Chapter 3
Chemoinformatics Analysis and Structural Similarity Studies of Food-Related Databases

Karina Martinez-Mayorga, Terry L. Peppard, Ariadna I. Ramírez-Hernández, Diana E. Terrazas-Álvarez and José L. Medina-Franco

Chemoinformatics approaches to problem solving are commonly used in both academia and industry, and while a major focus is the pharmaceutical industry, many other sectors of the chemical industry lend themselves to it equally well. The chemoinformatic concepts, thoroughly discussed in Chap. 1 of this book, are general and can also be applied to address problems frequently encountered in food chemistry. A general strategy when applying these computational methods is to replace biological activity by a food-related property, for instance, flavor character or antioxidative activity. In many cases, the representation of the chemical structure remains the same (using, for example, molecular fingerprints, physicochemical and/or structure/substructure representations). In other words, structure-activity relationships (SAR) studies commonly conducted in medicinal chemistry for the purpose of drug discovery can be generalized to the study of structure–property relationships (SPR) for virtually any chemistry-related project [1]. Herein, we discuss representative and specific applications of methods used in chemoinformatics to mine data and characterize SPR information relevant to food chemistry. The chapter is organized into two major sections. First, we discuss exemplary applications of chemoinformatic analyses and characterization of the chemical space of compound databases. In this section, we cover major related concepts such as chemical space and molecular representation. The second section is focused on the application of similarity searching to food chemical databases.

J. L. Medina-Franco (✉) · K. Martinez-Mayorga
Departamento de Fisicoquímica, Instituto de Química, Universidad Nacional Autónoma
de México, Av. Universidad 3000, Mexico City 04510, Mexico

Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway,
Port St. Lucie, FL 34987, USA
e-mail: medinajl@unam.mx

T. L. Peppard
Robertet Flavors, Inc., 10 Colonial Dr. Piscataway, NJ 08854, USA

A. I. Ramírez-Hernández · D. E. Terrazas-Álvarez
Departamento de Fisicoquímica, Instituto de Química, Universidad Nacional Autónoma de
México, Av. Universidad 3000, Mexico City 04510, Mexico

© Springer International Publishing Switzerland 2014

K. Martinez-Mayorga, J. L. Medina-Franco (eds.), *Foodinformatics,*
DOI 10.1007/978-3-319-10226-9_3

3.1 Chemoinformatic Analyses

Chemoinformatics, “cheminformatics,” and “chemical information science” are different terms that have been coined for the common goal of applying informatics methods to solve chemical problems [2]. Chemoinformatics has also been defined as “a scientific field based on the representation of molecules as objects (graphs or vectors) in a chemical space” [3]. Further definitions are surveyed by Varnek and Baskin [3] and Willet [4]. Major aspects of chemoinformatics include the representation of chemical compounds, storing and mining information in databases, and generating and analyzing data [2].

Representation Molecular representation is at the core of chemoinformatics. There are two major types of representation: graphs and descriptor vectors. Graph-based approaches are applied to conduct structure and substructural analysis. These methods are easy to interpret and allow relatively straightforward communication with non-computational experts. Representations employing descriptor vectors are commonly used in chemoinformatics for database processing, clustering, similarity searching, and developing descriptive and predictive models of SAR; for example, QSPR/QSAR models and activity landscape models [1]. More than 5000 descriptors of different design have been developed [5]. The choice of descriptors used to analyze compound data sets gives rise to different chemical spaces.

In the food chemistry field, it has been recognized that there is a need for standardized food descriptions [6]. Food databases such as INFOODS contain free text. Representative databases relevant to the food chemistry field are presented in more detail in Chap. 9. Such databases require curation of their chemical structures as well as of the associated descriptions. Curation then involves the standardization of vocabulary, dictionaries to homogenize terms, and deletion of unnecessary wording. This is a tedious, but an important and necessary step. Relevant food databases not involving chemical structures are also in common use in the food industry. These databases may have different purposes, involving: cooking methods, ingredients, recipes, cuisine, and preparation location. In this context, the concept “food description” is used in a broad sense and applies to chemical and non-chemical databases. These databases allow for the sharing and exchange of food composition data. Some of the aspects that affect the quality of the information are: nutrient definitions, analytical methods used, and food description. The need for a “universal system” to describe and store food information has been recognized [6].

Another important aspect of food databases is that food and some food additives are, by nature, mixtures of components. For example, flavors frequently comprise or contain extracts of plants. Such mixtures and combinations of mixtures provide fertile ground for innovation. Similarly, in the search for bioactive molecules, natural products have been and continue to be a primary source of molecules with potential therapeutic effect. In fact, traditional medicine around the world is ancestral and still in use. An interesting example of this is the medicinal herb St John’s wort (*Hypericum Perforatum*) which is prescribed in some countries for the treatment for depression [7]. The chemical composition and pharmacological effect of the individual constituents have been characterized; however, the less dramatic side effects typically observed cf. standard antidepressant drugs seems to be related to the mixture's complexity.

With the aim of standardizing the description of food-related databases and its analysis, Haddad et al. [8], for example, used a structural representation consisting of 1664 odorants, and used this information for classifying odorants based on similarity measures, as explained later in this chapter.

Chemical Space The concept of chemical space has broad application not only in drug discovery but also in virtually any chemistry-related dataset. It has been pointed out that "unlike real physical space, a chemical space is not unique; each ensemble of graphs and descriptors defines its own chemical space" [3]. Chemical space has been directly compared to the cosmic universe and several definitions have been proposed in the literature [9]. For example, Virshup et al. [10] recently defined chemical space as "an *M*-dimensional Cartesian space in which compounds are located by a set of *M* physicochemical and/or chemoinformatic descriptors." Comparison of the chemical space of compound collections is important for library selection and design [11]. When designing new libraries, or screening existing libraries, it is relevant to consider the chemical space coverage of the new compounds, the structural novelty, and the pharmaceutical relevance. Systematic analysis of the chemical space of compound libraries, in particular, large collections, requires computational approaches [12]. As we recently pointed out, depending on project goals, a wide range of approaches have been developed to populate, mine, and select relevant areas of chemical space [13].

It is possible to draw a direct analogy between chemical space and flavor space. A thorough discussion of chemical space is described elsewhere [9], while a comprehensive discussion of flavor and fragrance-relevant chemical space is discussed by Reymond et al. in Chap. 2 of this book.

Chemical Databases Chemical libraries vary in nature, composition, and design, and each may serve one or more specific purposes. Compound collections used for virtual (*in silico*) screening include combinatorial libraries, commercial vendors' compounds, and natural products [14]. Molecular databases may contain hundreds, thousands, or even millions of molecules; these may be existing chemicals, or they may be hypothesized compounds, e.g., for later chemical synthesis. Libraries of existing compounds may be commercial, public domain, or proprietary.

Such chemical databases can be used for a wide variety of purposes, such as the development and systematic analysis of SAR [15] and identification of polypharmacology [16]. The constant increase in the number of molecules stored in compound databases [17] has led to the concept of chemical space (*vide supra*).

Repurposing or repositioning of chemical compounds is an approach to accelerate the identification of a new use for a compound with a pre-existing use. Repurposing can be achieved computationally or experimentally or by using a combination of the two approaches. In the pharmaceutical area, it is known as drug repurposing [18] and represents an application based on increasing evidence for the concept of *polypharmacology,* i.e., that observed clinical effects are often due to the interaction of single or multiple drugs with multiple targets [19]. Reviews and discussions are described in the literature in an integrated manner with related concepts such as polypharmacology, chemogenomics, phenotypic screening, and high-throughput *in vivo* testing [20].

A number of food phytochemicals and food-related molecular databases are available [21]. Food and food-related databases are described in more detail in Chap. 9 of this book. Major examples of public databases of chemical compounds annotated with biological activity for drug-discovery applications have been developed. Prominent examples include: BindingDB, ChEMBL, PubChem, and WOrld of Molecular BioAcTivity (WOMBAT). These databases and others described in Chap. 9 can be analyzed and compared for knowledge of chemical space coverage and potential repurposing, for example, using the concept of similarity searching.

Chemoinformatic Profiling of Chemical Databases Chemoinformatics has a fundamental role in the diversity analysis of compound collections and in the mining of chemical space. Chemoinformatic approaches designed to mine and navigate through the chemical space of compound collections is described in detail elsewhere (Chap. 1 of this book). The various approaches in conducting chemoinformatic characterization of compound libraries are mainly distinguished by the structural representations and criteria used to characterize the chemical libraries. Typically, compound databases are compared using physicochemical properties, molecular scaffolds, or structural fingerprints. Following the same or similar approaches to those used to characterize databases of interest in the pharmaceutical industry, it is possible to conduct analysis of food chemical databases.

Since these three major types of structural representation are focused on specific aspects of the structures, it is convenient to use more than one criterion for comprehensive analysis of the structural and property diversity of molecular databases. This is because each of these methods has its own strengths and weaknesses. For example, the use of whole molecule properties (holistic properties) has the advantage of being intuitive and straightforward to interpret. However, physicochemical properties do not provide information regarding structural patterns, and molecules with different chemical structures can have the same or similar physicochemical properties. Similar to physicochemical descriptors, chemotypes or scaffolds may be readily interpreted and enable easy communication with medicinal chemists and biologists. For example, scaffold analysis has led to concepts which are widely used in medicinal chemistry and drug discovery, e.g., “scaffold hopping” [22] and “privileged structures” [23]. One of the shortcomings of molecular scaffold analysis is a lack of information regarding structural similarity primarily due to the side chains cf. the inherent similarity or dissimilarity of the scaffolds themselves. An obvious solution is the analysis not only of the molecular frameworks *per se* but also of the side chains, the functional groups, and other substructural analysis strategies [24].

Molecular fingerprints are widely used and have been successfully applied to a number of chemoinformatic and computer-aided molecular applications. A challenge of some fingerprints is that they are more difficult to interpret. Also, it is well known that chemical space may be highly dependent on the types of fingerprints used to derive it. In order to reduce the dependence of chemical space on the choice of structure representation, several SAR/SPR studies have implemented consensus methods in order to combine the information encoded by different molecular representations. Use of multiple fingerprints and representations to derive consensus conclusions (e.g., *consensus activity cliffs*) has been proposed as a solution [1].

We have conducted a comprehensive chemoinformatic characterization of a subset of the Flavor and Extract Manufacturers Association (FEMA) Generally Recognized As Safe (GRAS) list of approved flavoring substances (discrete chemical entities only) [25, 26]. To this end, we employed a set of rings, atom counts (carbon, nitrogen, oxygen, sulfur, and halogen atoms), six molecular properties (octanol/water partition coefficient, polar surface area, numbers of hydrogen bond donors and acceptors, number of rotatable bonds, and molecular weight), and seven structural fingerprints of different design: MACCS keys radial fingerprints (also known as extended connectivity fingerprints), chemical hashed fingerprints (implemented in ChemAxon), atom pair (Carhart), fragment pair, pharmacophore fingerprints, and weighted Burden number. In that work, we considered a set of 2244 compounds based on the FEMA GRAS list, complete through GRAS 25 [26]. An early version of this GRAS database is briefly described in Peppard et al. [27]. This data set was compared to a database of 1713 approved drugs, two databases of natural products (with 2449 and 467 molecules, respectively) a set of 10000 commercial compounds, a database of 2116 flavors and scents, and a collection of 32357 compounds used in traditional Chinese medicine. It was concluded that the molecular size of the GRAS flavoring substances and the SuperScent database is, in general, smaller cf. members of the other databases analyzed. The lipophilicity profile of these two databases, a key property to predict human bioavailability, was similar to approved drugs. Using a visual representation of chemical space based on a principal component analysis based on the number of aromatic rings and six additional molecular properties, it was concluded that a large number of GRAS chemicals overlapped a broad region of the property space occupied by drugs. The GRAS list analyzed in that work has high structural diversity, comparable to approved drugs, natural products, and libraries of screening compounds (Table 3.1).

Table 3.1 Reference databases used to characterize and compare FEMA GRAS list (3–25) and SuperScent

Database	Content	Size
FEMA GRAS	Flavors	2244
AnalytiCon	Natural products	2449
Specs NP	Natural products	467
DrugBank	Approved drugs	1713
SpecsWD3	Approved drugs	10000
TCM	Natural products	32357
SuperScent	Flavors and fragrances	2116

3.2 Similarity Searching

Computational approaches, including those based on molecular modeling and chemoinformatics tools, are increasingly being used to help identify compounds with biological activity. In particular, *in silico* or virtual screening is a valuable means of focusing experimental efforts on filtered sets of compounds yielding a higher probability of having the desired biological activity [28]. The rationale here is that the information of the system encoded in the computational procedure will increase the probability of identifying compounds with biological activity. Hit identification using computational screening requires several interactive and iterative steps and requires a careful selection of the methods to be used. The selection of a particular approach depends on the aim of the project, the information available for the system, and the computational resources available. In addition, one needs to consider the inherent limitations of each step involved and computational cost.

Virtual screening methods can be roughly organized into two major groups, namely, ligand based and structure based [29]. Ligand-based approaches use structure-activity data from a set of known actives in order to identify candidate compounds for experimental evaluation. A common ligand-based approach is based on the molecular similarity concept, which states that structurally similar molecules are more likely to have similar biological activity [30]. Significant exceptions to this rule do occur, with so-called activity cliffs describing situations where compounds with similar structure have, unexpectedly, very different biological activity [31]. Other ligand-based methods include substructure, clustering, quantitative structure-activity relationships (QSAR), pharmacophore, and three-dimensional (3D) shape matching techniques [32].

Structure-based approaches use the 3D structure of the target, usually obtained from X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. However, in the absence of a receptor's 3D structural information, homology modeling [32] has successfully been used in virtual screening [33]. One of the most common structure-based methods is molecular docking. If information for both the experimentally active compound(s) and the 3D structure of the target are available, then the ligand- and structure-based virtual screening methods can be combined. Indeed, combining both methods increases the possibility of identifying active compounds [34].

Similarity searching is a typical ligand-based approach. Selection of the query or reference compounds in virtual screening is one of the crucial initial steps required for a successful outcome. Depending on both the dataset and the biological activity, it is possible that one or more reference compounds are associated with activity cliffs, i.e., that each might be a potential "activity cliff generator" [35]. An activity cliff generator is defined as a molecular structure that has a high probability of forming an activity cliff with molecules tested in the same biological assay. Since activity cliffs represent significant exceptions to the similarity principle, typically leading to erroneous results in similarity searching, it has recently been proposed that activity cliff generators be identified and removed from data sets before selecting reference compounds. Moreover, removal of activity cliff generators has been proposed as a general strategy, to be employed before developing predictive models such as those obtained with traditional QSAR, or other machine learning algorithms based on the similarity property principle [36].

Selection of chemical databases for similarity searching (or any other virtual screening approach) is another major component of the searching protocol. As mentioned in the previous section, a number of compound databases from different sources can be used. Notably, similarity searching can be applied to compound collections initially assembled for a different purpose, detailed above as repurposing. For example, Méndez-Lucio et al. recently conducted a 3D similarity search of DrugBank, a database of drugs approved for clinical use, with a distinct inhibitor of DNA methyltransferases, an emerging and promising epigenetic target for the treatment of cancer and other diseases [37]. The anti-inflammatory drug olsalazine was one of the most similar molecules to the reference compound, and it indeed showed hypomethylating activity based on a well-characterized live-cell imaging assay mediated by DNMT isoforms [38].

Information contained in databases is, in almost all cases, multivariate in nature; those related to food chemicals present particular challenges. One issue frequently encountered is that the chemical information is ambiguous. For example, materials may comprise a mixture of constituents, as in the case of essential oils; a mixture of isomers; or single components, but having incomplete stereochemical information. This adds to the unavoidable problem of missing information in chemical databases, such as protonation state of amino or carboxylic acid groups, prevalence of particular tautomers, etc. Moreover, these structural characteristics change depending on environment, for instance, when bound to a biological target (or targets). Since these are unavoidable and “dynamic” structural features, the preference is to ignore protonation states and consider the most stable tautomer for a given molecule.

When geometric isomers or stereoisomers are incompletely defined, one strategy is to consider all possible isomers in the computations. Alternatively, it is possible to use structural representations that do not take into account stereochemical information, although this will, of course, convey less chemical information. In the case of mixtures comprising multiple constituents, it is not possible to perform traditional chemoinformatic studies based on chemical structure (although there are studies that can be performed based purely on the nonstructural content of the databases). For such mixtures, e.g., essential oils, oleoresins, or other natural extracts, chemoinformatic studies can be performed if the composition and property description (organoleptic, biological activity, etc.) can be obtained for each constituent. In addition, the possibility of synergistic effects cannot be dismissed or, as in the case of St. John’s wort, reduce side effects (in the treatment of mood disorders) due to the composition of the herb.

Another aspect to consider when dealing with food chemical databases is the dimensionality and, often times, the non-standardized description of the chemicals. In such cases, it is necessary to first use dictionaries or lexicons to ensure the information is as homogeneous as possible. This process, which is part of the curation of the database, may require manual intervention in which case it may not be entirely unbiased. Curation also includes deletion of unnecessary wording and of duplicates.

Once these steps have been performed, the database may now have chemicals without description; these will be discarded.

A final consideration is that the cleaned-up database which contains more than one description for each chemical is multi-dimensional cf. databases of chemical compounds containing just one biological activity. A similar scenario can be seen in the case of chemical databases containing the results of multiple biological assays.

There are reports in the literature by us and also by others facing these challenges. For example, both Zarzo et al. (*vide infra*) and our group have discussed the curation and chemoinformatic description of odor and flavor databases, respectively. Regarding the analysis of chemical structures, we performed structural similarity of chemical structures based on fingerprint representations. In this arena, Sprous et al. [39], Pintore et al. [40], and Jensen et al. [41] have reported related studies.

Zarzo et al. [42] characterized an odor database; the first step consisted of encoding the odor description of the database in a dichotomic format, where 0 corresponded to the absence of a given descriptor, while 1 represented its presence. From those data, the authors were able to perform a descriptive analysis of the database and show the incidence of each descriptor in the database. They also demonstrated associations among descriptors, in other words, pairs of descriptors that repeatedly were used together in the database. Lastly, using principal component analysis on a selected subset of the database, the authors constructed the corresponding "odor space." The 2D graphical representation of this odor space organized descriptors in the same regions of the plot that are intuitively similar, such as fruity (pineapple, berry, peach, cherry, apple, etc.), floral (rose, sweet, other floral), etc. One of the outcomes of this work was the presentation of an odor space which provides useful information when training sensory panels for odor profiling.

We performed a chemoinformatic analysis of the FEMA-GRAS list (containing both chemical structures and associated sensory attributes), the first steps of which comprised the compilation and curation of the database [25]. After standardization of descriptive flavor terms using a recognized sensory lexicon (ASTM, American Society for Testing and Materials publication DS 66) and removal of unnecessary wording, the resultant database was analyzed for the incidence of descriptors and their associations using three independent methods: principal component analysis, clustering, and flavor descriptor relationships. We found that certain descriptors appear in the same region of the flavor space generated with the principal component analysis, as well as within nearby clusters when generating a clustering-based heat map, and also in a pair-wise analysis of descriptor associations. The correspondence of results obtained with these three methods gives confidence in the results.

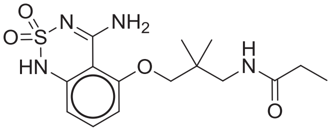
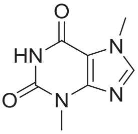
The concept of information content, commonly used in the field of chemoinformatics, has been applied to olfactory databases by Pintore et al. [40]. The challenge of establishing a standard olfactory description of chemicals is recognized by the authors. Two olfactory databases were compared, according to the consistency of odor description. Based on 2D representations, the authors applied several classification methods, along with corresponding means of validation. The authors related this consistency to the information content of the databases, and concluded that one of the main difficulties when working with odor databases is the subjectivity used, even by experts, to describe odor perception. Not surprisingly, this led to some wide discrepancies in descriptions of the same compound in the two databases. In this study, the 2D representations of the chemical structures included in the two databases were used to explore the consistency of the odor descriptions rather than to perform structural similarity with the aim of finding either similar compounds for structure–property relationships, or compounds with similar property profiles (biological activity, odor description, etc.).

Sprous and Salemme [39] reported a comparison of the FEMA GRAS compounds with compounds contained in the Drugbank database. The study was based on determining the chemoinformatic profile of the database (*vide supra*), computing the population of structural and physicochemical features, such as molecular weight, molecular flexibility, logP, logS, and numbers of acceptor, donor, acidic and basic atoms, etc. The authors concluded that, in general, GRAS compounds occupy a different and identifiable region of chemical space relative to pharmaceuticals. However, more recent subsets of the GRAS list, which contain fewer compounds from natural sources, are more diverse, thus expanding the chemical space occupied by compounds of previous versions of the FEMA/GRAS list.

Haddad et al. [8] developed a metric for odorant comparison based on a chemical space constructed from 1664 molecular descriptors. A refined version of this metric was devised following the elimination of redundant descriptors. The study included the comparison with models previously reported for nine datasets. The final, so-called multidimensional metric, based on Euclidean distances measured in a 32-descriptor space, was more efficient at classifying odorants cf. reference models previously reported. Thus, this study demonstrated the use of structural similarity for the classification of odors in multidimensional space.

In order to identify potential bioactivity among the food-flavoring components that comprise the FEMA GRAS list, we recently conducted ligand-based virtual screening for compounds with structures similar to approved antidepressant drugs [43]. The virtual screening was performed by means of fingerprint-based similarity searching. Valproic acid turned out to be the most similar antidepressant to a small number of GRAS compounds. Guided by the hypothesis that the inhibition of histone deacetylase-1 (HDAC1) may be associated with the efficacy of valproic acid in the treatment of bipolar disorder, we screened the GRAS compounds most similar to valproic acid for HDAC1 inhibition. The GRAS chemicals nonanoic acid and 2-decenoic acid inhibited HDAC1 at the micromolar level, with potency comparable to that of valproic acid. GRAS compounds likely do not exhibit strong enzymatic inhibitory effects at the concentrations typically employed in foods and beverages. As shown in that study, GRAS chemicals are able to bind, albeit weakly, to important therapeutic targets. Additional studies on bioavailability, toxicity at higher concentrations (GRAS flavor molecules being safe when used at or below the levels approved for foods and beverages) and off-target effects are warranted. The results of that work demonstrate that similarity searching followed by experimental evaluation can be used for rapid identification of GRAS chemicals with possible biological activity, with potential application for promoting health and wellness [43].

Table 3.2 GRAS flavor chemicals with highest similarity to known analgesics

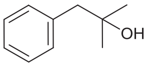
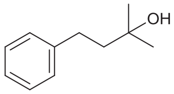
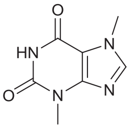
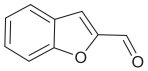
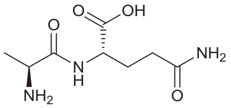
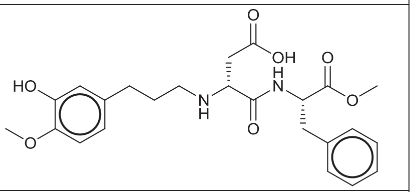
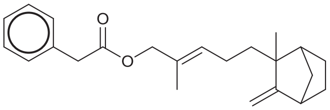
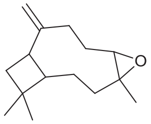
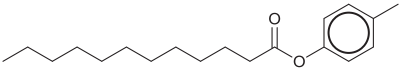
CAS #	Name	Structure
1093200-92-0	N-[(4-Amino-2,2-dioxido-1H-2,1,3-benzothiadiazin-5-yl)oxy)]-2,2-dimethyl-N-propylpropanamide	
83-67-0	Theobromine	

In two subsequent studies, again using structural similarity, we compared the FEMA GRAS list with analgesics and with compounds used as satiety agents. The list of analgesics comprised ten structurally diverse molecules currently used in the clinic. A total of eight satiety agents were identified in the literature, and these were used for similarity searching. The satiety agents included those currently used in the clinic, as well as those still in clinical trials.

In both studies, reference compounds were compared with the FEMA GRAS list using three software programs (MOE, ChemAxon, and PowerMV), with a total of seven structural representations. Compounds identified by different programs and representations were chosen as consensus compounds for further study. Then, a chemical space was constructed based on physicochemical properties. Nearest neighbors were identified based on Euclidian distances considering all the dimensions (properties). Based on the comparison of structural features and physicochemical properties, two FEMA GRAS compounds (listed on Table 3.2)were identified as similar to the reference analgesics. In the second study, a total of nine FEMA GRAS compounds were identified as similar to those used as reference satiety agents (see Table 3.3). For compounds having a known mode of action, *in vitro* studies using the identified GRAS chemicals could help determine whether or not they may have a satiety or analgesic effect in humans. However, it must be borne in mind that biological effects, in the large majority of cases, result from complex and multiple interactions in the body, as already described above in the area of polypharmacology.

Phytochemicals derived from eatable plants represent a remarkable source of bioactive compounds. In a recent study, Jensen et al. [41] performed a high-throughput analysis of phytochemicals in order to uncover associations between diet and health benefits using text mining and chemoinformatic methods. The first step of that study involved the extraction of associations between the terms of plants and phytochemicals, analyzing 21 million abstracts in PubMed/MEDLINE covering the period 1998–2012. This information was merged with the Chinese Natural Product Database and the Ayurveda dataset, which was also curated by the authors. The final dataset contained almost 37000 phytochemicals. A remarkable outcome

Table 3.3 GRAS flavor chemicals with highest similarity to known satiety agents

CAS #	Name	Structure
100-86-7	2-mehtyl-1-phenylpropan-2-ol	
103-05-9	2-Methyl-4-phenyl-2-butanol	
83-67-0	Theobromine	
4265-16-1	2-Benzofurancarboxaldehyde	
39537-23-0	L-Alanyl-L-glutamine	
714229-20-6	Advantame	
1323-75-7	(2Z)-2-Mehtyl-5-{2-methyl-3-methylidenebicyclo[2.2.1]heptan-2-yl}pent-2-en-1-yl 2-phenylacetate	
1139-30-6	(1R,4R,6R,10S)-9-Methylene-4,12,12-trimethyl-5-oxatricyclo[8.2.0.04,6]dodecane	
10024-57-4	(4-Methylphenyl) dodecanoate	

of that work is the structured and standardized database of phytochemicals associated with medicinal plants. As claimed by the authors, their approach facilitates the identification of novel bioactive compounds from natural sources, and the repurposing of medicinal plants for diseases other than those traditionally used for, with the added benefit that the information collected can help elucidate mechanism of action [41]. As a case study, the authors applied structural similarity searching in order to find molecules in their compiled database of phytochemicals with activity against a protein involved in the colon cancer pathway or a colon cancer drug target; the reference compounds were those reported in the ChEMBL database. A set of molecules from this study have not only reported health benefit against colon cancer but also verified activity against colon cancer protein targets.

The studies here described exemplify the application of the concepts and methodologies widely used in pharmaceutical settings, such as of data mining, diversity analysis, polypharmacology, repurposing, and similarity searching, in databases containing food additives and phytochemicals.

Acknowledgments K.M-M. thanks the Institute of Chemistry-UNAM and DGAPA-UNAM for funding (PAPIIT IA200513). The authors also wish to thank Robertet Flavors for permission to publish this chapter.

References

1. Medina-Franco JL, Yongye AB, López-Vallejo F (2012) Consensus models of activity landscapes. In: Matthias D, Kurt V, Danail B (eds) Statistical modeling of molecular descriptors in QSAR/QSPR. Wiley-VCH, Weinheim, pp 307–326
2. Engel T (2006) Basic overview of chemoinformatics. J Chem Inf Model 46:2267–2277
3. Varnek A, Baskin II (2011) Chemoinformatics as a theoretical chemistry discipline. Mol Inf 30:20–32
4. Willett P (2011) Chemoinformatics: a history. WIREs Comput Mol Sci 1:46–56
5. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
6. Pennington JT (2006) Issues of food description. Food Chem 57:145–148
7. Caccia S, Gobbi M (2009) St. John's wort components and the brain: uptake, concentrations reached and the mechanisms underlying pharmacological effects. Curr Drug Metab 10:1055–1065
8. Haddad R, Khan R, Takahashi YK, Mori K, Harel D, Sobel N (2008) A metric for odorant comparison. Nat Methods 5:425–429
9. Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C (2008) Visualization of the chemical space in drug discovery. Curr Comput Aided Drug Des 4:322–333
10. Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN (2013) Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. J Am Chem Soc 135:7296–7303
11. Fitzgerald SH, Sabat M, Geysen HM (2006) Diversity space and its application to library selection and design. J Chem Inf Model 46:1588–1597
12. Akella LB, DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol 14:325–330
13. Medina-Franco JL, Martinez-Mayorga K, Meurice N (2014) Balancing novelty with confined chemical space in modern drug discovery. Expert Opin Drug Discov 9:151–165

14. Harvey AL (2008) Natural products in drug discovery. Drug Discov Today 13:894–901
15. Scior T, Bernard P, Medina-Franco JL, Maggiora GM (2007) Large compound databases for structure-activity relationships studies in drug discovery. Mini Rev Med Chem 7:851–860
16. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690
17. Gozalbes R (2011) Rational generation of focused chemical libraries: an update on computational approaches. Comb Chem High Throughput Screen 14:428–428
18. Ashburn TT, Thor KB (2004) Drug repositioning: Identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3:673–683
19. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24:805–815
20. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA (2013) Shifting from the single to the multi target paradigm in drug discovery. Drug Discov Today 18:495–501
21. Scalbert A, Andres-Lacueva C, Arita M, Kroon P, Manach C, Urpi-Sarda M, Wishart D (2011) Databases on food phytochemicals and their health-promoting effects. J Agric Food Chem 59:4331–4348
22. Schneider G, Neidhart W, Giller T, Schmid G (1999) Scaffold-hopping by topological pharmacophore search: a contribution to virtual screening. Angew Chem Int Ed 38:2894–2896
23. Duarte CD, Barreiro EJ, Fraga CA (2007) Privileged structures: a useful concept for the rational design of new lead drug candidates. Mini Rev Med Chem 7:1108–1119
24. Villar HO, Hansen MR, Kho R (2007) Substructural analysis in drug discovery. Curr Comput Aided Drug Des 3:59–67
25. Martínez-Mayorga K, Peppard TL, Yongye AB, Santos R, Giulianotti M, Medina-Franco JL (2011) Characterization of a comprehensive flavor database. J Chemom 25:550–560
26. Medina-Franco JL, Martínez-Mayorga K, Peppard TL, Del Rio A (2012) Chemoinformatic analysis of GRAS (Generally Recognized as Safe) flavor chemicals and natural products. PLoS One 7:e50798
27. Peppard TL, Le M, Pandya RN (2008) Prediction tool for modern flavor development. In: Hofmann T, Meyerhof W, Schieberle P (eds) Recent Highlights in Flavor Chemistry & Biology. Proceedings of the 8th Wartburg Symposium on flavour chemistry and biology. Deutsche Forschungsanstalt für Lebensmittelchemie, Garching, pp 374–378
28. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK (2012) Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 52:867–881
29. Alvarez J, Shoichet B (2005) Virtual screening in drug discovery. Taylor & Francis Group, LLC CRC Press, Boca Raton
30. Maldonado AG, Doucet JP, Petitjean M, Fan BT (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers 10:39–79
31. Maggiora GM (2006) On outliers and activity cliffs-why QSAR often disappoints. J Chem Inf Model 46:1535
32. Villoutreix BO, Renault N, Lagorce D, Sperandio O, Montes M, Miteva MA (2007) Free resources to assist structure-based virtual ligand screening experiments. Curr Protein Pept Sci 8:381–411
33. Radestock S, Weil T, Renner S (2008) Homology model-based virtual screening for GPCR ligands using docking and target-biased scoring. J Chem Inf Model 48:1104–1117
34. Kruger DM, Evers A (2010) Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. Chemmedchem 5:148–158
35. Mendez-Lucio O, Perez-Villanueva J, Castillo R, Medina-Franco JL (2012) Identifying activity cliff generators of PPAR ligands using SAS maps. Mol Inf 31:837–846
36. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F (2014) Activity cliffs in drug discovery: Dr. Jekyll or Mr. Hyde? Drug Discov Today 19:1069–1080
37. Rius M, Lyko F (2012) Epigenetic cancer therapy: rationales, targets and drugs. Oncogene 31:4257–4265

38. Méndez-Lucio O, Tran J, Medina-Franco JL, Meurice N, Muller M (2014) Towards drug repurposing in epigenetics: olsalazine as a novel hypomethylating compound active in a cellular context. ChemMedChem 9:560–565
39. Sprous DG, Salemme FR (2007) A comparison of the chemical properties of drugs and FEMA/FDA notified GRAS chemical compounds used in the food industry. Food Chem Toxicol 45:1419–1427
40. Pintore M, Wechman C, Sicard G, Chastrette M, Amaury N, Chretien JR (2006) Comparing the information content of two large olfactory databases. J Chem Inf Model 46:32–38
41. Jensen K, Panagiotou G, Kouskoumvekaki I (2014) Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. PLoS One 10:e1003432
42. Zarzo M, Stanton DT (2006) Identification of latent variables in a semantic odor profile database using principal component analysis. Chem Senses 31:713–724.
43. Martinez-Mayorga K, Peppard TL, López-Vallejo F, Yongye AB, Medina-Franco JL (2013) Systematic mining of generally recognized as safe (GRAS) flavor chemicals for bioactive compounds. J Agric Food Chem 61:7507–7514