JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

REVIEW

pubs.acs.org/JAFC

Databases on Food Phytochemicals and Their Health-Promoting Effects

Augustin Scalbert,*[,†] Cristina Andres-Lacueva,[§] Masanori Arita,[#] Paul Kroon,[⊥] Claudine Manach,[⊗] Mireia Urpi-Sarda,[§] and David Wishart[△]

[†]Nutrition and Metabolism Section, Biomarkers Group, International Agency for Research on Cancer (IARC), 150 cours Albert Thomas, F-69372 Lyon Cedex 08, France

[§]Nutrition and Food Science Department, XaRTA INSA, INGENIO—CONSOLIDER Program, Fun-C-Food CSD2007-063/AGL200913906-C02-01, Pharmacy School, University of Barcelona, Avinguda Joan XXIII s/n, 08028 Barcelona, Spain

[#]RIKEN Plant Science Center and Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, 113-0033 Tokyo, Japan

[⊥]Institute of Food Research, Colney Lane, NR4 7UA Norwich, United Kingdom

[⊗]INRA, Centre de Recherche de Clermont-Ferrand/Theix, and Université Clermont 1, UFR Médecine, UMR1019, Unité de Nutrition Humaine, 63122 Saint-Genès-Champanelle, France

[△]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2E8

ABSTRACT: Considerable information on the chemistry and biological properties of dietary phytochemicals has accumulated over the past three decades. The scattering of the data in tens of thousands publications and the diversity of experimental approaches and reporting formats all make the exploitation of this information very difficult. Some of the data have been collected and stored in electronic databases so that they can be automatically updated and retrieved. These databases will be particularly important in the evaluation of the effects on health of phytochemicals and in facilitating the exploitation of nutrigenomic data. The content of over 50 databases on chemical structures, spectra, metabolic pathways in plants, occurrence and concentrations in foods, metabolism in humans and animals, biological properties, and effects on health or surrogate markers of health is reviewed. Limits of these databases are emphasized, and needs and recommendations for future developments are underscored. More investments in the construction of databases on phytochemicals and their effects on health are clearly needed. They should greatly contribute to the success of future research in this field.

KEYWORDS: phytochemicals, foods, metabolism, health, databases, bioinformatics, nutrigenomics

INTRODUCTION

The composition of foods cannot be reduced to the sum of macronutrients and the 40 or so essential micronutrients they contain. Foods also contain a large number of other compounds that, although not essential, also influence health: Some can be toxic, others are thought to be beneficial for health. In particular, several hundreds of phytochemicals such as polyphenols, carotenoids, glucosinolates, phytates, saponins, amines, or alkaloids have been identified in foods of plant origin. Some of these compounds may contribute to explain the beneficial health effects of the consumption of fruits and vegetables or whole grain cereals. Understanding their role in nutrition is a major challenge for the nutritionists of the 21st century.[1] It requires full knowledge on their chemistry, occurrence in foods, metabolism and bioavailability, biological properties, and effects on health or surrogate markers of health. None of this information should be ignored when their role in nutrition is evaluated.

The volume of information, the diversity of experimental approaches and methods, the diversity of reporting formats, and the scattering of the information in tens of thousands publications all make the exploitation of this information very difficult. Furthermore, phytochemicals are not present in isolation in foods. Their properties very much depend on complex interactions within the food matrix and with various targets in the human body. Nutrigenomic approaches able to simultaneously characterize the effects of phytochemicals on a large number of genes, proteins, or metabolites appear particularly adapted to the exploration of health effects of phytochemicals.[2,3] Furthermore, metabolomics should also allow the simultaneous measurement of exposure to a large number of dietary phytochemicals.[4−6]

The capacity for biologists and chemists to generate gigabytes of information on a daily basis is having a profound impact on the way that scientific information is being stored or delivered. Whereas most scientific data are still presented in scientific journals and the majority of high-level scientific knowledge is still published in textbooks, it is becoming increasingly obvious that today's publishing industry cannot keep up with the pace of scientific advancement and the quantity of data that the scientific community would like to publish. These publishing bottlenecks are beginning to be cleared through the introduction of a new and very important kind of scientific archive: the database.

Received: February 10, 2011
Revised: March 20, 2011
Accepted: March 28, 2011
Published: March 28, 2011

Table 1. Key Data Fields in the "Perfect" Nutrient/Phytochemical Database

data category	specific data content
nomenclature	chemical name, common name, synonyms, IUPAC name, InChI, CAS Registry No., other database identifiers
description	text description of compound covering history, utility, discovery, biological role
structure	structure image, Mol file, SDF file, SMILES strings, chemical formula
chemical class or ontology	chemical kingdom, class, family, subclass, or related ontology
physicochemical data	molecular weight, LogP, pK_a, water solubility, IR spectra, NMR spectra, EI-MS spectra, GC indices, MS/MS spectra
taxonomy/origin	genus, species, and common names of plant(s) or organism(s) of origin
physiological effect	role in human nutrition, health, physiology, disease prevention or mitigation, test concentrations
health studies and claims	references to preclinical and clinical trial studies, synopsis of claims, tested cell lines or organisms, assays, test concentrations, sample number, significance
protein target(s)	names, protein sequences, gene sequences, gene location, functions, gene ontology of human targets
biosynthesis/synthesis	pathways, descriptions, enzymes, starting compounds associated with biosynthesis or organic synthesis
source content/concentration	concentration or abundance in different plant parts, list of known plants or food sources containing compound
metabolism	pathways, descriptions, and enzymes associated with human metabolism and elimination
metabolites	names, chemical formulas, and structures of known human metabolites
human content/concentration	concentration or abundance of compound (and known metabolites) in different biofluids and tissues

Simply stated, a database is a repository of data. More formally, a database is defined as a consolidated, integrated collection of conceptually related data records covering one or more subject areas. The data in a database can consist of text, numbers, images, or combinations of all three data types. Databases come in many different formats and sizes; they may be small (a few hand-written pages stuck in a file folder) or large (thousands of terabytes stored on large disk drives). Obviously, most of today's scientific databases are electronic. Electronic databases typically consist of software-based "containers" that are designed to collect and store data so that users can automatically retrieve, add, update, or delete data.

Databases tend to fall into two main categories: (1) archival or (2) curated. Archival databases are designed to capture all data of a certain type, regardless of its quality, redundancy, or utility, much like a security camera captures random images at predefined time intervals. Often, archival databases consist of large quantities of machine-processed data of questionable quality provided by many contributors. Examples of well-known archival databases in the life sciences include PubChem,[7] GenBank,[8] the Gene Expression Omnibus,[9] and the Protein Data Bank.[10] On the other hand, curated databases are designed to capture high-quality data entered and vetted by a knowledgable curator or curatorial staff, much like a museum acquires high-quality items based on expert suggestions and evaluations. Most curated databases consist of modest quantities of high-quality, manually extracted or measured data. Examples of curated databases in the life sciences include MassBank,[11] KEGG,[12] UniProt,[13] and HMDB.[14]

Life science databases may contain a variety of scientifically relevant data including sequence, structure, function, taxonomy, nomenclature, physicochemical property, concentration, or any combination of the just-mentioned data types. Within the field of nutrition and phytochemical research, there are very specific needs for certain types of data. Table 1 provides a list of the data fields and data types that should ideally exist in a nutrition/phytochemical database. These include descriptive (i.e., biological properties), chemical, structural, spectral, nomenclature, methodological, taxonomic, and composition data. The chemical, structural, nomenclature, methodological, and spectral information is particularly important for analytical chemists and metabolomics specialists. The descriptive, taxonomic, and composition data are particularly important for nutritionists, botanists, and natural product chemists. Unfortunately, many nutrient databases provide only one or two of these data fields. For instance, of the approximately 150 food composition databases found around the world, most provide only taxonomic and nutrient composition data.[15]

DATABASES ON PHYTOCHEMICAL STRUCTURES AND CLASSIFICATION OF PHYTOCHEMICALS

Table 2 provides a list of some of the better-known or more comprehensive phytochemical databases. Of the 21 databases we could identify, some provide structures and physical properties eventually with taxonomic data, whereas others give mainly food composition data with relatively minimal structural data. For example, PubChem is largely a chemical structure databases. Others, such as KEGG,[12] KNApSAcK,[16] and the Dictionary of Food Compounds[17] provide some chemical data and also offer a strong taxonomic component. Still others, such as Dr. Duke's Phytochemical database[18] and the USDA Food Composition databases,[19,20] are strictly nutrient composition databases. The one database that comes reasonably close to being the "ideal" nutrient/phytochemical database is Phenol-Explorer.[21,22] This particular database contains chemical, nomenclature, methodological, taxonomic, and composition data and offers full traceability of data sources. However, it still lacks important descriptive, structural, spectral, and clinical data.

Phytochemicals (in foods) can be classified in any number of ways, on the basis of their chemical structure, botanical origin, biosynthesis, or biological properties. The presence of characteristic structural motifs or chemical functions determines their belonging to a particular class: 2-phenyl-1,4-benzopyrone for flavonoids, phenolic groups in polyphenols, phytosterols with their steroid structure hydroxylated in the 3-position of the A-ring, alkaloids containing nitrogen atoms in complex and highly diverse structures, etc. Phytochemical classification may also derive from their biosynthetic origin, like "true alkaloids" derived from amino acids or terpenoids resulting from the condensation of a varying number of isoprene units formed through the mevalonate pathway. As a result, most phytochemical classification schemes are based on chemical structure definitions.

Table 3 provides characteristic examples and a general classification scheme for most major phytochemicals found in foods

Table 2. Phytochemical Databases and Resources of Interest to Food Scientists

database	domain	phytochemicals and other compounds	type of information	type of database	URL	ref
PubChem	all organisms	>26 million unique chemicals, synthetic and natural	structures, physical properties, literature links	open access, queryable, downloadable	http://pubchem.ncbi.nlm.nih.gov/	7
ChEBI	all organisms	>580,000 compounds, synthetic and natural	structures, physical properties, literature links	open access, queryable	http://www.ebi.ac.uk/chebi	73
eMolecules	all organisms	8 million compounds, synthetic and natural	commercial suppliers	open access, queryable, downloadable	http://www.emolecules.com/	
Dictionary of Natural Compounds	all organisms	170,000 compounds	structures, physical properties, literature links	commercial, queryable	http://dnp.chemnetbase.com	74
KEGG (Kyoto Encyclopeida of Genes and Genomes)	all organisms	16,054 compounds in 1100+ organisms	structures, physicochemical properties, pathways, occurrence in species	open access, queryable, downloadable	http://www.genome.jp/kegg/	36
MetaCYC	all organisms	>8700 compounds in 1914 organisms	structures, physicochemical properties, pathways, occurrence in species	open access, queryable	http://metacyc.org/	24
KNApSAcK	plants	7462 compounds from 6,324 species	structures, occurrence in plant species	open access, queryable, downloadable	http://kanaya.naist.jp/KNApSAcK/	16
Dr. Duke's Phytochemical and Ethnobotanical Databases	plants	8500 phytochemicals	occurrence in plants, content in plants, biological properties	open access, queryable	http://www.ars-grin.gov/duke/	18
Dictionary of Food Compounds	foods	30,000 natural food components and food additives	structures, physicochemical properties	commercial, queryable	CD ROM	17
USDA What's In The Foods You Eat	foods	63 fatty acids, vitamins, minerals, carotenoids, methylxanthines in 13,000 foods commonly eaten in the U.S.A.	content in foods[a]	open access, queryable, downloadable	http://www.ars.usda.gov/Services/docs.htm?docid=17032	

Table 2. Continued

database	domain	phytochemicals and other compounds	type of information	type of database	URL	ref
USDA−NCC database for carotenoids	foods	5 carotenoids from 215 foods	content in foods[a]	open access, downloadable	http://www.nal.usda.gov/fnic/foodcomp/Data/car98/car98.html	75
USDA National Nutrient Database	foods	2 methylxanthines in 4159 foods	content in foods[a]	open access, downloadable	http://www.ars.usda.gov/Services/docs.htm?docid=18879	
USDA database for flavonoids	foods	26 flavonoid aglycones from 231 foods	content in foods[a]	open access, downloadable	http://www.nal.usda.gov/fnic/foodcomp/Data/Flav/flav.html	
USDA database for proanthocyanidins	foods	6 proanthocyanidin fractions from 135 foods	content in foods[a]	open access, downloadable	http://www.nal.usda.gov/fnic/foodcomp/Data/PA/PA.html	
USDA−Iowa State University Database for isoflavones	foods	6 isoflavone aglycones from 128 foods	content in foods[a]	open access, downloadable	http://www.nalusda.gov/fnic/foodcomp/Data/isoflav/isoflav.html	
VENUS	foods	7 phytoestrogens from 791 foods	content in foods			76
USDA National Nutrient Database	foods	3 phytosterols in 529 foods	content in foods[a]	open access, downloadable	http://www.ars.usda.gov/Services/docs.htm?docid=18879	
Phenol-Explorer	foods	502 polyphenols from 452 foods	structures, content in foods,[a] metabolism	open access, queryable, downloadable	www.phenol-explorer.eu	21
EuroFIR-BASIS	foods	256 phytochemicals in 199 foods	content in foods, biological properties	membership, queryable	http://ebasis.eurofir.org/	65
Human Metabolome Database	humans	8147 human metabolites including some phytochemical metabolites	structures, physicochemical properties, spectral data, pathways, concentrations in human tissues, biological properties and literature links	open access, queryable, downloadable	http://www.hmdb.ca	14

[a] Composition values are means of original content data collected in one or more data sources.

Table 3. Chemical Classification of Major Phytochemicals

category	chemical class	chemical subclass	example
carbohydrates	monosaccharides		fructose
	disaccharides		sucrose
	oligosaccharides		amylose
	sugar alcohols		sorbitol
organic acids and lipids	short-chain organic acids	aldonic acids	ascorbic acid
		aldaric acids	tartaric acid
	fatty acids and lipids	omega-6 fatty acids	arachidonic acid
	alkanes and related hydrocarbons	waxes	nonacosane
	sulfur compounds	thiosulfinates	allicin
nitrogen-containing compounds	amines	benzylamines	capsaicin
		phenylethylamines	ephedrine
		tryptamines	psilocybin
	cyanogenic glycosides		amygdalin
	glucosinolates	aliphatic glucosinolates	sulforaphane
			sinigrin
		aromatic glucosinolates	glucobrassicin
	purines	xanthines	caffeine
	miscellaneous nitrogen compounds	indole alcohols	indole-3-carbinol
alkaloids	pyridine alkaloids		trigoneline
	betalain alkaloids	betacyanins	betanin
		betaxanthins	indicaxanthin
	indole alkaloids	ergolines	ergine
		yohimbans	reserpine
		tryptolines or β-carbolines	harman
			vinblastine
	indolizidine alkaloids		swaisonine
	pyrrolidine alkaloids		nicotine
	quinoline alkaloids		quinine
	isoquinoline alkaloids		berberine
		morphinans	morphine
	steroidal alkaloids		solanidine
		saponins	solanine
	tropane alkaloids		atropine
phenolics	flavonoids	anthocyanins	cyanidin
		flavanols	theaflavin
			procyanidin B2
		flavonols	quercetin
		dihydroflavonols	taxifolin
		flavones	apigenin
		isoflavonoids	genistein
		flavanones	naringenin
		dihydrochalcones	phloretin
	phenolic acids	hydroxybenzoic acids	gallic acid
			pentagalloylglucose
			anacardic acid
		hydroxycinnamic acids	ferulic acid
	lignans		pinoresinol
	coumarins		coumarin
		coumestans	coumestrol
		furanocoumarins	psoralen
	phenols	alkylphenols	4-ethylguaiacol
			5-heptadecylresorcinol
		methoxyphenols	guaiacol
			tyrosol
	phenylpropanoids	benzodioxoles	apiole
		curcuminoids	curcumin
		hydroxyphenylpropenes	eugenol
	quinones	benzoquinones	maesanin
		naphthoquinones	phylloquinone
		anthraquinones	rubiacardone A
	stilbenoids		resveratrol
	xanthones		mangostin
terpenoids	monoterpenoids		limonene
		phenolic terpenes	thymol
	sesquiterpenoids		farnesol
	diterpenoids		cafestol
	triterpenoids	phenolic terpenes	vitamin E
		saponins	ursolic acid
		phytosterols	campesterol
	tetraterpenoids	carotenoids	β-carotene

largely derived from the one proposed by Harborne.[23] This classification covers most of the ~20,000 phytochemicals identified in ~7000 edible plants.[17]

The classification schemes adopted in Table 3 are not without problems. Indeed, it is quite possible to have the same phytochemical classified into multiple categories. For instance, some phenolic terpenes such as oleuropein can be classified as terpenes or phenolic compound as they contain substructures for two different classes. Alternately, structurally different phytochemicals can be classified into the same category. For instance, phytoestrogens, a class defined on the basis of their bioactivity, can include such widely different chemicals as isoflavones, lignans, and coumestans. These discrepancies serve to underscore the need for shared classification for phytochemicals based on their chemical structures. Furthermore, even when there is agreement about structure similarities and classification, there is often some disagreement in structure-naming conventions.

There are at least three major chemical or metabolite databases that have developed reasonably useful chemical classification schemes. These databases include the HMDB,[14] the "Cyc" databases,[24] and ChEBI.[25] Each database uses its own classification scheme, although there is some general similarity. For instance, the HMDB[14] uses a hierarchical chemical classification scheme that is based on (1) kingdoms, (2) superclasses, (3) classes, (4) subclasses, and (5) chemical constituents. There are approximately 4 kingdoms, 30 superclasses, 300 classes, and 400 subclasses in this particular scheme. This classification scheme has been used to classify 8000 compounds in the HMDB and 1500 compounds in DrugBank.[26] The classification work done by the curatorial staff of the HMDB and DrugBank represents one of the largest chemical taxonomic classification efforts undertaken to date.

In contrast to the HMDB, the "Cyc" databases[27] use a slightly different hierarchical chemical classification scheme applied to a somewhat smaller number of compounds. However, the "Cyc" classification scheme has been applied to many more phytochemicals than HMDB and appears to be quite robust and well-designed. ChEBI[25] has also embarked on a systematic chemical classification effort using a carefully defined chemical ontology. An ontology is defined as a formal representation of a set of concepts about a subject and the relationships between those concepts. Ontologies are used to reason about the properties of a particular entity or subject and may be used to define/describe that entity or that subject. The ChEBI ontology does not quite fit with the conventional classification or taxonomic ideas that many chemists use, but it does have a logic and a rigor that make it very useful for computer-based searching and relational database development.

To date, all chemical classification, chemical systematics, or chemical ontology efforts have been done manually. Although this ensures some degree of rigor; if done by experts, manual classification is subject to the usual problems of human variability, differing definitions, and differing preferences. Furthermore, given that there are hundreds of thousands of known natural products, it is also clear that manual classification is not going to be possible for the vast majority of these compounds. Clearly, what is needed is a mechanism to automatically "compute" chemical classes and chemical ontologies for natural products. In other words, a computer program needs to be developed that can take a chemical structure file and then accurately identify what chemical class this compound belongs to and what kind of descriptors (ontological terms) are most suitable for that compound. An interesting classification of flavonoids based on substitution patterns of the different rings in their structure has been proposed and permits an easy recognition of the 6850 compounds known in plants (www.metabolome.jp/software/FlavonoidViewer/viewer).

If we consider other fields that have to deal with large numbers of entities, such as botany, microbiology, or zoology (~1 million species), genomics or proteomics (hundreds of millions of sequences), or structural biology (65,000 protein structures), all of them have developed automatic or semiautomatic classification schemes to group, cluster, or classify the entities they study. These classification schemes have revealed important insights into evolutionary processes, identified unexpected biological/physiological connections, explained novel or seemingly unrelated functions, and helped predict the existence of previously undiscovered entities. Classification schemes and ontologies also provide a common language or a common framework that allows thousands of scientists from diverse backgrounds to communicate easily and effectively. Certainly if natural product chemists could adopt a robust ontology or establish a consistent chemical classification scheme, then potentially the same positive impact could be seen in the fields of phytochemistry and natural product chemistry as well.

■ DATABASE RESOURCES FOR PHYTOCHEMICAL SPECTRA

Phytochemicals are often complex organic molecules that must normally be identified through mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. Identities of known phytochemicals are typically confirmed by comparing their mass or NMR spectra to the spectra of authentic standards. Novel or "unknown" phytochemicals must be identified through a combination of elemental analysis and MS and NMR spectroscopy. The availability of reference NMR or MS spectra of pure, authentic compounds is particularly important for the routine and rapid identification of phytochemicals in foods or beverages. Because of the importance of reference spectra to phytochemical research, it stands to reason that any "modern" phytochemical database should ideally contain reference MS or NMR spectra corresponding to each of the phytochemicals it contains. These reference spectra should be fully assigned (chemical shifts, mass fragments), viewable (as images), downloadable, and searchable. They should also have information about how the spectra were collected, including details on instrument type and model, instrumentation parameters, solvent, derivatization protocols, fragmentation energies, etc. This information is necessary so that other scientists can attempt to reproduce the data, if required. Likewise, the "raw" spectral data should also be available for download so that users may be able to process or inspect the data using their own software. Furthermore, the spectral data should be in a format that is easily exchanged or easily processed by commonly (or freely) available software. In the case of GC-MS data, the most common exchange format is the NIST format;[28] in the case of LC-MS data, this is the NetCDF format;[29] and in the case of NMR, this is either the CML (chemical markup language) or NMR-STAR format.[30,31]

As yet, there is no dedicated phytochemical database that meets all of these spectral archiving criteria. On the other hand, there are a number of spectral databases (containing at least some phytochemical entries) that do meet most of these requirements. Table 4 lists a number of dedicated NMR, GC-MS, and LC-MS

Table 4. Spectral Databases for Phytochemicals and Metabolites

database	content	type	URL	ref
HMDB	1824 1D and 2D NMR, 2560 MS/MS, 200 GC-MS spectra of metabolites	open access, queryable, downloadable	http://www.hmdb.ca	14
NMRShiftDB	25,100 NMR spectra of 21,500 natural products and organic compounds	open access, queryable, downloadable	http://www.ebi.ac.uk/nmrshiftdb	31
METLIN Metabolite Databse	4282 metabolite MS/MS spectra from 3156 metabolites	open access, queryable, downloadable	http://metlin.scripps.edu/	77
Madison Metabolomics Consortium Database (MMCD)	6218 1D and 2D ^{13}C and ^{1}H NMR spectra for 1840 metabolites	open access, queryable, downloadable	http://mmcd.nmrfam.wisc.edu/	78
NAPROC-13	^{13}C NMR spectra from >6000 natural products	open access, queryable	http://c13.usal.es/	32
BioMagResBank (BMRB – Metabolomics)	^{1}H and ^{13}C NMR spectra (1D and 2D) of 270 plant and animal metabolites	open access, queryable, downloadable	http://www.bmrb.wisc.edu/metabolomics/	30
Fiehn Metabolome Library (BinBase)	GC-MS spectra with RI data for 700 metabolites	commercial	http://www.chem.agilent.com http://www.leco.com	
Manchester Metabolome Database (MMD)	GC-MS and MS/MS data on 1065 metabolites	open access, queryable, downloadable	http://dbkgroup.org/MMD/	34
Spectral Database for Organic Compounds (SDBS)	24,000 EI-MS spectra, 28,000 NMR spectra from 34,000 organic compounds	open access, queryable, downloadable (partial)	http://riodb01.ibase.aist.go.jp/sdbs	
Golm Metabolome Database	GC-MS spectra for 500 plant metabolites	open access, queryable	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html	33
MassBank	MS/MS and EI-MS spectra from 12,000 organic compounds	open access, queryable, downloadable	http://www.massbank.jp/	11
NIST MS Library and GC Retention Index Database	EI MS spectra for 192,000 compounds and RI values for 21,000 compounds	commercial	http://www.nist.gov/	
HaveItAll CNMR-HNMR Library	438,000 ^{13}C NMR and 30,000 ^{1}H NMR spectra of organic compounds	commercial	http://www.bio-rad.com	
ACD Laboratories Aldrich NMR Library	^{13}C and ^{1}H NMR spectra for 35,000 compounds	commercial	http://www.acdlabs.com	
ACD Laboratories HNMR DB & CNMR DB	^{13}C and ^{1}H NMR spectra for >200,000 compounds	commercial	http://www.acdlabs.com	
Thermo Scientific Fragment Library	19,000 literature-derived MS fragment trees	commercial	http://www.thermo.com http://www.highchem.com	

databases that may be particularly useful for phytochemical analysis and identification. In many cases, the spectra contained in these open access spectral databases can be easily imported into existing phytochemical or nutrient databases. Unfortunately, despite their ready availability, this has not yet happened.

With regard to NMR spectral resources for phytochemicals and other natural products, there are at least seven freely available resources and at least three commercial databases (see Table 4). The two largest are NAPROC-13[32] and NMRShiftDB.[31] Both of these databases appear to have a fairly substantial collection of natural product and phytochemical spectra under a variety of solvent conditions. Because of the large spectral dispersion, the relative chemical shift invariance, and the simplicity of ^{13}C NMR spectra, most analytical chemists prefer to use ^{13}C NMR for the identification of phytochemicals, phytochemical metabolites, and other natural products. In this regard, NAPROC-13, which is a ^{13}C NMR database of natural products, probably represents the richest NMR resource for phytochemists and phytochemical databases.

With regard to GC-MS spectral resources for phytochemicals and other natural products, the most widely used database is the NIST database. The latest release contains EI-MS spectra for 192,100 compounds and retention index (RI) values for 121,800 compounds. Unfortunately, many of the NIST compounds are not natural products or phytochemicals. Four other databases, albeit somewhat smaller in size, also provide some GC-MS data for phytochemical identification. These are the Golm Metabolome Database,[33] the Manchester Metabolome Database,[34] the Fiehn Metabolome Database (FiehnLib),[35] and the HMDB.[14]

LC-MS or LC-MS/MS techniques offer much greater sensitivity than NMR or GC-MS does. Unfortunately, LC-MS methods often lack the consistency or reproducibility that characterizes GC-MS or NMR. This makes compound identification via spectral matching quite difficult. For instance, differences in column geometry, column packing, and solvent elution protocols can lead to profound differences in elution times for the same compound. Likewise, differences in collision energies (for MS/MS) along with differences in ionization techniques (MALDI versus electrospray) or instrument configuration [ion trap, Fourier-transformed ion cyclotron resonance (FTICR), triple quad] can lead to significantly different mass spectra for the same compound. This has made it difficult to develop reliable instrument-independent LC-MS databases. Nevertheless, some efforts are being made to overcome these problems, and a number of LC-MS spectral databases are beginning to appear. Some are relatively instrument independent, such as MassBank, which contains spectra obtained with different instruments[11], whereas a number of commercial databases are specific to a restricted set of instruments. In the area of phytochemical research, there is a tendency for many MS specialists to create their own "private" library of LC-MS spectra that is specific to their own instrument. Although this is not an ideal solution, until more widespread LC-MS standards can be established, this may be the best option for the time being.

Many of these databases are not particularly focused on phytochemicals, and it is difficult to evaluate the extent of coverage for phytochemicals in these tools. Conversely, some other databases are focused on some particular classes of phytochemicals. MS-MS Fragment Viewer (http://webs2.kazusa.or.jp/msmsfragmentviewer/) is a spectral database for flavonoids having MS, MS^2, and photodiode array spectra for 116 pure compounds with structures of the MS^2 fragments.

DATABASES ON PHYTOCHEMICAL METABOLIC PATHWAYS IN PLANTS

Pathway databases are expected to provide biosynthetic/degradation routes of metabolites to visually introduce their functional roles. Because description of metabolic pathways requires detailed knowledge on related enzymes and metabolites, extensive expertise is necessary for the design and maintenance of pathway databases. Each database takes a different strategy to compile pathway knowledge and exhibits unique characteristics depending on its expected usage. From users' perspective, we here categorize them into three types: comprehensive pathway databases, specialized pathway databases, and community-based approaches to accumulate pathway knowledge.

Comprehensive Databases. Comprehensive databases are online counterparts of the classic biochemical wallcharts (Roche's and Sigma's versions are famous; see Table 5 for their online information), covering all pathways of multispecies in a single map. The KEGG database is well-known for its comprehensiveness and provides the pathway knowledge in a downloadable format for over 1200 fully sequenced organisms.[36] Most genomes are bacterial, and for plants seven higher species are included (thale cress, black cottonwood, castor bean, wine grape, Japanese rice, sorghum, and maize) as of January 2011. Its pathway reconstruction is semiautomated: about 160 manually designed pathway charts are prepared as the reference information, on which precomputed results of genome-wide homology search can be projected for a specific organism on users' demand. The functional assignments for genes in each species are based on EC numbers of enzymes. Therefore, it provides a genome-centric view of computationally predicted metabolic network. In the past few years, plant-specific information has been actively compiled in the KEGG Plant page (Table 5). In this portal, the "category maps" covering plant secondary metabolites are drawn with molecular structures and are useful for beginners to grasp the biosynthetic overview of phytochemicals.

The KEGG database represents a semiautomatic annotation. The representative of manual curation is the Cyc database families, the information of which is summarized at the Plant Metabolic Network (PMN) and Gramene Pathway (GP) Web sites. In these Cyc projects, the general repository of reference pathways is called the MetaCyc database, and plant-specific pathways are compiled as the PlantCyc database.[24] Well-known species-specific versions are AraCyc for *Arabidopsis thaliana* (thale cress, Brassicaceae) at PMN,[37] RiceCyc for *Oryza sativa* ssp. *japonica* (rice, Poaceae/Gramineae) at GP, LycoCyc for *Solanum lycopersicum* (tomato, Solanaceae) at Sol Genomics Network,[38] and MedicCyc for *Medicago trancatula* (barrel clover, Fabaceae/Leguminosae) at Noble Foundation.[39] To construct such a site, all pathways in the MetaCyc database are computationally matched against genomic information as in the KEGG database in the first place. Predicted pathways then undergo an extensive manual curation using literature to improve quality and the coverage of experimentally verified pathways. Therefore, although starting from an automated prediction, each Cyc database becomes a biochemical corpus of expert knowledge gradually increased with time. The AraCyc, by far the most well curated database among plant Cycs, contains 400 pathway pages with a total of 3400 references. It must be noted, however, that the definition of "pathway" is different among database projects. The Cyc projects tend to represent shorter pathway fragments for detailed annotations, whereas the KEGG emphasizes visual

Table 5. Primary Pathway Resources on the Internet

database	content	pathway format[a]	URL
Roche Biochemical Pathways	digitized version of the paper wallchart	PNG	http://www.expasy.ch/cgi-bin/show_thumbnails.pl
IUBMB−Sigma Metabolic Pathways Chart	smaller charts are freely available online	PNG, SVG, PDF	http://www.iubmb-nicholson.org/
KEGG Plant	portal for metabolic maps, phytochemicals and crude drugs	PNG	http://www.genome.jp/kegg/plant/
Plant Metabolic Network	portal for PlantCyc (general), AraCyc (thale cress), and PoplarCyc (poplar)	dynamic HTML	http://www.plantcyc.org/
Gramene Pathway	portal for RiceCyc (rice) and SorghumCyc (sorghum)	dynamic HTML	http://www.gramene.org/pathway/
SolCyc (Sol Genomics Network)	portal for LycoCyc (tomato), PotatoCyc (potato), CapCyc (pepper), CoffeaCyc (coffee), PetuniaCyc (petunia), NicotianaCyc (tobacco), and SolaCyc (eggplant)	dynamic HTML	http://solgenomics.net/tools/solcyc/
MedicCyc	annotations for barrel clover	dynamic HTML	http://mediccyc.noble.org/
SoyCyc (SoyBase)	annotations for soy	dynamic HTML	http://www.soybase.org:8082/
MapMan	interactive visualization for plants	PNG	http://mapman.gabipd.org/web/guest/mapman
Kappa View	interactive visualization for plants	SVG	http://kpv.kazusa.or.jp/kpv4/
IUBMB Enzyme Nomenclature	terpene synthesis (EC 5.3.3.2) and sterol synthesis (EC 5.5.1.9)	PNG	http://www.chem.qmul.ac.uk/iubmb/enzyme/
BioCarta	mainly proteomic pathways for human	GIF by FreeHand (Adobe)	http://www.biocarta.com/genes/index.asp
WikiPathways	mainly proteomic pathways	SVG by PathVisio	http://wikipathways.org/index.php/WikiPathways http://www.pathvisio.org/

[a] Abbreviations: PNG, portable network graphics; SVG, scalable vector graphics; PDF, portable document format; GIF, graphics interchange format.

effect and excels in its summary views. For this reason, the number of pathways does not scale to the coverage of pathway knowledge.

Specialized Databases. Because no comprehensive approach can cover everything, there is room for smaller database projects. One typical demand is overlaying locally measured data, for example, gene expression or metabolite concentration, on metabolic maps. MapMan is a software system to project quantitative information on metabolic maps, designed primarily for *A. thaliana*.[40] The whole system including 60 metabolic maps is freely downloadable from its Web site, whereas the software is also available as a Web application program. Its metabolic maps, provided in the portable network graphics (PNG) format, are simple and easier to understand than more comprehensive KEGG maps. Likewise, KaPPA-View is designed to overlay quantitative information on its 130 metabolic maps.[41] This Web-based system supports many interactive features such as upload of user-defined pathways or correlation data in the Excel format (Microsoft, Redmond WA). Metabolic maps are provided in the scalable vector graphics (SVG) format, and users can download, edit, and reupload them using either a separately provided free drawing editor or any commercial editor such as Illustrator (Adobe, San Jose, CA). For informatics experts, superposition of user-defined data is achievable for the KEGG maps through its Simple Object Access Protocol/Web Service Definition Language (SOAP/WSDL) interface. However, customization of the KEGG requires some programming skill on the users' side and is not easy for everybody.

Enzyme Nomenclature by International Union of Biochemistry and Molecular Biology (IUBMB) is known as the authoritative source of the EC hierarchy for enzymatic reactions and other nomenclatures, but what is less known is its pathway resource for the terpene synthesis (accessible from the EC 5.3.3.2 entry) and the phytosterol synthesis (accessible from the EC 5.5.1.9 entry) in combination with EC annotation. Although the number of available pathways is limited, pathway information with reaction scheme (i.e., the movement of electrons) is quite useful. Nomenclature for lignans, carotenoids, retinoids, and other vitamins is also available from its parent directory and is used as the standard for phytochemical namings.

Community-Based Repositories. Because metabolic pathways are a culminated form of biochemical knowledge, pathway databases require tremendous construction and maintenance work. To alleviate such cost, at least in part, a few Wiki-based projects have been proposed. Well-known repositories for biological pathways are BioCarta and WikiPathways.[42] Both repositories provide drawing aids to standardize pathway views and the degree of annotation and encourage users to contribute pathway information.

The statistics of contribution reflect the number of researchers in each field, and many pathways on community-based sites describe proteomic networks (e.g., cell signaling) in humans and animals. Only a few contributions are related to plant metabolism as of November 2010. Community-based design conforms to the academic method of knowledge compilation, but no such sites can offer a clear incentive for busy researchers to join and contribute. The systemic analysis and invention of incentives for collaborative effort are necessary to maintain and expand the pioneering success.[43]

DATABASES ON PHYTOCHEMICAL CONCENTRATIONS IN FOODS

It is important to know precisely the concentrations of phytochemicals in foods to understand, master, and eventually improve technological, biological, and nutritional properties of the many foods consumed with the diet. This information is most notably needed to determine phytochemical intake in different populations and to study associations with health and disease outcomes in epidemiological studies. In contrast to concentrations of macronutrients, vitamins, and minerals found in most food composition tables, information on phytochemical composition is still largely scattered in the literature. A common repository for phytochemical content in foods is highly desirable. There are several major challenges in developing a food composition table for phytochemicals. These include the structural diversity of the compounds, the large number of dietary sources, the large variability in content for a given source, the diversity of analytical methods, and, in some cases, the lack of suitable analytical methods.

Furthermore, most analytical methods for phytochemicals are not standardized. Phytochemicals are generally analyzed by LC or GC with a UV or MS detector. However, measured content values may vary according to the protocol used to collect, store, treat, and analyze the samples. Data quality should also be evaluated according the analytical method used, which should be carefully documented in the original data sources. In particular, curatorial staff should ensure that data from various sources are comparable in terms of sample extraction (a hydrolysis is sometimes used to liberate photochemicals from the plant matrix) and analysis (standards, etc.).

Various authors have analyzed a limited number of phytochemicals in tens and sometimes hundreds of foods commonly consumed in a given country. Samples are collected according to a proper sampling plan to limit possible bias that may result from genetic, geographical, or environmental variability,[44] and a specific analytical method is applied to estimate the phytochemicals of interest in these samples. Small databases have thus been produced for, for example, 7 phytosterols in 87 foods,[45] 6 catechins in over 50 food items,[46] or 8 phytoestrogens in 240 English foods.[47] However, due to the considerable diversity of food phytochemicals, of methods needed to analyze them, and of foods consumed throughout the world, the construction of a food composition database for all food phytochemicals is an impossible task for a single laboratory. More comprehensive databases have then been built by curation of composition data collected from a large number of peer-reviewed publications (Table 2). These databases contain either original content data as collected from data sources or mean content values calculated from multiple original content data. One database (Phenol-Explorer) also provides all original data with the corresponding literature sources used to calculate mean values.

Mean content values should be considered more representative of the average content of a phytochemical due to the large content variability described above, unless a proper sampling plan is applied to obtain samples characterizing the average diet in a given population or country. However, such sampling plans are costly and not often implemented. The number of original data used to calculate the mean should then be large enough to obtain mean content values close to that of an average sample of the food considered. The USDA databases and Phenol-Explorer provide the number of sample analyses and the number of studies used to calculate mean content values, both essential parameters to evaluate the quality of mean content values.[21,48]

The quality of food composition data varies widely from one database to another. Dr. Duke's Phytochemical and Ethnobotanical Databases give estimates for over 8000 compounds in different organs of a large number of plant species, with very few details on the source of the information and no information on the analytical methods used (Table 2). The USDA databases provide detailed information on contents of carotenoids, methylxanthines, flavonoids, and phytoestrogens (48 compounds in total) in a large number of foods. Data sources are peer-reviewed journals and unpublished data from USDA and food industries. Phenol-Explorer is the most complete database for dietary polyphenols. Over 60,000 original data have been compiled and evaluated, and average content values have been calculated for more than 500 polyphenols (flavonoids, phenolic acids, lignans, and stilbenes) in 450 foods.[22] Unique features of Phenol-Explorer are that different content values are provided according to different types of analytical methods and that all original data used to calculate mean content value can be retrieved on the Web site. Various queries can be made to calculate contents of polyphenols as aglycones or total by classes and subclasses. Text information on polyphenols in the different food groups is also available.

Databases for dietary supplements containing bioactive supplements will also be important in the future due to their widespread and increasing consumption. Existing databases mainly contain data on minerals and vitamins, but some bioactive phytochemicals have also been considered of highest priority to be included in such databases.[49] These are caffeine, lycopene, soy isoflavones, and *Ginkgo biloba* extracts. Phytochemical content compiled in the database can be based on the label information.[49] However, label information is often not accurate. Phytoestrogen content is commonly overestimated, and this advocates for the analysis of the most largely consumed dietary supplements.[50,51] Composition data will also have to be included in databases together with the label information as in the NHANES-DSLD database (National Health and Nutrition Examination Survey's Dietary Supplement Label Database).[52]

A number of databases on food functionality also exist. Food functionality is linked to food composition, and it sometimes provides information not easily collected by direct food analysis of individual phytochemicals. One such property often measured on foods is their antioxidant capacity, measured by assays such as the oxygen radical absorbance capacity (ORAC) and ferric reducing antioxidant power (FRAP) assays. The antioxidant capacity is linked to the presence in foods of reducing compounds (in the chemical sense) such as polyphenols, ascorbic acid, vitamin E, or carotenoids. Whether such assays help to predict health benefits is still questionable,[53] but these assays are still largely used to promote the merits of various foods rich in the so-called antioxidants, and ORAC, FRAP, or Folin values have been collated in several databases as the result of direct food analysis[54,55] or compilation of data from the scientific literature.[22]

Contents of nutrients in foods are influenced by cooking and processing. Not all cooked or processed foods are found in food composition database, and it is common practice to apply retention factors to nutrient contents in raw foods to calculate contents in cooked or processed foods from those in raw foods.[56] Retention factors are available for a number of common nutrients, vitamins, minerals, and protein, in the USDA database and a few European food composition tables,[57] but very limited information on phytochemicals can be found. The only table so far available gives retention factors for 5 carotenoids in 280 foods.[58]

■ DATABASE RESOURCES ON PHYTOCHEMICAL METABOLITES IN HUMANS AND ANIMALS

Once ingested, phytochemicals are absorbed and found in their native form in blood and tissues. They are also extensively metabolized in tissues and by the microbiota in the gut. Polyphenols form glucuronide, methyl, and sulfate conjugates and are degraded in the colon into low molecular weight compounds such as phenolic acids.[59] Glucosinolates are hydrolyzed to isothiocyanates and indoles; the former is further metabolized to mercapturic acids and the latter, condensed to form indole acids in the stomach.[60] Tocopherols are hydroxylated and oxidized and metabolites further conjugated to sulfate, glucuronide, and glucoside groups.[61] Carotenoids undergo isomerization and are eventually cleaved into two retinal molecules.[62,63]

Some of these phytochemical metabolites can be found in the phytochemical and spectral databases described above (Tables 2 and 4): PubChem, Chemspider, ChEBI, eMolecules, KEGG, MetaCYC, HMDB, MassBank, Madison Metabolomics Consortium Database (MMCD), and METLIN. They include molecular weight, molecular formula, structure, name, and synonyms as well as NMR and mass spectra. Spectra are often missing in these databases due to the lack of commercial standards. When only limited knowledge on phytochemical metabolites is available, the structure and spectra of metabolites can be predicted using in-silico prediction tools. For example, Meteor is one such expert-based system designed to predict the most likely phase I and phase II metabolites of any compound from its chemical structure. These tools are commonly used in pharmacology but rarely in nutrition. Such information should also be included in metabolite databases. Both established and predicted data would be particularly valuable to interpret results of metabolomic studies aiming at the identification of new biomarkers for plant food consumption or phytochemical exposure.

These databases are also expected to provide data on the occurrence and range of concentrations of metabolites in biofluids and tissues in both humans and experimental animals as well as data on metabolic pathways. Whereas descriptions of metabolic pathways for lipids, proteins, amino acids, sugars, and hormones are well-known and summarized in several databases such as KEGG, Reactome (www.reactome.org), and PharmGKB (www.pharmgkb.org/), no detailed information on pathways for phytochemical metabolism is available in these databases. HMDB is, to our knowledge, the only database containing concentrations of phytochemical metabolites in human biofluids. However, the number of compounds is still limited. For compounds such as quercetin or catechin, it contains only concentrations for aglycones as measured after enzymatic or acid hydrolysis of plasma and urine and no concentrations for conjugated metabolites or microbial metabolites.

Ideally, a database on phytochemical metabolites should include all metabolites identified in intervention studies with an isolated phytochemical or phytochemical-rich extract or food. All metadata on the intervention study should be included: study design, a detailed description of the phytochemical source (more particularly, the nature and concentrations of the phytochemicals) and of the control, the dose ingested, the period of intervention, the subject characteristics, the timing of the biofluid and tissue collection, the description of the analytical methods, and the concentrations of the phytochemical metabolites measured at different time points. Data obtained on experimental animals should also be included, in particular when human data are missing. Data on animals fed isolated phytochemicals are particularly useful to establish metabolic pathways with sufficient certainty. No such database exists today. A new module of the Phenol-Explorer database[21] is under construction. It will include all available information on about 380 polyphenol metabolites so far described in the literature (J. Rothwell, M. Urpi-Sarda, C. Andres-Lacueva, and A. Scalbert, unpublished data).

■ DATABASE RESOURCES ON BIOLOGICAL PROPERTIES OF PHYTOCHEMICALS

Today, the greatest interest in phytochemicals lies not in their chemical properties but in their biological or health-promoting properties. As a consequence, there is a growing expectation that phytochemical databases should include not just information on chemical structures, chemical names, and chemical descriptions but also quantitative data on the physiological effects of phytochemicals or their metabolites. Unfortunately, physiological effect data are not as easily compiled or presented as chemical or nomenclature data. In particular, biological effects cannot be presented in a compact, quantitative form such as a structure, a molecular formula, an IUPAC name, or an NMR spectrum. Rather, biological effects have to be described in human-readable sentences or an agreed-upon ontology (using terms such as "antioxidant", "anti-cancer", or "anti-inflammatory"). Beyond providing a simple indication of the presumptive physiological effect, this claim also has to be backed up by some supporting information. These supporting data should include the original reference, a synopsis of the study, the testing conditions, the test system (type of cells, organs, or animals), the type of assay(s), the phytochemical concentrations, the phytochemical metabolites (if detected), the type of effect (beneficial or toxic), the degree of the biological effect, the number of samples, the statistical significance of the effect, and an "external" assessment of the quality or reliability of the study. This is not a task that can be easily automated. Indeed, the only way that biological effect data can be properly compiled (at least for now) is for expert curators to manually scan through the relevant papers, books, and journals and to enter these data manually using a laboratory information management system (LIMS).

Compiling this kind of information presents an enormous challenge for the phytochemical and nutritional chemistry community. Over the past two decades thousands of studies on the health-promoting or beneficial effects of phytochemicals have been published. These studies have used a wide range of enzyme assays, cell assays, perfused organ models, and animal (rodent) models. Unfortunately, as most researchers now know, there is considerable variability in the quality, testing conditions, and claims made in these studies.[64] In far too many cases, phytochemical/nutrient studies are poorly controlled and restricted to very simple cellular or in vitro assays that have little relevance to physiological conditions. In many other cases the concentrations used to generate a detectable effect are many tens or hundreds of times higher than what could be achieved through normal food consumption. In other cases, only the presumptive beneficial effects have been measured, but no assessment of the toxicity or potential adverse side effects has been attempted.[64] This is why it is critical that nutrient/phytochemical databases of the future include an appropriate amount of information (i.e., conditions, system, effect, assays, sample number, concentrations, significance, quality, etc.) about any measured biological or physiological effects.

Performing this kind of systematic compilation of the biological effects measured in phytochemical/nutrient studies will certainly allow researchers improved access and improved opportunities to comparatively assess phytochemical effects. Likewise, allowing users to search phytochemical databases for physiological effect terms (such as antioxidant or anti-cancer) or for study quality measures (poor, good, excellent) or for assay conditions (cell types, animal types) will also allow meta-studies to be far more conveniently performed. A common database on biological properties would also allow the sorting of results of in vitro and animal studies according to their nutritional relevance for humans and the identification of the studies carried out with the lowest doses (closer to nutritional exposure) and with the main phytochemical metabolites, rather than food native phytochemicals as most commonly done. The other benefit to compiling this kind of information into a centralized, open-access database is that it will help researchers to improve the design and scope of their own in vitro or in vivo studies.

■ DATABASE RESOURCES FOR CLINICAL TRIALS WITH PHYTOCHEMICALS

Data from interventional randomized clinical trials is the gold standard of evidence when the effects of a particular dietary intervention on disease risk and the safety of foods and food components in humans are assessed. For phytochemicals, useful data indicating the effect of the phytochemical on disease risk can be obtained if such interventions are conducted with appropriate controls, whether the interventions are with isolated compounds, phytochemical-rich extracts, or phytochemical-rich foods. Although data from individual interventional trials can be useful, the most powerful assessments come in the form of systematic reviews, wherein all of the available data from all of the trials that have appropriately investigated the effects of a particular phytochemical on the risk of a disease are meta-analyzed to increase the statistical power. Evidence from randomized controlled clinical trials and meta-analyses of multiple such trials is routinely used by researchers as well as by expert groups working for local and federal government health departments such as the U.S. Food and Drug Administration and food safety authorities such as the European Food Safety Authority or the World Health Organisation to underpin policy decisions and public health advice. Clinical trial data are also used by the food and supplements industries in support of health claims and as evidence of product safety. The number of publicly reported interventional clinical trials assessing the effects of phytochemicals on human health is increasing rapidly, but the reports of such trials are often difficult to find and the outcome data are often difficult to extract or not reported at all. Database resources that facilitate rapid searching and retrieval of data from such trials is highly desirable.

What Information Do These Databases Need To Contain? For interventional clinical trials, there are a large number of important variables that need to remain closely associated with the outcome data to retain the context. For example, it is not sufficient for a clinical trial database to merely contain data for the following headings: "food or phytochemical", "outcome measure", and "change from control". It is imperative that many other details including the study design, the dose, the form (i.e., whole food, crude extract, pure compound), the period of intervention, the subject characteristics including the measurement values at baseline, and the timing of the measurements are included if the data are to be suitable for users to be able to assess the trial suitability and quality. These criteria are routinely assessed as inclusion/exclusion criteria during the data retrieval steps of systematic reviews of clinical trial data. The criteria set out in the "CONSORT Statement", which is an evidence-based, minimum set of recommendations for reporting RCTs, are also a useful guide (http://www.consort-statement.org/home/). It is important that all of these variables are included in clinical trial outcome databases so that different users can apply their own selection criteria to extract the data that are relevant to their needs. A link to the original published paper, when available, is also essential. The important data fields for a phytochemical clinical trial outcome database are shown in Table 6.

Currently Available Resources for Phytochemical Clinical Trials. It is considered good scientific practice that all of the details of all clinical trials (that is, biomedical or health-related research studies in human beings that follow a predefined protocol) are registered in an open-access registry. This has become widespread practice, partly because, in many cases, there is a requirement for trials to be registered, for example, as a stipulation of the funding body, trial sponsor, or the journal in which the data are to be published. For example, the U.S. National Institutes of Health host a site that serves as a registry of federally and privately supported clinical trials conducted in the United States and around the world (www.ClinicalTrials.gov). An ability to search for all of the clinical trials that have been

Table 6. Fields Required in a Phytochemical Clinical Trial Outcome Database

data category	specific data content
study design	description of trial design such as parallel or crossover, randomized
subjects	subject inclusion and exclusion criteria. subject baseline characteristics by group
blinding	who was blinded to the interventions (participants, clinical staff, staff assessing outcomes, statistician)
interventions	details of what was provided (dose, form) and when it was provided or ingested by study participants details of placebo or control diet
outcome measures	when measurements were made, what measurements were made, the methods used to make the measurements
numbers analyzed	number of participants per group for which the outcome data were calculated
outcomes	baseline measurement, effect size, and precision (e.g., 95% confidence interval, standard error, standard deviation)

conducted is important because in some cases, and for various reasons, clinical trial outcome data are not reported publicly, and such information may be interrogated to determine the likelihood of reporting bias.

In terms of currently available information on outcomes, it is in the form of (i) systematic reviews (meta-analyses) that are typically published in peer-reviewed journals, (ii) nonsystematic reviews that are also typically published in peer-reviewed journals, and (iii) Web-based databases (Table 7). The Chemoprevention of Colorectal Cancer Database includes data on β-carotene, but this is the only true phytochemical in this database. eBASIS is an online fully searchable database resource that contains a description of 445 clinical trials on 144 food bioactive compounds and their effects on 56 biomarkers mainly related to cardiometabolic and bone health outcomes.[65]

DATABASE DESIGN: RECOMMENDATION FOR FUTURE DEVELOPMENTS

Most databases are constructed using certain well-defined schemes or architectures. Simple databases consist of a single table or list. More complex databases are relational, meaning that the data are organized as a set of multiple, formally described tables allowing the data to be accessed or reassembled in many different ways without having to reorganize the database tables. Because of their flexibility, relational databases now dominate the world of electronic databases and are found in every area of business, finance, art, design, entertainment, engineering, and science.

Table 7. Current Sources of Phytochemical Clinical Trial Data

name	description	URL or ref
Current Controlled Trials	online resource for searching for clinical trials across multiple registers (including U.S. ClinicalTrials.gov and U.K. NHS)	http://www.controlled-trials.com/
ClinicalTrials.gov	online registry for clinical trials established by the U.S. National Institutes of Health, used worldwide	http://clinicaltrials.gov/
systematic reviews	published peer-reviewed papers in the scientific literature	79−82
nonsystematic reviews	publications (often peer-reviewed) that review all existing published data for a phytochemical (group) and a disease or specific outcome measure	e.g., 83−85
Chemoprevention of Colorectal Cancer	online database of agents and diets ranked by efficacy including a systematic review of experimental studies (men, rats, mice)	http://www.inra.fr/internet/Projets/reseau-nacre/sci-memb/corpet/indexan.html
eBASIS (BioActive Substances in food Information System)	online database of biological activity data including clinical trial outcomes for phytochemicals currently available via subscription	http://www.polytec.dk/ebasis/Default.asp; 65

A database is only as useful as the data that it contains. Obviously the more relevant or current the data, the more useful the database will be. In an effort to keep their databases relevant, many database developers and curators spend a considerable amount of time acquiring data or developing methods to acquire high-quality data. For scientific databases, data quantity, quality, and currency are of paramount importance. Consequently, automated data retrieval, data validation, or data deposition systems often play an important role in scientific data acquisition or data compilation. Archival databases such as PubMed, GenBank,[8] Protein Data Bank,[10] and BioMagResBank[30] have very elaborate, highly automated data management systems to handle submissions, validate entries, track files, and store information. The design and construction of these automated or semiautomated data acquisition systems represent a challenge that is unique to each database and is far beyond the scope of this paper.

Curated databases, on the other hand, tend to be the product of manual labor by a single curator or a team of curators. For these databases data acquisition and data entry are not automated, but rather data are usually manually searched, read, assessed, entered, and validated. Some automated text mining systems, such as Textpresso[66] or PolySearch,[67] can help simplify the task of finding relevant text or papers. However, these autolocated papers or abstracts must still be manually read and the data manually extracted and entered. In addition to text mining systems, there are also data entry systems (called laboratory information systems or LIMS) or commercial database packages (such as MS-ACCESS, Oracle) that can be used to facilitate data entry and compilation.

Once the data of interest have been acquired. there are some general rules on how to assemble these data into a high-quality scientific database. These rules follow an easy-to-remember acronym: A-C-Q-U-I-R-E. In particular, every scientific database should be Accessible, Comprehensive, Queryable, User-friendly, Interactive, Referenced, and Expandable.

Accessible. The fundamental reason to create a database is to make its contents readily accessible. Accessibility is the key to the success of almost any scientific database as there is a widespread (and justified) belief that publicly funded scientific data must be freely available to the public. As a result, the vast majority of life science databases and a growing number of chemical databases are being converted into freely available resources that can be easily accessed or downloaded over the Web without passwords or logins. Open accessibility has many benefits, not the least of which is increased visibility. Indeed, high-quality, open-access databases often receive millions of Web hits, thousands of downloads, and hundreds of citations a year. Given the importance of phytochemicals in food and nutrition research, a well-designed, Web-accessible phytochemical database could certainly be very popular across many communities.

Comprehensive. A high-quality database must also be comprehensive. Not only should a database provide comprehensive data coverage of a given field or topic, but it should also contain a wide diversity of data types. In particular, good scientific databases typically contain a good mixture of text, numeric, graphical, and image data. For instance, the GeneCards database[68] is an excellent example of a comprehensive life science database. It contains a rich mixture of text, numbers, charts, graphs, and chromosome maps. An equally comprehensive mix of data types (pictures, graphs, charts, numbers, and text) can be found in the Protein Data Bank[10] or DrugBank.[26] As a general rule, comprehensive databases typically have 30−100 data fields for each entry. Unfortunately, many of today's phytochemical or nutritional databases contain only 5−10 data fields and are primarily restricted to textual data.

Queryable. A database is not of much use if it cannot be queried or searched. Better databases support a wide range of searches, from simple text matching to complex Boolean queries (AND, OR, NOT). Some of the best-designed databases support partial text matching, wild-card characters, and automated synonym searches. A few even provide suggestions for misspelled words. A growing number of databases also support data-field specific queries. This allows users to look in only specified parts of the database for certain numbers, names, or images. Many chemical databases also support structure similarity searches (using subgraph isomorphism or Tanimoto scores[69]), molecular weight searches, chemical formula searches, and SMILES string searches. Likewise, many food composition databases also support queries by nutrient content type, content ranges, food types, and plant/taxonomic identifiers.

User-Friendly. A database needs to be designed so that almost anyone can use it. Indeed, a key question that every database developer must ask is: Could my 80-year old parent/grandparent use it? If the answer is no, then the database is probably not sufficiently user-friendly. Unfortunately, too many databases are being built without user-friendliness as a high priority. It is not unusual to find a "public" database that is so poorly designed that only those who know the database's specialized accession numbers can access its data or attempt to view its content. User-*un*friendly databases, no matter how valuable or rich the content, are almost never used. At a minimum, user-friendly databases should always be "browsable", meaning that if users do not quite know what they are looking for or how to look for it, they can simply scan through the contents. Once a user has browsed the content, then he or he is usually better able to make specific queries. Bookstores, libraries, and magazine racks in stores are all examples of user-friendly and easily browsed data repositories. A good electronic database should offer the same kind of browsability. User-friendliness also refers to how easy it is to use the database query system. Given that few database users are versed in relational database queries or structured-query language (SQL), it is essential to design a database interface so that complex queries can be performed through simple pull-down menus or clickable boxes using plain language.

Interactive. Thanks to the Web and the hyperlinking capabilities of HTML, electronic databases are becoming increasingly interactive. Interactivity means that a database is "clickable". In other words, users can use a mouse, a stylus, or a keyboard to type in queries, select menu options, expand views, manipulate images, hyperlink to other data files, or connect seamlessly to other (related) databases. Interactivity is an important component of user-friendliness, but is also an important part of interconnectability. Databases should not be developed as isolated "data islands". The linking of other high-quality data resources to an existing database adds value not only to the database being built but to all of the databases to which it connects. Most life science databases have hyperlinks to at least four or five other databases. In some cases hyperlinks to more than 20 databases are not uncommon. Hyperlinking to other databases allows users to see complementary data or to obtain additional information in a quick and easy manner. Hyperlinking also simplifies things for database curators/developers as they do not have to worry about compiling data or covering areas in which they have little interest or expertise.

Referenced. A database needs to be reliable. This reliability comes from acquiring data that are fully and properly referenced. A database without references, data sources, or citations is intrinsically unreliable. Obviously, some databases may consist of mostly unpublished experimental measurements or experimental observations. Likewise, other data types can be computationally predicted (i.e., predicted LogP, pK_a, or molecular weight). In these cases, general references to the methods, techniques, or programs used to generate the data must still be made. Proper references ensure that the data can be regenerated or reproduced. References also allow users to investigate the data sources for further information or further clarification. Having data that are properly referenced also helps safeguard a database against one of the biggest problems in databases today, data entry errors. In particular, references allow both internal curators and external users to validate what has been entered.

Expandable. Databases should never be viewed as static entities. Certainly in the life sciences new information is being discovered all the time. As a result, databases, especially scientific databases, must be designed so that they can be continually expanded and updated. Not only must they be designed to accept additional entries, they must also be designed to accept additional data types or additional data fields. If a database architecture is chosen that does not provide this kind of flexibility, then data acquisition bottlenecks can quickly develop, leading to countless problems down the road. Expandability also refers to the capacity to expand or enhance a database's querying capabilities, its design, its layout, and its user-friendliness. Many databases have set release dates that essentially "force" database curators into a routine of continually expanding and enhancing their databases. If one assumes that there will only be a release 1.0 for a given database, then it is almost certain that the database will soon become extinct or obsolete.

CONCLUSIONS

The information available today on phytochemicals, from chemistry and occurrence in foods to biological and health effects, is considerable, but this knowledge scattered in various literature sources is often underexploited, if not ignored. Part of this information has been included in the various databases reviewed here, and this contributes to make data more easily accessible and exploitable. With no doubt, databases are major factors of progress not only to speed the pace of research but also to make possible experiments that would otherwise not be possible. For example, comprehensive phytochemical spectra databases will allow the rapid identification of biomarkers in highly complex fingerprints such as those obtained in metabolomics experiments, a process that is still one of the main bottlenecks in such experiments. Food composition databases for phytochemicals should stimulate epidemiological research to further explore links between intake and metabolic, physiological, or health outcomes. Databases on clinical trials will allow better evaluation of the evidence on health effects of phytochemicals needed to define the still missing nutritional recommendations for phytochemicals. Another important application of phytochemical databases will be to better define priorities for research, based on predictive computational algorithms developed to estimate more accurately phytochemical intake and predict tissular exposure and biological and health properties. New hypotheses can be generated and tested theoretically by modeling or experimentally.

However, the ideal information system on phytochemicals is still missing, due to both insufficient data coverage in current databases (electronic resources are particularly scarce in the field of nutrition) and the lack of a unified system able to combine data from traditionally unrelated sources and to link databases with different structures and data types. It will be important to expand fields covered by these databases, for example, to include less studied classes of phytochemicals or new biological properties as they are discovered. It will also be essential to link the various databases to more easily connect information from different disciplines curated in different parts of the world.[70] Common ontologies and methods should be shared to collect, evaluate, analyze, and retrieve data, to guarantee easy and reliable connections between databases.

Beyond such technical issues, adequate financial support will be needed to cover the high costs of phytochemical data curation.[71] The collection of high-quality data is still largely done manually, and this requires high-level expertise in each of the areas covered. Unfortunately, biocuration of such data is still often not considered as a priority for many food scientists and nutritionists, as well as for funding bodies. This attitude should change to generate these key database resources, which should be seen as a new infrastructure needed for future experiments.

Development and implementation of new bioinformatic methods, such as automatic annotation of original literature sources, may speed biocuration processes and reduce corresponding costs. "Wiki" projects may facilitate community efforts provided that methods for data collection and evaluation are agreed upon and shared by all curators. Eventually, one might hope that journal editors or publishers in the phytochemistry and nutrition fields will encourage or even require that authors submit their data to one or more phytochemical databases as part of the publication requirements.[72] Certainly simultaneous database submission and publication has already become a common practice in the field of genomics (for gene sequences), transcriptomics (for microarray studies), and structural biology (for X-ray structures). A large concerted community effort remains to be organized to facilitate collection and exploitation of information on phytochemicals and their effects on health and to bring bioinformatics to the forefront in food science and nutrition research.

AUTHOR INFORMATION

Corresponding Author

*Phone: + 33 (0)4 72 73 80 95. E-mail: scalberta@iarc.fr.

REFERENCES

(1) Doets, E. L.; de Wit, L. S.; Dhonukshe-Rutten, R. A. M.; Cavelaars, A.; Raats, M. M.; Timotijevic, L.; Brzozowska, A.; Wijnhoven, T. M. A.; Pavlovic, M.; Totland, T. H.; Andersen, L. F.; Ruprich, J.; Pijls, L. T. J.; Ashwell, M.; Lambert, J. P.; Van't Veer, P.; De Groot, L. Current micronutrient recommendations in Europe: towards understanding their differences and similarities. *Eur. J. Nutr.* **2008**, *47*, 17–40.

(2) Fuchs, D.; Vafeiadou, K.; Hall, W. L.; Daniel, H.; Williams, C. M.; Schroot, J. H.; Wenzel, U. Proteomic biomarkers of peripheral blood mononuclear cells obtained from postmenopausal women undergoing an intervention with soy isoflavones. *Am. J. Clin. Nutr.* **2007**, *86*, 1369–1375.

(3) Fardet, A.; Llorach, R.; Martin, J.-F.; Besson, C.; Lyan, B.; Pujos, E.; Scalbert, A. A liquid chromatography-quadrupole time-of-flight (LC-QTOF)-based metabolomic approach reveals new metabolic

effects of catechin in rats fed high-fat diets. *J. Proteome Res.* **2008**, *7*, 2388–2398.

(4) Mennen, L. I.; Sapinho, D.; Ito, H.; Galan, P.; Hercberg, S.; Scalbert, A. Urinary excretion of 13 dietary flavonoids and phenolic acids in free-living healthy subjects — variability and possible use as biomarkers of polyphenol intake. *Eur. J. Clin. Nutr.* **2008**, *62*, 519–525.

(5) Manach, C.; Hubert, J.; Llorach, R.; Scalbert, A. The complex links between dietary phytochemicals and human health deciphered by metabolomics. *Mol. Nutr. Food Res.* **2009**, *53*, 1303–1315.

(6) Perez-Jimenez, J.; Hubert, J.; Ashton, K.; Hooper, L.; Cassidy, A.; Manach, C.; Williamson, G.; Scalbert, A. Urinary metabolites as biomarkers of polyphenol intake in humans — a systematic review. *Am. J. Clin. Nutr.* **2010**, *92*, 801–809.

(7) Wang, Y. L.; Xiao, J. W.; Suzek, T. O.; Zhang, J.; Wang, J. Y.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.

(8) Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.* **2009**, *37*, D26–D31.

(9) Barrett, T.; Troup, D. B.; Wilhite, S. E.; Ledoux, P.; Rudnev, D.; Evangelista, C.; Kim, I. F.; Soboleva, A.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Muertter, R. N.; Edgar, R. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.* **2009**, *37*, D885–D890.

(10) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **2007**, *35*, D301–D303.

(11) Taguchi, R.; Nishijima, M.; Shimizu, T. Basic analytical systems for lipidomics by mass spectrometry in Japan. *Lipidomics Bioactive Lipids: Mass-Spectrom.-Based Lipid Anal.* **2007**, *432*, 185–211.

(12) Okuda, S.; Yamada, T.; Hamajima, M.; Itoh, M.; Katayama, T.; Bork, P.; Goto, S.; Kanehisa, M. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **2008**, *36*, W423–W426.

(13) Schneider, M.; Lane, L.; Boutet, E.; Lieberherr, D.; Tognolli, M.; Bougueleret, L.; Baiyoch, A. The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. *J. Proteomics* **2009**, *72*, 567–573.

(14) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J. G.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; De Souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y. P.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603–D610.

(15) Church, S. M. The history of food composition databases. *Nutr. Bull.* **2006**, *31*, 15–20.

(16) Shinbo, Y.; Nakamura, Y.; Altaf-Ul-Amin, M.; Asah, H.; Kurokawa, K.; Arita, M.; Saito, K.; Ohta, D.; Shibata, D.; Kanaya, S. KNApSAcK: a comprehensive species-metabolite relationship database. In *Plant Metabolomics*; Biotechnology in Agriculture and Forestry 57; Springer: New York, 2006; pp 165−181.

(17) Yannai, S. *Dictionary of Food Compounds*; Chapman & Hall/CRC Press: Boca Raton, FL, 2004.

(18) Duke, J. A. *Handbook of Phytochemical Constituents of GRAS Herbs and Other Economic Plants*; CRC Press: Boca Raton, FL, 2001; 654 pp.

(19) USDA Database for the Flavonoid Content of Selected Foods — release 2.1, 2007; http://www.ars.usda.gov/Services/docs.htm?docid=6231.

(20) Harnly, J. M.; Doherty, R. F.; Beecher, G. R.; Holden, J. M.; Haytowitz, D. B.; Bhagwat, S.; Gebhardt, S. Flavonoid content of U.S. fruits, vegetables, and nuts. *J. Agric. Food Chem.* **2006**, *54*, 9966–9977.

(21) Neveu, V.; Perez-Jimenez, J.; Vos, F.; Crespy, V.; du Chaffaut, L.; Mennen, L.; Knox, C.; Eisner, R.; Cruz, J.; Wishart, D.; Scalbert, A. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database* **2010**, doi: 10.1093/database/bap024.

(22) Perez-Jimenez, J.; Neveu, V.; Vos, F.; Scalbert, A. Systematic analysis of the content of 502 polyphenols in 452 foods and beverages: an application of the Phenol-Explorer database. *J. Agric. Food Chem.* **2010**, *58*, 4959–4969.

(23) Harborne, J. B.; Baxter, H.; Moss, G. P. *Phytochemical Dictionary — A Handbook of Bioactive Compounds from Plants*; Taylor & Francis: London, U.K., 1999; p 976.

(24) Caspi, R.; Altman, T.; Dale, J. M.; Dreher, K.; Fulcher, C. A.; Gilham, F.; Kaipa, P.; Karthikeyan, A. S.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Paley, S.; Popescu, L.; Pujar, A.; Shearer, A. G.; Zhang, P.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2010**, *38*, D473–D479.

(25) Degtyarenko, K.; De Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcantara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2008**, *36*, D344–D350.

(26) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

(27) Caspi, R.; Foerster, H.; Fulcher, C. A.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S. Y.; Shearer, A. G.; Tissier, C.; Walk, T. C.; Zhang, P.; Karp, P. D. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **2008**, *36*, D623–D631.

(28) Ausloos, P.; Clifton, C. L.; Lias, S. G.; Mikaya, A. I.; Stein, S. E.; Tchekhovskoi, D. V.; Sparkman, O. D.; Zaikin, V.; Zhu, D. The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 287–299.

(29) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787.

(30) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Wenger, R. K.; Yao, H. Y.; Markley, J. L. BioMagResBank. *Nucleic Acids Res.* **2008**, *36*, D402–D408.

(31) Steinbeck, C.; Kuhn, S. NMRShiftDB — compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **2004**, *65*, 2711–2717.

(32) Lopez-Perez, J. L.; Theron, R.; del Olmo, E.; Diaz, D. NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. *Bioinformatics* **2007**, *23*, 3256–3257.

(33) Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmuller, E.; Dormann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A. R.; Steinhauser, D. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **2005**, *21*, 1635–1638.

(34) Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst* **2009**, *134*, 1322–1332.

(35) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **2009**, *81*, 10038–10048.

(36) Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **2010**, *38*, D355–D360.

(37) Rhee, S. Y.; Zhang, P.; Foerster, H.; Tissier, C. AraCyc: overview of an Arabidopsis Metabolism Database and its applications for plant research. In *Plant Metabolomics*; 2006; pp 141−154.

(38) Mueller, L. A.; Solow, T. H.; Taylor, N.; Skwarecki, B.; Buels, R.; Binns, J.; Lin, C. W.; Wright, M. H.; Ahrens, R.; Wang, Y.; Herbst, E. V.; Keyder, E. R.; Menda, N.; Zamir, D.; Tanksley, S. D. The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond. *Plant Physiol.* **2005**, *138*, 1310–1317.

(39) Urbanczyk-Wochniak, E.; Sumner, L. W. MedicCyc: a biochemical pathway database for *Medicago truncatula*. *Bioinformatics* **2007**, *23*, 1418–1423.

(40) Thimm, O.; Blasing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Kruger, P.; Selbig, J.; Muller, L. A.; Rhee, S. Y.; Stitt, M. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37*, 914–939.

(41) Tokimatsu, T.; Sakurai, N.; Suzuki, H.; Ohta, H.; Nishitani, K.; Koyama, T.; Umezawa, T.; Misawa, N.; Saito, K.; Shibata, D. KaPPA-View. A web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiol.* **2005**, *138*, 1289–1300.

(42) Pico, A. R.; Kelder, T.; van Iersel, M. P.; Hanspers, K.; Conklin, B. R.; Evelo, C. WikiPathways: pathway editing for the people. *PLoS Biol.* **2008**, *6*, 1403–1407.

(43) Arita, M. What can metabolomics learn from genomics and proteomics? *Curr. Opin. Biotechnol.* **2009**, *20*, 610–615.

(44) Holden, J. M. Food sampling strategies for energy intake estimates. *Am. J. Clin. Nutr.* **1995**, *62*, 1151S–1157S.

(45) Normen, L.; Ellegard, L.; Brants, H.; Dutta, P.; Andersson, H. A phytosterol database: fatty foods consumed in Sweden and the Netherlands. *J. Food Compos. Anal.* **2007**, *20*, 193–201.

(46) Arts, I. C.; van de Putte, B.; Hollman, P. C. Catechin contents of foods commonly consumed in The Netherlands. 1. Fruits, vegetables, staple foods, and processed foods. *J. Agric. Food Chem.* **2000**, *48*, 1746–1751.

(47) Kuhnle, G. G. C.; Dell'Aquila, C.; Aspinall, S. M.; Runswick, S. A.; Joosen, A. M. C. P.; Mulligan, A. A.; Bingham, S. A. Phytoestrogen content of fruits and vegetables commonly consumed in the UK based on LC-MS and ^{13}C-labelled standards. *Food Chem.* **2009**, *116*, 542–554.

(48) USDA Database for the flavonoid content of selected foods — release 2, 2006; http://www.ars.usda.gov/Services/docs.htm?docid=6231.

(49) Dwyer, J. T.; Picciano, M. F.; Betz, J. M.; Fisher, K. D.; Saldanha, L. G.; Yetley, E. A.; Coates, P. M.; Milner, J. A.; Whitted, J.; Burt, V.; Radimer, K.; Wilger, J.; Sharpless, K. E.; Holden, J. M.; Andrews, K.; Roseland, J.; Zhao, C.; Schweitzer, A.; Harnly, J.; Wolf, W. R.; Perry, C. R. Progress in developing analytical and label-based dietary supplement databases at the NIH Office of Dietary Supplements. *J. Food Compos. Anal.* **2008**, *21*, S83–S93.

(50) Nurmi, T.; Mazur, W.; Heinonen, S.; Kokkonen, J.; Adlercreutz, H. Isoflavone content of the soy based supplements. *J. Pharm. Biomed. Anal.* **2002**, *28*, 1–11.

(51) Thompson, L. U.; Boucher, B. A.; Cotterchio, M.; Kreiger, N.; Liu, Z. Dietary phytoestrogens, including isoflavones, lignans, and coumestrol, in nonvitamin, nonmineral supplements commonly consumed by women in Canada. *Nutr. Cancer—Int. J.* **2007**, *59*, 176–184.

(52) Roseland, J. M.; Holden, J. M.; Andrews, K. W.; Zhao, C.; Schweitzer, A.; Harnly, J.; Wolf, W. R.; Perry, C. R.; Dwyer, J. T.; Picciano, M. F.; Betz, J. M.; Saldanha, L. G.; Yetley, E. A.; Fisher, K. D.; Sharpless, K. E. Dietary supplement ingredient database (DSID): preliminary USDA studies on the composition of adult multivitamin/mineral supplements. *J. Food Compos. Anal.* **2008**, *21*, S69–S77.

(53) Hollman, P.; Cassidy, A.; Comte, B.; Hatzold, T.; Heinonen, M.; Richling, E.; Serafini, M.; Scalbert, A.; Sies, H.; Vidry, S. Antioxidant activity of polyphenols and cardiovascular health: application of the PASSCLAIM criteria. *J. Nutr.* **2010**, doi: 10.3945/jn.110.131490.

(54) Wu, X. L.; Beecher, G. R.; Holden, J. M.; Haytowitz, D. B.; Gebhardt, S. E.; Prior, R. L. Lipophilic and hydrophilic antioxidant capacities of common foods in the United States. *J. Agric. Food Chem.* **2004**, *52*, 4026–4037.

(55) Carlsen, M.; Halvorsen, B.; Holte, K.; Bohn, S.; Dragland, S.; Sampson, L.; Willey, C.; Senoo, H.; Umezono, Y.; Sanada, C.; Barikmo, I.; Berhe, N.; Willett, W.; Phillips, K.; Jacobs, D.; Blomhoff, R. The total antioxidant content of more than 3100 foods, beverages, spices, herbs and supplements used worldwide. *Nutr. J.* **2010**, *9*, 3.

(56) Reinivuo, H.; Bell, S.; Ovaskainen, M. L. Harmonisation of recipe calculation procedures in European food composition databases. *J. Food Compos. Anal.* **2009**, *22*, 410–413.

(57) Bell, S.; Becker, W.; Vásquez-Caicedo, A. L.; Hartmann, B. M.; Møller, A.; Butriss, J. *Report on Nutrient Losses and Gains Factors used in European Food Composition Databases*; Workpackage 1.5 on Standards Development, on behalf of the EuroFIR consortium; 2006.

(58) USDA table of nutrient retention factors — release 6, 2007; http://www.ars.usda.gov/Services/docs.htm?docid=9448.

(59) Manach, C.; Williamson, G.; Morand, C.; Scalbert, A.; Remesy, C. Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies. *Am. J. Clin. Nutr.* **2005**, *81*, 230S–242S.

(60) Verkerk, R.; Schreiner, M.; Krumbein, A.; Ciska, E.; Holst, B.; Rowland, I.; De Schrijver, R.; Hansen, M.; Gerhauser, C.; Mithen, R.; Dekker, M. Glucosinolates in *Brassica* vegetables: the influence of the food supply chain on intake, bioavailability and human health. *Mol. Nutr. Food Res.* **2009**, *53*, S219–S265.

(61) Freiser, H.; Jiang, Q. Optimization of the enzymatic hydrolysis and analysis of plasma conjugated γ-CEHC and sulfated long-chain carboxychromanols, metabolites of vitamin E. *Anal. Biochem.* **2009**, *388*, 260–265.

(62) Maiani, G.; Caston, M. J. P.; Catasta, G.; Toti, E.; Cambrodon, I. G.; Bysted, A.; Granado-Lorencio, F.; Olmedilla-Alonso, B.; Knuthsen, P.; Valoti, M.; Bohm, V.; Mayer-Miebach, E.; Behsnilian, D.; Schlemmer, U. Carotenoids: actual knowledge on food sources, intakes, stability and bioavailability and their protective role in humans. *Mol. Nutr. Food Res.* **2009**, *53*, S194–S218.

(63) Franssen-van Hal, N. L. W.; Bunschoten, J. E.; Venema, D. P.; Hollman, P. C. H.; Riss, G.; Keijer, J. Human intestinal and lung cell lines exposed to β-carotene show a large variation in intracellular levels of β-carotene and its metabolites. *Arch. Biochem. Biophys.* **2005**, *439*, 32–41.

(64) Espin, J. C.; Garcia-Conesa, M. T.; Tomas-Barberan, F. A. Nutraceuticals: facts and fiction. *Phytochemistry* **2007**, *68*, 2986–3008.

(65) Gry, J.; Black, L.; Eriksen, F. D.; Pilegaard, K.; Plumb, J.; Rhodes, M.; Sheehan, D.; Kiely, M.; Kroon, P. A. EuroFIR-BASIS — a combined composition and biological activity database for bioactive compounds in plant-based foods. *Trends Food Sci. Technol.* **2007**, *18*, 434–444.

(66) Muller, H. M.; Kenny, E. E.; Sternberg, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* **2004**, *2*, 1984–1998.

(67) Cheng, D.; Knox, C.; Young, N.; Stothard, P.; Damaraju, S.; Wishart, D. S. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* **2008**, *36*, W399–W405.

(68) Rebhan, M.; ChalifaCaspi, V.; Prilusky, J.; Lancet, D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* **1997**, *13*, 163–163.

(69) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.* **2002**, *16*, 521–533.

(70) Kind, T.; Scholz, M.; Fiehn, O. How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS ONE* **2009**, *4*.

(71) Howe, D.; Costanzo, M.; Fey, P.; Gojobori, T.; Hannick, L.; Hide, W.; Hill, D. P.; Kania, R.; Schaeffer, M.; St Pierre, S.; Twigger, S.; White, O.; Yon Rhee, S. Big data: the future of biocuration. *Nature* **2008**, *455*, 47–50.

(72) Lemay, D. G.; Zivkovic, A. M.; German, J. B. Building the bridges to bioinformatics in nutrition research. *Am. J. Clin. Nutr.* **2007**, *86*, 1261–1269.

(73) de Matos, P.; Alcantara, R.; Dekker, A.; Ennis, M.; Hastings, J.; Haug, K.; Spiteri, I.; Turner, S.; Steinbeck, C. Chemical entities of biological interest: an update. *Nucleic Acids Res.* **2010**, *38*, D249–D254.

(74) Buckingham, J. *Dictionary of Natural Products*; CRC Press: Boca Raton, FL, 1993; 8584 pp.

(75) Holden, J. M.; Eldridge, A. L.; Beecher, G. R.; Marilyn Buzzard, I.; Bhagwat, S.; Davis, C. S.; Douglass, L. W.; Gebhardt, S.; Haytowitz, D.; Schakel, S. Carotenoid content of U.S. foods: an update of the database. *J. Food Compos. Anal.* **1999**, *12*, 169–196.

(76) Kiely, M.; Faughnan, M.; Wahala, K.; Brants, H.; Mulligan, A. Phyto-oestrogen levels in foods: the design and construction of the VENUS database. *Br. J. Nutr.* **2003**, *89* (Suppl. 1), S19–S23.

(77) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. METLIN — a metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27*, 747–751.

(78) Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* **2008**, *26*, 162–164.

(79) Hooper, L.; Kroon, P. A.; Rimm, E. B.; Cohn, J. S.; Harvey, I.; Le Cornu, K. A.; Ryder, J. J.; Hall, W. L.; Cassidy, A. Flavonoids, flavonoid-rich foods, and cardiovascular risk: a meta-analysis of randomized controlled trials. *Am. J. Clin. Nutr.* **2008**, *88*, 38–50.

(80) Baker, W. L.; Baker, E. L.; Coleman, C. I. The effect of plant sterols or stanols on lipid parameters in patients with type 2 diabetes: a meta-analysis. *Diabetes Res. Clin. Pract.* **2009**, *84*, e33-7.

(81) Desch, S.; Schmidt, J.; Kobler, D.; Sonnabend, M.; Eitel, I.; Sareban, M.; Rahimi, K.; Schuler, G.; Thiele, H. Effect of cocoa products on blood pressure: systematic review and meta-analysis. *Am. J. Hypertens.* **2009**, *23*, 97–103.

(82) Li, S.-H.; Liu, X.-X.; Bai, Y.-Y.; Wang, X.-J.; Sun, K.; Chen, J.-Z.; Hui, R.-T. Effect of oral isoflavone supplementation on vascular endothelial function in postmenopausal women: a meta-analysis of randomized placebo-controlled trials. *Am. J. Clin. Nutr.* **2010**, *91*, 480–486.

(83) Williamson, G.; Manach, C. Bioavailability and bioefficacy of polyphenols in humans. II. Review of 93 intervention studies. *Am. J. Clin. Nutr.* **2005**, *81*, 243S–255S.

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

(85) Ostertag, L. M.; O'Kennedy, N.; Kroon, P. A.; Duthie, G. G.; Roos, B. d. Impact of dietary polyphenols on human platelet function — a critical review of controlled dietary intervention studies. *Mol. Nutr. Food Res.* **2010**, *54*, 60–81.