How to Valorize Biodiversity? Let’s Go Hashing, Extracting, Filtering, Mining, Fishing

Quoc Tuan Do¹, José L. Medina-Franco², Thomas Scior³, Philippe Bernard¹

¹ Greenpharma S. A. S., Orléans, France
² Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Mexico City, Mexico
³ Department of Pharmacy, Benemérita Universidad Autónoma de Puebla, Puebla, México

Abstract

Nature was and still is a prolific source of inspiration in pharmacy, cosmetics, and agro-food industries for the discovery of bioactive products. Informatics is now present in most human activities. Research in natural products is no exception. In silico tools may help in numerous cases when studying natural substances: in pharmacognosy, to store and structure the large and increasing number of data, and to facilitate or accelerate the analysis of natural products in regards to traditional uses of natural resources; in drug discovery, to rationally design libraries for screening natural compound mimetics and identification of biological activities for natural products. Here we review different aspects of in silico approaches applied to the research and development of bioactive substances and give examples of using nature-inspiring power and ultimately valorize biodiversity.

Abbreviations

- ADMET: absorption, distribution, metabolism, excretion, toxicity
- ANN: artificial neural networks
- BBB: blood-brain barrier
- COX: cyclooxygenase
- 1D, 2D, 3D: one-, two-, three-dimensional
- DNMT: DNA methyltransferase
- EPS: electrostatic potentials
- FAK: focal adhesion kinase
- FEMA: Flavor and Extract Manufacturers Association
- GRAS: generally recognized as safe
- HERG K+: human ether-a-go-go-related gene potassium
- HDAC1: histone deacetylase-1
- HTS: high-throughput screening
- MOE: molecular operating environment
- NABATIVI: Novel Approaches to Bacterial Target Identification Validation and Inhibition
- PCA: principle component analysis
- PD: pharmacodynamic
- PESD: properties encoded shape distributions
- PK: pharmacokinetic
- PLA2: phospholipase A2
- PPAR: peroxisome proliferator-activated receptors
- QSAR: quantitative structure-activity relationship
- R&D: research and development
- SAR: structure-activity relationships
- SPID: structure-promiscuity index difference
- SVM: support vector machine
- UFSR: ultrafast shape recognition
- vHTS: virtual HTS
- VLS: virtual library screening
- ZINC: free database of commercially available compounds for virtual screening

Introduction

“Biological diversity’ or ‘biodiversity’ means the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are a part; this includes diversity within species, between species and of ecosystems” [1]. Biodiversity is endangered by human activities, and its decline in some regions is exacerbated by climate changes. The extent of such modifications of the environment depends on complex criteria including geographical, environmental, political, and societal conditions [2, 3].
This makes global protection policies, though necessary, very difficult to implement. Why and how to protect biodiversity? Some encouraging attempts have been made by regional actors to stimulate industry, nonprofit, and academic research in chemical and life sciences [4], national governments for biodiversity protection and societal issues [5] or, at a continental level such as the European FP7 project Marex composed of nine European Union countries and four developing countries, to explore the possible industrialization of bioactive substances from marine resources [6]. Ethical and utilitarian arguments are the common points from the numerous existing examples. Sustainable developments can be envisaged to valorize biodiversity (i.e., to estimate economic value, to highlight its value and/or to increase its value) in a variety of domains such as biofuels to replace fossil energy [7], materials, e.g., batteries made with emodin derivatives [8], biomimetics, i.e., nature as a source of inspiration to design new materials, processes, etc... [9], cosmetics [10], functional foods [11], or pharmacy [12].

Historically, natural products, e.g., plants, have been a source of food and medicine; as a matter of fact, in ancient civilizations, the two things are “interchangeable” according to Hippocrates. Therefore, a lot of knowledge was accumulated as evidenced by traditional Chinese medicine’s so-called “Yellow Emperor’s Inner Classic” or Dioscorides’ “De Materia Medica”, to cite a few. Nowadays, with the evolution of analytical techniques [13] and the fast development and advances of computers, the amount of data about natural products has grown drastically [14–16]. This allows for the emergence of new strategies to valorize the natural products, such as reverse pharmacognosy [17,18], by working with natural flavor molecules [19,20], by relating traditional products, such as reverse pharmacognosy [17,18], by working with natural flavor molecules [19,20], by relating traditional medicine concepts to modern Western medicine pathologies [21], or by using ancestral knowledge as a starting point for scientific investigations [22]. In this review, we focus on the contributions of natural products in drug discovery aided by in silico techniques with a particular emphasis on the most frequently used approaches such as database mining, systematic screening using similarity searching and molecular docking, and inverse docking techniques.
intervention. The ultimate goal of pharmaceutical profiling is to predict potential drawbacks concerning critical issues of PD and PK as well as the development of trial dosage forms or final delivery systems as early as possible and to evaluate drug usage and security risks in general [23].

Medicinal chemistry textbooks contain some popular rules of thumb like the empiric replacement patterns for chemical groups known as biostereo-isomerism, the traffic light scheme for洛尔’s “Oral PhysChem Score” [32] or “Lipinski’s rule of five” [33], which can be embedded in drug-likeness screens or biopharmaceutical profiling efforts (© Table 1).

### Table 1: Empirical descriptors or patterns for a typical biopharmaceutical profile.

<table>
<thead>
<tr>
<th>Descriptors/parameters/patterns or features</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>MW &lt; 500</td>
<td>[33]</td>
</tr>
<tr>
<td>logP &lt; 5</td>
<td>[33]</td>
</tr>
<tr>
<td>Hydrogen bond donors (HBD) &lt; 5</td>
<td>[33]</td>
</tr>
<tr>
<td>Hydrogen bond acceptors (HBA) &lt; 10</td>
<td>[33]</td>
</tr>
<tr>
<td>Number of rotatable bonds (nb) &lt; 10</td>
<td>[34–36]</td>
</tr>
<tr>
<td>Solubility logS at pH 6.5 &gt; 10 mg/L</td>
<td>[32]</td>
</tr>
<tr>
<td>Topological polar surface area (TPSA) &lt; 140 Å²</td>
<td>[37, 38]</td>
</tr>
<tr>
<td>Aromatic rings &lt; 4</td>
<td>[39]</td>
</tr>
<tr>
<td>GI tract and BBB permeability decreases with a lower log D &lt; 0</td>
<td>[40]</td>
</tr>
<tr>
<td>Water solubility and renal excretion increases with a lower log D &lt; 0</td>
<td>[40]</td>
</tr>
<tr>
<td>Water solubility and membrane permeability are “drug-like” in a range of log D between 0 to 3 units (0 &lt; log D &lt; 3)</td>
<td>[40]</td>
</tr>
<tr>
<td>Water solubility is increased by polar groups, hydrogen bonding, or dissociation into ions or permanent ionization (cations, anions)</td>
<td>[40]</td>
</tr>
<tr>
<td>Potency increases with higher log D &gt; 5</td>
<td>[40]</td>
</tr>
<tr>
<td>Hepatic biotransformation by CYP450 increases with a higher log D &gt; 5</td>
<td>[40]</td>
</tr>
<tr>
<td>Water solubility, oral absorption, and bioavailability tend to decrease with a higher log D &gt; 5</td>
<td>[40]</td>
</tr>
<tr>
<td>Water solubility increases, and lipophilicity and membrane permeability (by passive diffusion of a given drug) diminish</td>
<td>[40]</td>
</tr>
<tr>
<td>Esters and other produg solutions increase lipophilicity if the acidic drug is too hydrophilic</td>
<td>[40]</td>
</tr>
</tbody>
</table>

### Table 2: Natural product database.

<table>
<thead>
<tr>
<th>Database name</th>
<th>Accessibility</th>
<th>Data types</th>
<th>Advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>AfroDB [60]</td>
<td>Freely accessible from the supplementary information of [60]</td>
<td>1000 Compounds; physicochemical data and ADMET properties</td>
<td>Comprehensive predicted data</td>
<td>No data on plants</td>
</tr>
<tr>
<td>ChemNetBase [62]</td>
<td>Searches are free; results browsing under license <a href="http://dnp.chemnetbase.com">http://dnp.chemnetbase.com</a></td>
<td>170 000 Natural compounds</td>
<td>Very comprehensive; frequently updated</td>
<td>Lack of organism data; commercial database</td>
</tr>
<tr>
<td>Dr Duke’s database [63]</td>
<td>Freely accessible <a href="http://www.ars-grin.gov/duke">http://www.ars-grin.gov/duke</a></td>
<td>7500 Molecules, 2000 organisms, 2200 traditional uses; biological activities</td>
<td>Many ways to query the database; huge amount of data</td>
<td>Lack of molecule data (structures, etc.); not updated since 1998</td>
</tr>
<tr>
<td>GPDB [10]</td>
<td>Greenpharma internal search</td>
<td>140 000 Compounds, 160 000 organisms, 4360 targets, 10 000 activities, 1000 traditional uses</td>
<td>Rich query system; structural searches; numerous links between data</td>
<td>Lack of data, but very frequently updated</td>
</tr>
<tr>
<td>KNApSACK [64]</td>
<td>Freely accessible <a href="http://kanaya.naist.jp/knapsack.jsp/top.html">http://kanaya.naist.jp/knapsack.jsp/top.html</a></td>
<td>51 000 Molecules, 22 000 organisms, 110 000 metabolite/species pairs</td>
<td>Frequently updated; query to database can be implemented in software</td>
<td>No structural search</td>
</tr>
<tr>
<td>Nap Alert [65]</td>
<td>Searches are free, but pay per view for results report <a href="http://www.napralert.org">http://www.napralert.org</a></td>
<td>200 000 Publications annotated; organisms; molecules; biological activities; ethnomedical and medicinal data</td>
<td>Very comprehensive; frequently updated</td>
<td>Lack of molecule data; commercial database; lack of flexibility in results presentation</td>
</tr>
<tr>
<td>Pfaf [66]</td>
<td>Freely accessible <a href="http://www.pfaf.org">http://www.pfaf.org</a></td>
<td>7000 Plants; traditional uses; medical and edible quality scores</td>
<td>Seldom used and original plants; highly suited to RPG</td>
<td>No molecule data</td>
</tr>
<tr>
<td>Supernatural [67]</td>
<td>Freely accessible <a href="http://bioinformatics.charite.de/supernatural">http://bioinformatics.charite.de/supernatural</a></td>
<td>46 000 Natural compounds; molecular characteristics; supplier data</td>
<td>Similarity searches</td>
<td>No organism data</td>
</tr>
<tr>
<td>TCM-ID [68]</td>
<td>Freely accessible <a href="http://bidd.nus.edu.sg/group/TCMsite/Default.aspx">http://bidd.nus.edu.sg/group/TCMsite/Default.aspx</a></td>
<td>12 000 Compounds; 1100 plants; 1200 TCM formula</td>
<td>Interesting relation with TCM-molecules</td>
<td>Cannot be exported</td>
</tr>
<tr>
<td>UNPD [69]</td>
<td>Free accessible <a href="http://pkuxj.pku.edu.cn/UNPD">http://pkuxj.pku.edu.cn/UNPD</a></td>
<td>200 000 Compounds</td>
<td>Largest noncommercial and freely available database for natural products</td>
<td>No data on organisms</td>
</tr>
</tbody>
</table>
Computer-based profiling

Recently, a theoretical study demonstrated the toxicological characterization of a series of chemicals with cheminformatics. To this end, the cytotoxicity profile was estimated on the basis of structural molecular fragments to identify several moieties that can be regarded as bearing cell toxicity (cytotoxicophores) [41]. The detection of fragments with proven or alleged toxic properties, so-called toxicophores, can be carried out on the Web-based server Ochem (https://ochem.eu/) via a link to ToxA- lerts. The server also provides QSAR modeling. Another helpful, almost all-in-one solution is Vega ZZ for 3D model generation, bi- molecules, manual docking, empirical molecular mechanics force field calculations or semiempirical quantum mechanics, to list only a few features [42]. The use is free for public universities and not-for-profit research institutes. As a general rule for the software novice, computed values can be used with confidence if the compound lies within the applicability domain (scope) of a program [43,44]. Software algorithms looking up databases or parameter sets when applying empirical equations are more sus- ceptible than first principle ab initio methods. The latter are not foolproof either and can fail, too, if the underlying theory does not reflect natural processes. In general, conventional small or- ganic molecules are more likely to be in the applicability or “cali- bration” range. Other structures fall short of expectations be- cause they possess noncanonical electronic constellations, like carbamoyl, azid, nitro, sulfon, and metal organic groups or they are hydrazones, thioesters, etc. Sometimes, electronic, mesomer- ic effects depend on the conformation (between bridged aryl rings), and pK_a predictions tend to fail. Recent approaches suc- cessfully applied classification models for drug profiling in combi- nation with public databases (PubChem [27,45], AntiMarin da- tabase [46]). The success was documented for some modeled ac- tivities that could be found in the literature and thereby con- firmed [47].

Newman and Cragg have reviewed the contributions of natural products as sources of new drugs for three decades from 1981 to 2010 [12,48,49]. They are still important because they provide the final entity or are starting points in drug discovery (mimetics, derivatives, botanicals, etc.), particularly in oncology and infec- tion domains. To date, only one de novo drug obtained from com- binatorial chemistry has been approved during the reviewed pe- riod. What makes natural products so successful? What lessons can the medicinal chemist learn from natural products and their properties?

In the pharmaceutical industry the attrition rate remains very high [50], particularly at the later stage of expensive clinical trials. Therefore a “fail early, fail cheap” paradigm represents an attrac- tive strategy. Many investigators have tried to capture the es- sence of existing drugs to extrapolate physicochemical criteria with the ease of implementation along the drug discovery work- flow. Table 1 lists examples of the different empirical descrip- tors derived from statistical mining of drug databases. Thanks to their numerical nature they can be used as prefilters in virtual li- brary screenings, QSAR studies [28,51], or in compound selection in general. However, one must be cautious about their use. For in- stance, Lipinski’s rule of five of “drug likeness” (oral delivery and passive absorption mechanisms) [33], though widely used, is not applicable to natural products and does screen out many drugs derived from natural products. Keller et al. [52] hypothesized that natural compounds may have evolved over millennia to take ad- vantage of active transport or gained specific conformations suited to passive transport. For Kellenberger et al., the reason of natural drugs may be the similarity of interactions of natural products with biosynthetic enzymes and therapeutic targets [53]. As a large number of drugs are derived from natural sources, we see an overlap of the drug chemical space and the natural compound space [54]. Indeed, several authors have advocated mimicking certain physicochemical profiles of natural com- pounds to synthesize compounds that are more diverse and bio- logically relevant [54–57], e.g., a reduced number of nitrogen atoms or aromatic rings, the presence of nonaromatic, polycyclic core structures, etc.

HTS assays can be perturbed by certain chemical features, gener- ating a false positive. Rishton surveyed such reactive functional groups in [58]: among them some can be frequently found in ac- tive natural compounds and drugs, such as aldehydes, aliphatic esters and ketones, epoxides, 1,2-dicarbonyl compounds (tanshi- nones), Michael acceptors (chalcones), peroxides (artemisin deriva- tives), and disulfides (glutathione disulfides). Some natural compounds may therefore be filtered out because they are not suitable for HTS.

Data Mining

The scientific literature search, storing, and exploitation are im- perative and have come to terms with data mining in the modern ages of electronic information technologies. Information that is produced and kept in-house (corporate data sources) is not pub- licly available. Proprietary data can be used by costumers on a commercial basis, while other sources lay open on the Internet (free web services). Helpful web sites to assist the profiling phases are scientific journals, patents, and bioinformatics ser- vices dealing with genomics, proteomics, and metabolomics. The literature survey compiles relevant data on comparative or disease-associated genetics, pharmacogenetics, pharmacody- namics, pharmacokinetics, metabolic and cellular signaling path- ways, in vitro and ex vivo (cell based) pathophysiological models, etc. [59]. Databases on traditional usage of natural substances are also booming with descriptions from the organisms to the mole- cules with folk uses, biological data, and predicted properties. Ethnopharmacology data offers valuable “clinical” observations that can guide the drug discovery process (see Table 2 for ex- amples of database). Bernard et al. [19] gathered a list of plants used in several populations of Latin America in cases of insect stings or snakebites to find new anti-inflammatory agents target- ing PLA2. Extracts of plants used by several populations for these ailments were tested in priority on PLA2. By searching for com- pounds that were common to the active extracts in their in- house data, the authors could narrow the possible can- didates and perform docking on PLA2 to retrieve betulin and betu- linic acid as potent inhibitors of PLA2. This prediction was further validated by in vitro binding tests. This work demonstrated the usefulness of computer-based analysis of ancestral knowledge to guide and accelerate the modern drug discovery process. More- over, it also enabled experiments to prove some intuitive rela- tions between folk medicine concepts (insect stings) and modern medicine pathology (inflammation). Rollinger et al. [21] have demonstrated the efficiency on combining in silico techniques with ethnopharmacological knowledge. Molecules from plants listed in Dioscorides “De Materia Medica” as having “anti-inflam- matory properties” were screened on structure-based pharma- cophore models of COX 1 and 2. The hit rate using this procedure was about 100% higher compared to the same virtual screening
but on molecules from databases comprising marketed and development drug substances or natural compounds.

More recently, an interesting initiative by Ntie-Kang et al. [60] offered access to a database of more than 1000 natural compounds isolated from African medicinal plants, called AfroDB. The authors calculated numerous descriptors of drug-, lead-, and fragment-likeliness and ADMET. They predicted the following ADMET properties: bioavailability, BBB penetration, dermal penetration, plasma-protein binding, metabolism, and blockage of the HERG K+ channel. These parameters were also made available to the research community to help with compound selection, comparison, and virtual screening. The p-ANAPL library [61] containing most of the compounds of AfroDB was supplied upon request to the authors for in vitro validation. This type of initiative will no doubt encourage drug discovery from African plants and collaborations to valorize these resources.

The wealth of structural data on natural compounds allows investigators to compare drugs with natural compounds, and vice versa, to derive SAR and, subsequently, to deduce putative biological activities. Similarity searching is a fertile approach in drug discovery.

**Similarity Searching**

Similarity-based screening or similarity searching is a typical ligand-based approach that can be conducted without prior knowledge of the 3D structure of the target. This approach is based on the notion that similar compounds have similar activity [70]. Remarkable exceptions to this concept are the “activity cliffs”, i.e., similar compounds with an unexpectedly high activity difference. The interested reader is referred to reviews that address in detail the role of activity cliffs in medicinal chemistry and elaborate on the computational approaches to identify them [71,72].

Similar to other computational screening efforts, similarity searching should be part of an iterative process that involves the prediction, experimental testing of selected compounds, and design of new chemical data sets based on the structure of the experimental hits. Also, if enough information of the system is available, e.g., 3D coordinates of the target, similarity searching should be combined with other ligand-based and/or structure-based methods. The selection of a particular approach or set of methods depends on the aim of the project, the information of the system, and the computational resources available. Moreover, one needs to consider the inherent limitations of each step involved and the associated computational cost.

In natural products research, the combination of computational approaches has been emphasized by Yue et al., who have recently discussed progress on the target profiling of natural products using experimental (genomics and proteomics) and computational approaches [73]. In that review, Yue et al. emphasized the convenience of integrating various methods, such as inverse docking (docking compounds across different targets), mapping ligand-target profiling space, and network analysis. Similarity searching can be combined with other current major strategies in drug discovery such as drug repurposing. A recent example of this successful synergy is the similarity searching of a database of approved drugs that led to the identification of olusalazine, an anti-inflammatory drug approved for the treatment of ulcerative colitis, as a novel DNA hypomethylating agent [74]. Comprehensive reviews of virtual screening that cover methods, successful applications, pitfalls, and workarounds are published elsewhere [75–78]. Advances in the progress in the virtual screening of NPs have also been presented [79–85]. Any similarity searching involves several essential components, which are briefly outlined below.

A) One or more query or reference molecules that are compared against a molecular data set. The reference molecule is typically a chemical structure that can be represented in 2D or 3D. In general, in similarity searching, a notable advantage of 2D over 3D approaches is computational speed since most 2D methods (with the exception of those using chemical graphs) do not require costly structure alignments. In contrast, many but not all 3D methods require such alignments [86]. Moreover, 3D approaches have to deal with the conformational flexibility of the molecules, which, in many instances, give rise to multiple low-energy conformers. Diverse solutions have been proposed to alleviate this problem [87]. Currently, most 3D similarity searching studies use a single low-energy conformer (usually the global minimum or other representative 3D conformation). This, in any case, raises the question if such a conformation is biologically significant [88].

The performance of 2D and 3D similarity approaches has been compared directly in a number of applications, including virtual screening [89–92]. Since 3D similarity searching should incorporate, at least in principle, more accurate features than 2D methods, it would be expected that the results obtained from 3D methods should be more reliable than those obtained by 2D methods. However, in many instances, 2D approaches have outperformed 3D approaches, although it has been noted that this superiority is somewhat case-dependent [92].

Depending on both the data set and the biological activity, it is feasible that one or more reference compounds are associated with activity cliffs. In other words, they may be an “activity cliff generator” (defined as a molecular structure that has a high probability of forming an activity cliff with molecules tested in the same biological assay) [93]. Since, as discussed above, activity cliffs are exceptions to the similarity principle and lead to misleading results in similarity searching, it has been proposed that activity cliff generators be identified and removed from the data sets before selecting the reference compounds. In addition, the removal of activity cliff generators has been proposed as a general approach 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 [94].

B) Another element in similarity searching is the compound database. Compound databases have been reviewed elsewhere, including collections of natural products in the public domain [95]. A current trend in screening libraries for drug discovery is to balance chemical novelty with confined chemical space [96]. In this context, natural product databases (and natural product derivatives) are excellent sources for virtual screening as they expand the currently known medicinal chemistry space [97]. The “expansion” is associated in part with molecular complexity. This feature makes natural product databases attractive to identify compounds with a high selectivity towards molecular targets (including a target family) and can be ideal resources to identify “master key” compounds that selectively bind to a series of targets in order to yield a desired clinical effect [31]. Examples of specific and appealing regions in chemical space covered by databases of natural products include peptides and macromolecules [96].

C) A third and critical component in similarity searching is chemical representation, which is at the core of virtually any chemoin-
forms for which the chemical space is too large for complete search. The chemical space is typically associated with the possible presence of rotatable bonds connecting them, and in some cases includes stereochemical features, but they do not explicitly depict the 3D structures of molecules. Descriptors, as their name implies, are associated with the 3D structures of molecules. Descriptors are nominally classified as 1D, 2D, or 3D. 1D descriptors are commonly related to whole molecule properties such as molecular weight, logP, solubility, number of hydrogen bond donors, number of rotatable bonds, etc. 2D descriptors are associated with the topological structure of molecules as typically depicted in chemists’ drawings. This type of representation shows the atoms, the bonds connecting them, and in some cases includes stereochemical features, but they do not explicitly depict the 3D structures of molecules. 3D descriptors, as their name implies, are associated with the 3D structures of molecules. Todeschini and Consonni have assembled a comprehensive list of the descriptors used in chemical informatics applications.

Despite the fact that many descriptors are available, it is highly unlikely that a single representation and set of descriptors will capture all of the many different aspects of molecular and chemical information. Therefore, in order to reduce the impact of the dependence of chemical representation in similarity searching, it has been proposed to use several methods and then combine the solutions. This is called “data fusion”, and the group of Willet is a pioneer in this field. A recent exhaustive study conducted by Holliday et al. provides strong evidence that suggests that fusion-based approaches to similarity searching yield improved results over single-search-based similarity methods. Following a similar approach, the use of several molecular representations and then the combination of such representations has been implemented in different areas of chemoinformatics, including activity landscape modeling. In the latter, the term “consensus activity cliffs” have been proposed.

A fourth component of similarity-based virtual screening is a similarity measure which, in turn, depends on three elements: (1) the representation used to encode the desired molecular and chemical information, (2) whether and how much information is weighted, and (3) the similarity function, also called the similarity coefficient, that maps the set of ordered pairs of representations onto the unit interval of the real line.

Using the components of similarity searching outlined above, different groups have been using similarity searching alone or in combination with other computational approaches to uncover bioactive compounds from natural products. Examples of recent investigations are summarized in Table 3 and described in the following paragraphs.

Gusach et al. used a combination of computational methods to identify five PPAR-γ partial agonists from a compound collection with more than 89 000 natural products and natural product derivatives from ZINC. The authors of that work implemented a sequential or cascade virtual screening approach using a set of ADMET filters, structure-based pharmacophore screening, molecular docking, electrostatic, and fingerprint-based similarity analysis. A total of ten compounds with different chemical scaffolds were selected for experimental validation using in vitro assays. All five compounds were confirmed as PPAR-γ partial agonists.

Also in a combined approach, Guasch and Medina-Franco and Yoo implemented a sequential computational screening of five compound libraries to identify candidate compounds for testing as potential inhibitors of DNMT1. The reference molecule was a known DNMT inhibitor recently identified from HTS whose chemical structure was made publicly available in PubChem. The compound databases screened included two collections of natural products, a DNMT-focused library, a general screening collection, and a set of approved drugs. Similarity searching was performed using the widely used MACCS keys (166 bits) as implemented in MOE. The molecular similarity was computed using two measures, Tanimoto and Tanimoto-substructure. Of note is that Tanimoto-substructure takes into account the putative different sizes of the query molecule and the compounds in the databases screened. Compounds selected from similarity searching were subject to docking with a crystallographic structure of human DNMT1 using a validated docking protocol. At least 108 molecules with promising DNMT1 inhibitory activity were identified. The chemical structures of the computational hits were disclosed to encourage the research community working on epigenetics to experimentally test the enzymatic and demethylating activity in vivo.

### Table 3: Representative and recent studies using similarity searching to uncover bioactive compounds in natural products and related compounds.

<table>
<thead>
<tr>
<th>Study</th>
<th>Similarity searching method used</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential virtual screening of ZINC natural compounds identifies five compounds as PPAR-γ partial agonists.</td>
<td>Electrostatic and fingerprint-based similarity analysis combined with ADMET and structure-based filtering.</td>
<td>[103]</td>
</tr>
<tr>
<td>Sequential docking-based virtual screening followed by similarity searching to select promising inhibitors of DNMT1 in two natural products collections.</td>
<td>Fingerprint-based similarity searching using MACCS keys and the similarity coefficients Tanimoto and Tanimoto-substructure.</td>
<td>[104]</td>
</tr>
<tr>
<td>Searching of GRAS compounds to uncover compounds similar to approved antidepressants. Identification of nonanoic acid and 2-decenonic acid (similar to valproic acid) as inhibitors of HDAC1.</td>
<td>Fingerprint-based similarity searching with MACCS keys/Tanimoto.</td>
<td>[19]</td>
</tr>
<tr>
<td>Structural comparison of the FEMA GRAS list with analgesics and with compounds used as safety agents.</td>
<td>Comparison based on physicochemical properties and seven structural representations obtained from three different software programs.</td>
<td>[105]</td>
</tr>
<tr>
<td>Similarity searching to identify compounds in a compiled database of phytochemicals with activity against a protein involved in the colon cancer pathway or a colon cancer drug target.</td>
<td>Text mining in PubMed abstracts led to the collection of more than 20 000 diverse chemical structures present in the human diet. Authors systematically explore their numerous targets using chemoinformatics methods.</td>
<td>[106]</td>
</tr>
</tbody>
</table>

Do QT et al. How to Valorize... Planta Med 2015; 81: 436–449
Feng et al. [109] used chemoinformatics analysis based on Lipinski’s rule-of-five, ChemGPS-NP [110] principal component analysis, and chemical clustering to compare a set of antitrypanosomal marine natural products with approved drugs to prioritize products with a similar profile as the reference drugs. GRAS compounds are largely comprised of natural products. A recent and notable application of similarity searching of GRAS compounds for bioactive compounds is represented by the work of Martinez-Mayorga et al. In that work, the authors searched for similar structures to approved antidepressant drugs in the food flavoring components in the FEMA GRAS list [19]. The virtual screening was conducted using fingerprint-based similarity searching with the MACCS keys and the Tanimoto coefficient. Hit compounds in the FEMA GRAS list were chosen as the most similar compounds (ranked with the highest similarity values) to any of the 32 approved antidepressant drugs. Selected compounds represented the “nearest neighbors” of the approved antidepressants. Valproic acid was the most similar antidepressant to GRAS molecules. Based on the knowledge that the inhibition of HDAC1 could be related to the efficacy of valproic acid in the treatment of bipolar disorder, Martinez-Mayorga et al. screened the GRAS compounds most similar to valproic acid for HDAC1 inhibition. The GRAS compounds nonanoic acid and 2-decenoic acid inhibited HDAC1 at a micromolar level with a potency comparable to that of valproic acid. Of note is that the GRAS chemicals were not expected to have strong enzymatic inhibitory effects at the concentrations typically employed in flavor formulations designed for use in foods and beverages. However, as shown in that work, GRAS chemicals were able to bind to a relevant therapeutic target. That study also served as a proof-of-principle of the feasibility of exploring the FEMA GRAS flavoring list using computational methods as a potential source of biologically active molecules. In addition, the study demonstrated that similarity searching followed by experimental evaluation could be used for rapid identification of GRAS chemicals with potential bioactivity [19].

In two subsequent and separate studies, Martinez-Mayorga et al. employed structural similarity to compare the FEMA GRAS list with analgesics and with compounds used as satiety agents [105]. The list of analgesics used as query molecules contained ten structurally diverse molecules currently used in clinics. A total of eight satiety agents were identified in the literature, which were used as reference compounds for similarity searching. The similarity agents included those currently used in clinics, as well as those still in clinical trials. In both studies, reference compounds were compared with the FEMA GRAS list using a total of seven structural representations obtained from three different software programs, MOE, ChemAxon, and PowerMV. Compounds identified by different programs and representations were chosen as consensus compounds for additional studies. 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 were selected as being similar to the reference analgesics. In the second study, a total of nine FEMA GRAS molecules were identified as being similar to those used as reference satiety agents. 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 considered that in the large majority of cases biological effects result from complex and multiple interactions in the body [105].

As previously discussed in this review, phytochemicals derived from edible plants are notable sources of bioactive molecules. In a recent study, Jensen et al. [106] performed a high-throughput analysis of phytochemicals in order to reveal associations between diet and health benefits using text mining and chemoinformatic methods. The first step of that work was the retrieval of associations between the terms plants and phytochemicals from 21 million abstracts in PubMed/MEDLINE during the period 1998–2012. This information was merged with the Chinese National Product Database and the Ayurveda data set, which was also curated by the authors. The final data set included nearly 37000 phytochemicals. A major outcome of that study is the structured and standardized database of phytochemicals associated with medicinal plants. The authors pointed out that their approach facilitates the identification of novel bioactive compounds from natural sources, and the repurposing of medicinal plants for diseases other than those for which they are traditionally used, with the added benefit that the information collected can help elucidate a mechanism of action [106]. As a case study, Jensen et al. conducted 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 ChEMBL. A set of molecules from this study not only showed reported health benefits against colon cancer, but activity was also verified against colon cancer protein targets [106].

**Polypharmacology and Chemogenomics in Natural Products Research**

The increasing awareness that a drug may have its clinical effect through the interaction of multiple targets (called “polypharmacology”) is changing the drug discovery paradigm from a single target to a multi-target approach [31]. This change is enriching chemogenomics data sets that capture ligand-target relationships [111]. As a consequence, a number of computational and experimental approaches are being developed to generate, store, analyze, mine, and visualize target-ligand interactions that define chemogenomic spaces [112–114].

Using the literature reports, the identification of the pharmacological evaluation of compounds (in particular with novel chemical structures) isolated from natural sources is frequent. The pharmacological evaluation usually includes a handful of biological endpoints. In light of the generation of chemogenomics data sets, natural products are being evaluated systematically across a large number of biological endpoints, and the screening data is being released to the public. A representative example of a chemogenomics data set that contains natural products is the large microarray data released by Clemons et al. [115]. In that work, the authors evaluated the binding specificity of 2477 natural products (which were part of a larger collection with 15 000 compounds) across 100 sequence-unrelated proteins. The authors released the results of the screening to the public domain (the interested reader has access to the screening data along with the chemical structures in the paper of Clemons et al. [115]). The microarray data set has been analyzed with chemoinformatic approaches with the goal of elucidating the SAR; in particular to uncover structural characteristics related to the selectivity or promiscuity of the molecules using fingerprint or substructure rep-
resentations [116–118]. For instance, Yongye and Medina-Franco developed the SPID metric to quantify and uncover specific structural changes that have a significant impact on the number of proteins to which a compound binds [116]. In a subsequent publication, Dimova et al. reported an analysis of the same data set using matched molecular pairs [119] to identify single-site substitutions that are associated with large magnitude differences in apparent compound promiscuity. The results of Dimova et al. further confirmed the results of Yongye and Medina-Franco previously published in that promiscuity can be induced by small chemical substitutions.

Docking

The concept of one disease/one target was a milestone in modern molecular medicine because it enabled the simplification of complex in vivo symptoms and related them to simple in vitro models. Though this paradigm is shifting to multitargets [31] as our knowledge progresses, this reductionist approach did prove successful in many diseases. To better understand the molecular mechanism of action of molecules on their biological targets, several methods were developed to determine the 3D structure of these proteins, e.g., X-ray, NMR, and electron microscopy. During the past decades, the number of solved protein crystallographic structures grew exponentially, and now tops at 94,000 structures (statistics from the PDB homepage [120]). At the same time, computer power has also increased dramatically. Molecular modeling software could then be developed to exploit these types of data. The first docking software was DOCK [121]. Docking refers to methods that predict the orientation of a molecule bound to another. The stabilities or the affinities of the resulting complex are estimated by a mathematical or scoring function [122]. Many different strategies and algorithms for docking exist, e.g., AutoDock [123], FlexX [124], Glide [125], Gold [126], and Surflex [127], to predict the positioning of molecules into the protein-binding site. Authors have also studied scoring methods to improve the hit rate. Many are related to their cognate docking software [refer to the review by Li et al. [128] describing 20 scoring functions]. Because different docking software and scoring functions have different strengths and weaknesses, several authors tested the combination of docking and scoring methods to find the optimal procedure [129], while others proposed to use consensus scoring to accommodate the weaknesses [130–133]. The scoring of the predicted poses from docking will be performed by several scoring functions, not only by one. Predictions well scored by multiple scoring functions will be better ranked. Interested readers can refer to [134] for a review of several consensus-scoring methods. As the scoring is dependent on the pose predictions, authors have also worked on improving this step by using consensus docking. It consists in retaining the poses predicted by a majority of docking software [135–137]. In an ideal case, the software can be selected because the natural substrate is structurally related to either a ligand or a receptor, or even both, which belong to the software’s calibration set resulting in a higher confidence that the computed solutions are trustworthy [138]. Sometimes docking problems arise when the target receptor is a constitutively inactive mutant or exists in unliganded states (inactive vs. active); it could also be under allosteric control (conformational modulation) [139].

Natural products remain a large source of active products and also an inspirational source for medicinal chemists; most of the resources, particularly from the microorganisms, are underexploited [12]. Structure-based techniques constitute a possible way to find new applications to these natural products. The majority of drugs in oncology and biocide products are derived from natural products. It is not surprising that many docking studies with natural products fall in these therapeutic domains. Thiagarajan et al. [140] targeted FAK by docking a library of 109 natural products. Four selected candidates showed activity of C6 glioma and N18 neuroblastoma cell lines by promoting apoptosis. Medina-Franco and Yoo [104] screened by combining structure-based pharmacophore filtering and docking on DNMT with a library composed of natural products, approved drugs, a DNMT-focused library, and general screening compounds. One hundred and eight potential hits were disclosed to the scientific community for experimental validation. Hussain et al. [141] adopted a docking strategy coupled to a 3D-QSAR to predict the activity of the analogues of aplyronine A that bind to actin. Their models may be helpful in designing more efficient and tolerable antitumor agents.

Docking may be used to assess the binding mode of natural products and subsequently guide the design of more potent candidates. For instance, the comparative docking of forskolin (activator) and labd-13(E)-ene-8a,15-diol diterpene (inhibitor) into the active site of adenylyl cyclase revealed important features in the binding mode of the activator and the inhibitor, allowing for the design of potent cytotropic and cytostatic agents against cancer cells [142].

Due to antibiotic multiresistant bacteria, finding a new class of antibiotics with a new mode of action has become of paramount importance. This can be evidenced by numerous public fundings at national or international levels. A list of the multimillion Euros projects financed by the EU can be found in [143]. It is noteworthy that the NABATIVI project [144] succeeded in finding a peptidomimetic product with a new mode of action targeting a membrane receptor [145]. This product is in clinical trial phase II. Docking techniques were extensively applied not only to discover new antibiotics but also antiviral, antifungal, or antiprotozoan products. One strategy that is successful so far is to target essential bacterial genes, whose inhibition will kill the microorganisms. An example of a structure-based screening on an essential gene such as the filamenting temperature-sensitive mutant Z (FtsZ) provides promising leads [146]. Other authors also performed docking studies to demonstrate the bacterial potentiality of xanthone derivatives [147]. An interesting use of docking was exemplified by Harris et al. [148]. They performed docking on the bacterial essential enzyme peptidyl-tRNA hydrolase to identify possible active compounds and guide their activity-directed isolation to discover antibacterial molecules from an ethanol bark extract of Syzygium japonini. For examples of drug discovery from natural compounds using docking to target viruses, fungi, and protozoan parasites, the reader is invited to consult the following respective works: [149] reviewed several in silico approaches to tackle urgent threats caused by new viruses or their variants (HIV, SARS, etc.) and how helpful computational techniques were to disclose the antiviral properties of natural products; docking studies helped to hypothesize the mechanism of action of antifungal pyanocoumarin derivatives in [150, 151]; the authors performed docking studies with geldanamycin targeting the HSP90 homolog proteins of pathogenic protozoans Plasmodium falciparum, Leishmania donovani, Trypanosoma brucei, and Entamoeba histolytica. This work allowed for designing
Inverse docking plays a key role in the concept of “reverse pharmacognosy” introduced by Do and Bernard [161] and extended by Blondeau et al. [17]. Pharmacognosy starts with natural sources (e.g., extracts of plants and microorganisms) and thanks to activity-guided fractionation, identifies the molecule(s) responsible for a biological activity. Conversely, reverse pharmacognosy begins with a natural molecule and, thanks to inverse docking, identifies putative targets of interest. The predictions are then validated with related in vitro assays. Thanks to a database linking molecules and the organisms producing them, we can identify new applications for plants, for example, with the mode of action at the molecular level (ligand-protein interactions). With this approach, Do et al. could identify protein interacting partners for epsilon-viniferin from Vitis vinifera, which inhibits phosphodiesterase 3 and 4 [162], and for meranzin from Limnocritus littoralis, which blocks COX 1 and 2, and activates PPAR-γ [163]. Thus, extracts from these two plants at an adequate concentration of the active molecules may be used in indications involving the described proteins.

Other scenarios of inverse docking were described for pharmacological profiling of natural products [164], either to understand the mode of action as well as repurpose molecules, e.g., tanshinone IIa [165], or to evaluate the toxicity profile [166].

Only a few software programs have been developed for inverse docking, but the field is gaining more and more attention, as we can notice through the development of tools based on an existing docking engine or on a specific software: Invdock [152], iRAISE [167], Mdock [168], Selnergy [161] based on the Surfex programme, Tarfisdock [169] based on the DOCK programme, and TarSearch-X [170]. Inverse docking is not yet mature technology but should mutually fertilize other approaches, e.g., chemogenomics and bioinformatics.

It should be mentioned that inverse docking is one out of the several techniques available to conduct target fishing. Other common approaches such as data mining and similarity searching (see above) are extensively used to explore putative targets of bioactive compounds. In similarity searching, targets are represented by their ligands and query molecules are compared with the known ligands. Based on the concept of SAR, similar molecular structures will certainly have similar biological activities.
Thus, by finding similar ligands to query structures, one can relate the query compounds to the ligands’ targets. Databases such as DrugBank, PubChem, and ChEMBL [171] are key to have as many as possible interaction pairs of ligands and targets. Machine learning techniques (e.g., Support Vector Machine, Neural Networks [172, 173]) are also popular to identify the relationship of molecules and possible targets. These systems are usually trained with a training set of known pairs of ligands/proteins based on descriptors, then validated with an external validation set (known pairs of ligands/proteins not used to build the models). We will evoke in the next paragraph the different types of descriptors. Structure-based and other computational approaches for target fishing are reviewed elsewhere [80,174].

Pharmacophoric and Other Descriptors in Virtual Screening

A crucial point for the success of virtual screening is the design of the filter layer that constitutes the similarity patterns to retrain the potential candidates and discard the reminder. Many virtual screening programs have special graphical or scripting methods to write such filter definitions. Sometimes they consist of physicochemical properties, for example, “filter out all compounds with pKₐ greater than ... and/or with aryl rings ... and/or with a nonpolar surface larger than”. On occasion, they also describe a substructure of the scaffold common to all or almost all expected hit compounds. The underlying assumption is twofold: (1) the existence of a pharmacophore, “an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response” [78,175] and (2) similar chemicals have similar biological activities [176]. Ligand-based virtual screening for structures with similar pharmacophoric patterns has become a successful method to identify potential drug candidates. Some of the latter find their role as lead compounds for lead expansion, lead hopping, and scaffold hopping in the desired therapeutic area. When 3D structures or homology models of the target protein are available, protein-based screening can be carried out. The pharmacophore defines spatial requirements like interatomic distances, angles, or the location of particular properties, and ionic sites as well as other descriptors that depend on the spatial coordinates. Such 3D information renders the screening query (filter) more precise but also more error-prone. Hence, not unexpectedly, researchers noticed that virtual screening based on 2D fingerprints (filter concepts based on atoms and bonds and their connectivity but without spatial coordinates) could be more successful than 3D pharmacophore. The authors recommended the combination of 2D and 3D descriptors [176]. Sometimes the sheer number of conformations under which compounds are collected is so overwhelming that the docking, screening, or simply identifying relationships between compounds based on their shape similarities risks overthrowing the computer resources at hand. Thus, screening for whole molecules, their side chain substituents, or their central scaffolds by conformationally-independent topomer similarities becomes a useful strategy. The partitioning of solutes in liquids along with surface defining descriptors, like nonpolar surface area, solvent accessibility, etc., is commonly applied in ADMET prediction models or in studies of membrane crossing, transport into cell compartments, or diffusion kinetics. The partial charges of the compounds can be calculated and projected as isocontour lines in the space surrounding the molecule of interest. TARIS is an approach based on such molecular EPS. The classes and types of descriptors are far too many, thereby lying beyond the scope of this review [177] (refer to topics in cheminformatics, e.g., GETAWAY, 3D MoRSE, MS-WHIM, FEPOS circular fingerprints, MACCS keys, or graph-based multi-point pharmacophore as well as the so-called ROCs shape descriptors [178,179]). ROC profiles (receiver operating characteristics) show a sort of hit enrichment in the final solution list against other compounds, e.g., decoys for testing (benchmarking), docking, and screening simulations. They have highly similar structures but are biologically inactive. A well-performing method should discard them from the hit list [180]. Although the number of descriptors used in a study may end up in the thousands, the right choice remains a challenging task of its own kind. Apparently, descriptors fail in reflecting in exactly which item the molecules resemble each other. PCA, a sort of statistical factorial analysis, simplifies the level of data complexity to a minimum set of orthogonal (independent) diagram axes (factors or components). PCA eventually sheds some light and explains some of the failures and downside when molecules are screened [181].

When needing to save time and running costs, HTS can be elegantly simulated in silico by screening virtual libraries (vHTS) [182]. To this end, vHTS descriptors have been developed, which do not need a lengthy superposition of data set molecules (for comparison) like PESD or UFSR. In addition, other techniques have been developed, for instance, 2D and higher dimensional QSAR, SVM, rule-based methods, or ANN [28]. ANN outperforms rule-based pharmacophore screens in those cases when decision taking in a straightforward manner is not behaving well, or tenets are ill-designed or believed to be just “better than no rule at all”. Such rule-based pharmacophore screens generally make use of binary variables (simple “Yes/No” criteria) or integer values (the Lipinski’s rule-of-five, more than five hydrogen bonds, etc.). The architecture of ANNs is based on a multilayer of criteria with individual weight put on them by training, i.e., probabilities of their contribution to yield the right answer which is a priori known during ANN training. What the ANN learned (as a black box to the user) through the training set of known cases is then applied to the test set. Although very complex phenomena can be handled, wrong answers emerge, mostly in cases when the molecule has unforeseeable characteristics.

Conclusions and Perspectives

Thanks to the large amount of information accumulated on natural product research, in silico techniques related to chemoinformatics, database mining, and molecular modeling facilitate the use of this information to further valorize natural products as a source and/or inspiration of drugs. In silico approaches enable the characterization of their physicochemical profile, analysis of chemical diversity, coverage of chemical space, and uncovering of trends in their SAR. The outcome of such analyses is valuable to guide medicinal chemistry efforts to optimize their properties or inspire the synthesis of novel scaffolds. Molecular modeling approaches, either ligand based or structure based, coupled with experimental methods, constitute techniques of choice to identify putative biological properties for natural products in a systematic manner and thus find ways to valorize them. To this end, numerous authors have applied computational structure similarity techniques to the GRAS list compounds [19] to repurpose them
as potential functional foods or use reverse pharmacognosy to find new uses for the molecules and their sources [18]. Although in this review we only examine health-related aspects of natural product utility, many applications can be found in numerous domains, such as material science and energy engineering among others. With new insights in microorganism biomes, the possibilities offered by Nature become even more tremendous [12], and preserving biodiversity has never been so crucial even at the restrictive economical point of view. It is anticipated that in silico approaches will continue to be part of the research to study and further potentiate the use of biodiversity.

Acknowledgements

We thank Dr. Karina Martinez-Mayorga for her insightful discussions, stimulating ideas, and fruitful conversations.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

4 Camp DJ, Newman S, Pham NB, Quinn RJ. Nature Bank and the Queensland Landmark Library: unique international resources at the Ekitis Institute for Drug Discovery, Comb Chem High Throughput Screen 2014; 17: 201–209
7 Burton RA, Fincher GB. Plant cell wall engineering: applications in biofuel production and improved human health. Curr Opin Biotechnol 2014; 26: 79–84
35 Blake JF. Examination of the computed molecular properties of compounds selected for clinical development. Biotechniques 2003; 34: S16–S20
39 Ritchie TJ, Macdonald SJ. The impact of aromatic ring count on compound developability—are too many aromatic rings a liability in drug design? Drug Discov Today 2009; 14: 1011–1020

Do QT et al. How to Valorize... Planta Med 2015; 81: 436–449

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.
Rishton GM.

Loub WD, Farnsworth NR, Soejarto DD, Quinn ML.


Bergström CA, Holm R, Jørgensen SA, Andersson SBA, Artursson P, Beato S, Ehrman TM, Barlow DJ, Hylands PJ.


Kellenberger E, Hofmann A, Quinn RJ. Similar interactions of natural products with biosynthetic enzymes and therapeutic targets could explain why nature produces such a large proportion of existing drugs. Nat Prod Rep 2011; 28: 1483–1492


Ortholand JY, Ganesan A.

Messer R, Fuhrer CA, Häner R.

McAllister M, Muenster U, Thinnes J, Taylor R, Mullertz A.


Geldenhuys WJ, Bishayee A, Darvesh AS, Carroll RT.

Ma DL, Chan DSH, Leung CH.

Zhang Q, Muegge I.

Kalászi A, Szisz D, Imre G, Polgár T.


Medina-Franco JL. Advances in computational approaches for drug discovery based on natural products. Rev Latinoam Quim 2013; 41: 95–110


tion site of focal adhesion kinase (Y397) by molecular docking. Biochim Biophys Acta 2013; 1830: 4091–4101


146 Hussain A, Melville JL, Hirst JD.


This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.