Chapter 2

Prediction of Medicinal Properties Using Mathematical Models and Computation, and Selection of Plant Materials

Sanjoy S. Ningthoujam*, Anupam D. Talukdar*, Satyajit D. Sarker†, Lutfun Nahar†, Manabendra D. Choudhury*

**Assam University Silchar, Cachar, India, †Liverpool John Moores University, Liverpool, United Kingdom*

Chapter Outline

2.1. INTRODUCTION

Use of plant-based materials dates back to early periods of human existence with several ancient records that provide formulations and evidence of phytotherapy. The knowledge derived from ancient and traditional systems of medicine has now transcended to the modern pharmaceutical industry. The focus of any phytochemical research is often to discover new drugs or drug leads from medicinal plants. One of the important issues in medicinal plant research is the appropriate selection of target plant species that may provide lead to new drugs.

Computational Phytochemistry. https://doi.org/10.1016/B978-0-12-812364-5.00002-X

Throughout the history of drug discovery from plants, serendipity has played a significant role (Kinghorn, 1994; Kubinyi, 1999). For example, the discovery of dicoumarol from fatal cattle poisoning from *Melilotus officinalis* was simply a serendipitous discovery (Kubinyi, 1999), which then led to the development of the well-known anticoagulant warfarin (Fig. 2.1). Conducting research without any working hypotheses may produce such unexpected discoveries, but the chances of success are much slimmer than any targeted approach. Research on medicinal plants, thus, requires a thorough knowledge of their various properties that may reflect the hitherto unknown medicinal properties from the plants. Challenges lie in devising appropriate methods to uncover the existing or potential medicinal properties as well as selecting the right plants that may fulfil these criteria.

A plant is said to be 'medicinal' if it possesses certain medicinal or curative properties against any ailment or group of ailments. Efficacies of the herbal medicine or phytotherapy in their treatment of several diseases may sometimes be linked to placebo effects, but often involve active natural products mostly of low molecular weight that possess 'drug-like' properties. Earliest known investigation of bioactive plants dates back to 3000 BCE with Egyptians scrolls detailing these plants along with their medicinal properties. Modern study of bioactive compounds isolated from living organisms for therapeutic purposes began around 200 years ago with the isolation of morphine by F.W. Serturner (Schmitz, 1985). After this, there was no turning back, but to accelerate the process of phytochemical discoveries with tremendous advances in phytochemical methods and medicinal chemistry and allied disciplines.

Plant-based medicines have contributed to (and have been continuously doing so) the advancement of modern medical treatments and provision of new drug candidates. However, sometimes the progress in medicinal plant research has somehow been negatively impacted by the introduction of various modern technology-based developments in synthetic medicinal chemistry, e.g., combinatorial chemistry, and by the sheer completion in library-based high-throughput-screening (HTS) process in modern drug discovery scene. However, the inherent chemical diversity and structural novelty that natural products offer are the best, and for this very reason, natural products or drug discovery from plant remains as one of the main sources of new drugs. One of the bottlenecks

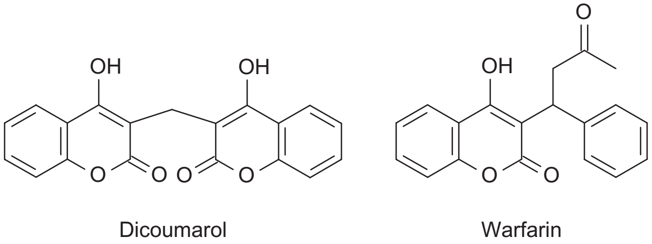


FIG. 2.1 Dicoumarol from *Melilotus officinalis*, and the well-known anticoagulant warfarin.

in phytochemical drug discovery is probably the arduous protocols, and in addition, the overall cost in conventional plant drug discovery methodologies sometimes can be prohibitive for any drug discovery initiative when it comes to cost-effectiveness. To mitigate some of these issues, various stages of plant-based drug discovery programmes require much smarter approach and incorporation of new computational approaches coupled with mathematical models.

Thousands of structurally diverse bioactive compounds have evolved during plant development and evolutionary processes, sometimes to offer plant the necessary protection against herbivores and pathogens, while some others to serve as signal compounds to facilitate reproduction, as antioxidants and UV protectants. Isolation and analysis of potential bioactive phytochemicals may include generation of a hypothesis of the target receptor for a particular disorder and the subsequent screening of the *in vitro* and/or *in vivo* biological activities of the candidate drug. The major challenge in phytochemistry is to describe and understand the diversity of these molecules, their modes of action, and determination of their natural combinations found in plants (Sarker and Nahar, 2012; Wink, 2015). However, right at the beginning of any plant-based drug discovery programme, probably the most important task is to appropriately select the medicinal plants from a vast array of plants available on the earth that may possess expected or desired bioactivity.

Success of any drug discovery programme often depends on accurate data on pharmacokinetics and metabolism. Initiation of absorption, distribution, metabolism, excretion, and toxicity screening has contributed to the success rate of compounds during clinical trials. Pharmacokinetic parameters provide information for future experiments involving animal model and clinical studies for selection of the dose levels and frequency of administration. Apart from these, various techniques and approaches have been attempted to predict potential medicinal activity of plants. One of such attempts is application of phylogenetic methods and chemotaxonomic understanding to determine the pattern of evolution of various groups of specialized metabolites and deriving a correlation between phylogeny and biosynthetic pathways (Rønsted et al., 2012). There are also attempts to correlate the taste of medicinal plants with their ethnopharmacological activities (Gilca and Barbulescu, 2015). These novel or improvised methods are increasingly using various mathematical modelling and computational approaches such as regression analysis, data mining, or analysing structure-activity relationships (see Chapters 1 and 7).

The fundamental aim of this chapter is to present an overview of methods and processes involved in plant selection by utilizing various mathematical modelling and computational techniques.

2.2. MATHEMATICAL MODELS

A mathematical model can be defined as a description of a system using mathematical concepts and language to facilitate proper explanation of a system or to study the effects of different components and to make predictions on patterns of behaviour (Abramowitz and Stegun, 1968). The process of constructing a mathematical model is often called mathematical modelling (Press et al., 1987).

Mathematical modelling is known by various names, such as, predictive modelling, simulation, or decision analysis. A traditional mathematical model comprises four major elements:

1. governing equations;
2. defining equations;
3. constitutive equations; and
4. constraints.

Mathematical models depend on advanced computational tools and can simulate medical outcomes under some given parameters. Some common methodologies are the Markov Chain and Monte Carlo simulations. Mathematical modelling can be applied for predicting outcomes. It is particularly helpful when limitations like a rare event prohibit repeating actual studies or expanding research on clinical trials. Innovative use of this technique includes estimation of missing data points. While common strategies for replacement of missing values include a point of central tendency (e.g., mean or median), these methods usually have cut off criterion for the minimum allowable proportion. There are technical limitations in preserving the variance.

The Markov Chain was first used in the 1940s to model nuclear reactions (McKean, 1966). It is a series of conditional probabilities in a fixed dependent order. This technique was generalized from its limited applications to different disciplines, where one could not derive a single probability function. A Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. Simply, a process satisfies the Markov property only if one can predict the future of the process based solely on its present state just as well as one could know the process's full history. A Markov chain is a type of Markov process that has either discrete state space or discrete index set, often representing time, but the precise definition of a Markov chain may vary.

Monte Carlo simulation is a series of random draws, simulating an event within the known parameters of the probability distribution of the event. It is a computerized mathematical technique or algorithm that allows people to account for risk in quantitative analysis and decision-making. Monte Carlo simulation offers the decision-maker with a range of possible outcomes and the probabilities they will occur for any choice of action. This simulation technique came as a useful application in the time of Markov Chain processes. In principle, Monte Carlo methods can be used to solve any problem having a probabilistic interpretation. By the law of large numbers, integrals described by the expected value of some random variables can be approximated by taking the empirical mean of independent samples of the variables. When the probability distribution of the variable is parameterized, mathematicians often use a Markov Chain

Monte Carlo (MCMC) sampler (Del Moral et al., 2006; Kroese et al., 2014). MCMC estimated value preserves the actual variance. Monte Carlo simulation has several advantages over deterministic or 'single-point estimate' analysis. Some of those advantages are:

1. *Probabilistic results*: Results not only display what could happen, but also how probable each outcome is.
2. *Graphical results*: Monte Carlo simulation-generated data can be easily presented in graphs of different outcomes and their chances of occurrence. This is particularly important for informing findings to other stakeholders.
3. *Sensitivity analysis:* Deterministic analysis sometimes makes it difficult to see which variables influence the outcome the most. However, in Monte Carlo simulation, it is easy to observe which inputs have the biggest effect on bottom-line results.
4. *Scenario analysis*: It is extremely difficult to model different combinations of values for different inputs to see the effects of truly different scenarios in deterministic models, but Monte Carlo simulation clearly demonstrates correlations between inputs and several values when certain outcomes are achieved.
5. Correlation of inputs: In Monte Carlo simulation, it is possible to model interdependent relationships between input variables.

An enhancement to Monte Carlo simulation is the use of Latin Hypercube sampling (LHS), which samples more accurately from the entire range of distribution functions. LHS, first introduced by McKay in 1979, is a statistical method for generating a near-random sample of parameter values from a multidimensional distribution (McKay et al., 1979; Tang, 1993). The sampling method is often used to construct computerized experiments or for Monte-Carlo integration.

Many mathematical modelling approaches, including simulated data, have been applied in determining the medicinal properties of plants. Some examples of the applications of mathematical modelling in predicting medicinal properties and plant selection are discussed in Section 2.4.

2.3. COMPUTATIONAL MODELS IN DRUG DISCOVERY

A computational model is a mathematical model in computational science that requires extensive computational resources to study the behaviour of a complex system by computer simulation. Thus, computational modelling refers to the use of computers to simulate and study the behaviour of complex systems using mathematics, physics, and computer science. A computational model may contain numerous variables that characterize the system under investigation.

Computer-aided drug discovery (CADD) methods contribute significantly to the development of therapeutically important small molecules, either from synthetic or natural sources (Song et al., 2009). CADD methods significantly decrease the number of compounds necessary to screen, while retaining the same level of lead compound discovery. Many compounds predicted to be inactive can easily be skipped, and those predicted to be active can be prioritized, thus reducing the cost and workload of a full HTS screen without compromising lead discovery. CADD methods increase the hit rate of novel drug compounds as it uses a much more targeted search than traditional HTS and combinatorial chemistry. It not only aims to explain the molecular basis of therapeutic activity, but also does help predict possible derivatives for improved activity. Mainly the methods can be classified as structure-based or ligand-based methods (Sliwoski et al., 2014).

Structure-based methods rely on the knowledge of the target protein structure to estimate interaction energies for all compounds tested. On the other hand, ligand-based CADD utilizes the knowledge of known active and inactive molecules through chemical similarity searches or construction of quantitative structure-activity relationships (QSAR models). Important tools, e.g., target/ligand databases, homology modelling, and ligand-fingerprint methods, are necessary for successful implementation of various computer-aided drug discovery/design methods in any modern drug discovery programme. Computational methods for toxicity prediction and optimization for favourable physiologic properties are also parts of modern drug discovery and design protocols. Various approaches of computer-aided drug design can be represented by the following figure (Fig. 2.2) (Aparoy et al., 2012).

Many mathematical modelling approaches, including simulated data, have been used to determine the medicinal properties of plants. Computational methods are powerful knowledge-based approach that helps to select plant material or natural products with a high likelihood for biological activity. These methods can also offer rationalization of biological activity of natural products. *In silico* simulations can be used to propose protein ligand-binding characteristics for

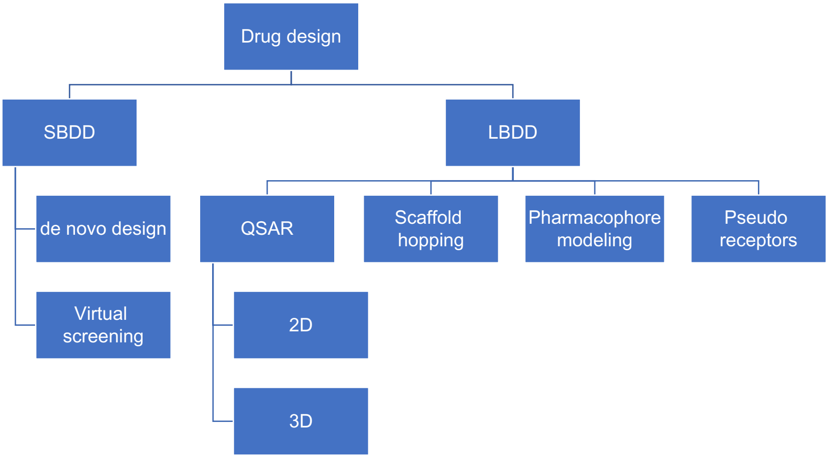


FIG. 2.2 Types of drug design.

molecular structures, e.g., known constituents of a plant material. Compounds that perform well in *in silico* predictions can be used as promising starting materials for experimental work. Some examples of the applications of computational modelling in predicting medicinal properties and plant selection are discussed in Section 2.4.

2.3.1 Structure-Based CADD

In principle, structure-based CADDs are similar to HTS in that both target and ligand structure information is essential (Douguet et al., 2005). Structure-based approaches include ligand docking, pharmacophore, and ligand design methods. Structure-based CADDs rely on the knowledge of the target protein structure to calculate interaction energies for all compounds tested (Sliwoski et al., 2014).

In structure-based drug discovery approaches, therapeutics are designed based on the knowledge of the target structure (Leelananda and Lindert, 2016). This approach depends on the ability to determine and analyze the three-dimensional structures of biological molecules. It is based on the hypothesis that a molecule's ability to interact with a specific protein and exert a desired biological effect depends on its ability to favourably interact with a particular binding site on that protein (Sliwoski et al., 2014). Molecules sharing favourable interactions will possess similar biological effects. Therefore, novel compounds can be determined through analysis of a protein's binding site. Prerequisite for this approach is structural information that can be accessed for target databases. One of the important requirements is the ability to rapidly determine potential binders to the target of biological interest. Computational models are applied for rapid screening of a large compound library and determination of potential binders through modelling, simulation, and visualization techniques.

The ideal starting point for docking is the determination of a target structure that is experimentally confirmed through X-ray crystallography or NMR techniques. Evaluation of appropriate binding pocket is usually performed through the analysis of known target–ligand co-crystal structures. Alternative method is to use *in silico* methods for identifying novel binding sites. When the experimental structures are not available or absent, computational models are utilized for predicting the 3D structure of the target proteins. Target structure may be predicted based on a template with a similar sequence by the process called comparative modelling. It is based on the belief that protein structure is better conserved than sequence that is proteins with similar sequences have similar structures. In essence, comparative modelling involves the following steps:

1. Identification of related proteins to serve as template structures
2. Sequence alignment of the target and template proteins
3. Copying coordinates for confidently aligned regions
4. Constructing missing atom coordinates of target structures
5. Model refinement and evaluation. The process can be automated through computer programmes, e.g., PSIPRED, MODELER, etc.

One of the most significant approaches is the homology modelling, where the template and target proteins share the evolutionary origin. Homology modelling is a popular computation method for predicting the 3D coordinates of structures. Homology modelling, also known as comparative modelling of protein, actually refers to constructing an atomic-resolution model of the target protein from its amino acid sequence and an experimental 3D structure of a related homologous protein, which is commonly referred to as the template. The basis of this approach is the fact that evolution-related proteins often share similar structures. The protein structure generally remains more conserved than the sequence during evolution. As such, understanding structures having amino acid sequences similar to the target sequence of interest may assist in predicting the target structure, function, and possible binding and functional sites (Leelananda and Lindert, 2016). Application of homology modelling has emerged as the main alternative to get a 3D representation of the target in the absence of crystal structures (Aparoy et al., 2012). Combination of homology modelling and docking studies has contributed to identification of oxidosqualene cyclases associated with primary and secondary metabolism of *Centella asiatica* (Kumar et al., 2013) and understanding the structure and function of chalcone synthase protein from *Coleus forskohlii*.

Computational tools have become essential in binding site detection and characterization, which are fundamental to identification of activity of any drug or bioactive molecule. Binding sites can be detected from co-crystal structures of the target or a closely related protein. In the absence of a co-crystal structure, mutational studies can be used to identify ligand-binding sites. Computational methods are used, when there is absence of binding sites or there is need for identification of new binding sites. Computational methods can be divided into three general groups:

1. Geometric algorithms to find shape concave invaginations in the target
2. Methods based on energetic consideration
3. Methods considering dynamics of protein structures

Optimal interaction of a ligand with a target can be identified through steric and electronic features derived from a pharmacophore model. Such models are usually defined by hydrogen bond acceptors, hydrogen bond donors, basic groups, acidic groups, partial charge, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. Pharmacophore model can be used for querying database for bioactive compounds as well as for guiding design of new compounds. Analysis of the target binding site or study of target-ligand complex structure is used for performing structure-based pharmacophore methods. Screening for natural product inhibitors of acetylcholinesterase and cyclooxygenase using protein-based pharmacophores led to the identification of scopoletin as potent AChE (acetylcholinesterase) inhibitor and sanggenons as a potential COX inhibitor (Fig. 2.3) (Barlow et al., 2012). Molecular docking studies established that sieboldigenin could bind to the active site of soybean lipoxygenase and reduce carrageenan-induced paw oedema. This sterol is found

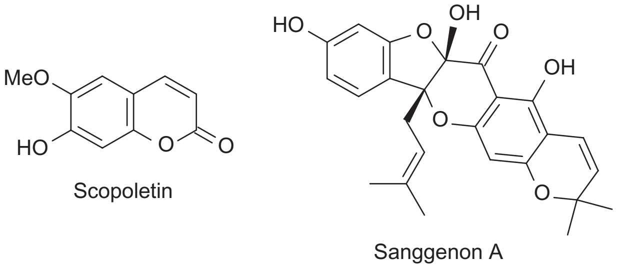


FIG. 2.3 Scopoletin, and sanggenon A.

in several species of Smilax, which is traditionally used in arthritic and skin ailments (Barlow et al., 2012).

In recent years, screening of bioactive compounds based on target recognition has become quite popular among researchers. However, these methods can hardly be effective in direct screening of bioactive compounds from plant extracts, which are often complex mixtures of many compounds. General procedures follow the strategy of 'Isolation—Structure Identification—Activity confirmation'. One novel protocol allowed determining structural information of bioactive compounds without isolating the ligand(s) molecules experimentally by using NMR spectroscopy technique (Tang et al., 2012).

2.3.2 Ligand-Based CADD

Ligand-based methods use only ligand information for predicting activity depending on its similarity/dissimilarity to previously known active ligands. Ligand-based methods include ligand-based pharmacophores, molecular descriptors, and quantitative structure-activity relationships. Ligand-based CADDs exploit the knowledge of known active and inactive molecules through chemical similarity searches or construction of predictive, quantitative structure-activity relation (QSAR) models (Acharya et al., 2011; Sliwoski et al., 2014).

Ligand libraries are usually constructed by enriching ligands having desirable physiochemical properties suitable for the target of interest. Though there are various docking algorithms available, docking of millions of compounds requires considerable resources. As such, time can be saved by elimination of non-drug like unstable or unfavourable compounds. One of the important parameters selected for study is drug likeness, which is commonly evaluated using Lipinski's rule of five (Pfizer's rule of five) (Lipinski et al., 2001). The rule generally states that an orally active drug should have no more than one violation of the following criteria based on multiples of five:

1. maximum of five hydrogen bond donors;
2. maximum of 10 hydrogen bond acceptors (all oxygen and nitrogen atoms);
3. molecular mass of less than 500 Da;
4. an octanol-water partition coefficient of not greater than five.

If two or more of these conditions are violated, adsorption will be compromised. To improve the predictions of drug likeness, the rules have seen many extensions, such as polar surface area no greater than 140 Å^2, molecular weight ranging from 180 to 500, molar refractivity from 40 to 130, partition coefficient in −0.4 to +5.6 range, etc. Before initial virtual HTS, molecules are filtered based on predicted ADMET properties. These predictions depend on statistical and machine-learning approaches, molecular descriptors, and experimental data to model biological processes such as oral bioavailability, intestinal absorption, permeability, half-life time, and distribution in human blood plasma (Sliwoski et al., 2014). In any drug discovery process, lipophilicity and molecular weight are often increased to improve the affinity and selectivity of the drug candidate. As a result, it is often difficult to maintain strict drug likeness as per RO5 during hit and lead optimization. It has been proposed that members of libraries should be biased toward lower molecular weight and lipophilicity. The rule of five has been extended to the rule of three for defining lead-like compounds as per following criteria (Congreve et al., 2003):

1. octanol-water partition coefficient log P not greater than 3;
2. molecular mass less than 300 Da;
3. not more than three hydrogen bond donors;
4. not more than three hydrogen bond acceptors; and
5. not more than three rotable bonds.

Lipinski's fifth rule states that the original four rules do not apply to natural products nor to any molecule that is recognized by an active transport system for considering 'druggable chemical entities' (Newman and Cragg, 2012).

Compound libraries (Wessjohann, 2000; Geysen et al., 2003) are usually enriched for a particular target or group of targets (see Chapter 5). Physiochemical filters determined from observed ligand-target complexes are used for enriching such libraries by searching for ligands that are similar to known active ligands. As molecules are flexible in solvent environment, their representation of conformational flexibility remains important criteria for determining their potentials. These conformations of protein and ligands are usually precomputed using computational simulation or knowledge-based methods (Foloppe and Chen, 2009).

Ligand-based computer-aided drug design involves the analysis of ligands that can interact with a target molecule. The methods require collection of reference structures collected from compounds interacting with the target of interest. Objective of this activity is to represent these compounds with their physicochemical properties that determine desired interactions. There are two main approaches of ligand-based drug designing methods—(a) selection of compounds based on chemical similarity to known actives using some similarity measure and (b) construction of a QSAR model.

For these analyses, molecular properties are converted to numerical vectors for descriptors. Conversion is required to ensure that descriptions of molecules have a constant length independent of size. Representation of information encoded in the molecular structure with one or more numbers is called molecular descriptors. These characteristics are used to establish quantitative relationship between structures and properties, biological activities, and other experimental properties. To date, more than 2000 molecular descriptors that encode the molecular features have been reported. Molecular descriptors can be classified according to their dimensionality, i.e. the representation of molecules from which descriptor values are computed.

1. One-dimensional (1D) descriptors capture bulk properties, i.e. molecular weight, molar refractivity, log$P$ (logarithm of the octanol/water partition coefficient), etc.
2. Two-dimensional (2D) descriptors describe properties that can be computed from two-dimensional representation of molecules, such as number of atoms, number of bonds, connectivity indices, etc.
3. Three-dimensional (3D) descriptors depend on conformations of molecules, i.e. solvent accessible surface areas, principal moment of inertia, van der Waals volume, etc.

Some of the descriptors derived from 3D structures may require analysis of many molecular conformations if biologically active conformations are usually not known from previous experiments. Common 3D descriptors may include pharmacophore type representation of molecules, where features known or thought to be responsible for biological activity are mapped to positions in a molecule. Molecular descriptors may be divided according to their 'nature' into:

1. constitutional (fragment additive and reflect mostly the general properties of the compound);
2. topological (which are calculated using the mathematical graph theory applied to the scheme of atoms connections of the structure);
3. geometrical;
4. electronic; and
5. quantum-chemical (the last three are derived from the results of empirical schemes or molecular orbital calculations).

Among various approaches of ligand-based CADD, application of quantitative structure-activity relationship (QSAR) has contributed significantly in the development of predictive models. QSAR methods are based on the assumption that the quantitative understanding of the role of molecular structure governed the biological or other attributes. The method tries to enumerate how a fragment or sub-structure could result in a certain activity. In many cases, SAR (structure-activity relationships), involving enumeration of a fragment or substructure in their biological activity, and QSAR, which quantified the descriptors, are collectively referred to as (Q)SAR (Puzyn et al., 2010). Successful creation of QSAR models demands fulfilment of the following conditions:

1. consensus data on the structures and biological activity of studied compounds;
2. extracting descriptors for the presentation of structures;
3. machine-learning methods, either multiple linear regressions, neural networks, random forest, similarity, support vector machine, etc.

QSAR models developed because of homogenous data are known as local models and traditionally used for the optimization of hit or lead compounds. On the other hand, QSAR models are developed based on heterogeneous data and are considered as global models with a wide applicability domain. Global models may be used for virtual screening, prediction of biological activity, and target fishing. QSAR has large potentials across industry, academia, and regulatory agencies. Some of the potential uses include identification of new leads with pharmacological, biocidal, or pesticidal activity, prediction of toxicity, rational design of desirable products and selection of compounds with optimal pharmacokinetic properties, etc.

If the researchers tend to determine the potential targets of new chemical entity, the following tools can be used for studying biological activity—(a) pair similarity with known compounds, e.g., Tanimoto coefficient, (b) docking, e.g., INVDOCK, (c) pharmacophore-based virtual screening, and (d) classification prediction based on Bayesian statistics and substructure descriptors or fingerprints.

Successful prediction of the properties of all chemical entities including phytochemicals depends on the data on which they are based, the technique to develop the model, and the overall quality of the information including the item to be modelled. Generally, two types of information are required for a model (the effect to be modelled and descriptors on the chemicals) and a technique(s) to formulate the relationship(s). The data to be modelled in QSAR may be denoted by the X-matrix and the descriptors as the Y-matrix (Table 2.1). By using this matrix, various types of relationship may be established by statistical machine-learning techniques. A QSAR is based on a continuous endpoint where activity (X) is a function of one or more descriptors (Y).

The development of SAR is associated with identification of a firm basis of relationship. If a compound is identified to elicit a particular effect, and the structural determinant is recognized, then the structural fragment can be determined. It may be flagged as a 'structural alert' that can be coded into software. Greater the number of compounds with the same structural determinant demonstrating the same effect, greater will be the confidence that the flag is associated with that particular effect. Development of SAR model is usually more appropriate for qualitative (such as yes/no, active/inactive, presence of toxicity/absence of toxicity, etc.) endpoint.

Successful implementation of QSAR depends on selection of appropriate statistical and machine-learning algorithms supplemented with powerful

TABLE 2.1 Typical Data Matrix for QSAR

Chemical Identifier	Activity (X)	Property/Descriptor (Y)				
		Fragment 1	Fragment 2	Fragment 3	…	Fragment n
Molecule i	X_i	Y_{1i}	Y_{2i}	Y_{3i}	…	Y_{ni}
Molecule ii	X_{ii}	Y_{1ii}	Y_{2ii}	Y_{3ii}	…	Y_{nii}
Molecule iii	X_{iii}	Y_{1iii}	Y_{2iii}	Y_{3iii}	…	Y_{niii}
….	…	…	…	…	…	…
Molecule *n*	Xn	Y_{1n}	Y_{2n}	Y_3n	…	Y_{nn}

computational tools. In the last few decades, multiple linear regression (MLR) is one of the popular methods to derive linear mapping. However, MLR methods have several limitations of multicollinear, overfitting issues, and non-linear relationship, thereby making the researchers to look to other alternative methods. As such, various methods such as neural networks, genetic algorithms, support vector machine, and random forests are applied in the QSAR analysis.

Over the last few decades, several QSAR models have been attempted to explain or describe the potentials of traditionally used medicinal plants. However, application of QSAR methods in herbal formulae, particularly used in Traditional Chinese Medicine and Ayurvedic Systems, is somewhat limited as structure and composition of all compounds in these formulae are not completely known (Wang et al., 2006). Thus, QSAR method cannot be directly applicable for prediction of bioactivity of polyherbal medicine. Despite that, variation of biological activity of herbal medicine is also associated with the variation of their chemical composition. Considering this relationship, another relationship called quantitative composition activity relationship (QCAR) has been proposed to establish relationship between chemical composition and biological activity (Cheng et al., 2006). This method applies the same mathematical model used in QSAR studies to derive quantitative relationship of the composition bioactivity of the herbal components. One of the advantages of this method is deriving optimal combination of herbal medicine (Wang et al., 2006).

Molecular fingerprint-based technique is one approach more qualitative in nature as compared to other LB-CADD approaches (Sliwoski et al., 2014). Molecular fingerprints are representation of molecular structure and properties encoded as binary bit strings whose settings produce a bit ‘pattern’ characteristic of a given molecule (Hert et al., 2004). Fingerprints may provide different sets of molecular descriptors, structural fragments, and possible connectivity pathways through a molecule or different types of pharmacophores. There are several methodologies for representing chemical binary information. For instance, path-based approach, key-based fingerprint, dictionary approach, and SMARTS pattern matching. Molecular fingerprint-based techniques are used to represent molecules for rapid structural comparison of phytochemicals. These approaches depend more on chemical structure and are less computationally expensive than pharmacophore mapping or QSAR models. Fingerprint-based methods provide equal treatment to all parts of the molecule and avoid focus only on parts of a molecule considered to have important role in bioactivity. Screening of phytochemicals using a molecular fingerprint based on the HIV protease inhibitor, saquinavir, led to the discovery of a potential anti-HIV agent leucovorin. Molecular dynamic studies revealed the favourable binding of this compound to the protease active site (Barlow et al., 2012). Combination of molecular fingerprint-based method with docking studies led to the discovery of aurantiamide acetate from *Artemisia annua* (Fig. 2.4) as an inhibitor of severe acute respiratory syndrome coronavirus main proteinase (Wang et al., 2007).

In ligand-based-CADD, machine-learning algorithms are used to be trained to identify patterns in data and for predictions on test data sets. One of the common algorithms is support vector machine (SVM) that is being usually used for classification of sets of biological data (Leelananda and Lindert, 2016). Other significant candidates are Random Forest (Svetnik et al., 2003) and Artificial Neural Network (Wang, 2003).

2.3.3 Network Pharmacology

Network pharmacology is the new paradigm in the drug discovery and development (Hopkins, 2008) and offers enhanced understanding of drug action. It applies network analysis to determine the set of proteins most critical in any disease, and then chemical biology to identify molecules capable of targeting that set of proteins. By addressing the true complexity of disease and by seeking to harness the ability of drugs to influence many different proteins, network pharmacology differs from conventional drug discovery approaches, which have usually been based on highly specific targeting of a single protein. Network pharmacology has the potential to provide new treatments for complex diseases, where conventional approaches have failed to deliver satisfactory therapies.

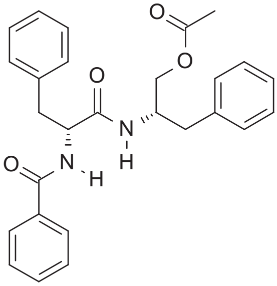


FIG. 2.4 Aurantiamide acetate from *Artemisia annua.*

With the advancement of bioinformatics, systems biology, and polypharmacology, network-based drug discovery has provided promises toward cost-effective drug discovery and development in traditional herbal medicine. Network analysis is the study of molecular interactions and is related with the mathematical field of graph theory in which the assembly of pairwise connections (edges) between discrete objects (nodes) coalesces to form a network or graph (Arrell and Terzic, 2010). Biological networks can be derived from:

1. *de novo* through direct experimental interactions;
2. applying known interactions to an -omic dataset; or
3. reverse engineering to generate a subset of networks *ab initio* that predict the dynamics under study.

In the past, the concept of designing selective ligands to avoid unwanted side effects was a major issue in drug discovery. With the emergence of more complex drug action, often it was discovered that there may be many drugs for each drug target as well as single drug that can fit to multiple drug targets (Hopkins, 2007). Network pharmacology tries to understand this complex relationship along with validating target combinations and optimizing multiple structure-activity relationships. Network pharmacology is a system biology-based methodology that tries to exploit the pharmacological mechanism of drug action in the biological networks. In the application of network analysis in herbal medicine, a network is a mathematical and computable representation of various connections between herbal formulae and diseases in a complex biological systems (Li and Zhang, 2014). The scope of this new approach includes study of:

1. theories, algorithms, models, and software of network pharmacology;
2. network construction and interactions prediction;
3. theories and methods on dynamics, optimization, and control of pharmacological networks;
4. network analysis of pharmacological networks, including flow balance analysis, topological analysis, network stability;
5. various pharmacological networks and interactions;
6. factors that affect drug metabolism;
7. network approach for searching targets and discovering medicines; and
8. big data analytics of network pharmacology (Zhang, 2016).

One of the important contributions of network pharmacology is changing the perspective from 'one target, one drug' strategy to a novel version of the 'network target multi components' strategy, which is perfectly applicable to Traditional Chinese Medicine and to Ayurvedic medicine system. In one of the pioneering works in 1999, the Chinese researcher Li proposed that there was a possible relationship between Traditional Chinese Medicine syndrome and molecular networks and established a network-based TCM research strategy in 2007 (Li and Zhang, 2014). Subsequently he also proposed a new concept of network target approach in the research of herbal medicine.

In the biological network approach, a node represents either (a) a gene, gene product, or any biological entity in the biomolecular network, gene regulatory network, genetic interaction, metabolic network, and signaling network, (b) an herb, herb ingredient, or drug, or (c) a clinical phenotype of a disease in the network. An edge represents an association, interaction, or any other well-defined relationship. The degree of a node is represented by the number of edges connected to it, while the betweenness of a node is the number of shortest paths that can traverse through a given node. Network parameters such as betweenness, degree, shortest path, and modules are used to measure the targeted key proteins or protein interactions.

Potential of network pharmacology lies in its multidisciplinary approach that integrates a large amount of information to make new discoveries by combining both computational and experimental approaches. Main computational approach includes graph theory, statistical methods, data mining, modelling, and information visualization methods. Network pharmacology can be used to identify active herbal ingredients and synergistic combinations as well as contribution to rational design and optimization of drug discovery process from herbal formulae. Application of network mapping to a wide array of drugs to protein targets both before and after modelling with chemical drug-ligand interactions helps in prediction of new targets. It also enables identification of primary sites of action and off-target proteins as explanations for well-known side effects, with new and unexpected drug binding revealed across major categories of proteins unrelated by sequence or structure (Arrell and Terzic, 2010).

2.4. SELECTION OF MEDICINAL PLANTS

Documentation and analysis of legacy knowledge about medicinal plant provide certain advantages in identification and designing pharmacological products from plants. Identification and selection of medicinal plants for drug discovery studies is a challenging task. In medicinal plants research, what type of plants to be selected and what would be the right criteria still remain the enigma of the scientists. Conventionally, targeted approach (Mann et al., 2000) is favourable, where certain medicinal plants are prioritized, and even though expensive, depends on generation of working hypothesis and performing experimental studies on the bases of the hypothesis. In this approach, careful selection and choosing the right criteria for the targeted botanical species usually determine the outcome. Considering the diversity of the higher plant, selection of candidate species for the bioprospecting programme is not an easy task. Out of total number of higher plants (estimated 400,000 angiosperms and 1000 gymnosperms), only about 6% have been screened for biological activity and about 15% for phytochemical properties (Rates, 2001). Various research institutes and pharmaceutical industries have taken up different approaches that can be categorized into four broad groups—ethnobotany-directed, random selection, chemotaxonomic, and integrated approach, respectively. Approaches of ethnobotany-directed, chemotaxonomic, and integrated approach can be part of targeted approach (Yea et al., 2016).

2.4.1 Ethnobotany-Directed Drug Discovery

As the traditional medicine has been used for many centuries in different cultures, medicinal plants have attracted a lot of attention as a source of medicinal products. This approach assumes that the traditional use of plants can provide strong clues to the biological activities of the plants (Cox and Balick, 1994). In the drug discovery procedure, plants possessing potential medicinal properties are recognized through ethnobotanical field studies or literature. Such plants are further investigated for bioactive properties in the laboratories for pure compound isolation. Preparation procedure of traditional recipe may provide an indication of best extraction protocol. Formulation method may provide primary information about pharmacological activity, optimal doses, and suitable mode of application of the future drug (Rates, 2001). Ethno-directed approach was initiated by scientists like Luis Lewin, Carl Hartwich, Alexander Tschirch, and Richard Evans Schultes by applying molecular interpretation of the pharmacologically active plants in 19th and early part of 20th century (Gertsch, 2009). During the early phase of ethno-directed approach, anthropologists worked in tandem with chemists and pharmacologists resulting in isolation of various drug molecules such as caffeine and quinine (Fig. 2.5). However, in spite of the richness of the ethnopharmacological surveys worldwide and diversity of traditional knowledge, many of the collected data could not be translated successfully into bioprospecting programmes (Albuquerque et al., 2014).

One classical example is Shaman Pharmaceuticals, a bioprospecting company established in 1987 that failed to deliver any blockbuster drugs in spite of wealth of ethnopharmacological knowledge and subsequently went bankrupt in 2001. It is a fact that only a few significant contributions have been made by ethno-directed approach in the last few decades (Gertsch, 2009). Research problems like inadequate design for data collection, misinterpretation of the role of medicinal play in the traditional medicine system, unfavourable influence of

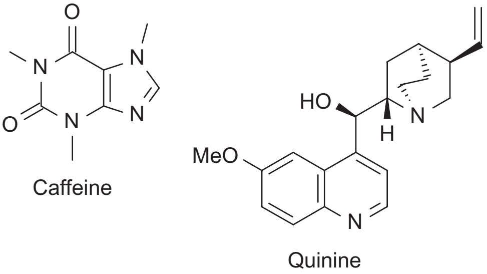


FIG. 2.5 Caffeine, curare, and quinine.

sampling, and application of irrelevant informant consensus indices hampered ethnobotanical-directed drug discovery programmes, individually or in combination (Albuquerque et al., 2014).

In the conventional ethnobotanical research, application of computational tools along with statistical analysis is a new approach, but still rare. Regression analysis of medicinal plant families is a unique method for predicting plant families with large number of medicinal plants. The method was first introduced by Moerman in his classical work (Moerman, 1991). The work concerned prediction of the families having the large number of potential medicinal plants based on linear regression analysis. After that work, there have been many applications of the approach as well as many derivations including Bayesian approach, which challenges the earlier regression methods (Bennett and Husby, 2008; Moerman, 2012). However, application of regression analysis still remains a popular method and helps in identification of plant orders and families favoured by traditional healers among ethnomedicinal plants of South Africa, proving that the use of these plants by traditional healers is not random (Douwes et al., 2008). There was an attempt to apply mathematical and logical method of replacing rare herbs and simplifying traditional Chinese medicine formula and its applicability in the perspective of pathway enrichment analysis (Fang et al., 2013).

2.4.2 Chemotaxonomic and Ecological Approach

The knowledge that a particular group of plants contains a particular group of natural product may help in predicting the presence of similar or related compounds in phylogenetically related species (Rates, 2001); this is chemotaxonomy-guided approach. Chemical plant taxonomy, or simply chemotaxonomy of plants, focuses on the classification of plants based on their chemical composition, i.e. secondary metabolites. The selection highlights the chemical taxonomy of acetylenic compounds, the distribution of fatty acids in plant lipids, distribution of aliphatic polyols, cyclitols, plant glycosides, and alkaloids. Chemotaxonomy is a method of biological classification based on similarities in the structure of certain compounds produced by the organisms in question, e.g., plants. As proteins are more closely controlled by the genes and less subjected to natural selection than are anatomical features, they are more reliable indicators of genetic relationships or phylogeny. This approach may become significant, when a particular compound class is desirable with known biological activity. A good example of this approach is targeting *Datura stramonium* for tropane alkaloids, with the knowledge that *Atropa belladonna* contains the alkaloid hyoscyamine, subsequently leading to the discovery of similar alkaloid hyoscine (Fig. 2.6) (Heinrich et al., 2012).

Similarly, the discovery of the alkaloid febrifugine (Fig. 2.7) from *Hydrangea macrophylla*, a native Japanese plant, was the result of targeting this plant because of its taxonomic status as a member of the Hydrangeaceae

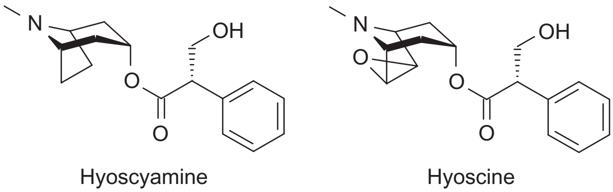


FIG. 2.6 Structures of hyoscyamine and hyoscine.

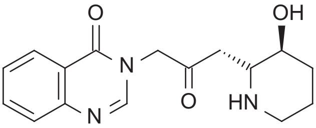


FIG. 2.7 Febrifugine from *Hydrangea macrophylla.*

family (Heinrich et al., 2012). Presence of desirable chemicals may rely on the knowledge of toxicity of particular plant (Rates, 2001), ecology, or plant-pathogen relationships (Ottmann et al., 2012). In plant-pathogen interactions, molecular frameworks of natural products play significant role in host colonization or pathogen immunity. These molecules are the outcome of co-evolutionary process enriched with biological activity. Investigation of the mode of action of these natural product classes may provide significant outcome as novel molecule and discovery of novel concepts on how living systems can be manipulated with small molecules (Ottmann et al., 2012).

In ecological study-based approach, scientists select plants that occupy a particular habitats or display characteristics, indicating they produce molecules possessing desirable properties. This approach is based on the ecological plant defence theory (Coley et al., 2003). For instance, absence of predation might suggest presence of toxic chemicals. Many phytochemicals that are toxic to insects also exhibit biological activity in humans and might be exploited for therapeutic applications. Observation of the plant's environment that reflects the toxicological properties of the plant led to the isolation of many antibacterial drugs. The ecological approach to select plant material relies on the observation of interactions between organisms and their environment that might lead to the production of bioactive natural compounds. The hypothesis underpinning this approach is that secondary metabolites, e.g., in plant species, possess ecological functions that may have also therapeutic potential for humans. For example, metabolites involved in plant defence against microbial pathogens may be useful as antimicrobials in humans, or secondary products defending a plant against herbivores through neurotoxic activity could have beneficial effects in humans due to a putative central nervous system activity (Barbosa et al., 2012).

Major potential limitation of this approach lies in the classification of medicinal plant use (Ernst et al., 2016). Before analysing the medicinal plants in a phylogenetic context, medicinal plant documented or collected need to be classified according to the diseases used to treat in ethnomedicinal system. Some common widely used classifications are International Classification of Diseases (ICD) of the WHO and the classification of Cook developed as Economic Botany Data Collection Standards (Cook, 1995). These classification systems could not fully capture the complexity and idiosyncrasy of local healthcare systems. At the same time, these systems are based on categories reflecting systems of the human body or symptoms. These systems provide little information in disease etiology and potential underlying biological activity of the medicinal plants. In recent years, more extensive studies in cellular and molecular mechanisms underlying diseases have been carried out providing more information on disease etiology. Alternative approaches in phylogenetic systems emerge by using classification based on modulating the disease response (Ernst et al., 2016). For the taxonomic classification, the current default classification is usually the Angiosperm Phylogeny Group IV (APG IV, 2016). It is because categorization and assemblage of plant species within a particular category or sub-categories needs to be based on phylogenetic relationship and APG being the most common among the practicing taxonomists. Theoretically, of the various chemical compounds used, most reliable are the semantides (DNA, RNA, and proteins) that provide more reliable taxonomic information. However, in the practical application, the approach is far from perfection and many researchers are still trying on other compound types. One such instance is that the application of graph-clustering algorithm on the metabolite content of the plant led to the successful classification of 217 plants in Japan (Liu et al., 2017). The approach provides successful result even in incomplete metabolite data by obtaining consistent relationship between plant clusters and known evolutional relationship of plants. This finding led to the application of predictive power of metabolite content in exploring medicinal properties in plants. As such, apart from establishing correlation between the plant group and chemical properties, development of reliable cluster analysis with visual representation of dendrogram remains the fundamental step. All these processes need selection of appropriate clustering algorithms with application of computational tools.

2.4.3 Random Approach

Random Approach was popular in 60s, but with limited results. It does not require any computational or mathematical input whatsoever. In this approach, plants are collected regardless of any previous knowledge of their phytochemical or biological activity. This approach relies on availability of plants and is purely serendipitious in nature (Heinrich et al., 2012). It requires a lot of investment in terms of money, time, and sheer amount of luck. This approach has made effective contributions to the development of drugs for many diseases (Albuquerque et al., 2014). There are two approaches in the random screening. In the first approach, plants are screened for selected class of compounds like alkaloids, flavonoids, coumarins, or lignans. This approach usually does not provide any idea of the biological efficacies. Second approach screens randomly selected plants for selected bioassays, through focused screening as well as general screening. The Central Drug Research Institute, India, started this approach three decades back. Though the institute has screened about 2000 plants for biological efficacy, the screening could not provide any new drug (Katiyar et al., 2012). In the United States, the National Cancer Institute of National Institute of Health screened about 35,000 plants for anticancer activity spanning two decades from 1960 to 1982, resulting in discovery of chemotypes including those of taxanes and camptothecin (Fig. 2.8) (Cragg and Newman, 2005). Their development into clinically active agents spanned about 30 years.

2.4.4 Integrated Approach

This approach is also called knowledge or information-driven approach and takes into consideration ethnobotanical, random, and chemotaxonomic approach for selecting the medicinal plants (Katiyar et al., 2012; Lin et al., 2015). Computational and mathematical tools are extensively applied in this approach. Related information for a particular species are integrated into a database for prioritizing the screening process. Hypothesis generation and subsequent analysis requires careful assembly, overlay, and comparison of data from divergent sources. Importance of database-driven information sharing in drug discovery can be demonstrated by large-scale production of taxol (Fig. 2.8). In 1962, a team of researchers in National Cancer Institute in US discovered that extracts of Pacific yew (*Taxus brevifolia*) contained cytotoxic activity. In 1977, the team confirmed the bioactive component of the extract as paclitaxel, also known by its trade name taxol. After starting clinical phase I in 1984 against number of cancer types, taxol was approved by the FDA for the treatment of ovarian cancer and breast cancer. However, supply of paclitaxel was a major challenge as this compound is found in the thin bark of *Taxus* in extremely low concentration. The bark from a single tree could provide only a single dose for clinical trial leading to the destruction of whole plant. Large-scale production of taxol

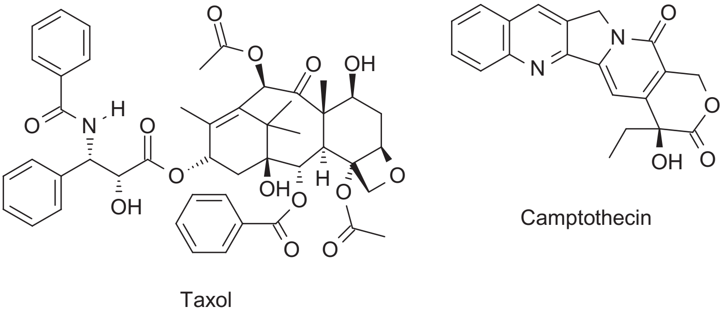


FIG. 2.8 Taxol and camptothecin.

was made by developing several pathways to derive 10-deactyl-baccatin III, a non-cytotoxic precursor of taxol. This precursor was initially isolated in France (Institute de Chimie des Substances Naturelles, Gif-sur-Yvette) from the needles of European yew *Taxus baccata* by the French scientists (Raviña, 2011). One kilogram of fresh needles can provide 1 g of precursor, making it possible for large-scale supply of taxol. In this way, modern drug discovery approaches relied on various information resources.

Curation and compilation of various information from different sources demand access and sharing of different data curation services and databases. Statistically, processing of this varied information through multivariate analysis (e.g., Principal Component Analysis, Discriminant Function Analysis) provided the potential for understanding the pattern of medicinal properties in several target species. Details of curation of medicinal plant and their potentials are discussed in the following sections.

2.5. ROLE OF MEDICINAL PLANTS DATABASES

Phytochemicals, and natural products in general, are recognized to possess characteristics of high chemical diversity, biochemical specificity, and other properties that make them favourable as lead structures for drug discovery programmes. However, assessing the diverse chemical space efficiently and effectively is still impractical in terms of resource and time. It is expected that the application of computational approaches for the identification of bioactive phytochemicals can accelerate the drug discovery by exploring on the chemical space covered by these molecules or on the application of natural product (phytochemical) libraries (see Chapter 5) in ligand-based and target-based virtual screening. Major problems usually encountered in the *in silico* studies of the biological activity of natural products include unavailability of large natural products or phytochemical databases having adequate structural and biological information.

Medicinal plant databases curate the information about plants covering a large spectrum of plant properties including the formulae of traditional medicine (Ningthoujam et al., 2012). Dozens of databases and Internet resources are available on the internet providing various types of information for the last decades. Development of an inclusive database with information about classification, activity, and ready to dock library of medicinal plant compounds is essential for drug designing using resources of medicinal plants (Mumtaz et al., 2017). If one particular database could not provide a complete picture of a medicinal plant, data mining and sharing from different resources may be utilized. Such items need unique identification number for a particular plant or a particular entity. These databases are required to provide phytochemical and pharmacological information on medicinal plants. Number of medicinal plant databases increases year by year with specialized or comprehensive information giving opportunity for the study of plants and the utilization of traditional knowledge. For instance, development of Global Natural Products Social Molecular Networking curated information and enabled sharing of raw, processed, or identified tandem mass spectrometry data (Wang et al., 2016). Another example of curating specialized data are plant protein interaction data, such as IntAct, The Arabidopsis Information Resource and BioGRID, etc. (Lee et al., 2010). Data curated on medicinal plant databases need to be comprehensive as far as possible to serve as important resources for drug discovery studies. Aggregation of these data allows researchers to visualization, data mining, and further analysis to produce new insights. Aggregations would be unachievable if the data are dispersed within largely inaccessible formats (Rodriguez et al., 2009). Challenges encountered during aggregating data arising from different formats can be mitigated by ontological linking as well as introduction of noSQL data model (Ningthoujam et al., 2014). With the rapidly expanding information derived from various analytical and exploratory activities, the role of medicinal plant databases is also progressively increasing to house all these available information. Availability of comprehensive information about a particular plant species or plant groups would accelerate the analysis and prediction of their medicinal properties.

2.6. TOOLS AND TECHNIQUES

Various tools and techniques are used to explore medicinal properties through data curated in medicinal plants databases as well as analytical methods such as QSAR, molecular modelling, and virtual screening (Lagunin et al., 2014). Software and tools that can be used for virtual screening and identification of potential mechanism of action of herbal constituents can be categorized (Barlow et al., 2012) as shown in Table 2.2.

2.7. ROLE OF DATA MINING IN MEDICINAL PLANT SELECTION

Data mining is the process of sorting through large datasets to identify patterns and establish relationships to solve problems through data analysis by using machine-learning and statistical methods (Afendi et al., 2013; Yea et al., 2016). Data mining methods use various kinds of information obtained from sources such as bibliographic literature, experimental data, clinical data, and curated data. Vast amount of data stored in these databases are screened to identify potential natural products. In the data mining approaches, random selection approach does not consider taxonomic affinities, ethnomedicinal contexts, or other intrinsic qualities. However, random screening is associated with extremely low probability of discovery of useful compounds (Yea et al., 2016). Considering the limitations of random selection, some advanced methods have been proposed. For instance, a simple scoring system for searching the local alternatives to *Phytolacca dodecandra* was developed in ways that are more

TABLE 2.2 Uses and Tools and/or Algorithms Important in Computational Methods of Herbal Medicine

Methods	Prerequisites	Use	Tools/Algorithms
Ligand-based screening	Knowledge of compounds with known activity	To identify putatively active compounds	Classification/regression trees (including Random Forest), linear discriminant analysis, artificial neural networks, support vector machines
Pharmacophore (ligand-based or target-based)	• 3D structures of known ligands to chosen targets (Ligand-based) • known 3D structures of target protein(s) ideally known as 3D structure(s) of known complex(es) (Target-based)	To identify putative active compounds	LigandScout, Schrödinger's Phase program, Accelrys's Discovery Studio Catalyst, etc.
Docking	Known 3D structure (s) of target proteins	To 'dock' potential small molecule ligands into protein active sites, optimizing their topographical and chemical complementarity, and scoring their interaction	FlexX, Gold, Dock, Glide, MolDock, AutoDock, LigandFit, etc.
Pattern recognition		Post-screening analyses (involving dimensionality reduction)	Principle components analysis (PCA), multidimensional scaling, self-organizing maps, various forms of cluster analysis, etc.
Proteomics and genomics data visualization and analysis		Application-specific programs for statistical processing and visualization of data output from DNA micro-array experiments, MS proteomics experiments, etc.	

complex. Despite the applications of simple scoring, regression analysis, or a logical formula method in data mining from random selection, successful mining of vast body of information and knowledge pertaining to biology, medicine, and botany is far from complete. Still today, mining of biomedical data to unearth knowledge or generate hypothesis is an active research field. One major innovation is inclusion of semantic information of the Medical Subject Headings (MeSH) thesaurus to cluster documents of MEDLINE database (Yea et al., 2016). In the approach, three categories containing terms related to herbal compounds, efficacy, toxicity, and the metabolic processes, were selected and subjected to similarity measurement method. Application of this novel approach in data mining could predict herbs by 500% more accurately with similar efficacy as compared to random selection.

Association rule mining is one of the powerful tools to derive the relationship between different factors with the properties of Chinese traditional medicines. As the data mining aims at extracting structured information or discovering new knowledge from large data, one of the prerequisites is data availability (Lee, 2015). So, data mining techniques are intricately related with the advancement in technology and curation of medicinal plant databases.

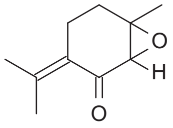
2.8. SAFETY CONSIDERATIONS

Medicinal plants, though considered to be of lower risk, are not completely free from the possibility of toxicity or other adverse effects. Apart from inherent toxicity, adverse effects of the herbal preparations may come from contamination of products with toxic metals, adulteration, misidentification, or substitution of herbal ingredients and improper processing (Jordan et al., 2010). There may be interactions between drugs, foods, and other herbal products if taken concomitantly. Considering these aspects, there is increased concern on the safety assessment of herbs with various protocols and guidance documents have been issued. Documents issued by the International Life Sciences Institute, the Union of Pure and Applied Chemistry, the European Medicines Agency, and the European Food Safety Authority discuss the assessment of safety of herbs for using in foods and medicines. Assessment of safety of herbal products, either as pre-market assessment or post-market surveillance, is subject to challenges. Data deficiencies with regard to quantity and the quality of information are the major factors. For efficient assessment, proper information of adverse reactions, ideal product quality, composition of herbal formulae, and toxicity of the constituents are required. Integration of all these parameters can reduce the uncertainty in decision-making and can be fulfilled if all the available information are in a knowledge base. Development of these knowledge base requires integration of various data sources and mapping different information (e.g., toxicity, bioassay, herbal formulae, and contraindications) arising from diverse domains.

Another dimension in safety consideration is application of predictive toxicology (computer-assisted) to screen and assess the potential toxicity of chemicals. Predictive toxicology contributes to the aim in transforming toxicology testing from primarily observational science to a truly predictive science for the benefit of drug development, chemical risk assessment, and food safety, using mathematical and computational tools. Predictive toxicology deals with the development of new non-animal tests that do not simply duplicate existing animal tests, but offer a new scientific basis for safety testing. It reflects a significant shift away from adverse effects observed in experimental animals, sometimes at high doses, to analysing the effects of chronic exposures to low concentrations on cells and organ systems. This approach offers the potential for reliable, reproducible, faster, and more cost-effective safety assessment in both new product development, e.g., new drug development, and eventually in regulatory testing and is advantageous when the numbers of individual chemicals to be screened far exceed the capacity for assessment.

Prediction of toxicological property uses the computational toxicology methods such as QSAR to assessment environmental chemicals. Various QSAR models that could predict LD_{50} in rats, mutagenicity and carcinogenicity of chemicals. Interests have increased on computational predictions of toxicological properties. *In silico* methods have been used to predict cytotoxic activity of sesquiterpene lactones in members of the Asteraceae family. Fernandes et al. (2008) used artificial neural network to examine these compounds with regard to their cytotoxic potential. One landmark discovery was made by Valerio by using a QSAR model for rodent carcinogenicity. Di Sotto et al. (2017) have reported genotoxicity assessment of peperitenone oxide, a natural flavouring agent also known as rotundifolone (Fig. 2.9), based on an integrated *in vitro* and *in silico* evaluation protocol.

In *in silico* part of the study, the computational prediction of genotoxicity was carried out using the Toxtree and VEGA tools. Computational prediction for piperitenone oxide agreed with the toxicological data and highlighted the presence of the epoxide function and the α,β-unsaturated carbonyl as possible structural alerts for DNA damage. However, it was felt that an improvement of the toxicological libraries for natural occurring compounds was essential to augment the applications of the *in silico* models to the toxicological predictions.



Piperitenone oxide

FIG. 2.9 Piperitenone oxide, a naturally occurring flavouring agent.

2.9. CONCLUSION

Study of pleiotropic pharmacological potential of the natural products derived from medicinal plants may be possible with the availability of medicinal plant databases that stored information on chemical structure and therapeutic uses. Modelling may provide answers to hitherto unknown problems and greatly expand our knowledge base from actual study data. Initially, scientific community suffered from lack of large data sets particularly from curated biological activity. These limitations have been overcome, to some extent, with the availability of many open access initiatives like PubChem, DrugBank, ChemBank, and ChemSpider. Nevertheless, problem still persists, as there are limitations in managing high capacity data in sync with generated big data and the ability to transform these data into meaningful knowledge. Data integration from divergent sources at different platform, coupled with increasingly complex multidisciplinary approaches, increased the need of data analysis and interpretation.

REFERENCES

Abramowitz, M., Stegun, I.A., 1968. Handbook of Mathematical Functions. Dover Publications, New York.

Acharya, C., Coop, A., Polli, J.E., MacKerell, A.D., 2011. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr. Comput. Aid. Drug Design 7, 10–22.

Afendi, F.M., Ono, N., Nakamura, Y., Nakamura, K., Darusman, L.K., Kibinge, N., Morita, A.H., Tanaka, K., Horai, H., Altaf-Ul-Amin, M., Kanaya, S., 2013. Data mining methods for omics and knowledge of crude medicinal plants toward big data biology. Comput. Struct. Biotechnol. J. 4, e201301010.

Albuquerque, U.P., Medeiros, P.M.D., Ramos, M.A., Júnior, W.S.F., Nascimento, A.L.B., Avilez, W.M.T., Melo, J.G.D., 2014. Are ethnopharmacological surveys useful for the discovery and development of drugs from medicinal plants? Rev. Bras 24, 110–115.

Aparoy, P., Kumar Reddy, K., Reddanna, P., 2012. Structure and ligand based drug design strategies in the development of novel 5-LOX inhibitors. Curr. Med. Chem. 19, 3763–3778.

APG IV, 2016. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181, 1–20.

Arrell, D.K., Terzic, A., 2010. Network systems biology for drug discovery. Clin. Pharmacol. Therap. 88, 120–125.

Barbosa, W.L.R., Do Nascimento, M.S., Do Nascimento Pinto, L., Maia, F.L.C., Sousa, A.J.a.D., JúNior, J.O.V.C.R.S., Monteiro, M.C.M., De Oliveira, D.R., 2012. Selecting medicinal plants for development of phytomedicine and use in primary health care. In: Bioactive Compounds in Phytomedicine. InTech, London, UK.

Barlow, D.J., Buriani, A., Ehrman, T., Bosisio, E., Eberini, I., Hylands, P.J., 2012. *In-silico* studies in Chinese herbal medicines' research: evaluation of *in-silico* methodologies and phytochemical data sources, and a review of research to date. J. Ethnopharmacol. 140, 526–534.

Bennett, B.C., Husby, C.E., 2008. Patterns of medicinal plant use: an examination of the Ecuadorian Shuar medicinal flora using contingency table and binomial analyses. J. Ethnopharmacol. 116, 422–430.

Cheng, Y., Wang, Y., Wang, X., 2006. A causal relationship discovery-based approach to identifying active components of herbal medicine. Comput. Biol. Chem. 30, 148–154.

Coley, P.D., Heller, M.V., Aizprua, R., Araúz, B., Flores, N., Correa, M., Gupta, M., Solis, P.N., Ortega-Barría, E., Romero, L.I., 2003. Using ecological criteria to design plant collection strategies for drug discovery. Front. Ecol. Environ. 1, 421–428.

Congreve, M., Carr, R., Murray, C., Jhoti, H., 2003. A 'rule of three' for fragment-based lead discovery? Drug Discov. Today 8, 876–877.

Cook, F.E.M., 1995. Economic Botany Data Collection Standard: Prepared for the International Working Group on Taxonomic Databases for Plant Sciences (TDWG). Kew, Royal Botanic Gardens, Kew.

Cox, P.A., Balick, M.J., 1994. The ethnobotanical approach to drug discovery. Sci. Am. 270, 82–87.

Cragg, G.M., Newman, D.J., 2005. Plants as a source of anti-cancer agents. J. Ethnopharmacol. 100, 72–79.

Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers. J. Royal Stat. Soc. Ser. B (Stat. Method.) 68, 411–436.

Di Sotto, A., Di Giacomo, S., Abete, L., Bozovic, M., Parisi, O.A., Barile, F., Vitalone, A., Izzo, A.A., Ragno, R., Mazzanti, G., 2017. Genotoxicity assessment of piperitenone oxide: an *in vitro* and *in silico* evaluation. Food Chem. Toxicol. 106, 506–513.

Douguet, D., Munier-Lehmann, H., Labesse, G., Pochet, S., 2005. LEA3D: a computer-aided ligand design for structure-based drug design. J. Med. Chem. 48, 2457–2468.

Douwes, E., Crouch, N.R., Edwards, T.J., Mulholland, D.A., 2008. Regression analyses of southern African ethnomedicinal plants: informing the targeted selection of bioprospecting and pharmacological screening subjects. J. Ethnopharmacol. 119, 356–364.

Ernst, M., Saslis-Lagoudakis, C.H., Grace, O.M., Nilsson, N., Simonsen, H.T., Horn, J.W., Ronsted, N., 2016. Evolutionary prediction of medicinal properties in the genus *Euphorbia* L. Sci. Rep. 6, 30531.

Fang, Z., Zhang, M., Yi, Z., Wen, C., Qian, M., Shi, T., 2013. Replacements of rare herbs and simplifications of traditional Chinese medicine formulae based on attribute similarities and pathway enrichment analysis. Evid. Based Complement. Alternat. Med. 2013, 136732. (9 pages).

Fernandes, M.B., Scotti, M.T., Ferreira, M.J., Emerenciano, V.P., 2008. Use of self-organizing maps and molecular descriptors to predict the cytotoxic activity of sesquiterpene lactones. Eur. J. Med. Chem. 43, 2197–2205.

Foloppe, N., Chen, I.-J., 2009. Conformational sampling and energetics of drug-like molecules. Curr. Med. Chem. 16, 3381–3413.

Gertsch, J., 2009. How scientific is the science in ethnopharmacology? Historical perspectives and epistemological problems. J. Ethnopharmacol. 122, 177–183.

Geysen, H.M., Schoenen, F., Wagner, D., Wagner, R., 2003. Combinatorial compound libraries for drug discovery: an ongoing challenge. Nat. Rev. Drug Discov. 2, 222–230.

Gilca, M., Barbulescu, A., 2015. Taste of medicinal plants: a potential tool in predicting ethnopharmacological activities? J. Ethnopharmacol. 174, 464–473.

Heinrich, M., Barnes, J., Gibbons, S., Williamson, E.M., 2012. Fundamentals of Pharmacognosy and Phytotherapy. Elsevier Health Sciences, Edinburg.

Hert, J., Willet, P., Wilton, D.J., 2004. Comparison of fingerprint-based methods for virtual screening uning multiple bioactive reference structures. J. Chem. Inf. Model. 44, 1177–1185.

Hopkins, A.L., 2007. Network pharmacology. Nat. Biotechnol. 25, 1110–1111.

Hopkins, A.L., 2008. Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. 4, 682–690.

Jordan, S.A., Cunningham, D.G., Marles, R.J., 2010. Assessment of herbal medicinal products: challenges, and opportunities to increase the knowledge base for safety assessment. Toxicol. Appl. Pharmacol. 243, 198–216.

Katiyar, C., Gupta, A., Kanjilal, S., Katiyar, S., 2012. Drug discovery from plant sources: an integrated approach. AYU 33, 10.

Kinghorn, A.D., 1994. The discovery of drugs from higher plants. In: Gullo, V.P. (Ed.), Discovery of Novel Natural Products With Therapeutic Potential. Newnes, Boston.

Kroese, D.P., Brereton, T., Taimre, T., Botev, Z.I., 2014. Why the Monte Carlo method is so important today. Wiley Int. Rev. Comput. Stat. 6, 386–392.

Kubinyi, H., 1999. Chance favors the prepared mind-from serendipity to rational drug design. J. Recept. Sig. Transd. 19, 15–39.

Kumar, V., Kumar, C. S., Hari, G., Venugopal, N. K., Vijendra, P. D., B, G. B., 2013. Homology modeling and docking studies on oxidosqualene cyclases associated with primary and secondary metabolism of *Centella asiatica*. SpringerPlus 2**,** 189.

Lagunin, A.A., Goel, R.K., Gawande, D.Y., Pahwa, P., Gloriozova, T.A., Dmitriev, A.V., Ivanov, S.M., Rudik, A.V., Konova, V.I., Pogodin, P.V., Druzhilovsky, D.S., Poroikov, V.V., 2014. Chemo- and bioinformatics resources for *in silico* drug discovery from medicinal plants beyond their traditional use: a critical review. Nat. Prod. Rep. 31, 1585–1611.

Lee, S., 2015. Systems biology—a pivotal research methodology for understanding the mechanisms of traditional medicine. J. Pharm. 18, 11–18.

Lee, K., Thorneycroft, D., Achuthan, P., Hermjakob, H., Ideker, T., 2010. Mapping plant interactomes using literature curated and predicted protein–protein interaction data sets. Plant Cell 22, 997–1005.

Leelananda, S.P., Lindert, S., 2016. Computational methods in drug discovery. Beilstein J. Org. Chem. 12, 2694.

Li, S., Zhang, B., 2014. Traditional Chinese medicine network pharmacology: theory, methodology and application. Chin. J. Nat. Med. 11, 110–120.

Lin, W.-C., Wen, C.-C., Chen, Y.-H., Hsiao, P.-W., Liao, J.-W., Peng, C.-I., 2015. Integrative approach to analyze biodiversity and anti-inflammatory bioactivity of Wadelia medicinal plants. PLoS One 10, e0129067.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. Adv. Drug Deliv. Rev. 46, 3–26.

Liu, K., Abdullah, A.A., Huang, M., Nishioka, T., Altaf-Ul-Amin, M., Kanaya, S., 2017. Novel approach to classify plants based on metabolite-content similarity. Biomed. Res. Int. 2017, 12.

Mann, D.R.A., Da Rocha, A.B., Scheartsmann, G., 2000. Anti-cancer drug disovery and development in Brazil: targeted plant collection as a rational strategy to acquire candidate ant-cancer compounds. Oncologist 5, 185–198.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245. (JSTOR Abstract)|*format*= *requires* |*url*= (*help*) - American Statistical Association.

Mckean, H.P., 1966. A class of Markov processes associated with nonlinear parabolic equations. Proc. Natl. Acad. Sci. U. S. A. 56, 1907–1911.

Moerman, D.E., 1991. The medicinal flora of native North America: an analysis. J. Ethnopharmacol. 31, 1–42.

Moerman, D.E., 2012. Commentary: regression residual vs. Bayesian analysis of medicinal floras. J. Ethnopharmacol. 139, 693–694.

Mumtaz, A., Ashfaq, U.A., Ul Qamar, M.T., Anwar, F., Gulzar, F., Ali, M.A., Saari, N., Pervez, M.T., 2017. MPD3: a useful medicinal plants database for drug designing. Nat. Prod. Res. 31, 1228–1236.

Newman, D.J., Cragg, G.M., 2012. Natural products as sources of new drugs over the 30 years from 1981 to 2010. J. Nat. Prod. 75, 311–335.

Ningthoujam, S.S., Talukdar, A.D., Potsangbam, K.S., Choudhury, M.D., 2012. Challenges in developing medicinal plant databases for sharing ethnopharmacological knowledge. J. Ethnopharmacol. 141, 9–32.

Ningthoujam, S.S., Choudhury, M.D., Potsangbam, K.S., Chetia, P., Nahar, L., Sarker, S.D., Basar, N., Talukdar, A.D., 2014. NoSQL data model for semi-automatic integration of ethnomedicinal plant data from multiple sources. Phytochem. Anal. 25, 495–507.

Ottmann, C., Van Der Hoorn, R.A., Kaiser, M., 2012. The impact of plant-pathogen studies on medicinal drug discovery. Chem. Soc. Rev. 41, 3168–3178.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1987. Numerical Recipes 3rd Edition: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Puzyn, T., Leszczynski, J., Cronin, M.T., 2010. Recent Advances in QSAR Studies: Methods and Applications. Springer Science & Business Media.

Rates, S., 2001. Plants as source of drugs. Toxicon 39, 603–613.

Raviña, E., 2011. The Evolution of Drug Discovery: From Traditional Medicines to Modern Drugs. John Wiley & Sons.

Rodriguez, H., Snyder, M., Uhlén, M., Andrews, P., Beavis, R., Borchers, C., Chalkley, R.J., Cho, S.Y., Cottingham, K., Dunn, M., Dylag, T., Edgar, R., Hare, P., Heck, A.J.R., Hirsch, R.F., Kennedy, K., Kolar, P., Kraus, H.-J., Mallick, P., Nesvizhskii, A., Ping, P., Pontén, F., Yang, L., Yates, J.R., Stein, S.E., Hermjakob, H., Kinsinger, C.R., Apweiler, R., 2009. Recommendations from the 2008 international summit on proteomics data release and sharing policy—the Amsterdam principles. J. Proteome Res. 8, 3689–3692.

Rønsted, N., Symonds, M.R.E., Birkholm, T., Christensen, S.B., Meerow, A.W., Molander, M., Mølgaard, P., Petersen, G., Rasmussen, N., Van Staden, J., Stafford, G.I., Jäger, A.K., 2012. Can phylogeny predict chemical diversity and potential medicinal activity of plants? A case study of Amaryllidaceae. BMC Evol. Biol. 12, 182.

Sarker, S.D., Nahar, L., 2012. Natural Products Isolation, third ed. Humana Press, Springer-Verlag, USA.

Schmitz, R., 1985. Friedrich Wilhelm Sertürner and the discovery of morphine. Pharm. Hist. 27, 61–74.

Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E.W., 2014. Computational methods in drug discovery. Pharmacol. Rev. 66, 334–395.

Song, C.M., Lim, S.J., Tong, J.C., 2009. Recent advances in computer-aided drug design. Brief. Bioinform. 10, 579–591.

Svetnik, W., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Info. Model. 43, 1947–1958.

Tang, B., 1993. Orthogonal array-based latin hypercubes. J. Am. Stat. Assoc. 88, 1392–1397.

Tang, Y., Shang, Q., Xiang, J., Yang, Q., Zhou, Q., Li, L., Zhang, H., Li, Q., Sun, H., Guan, A., Jiang, W., Gai, W., 2012. Integration of screening and identifying ligand(s) from medicinal plant extracts based on target recognition by using NMR spectroscopy. Protoc. Exchange. https://doi.org/10.1038/protex.2012.060.

Wang, S.-C., 2003. Artificial neural network. In: Interdisciplinary Computing in Java Programming—Part of the Springer International Series in Engineerings and Computer Science. vol. 743. Springer-Verlag, New York, NY, pp. 81–100.

Wang, Y., Wang, X., Cheng, Y., 2006. A computational approach to botanical drug design by modeling quantitative composition-activity relationship. Chem. Biol. Drug Des. 68, 166–172.

Wang, S.-Q., Du, Q.-S., Zhao, K., Li, A.-X., Wei, D.-Q., Chou, K.-C., 2007. Virtual screening for finding natural inhibitor against cathepsin-L for SARS therapy. Amino Acids 33, 129–135.

Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu, W.-T., Crusemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderon, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C.-C., Floros, D.J., Gavilan, R.G., Kleigrewe, K., Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., Mclean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A.M., Sedio, B.E., Klitgaard, A., Larson, C.B., Boya, P.C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'neill, E.C., Briand, E., Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., Mcphail, K.L., Nielsen, K.F., Vuong, L., Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R., Gurr, J., Rodriguez, A.M.C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P.-M., Phapale, P., 2016. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat. Biotechnol. 34, 828–837.

Wessjohann, L.A., 2000. Synthesis of natural-product-based compound libraries. Curr. Opin. Chem. Biol. 4, 303–309.

Wink, M., 2015. Modes of action of herbal medicines and plant secondary metabolites. Medicines 2, 251–286.

Yea, S.J., Seong, B., Jang, Y., Kim, C., 2016. A data mining approach to selecting herbs with similar efficacy: targeted selection methods based on medical subject headings (MeSH). J. Ethnopharmacol. 182, 27–34.

Zhang, W., 2016. Network pharmacology: a further description. Network Pharmacol. 1, 1–14.