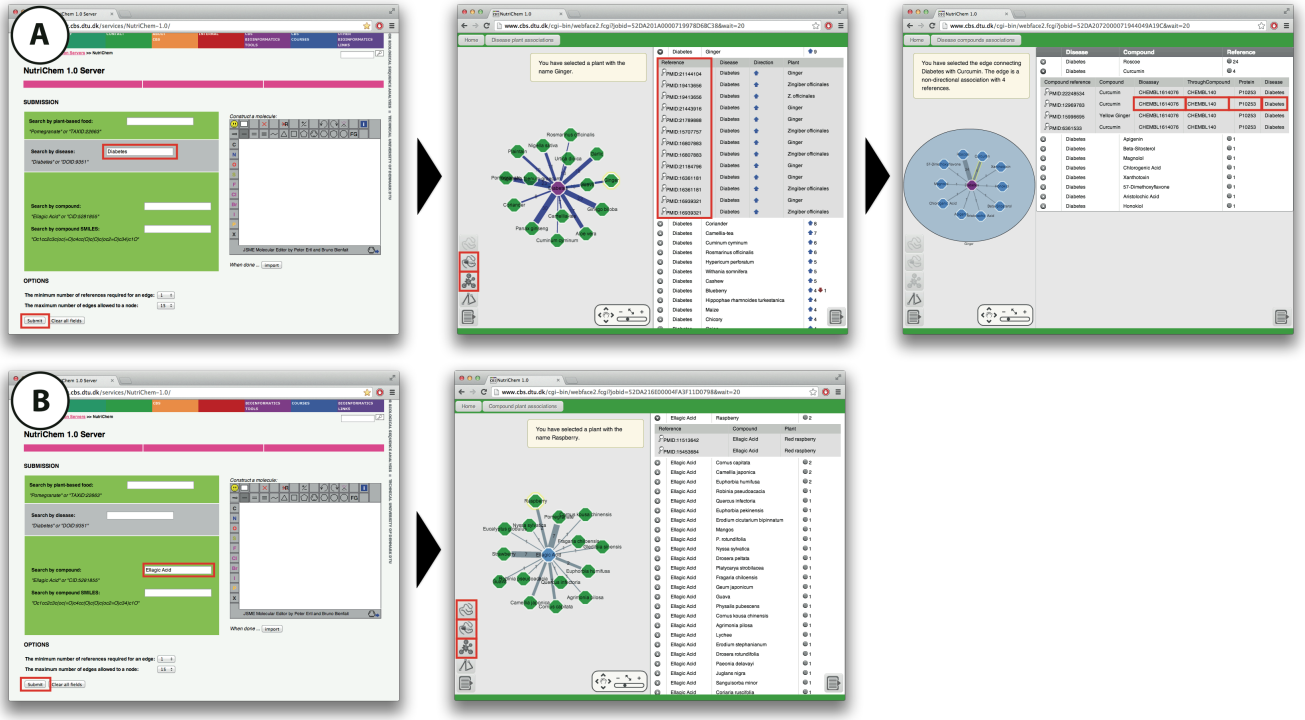


Figure 12: (A) Diabetes as query returns a network with 15 foods that have a preventive association against the disease. On the right-hand side the user can directly access the relevant references in PubMed in support of the diabetes-food associations, by clicking on the respective PMID. Clicking on any of food names on the network enables the action buttons "Get compound activities to disease proteins" and "Get predicted compound activities". On the right side of the panel the user can click on an arrow, which expands the results and displays the experimental activity data related to each compound. The user has again the option to click either on the bioassay, the compound or the protein IDs, which will open in a new tab the respective ChEMBL and UniProt pages. (B) Ellagic acid as query returns 15 foods that have been associated with it in the literature. On the right-hand side the user can directly access the relevant references in PubMed in support of the Ellagic acid-food associations, by clicking on the respective PMID. Clicking on any of the food names enables the three action buttons, as described above.

b) Disease as query

The user can query "diabetes" which returns a network of 15 foods that have a preventive association with diabetes (Figure 12A). On the right-hand side the user can directly access the relevant references in PubMed in support of the diabetes-food associations, by clicking on the respective PMID. Clicking on any of the food names on the network enables the action buttons "Get compound activities to disease proteins" and "Get predicted compound activities", as described above. Ginger, for example, has been associated in the literature with 10 compounds with experimental biological activity against a diabetes-related target. On the right side of the panel the user can click on an arrow, which expands the results and displays the experimental activity data related to each compound. For example, by expanding the results for Curcumin, we see that it targets P10253 (lysosomal alpha-glucosidase), a clinical trial target against diabetes mellitus type 2 according to TTD (Zhu et al., 2012). The user has again the option to click either on the bioassay, the compound or the protein IDs, which will open in a new tab the respective ChEMBL and UniProt pages.

c) Compound as query

The user can query "ellagic acid" which returns a network of 15 foods that have the compound associated (Figure 12B). Ellagic acid as query returns 15 foods that have been associated with it in the literature. On the right-hand side the user can directly access the relevant references in PubMed in support of the Ellagic acid-food associations, by clicking on the respective PMID. Clicking on any of the food names enables the three action buttons that allow the use to perform the steps described above.

Conclusion

The need for a more complete assessment of the environmental factors in epidemiological studies gave birth to a new –ome, the exposome. Here, we envisage elucidating the link between diet, molecular biological activity and diseases by developing a database source that translates the diet-exposome from concept to utility. Our methodology for better delineating the prevention of human diseases by nutritional interventions relies heavily on the availability of information related to the small molecule constituents of our diet. We expect to maintain and expand NutriChem regularly. As the next thing in the pipeline, we intend to integrate in NutriChem biological activity data of marketed drugs, which will make possible to study the effect of diet on drug properties related to their pharmacokinetics and pharmacodynamics.

Chapter III: Developing a molecular roadmap of drug-food interactions

Developing a Molecular Roadmap of Drug-Food Interactions

K. Jensen[1], B. Ni[2], G. Panagiotou[2,*], I. Kouskoumvekaki[1,*]

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark

2 Systems Biology & Biotechnology Group, School of Biological Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong

*** E-mail: Correspondence gipa@hku.hk, irene@cbs.dtu.dk**

Abstract

Recent research has demonstrated that consumption of food -especially fruits and vegetables- can alter the effects of drugs by interfering either with pharmacokinetic or pharmacodynamic processes. Despite the recognition of these drug-food associations as an important element for successful therapeutic interventions, a systematic approach for identifying, predicting and preventing potential interactions between food and marketed or novel drugs is not yet available. The overall objective of this work was to gain a global knowledge on the interference of ~ 4,000 dietary components present in ~1800 plant-based foods with the pharmacokinetics and pharmacodynamics processes of medicine, with the purpose of elucidating the molecular mechanisms involved. By developing a systems chemical biology approach that integrates data from the scientific literature, online databases and a recently developed *in-house* database we revealed novel associations between diet and dietary molecules with drug effect targets, metabolic enzymes, drug transporters and carriers currently deposited in DrugBank. Moreover, we identified disease areas, e.g. cancer and neurological diseases, as well as drug effect targets e.g. carbonic anhydrase family, kappa- and delta- type opioid receptors and 5-hydroxytryptamine receptors, that are most prone to the negative effects of drug-food interactions, showcasing a platform for making recommendations in relation to foods that should be avoided under certain medications. Lastly, by investigating the correlation of gene expression signatures of foods and drugs we were able to generate a completely novel drug-diet interactome map.

Introduction

Drugs and plant-based foods (i.e. fruits, vegetables and beverages derived from them, referred to simply as "food" throughout the rest of the document) share an intricate relationship in human health and have a complementary effect in disease prevention and therapy. In many diseases, such as hypertension, hyperlipidemia, and metabolic disorders, dietary interventions play a key part in the overall therapeutic strategy (Chan, 2013). But there are also cases, where food can have a negative impact on drug therapy and constitute a significant problem in clinical practice. Recent research has demonstrated that foods are capable of altering the effects of drugs by interfering either with pharmacokinetic or pharmacodynamic processes (Yamreudeewong et al., 1995). Pharmacokinetics includes the Absorption, Distribution, Metabolism and Excretion of drugs, commonly referred to jointly as ADME. Pharmacodynamic processes are related to the mechanisms of drug action, hence the therapeutic effect of drugs; interactions between food and drugs may inadvertently reduce or increase the drug therapeutic effect (Schmidt and Dalhoff, 2002). Until not long ago, most knowledge about drug-food interactions derived mostly from anecdotal experience, but recent scientific research can demonstrate examples, where food is shown to interfere with the pharmacokinetics and pharmacodynamics of drugs via a known mechanism of interaction: an inhibitory effect of grapefruit juice on Cytochrome P450 isoenzymes (e.g. CYP3A4) that leads to increased bioavailability of drugs e.g. felodipine, cyclosporin and saquinavir and potential symptomatic toxicity has been reported (Seden et al., 2010); green tea reduces plasma concentrations of the β-blocker nadolol, possibly due to inhibition of Organic Anion Transporter Polypeptide 1A2 (OATP1A2) (Misaka et al., 2014); activity and expression of P-glycoprotein (P-gp), an ATP-driven efflux pump with broad substrate specificity, can be affected by food phytochemicals, such as quercetin, bergamottin and catechins, which results in altered absorption and bioavailability of drugs that are Pgp substrates (Zhou et al., 2004b); an antagonistic interaction of anticoagulant drug warfarin with vitamin K_1 in green vegetables (e.g. broccoli, Brussels sprouts, kale, parsley, spinach), whereby the hypoprothrombinemic effect of warfarin is decreased and thromboembolic complications may develop (Yamreudeewong et al., 1995); sesame seeds have also been reported to negatively interfere with the tumor-inhibitory effect of Tamoxifen (Sacco et al., 2008). Judging from the examples above, under most *in vitro* drug-food interaction studies, food is either treated as a black box, or the study focuses on few, well-studied compounds, such as polyphenols, lipids and nutrients.

Our main hypothesis in the current work is that the interference of food on drug pharmacokinetic or pharmacodynamic processes is mainly exerted at the molecular level via natural compounds in food that are biologically active towards a wide range of proteins involved

in drug ADME and drug action. The hypothesis is certainly supported by the large number of natural compounds that have reached the pharmacy shelves as marketed drugs. Hence, the more information we gather about these natural compounds, such as molecular structure, experimental and predicted bioactivity profile, the greater insight we will gain about the molecular mechanisms dictating drug-food interactions, which will help us identifying, predicting and preventing potential unwanted interactions between food and marketed or novel drugs.

Unlike drug bioactivity information that has already been made available for system-level analyses via databases such as ChEMBL (www.ebi.ac.uk/chembldb/) and DrugBank (http://www.drugbank.ca/), biological activity data and source origin information of natural compounds present in food are scarce and unstructured. To this end, we have developed a database generated by text mining of 21 million MEDLINE abstracts that links plant-based foods with their small molecule components, experimental bioactivity data and human disease phenotypes (Jensen et al., 2014). In the present work, we are exploring this resource for links between the natural compound chemical-space of plant-based foods with the drug target space. By integrating protein-chemical interaction networks and gene expression signatures we provide the foundation for understanding mechanistically the effect of eating behaviors on therapeutic intervention strategies.

Results

The drug-like chemical space of the plant-based diet

Our *in-house* database (Jensen et al., 2014), consists of 1,772 plant-based foods associated with ~8,000 unique compounds. Information on the bioactivity profile exists for less than half of these food compounds (Figure 13A). Initially, in order to map the interactions associated with diet and drug treatment on the therapeutic target space, we studied the activity profile of the 3,799 phytochemicals with experimental bioactivity data. We identified 463 phytochemicals with bioactivity at the range of drug activity against 207 drug targets, 18 enzymes, 7 transporters and 3 carriers currently deposited in DrugBank. As shown in Figure 13B, foods that are routinely part of our diet, such as strawberry, tomato, celery and maize, are involved via the bioactive phytochemicals they contain in high number of interactions with targets within these four categories (Figure 13B). This illustrates that ignoring the complete phytochemical content of a food and focusing on a couple of "hot" molecules, a strategy widely applied in traditional food research, will never reveal the true magnitude of drug-food associations.

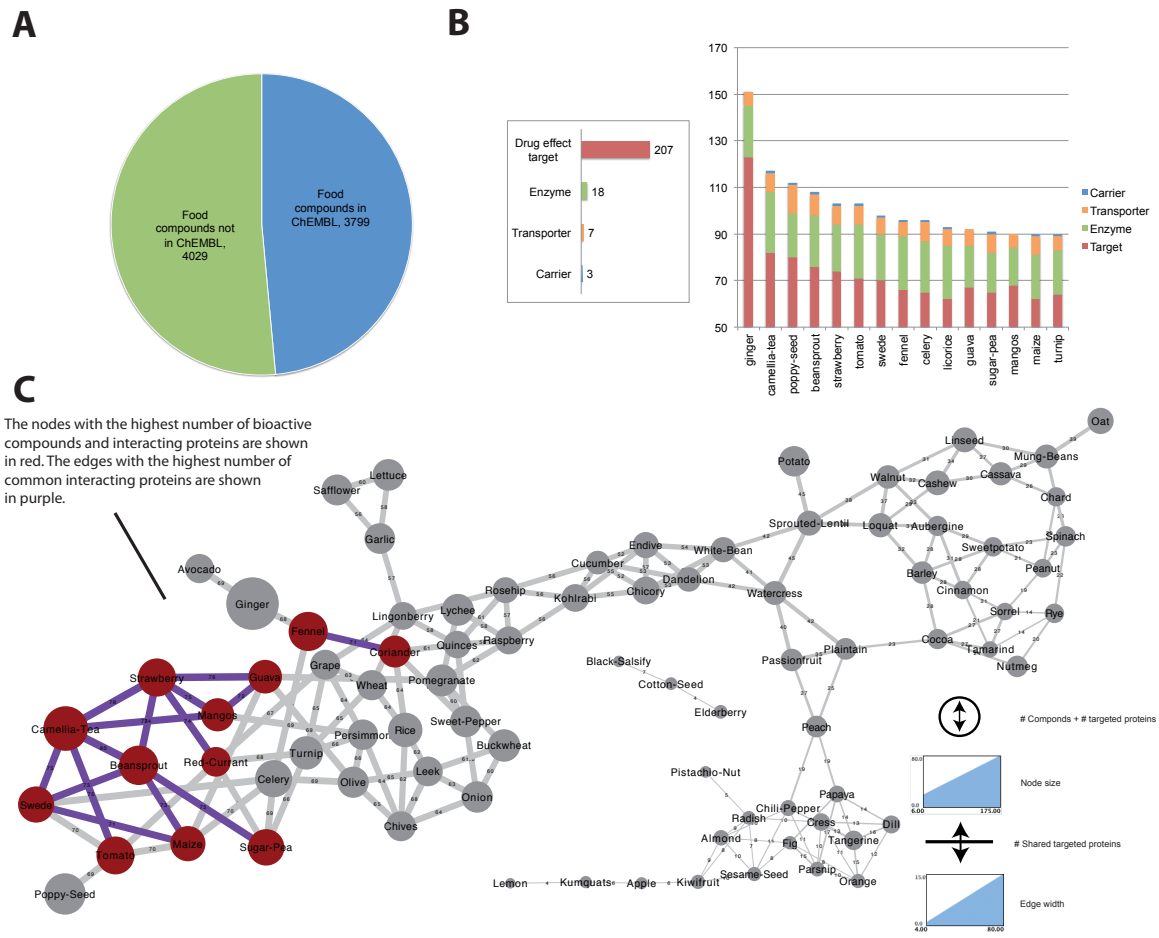


Figure 13: (A) Number of plant-based food compounds in our database with (blue) and without (green) experimental bioactivity information in ChEMBL. (B) Left: Total number of drug effect targets, enzymes, transporters and carriers from DrugBank targeted by food compounds, based on biological activity data deposited in ChEMBL. Right: The plant-based foods with the most interactions to drug carriers, transporters, enzymes and effect targets. The plot shows the 15 most interacting foods within these four categories. (C) Network of foods that share protein targets with drugs. Node size reflects the number of bioactive compounds (phytochemicals) and targeted proteins for a given food. Edge width reflects the number of common targeted proteins between two foods. For visualization purposes, only the 5 strongest edges for each node are shown. The food network with the widest edges is highlighted.

Ginger's phytochemical profile appears as the most biologically active, targeting in total 151 proteins, most of which are targets associated with drug pharmacodynamics. This comes with no surprise since ginger has been positively associated in the scientific literature with at least 87 human disease phenotypes (Jensen et al., 2014). It should be pointed out that the 15 highly interacting foods shown in the figure are not necessarily the best characterized in terms of number of assigned phytochemicals. The number of bioactive phytochemicals in them ranges from 18 for mango to 42 for camellia-tea, while foods like licorice and rhubarb, for example, contain similar number of bioactive compounds (33 and 24 respectively) without, however, targeting as many proteins within these four categories. This indicates that specific structural characteristics of the food components are dictating drug-food interactions.

In order to further hone in the dietary habits that augment the impact on drug efficiency we created a network that relies on the number of protein targets shared between different foods. As shown in Figure 13C several sub-networks of foods target the same protein space, a property that could be taken into account when drugs targeting these proteins are prescribed. For example, safflower, lettuce and garlic are forming a small sub-network sharing more than 55 proteins targeted by their food compounds. The most highly influential food group consists of guava, mango, strawberry, beansprout, camellia-tea, swede and tomato, with the average number of shared protein targets to be more than 70. Papaya, orange, dill, tangerine, cress and chili pepper together with a few more foods form an isolated module targeting a separate protein target space. In all the food clusters of Figure 13C it is apparent that there is no phenotypic or higher level taxonomic characteristic of the foods that could be used to predict the shared interactions with the therapeutic space; this pattern has been revealed from the knowledge of their phytochemical space.

Effect of drug-food interactions on drug pharmacodynamics and pharmacokinetics

To get an insight of the pharmacodynamics processes that are mostly affected by the bioactive phytochemicals of our diet we zoomed in the interactions with the drug effect targets. Comparing Figure 14A, which presents the foods with the highest number of interactions with targets involved in drug pharmacodynamics, with Figure 13B that relies on all protein targets of a drug (effect target, transporter, carrier and metabolic enzyme), we notice that rice and avocado have replaced maize and licorice in the top-15 list. Furthermore, categorizing drug effect targets based on their human disease association, demonstrates the broad spectrum of disease treatments that may be affected by dietary habits.

As shown in Figure 14A, the drug effect targets for 13 disease categories, ranging from neurological and cardiovascular to infectious and immunological diseases, could be potentially altered by food components. It is particularly interesting that cancer-related proteins are highly targeted by dietary molecules; since cancer is still one of the most deadly diseases, patients are willing to follow alternative therapeutic approaches, most often concomitantly with standard drug treatment, such as adopting a "healthy diet" that usually consists of fruits and vegetables. While this approach could be beneficial prior the onset of disease as a preventive measure, it appears that it should be adopted with caution when a patient is under drug therapy, as it may interfere with the therapeutic effect of the drug.

Another observation from Figure 14A, not surprising due to the well-known protective role of plant-based diet against these diseases, is that cardiovascular and gastrointestinal drug effect targets are highly associated with dietary molecules. Furthermore, looking into the association between food and drug effect targets at a biological process level reveals a wide range of functions that are targeted by food components (Figure 14B). Nevertheless, our analysis points to that food "shows a preference" towards a specific drug effect target space that is significantly overrepresented (Student's t-test, $p<0.05$) with proteins involved in signal transduction, immune system and developmental processes (Figure 14B).

Having identified the foods that interact the most with drug effect targets and the biological processes that these targets participate in, we took a step further and zoomed into the level of individual proteins. We constructed a network (Figure 14C) for each disease category, which links drug effect targets based on the drug-food pairs that they share. For example in cancer, drug targets of the carbonic anhydrase family are tightly connected, as they share a large number of drug-food pairs. Looking into the neurological diseases we could identify a tight connection between the kappa- and delta- type opioid receptors, while for cardiovascular diseases a network of the 5-hydroxytryptamine receptors is highly targeted by the same drug-food pairs. Naturally, since many of the drug effect targets are shared between different disease classes, some of these networks were observed in more than one disease category. Nevertheless, similarly to our observations above at the biological process level, our analysis here revealed that drugs developed for certain protein targets are more prone to be affected by diet than others.

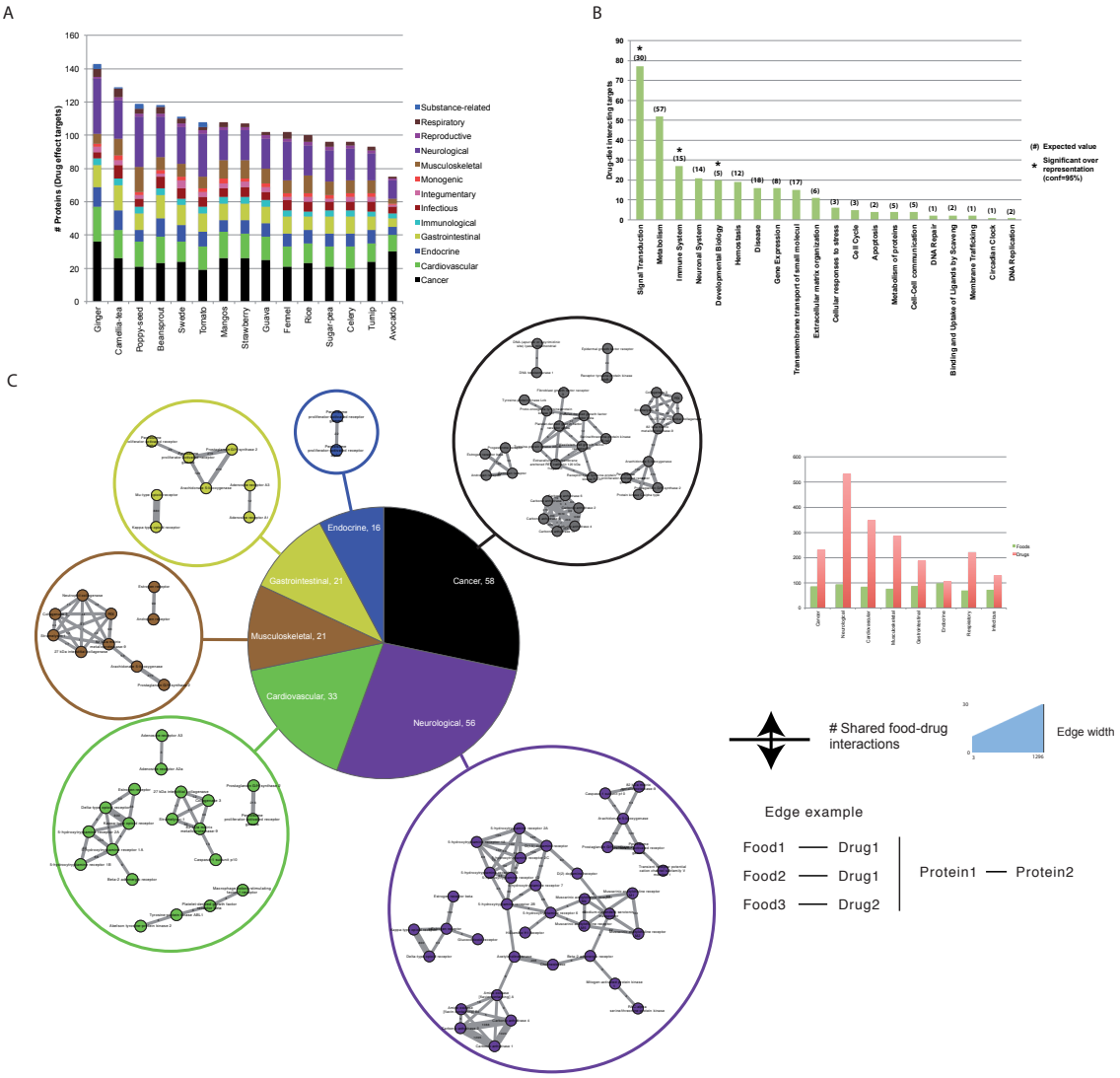


Figure 14: (A) The plot shows the plant-based foods with the highest number of interactions with drug effect targets and their associated human disease classification. (B) The number of drug effect targets affected by food, annotated to different biological systems. The expected number of targets in each biological category was calculated as: exp = (tpc / tdt) * tpa, where, tpc: the total number of drug effect targets from DrugBank in a biological category, tdt: the total number of drug effect targets in all biological categories (1,806 proteins) and tpa: the total number of drug effect targets that participate in drug-food interactions based on our analysis (186 proteins). (C) Networks of drug effect proteins affected by food, per human disease class, shown for the 6 most dominant classes. Two drug effect targets are connected when there are at least 3 drug-food pairs with biological activity against both proteins. We show only the 5 strongest edges for each node. The numbers for each disease class inside the pie correspond to the total number of drug effect targets that are affected by food. The bar-plot to the right shows the total number of foods and drugs interacting with drug effect targets in each disease class.

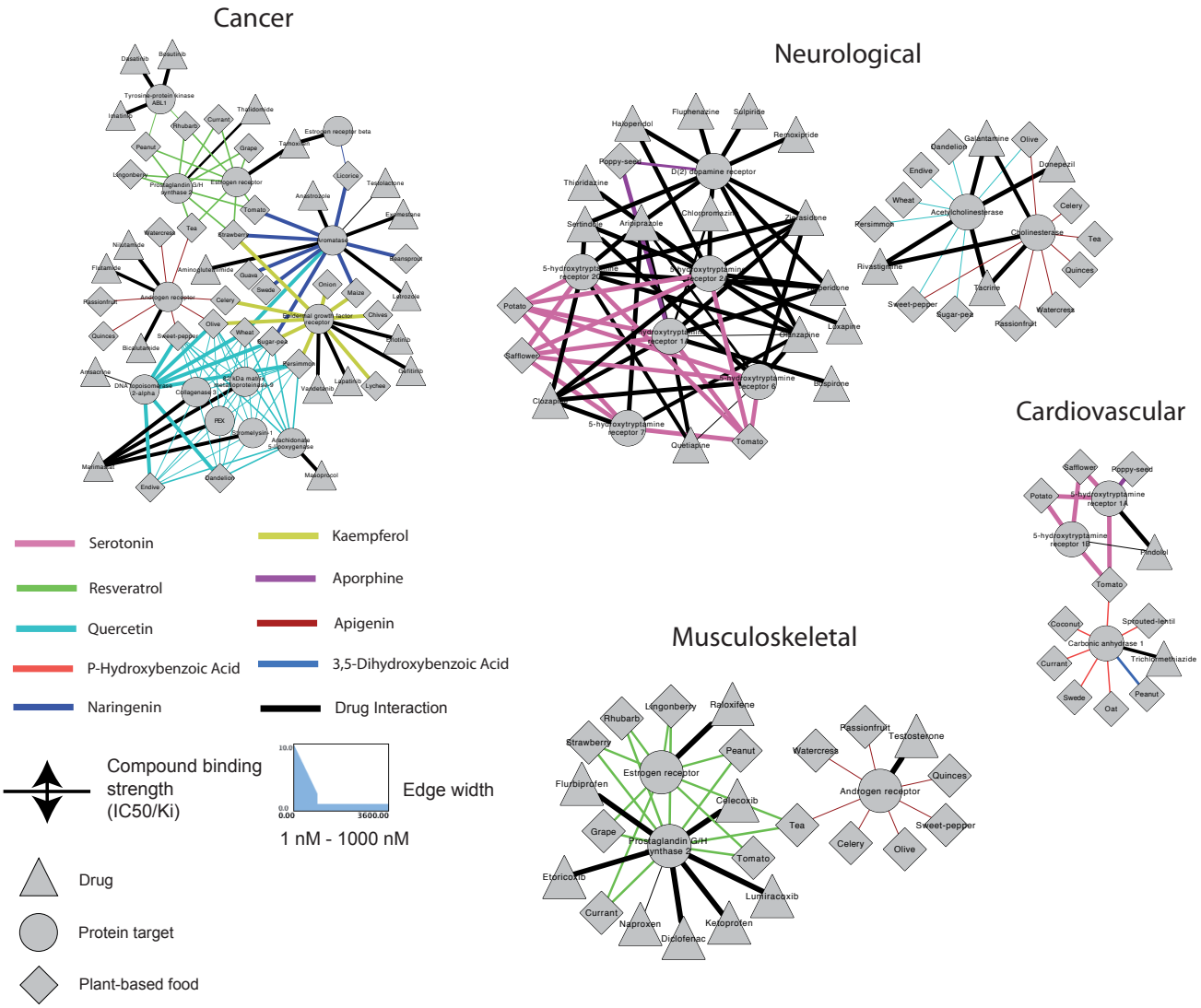


Figure 15: Disease-specific networks of drug-food interactions on proteins affecting drug pharmacodynamics. Node shape denotes drug effect target (circle), drug (triangle) and food (diamond). Edge color highlights the food compound (phytochemical) that shows the highest binding activity to the effect target. Edge width denotes the biological activity (Ki, IC50) and ranges between 1nM to 1000 nM. For visualization purposes, only 3 drugs and 3 foods with the highest biological activities are shown for each drug effect target.

In order to attempt a molecular level understanding of identified drug-food interactions we selected case studies from five disease categories and highlighted the food components with the highest binding affinity to the drug effect targets. As shown in Figure 15, aromatase, a protein targeted by five anticancer drugs (Anastrozole, Testolactone, Exemestane, Letrozole and Aminoglutethiumide) is also targeted by naringenin (IC50=2.9nM), a compound found in licorice, beansprout and maize, among others.

Kaempferol, present in lychee, onion, strawberry and other common foods has a high binding affinity (IC50=3.0nM) for the epidermal growth factor receptor, effect target for Lapatinib, Gefitinib, Vandetanib and Erlotinib anticancer drugs. For neurological diseases, serotonin, present in sunflower, potato and tomato interacts strongly (Ki=1.1nM) with the 5-hydroxytryptamine receptors targeted by several drugs (e.g. Loxapine, Buspirone, etc.), whereas, aporphine, present in poppy-seed, binds to the D(2) dopamine receptor (Ki=527nM), drug effect target of Ofremoxipride, Sulpiride and other drugs (Figure 15).

P-hydroxybenzoic acid, a compound naturally found in coconut, currant, sprouted lentil, swede and other foods, binds strongly (Ki=920nM) to carbonic anhydrase, effect target of the cardiovascular drug Trichlormethiazide. Lastly, resveratrol, a compound that earned its prophylactic reputation against cardiovascular diseases due to its presence in red wine (represented by grape in Figure 15), was found in our analysis to interfere with the activity of several drugs (Diclofenac, Raloxifene, etc.) that target either the estrogen receptor or prostaglandin G/H synthase 2 (involved in musculoskeletal diseases).

Turning our focus to the effects of diet on the pharmacokinetics of drugs, we studied the interactions of food components with proteins involved in drug ADME. Figure 16 illustrates the corresponding drug-food interaction networks for metabolic enzymes and transporters, where apigenin, quercetin, naringenin, resveratrol and nicotinic acid are the food molecules with the strongest binding affinities. These compounds are found in more than 40 foods; hence, a complete understanding of their interaction profile with drug ADME proteins is of utmost importance. For acetylcholinesterase and cholinesterase, two proteins involved in the metabolism of several drugs, e.g. Neostigmine (myasthenia gravis), Isoflurophate (glaucoma), Donepezil (dementia), Galantamine (dementia) the binding affinity of food components quercetin and apigenin is lower than this of the drugs, but still in the range of measured activities for these metabolic targets. In other cases, such as aromatase, involved in the metabolism of Anastrozole, Letrozole, Exemestane and Aminoglutethimide, drugs used against breast cancer, we encounter bioactive food components with stronger activities. Naringenin, a compound found in licorice, sugar pea, guava and others, has a binding affinity of IC50=1000nM against aromatase, comparable with the actual drug's. Similarly, the ribosyl-dihydronicotinamide dehydrogenase, involved in the metabolism of primaquine, is targeted from resveratrol with binding affinity higher than that of the respective drug (IC50=450nM) (Figure 16). Resveratrol interacts as well with the multidrug resistance protein 1, involved in the transport of several cancer drugs (Tamoxifen, Vinblastine) and other types of drugs, such as the antiretroviral drug Nelfinavir or Haloperidol an antipsychotic medication.

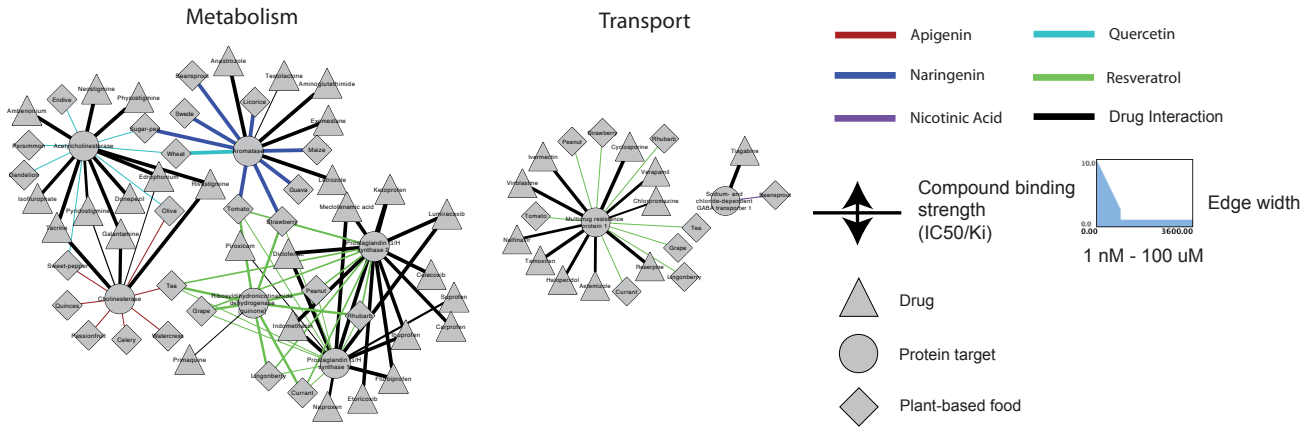


Figure 16: Networks of drug-food interactions affecting drug pharmacokinetics. Node shape denotes enzyme/transporter (circle), drug (triangle) and food (diamond). Edge color highlights the food compound (phytochemical) that shows the highest binding activity to the protein target. Edge width denotes the biological activity (Ki, IC50) and ranges between 1nM to 1000 nM. For visualization purposes, only 3 drugs and 3 foods with the highest biological activities are shown for each protein target.

Evaluation of drug-food interactions through their gene expression signatures

Despite a thorough investigation of the interaction network formed by the bioactive compound space of diet and the drug activity space, the obtained results of possible drug-food interactions heavily rely on the available data related to the phytochemical content of food as well as the activity of these molecules on human proteins. To overcome the barrier of data incompleteness we decided to compare the gene expression signatures of diet with the ones of FDA approved drugs, looking for correlated and anti-correlated profiles. The statistical analysis was performed using the Connectivity Map (Lamb et al., 2006) which includes gene expression signatures from 1,309 compounds, both FDA-approved drugs and bioactive compounds. We found gene expression data for 9 foods that are linked in our *in-house* database with 390 unique compounds (Figure 17). These 390 compounds are chemically similar to both CMap compounds as well as FDA approved drugs currently present in DrugBank (Figure 17A). We could retrieve 5,171 protein targets (direct and indirect) for these compounds, where "disease" is the most enriched pathway with 538 protein targets involved (Figure 17B). Other significantly enriched pathways include cell cycle, developmental biology and apoptosis. Looking into the gene expression profiles of these 9 foods, we notice that 9,072 genes are significantly differentially expressed (FDA corrected moderated t-test, $p<0.05$) between the control samples and the diet interventions.

Interestingly, when these gene targets were further analyzed, the biological processes that were found significantly enriched have a high similarity with the protein space targeted by the food compounds (Figure 17B vs Figure 17C). However, what we also observed was that from the 5,171 proteins targeted by the food components, only 2,653 were found in the significantly differentially expressed (DE) gene list. In our attempt to understand the reasons behind this discrepancy we selected a sub-set of 56 food molecules that were targeting proteins from both groups; the ones that showed a significantly different expression level and the ones that did not (non-DE). We compared the binding affinity for those compounds having both DE targets and non-DE targets as well as the protein connectivity of their targets. While the protein connectivity analysis did not yield a statistically significant difference between the two groups, the IC50 values were significantly lower (Wilcoxon rank-sum test p value< 0.05) for the compounds targeting the non-DE group of genes.

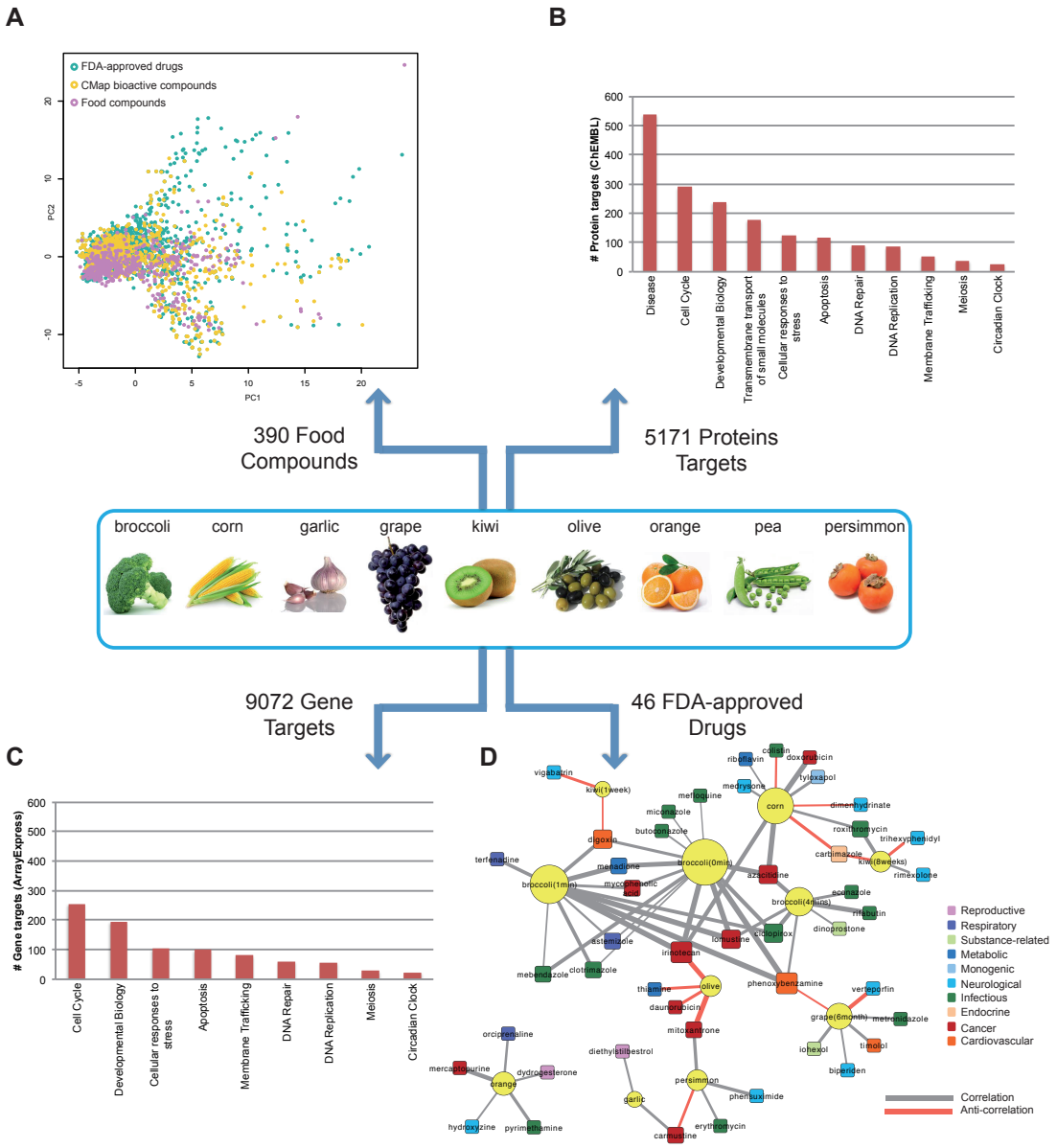


Figure 17: Comparative analysis of food and drug gene expression signatures. In total 9 high-quality gene expression signatures of plant-based foods could be retrieved (A) Chemical similarity comparison between the food compounds (390), the bioactive compounds present in ConnectivityMap (1,309) and FDA-approved drugs in DrugBank. (B) Pathway enrichment analysis of the proteins targeted (directly and indirectly, see Materials and Methods) by the food compounds. (C) Pathway enrichment analysis of the 9,072 genes that were found collectively significantly differentially expressed in the pair-wise comparisons of each food with the respective control. (D) Drug-food interactions based on the food gene expression signatures submitted to CMap. Foods (yellow nodes) and drugs (squares with different colors based on disease classification) are connected if they show a correlated (grey edge) or anti-correlated (orange edge) gene expression signature. The width of the edge indicates the significance level of the observed correlation.

Using as an input the gene expression signatures of each of the nine foods, we retrieved 133 CMap compounds, of which, 46 FDA approved drugs that have a significant correlated or anti-correlated profile (Figure 17D). These 46 FDA approved drugs were further mapped to disease categories showing that mostly drugs used against infectious diseases (13), cancer (9) and neurological diseases (9) induce a gene expression signature that can be either enhanced or reversed by diet. In the drug-food interaction network shown in Figure 17D, broccoli has the highest number of connections with drugs. Interestingly, all connections between broccoli and drugs are correlation-based, most of which display strong correlation coefficients. This finding sheds some additional light -from a mechanistic point of view- on the well-known beneficial effect of broccoli on human health. Orange and garlic induce a gene expression profile that is highly correlated with drugs used in cancer (Carmustine, Mercaptopurine) and reproductive disorders (Dydrogesterone, Diethylstilbestrol); orange specifically, is highly correlated with the activity of Orciprenaline (drug against a metabolic disease), Pyrimethamine (drug against an infectious disease) and Hydroxyzine (drug against a neurological disease). One notable case is olive oil; olive oil induces a gene expression signature highly anti-correlated with the anticancer drugs Mitoxanthrone, Irinotecan and Daunorubicin. Mitoxanthrone and Daunorubicin are typically used against leukemia, where olive oil has not demonstrated any beneficial effect. Irinotecan, on the other hand, is a drug used against colon cancer, a disease which several studies suggest that olive oil has actually a prophylactic effect on. This seemingly contradictory observation could be interpreted from another angle; foods that have the potentially to reverse -to some degree- the global effect of drugs, especially anticancer drugs that trigger severe side effects, without reducing the therapeutic action of the drug, could play a significant role in the life-quality of patients under treatment.

Discussion

Herbal therapies have been used for a variety of symptoms for thousands of years while recently there has been a drastic growth in the consumption of herbs and natural supplements with health benefits. In relation to AIDS and cancer patients especially, two life-threatening diseases where classical drug treatment does not always have a guaranteed effect, the use of both multiple prescription drugs and herbal supplements is very prevalent (Richardson et al., 2000; Sparber et al., 2000). Given that components of herbs and natural supplements interact with human proteins in a similar manner as drugs, there is a high potential for altering drug efficiency. Furthermore, phytochemicals are abundant in our diet and have been shown *in vitro* to influence human proteins and cell-cultures.

Several have demonstrated activity against the same proteins and biological systems as drugs, and thus, potentially influence their pharmacokinetics and pharmacodynamics behavior when consumed concomitantly. In example of sesame seed that has been reported to negatively interfere with the tumor-inhibitory effect of Tamoxifen (Sacco et al., 2008), the protein responsible for the therapeutic effect of Tamoxifen is the estrogen receptor (P003372), which is also targeted by a number of different bioactive phytochemicals present in sesame, including beta-sitosterol (Chang et al., 2005; Hu et al., 2007). According to our *in-house* database, which links food with its molecular components, besides sesame seed, beta-sitosterol is a phytochemical component of guava, onion, pomegranate, turnip, fennel, celery and kiwifruit - all common foods of our diet that could also be potentially involved in interactions with Tamoxifen negatively affecting its therapeutic activity. As another example, health professionals recommend to patients under medication against high blood pressure to avoid consumption of licorice (http://www.ehow.com/list_5798754_foods-avoid-taking-beta_blockers.html, Accessed May 30). The mechanism of this drug-food interaction is not yet clarified, although it has been occasionally attributed to the presence of glycyrrhizin. Our analysis points though towards the phytochemical liquiritin contained in licorice. Liquiritin has been found to interact with the beta-2 adrenergic receptor (P07550; Bioassay CHEMBL1738166), which is the primary target of beta-blockers, such as Penbutolol, a drug against hypertension.

The overall objective of this work was to gain knowledge on the interference of dietary components with the pharmacokinetics and pharmacodynamics processes of medicine, with the purpose of elucidating the molecular mechanisms involved. To the best of our knowledge this is the first time of such a scale integration of data from the literature, online public and *in-house* databases coupled with gene expression analysis to study the effect of natural bioactive compounds from foods on proteins related to drug bioavailability and therapeutic effect. The novelty of this platform is that it takes into account the global effects of food, propelled by its rich natural compound content, increasing the level of confidence of the scientific community and medical professionals when making recommendations for foods that should be avoided under certain medications. We identified clusters of foods that target the same therapeutic space as drugs, a property that could potentially increase the chances for severe alterations of the drug activity if these foods are consumed concomitantly. We also identified a large number of food components that are potentially involved in yet not documented drug-food interactions supporting the notion that ignoring the complete chemical background of a food is a missing link for obtaining a holistic view of the effect of dietary habits. From a methodological point of view we believe that including the actual bioactivity values of the food components against drug targets allowed us to go beyond a simple enumeration of interactions to a more comprehensive and possibly accurate mapping of food-drug associations. Our food-drug interaction network also

revealed that therapeutic interventions for every disease category could be potentially affected to some degree by diet, even though specific disease areas, e.g. cancer and neurological diseases, are most prone to the negative effects of drug-food interactions than others. Lastly, we believe that we have demonstrated with several examples the power of a systems-level analysis to answer the two most important questions for patients and clinicians: (1) which foods should be potentially avoided from a patient under treatment, and (2) which is the underlying mechanism behind these drug-food interactions. However we should also keep in mind that, since many of the food compounds that are strong binders to therapeutic targets are very common in our diet, it will certainly be a daunting task to actually design diets that will not include such compounds. Thus, adding in the network analysis the actual concentration in food of each bioactive compound would give a more accurate picture of the extent and severity of these drug-food interactions.

Materials and Methods

The food-drug interaction space

The plant-based food compounds and their chemical structures were retrieved form our *in-house* database (Jensen et al., 2014). FDA-approved small molecule drugs were retrieved from DrugBank v.3. Food compounds and drugs were mapped to their protein interactions using ChEMBL v.16. Binding activities were retrieved from ChEMBL Bioassays. Protein targets were categorized into "Drug effect target", "Enzyme", "Transporter" and "Carrier", following the DrugBank categorization. For a food compound to be considered active against a protein target, it had to bind within the range of the drugs targeting the same protein. For proteins were the binding activity of the drugs is unknown, the binding activity of the food compound was compared to the binding activities of proteins from the same category (i.e. drug effect target, enzyme, transporter or carrier). Drug effect proteins were mapped to disease categories using the Therapeutic Target Database (Zhu et al., 2012) and the Human Disease Ontology (LePendu et al., 2011). Categories were selected as the third level in the human disease ontology. Drug effect proteins were assigned to biological systems using Reactome (http://www.reactome.org).

Gene expression signature comparison

Chemical similarity between food compounds, CMap bioactive compounds and FDA-approved drugs

SMILES strings of food compounds were retrieved from PubChem (Bolton et al.), while SMILES of the CMap bioactive compounds and FDA-approved small molecule drugs were retrieved from Connectivity Map build 02 (Lamb et al., 2006) and DrugBank 3.0 (Knox et al., 2011), respectively. Based on the chemical structures, molecular or physical descriptors were calculated for each compound using the RDKit plugin (http://www.rdkit.org) in KNIME (Berthold et al., 2008), including a 1024-bit Morgan circular fingerprint, Topological Polar Surface Area (TPSA), octanol/water partition coefficient (SlogP), Molecular Weight (MW), number of Lipinski hydrogen bond acceptors (HBA) and donors (HBD). Afterwards, a matrix of compound descriptors with 1029 columns was constructed, in which each row represented a food compound, a CMap bioactive compound, or an FDA-approved drug and a principle component analysis (PCA) using R was performed.

Retrieval of direct and indirect protein targets of the food compounds

Firstly, food compounds were mapped to exact InChI key matches in ChEMBL and similar ChEMBL compounds using the Morgan circular fingerprint. The Tanimoto Coefficient (TC) was calculated based on Morgan fingerprint. Two of the compounds were similar if TC $\geq$0.85 and their difference in molecular weight was lower than 50 g/mol. Next, the interactions of food compounds and proteins were built by searching in ChEMBL the protein targets of those exactly matched or similar ChEMBL compounds. The bioactivities were filtered based on the following thresholds: for K_i, K_d, IC_{50} and EC_{50}, pchembl_value larger than 6; for inhibition, measurement value greater than 30%; for potency, measurement value lower than 50 μM. To deal with the multiple measurements of the same compound on the same protein, we calculated a frequency of "positive" measurements (served as evidence of compound-protein interaction) among all candidate measurements. Only chemical-protein interactions with a frequency of higher than 0.5 were considered confident and were used for further analysis. In addition to chemical-protein interactions from ChEMBL, we also included first-degree protein-protein interaction (PPI) partners (a confidence score higher than 400) from STRING 9.1 (Franceschini et al., 2013), to expand further our protein target space of food compounds. Note that this was only done for the protein targets of food compounds matching exactly to ChEMBL compounds. Only those PPI

from human, rats and mice were included. After obtaining protein targets for food compounds, protein targets in rats and mice were mapped to their human orthologous proteins through Ensembl Biomart Homolog Service (Flicek et al., 2014).

Microarray data extraction and analysis

ArrayExpress database (Rustici et al., 2013) was first queried with the list of 126 edible plants. Besides microarray data from human, experiments from rats and mice were also included. After manual inspection and quality check, 17 gene expression microarray experiments remained for downstream analysis, which corresponded to 12 unique foods that had available chemical composition information in the *in-house* database. Differential expression analysis of microarray data was conducted with R Bioconductor *limma* package (Smyth, 2004). A *p*-value of 0.05 was used as the cutoff when selecting significantly differentially expressed (DE) genes (Yoav and Hochberg, 1995). For each food, the list of DE genes was split into two lists of up and down-regulated genes, referred to as "tag lists" in Connectivity Map (Lamb et al., 2006). The genes in tag lists were converted to the CMap-compatible probe-set IDs from Affymetrix GeneChip Human Genome U133A Array. For DE genes in rats and mice, the human homolog genes were obtained through Ensembl Biomart Homolog Service (Flicek et al., 2014) before mapping to probe-set IDs. The paired tag lists were used to query Connectivity Map to reveal the correlation or anti-correlation relationship between foods and drugs. Based on the output from Connectivity Map, CMap drugs or bioactive compounds were considered to correlate or anti-correlate with foods if they had an absolute enrichment score higher than 0.75, permutation *p*-value less than 0.01 and a non-null percentage above 80%.

Chapter IV: Exploring Mechanisms of Diet-Colon Cancer Associations through Candidate Molecular Interaction Networks

Exploring Mechanisms of Diet-Colon Cancer Associations through Candidate Molecular Interaction Networks

David Westergaard[1], Jun Li[1], Kasper Jensen[2], Irene Kouskoumvekaki[2], Gianni Panagiotou[1,*]

[1]School of Biological Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong

[2]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark

***E-mail: Correspondence gipa@hku.hk**

Abstract

Background: Epidemiological studies in the recent years have investigated the relationship between dietary habits and disease risk demonstrating that diet has a direct effect on public health. Especially plant-based diets -fruits, vegetables and herbs- are known as a source of molecules with pharmacological properties for treatment of several malignancies. Unquestionably, for developing specific intervention strategies to reduce cancer risk there is a need for a more extensive and holistic examination of the dietary components for exploring the mechanisms of action and understanding the nutrient-nutrient interactions. Here, we used colon cancer as a proof-of-concept for understanding key regulatory sites of diet on the disease pathway.

Results: We started from a unique vantage point by having a database of 158 plants positively associated to colon cancer reduction and their molecular composition (~3,500 unique compounds). We generated a comprehensive picture of the interaction profile of these edible and non-edible plants with a predefined candidate colon cancer target space consisting of ~1,900 proteins. This knowledge allowed us to study systematically the key components in colon cancer that are targeted synergistically by phytochemicals and identify statistically significant and highly correlated protein networks that could be perturbed by dietary habits.

Conclusion: We propose here a framework for interrogating the critical targets in colon cancer process and identifying plant-based dietary interventions as important modifiers using a systems chemical biology approach.

Our methodology for better delineating prevention of colon cancer by nutritional interventions relies heavily on the availability of information about the small molecule constituents of our diet and it can be expanded to any other disease class that previous evidence has linked to lifestyle.

Introduction

Nutrition is the cornerstone of an individual's lifestyle, thus, understanding how diet influences metabolic regulation and how dietary interventions can improve health is a key scientific goal. At the same time, diet has a major influence on the overall quality of life beyond the prevention of diseases. Thus, even though the personalized approach to diet is the logical transition – much like the transition from pharmacology to personalized medicine – this task is extraordinary complex (Kaput, 2008; Kussmann and Fay, 2008). Most foods are composed of hundreds of bioactive compounds, often interacting synergistically, with varied concentrations and several biological targets with different affinities and specificities. However, since nutritional trials are not designed as mechanism-based preclinical studies little is known about these molecular targets. Observational studies performed on populations with distinct diet preferences, often coupled with sophisticated statistical analyses, have offered some associations between certain diseases and foods (Hutter et al., 2012; Nanri et al., 2011; Shiels et al., 2011). Even though in some cases, these kind of studies have led to further mechanistic investigations that resulted in elucidation of the bioactive natural compounds, they suffer in two ways: (*a*) they are phenomenological in nature, (*b*) they are very restricted in the confined disease phenotype space that is under study.

On the other hand, numerous successful stories of FDA-approved therapeutic agents derived from phytochemicals: *Hycamtin* (GlaxoSmithKline), *Navelbine* (Pierre Fabre Pharmaceuticals Inc.), *Taxol* (Bristol-Myers Squibb), *Taxotere* (Sanofi-aventis) among others, have propelled the research on how dietary bioactive compounds interact with target proteins and perturb key signaling pathways. Thus, the second approach that we often encounter in nutritional studies is the *in vivo* or *in vitro* investigation of the beneficial effects of common phytochemicals, such as specific polyphenols that are present in a variety of fruits, vegetables, and other dietary botanicals. An abundant literature has shown that polyphenols can, among others, trigger apoptosis in cancer cells through the modulation of a number of key elements in cellular signal transduction pathways linked to apoptosis (caspases, *bcl-2* genes) (Manson, 2003) and modulate epigenetic alterations in cancer cells (Katada et al., 2012). However, in such studies the major limitation is that the possible therapeutic value of the phenolic compound of interest is evaluated while ignoring the chemical background of the diet, which is probably one of the reasons for contradictory results from different research groups on the same compounds.

Therefore, there is a clear need for more systematic studies to identify those dietary factors that influence the most, reveal their synergistic interactions, and uncover the mechanisms of action.

Cancer research in the last 20 years has brought to the fore a dramatic amount of information at the molecular level, leading to an overwhelming number of possible pharmaceutical targets for drug discovery. However, the redundancy and interconnection of the many regulatory pathways that are involved in cell replication, growth and apoptosis, as well as the capacity of mutations in cancer cells, is a significant barrier for drug development using targeted approaches.

These hard limitations encountered with the targeted approach are contributing to prompt us to reconsider the global fight against cancer, especially for cancers that are proven to be intrinsically or partly related to lifestyle die). Beyond the specific aspect of cancer prevention, understanding the capacity of the body to maintain health homeostasis is a genuine subject of study for which a methodological approach needs to be considered. Taken separately each regulatory cascade interaction may not help framing an operational understanding of health homeostasis whereas a more global view, where the concomitant activity of the largest number of targets with respect to the wave of external agent exposure, such as dietary molecules, could be scrutinized as a complex interaction network. There is a general consensus that in the new era of nutritional research, systems analysis of normal and nutrient-perturbed signaling networks is required for identifying critical network nodes targeted through nutritional intervention of either preventive or therapeutic fashion (Fu et al., 2010). Borrowing chemoinformatics methods, well established and widely used in drug discovery research, could help us understand the complex interaction network between dietary small molecules and biological systems. In line with this, we present here a systems chemical biology approach that provides a fundamental foundation for understanding, which processes involved in the onset, incidence, progression and severity of colon cancer are modulated by dietary components. We selected colon cancer as a case study not only because it is one of the most aggressive cancers and the fourth most commonly diagnosed, but also because colon cancer seems not to be a consequence of aging but of eating behavior (Hambly et al., 2002; Tammariello and Milner, 2010). Nevertheless, the methodology proposed here is applicable to any large-scale diet-disease association study, where information about the small molecule constituents of the diet is available.

Results

The chemical space of diet associated with colon cancer

Using an *in-house* database developed by Jensen *et al.,* (2014) (Jensen et al., 2014) through text mining of 21 million abstracts present in PubMed, we investigated here the role of dietary small molecules present in plants (edible and non-edible) with an established phenotype against colon cancer. As shown in Figure 18A 158 plants that have been positively associated in the literature with colon cancer, with 39 of them being part of a common diet, e.g. celery, garlic, thyme, among others (Figure 18A). From our *in-house* database we could also retrieve molecular information for these plants for 3,526 unique phytochemicals. It is quite interesting that despite the fact that all these plants have been positively associated with colon cancer reduction, the majority of phytochemicals (2,023) are plant specific (found only in one plant). The number of compounds associated with each plant varied considerably (Figure 18A), with a median value of 41 compounds per plant and reaching as high as 392 compounds for ginseng. Not surprisingly, as shown in Figure 18A, common foods have been studied more thoroughly than other plants and have a median value of 129 compounds per food. Nevertheless, a group of phytochemicals has been detected in a very large number of plants associated with colon cancer; quercetin (N=51), gallic acid (N=43), vitamin P (N=43), gamma-sitosterol (N=42) and kaempferol-3-O-rutinoside (N=38). Since our objective was to obtain a mechanistic understanding of the association between plants and colon cancer we examined how many of the 3,526 phytochemicals are present an exact match in the ChEMBL database, one of the largest repositories of chemical-protein interaction data. As shown in Figure 18A, roughly for one third of the compounds present in each plant there are available experimental data for their interactions with the human protein space. We could find, in total, an exact match in ChEMBL for 1,663 phytochemicals, while 887 have a chemical similarity (TC > 0.85 & a difference in molecular weight less than 50 g/mol) with at least one other chemical in ChEMBL. For the remaining 976 compounds no information is available for possible protein targets. The chemical space of the phytochemicals from plants associated with colon cancer was evaluated using 1,027 chemical descriptors. To obtain a holistic view of this chemical space we compared it with FDA approved drugs and the human metabolites involved in the colon metabolic network developed by Agren *et al.* (Agren et al., 2012). Interestingly, a large percentage of the plant phytochemicals shows a high degree of chemical similarity with metabolites of the human colon metabolic network pointing out that the therapeutic effect of these plants could be mediated at a metabolic level (Figure 18B); however, we should not overlook the high similarity between FDA drugs and plant phytochemicals and especially anticancer drugs (Figure 18B).

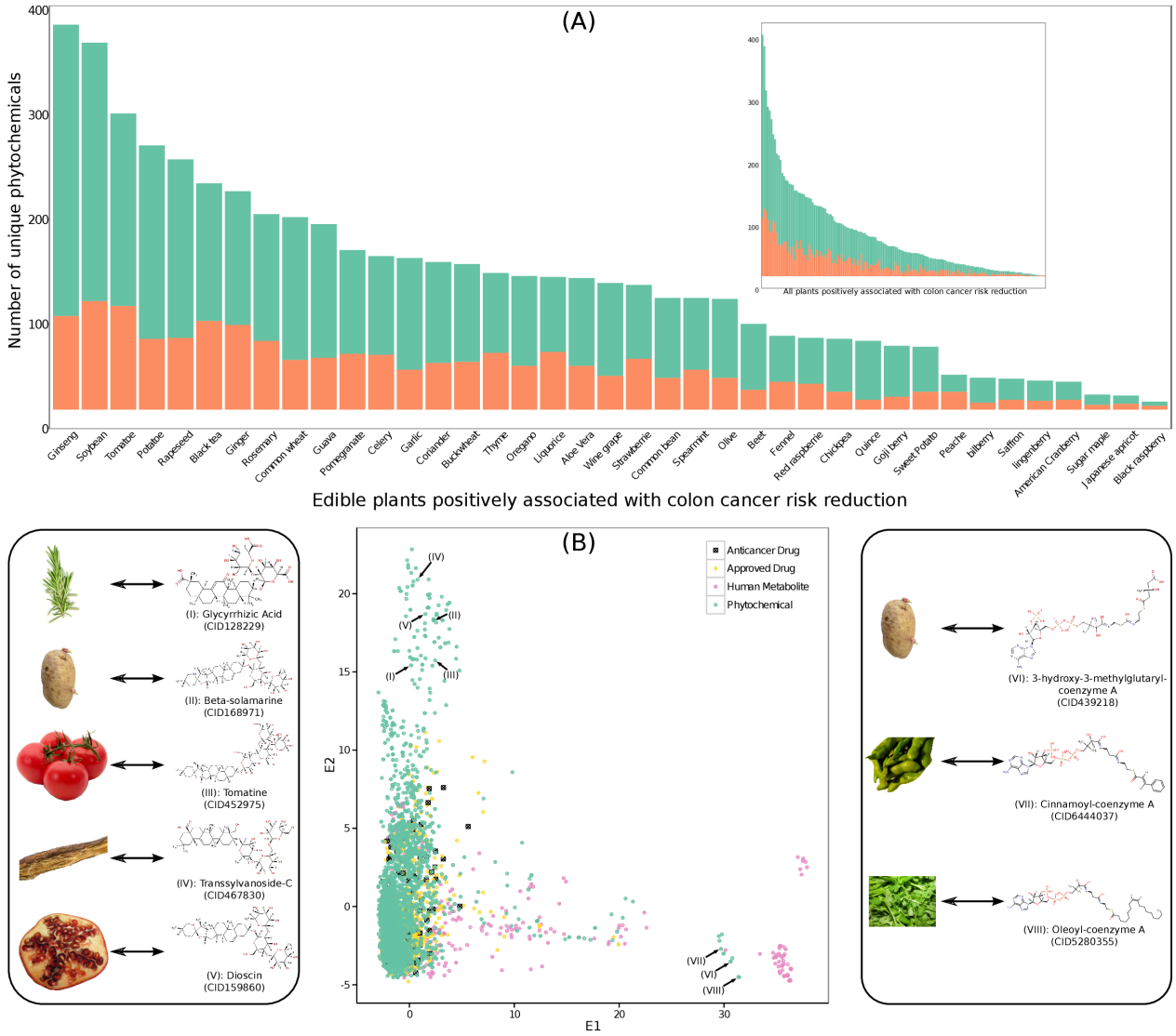


Figure 18: (A) A sub list containing 39 plants that are considered typical for western diet; the embedded figure shows the compound distribution as found in the 159 edible and non-edible plants. Green indicates total number of compounds, while orange indicates compounds that were found as exact match in CHEMBL; (B) Clustering of the plant compounds (green), human colon metabolites (yellow), FDA approved drugs (purple) and anticancer FDA approved drugs (black cross) based on 1027 chemical descriptors. Selected compounds ((i)-(vii)) from the plants chemical space that show no or low chemical similarity with the other groups of compounds are shown on the left.

On the other hand a large number of phytochemicals (left upper corner of Figure 18B) has a very unique chemical profile with no similarities to either the drug space or the colon metabolic network. Examples of such compounds (Figure 18B) are glycyrrhizic acid (in rosemary), beta-solamarine (in potato), tomatine (in tomato), transsylvanoside C (in liquorice) and diocine (in pomegranate). The above compounds are present in just a handful of edible and non-edible plants that have been associated to colon cancer. In the lower right part of Figure 18B we find compounds with structural similarities solely with approved drugs, e.g. 3-hydroxy-3-methylglutaryl-coenzyme A (in potato), cinnamoyl-coenzyme A (in soybean) and oleoyl-coenzyme A (in coriander). The source of these compounds is again restricted to a few edible and non-edible plants.

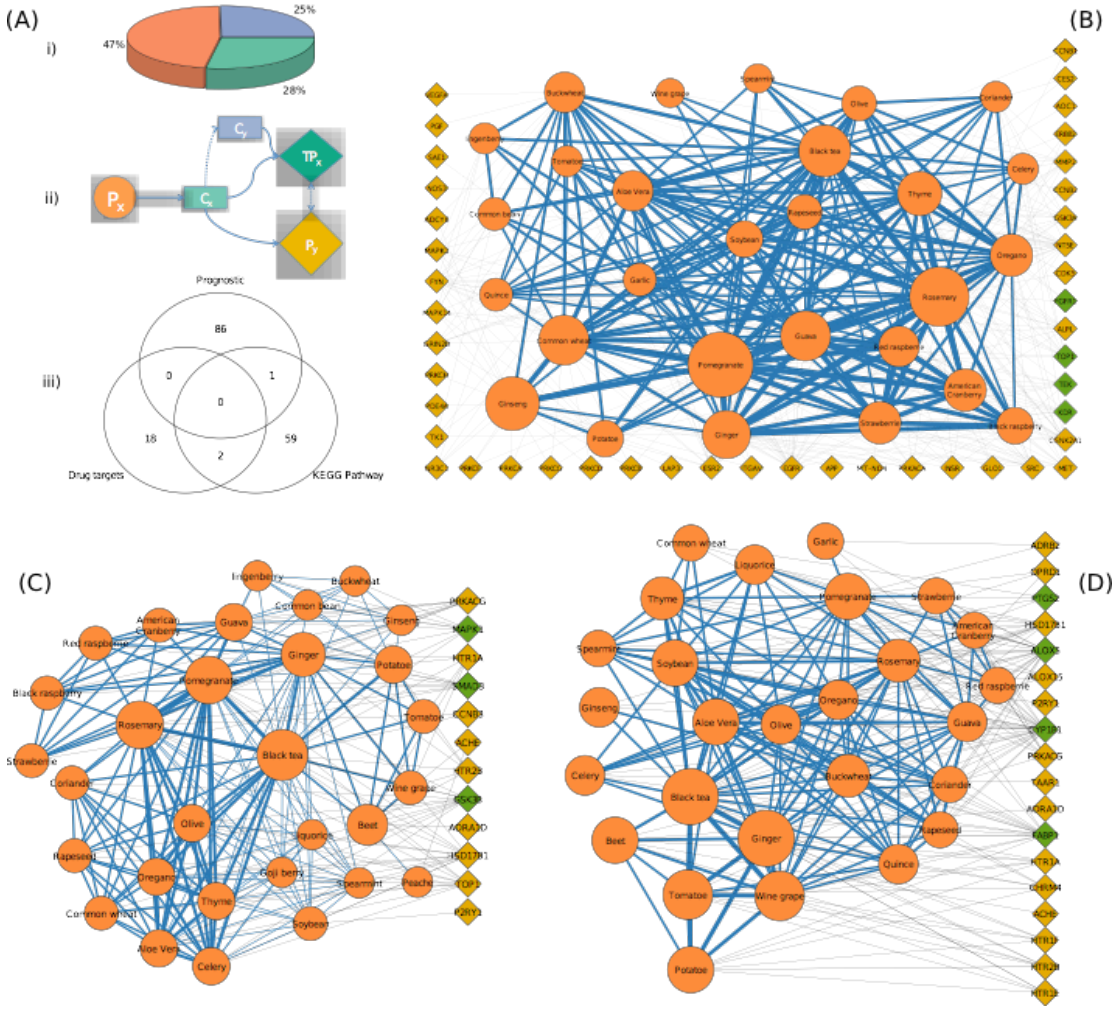


Figure 19: (A) (i) The pie chart offers information for the compounds present in these edible plants; phytochemicals found as exact match in ChEMBL (orange), phytochemicals having at least one similar small molecule present in ChEMBL that do not fulfill any of the above criteria (green); (ii) A graphical representation of how the interactome map between edible plants and candidate colon cancer protein targets was generated: (a) an edible plant (orange circle; Px) contains a

phytochemical (box; Cx) which interacts directly with a target protein (green diamond; TPx) or through a first-degree neighbor (yellow diamond; Py) of TPx, (b) the phytochemical Cx is structurally similar with a compound (Cy) that interacts with the target protein (TPx). Straight lines represent verified interactions while dashed lines represent predictions; (iii) The Venn diagram shows the overlap between the proteins in the three candidate colon cancer target sets; (B) A plant-protein interaction network based on the interactions between phytochemicals, FDA approved colon cancer drug targets and their first-degree neighbors. The size of the plant node is proportional to the number of proteins that its molecular components target. The width of the edge connecting two plants reflects how many protein targets the plants share. (C) A plant-protein interaction network using as a candidate target space the KEGG colon cancer pathway. (D) A plant-protein interaction network using as a candidate target space the colon cancer prognostic gene signatures (Oh et al., 2012). The color of the nodes in (B)-(D) is according to (Aii).

An interactome map of candidate colon cancer targets and diet

To unravel the interactions associated with diet and colon cancer we studied the complex interaction networks of the small molecules present in the 158 plants and the candidate colon cancer target space. By "candidate" here we mean proteins that are potentially involved in the onset and development of colon cancer and fall in one of the following categories (Figure 19A), (i) proteins that are established targets of the FDA approved colon cancer drugs (N=20) and their first degree neighbors (N=1,224); ii) proteins that participate in the KEGG colon cancer pathway (N=62) and their first degree neighbors (N=1,588); iii) proteins characterized by Oh *et al.,* (Oh et al., 2012) as colon cancer prognostic signature (N=87) and their first degree neighbors (N=870). First-degree protein neighbors were included since cancer-related proteins are more likely to act as hubs in protein interaction networks. This feature of cancer proteins makes them amenable to activity disturbances through ligand binding to their protein neighbors. To ensure the biological validity of the interactions in the context of colon cancer only proteins for which there is positive evidence of expression in the colon according to the Human Protein Atlas (Uhlen et al., 2010) were included. As shown by the Venn diagram in Figure 19A there is very low overlap between the three categories (not taking into account the first-degree neighbors). The total number of unique proteins is 181 (and 1,708 unique first degree neighbors).

For the association of the plants with the candidate colon cancer protein space through their molecular components we considered direct (a) and indirect (b) interactions: (a) a plant

contains a compound that was found as an exact match in ChEMBL to interact with either the set of 181 or 1708 proteins; (b) a plant contains a compound structurally similar to a compound in ChEMBL that interacts with the set of 181 proteins. Only high confidence interactions, either chemical-protein or protein-protein, were kept (see Materials and Methods).

(i) In total, 105 plants (33 edible) were found to interact with 4 target proteins (TEK, KDR, FGR1 and TOP1) of the FDA approved colon cancer drugs and 43 of their first-degree neighbors. The mean number of proteins from this category targeted by each plant was 6 (9 if we restrict the analysis to the common edible). The most targeted proteins of the compounds from edible plants were EGFR (targeted by 26 edible plants or 79% of the total edible plants in our database), NT5E (24 edible plants), ESR2 (20 edible plants), CSNK2A1 (17 edible plants) and FYN (17 edible plants). None of the above is an FDA colon cancer drug target but all are first degree neighbors of the drug targets. Similar results were observed when looking into the most targeted proteins of the non-edible plants. The interaction network between common edible plants and proteins from this category is shown in Figure 19B. The FDA approved drugs used against 4 proteins that are targeted by small molecules present in the edible plants, are irinotecan (TOP1) and regorafenib (TEK, KDR, FGR1). Pomegranate ($N_{small\ molecules}$=13, targeting $N_{proteins}$=23), rosemary ($N_{small\ molecules}$=13, targeting $N_{proteins}$=20), black tea ($N_{small\ molecules}$=12, targeting $N_{proteins}$=16), ginseng ($N_{small\ molecules}$=14, targeting $N_{proteins}$=17) and wheat ($N_{small\ molecules}$=10, targeting $N_{proteins}$=11) are the common foods of our diet with small molecules targeting the most this protein space. The mean connectivity ratio for the edible plants of Figure 19B is 3.7 (calculated as the sum of all edge weights divided by the number of edge weights).

(ii) Similarly, we explored the interaction pattern of the phytochemicals in the 158 plants with the proteins in the KEGG colon cancer pathway and their first-degree neighbors. In total 12 proteins are targeted by 73 plants through 32 unique small molecules. From the common edible plants (Figure 19C) 11 were found to have compounds targeting 3 proteins (MARK1, SMAD3, GSK3B) of the KEGG colon cancer pathway and 28 targeting the remaining 9, which are first-degree neighbors of the proteins in the KEGG colon cancer pathway. Black tea ($N_{small\ molecules}$=5, targeting $N_{proteins}$=7), ginger ($N_{small\ molecules}$=4, targeting $N_{proteins}$=5), rosemary ($N_{small\ molecules}$=5, targeting $N_{proteins}$=6) and pomegranate ($N_{small\ molecules}$=4, targeting $N_{proteins}$=6) are the common foods of our diet with small molecules targeting the most this protein space. Interestingly, the pattern we observe in Figure 19B,C does not depend on the actual number of compounds present in each plant. Soybean, tomato, potato and guava, among others, are edible common plants with a large number of compounds however, none of them appears to target a large part of the candidate colon cancer protein space.

The proteins from this category that are targeted the most from the edible common plants are HSD17B1 (43% of edible plants), TOP1 (37%), GSK3B (37%), SMAD3 (27%) and PRKACG (27%). GSK3B and SMAD3 are proteins involved in the KEGG colon cancer pathway and the remaining three are first-degree neighbors. The mean connectivity ratio of the network of Figure 19C is 1.7.

(iii) The last candidate colon cancer protein space under study was the prognostic signatures based on gene expression data (Oh et al., 2012). From the 87 proteins and their 870 first-degree neighbors that this targeted space consists of, only 5 proteins designated as prognostic signatures and 17 of their first-degree neighbors are targeted by the chemical space of the plants. In this category we found most of the edible plants: 41 small molecules from 37 edible plants target 4 proteins that constitute colon cancer prognostic signatures and 14 that are their first-degree neighbors. The edible plants with the highest activity in this target space are black tea, ginger, tomato and grape, which interestingly share none of the active compounds. The most targeted proteins are CYP1B1, ALOX5 and FABP1 and the mean connectivity ratio of the network of Figure 19D is 3.9.

The candidate colon cancer protein space that is targeted by plants with an established phenotypic effect against colon cancer allowed us to get a better insight on the biological and network properties of these 79 proteins. By using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al., 2009), the 79 proteins were annotated according to the protein domains as found in the InterPro database (Hunter et al., 2012), and subsequently tested for enrichment. Then these domains were clustered using the integrated functional clustering in DAVID (default settings). Only clusters in which at least one InterPro family had a p-value < 0.05 (corrected for multiple testing using the Benjamini-Hochberg approach) were considered significantly enriched. We found three enriched clusters, in which the cluster with the highest enrichment score consisted mostly of kinase domains. The other two domains are related to cell division (Cyclin domains) and growth (EGF receptor domain). We also computed the intra-cluster distance between two proteins as the average shortest path distance between all pairs in this set of 79 proteins and we found a value of 2.2 (with the longest to be 5). Even though we started with three discrete candidate colon cancer target sets, the topological coefficients reveal a strong communication between the proteins, which may offer an explanation why, despite the differences in the observed candidate target space of each plant, they all produce the same phenotypic effect.

Seventy-two (72) plants, of which 28 edible (e.g. celery, thyme, coriander, oregano, olive, ginger among others) were found to contain compounds that target proteins from all three candidate colon cancer protein space. 97 plants; 35 edible- target proteins from two protein spaces and 117 plants; 37 edible target proteins from one. For 41 of the plants, of which 2 were common

edible plants (chickpea and sugar maple). We found no compounds to interact with any of the four categories of the candidate colon cancer target space. The number of the chemical compounds associated to chickpea and sugar maple could not explain this observation since these two edible plants are not the least characterized (Figure 18B). On the other hand, the average number of compounds (14.6) assigned to the non-edible plants showing no interactions with the colon cancer targets was significantly lower than the rest of the plant space of our study.

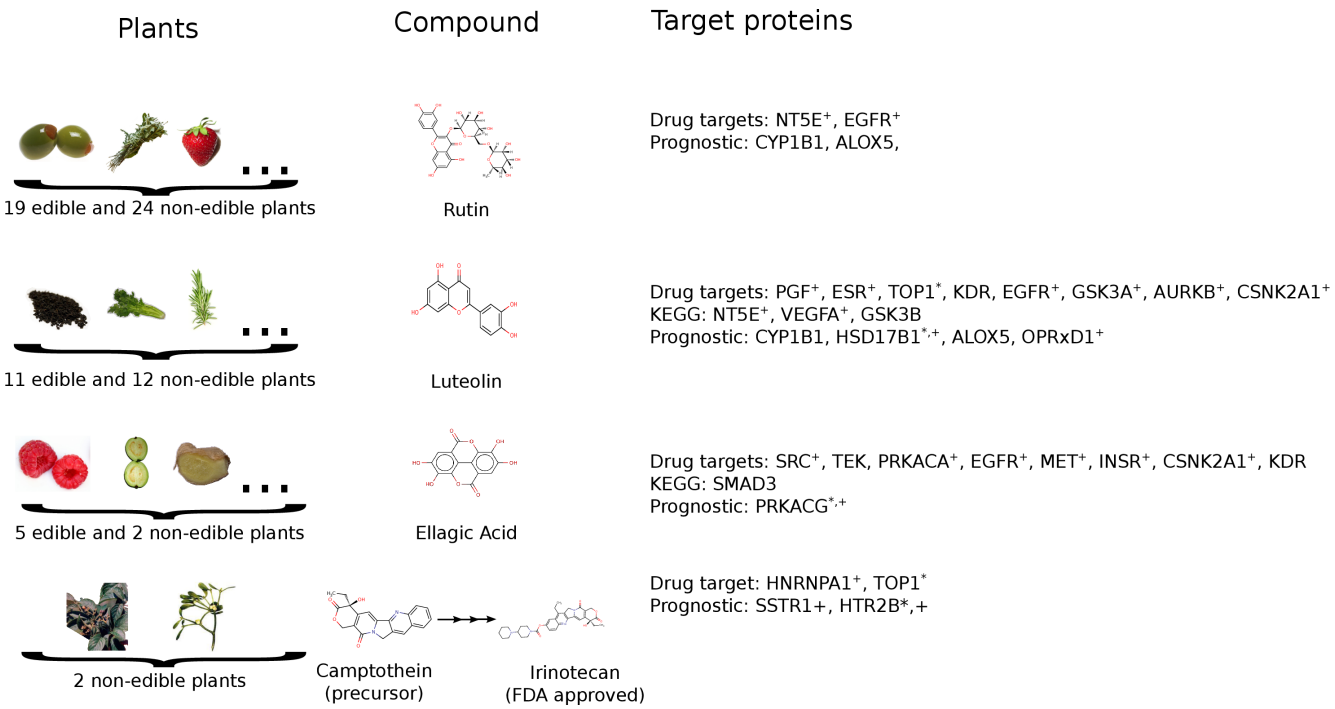


Figure 20: Phytochemicals present in edible and non-edible plants with significant interest in connection to the candidate colon cancer target space.
*: Target also found in KEGG Pathway
+: Protein is a Px (see Figure 19)

The "hot" colon cancer space

The next step was to take a closer look on these phytochemicals and proteins that could hold the key for understanding the mechanism behind the positive phenotype of the particular edible and non-edible plants against colon cancer. In Figure 20 we present some of the chemical structures that caught our attention in this analysis. For example, rutin, is a compound present in 19 edible (olive, thyme, strawberry, among others) and 24 non-edible plants. Rutin targets two proteins, namely NT5E and EGFR, which both interact with established colon cancer drug targets and two proteins that are part of the colon cancer prognostic signature gene set (CYP1B1 and ALOX5). Another interesting compound is luteolin, a compound found in 11 edible (black tea, celery, rosemary, among others) and 12 non-edible plants. Luteolin has an interesting

interaction network that includes 15 proteins from three candidate colon cancer target spaces, making it undoubtedly one of the most "hot" dietary compounds. Ellagic acid, present in 5 edible (strawberry, guava, ginger, rosemary and cranberry) and 2 non-edible plants targets as well proteins from all three candidate protein sets (Figure 20). Another interesting compound is camptothecin that was found in 2 non-edible plants. Camptothecin is actually a precursor of irinotecan, a drug primarily used for the treatment of colon cancer.

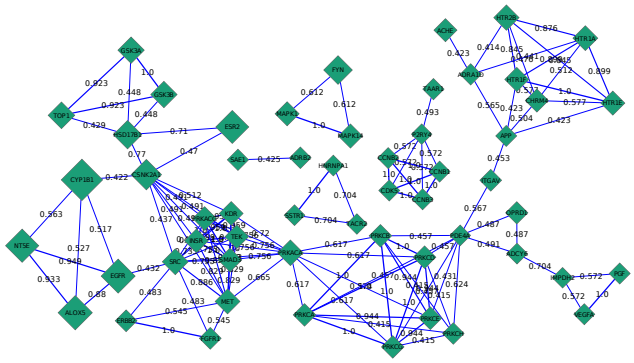
(A)

Edible Plants	X_1	X_2	X_3	X_4	...	X_n	Efficacy
Pomegranate	0	0	1	1	...	0	0.75
Rosemary	1	0	1	0	...	1	0.68
Black tea	1	1	1	1	...	1	0.65
Ginger	0	0	0	1	...	1	0.58
...	...	...	...	...	...	...	...
Fennel	0	0	0	0	...	1	0.02

(B)

Edible Plants	X_1	X_2	X_3	X_4	...	X_n	Efficacy
Black Tea	5	10	2	3	...	8	0.45
Rosemary	2	0	1	0	...	5	0.42
Pomegranate	0	0	8	1	...	0	0.41
Ginseng	0	0	0	3	...	1	0.35
...	...	...	...	...	...	...	...
Japanese Apricot	0	0	0	0	...	1	0.02

(C)



	$y=1$	$y=0$	total
$x=1$	n_{11}	n_{10}	$n_{1\bullet}$
$x=0$	n_{01}	n_{00}	$n_{0\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}}$$

$$\phi^2 \cdot n = \chi^2$$

Figure 21: The two tables show the efficacy of each plant (for details see Materials and Methods) calculated using either (A): a binary index; 0 = no phytochemicals in the plant interact with the protein (protein in the tables are indicated as X1-Xn), 1 = at least one phytochemical in the plant interacts with the protein, or (B): the number of unique phytochemicals present in the plant that interact with the protein. (C): A network of candidate colon cancer proteins connected by the phi correlation coefficient. Labels on the edges are the phi correlation coefficients. The network shows only significant correlations ($p < 0.05$, Bonferonni corrected). Correlation coefficients are calculated using the knowledge of plant-protein interactions, and thus a large correlation coefficient indicates that these two proteins are targeted by a common set of plants. The node size of is proportional to the actual number of plants targeting the protein. The equations used to calculate is shown on the right, in which the squared phi coefficient is related to the Chi-squared test statistic by the number of samples.

Subsequently we calculated the efficacy *E(P)* of each plant as the ratio of the number of candidate colon cancer proteins targeted by plant *P* to the total number of candidate colon cancer proteins targeted by all plants (79 in total). As shown in Figure 21A the highest efficacy was observed for pomegranate (59 targeted proteins, E=0.75), rosemary (54 targeted proteins, E=0.68), black tea (51 targeted proteins, E=0.65) and ginger (46 targeted proteins, E=0.58) from the edible plant space and *Gingo biloba* (E=1), *Butea monosperma* (E=0.48), *Withania somnifera* (E=0.48) and *Galphimia glauca* (E=0.43) from the non-edible plant space. If we take into consideration the actual number of compounds present in the plants that target each of these proteins the picture is slightly different (Figure 21B). The most promising edible plant interacting with the candidate colon cancer protein space now appears to be black tea with an efficacy of E=0.45, while ginseng reaches an efficacy of E=0.35. From the non-edible plants *Ginkgo biloba* shows the highest weighted efficacy followed by *Withania somnifera* and *Butea monosperma*.

In the last part of the analysis we tried to develop the necessary statistical framework in order to achieve the main objective of our study; to actually reveal the protein space that may explain the observed anti-cancer phenotype of these edible and non-edible plants. Due to its complex chemical background the way that diet induces particular phenotypes must be fundamentally different from the drug mode of action (one compound - one, or more target(s)). We calculated the correlation coefficients based on the plant-protein interaction patterns of Figure 21A. In Figure 21C we have only included the significant correlations ($p < 0.05$, Bonferonni corrected) that show some very interesting patterns in the candidate colon cancer target space. There are several small networks of proteins that are consistently targeted, or avoided by these plants. The significance of this analysis is not that it further reduces the candidate colon cancer target space to 55 proteins (Figure 21C) but mainly that it allows to formulate hypotheses on the sets of proteins that could be of significant interest as potential targets, either through dietary interventions or polypharmacology. The smallest networks consist of only three proteins (e.g. MAPK1/FYN/MAPK14, HNRNPA1/SSTR1/TACR2), whereas the largest one of 41 proteins. However, as shown in Figure 21C, this large network could be viewed as five sub-networks that was bridged via PDE4A (connecting three sub-networks) and PRKACA (connecting two sub-networks). As can also be seen in Figure 21C, there are a lot of edges with high correlation coefficients, but when taking the node size into account (the number of plants actually targeting a protein) the number of interesting clusters decreases. For instance, in the TOP1/GSK3A/GSK3B cluster all edges have correlation coefficients > 0.923 and a relatively large node size, indicating that these three proteins are targeted by a lot of plants as a group (indicated by the high correlation coefficient).

Another cluster in Figure 21C with high correlation coefficients and a relatively large node size is the NT5E/ALOX5/EGFR. CYP1B1, despite being the largest node in this network, shows very poor correlation (max(phi_CYP1B1)=0.563). This is probably due to the fact that nearly all compounds in the drug development pipeline are screened against this target in ADME assays. In that sense this target is most likely not as interesting as the aforementioned protein clusters when it comes to explaining the observed colon cancer phenotypes of particular plants as the result of synergistic interactions of small molecules.

Metabolic regulation by dietary components

Two of our previous observations, the fact that the majority of the plant phytochemicals appear structurally similar to the assigned metabolites in the human colon metabolic network (Figure 18B) and that 43 plants with a known phenotype against colon cancer have no compounds interacting with the candidate colon cancer protein space, was the motivation to study the possible metabolic regulation triggered by dietary components. The colon metabolic network consists of 2,934 metabolites and 1,773 enzymes involved in 3,060 reactions. In the 158 plants that have been positively associated with colon cancer reduction there are 122 phytochemicals that are exact match to one human metabolite and 13 more that are structurally similar to 10 human metabolites. We make the assumption that these phytochemicals perturb a human metabolic reaction in the colon only if they appear in the enzymatic reaction as substrates. Based on this, we found in total 570 metabolic reactions in colon to be affected by plants. From the edible plant space, soybean, rapeseed, potato and ginseng was the ones with the highest influence in the colon metabolic regulation by perturbing 421, 225, 210 and 196 of metabolic reactions, respectively. If we look specifically at the 43 plants that was not linked with any of the candidate colon cancer protein space, we see that only chickpea contains phytochemicals that are involved in 76 metabolic reactions in the colon.

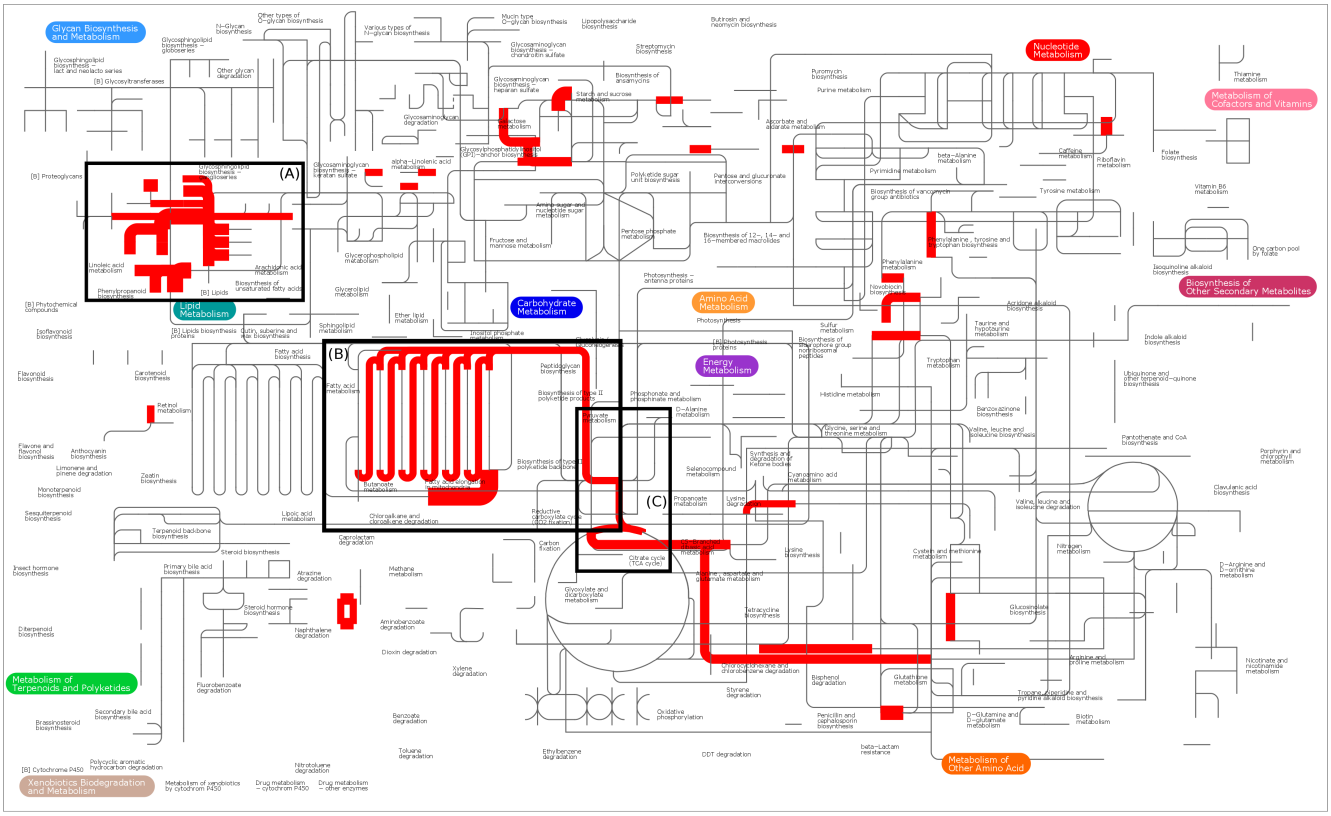


Figure 22: The metabolic pathways of the human colon metabolic network (Agren *et al.,* 2012) influenced mostly by phytochemicals present in plants associated with colon cancer. We highlighted the reactions in pathways that contain metabolites as substrates present in more than 20 plants. The width of the edge is proportional to the number of plants targeting a reaction in that pathway. We have zoomed in on lipid metabolism (A), fatty acid (B), pyruvate metabolism and TCA cycle (C).

The observation that the regulation of the human metabolic network is under the control of signaling pathways often altered in cancer has shifted a lot of attention to cancer metabolism (Hsu and Sabatini, 2008). This has actually revealed the therapeutic potential of metabolic targets in cancer with important implications in the development of anticancer drugs. From Figure 22 we could actually get a visual representation of the metabolic processes in colon that are mostly targeted by the plants associated with colon cancer. Interestingly, the most targeted parts of the colon metabolic network are the lipid, fatty acid and pyruvate metabolism as well as the TCA cycle. Our findings are to a great extent in agreement with the analysis performed recently by Hu *et al.* (Hu et al., 2013), who used gene expression profiles gathered over the last decade to investigate the global shift in metabolic gene expression between and within cancers, including colon cancer. In their study, tumor-induced mRNA expression changes in lipid metabolism and fatty acid biosynthesis were associated with several cancer types. Even more interesting were the findings on colon cancer that were further validated by measurement of metabolite levels; there

was observed a significant decrease in citrate concentration in tumor samples as well as a down-regulation of the pyruvate dehydrogenase complex that controls the majority of glucose carbon flux into the TCA cycle. Monitoring the levels of the TCA cycle intermediates in colon cancer patients after introducing specific dietary interventions could offer additional evidence for the mechanism that associates plants with colon cancer.

Discussion

The term ´´exposome´´, which is used to describe the totality of environmental exposures (e.g. diet, air pollutants, lifestyle factors) over the life course of an individual, has been proposed as a critical entity for disease etiology that complements genome information (Brook et al., 2010; Heinrich, 2011; Wild, 2011). Diet is certainly one of the most dynamic expressions of the exposome and one of the most challenging to assess its effects in health homeostasis and disease development, due to its many components and their temporal variation. Recognize, understand and interpret the interplay between diet and biological systems may contribute to the weight of evidence for assigning causality to a diet-disease association. Therefore, in order to open up new avenues to disease prevention through diet interventions it is crucial to provide insights into the mechanisms by which exposure to the chemical space of food might be exerting its effects.

Towards this direction we used colon cancer as a proof-of-concept for developing the necessary toolbox for a more cohesive view of diet exposure. From our systematic analysis of the candidate colon cancer target space, consisting of ~1,900 proteins, we identified a sub-set of 79 and further reduced it to 55 proteins that may reflect the mechanism by which the small molecule constituents synergistically define a food's anti-cancer activity. This is in our opinion the most important contribution of our study; we go beyond the one compound-one target paradigm that has been extensively used in drug discovery and is often borrowed to explain the mode of action of dietary interventions. In contrast, here we identified statistically significant small protein clusters, from a pre-defined candidate colon cancer related space (avoiding in that way noise from uncurated protein interactions), that are targeted by dietary small molecules in a highly correlated manner. We have demonstrated that plants with different molecular profile can be associated to colon anticancer activity, as long as their protein targets are part of the same disease space.

Furthermore, we attempted to rank the efficacy of the plants associated to colon cancer using a simple scoring system. Taking again into consideration all the compounds present in each plant and their interaction profile with what we called "the hot" protein colon cancer space (consisting of 79 proteins) we found black tea, rosemary, pomegranate and ginseng leading the list of edible plants. It would certainly be very interesting to perform a comparative study, using a

model animal system for colon cancer with edible plants that are ranked high and low in our list and verify in what degree these predictions stand true.

Actually this list can be further expanded to any other edible or non-edible plant without known association to colon cancer as long as the chemical profile of the plant is adequately defined. One of the major limitations in phenotypic screening studies is that it is practically impossible to test all foods against all disease phenotypes. However, analyses like the one performed here can lead to the identification of more foods with similar phenotypic effect based on the protein target space of their molecular components. Thus, our methodology for better delineating the prevention of human diseases by nutritional interventions relies heavily on knowing the small molecule constituents of our diet. While until recently this was a major obstacle to perform nutritional systems chemical biology studies, we have contributed significantly in this direction (Jensen et al., 2014) by developing a state-of-the art database (currently *in-house* but soon part of it will be publicly available) with information on 16,102 plants, their small molecules constituents (20,654) and the human disease phenotypes (1,592) associated with these plants. This database offers a unique platform for performing global analysis of our diet-exposome for elucidating the synergistic interactions of the small molecules that yield specific phenotypes and their protein targets and hopefully will contribute in the future towards personalized nutrition based on the disease risk of the individual.

Last but not least, we should acknowledge the limitations of our study, mainly attributed to data incompleteness in relation to the phytochemical content of plants, their therapeutic effect on diseases, as well as the activity of phytochemicals on human proteins. Even though our database contains 20,000 phytochemicals, this is still just a fraction of the natural compound space, which is estimated to be more than 150,000 compounds. Few plants have undergone a complete phytochemical profiling, while the majority has either been studied for specific compounds, if at all. In addition, the biological activity of natural compounds and plants is typically tested experimentally against few, selected proteins or disease phenotypes. Thus, the protein space and the phytochemicals identified in our study as the major players in the colon cancer interaction network, are based on the to date available information in PubMed and may be further revised in the future, as new knowledge on the medicinal properties of plants and their natural compound constituents is going to emerge.

Materials and Methods

Plant, phytochemical and protein target data

In the study of Jensen *et al.* (Jensen et al., 2014) we applied text mining and Naïve Bayes classification to assemble the plant-phytochemical and plant-disease associations. The 158 plants that were used in this study are the ones showing the highest probability (p=1) of a positive association with colon cancer. From the same *in-house* database we extracted the chemical composition of each plant (3,526 unique phytochemicals) and after standardization, by removing salts, ions and hydrogen atoms, an InChi key was generated for unique identification. Proteins forming the candidate colon cancer target space were retrieved from three different sources: (i) from the National Cancer Institute we retrieved all drugs approved by the FDA for treatment of colon cancer (http://www.cancer.gov/cancertopics/druginfo/colorectalcancer). The protein target of each drug was extracted from DrugBank database (Knox et al., 2011); (ii) from the KEGG Pathway Database (Kanehisa et al., 2012) all proteins from the colon cancer pathway (KEGG Pathway id: hsadd05210) were retrieved; (iii) the colon cancer prognostic signature gene set of 87 mRNA transcripts was taken from Oh *et al.,* (Oh et al., 2012). In addition, we included first-degree neighbors of all the proteins falling in (i) to (iii) using STRING 9.1 (Franceschini et al., 2013). In STRING each interaction is assigned a score based on evidence; here we applied a medium confidence threshold (score > 400). To ensure the biological validity of the interactions in the context of colon cancer only proteins for which there was positive evidence of expression in the colon according to the Human Protein Atlas (Uhlen et al., 2010) were included. Protein-protein interactions not derived from *Homo sapiens, Rattus norvegicus* and *Mus musculus* were removed.

Chemical-protein interactions

ChEMBL (Bento et al., 2014), a database of manually curated small molecule-protein bioactivities, quantified by a measured experimental value, was used for retrieving interactions of phytochemicals with proteins. The bioactivities were filtered according to (Kramer et al., 2012). In the present study, only K_i, IC_{50}, potency, inhibition, EC_{50} and K_d from experiments performed on proteins from *Homo sapiens, Rattus norvegicus* and *Mus musculus* were included. To accommodate for multiple measurements of the compound on the same protein, we calculated a probability (based on frequency) that the compound had an effect on the protein using the equation below:

$$P = \frac{Positive\ experiments}{All\ experiments}$$

A threshold was set as follows for the various kinds of pharmacological measurements: for Ki, EC50, IC50 and Kd, a compound was deemed to interact with the protein if the pChembl value (corresponding to the log10([M]) value) was greater than 5.5; for inhibition, a compound was deemed to interact with the target if the percentage value was greater than 20; for potency, a compound was deemed to interact with the target if the micro molar value was lower than 500μM. A single experiment was defined as "positive", i.e. the compound interacts with the protein, if the measured value was above the threshold. Only compounds for which the positive evidence outweighed the negative evidence (i.e. $P \geq 0.5$) were included for further analysis. The ChEMBL database was searched for both exact compounds using the InChI key and similar compounds using a Morgan circular based fingerprint and comparing compounds by the Tanimoto coefficient (Tc). Two compounds were deemed similar if $Tc \geq 0.85$ with a difference in molecular weight lower than 50 g/mol and were thus expected to show approximately the same behavior against the same set of proteins.

Chemical similarity between phytochemicals, drugs and metabolites of the colon metabolic network

The phytochemical space was compared to all approved drugs (retrieved from DrugBank (Knox et al., 2011) and human metabolites involved in reactions in the colon (Agren et al., 2012). For every compound, we computed a 1024 bit Morgan circular fingerprint, Molecular Weight (MW), Topological Polar Surface Area (TPSA) (Ertl et al., 2000) and Octanol/Water coefficient (SlogP) using the KNIME (Berthold et al., 2008) RDKit plugin. Using each descriptor, a matrix of 1027 columns was constructed, in which each row represented a drug, a human metabolite or a phytochemical. Each individual column was scaled to have mean = 0 and standard deviation = 1, to ensure no bias for further distance calculations. We calculated the Euclidian distance between each small molecule, and performed a classical multidimensional scaling (MDS) using the R built-in package cmdscale. Classical MDS is a dimensionality reduction technique, which aims to place objects in a lower dimensional space, keeping the between-object distance as close as possible to the original space. In this case, we choose to represent our 1027 dimensions (molecule features) in a 2 dimensional space.

Highly targeted protein space and plant efficacy

The pairwise correlation between each pair of proteins was calculated as the ϕ-coefficient. The ϕ-coefficient is a measure between -1 and 1 of correlation between two binary variables, and is related to the χ^2, as shown below:

$$\chi^2 = \phi^2 n$$

Where n is the total sample size. P-values were adjusted for multiple testing using the Bonferonni correction. Only correlations with adjusted p-value ≤ 0.05 were considered significant, however biological conclusions that rely on p-values, especially so close to the arbitrary cut-off of significance, should be interpreted with caution and the actual effect size should also be carefully examined before definitive conclusions are made.

For each plant *P*, the efficacy, *E,* of the plant was calculated as:

$$E(P) = \frac{Proteins\ within\ the\ colon\ cancer\ space\ targeted\ by\ plant\ P}{Total\ proteins\ (79)\ in\ target\ protein\ space}$$

Furthermore, we calculated a weighted efficacy, E_w, which takes into account the number of compounds targeting each protein:

$$E_w(P) = E(P) * \#Compounds$$

We scaled both the weighted and un-weighted efficacy values between 0 and 1, keeping the relative difference between plants.

Conclusion

In conclusion, by developing a systems chemical biology platform that integrates data from the scientific literature as well as online and in-house databases we revealed novel associations between dietary molecules with candidate colon cancer targets. Nevertheless, the methodology proposed here for understanding, which processes involved in the onset, incidence, progression and severity of colon cancer, are modulated by dietary components, and is applicable to any large-scale diet-disease association study, where information about the small molecule constituents of the diet is available.

Chapter V: Discovering novel anti-ovarian cancer compounds from our diet

Introduction

Ovarian cancer is the leading cause of death from gynecological disorders with an increasingly high incidence, especially in the western world (Hanna and Adams). It is among the five leading causes of death for women in developed countries (Kushi et al., 2012; La Vecchia, 2001). Epithelial ovarian cancer (EOC) comprises about 90% of all ovarian cancers and is associated with the epithelial cells on external surface of the ovary. The other 10% are stromal ovarian cancers (Schulz et al., 2004). With the diagnostic methods available today ovarian cancer is still hard to detect, thus many cases are discovered too late. We know only little about the biology of ovarian cancer. We do however know that hormonal, environmental and genetic factors have been implicated in the development. Several studies has shown that most ovarian cancers are environmental and avoidable (Kushi et al., 2012; La Vecchia, 2001).

Epidemiological studies suggests that some dietary factors may play a role in the development of ovarian cancer, so far most studies have shown up inconclusive (Bandera et al., 2009). The understanding of modifiable, contributing factors to reducing the incidence of the disease is therefore of outmost importance (Bandera et al., 2009; McCann et al., 2003). Little is known about the impact that lifestyle factors, such as diet, may have in the development of ovarian cancer (Bandera et al., 2009). It has been known for some time and shown in several epidemiologic studies that women have a significantly lower risk of developing ovarian cancer after having given birth (Hanna and Adams; Schulz et al., 2004). It is known that women in western culture are having babies much later in life then Asian women. Therefore one could speculate that the higher rate of ovarian cancer in the western world could be explained, to some extent, by women having children later in life. There are several studies that have shown a higher ovarian cancer risk for women with polycystic ovary syndrome or irregular menopause (Barry et al., 2014; Galazis et al., 2012; Smith et al., 2014). This has led to a different hypothesis; built upon the observation that phytoestrogen at doses typical of vegetarian or Asian diets influences the female menstrual cycle and could be an influencer of cancer risk. However, so far conducted clinical studies have shown considerable variation in the outcomes (Whitten and Naftolin, 1998). Phytoestrogens can change the ovarian cycle through ovarian, pituitary or hypothalamic actions. In addition, it has been shown that several bioflavonoids can affect the ovary, although the potential for direct ovarian effects may be limited (Whitten and Naftolin, 1998).

The literature on dietary components and ovarian cancer is limited. In general, reduced risks of ovarian cancer have been associated with higher intakes of vegetables (McCann et al., 2003). Especially green-leaf vegetable intake is more strongly associated with a decreased risk (Kushi et al., 1999; Zhang et al., 2002). Evidence that vegetable and fruit consumption reduces cancer risk has led to attempts to isolate the bioactive phytochemicals from these foods (Kushi et al., 2012). Phytoestrogens, bioflavonoids and lignans are known to have estrogenic as well as anti-estrogenic activity (McCann et al., 2003). However, studies on ovarian responses to phytoestrogens suggest suppression of estrogen and progesterone secretion in some cases and enhanced estradiol secretion in others (Whitten and Naftolin, 1998). This applies to most studies to date, where the effect of dietary components on ovarian cancer has been inconclusive (Chang et al., 2007). An issue for most of the studies is that their outcome is associated with relatively few phytochemicals such as isoflavones and isothiocyanates.

Many studies have concluded that even though a compound inhibits the development of many different cancer types, it may not be significantly associated with cancer of the ovary (Chang et al., 2007). The intake of isothiocyanates or foods high in isothiocyanates has not yet been significantly associated with ovarian cancer risk, neither has intake of macronutrients, antioxidant vitamins, or other micronutrients (Chang et al., 2007). For example, a study has shown that, compared with the risk for women who consumed less than 1 mg of total isoflavones per day, the relative risk of ovarian cancer associated with consumption of more than 3 mg/day was about half (Chang et al., 2007). It is however hard to determine with certainty that this difference is significant since there could be several other contributing factors (Chang et al., 2007).

Bioflavonoids such as quercetin have been shown to have anti-oxidant, anti-bacterial, anti-thrombotic, anti-inflammatory and anti-carcinogenic properties (McCann et al., 2003). For kaempferol consumption a significant 40% decrease in ovarian cancer incidence has been found for the highest compared to lowest quintile of kaempferol intake and a 34% decrease in incidence for the highest compared to lowest quintile of luteolin (Gates et al., 2007). Quercetin has also been shown to reduce the size of solid cancer cells line of the ovary, the amount required for an effect is however high -in the 2000 mg range-, and the length of treatment hard to determine (Thomasset et al., 2007).

It is hard to determine whether quercetin could be the phytochemical in the vegetarian-diet that reduces the risk of ovarian cancer. Some studies have shown that certain phytochemicals in our diet cause hormonal changes on the female menstrual cycle (Whitten and Naftolin, 1998): flaxseed lignans increased testosterone levels; soy-isoflavones decreases estrogen level together with the progesterone testerone levels; zearalenone decreases luteinizing while bourdon congeners decreases luteinizing and sex hormone binding globulin (Whitten and Naftolin, 1998).

Genistein action has been associated with precocious vaginal opening, follicles and fewer corpora luteal. However, high-doses are required for an effect (Whitten and Naftolin, 1998). The findings of clinical studies are mixed, and it is unclear whether isoflavones substantially reduce menopausal symptoms (Whitten and Naftolin, 1998). Although, several rat models have shown promising effects of phytochemicals on ovarian cancer, these findings have been hard to reproduce in humans. Human responses to phytoestrogens are highly variable in both quality and incidence of response (Whitten and Naftolin, 1998).

In the present study we search for novel phytochemicals from plant-based foods with activity against ovarian cancer, through text mining and a system-wide association of phytochemicals, foods and health benefits on human ovarian cancer, following the methodology described in our previously study (Jensen et al., 2014). In addition we evaluate the anti-ovarian cancer effects of some of the most significant compounds in a cell line study.

Methods

Prediction of phytochemicals' biological activity against ovarian cancer

Phytochemicals active against ovarian cancer were predicted using a fisher's exact test as described previously (Jensen et al., 2014). The compound composition of the plant-based foods of our diet and the health effect from their consumption were extracted from free textual data of MEDLINE abstracts. The plant – compound and plant – health benefit associations were extracted using natural language processing (Jensen et al., 2014). We then identified potential phytochemicals against ovarian cancer using a fisher's exact test and a 5% false discovery rate (Yoav and Hochberg, 1995). The likelihood of a plant compound to have biological activity against ovarian cancer is calculated based on the number of times the compound and the human health effect (in this case, ovarian cancer) are mentioned together with the same plant. Support for the predictions of bioactive compounds against ovarian cancer were sought through three different routes, as shown in Figure 23. i) We searched PubMed manually for reported bioactivities supporting the predicted health benefit. ii) We searched ChEMBL (Overington, 2009) using the structure of the compounds, for information about known protein targets linked to ovarian cancer in the therapeutic target database (Zhu et al., 2012). iii) We searched ChEMBL for compounds structurally similar to our predictions, with activity against ovarian cancer targets.

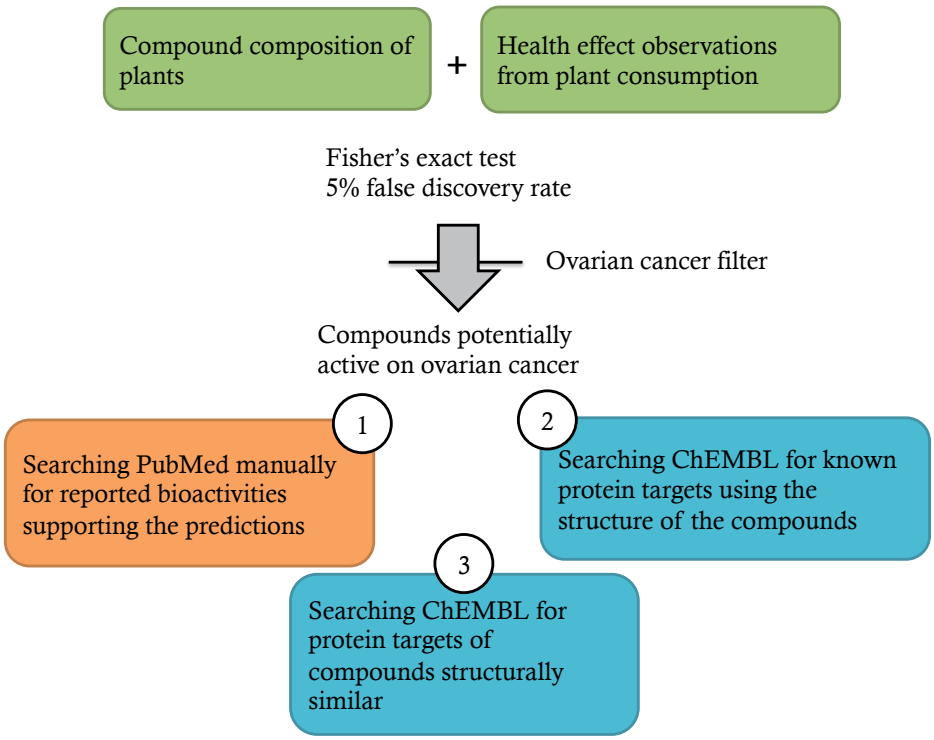


Figure 23: A flow diagram showing the discovery process and prioritization of novel anti-ovarian cancer compounds.

In-vitro evaluation of compounds

Melatonin, Protocathechuic acid, Curcumene, 6-O-alpha-L-Rhamnopyranosyl-Hyperin, 3-O-Sophorotrioside-3,4,5,7-tetrahydroxyflavone were purchased from Chengdu MUST bio-technology, China). D-(+)-Trehalose dehydrate, (-)-β-Pinene, (+)-Arabinogalactan, Abscisic acid were purchased from Sigma-Aldrich (USA). Tannic acid from our laboratory stock. All other chemical were purchased from Sigma-Aldrich (USA).

Growth inhibition effect of tested compounds on SKOV3 (ovarian cancer cell line) and HOSE6.3 (immortalized ovarian epithelial cell line) were determined by MTT assay. Briefly, 1,500 cells of each cell line were seeded into 96-well plates. After 24 hours, cells were treated with 100µl of fresh growth medium containing tested with designated concentrations for 48 hours. Cells were harvested for MTT assay. 20µl of MTT reagent containing 2.5 mg/ml MTT, and the plates were incubated for additional 4 hours to develop insoluble formazan product at 37oC. DMSO was added to each well to dissolve the formazan crystal formed in viable for spectrophotometric analysis at 570nm.

Results

Prediction of the phytochemicals' biological activity against ovarian cancer

We found several compounds with a p-value within the 5% false discovery rate. The predicted hits are divided in three categories: a) compounds that are structurally similar to a compound in ChEMBL targeting a known anti-ovarian cancer protein, b) compounds with no similar compounds in ChEMBL with such a target, c) compounds from plants used in Traditional Chinese Medicine, included in the Chinese Natural Product Database (Shen et al., 2003).

a) Compounds structurally similar to a compound with an anti-ovarian cancer target

The predicted anti-ovarian cancer activity of several of the compounds from the top 20 most significant hits could be verified in literature. Trilinolein is one compound with predicted anti-ovarian cancer activity at a low p-value (p-value < 10^{-11}). The compound is similar to fumaric acid, which interacts with the PPAR-gamma protein (P37231). Diphosphatidylglycerol has also predicted anti-ovarian cancer activity with a low p-value (p-value < 10^{-8}). This compound is similar to octanoic acid, which interacts with the LPA receptor-1 protein (Q92633).

A third example, the compound beta-sitostanol has predicted anti-ovarian cancer activity with p-value < 10^{-8}. This compound is similar to ChEMBL compound CHEMBL69891, which interacts with Cytochrome P450 19A1 (P11511). From a) group, were selected 5 compounds for experimental validation of the predicted activities shown in Table 4.

b) Compounds not similar to any compound with a known anti-ovarian cancer target

In the top 20 of most significant compounds, we find several compounds whose activities are confirmed in literature: anthocyanin (p-value < 10^{-11}), beta-glucan (p-value <10^{-9}), Saponin (p-value <10^{-3}) (Jafaar et al., 2014; Podolak et al., 2010; Schmitt and Stopper, 2001). From the two categories above we selected 10 compounds for experimental validation. From b) group, were selected 5 compounds for experimental validation of the predicted activities shown in Table 4.

c) Compounds from plants used in Traditional Chinese Medicine

We also identified several potential candidate compounds from the Chinese Natural Product Database (Shen et al., 2003) with potential anti-ovarian cancer activity. The identified compounds are shown in Table 5. Melatonin is one such compound from the Chinese Natural Product database predicted with anti-ovarian cancer activity (p-value < 10^{-10}). The compound is found in 8 plants, among others, in grape and ginkgo biloba, that are concomitantly mentioned in the literature to have anti-ovarian cancer activity, Gelseverine is another compound that we also find significantly associated with anti-ovarian cancer activity (p-value < 10^{-10}). The compound is found in 6 plants associated with anti-ovarian cancer activity (Abdul Wahab et al., 2004). Abscisic acid is another compound that we have associated with anti-ovarian cancer activity (p-value < 10^{-10}). The compound is found in 18 plants with known with anti-ovarian cancer activity (Nagle et al., 2010). The compound itself is also known to have anti-cancer activity (2010b). However, no activity on ovarian cancer has yet been reported.

Table 4 shows the compounds that are predicted associated with anti-ovarian cancer activity. The compound Arabinogalactan significantly associated with anti-ovarian cancer activity. The compound is structurally similar to the CHEMBL P16581, a compound that binds to the protein E-selectin (P16581). This compound is found in 9 foods, 3 medical plants and 1 other plant associated with ovarian cancer. The compound is proposed immune-stimulatory (Gannabathula et al., 2012). alpha,alpha-trehalose was found with anti-ovarian cancer activity. The compound is structurally similar with the compound CHEMBL461727 that binds to the protein E-selectin (P16581). The compound is found in 9 foods, 1 medical plant and 1 other plant associated with ovarian cancer.

Tannic acid was found with anti-ovarian cancer activity. The compound has features similar with CHEMBL1399656 that binds to hypoxia-inducible factor 1-alpha (Q16665). The compound is found in 12 foods, 5 medicinal plants and 5 other plants associated with anti-ovarian cancer activity. The table also shows some novel compounds, which no structural similarity with compounds has known to target proteins with anti-ovarian cancer activity. The compound beta-damascenone is one such compound found associated with anti-ovarian cancer activity (p-value < 10^{-7}). For this compound we find no structurally similarity with compound targeting known anti-ovarian cancer targets. The compound is found in 6 foods associated with anti-ovarian cancer activity. Ethyl-2-(diethoxyphosphinyl)-3-oxobutanoate (p-value < 10^{-6}) is another such compound found in 5 foods associated with anti-ovarian cancer activity. Beta-Pinene, is another compound found which associated with anti-ovarian cancer activity (p-value <10^{-3}). This compound is not

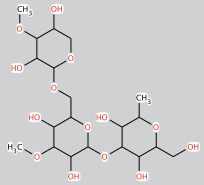
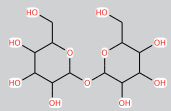
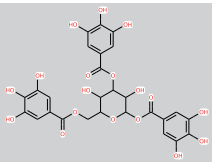
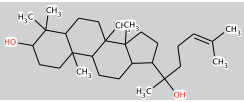
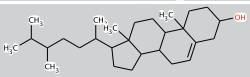
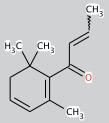
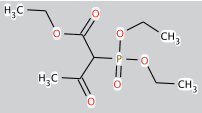
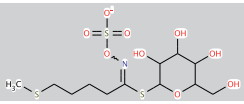
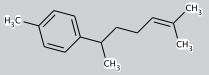
found structurally similar to compounds with known anti-ovarian cancer activity. It is found in 12 foods, 7 medicinal plants and 2 other plants associated with anti-ovarian cancer activity.

Table 5 shows the compounds associated with anti-ovarian cancer activity and with no structural similarity. Melatonin from the Chinese Natural Product database was associated with anti-ovarian cancer activity (p-value < 10^{-10}). The compound is found in 8 plants, among others, in grape and ginkgo biloba, that are concomitantly mentioned in the literature to have anti-ovarian cancer activity, Gelseverine is another compound that we also find significantly associated with anti-ovarian cancer activity (p-value < 10^{-10}). The compound is found in 6 plants associated with anti-ovarian cancer activity (Abdul Wahab et al., 2004). Abscisic acid is another compound that we have associated with anti-ovarian cancer activity (p-value < 10^{-10}). The compound is found in 18 plants with known with anti-ovarian cancer activity (Nagle et al., 2010). The compound itself is also known to have anti-cancer activity (2010b). However, no activity on ovarian cancer has yet been reported.

In vitro evaluation of compounds

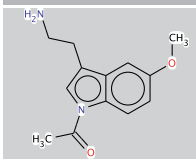
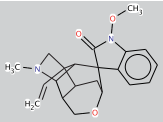
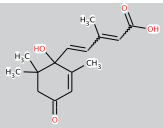
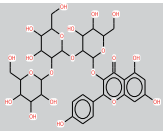
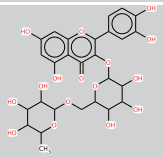
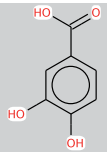
Figure 24 shows the results of the in-vitro assays. Four of the 11 compounds tested showed a lower cell viability then control (HOSE6.2). The compound Curcumene showed activity at 92,91uM (SKOV3) compared to 131.2 uM (HOSE6.3). 3-O-Sophorotrioside-3,4,5,7-tetrahydroxyflavone showed 18.88 uM (SKOV3) compared to 19.99 uM (HOSE6.3). Tannic acid showed 50.7 uM (SKOV3) compared to 98.2 uM (HOSE6.3) and Damascenone showed 32 uM (SKOV3) compared to 57.8 uM (HOSE6.3). The compounds have now been scheduled for mice studies.

Table 4: The table shows the top compounds that are associated with anti-ovarian cancer activity from group a) and b). The 5 last compounds has "No data" listed as evidence because we find no structural similarity with compounds known to target anti-ovarian cancer compounds (group b).

Name	Structure	Compound evidence	Plant evidence
Arabinogalactan		Similar structural features with CHEMBL311181** that binds to P16581	Found in 9 foods, 3 medicinal plants and 1 other plant associated with OC
alpha,alpha-trehalose		Similar structural features with CHEMBL461727** that binds to P16581	Found in 9 foods, 1 medicinal plant and 1 other plant associated with OC
Tannic acid		Similar structural features with CHEMBL1399656** that binds to Q16665	Found in 12 foods, 5 medicinal plants and 5 other plants associated with OC
Dammarenediol II		Similar structural features with CHEMBL1610940** that binds to Q16665	Found in 1 food, 4 medicinal plants and 1 other plant associated with OC
24-methylcholestero l		Similar structural features with CHEMBL69891** that binds to P11511	Found in 4 foods, 4 medicinal plants and 1 other plant associated with OC
beta-damascenone		No data	Found in 6 foods associated with OC
Ethyl-2-(diethoxyphosphinyl)-3-oxobutanoate		No data	Found in 5 foods associated with OC
beta-pinene		No data	Found in 12 foods, 7 medicinal plants and 2 other plants associated with OC
4-methylthiobutylglucosinolate		No data	Found in 1 food and 1 medicinal plant associated with OC
Curcumene		No data	Found in 1 food, 2 medicinal plants and 2 other plants associated with OC

* Search with this ID PubChem (http://pubchem.ncbi.nlm.nih.gov/) to see chemical structure
** Search with this ID ChEMBL (https://www.ebi.ac.uk/chembl/) to see chemical structure

Table 5: The table shows the compounds predicted associated with anti-ovarian cancer activity from group c).

Name	Structure	Plants (food or/and TCM) associated in the literature (PubMed) with ovarian cancer
melatonin		grape (TAXID:29760);ginkgo biloba f. epiphylla (TAXID:3311);azadirachta indica (TAXID:124943);ninjin (TAXID:4054);st. johns wort (TAXID:65561);rice (TAXID:4530);solanum tuberosum l. (TAXID:4113);olea europaea l. (TAXID:4146)
gelseverine		goetzea elegans (TAXID:70013);garovaglia elegans (TAXID:94572);goniothalamus elegans ast (TAXID:497607);gypsophila elegans m. bieb. (TAXID:146095);gomphrena elegans (TAXID:240060);gelsemium elegans (gardner and champ.) benth. (TAXID:427660)
abscisic acid		brassica rapa l. (TAXID:3711);grape (TAXID:29760);radish (TAXID:3726);brassica oleracea (TAXID:3712);ginkgo biloba f. epiphylla (TAXID:3311);ninjin (TAXID:4054);livistona chinensis (TAXID:115492);tea plant (TAXID:4442);camptotheca acuminata (TAXID:16922);salvia mihiorrhiza (TAXID:226208);dodonaea viscosa (TAXID:151065);solanum tuberosum l. (TAXID:4113);nicotiana benthamania (TAXID:4100);rice (TAXID:4530);cocos nucifera l. (TAXID:13894);pyrus malus (TAXID:3750);soya bean (TAXID:3847);olea europaea l. (TAXID:4146)
3-O-Sophorotriside-3,4',5,7-Tetrahydroxyflavone		radish (TAXID:3726);ginkgo biloba f. epiphylla (TAXID:3311);ninjin (TAXID:4054);dodonaea viscosa (TAXID:151065);lentils (TAXID:3864);soya bean (TAXID:3847);grape (TAXID:29760);acacia nilotica (TAXID:138033);curcuma domestica valeton (TAXID:136217);common oleander (TAXID:63479);tea plant (TAXID:4442);st. johns wort (TAXID:65561);cordyla madagascariensis r.vig. (TAXID:149643);pyrus malus (TAXID:3750);podophyllum hexandrum (TAXID:93608);brassica rapa l. (TAXID:3711);azadirachta indica (TAXID:124943);brassica oleracea (TAXID:3712);zingiber officinale (TAXID:94328);rice (TAXID:4530);olea europaea l. (TAXID:4146);rhodiola rosea (TAXID:203015);red stem pokeweed (TAXID:3527);pyrus balansae decne. (TAXID:23211)
6-O-alpha-L-Rhamnopyranosyl-Hyperin		radish (TAXID:3726);ginkgo biloba f. epiphylla (TAXID:3311);manioc (TAXID:3983);asparagus officinalis l. (TAXID:4686);solanum tuberosum l. (TAXID:4113);lentils (TAXID:3864);soya bean (TAXID:3847);grape (TAXID:29760);curcuma domestica valeton (TAXID:136217);common oleander (TAXID:63479);tea plant (TAXID:4442);st. johns wort (TAXID:65561);pyrus malus (TAXID:3750);podophyllum hexandrum (TAXID:93608);brassica rapa l. (TAXID:3711);azadirachta indica (TAXID:124943);brassica oleracea (TAXID:3712);zingiber officinale (TAXID:94328);platycodon grandiflorum (TAXID:94286);rice (TAXID:4530);olea europaea l. (TAXID:4146);rhodiola rosea (TAXID:203015);dodonaea viscosa (TAXID:151065);scutellaria barbata (TAXID:396367);pyrus balansae decne. (TAXID:23211)
protocathechuic acid		garden rhubarb (TAXID:3621);brassica rapa l. (TAXID:3711);soya bean (TAXID:3847);brassica oleracea (TAXID:3712);ginkgo biloba f. epiphylla (TAXID:3311);manioc (TAXID:3983);grape (TAXID:29760);st. johns wort (TAXID:65561);salvia mihiorrhiza (TAXID:226208);rice (TAXID:4530);solanum tuberosum l. (TAXID:4113);cremastra appendiculata (TAXID:459596);red stem pokeweed (TAXID:3527);cocos nucifera l. (TAXID:13894);pyrus malus (TAXID:3750);pyrus balansae decne. (TAXID:23211);olea europaea l. (TAXID:4146)

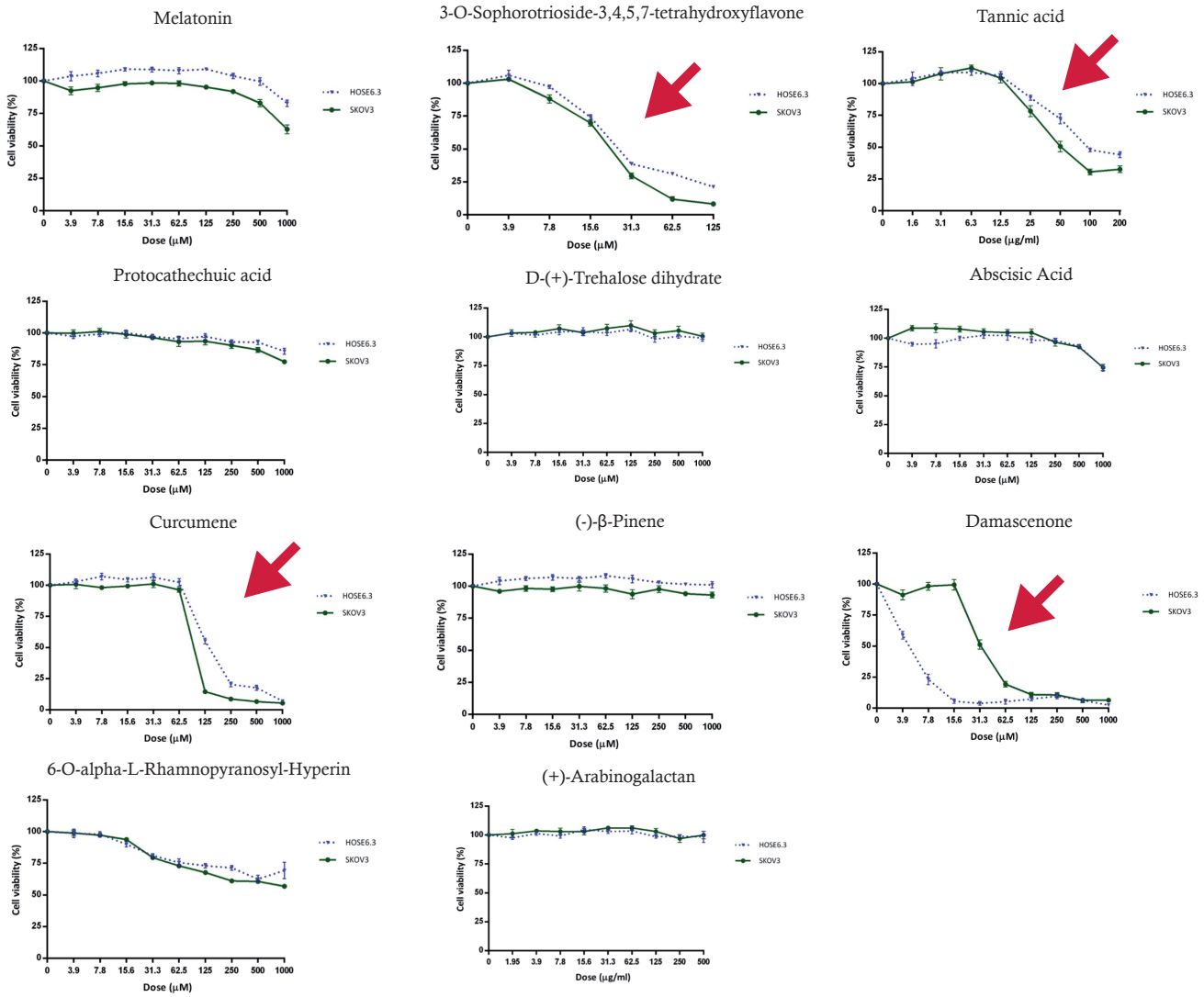


Figure 24: The figure shows the results of the in-vitro assays. Four of the 11 compounds tested showed a lower cell viability then control (HOSE6.2). The compound Curcumene showed activity at 92,91uM (SKOV3) compared to 131.2 uM (HOSE6.3). 3-O-Sophorotrioside-3,4,5,7-tetrahydroxyflavone showed 18.88 uM (SKOV3) compared to 19.99 uM (HOSE6.3). Tannic acid showed 50.7 uM (SKOV3) compared to 98.2 uM (HOSE6.3) and Damascenone showed 32 uM (SKOV3) compared to 57.8 uM (HOSE6.3).

Conclusion

We found several compounds with predicted biological activity at a p-value within the 5% false discovery rate. The predicted hits are divided in three categories: i) Compounds that are structurally similar to a compound in ChEMBL targeting a known anti-ovarian cancer protein. ii) Compounds with no similar compounds in ChEMBL with such a target. iii) Compounds from plants used in Traditional Chinese Medicine, included in the Chinese Natural Product Database. We found the compounds Arabinogalactan, alpha,alpha-trehalose and Tannic acid as compounds with potential anti-ovarian cancer activity. These compounds are structurally similar to compounds known to target anti-ovarian cancer targets. The compounds beta-damascenone, Ethyl-2-(diethoxyphosphinyl)-3-oxobutanoate and Beta-Pinene are compounds found potentially active on ovarian cancer with no structural similarity to compounds with known anti-ovarian cancer targets. From the Chinese Natural Product database we estimate that melatonin, Gelseverine and Abscisic acid with potential anti-ovarian cancer activity.

Four of the 11 compounds tested showed a lower cell viability then control (HOSE6.2). The compound Curcumene showed activity at 92,91uM (SKOV3) compared to 131.2 uM (HOSE6.3). 3-O-Sophorotrioside-3,4,5,7-tetrahydroxyflavone showed 18.88 uM (SKOV3) compared to 19.99 uM (HOSE6.3). Tannic acid showed 50.7 uM (SKOV3) compared to 98.2 uM (HOSE6.3) and Damascenone showed 32 uM (SKOV3) compared to 57.8 uM (HOSE6.3). The compounds have now been scheduled for mice studies.

Conclusion

In the first chapter we developed and described the method on which the data for the thesis was extracted from the raw textual data of MEDLINE. Food is a factor that exerts influence on human health on a daily basis. With this study we provide the molecular basis of the effect of food on health in the complete spectrum of human diseases and to suggest why and how diet and dietary molecules may represent a valuable tool to reinforce the effect of therapies and protect from relapse. Modulating the expression and the activity of enzymes, transcription factors, hormones and nuclear receptors is how food and its bioactive constituents modulate metabolic and signaling processes. Our systematized approach for connecting foods and their molecular components to diseases makes possible similar analyses as the one illustrated for colon cancer for approximately 2,300 disease phenotypes. In addition, it provides the phytochemical layer of information for nutritional systems biology studies with the aim to assess the systemic impact of food on health and make personalized nutritional recommendations.

In the second chapter we developed a systems chemical biology resource, NutriChem, to explore the medicinal value of diet. The need for a more complete assessment of the environmental factors in epidemiological studies gave birth to a new –ome, the exposome. We envisage elucidating the link between diet, molecular biological activity and diseases by developing a database source that translates the diet-exposome from concept to utility. Our methodology for better delineating the prevention of human diseases by nutritional interventions relies heavily on the availability of information related to the small molecule constituents of our diet.

In the third chapter we developed a molecular roadmap of drug-food interactions. The importance of drug-food interactions has long been recognized; a systematic approach for identifying, predicting and preventing potential interactions between food and marketed or novel drugs is not yet available. The overall objective of this work was to gain knowledge on the interference of dietary components with the pharmacokinetics and pharmacodynamics processes of medicine, with the purpose of elucidating the molecular mechanisms involved. We integrated data from the scientific literature, our in-house database, NutriChem, and online databases and using systems chemical biology approaches for the study of the effect of natural bioactive compounds from plant-based foods on proteins related to drug bioavailability and therapeutic effect. We identified phytochemicals and foods that are potentially involved in yet not documented drug-food interactions. Moreover, we identified disease areas that are most prone to the negative effects of drug-food interactions and made recommendation in relation to foods that should be avoided under certain medications.

In the fourth chapter of the thesis we proposed a framework for interrogating the critical targets in colon cancer progression and identifying plant-based dietary interventions as important modifiers of this process using a systems chemical biology approach. Our methodology for better delineating the prevention of colon cancer by nutritional interventions relies heavily on the availability of information about the small molecule constituents of our diet and it can be expanded to any other disease class that previous evidence has linked to lifestyle.

In the fifth chapter of the thesis is on the discovering of novel anti-ovarian cancer compounds. Here, we found the compounds Arabinogalactan, alpha,alpha-trehalose and Tannic acid as compounds with potential anti-ovarian cancer activity. In addition, we find beta-damascenone, Ethyl-2-(diethoxyphosphinyl)-3-oxobutanoate and Beta-Pinene potentially active on ovarian cancer. We estimate, from the Chinese Natural Product database, that Melatonin, Gelseverine and Abscisic acid could be active as well. Interestingly, 4 of the 11 compounds we tested in vitro, showed considerable lower cell viability then the control. This includes the compounds Curcumene, 3-O-Sophorotrioside-3,4,5,7-tetrahydroxyflavone, Tannic acid and Damascenone.

Future perspectives

A future perspective for this work is definitely to extent our work to cover the complete diet, i.e. to include meat, fish and their products. In this thesis we have focused on plant-based diets, due to our limited time and our limited resources. With more time and resources we should be able to extent the present work from covering plant-based diets to covering the complete human diet. It would also be interesting to investigate how diet affects our blood. Recently, interesting discoveries have been made about the role of blood in human health: young blood reverses age-related impairments in cognitive function and synaptic plasticity in mice (Villeda et al., 2014). In this perspective a better understanding of how diet affects our blood at a molecular level could be extremely interesting. It would also be interesting to investigate the molecular mechanisms underlying the relation between diets, calorie restriction and health (Testa et al., 2013).

This thesis represents current state-of-the-art nutritional systems biology. We believe our work in time will benefit people that suffer from chronic disorders and non-communicable diseases. In the thesis we demonstrate that the components of our diet function in many aspects similar to those of drugs. The concentration of the bioactive components in our diet is, however, less than when taking medicine. Diet cannot be used directly as a medical treatment but the components of our diet may however, modulate disease states and limit the need of medication. We also believe that our present work provides important mechanistic insight into the interaction of dietary components and drugs, useful for both the development of new drugs and treatment strategies.

We imagine a future where you could monitor your health condition from your blood using a nano/bio-chip. The chip could be connected with the internet through e.g. your smartphone providing health information on the go. From the knowledge developed in this thesis you could have an App that could tell you what to eat and when to eat it. This could be particularly useful for patients with chronic disorders, helping to move treatment of patients from hospitals to their homes and through diet limit the need of medication. This kind of service could also be useful during hospital treatment, as well as consist an important tool for the elderly under multiple prescription drugs, with which a wrong diet may interfere unfavorably.

Bibliography

Abdul Wahab, K., Ahmad, F.B., Din, L.B., Swee Hung, C., and Shiueh Lian, M. (2004). A Study of the in vitro cytotoxic activity of Gelsemium elegans using human ovarian and breast cancer cell lines. Trop Biomed 21, 139–144.

Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L.K., et al. (2012). KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant & Cell Physiology 53, e1.

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput. Biol. 8, e1002518.

Ali, N.A.A., Wursterb, M., Denkert, A., Arnold, N., Fadail, I., Al-Didamony, G., Lindequist, U., Wessjohann, L., and Setzer, W.N. (2012). Chemical composition, antimicrobial, antioxidant and cytotoxic activity of essential oils of Plectranthus cylindraceus and Meriandra benghalensis from Yemen. Nat Prod Commun 7, 1099–1102.

Bandera, E.V., Kushi, L.H., and Rodriguez-Rodriguez, L. (2009). Nutritional factors in ovarian cancer survival. Nutr Cancer 61, 580–586.

Barry, J.A., Azizia, M.M., and Hardiman, P.J. (2014). Risk of endometrial, ovarian and breast cancer in women with polycystic ovary syndrome: a systematic review and meta-analysis. Hum. Reprod. Update.

Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., et al. (2014). The ChEMBL bioactivity database: an update. Nucleic Acids Res. 42, D1083–1090.

Berry, M.W., and Kogan, J. (2010). Text Mining: Applications and Theory (John Wiley and Sons Ltd.).

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In Data Analysis, Machine Learning and Applications, (Springer Berlin Heidelberg), pp. 319–326.

Bolton, E.E., Wang, Y., Thiessen, P.A., and Bryant, S.H. PubChem: Integrated Platform of Small Molecules and Biological Activities. Annual Reports in Computational Chemistry 4.

Bravo, L. (1998). Polyphenols: chemistry, dietary sources, metabolism, and nutritional significance. Nutrition Reviews 56, 317–333.

Brook, R.D., Rajagopalan, S., Pope, C.A., 3rd, Brook, J.R., Bhatnagar, A., Diez-Roux, A.V., Holguin, F., Hong, Y., Luepker, R.V., Mittleman, M.A., et al. (2010). Particulate matter

air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. Circulation 121, 2331–2378.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 9, 207.

Chan, L.-N. (2013). Drug-nutrient interactions. JPEN J Parenter Enteral Nutr 37, 450–459.

Chang, E.T., Lee, V.S., Canchola, A.J., Clarke, C.A., Purdie, D.M., Reynolds, P., Anton-Culver, H., Bernstein, L., Deapen, D., Peel, D., et al. (2007). Diet and risk of ovarian cancer in the California Teachers Study cohort. Am. J. Epidemiol. 165, 802–813.

Chang, F.-R., Hayashi, K., Chua, N.-H., Kamio, S., Huang, Z.-Y., Nozaki, H., and Wu, Y.-C. (2005). The transgenic Arabidopsis plant system, pER8-GFP, as a powerful tool in searching for natural product estrogen-agonists/antagonists. J. Nat. Prod. 68, 971–973.

Chavez-Santoscoy, R.A., Gutierrez-Uribe, J.A., and Serna-Saldívar, S.O. (2009). Phenolic composition, antioxidant capacity and in vitro cancer cell cytotoxicity of nine prickly pear (Opuntia spp.) juices. Plant Foods Hum Nutr 64, 146–152.

Chen, X., Ji, Z.L., and Chen, Y.Z. (2002). TTD: Therapeutic Target Database. Nucleic Acids Res. 30, 412–415.

Cimino, J.J. (2001). Terminology tools: state of the art and practical lessons. Methods Inf Med 40, 298–306.

Cimino, J.J., and Zhu, X. (2006). The practical impact of ontologies on biomedical informatics. Yearb Med Inform 124–135.

Colombo, M., and Bosisio, E. (1996). Pharmacological activities of Chelidonium majus L (Papaveraceae). Pharmacological Research 33, 127–134.

Cowan, M. (1999). Plant products as antimicrobial agents. Clinical Microbiology Reviews 12, 564+.

Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A., and Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J. Chem. Inf. Comput. Sci. 32, 244–255.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344–350.

Demetriades, A.K., Wallman, P.D., McGuiness, A., and Gavalas, M.C. (2005). Low cost, high risk: accidental nutmeg intoxication. Emerg Med J 22, 223–225.

Dinicola, S., Cucina, A., Pasqualato, A., D'Anselmi, F., Proietti, S., Lisi, E., Pasqua, G., Antonacci, D., and Bizzarri, M. (2012). Antiproliferative and Apoptotic Effects Triggered by

Grape Seed Extract (GSE) versus Epigallocatechin and Procyanidins on Colon Cancer Cell Lines. Int J Mol Sci 13, 651–664.

Durant, J.L., Leland, B.A., Henry, D.R., and Nourse, J.G. (2002). Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42, 1273–1280.

Ertl, P., Rohde, B., and Selzer, P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J. Med. Chem. 43, 3714–3717.

Al-Fayez, M., Cai, H., Tunstall, R., Steward, W.P., and Gescher, A.J. (2006). Differential modulation of cyclooxygenase-mediated prostaglandin production by the putative cancer chemopreventive flavonoids tricin, apigenin and quercetin. Cancer Chemother. Pharmacol. 58, 816–825.

Ferguson, L.R., and Schlothauer, R.C. (2012). The potential role of nutritional genomics tools in validating high health foods for cancer control: broccoli as example. Mol Nutr Food Res 56, 126–146.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. Nucleic Acids Research 42, D749–D755.

Formisano, C., Rigano, D., Senatore, F., Piozzi, F., and Arnold, N.A. (2011). Analysis of essential oils from Scutellaria orientalis ssp. alpina and S. utriculata by GC and GC-MS. Nat Prod Commun 6, 1347–1350.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41, D808–815.

Frédérich, M., Marcowycz, A., Cieckiewicz, E., Mégalizzi, V., Angenot, L., and Kiss, R. (2009). In vitro anticancer potential of tree extracts from the Walloon Region forest. Planta Med. 75, 1634–1637.

Fu, W.J., Stromberg, A.J., Viele, K., Carroll, R.J., and Wu, G. (2010). Statistics and bioinformatics in nutritional sciences: analysis of complex data in the era of systems biology. J. Nutr. Biochem. 21, 561–572.

Galazis, N., Olaleye, O., Haoula, Z., Layfield, R., and Atiomo, W. (2012). Proteomic biomarkers for ovarian cancer risk in women with polycystic ovary syndrome: a systematic review and biomarker database integration. Fertil. Steril. 98, 1590–1601.e1.

Gannabathula, S., Skinner, M.A., Rosendale, D., Greenwood, J.M., Mutukumira, A.N., Steinhorn, G., Stephens, J., Krissansen, G.W., and Schlothauer, R.C. (2012). Arabinogalactan proteins contribute to the immunostimulatory properties of New Zealand honeys. Immunopharmacol Immunotoxicol 34, 598–607.

Gates, M.A., Tworoger, S.S., Hecht, J.L., De Vivo, I., Rosner, B., and Hankinson, S.E. (2007). A prospective study of dietary flavonoid intake and incidence of epithelial ovarian cancer. Int. J. Cancer 121, 2225–2232.

Gershenzon, J., and Dudareva, N. (2007). The function of terpene natural products in the natural world. Nature Chemical Biology 3, 408–414.

Goble, C., Bachhofer, S., Carr, L., De Roure, D., and Hall, W. Conceptual Open Hypermedia = The Semantic Web? (Hong Kong),.

Guarino, N., Masolo, C., and Vetere, G. (1999). OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems 14, 70–80.

Hambly, R.J., Saunders, M., Rijken, P.J., and Rowland, I.R. (2002). Influence of dietary components associated with high or low risk of colon cancer on apoptosis in the rat colon. Food Chem. Toxicol. 40, 801–808.

Hanna, L., and Adams, M. Prevention of ovarian cancer. Best Practice & Research Clinical Obstetrics and Gynaecology 20, 339–362.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., et al. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 41, D456–463.

Heinrich, J. (2011). Influence of indoor factors in dwellings on the development of childhood asthma. Int J Hyg Environ Health 214, 1–25.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). InChI - the worldwide chemical structure identifier standard. J Cheminform 5, 7.

Herrero, M. (2012). Foodomics: MS-based strategies in modern food science and nutrition. Mass Spectrometry 49–69.

Hettne, K.M., Stierum, R.H., Schuemie, M.J., Hendriksen, P.J.M., Schijvenaars, B.J.A., Mulligen, E.M. van, Kleinjans, J., and Kors, J.A. (2009). A dictionary to identify small molecules and drugs in free text. Bioinformatics 25, 2983–2991.

Hsu, P.P., and Sabatini, D.M. (2008). Cancer cell metabolism: Warburg and beyond. Cell 134, 703–707.

Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K., Cantley, L.C., Vander Heiden, M.G., and Vitkup, D. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nat. Biotechnol. 31, 522–529.

Hu, Y.-M., Ye, W.-C., Yin, Z.-Q., and Zhao, S.-X. (2007). [Chemical constituents from flos Sesamum indicum L]. Yao Xue Xue Bao 42, 286–291.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 40, D306–312.

Hutter, C.M., Chang-Claude, J., Slattery, M.L., Pflugeisen, B.M., Lin, Y., Duggan, D., Nan, H., Lemire, M., Rangrej, J., Figueiredo, J.C., et al. (2012). Characterization of gene-environment interactions for colorectal cancer susceptibility loci. Cancer Res. 72, 2036–2044.

International Conference on Information and Knowledge Engineering, A., Hamid R, and IKE '04, I.C. on I. and K.E. (2004). Proceedings of the International Conference on Information and Knowledge Engineering, IKE '04: Las Vegas, Nevada, USA, June 21 - 24, 2004 ([S.l.: CSREA Press).

Isaac, R.E., and Rees, H.H. (1984). Isolation and identification of ecdysteroid phosphates and acetylecdysteroid phosphates from developing eggs of the locust, Schistocerca gregaria. The Biochemical Journal 221, 459–464.

Itoh, A., Kumashiro, T., Yamaguchi, M., Nagakura, N., Mizushina, Y., Nishi, T., and Tanahashi, T. (2005). Indole alkaloids and other constituents of Rauwolfia serpentina. J. Nat. Prod. 68, 848–852.

Jafaar, Z.M.T., Litchfield, L.M., Ivanova, M.M., Radde, B.N., Al-Rayyan, N., and Klinge, C.M. (2014). □-D-glucan inhibits endocrine-resistant breast cancer cell proliferation and alters gene expression. Int. J. Oncol. 44, 1365–1375.

Jain, V., and Raut, D.K. (2011). Medical literature search dot com. Indian J Dermatol Venereol Leprol 77, 135–140.

Jensen, K., Panagiotou, G., and Kouskoumvekaki, I. (2014). Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level. PLOS Computational Biology.

Juan, M.E., Wenzel, U., Ruiz-Gutierrez, V., Daniel, H., and Planas, J.M. (2006). Olive fruit extracts inhibit proliferation and induce apoptosis in HT-29 human colon cancer cells. J. Nutr. 136, 2553–2557.

Kahan, J., Koivunen, M.-R., and Prud'Hommeaus, E. (2001). Annotea: An Open RDF Infrastructure for Shared Web Annotations. (Hong Kong),.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–114.

Kaput, J. (2008). Nutrigenomics research for personalized nutrition and medicine. Curr. Opin. Biotechnol. 19, 110–120.

Katada, S., Imhof, A., and Sassone-Corsi, P. (2012). Connecting threads: epigenetics and metabolism. Cell 148, 24–28.

Kim, S., and Wilbur, W.J. (2011). Classifying protein-protein interaction articles using word and syntactic features. BMC Bioinformatics 12 Suppl 8, S9.

Kim, S., Shin, S.-Y., Lee, I.-H., Kim, S.-J., Sriram, R., and Zhang, B.-T. (2008). PIE: an online prediction system for protein-protein interactions from text. Nucleic Acids Res. 36, W411–415.

Kim, S., Kwon, D., Shin, S.-Y., and Wilbur, W.J. (2012). PIE the search: searching PubMed literature for protein interaction information. Bioinformatics 28, 597–598.

Knekt, P., Kumpulainen, J., Järvinen, R., Rissanen, H., Heliövaara, M., Reunanen, A., Hakulinen, T., and Aromaa, A. (2002). Flavonoid intake and risk of chronic diseases. The American Journal of Clinical Nutrition 76, 560–568.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011). DrugBank 3.0: a comprehensive resource for “omics” research on drugs. Nucleic Acids Res. 39, D1035–1041.

Krallinger, M., Erhardt, R.A.-A., and Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. Drug Discov. Today 10, 439–445.

Kramer, C., Kalliokoski, T., Gedeck, P., and Vulpetti, A. (2012). The Experimental Uncertainty of Heterogeneous Public K i Data. Journal of Medicinal Chemistry 55, 5165–5173.

Kuhn, M., Szklarczyk, D., Franceschini, A., Campillos, M., von Mering, C., Jensen, L.J., Beyer, A., and Bork, P. (2010). STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Research 38, D552–6.

Kusari, S., Zühlke, S., and Spiteller, M. (2011). Correlations between camptothecin and related metabolites in Camptotheca acuminata reveal similar biosynthetic principles and in planta synergistic effects. Fitoterapia 82, 497–507.

Kushi, L.H., Mink, P.J., Folsom, A.R., Anderson, K.E., Zheng, W., Lazovich, D., and Sellers, T.A. (1999). Prospective study of diet and ovarian cancer. Am. J. Epidemiol. 149, 21–31.

Kushi, L.H., Doyle, C., McCullough, M., Rock, C.L., Demark-Wahnefried, W., Bandera, E.V., Gapstur, S., Patel, A.V., Andrews, K., Gansler, T., et al. (2012). American Cancer Society Guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. CA Cancer J Clin 62, 30–67.

Kussmann, M., and Fay, L.B. (2008). Nutrigenomics and personalized nutrition: science and concept. Personalized Medicine 5, 447–455.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313, 1929–1935.

Leonardi, T., Vanamala, J., Taddeo, S.S., Davidson, L.A., Murphy, M.E., Patil, B.S., Wang, N., Carroll, R.J., Chapkin, R.S., Lupton, J.R., et al. (2010). Apigenin and naringenin suppress colon carcinogenesis through the aberrant crypt stage in azoxymethane-treated rats. Exp. Biol. Med. (Maywood) 235, 710–717.

LePendu, P., Musen, M.A., and Shah, N.H. (2011). Enabling enrichment analysis with the Human Disease Ontology. J Biomed Inform 44 Suppl 1, S31–38.

Madhujith, T., and Shahidi, F. (2007). Antioxidative and antiproliferative properties of selected barley (Hordeum vulgarae L.) cultivars and their potential for inhibition of low-density lipoprotein (LDL) cholesterol oxidation. J. Agric. Food Chem. 55, 5018–5024.

Manson, M.M. (2003). Cancer prevention -- the potential for diet to modulate molecular signalling. Trends Mol Med 9, 11–18.

Martucci, W.E., Udier-Blagovic, M., Atreya, C., Babatunde, O., Vargo, M.A., Jorgensen, W.L., and Anderson, K.S. (2009). Novel non-active site inhibitor of Cryptosporidium hominis TS-DHFR identified by a virtual screen. Bioorg. Med. Chem. Lett. 19, 418–423.

Mas, S., Crescenti, A., Gassó, P., Deulofeu, R., Molina, R., Ballesta, A., Kensler, T.W., and Lafuente, A. (2007). Induction of apoptosis in HT-29 cells by extracts from isothiocyanates-rich varieties of Brassica oleracea. Nutr Cancer 58, 107–114.

McCann, S.E., Freudenheim, J.L., Marshall, J.R., and Graham, S. (2003). Risk of human ovarian cancer is related to dietary intake of selected nutrients, phytochemicals and food groups. J. Nutr. 133, 1937–1942.

McEntyre, J., and Ostell, J. (2003). The NCBI Handbook.

McGuinness, D.L. (1998). Ontological Issues for Knowledge-Enhanced Search. In Proc. of the Formal Ontology in Information Systems.

Mensinga, T.T., Sips, A.J.A.M., Rompelberg, C.J.M., van Twillert, K., Meulenbelt, J., van den Top, H.J., and van Egmond, H.P. (2005). Potato glycoalkaloids and adverse effects in humans: an ascending dose study. Regul. Toxicol. Pharmacol. 41, 66–72.

Miller, J.C., Gutowski, G.E., Poore, G.A., and Boder, G.B. (1977). Alkaloids of Vinca rosea L. (Catharanthus roseus G. Don). 38. 4'-Dehydrated derivatives. J. Med. Chem. 20, 409–413.

Misaka, S., Yatabe, J., Müller, F., Takano, K., Kawabe, K., Glaeser, H., Yatabe, M.S., Onoue, S., Werba, J.P., Watanabe, H., et al. (2014). Green tea ingestion greatly reduces plasma concentrations of nadolol in healthy subjects. Clin. Pharmacol. Ther. 95, 432–438.

Nagle, C.M., Olsen, C.M., Bain, C.J., Whiteman, D.C., Green, A.C., and Webb, P.M. (2010). Tea consumption and risk of ovarian cancer. Cancer Causes Control 21, 1485–1491.

Nanri, H., Nakamura, K., Hara, M., Higaki, Y., Imaizumi, T., Taguchi, N., Sakamoto, T., Horita, M., Shinchi, K., and Tanaka, K. (2011). Association between dietary pattern and serum C-reactive protein in Japanese men and women. J Epidemiol 21, 122–131.

Neveu, V., Perez-Jiménez, J., Vos, F., Crespy, V., du Chaffaut, L., Mennen, L., Knox, C., Eisner, R., Cruz, J., Wishart, D., et al. (2010). Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. Database : The Journal of Biological Databases and Curation 2010, bap024.

Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J., and Ananiadou, S. (2008). Kleio: a knowledge-enriched information retrieval system for biology. pp. 787–788.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: An open chemical toolbox. J Cheminform 3, 33.

Oh, S.C., Park, Y.-Y., Park, E.S., Lim, J.Y., Kim, S.M., Kim, S.-B., Kim, J., Kim, S.C., Chu, I.-S., Smith, J.J., et al. (2012). Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. Gut 61, 1291–1298.

Oprea, T.I., Tropsha, A., Faulon, J., and Rintoul, M.D. (2007). Systems Chemical Biology. Nature Chemical Biology 3, 447–450.

Overington, J. (2009). ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. Journal of Computer-Aided Molecular Design 23, 195–198.

Ovsiannikov, I.., Arbib, M.., and Mcneill, T.H. (1999). Annotation Technology. International Journal of Human-Computer Studies 50, 329–362.

Pafilis, E., O'Donoghue, S.I., Jensen, L.J., Horn, H., Kuhn, M., Brown, N.P., and Schneider, R. (2009). Reflect: augmented browsing for the life scientist. Nat. Biotechnol. 27, 508–510.

Panagiotou, G., and Nielsen, J. (2009). Nutritional systems biology: definitions and approaches. Annual Review of Nutrition 29, 329–339.

Pandi-Perumal, S.R., Srinivasan, V., Maestroni, G.J.M., Cardinali, D.P., Poeggeler, B., and Hardeland, R. (2006). Melatonin - Nature's most versatile biological signal. FEBS Journal 273, 2813–2838.

Paoloni-Giacobino, A. (2011). Post genomic decade--the epigenome and exposome challenges. Swiss Med Wkly 141, w13321.

Perkins, J. (2010). Python Text Processing with NLTK 2.0 Cookbook (Packt Publishing Ltd.).

Podolak, I., Galanty, A., and Sobolewska, D. (2010). Saponins as cytotoxic agents: a review. Phytochemistry Reviews 9, 425–474.

Polur, H., Joshi, T., Workman, C.T., Lavekar, G., and Kouskoumvekaki, I. (2011). Back to the Roots: Prediction of Biologically Active Natural Products from Ayurveda Traditional Medicine. Molecular Informatics 30, 181–187.

Pritchard, S.J., and Weightman, A.L. (2005). MEDLINE in the UK: pioneering the past, present and future. Health Info Libr J 22 Suppl 1, 38–44.

Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. Nat. Rev. Genet. 12, 529–541.

Rappaport, S.M. (2012). Discovering environmental causes of disease. J Epidemiol Community Health 66, 99–102.

Rappaport, S.M., and Smith, M.T. (2010). Environment and Disease Risks. Science 330, 460–461.

Richardson, M.A., Sanders, T., Palmer, J.L., Greisinger, A., and Singletary, S.E. (2000). Complementary/alternative medicine use in a comprehensive cancer center and the implications for oncology. J. Clin. Oncol. 18, 2505–2514.

Rocktäschel, T., Weidlich, M., and Leser, U. (2012). ChemSpot: A Hybrid System for Chemical Named Entity Recognition. Bioinformatics (Oxford, England) 1–9.

Rodriguez-Esteban, R., Iossifov, I., and Rzhetsky, A. (2006). Imitating manual curation of text-mined facts in biomedicine. PLoS Comput. Biol. 2, e118.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., et al. (2013). ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Res. 41, D987–990.

Sacco, S.M., Chen, J., Power, K.A., Ward, W.E., and Thompson, L.U. (2008). Lignan-rich sesame seed negates the tumor-inhibitory effect of tamoxifen but maintains bone health in a postmenopausal athymic mouse model with estrogen-responsive breast tumors. Menopause 15, 171–179.

Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C., Detmer, D.E., and Expert Panel (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 14, 1–9.

Sánchez-Hernández, L., Nozal, L., Marina, M.L., and Crego, A.L. (2012). Determination of nonprotein amino acids and betaines in vegetable oils by flow injection triple-quadrupole

tandem mass spectrometry: a screening method for the detection of adulterations of olive oils. J. Agric. Food Chem. 60, 896–903.

Santos, E.O., Lima, L.S., David, J.M., Martins, L.C., Guedes, M.L.S., and David, J.P. (2011). Podophyllotoxin and other aryltetralin lignans from Eriope latifolia and Eriope blanchetii. Nat. Prod. Res. 25, 1450–1453.

Scalbert, A., Manach, C., Morand, C., Rémésy, C., and Jiménez, L. (2005). Dietary polyphenols and the prevention of diseases. Critical Reviews in Food Science and Nutrition 45, 287–306.

Schadt, E.E., Bjorkegren, and M, J.L. (2012). NEW: Network-Enabled Wisdom in Biology, Medicine, and Health Care. Science Translational Medicine 4.

Schmidt, L.E., and Dalhoff, K. (2002). Food-drug interactions. Drugs 62, 1481–1502.

Schmitt, E., and Stopper, H. (2001). Estrogenic activity of naturally occurring anthocyanidins. Nutr Cancer 41, 145–149.

Schulz, M., Lahmann, P.H., Riboli, E., and Boeing, H. (2004). Dietary determinants of epithelial ovarian cancer: a review of the epidemiologic literature. Nutr Cancer 50, 120–140.

Seden, K., Dickinson, L., Khoo, S., and Back, D. (2010). Grapefruit-drug interactions. Drugs 70, 2373–2407.

Sharma, V., and Sarkar, I.N. (2013). Bioinformatics opportunities for identification and study of medicinal plants. Brief. Bioinformatics 14, 238–250.

Shen, J., Xu, X., Cheng, F., Liu, H., Luo, X., Chen, K., Zhao, W., Shen, X., and Jiang, H. (2003). Virtual screening on natural products for discovering active compounds and target information. Current Medicinal Chemistry 10, 2327–2342.

Shiels, P.G., McGlynn, L.M., MacIntyre, A., Johnson, P.C.D., Batty, G.D., Burns, H., Cavanagh, J., Deans, K.A., Ford, I., McConnachie, A., et al. (2011). Accelerated telomere attrition is associated with relative household income, diet and inflammation in the pSoBid cohort. PLoS ONE 6, e22521.

Singh, D., Gupta, R., and Saraf, S.A. (2012). Herbs-are they safe enough? an overview. Crit Rev Food Sci Nutr 52, 876–898.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech 25, 1251–1255.

Smith, E.R., Wang, Y., and Xu, X.-X. (2014). Development of a Mouse Model of Menopausal Ovarian Cancer. Front Oncol 4, 36.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3, Article3.

Sparber, A., Wootton, J.C., Bauer, L., Curt, G., Eisenberg, D., Levin, T., and Steinberg, S.M. (2000). Use of complementary medicine by adult patients participating in HIV/AIDS clinical trials. J Altern Complement Med 6, 415–422.

Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. Brief. Bioinformatics 6, 239–251.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research 39, D561–8.

Tammariello, A.E., and Milner, J.A. (2010). Mouse models for unraveling the importance of diet in colon cancer prevention. J. Nutr. Biochem. 21, 77–88.

Temple, N.J., and el-Khatib, S.M. (1987). Cabbage and vitamin E: their effect on colon tumor formation in mice. Cancer Lett. 35, 71–77.

Terry, P., Giovannucci, E., Michels, K.B., Bergkvist, L., Hansen, H., Holmberg, L., and Wolk, A. (2001). Fruit, vegetables, dietary fiber, and risk of colorectal cancer. J. Natl. Cancer Inst. 93, 525–533.

Testa, G., Biasi, F., Poli, G., and Chiarpotto, E. (2013). Calorie Restriction and Dietary Restriction Mimetics: a Strategy for Improving Healthy Aging and Longevity. Curr. Pharm. Des.

Thomasset, S.C., Berry, D.P., Garcea, G., Marczylo, T., Steward, W.P., and Gescher, A.J. (2007). Dietary polyphenolic phytochemicals--promising cancer chemopreventive agents in humans? A review of their clinical properties. Int. J. Cancer 120, 451–458.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. 28, 1248–1250.

UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 42, D191–198.

Vargas-Vera, M., Motta, E., and Domingue, J. (2002). MnM: Ontology Driven Semiautomatic and Automatic Support for Semantic Markup. (Spain),.

La Vecchia, C. (2001). Epidemiology of ovarian cancer: a summary review. Eur. J. Cancer Prev. 10, 125–129.

Villeda, S.A., Plambeck, K.E., Middeldorp, J., Castellano, J.M., Mosher, K.I., Luo, J., Smith, L.K., Bieri, G., Lin, K., Berdnik, D., et al. (2014). Young blood reverses age-related impairments in cognitive function and synaptic plasticity in mice. Nature Medicine.

Vineis, P., Khan, A.E., Vlaanderen, J., and Vermeulen, R. (2009). The impact of new research technologies on our understanding of environmental causes of disease: the concept of clinical vulnerability. Environ Health 8, 54.

Wedick, N., Pan, A., Cassidy, A., Rimm, E., Sampson, L., Rosner, B., Willet, W., Hu, F., Sun, Q., and van Dam, R. (2012). Dietary flavonoid intakes and risk of type 2 diabetes in US men and. The American Journal of Clinical Nutrition 95, 925–933.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36.

Whitten, P.L., and Naftolin, F. (1998). Reproductive actions of phytoestrogens. Baillieres Clin. Endocrinol. Metab. 12, 667–690.

Wild, C.P. (2011). Future research perspectives on environment and health: the requirement for a more expansive concept of translational cancer research. Environ Health 10 Suppl 1, S15.

Williams, A.J., Ekins, S., and Tkachenko, V. (2012). Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discovery Today 17, 685–701.

Wink, M. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. Phytochemistry 64, 3–19.

Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34, D668–672.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 36, D901–906.

Wu, H.C., Luk, R.W.P., Wong, K.F., and Kwok, K.L. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems 26, 1–37.

Yamreudeewong, W., Henann, N.E., Fazio, A., Lower, D.L., and Cassidy, T.G. (1995). Drug-food interactions in clinical practice. J Fam Pract 40, 376–384.

Yoav, B., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57, 289–300.

Zhang, H. (2004). The Optimality of Naive Bayes.

Zhang, M., Yang, Z.Y., Binns, C.W., and Lee, A.H. (2002). Diet and ovarian cancer risk: a case-control study in China. Br. J. Cancer 86, 712–717.

Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. (2004a). Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 20, 1178–1190.

Zhou, S., Lim, L.Y., and Chowbay, B. (2004b). Herbal modulation of P-glycoprotein. Drug Metab. Rev. 36, 57–104.

Zhu, F., Han, B., Kumar, P., Liu, X., Ma, X., Wei, X., Huang, L., Guo, Y., Han, L., Zheng, C., et al. (2010). Update of TTD: Therapeutic Target Database. Nucleic Acids Res. 38, D787–791.

Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., Zhang, L., Song, Y., Liu, X., Zhang, J., et al. (2012). Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Research 40, D1128–36.

(1982). MEDLARS and Health Information Policy.

(2010a). Entrez Programming Utilities Help.

(2010b). How natural drug, abscisic acid, fights inflammation. Science Daily.

(2014). PubMed Help.