BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 2. 238–250
Advance Access published on 15 May 2012
doi:10.1093/bib/bbs021

Bioinformatics opportunities for identification and study of medicinal plants

Vivekanand Sharma and Indra Neil Sarkar

Submitted: 3rd February 2012; Received (in revised form): 25th March 2012

Abstract

Plants have been used as a source of medicine since historic times and several commercially important drugs are of plant-based origin. The traditional approach towards discovery of plant-based drugs often times involves significant amount of time and expenditure. These labor-intensive approaches have struggled to keep pace with the rapid development of high-throughput technologies. In the era of high volume, high-throughput data generation across the biosciences, bioinformatics plays a crucial role. This has generally been the case in the context of drug designing and discovery. However, there has been limited attention to date to the potential application of bioinformatics approaches that can leverage plant-based knowledge. Here, we review bioinformatics studies that have contributed to medicinal plants research. In particular, we highlight areas in medicinal plant research where the application of bioinformatics methodologies may result in quicker and potentially cost-effective leads toward finding plant-based remedies.

Keywords: *medicinal plants; bioinformatics; drug discovery*

INTRODUCTION

Plants are a valuable resource for a variety of products that are important for human needs. Plant materials are used for many purposes including timber, food and medicine. With respect to medicine, the use of plant-based materials dates back to ancient civilizations. There are several ancient written records that provide evidence regarding use of plant sources of remedies [1, 2]. The knowledge from ancient systems of plant-based remedies has also been used by the modern pharmaceutical industry. There is thus an immense potential for discovery of new drugs from plants based on the ethno-medicinal data [3, 4]. About one-third of currently available drugs come from natural products that have a plant origin [5]. Even though plant-based remedies have much potential towards advancing modern medical treatments, research continues to lag behind (especially when compared to the interest in developing synthetic drugs for commercial use) [6]. This may be partly because conventional plant drug discovery methodologies can be slow and expensive [7]. Nonetheless, there may be utility to increase research in the area of medicinal plants. The available literature and resources in this area is generally scattered, which hinders the ability to readily leverage available information about medicinal plants. There are several computational approaches for analyzing the diversity of compounds. These approaches have played a significant role in computer-aided drug design [8]. The field of drug design and discovery from medicinal plant requires the application of such approaches for quicker and efficient progress so as to cope up with the continually demanding pharmaceutical needs.

Bioinformatics offers a suite of essential techniques for analyzing and interpreting huge volumes of information generated using molecular biology-based techniques. With the advancement of

Corresponding author. Indra Neil Sarkar, Center for Clinical and Translational Science, University of Vermont, 89 Beaumont Avenue, Given Courtyard N309, Burlington, VT 05405, USA. Tel: +1-802-656-8283; Fax: +1-802-656-4589; E-mail: neil.sarkar@uvm.edu
Vivekanand Sharma is a PhD student in the Department of Microbiology and Molecular Genetics at the University of Vermont. **Indra Neil Sarkar** is the Director of Biomedical Informatics in the Center for Clinical and Translational Science, as well as an Assistant Professor in Microbiology and Molecular Genetics as well as Computer Science at the University of Vermont. His research involves the development of biomedical informatics methods across the entire spectrum of life, from molecules to populations.

highthroughput techniques, such approaches have become essential in analyzing and integrating data to infer knowledge from a whole systems point of view. To increase our understanding of cellular processes associated with plants, an in depth analysis of genomic, proteomic and metabolomic information is required. Bioinformatics approaches offer essential tools for the identification of genes and pathways that may be associated with important bioactive secondary metabolites from medicinal plants [9].

This review focuses on describing potential applications of computational methodologies for the overall advancement of plant-based drug discovery. Different areas are explored where use of such approach can lead to valuable findings in a cost and time efficient manner. Aspects related to the integration of scattered information, analysis of molecular data, drug discovery and design, authentication and toxicology are discussed with focus on computational methods.

HARNESSING LEGACY KNOWLEDGE ABOUT MEDICINAL PLANTS

Usage of plant sources for medicinal purpose has existed for a long period of time. Early knowledge about medicinal uses of plants can be traced back to ancient civilizations, including those from Ancient Egypt, Mesopotamia and the Indus Valley [2, 10, 11]. There exist hundreds of ancient manuscripts about the ethnobotanical uses of herbs. This pre-existing knowledge base can be used in identification and designing pharmacological products from plants. Ethnobotanical knowledge bases may have immense potential to provide leads for modern day pharmaceutical drugs. For example, the identification of prostratin, which is used against HIV for its role in activating latent T-cell pools [12], was done using ethnobotanical work in Samoa [13].

A major challenge that exists for the development of ethnobotanical knowledge bases is that the current information related to plant substances for medicinal purposes is scattered and mostly embedded within literature in unstructured (i.e. natural language) form. Thus, it is important to develop techniques that can extract, store and present data in a useful format for subsequent data mining. Data mining techniques can then be enhanced to provide solutions for developing targeted bioprospecting and screening strategies. For retrospective knowledge, digitization of historical texts is a first step, such as being done by initiatives like the Biodiversity Heritage Library (http://www.biodiversitylibrary.org/) and the China-US Million Book Digital Library Project (http://www.cadal.zju.edu.cn). However, extracting potentially valuable information from these digitized texts does pose a significant challenge. Automated algorithms for character recognition can be used for extracting some information from the historical texts. This has been demonstrated by applying automated information extraction techniques on the 17th century herbal text, *Ambonese Herbal* [14]. The automated techniques were able to identify several medicinal plants that had valid evidence from contemporary sciences as well. For example, *Ceiba pentandra* was identified for its use in treating headaches. Cross-referencing this result with contemporary studies validated the analgesic and anti-inflammatory properties proposed in the historical text. Plant names can then be identified with string matching algorithms [15]. Additionally, extraction routines can also be used to identify treatments (passages of text in a document that provide a description for a particular taxon and its related features) [16]. There are a number of major linguistic challenges that are associated with analyzing old herbal texts. For example, the names used to refer to plants can vary across resources. The process of taxonomic name recognition has become an increasingly important area in named entity recognition [17], and is especially relevant for plant species name identification. Additionally, the terms or phrases used to describe the symptoms, diseases and treatment is diverse and can thus be hard to relate with those currently used. However, the extraction of information from text does not necessarily equate to semantic understanding. Natural language processing approaches can help overcome challenges faced in data parsing, extraction and organizing unstructured text as a structured data [18]. This then sets the framework for indexing the digitized text as usable data using ontology-based indexing systems [19]. Indexing of information in this type of a semantic way will potentially result in the integration of legacy knowledge into a form that can be used to complement contemporary knowledge sources. After legacy knowledge related to plants is identified and the appropriate treatments are extracted and semantically organized, a next step is to integrate with recent knowledge. Use of semantic-based technologies has been proposed for integrating heterogeneous

data sources, as described by Samwald *et al.* [20]. In the context of medicinal plants, their work has specifically shown how a semantic web approach can be used to find evidence for important pharmaceutical compounds from Chinese medicine to treat psychological disorders [20]. Another significant effort in this direction is the development of semantic e-Science infrastructure for Traditional Chinese Medicine (TCM) system. This semantic resource uses domain ontologies to support large-scale database integration from heterogeneous sources [21].

Historically, accumulated legacy data can provide a valuable resource to set the stage for the identification of novel pharmacological compounds. Extracting and organizing the huge magnitude of such data may highlight plants of interest in the context of biomedicine. Advances in data mining techniques offer some promise toward efficiently handling this process. Nonetheless, there will continue to be a need to integrate legacy data with contemporary knowledge to fully appreciate the potential of medicinal plants.

MEDICINAL PLANT DATABASES

The volume of information related to medicinal plants accumulated over the ages and those being generated by contemporary methodologies require a common platform for consolidated and integrated access. Several databases have been developed for cataloguing information related to one or more aspects of medicinal plants, such as ethnobotany, bioactive metabolites, pharmacological uses, genomic or transcript-based information, molecular targets of active ingredients, etc. In addition to data available through resources that are specific to medicinal plants, information related to medicinal plants can also be found embedded within the realm of general taxonomic, chemical and molecular data sources {e.g. at the National Center for Biotechnology Information (NCBI) [22], Kyoto Encyclopedia of Genes and Genomes (KEGG) [23] and KNApSAcK [24]}. Here, we provide some of the examples of databases that provide useful information related to medicinal plants. The International Ethnobotany Database (ebDB) is a noncommercial repository for ethnobotanical data supporting multilingual functionality (http://ebdb.org). This database contains a broad feature set and is designed specifically for ethnobotanical research providing complete location information, strong searching and data export features. NAPALERT is a database designed for identification and analysis of experimental data related to natural products including those from plant sources [25]. This database provides information through analysis of existing literature and its contained data. The United States Department of Agriculture (USDA) maintains a database for medicinal plants where short descriptions are provided for associated bioactive metabolites and their medicinal uses. This database, which is maintained by Dr Park, also provides a compilation of medicinal plants associated with six disorders: (i) inflammation; (ii) heart disease; (iii) obesity; (iv) hypertension; (v) kidney disease; and (vi) diabetes (http://www.pl.barc.usda.gov/usda_info/disease_intro.cfm?id=39). Dr Duke's Phytoch mical and Ethnobotanical Database (http://www.ars-grin.gov/duke/) is another database at the USDA that contains detailed information for medicinal plants, their taxonomy, active ingredients and therapeutic importance. A noteworthy compilation of TCM knowledge into an evidence base for scientific evaluation is the Traditional Chinese Medicine Information Database (TCM-ID) [26, 27]. This database provides comprehensive information related to TCM prescriptions, herbal constituents, molecular structure and functional properties of active ingredients, clinical indications as well as therapeutic and toxicity effects. TCM-ID has been evaluated for its ability to develop testable leads for drug discovery by Chen, *et al.* (2006) [28]. A resource that provides valuable information from pharmaceutical research perspective is the database for Herb Ingredient's Target (HIT) [29]. This database contains information related to therapeutic targets for herbal ingredients from TCM and has a user-friendly search interface (including keyword and similarity searches) providing links to other sources like the Therapeutic Target Database (TTD) [30], KEGG [23], Protein Data Bank (PDB) [31], Uniprot [32], Pfam [33], NCBI [22], TCM-ID [26, 27]. There are several other databases that are specific to geographic regions. For example, CMKb is a resource for Australian aboriginal medicinal knowledge base [34], which catalogs information related to taxonomy, phytochemistry, biogeography, biological activities and images of medicinal plants. The different species included in the database are linked to information sources from Integrated Taxonomic Information System, NCBI Taxonomy, Australia's Species links-integrated botanical information system and Google images. The bioactive metabolites are linked with PubChem

within the chemoinformatics module of CMKb. Functions for editing and visualization of molecules are also included. Another regional database is Raintree (http://rain-tree.com/ethnic.htm), which contains specific information about ethnic and therapeutic uses of plants specifically from Amazon rainforest. This database contains taxonomic data, phytochemical information, ethnobotanical data and links to clinical abstracts. The various menus and pages linked to database files are designed to provide available information to professionals as well as to those who are new to medicinal plants.

In addition to phytochemical data, there may be value in cataloging genomic or transcriptomic knowledge about medicinal plants. The NCBI at the United States National Library of Medicine hosts 'Plant Genome Central', which provides genome and Expressed Sequence Tags (ESTs) resources resulting from large-scale sequencing projects. One significant challenge with the volumes of EST data sets is that they are often poorly organized. Furthermore, EST data may be of poor quality and contain vector contamination. For transcriptome-based information related to plants, the EGENES database provides a platform for efficient analysis of plant ESTs by linking of genome information with higher order functional information [35]. The overall process involves sequence cleaning, repeat masking, vector masking, sequence assembling and KEGG annotation. EGENES attempts to capture all reactions and pathways in plants based on EST information. However, the utility of this resource is limited by the low number of plant species listed. An interesting effort specifically developed for medicinal plants is the Medicinal Plants Genomics Resource (MPGR: http://medicinalplantgenomics.msu.edu/). The goal of MPGR is to make available transcriptome and metabolome information from taxonomically diverse medicinal plant species. At present, this database contains assemblies of transcriptomes from 11 different medicinal plants.

There is a need to build a comprehensive resource for integrating legacy information with contemporary knowledge related to medicinal plants. Such a resource should provide a link between traditional sources with recent experimental evidence arising from on-going research endeavors relating to plant materials. For example, a catalog of potential medicinal plants could be linked with results from clinical trials involving plant-derived treatments and their associated toxicological studies.

TOOLS AND APPROACH TO ANALYZE THE MOLECULAR INFORMATION

A limited number of plants have whole-genome sequence data available. To date, the majority of genomics resources for plants have come from ESTs. Transcript-level information could be valuable to molecular biology-based research relative to medicinal plants. Transcriptome data has been used to identify putative genes and networks involved in secondary metabolite production in medicinal plants [36–38]. Analysis of transcriptome data can also be helpful in predicting transcription factors, response elements and effector genes involved in bioactive metabolite synthesis [39–41]. For example, ethylene responsive element binding genes were analyzed in *Salvia militiorrhiza* [42]. Another example is the identification of miRNAs, their targets and transcription factors involved in secondary metabolism pathways from *Salvia sclarea* L. [43]. Once EST data are generated and assembled, an essential next step is annotation. There are several resources like KEGG genes, SwissProt, TAIR, NCBI's nonredundant and nucleotide databases that provide a platform for annotation of sequence data. EST data can also be used for mining of molecular markers [44–46]. Identification of molecular markers can be used in studies involving linkage mapping, comparative genomics, identification of different species and distribution of genes on chromosomes [47–50]. Compared to other EST-based markers, Simple Sequence Repeat (SSR) markers have been shown to be most advantageous because of their multi-allelic nature, reproducibility, codominant inheritance, high abundance and extensive genome coverage [51]. SSRLocator is an example of a computational approach for detection and characterization of SSRs and mini-satellite motifs [52].

Bioinformatics approaches can be used to create coexpression networks from transcriptome data, providing possible leads to gene discovery in related plant species. In particular, the use of comparative genomics provides basis for exchange of information among the different species. Plant-specific data sets can be retrieved from PLEXdb [53], GEO [54] and EBI ArrayExpress [55]. Coupled with the study of coexpression networks, it may be possible to discover genes of interest and their function. For example, transcriptome data from barley have been collected and used to create a coexpression network [56]. Results from coexpression analyses were further

used to derive subnetworks ('modules') associated with biological functions, with particular emphasis given to identifying modules related to drought stress and cellulose biosynthesis. This genome scale sequence comparisons have been shown to reveal several Triticeae species-specific genes that are related to specific regulatory networks [56].

Pathway analysis can be valuable approach for identifying potential functional roles of genes. The KEGG is a resource that provides a platform for pathway analysis of secondary metabolites from several organisms [57]. The KEGG Drug database further provides information related to two types of molecular networks: (i) interaction of drugs with target molecules and (ii) biosynthetic pathways of natural products in various organisms. KEGG Drug contains chemical structures or components of prescription and Over-The-Counter (OTC) drugs as well as drugs from TCM [58]. This information could potentially be used for drug discovery from the genomes of plants. Another resource for pathway analysis of secondary metabolites indexed in KEGG is PathPred [59]. PathPred is web server that predicts pathways of multi-step reaction for a given query compound, starting with a similarity search against the KEGG COMPOUND database. This server was designed for pathways associated with microbial biodegradation of environmental compounds and biosynthesis of secondary plant metabolites. Nonetheless, PathPred reflects generalized reactions shared among structurally related compounds.

With a myriad of advances in 'omic' technologies, bioinformatics plays essential role in facilitating systems level understanding of metabolic processes. Integration of transcriptomic and metabolomic data facilitated by data mining techniques offers many opportunities to study metabolic pathways [60]. Expression patterns of intensities of ESTs and mass peaks classified by a batch-learning self-organizing maps revealed regulatory linkages among nutrient deficiency, primary metabolism and glucosinolate metabolism [61]. Gene–metabolite coexpression analysis led to identification of terpene synthase genes involved in volatile compound formation in cucumber [62]. Rischer *et al.* [63] analyzed a gene–metabolite coexpression network of the medicinal plant *Catharanthus roseus* to identify possible genes and metabolites associated with the biosynthesis of terpenoid indole alkaloids. Integrated gene–metabolite expression analyses have thus shown potential for examining metabolic regulation of nonmodel plants of potential medicinal value.

Bioinformatics provides essential mechanisms to analyze bulk information generated from highthroughput techniques. In particular, such approaches have made is possible for the identification of putative genes, pathways and networks involved in synthesis of bioactive metabolites in medicinal plants. In addition to facilitating the analysis of high-throughput data, bioinformatics approaches can be important for connecting scattered pieces of evidence into meaningful hypotheses thereby generating potential leads for experimental validation.

APPROACHES TO OPTIMIZE NATURAL PRODUCT USAGE AND REMEDIES

Plant extracts often possess more than one active ingredient that can make it difficult to fully understand the mechanisms of action for plant-based remedies. Furthermore, herbal drug remedies are generally prepared using a combination of herbs [64]. Taken together, these aspects pose serious challenges toward the rationalization of use of herbal remedies. Chemical fingerprinting and bioactive metabolite determination can provide some insight into the chemical and biological activity of such types of remedies [65, 66]. Characterization of chemical constituents and their respective bioactivities can be important for standardizing herbal therapeutic approaches. This aspect can also be important for quality control. Ideally, the optimum combination of active components should be determined before developing herbal remedies. Computational approaches can address such problems of optimization of herbal mixtures, accomodating the variation in chemical composition relative to the biological activity of components in herbal remedies. The relationship between chemical composition and biological activity can be analyzed quantitatively to predict the activity of herbal medicine. Such a relationship is known as Quantitative composition–activity relationship (QCAR)[67]. This approach is similar to the Quantitative structure–activity relationship (QSAR), which depends on structural information to predict the activity [68] (uses for QSAR are described in section 6). However, the structural information for the different chemical constituents of herbal drugs is generally not available, which renders QCARs as more applicable. QCAR models are applicable for determination of optimum combination of active ingredients in herbal remedies. However,

this approach seems to assume that activity of a mixture of herbs can be associated with a relatively small number of components. Although the prediction accuracy of QCAR model is higher than QSAR model, the overall accuracy still remains low. The utility of this model can be improved by more careful experimental design and precise bioassays [69]. Machine-learning techniques have been used to build QCAR models. A hybrid Genetic Algorithm-Artificial Neural Network approach to build a computational prediction QCAR model was proposed by Nayak *et al.* [70]. This model evaluated the optimal combination of constituents in a herbal-based remedy for diabetes. Additional approaches using Multiple Linear Regression, Artificial Neural Network and Support Vector Machines have also been shown to effectively model relationship between chemical composition and activity relationship [69]. In their study, they used their QCAR model (ANN based) to predict the optimal combination for components of Qi-Xue-Bing-Zhi-Fang (QXF), a TCM treatment of cardiovascular disease that decreases plasma lipid levels. The computational results were examined by administering the combinations to rats fed on a high cholesterol diet. The predicted optimal combination significantly reduced the total plasma cholesterol level, thereby validating the model predictions. Fuzzy logic-based approaches have also been proposed to identify the best possible remedy from thousands of candidate plants [71].

To develop a holistic understanding of the multi-component or multi-species nature of herbal remedies, network-based systems biology appoaches have also been proposed [72–75]. The details of network-based approaches in the context of TCM pharmacology and drug discovery has been reviewed previously by Zhao *et al.* [75]. This review also describes case studies that focus on demonstrating the utility of applying such approaches. Such network-based analyses may allow for the pharmacological evaluation of complex recipes leading to development and standardization of optimal remedies. The utility of network-based approaches can be best described through two examples.

In the first example of a network-based approach, Li *et al.* [73] proposed a technique for uncovering combination rules of TCM. They developed a Distance-based Mutual Information Model (DMIM) for identification of relationships of interest among herbs in different TCM formulae. This DMIM model involves a combination of mutual information entropy and between-herb-distance metrics to score herb interactions and constructs an herb network with the potential to uncover combination rules of TCM. In addition, they provide the concept of 'comodules' across multilayer herb–biomolecule–disease network to explore the combination mechanism. This network-based strategy identified strongly connected herbs and herb pairs with potential angiogenic activity. Results were experimentally evaluated using *in vitro* assays where DMIM extracted modules of herb and herb interactions displayed angiogenic activities and synergistic effects.

In a second example of making use of a network-based strategy, Li *et al.* [72] proposed synergistic combinations that could be used in a high-throughput way. In this work, they developed a Network target-based Identification of Multicomponent Synergy (NIMS) algorithm that consists of two components: Topology Score and Agent Score. The topology score is generated from the topological features of the background network related to disease condition and drug actions, while the agent score is based on the similarity of agent phenotypes. These scores are used to calculate a Synergy Score that is used to rank the synergistic effect of agent combinations. This NIMS model was used to prioritize synergistic combinations among 63 agents used in TCM with respect to their effect on angiogenesis. Based on the output, known synergistic combinations were highly ranked. The predicted combinations were further examined using *in vitro* assays for angiogenesis where the efficacy of a synergistic combination was scored. The scores from *in vitro* assays followed the same order as predicted by the NIMS model.

In the context of herbal remedy optimization, it may also be important to consider that many plant species have documented evidence for use in treating different diseases. Even for a single disease, there may exist a multitude of associated medicinal herbs. Additionally, several factors (e.g. availability, active compound, adverse effect, dose and price) can play a role in determining the suitability of a plant as possible effective medicinal source. There is great opportunity for the bioinformatics community to leverage heterogeneous data integration, machine-learning approaches and network-based approaches for identifying possible combinations of medicinal plants that might be used for efficacious treatments.

DISCOVERY OF NEW DRUG SOURCES AND BIOACTIVE METABOLITES

Numerous plant species have been used for a variety of medicinal purposes, including those that have become commercially important drugs (e.g. diosgenin from *Dioscorea nipponica* [76] and analgesic aspirin from willow bark (*Salix sp.*) [77]). There is thus immense potential that could be harnessed from the biodiversity within the plant kingdom. Traditional ethnobotanical patterns of uses have driven the identification and purification of many important compounds and drug formulations [3, 4]. Bioscreening of plant-based products guided by traditional knowledge has also provided an approach for discovery of potential new sources for drugs [3]. Computational approaches may be a cost-effective strategy to identify new plant sources and associated active ingredients, as well as used to uncover patterns in phytochemical data and drive the discovery of new pharmaceutical products. One such approach involves the combining of phylogenetic analyses with phytochemical studies. This approach is built on the premise that closely related plant species may share similar biochemical properties. Indeed, the distribution of medicinal properties has been shown to follow the phylogenetic pattern of plant species [78, 79]. This idea can be used to predict potential candidates while looking for new plant sources based on chemical similarities. The efficacy of such a phylogenetic-based approach has shown promise for identifying potential drugs from the *Pterocarpus* and *Narcissus* genera. For example, results of phylogenetic analyses indicate that therapeutic activity is significantly constrained by evolutionary history, which provides a basis for selection of target species that may warrant further evaluation [80, 81]. In a similar approach, a phylogenetic analysis of 32 taxa of Amaryllidaceae tribe Galantheae, 6 taxa of other Eurasian genera of Amaryllidaceae was carried out with *Phaedranassa dubia* as outgroup. The phylogenetic analysis was further coupled with alkaloid profiles for 18 taxa using gas chromatography–mass spectrometry (GC–MS) and an assay measuring inhibition of acetylcholinesterase (AChE) activity. AChE inhibitory activity was found in all investigated clades and could be correlated with alkaloid profiles of the plants. However, the lowest IC_{50} values were expressed by extracts containing either galanthamine or lycorine type compounds [82]. Phylogenetic analysis coupled with chemical and activity data from this study thus evaluates the important species within the tribe Galantheae with respect to their AChE inhibitory activity.

Based on the available knowledge related to secondary metabolites from plants, machine-learning approaches have also been shown to be effective [83, 84]. Neural networks have been applied in the classification and prediction of complex compounds in the field of chemistry [85]. The application of neural networks has also been used in commercial drug designing. However, there have been few published studies to date that have used this approach in study of plant-based products. A notable exception is a chemical taxonomy study in the *Asteraceae* genus that used neural networks to predict the occurrence of secondary metabolites [86]. Xue *et al.* [84] used a probabilistic neural network for classification of anticancer activity in molecules derived from plant extracts. In this study, a classification model was generated that describes the essential structural features for anticancer agents. The models were then used to understand the factors that govern the anticancer activity of molecules under consideration, with the goal to identify potential drugs from plants. They used a data set of 102 compounds screened for their *in vitro* anticancer activity in the human rhinopharyngocele cell line KB. Based on their anticancer activity (known ED_{50} values), the data set had compounds that could be grouped as higher, high, moderate and low activity agents. This data set was divided randomly into training, cross-validation and test sets. The training and cross-validation set were used to adjust the parameters of PNNs. The test set was then used to evaluate the performance of the trained network. The predictions for this test set were in conformity with experimental values and supported by 90.9% accuracy.

The QSAR-based approach has been widely used in drug designing and testing. QSAR-based models have been used in the prediction and identification of the bioactive secondary metabolites from medicinal plants [87–89]. For example, a QSAR-based study explored the immunomodulatory compounds from derivatives of coumarinolignoids [89]. This approach involved development of QSAR model and predicted the immunomodulatory activities ($logLD_{50}$ values) of three cleomiscosin molecules (A, B and C). The results from the prediction models were shown to be comparable with the experimentally determined log LD_{50} values in an *in vivo* toxicity model. QSAR studies in medicinal

plants have aimed at correlating the structural aspects of compounds with biological activities. They can thus provide insight into the possible mechanism of action of bioactive compounds. Collectively, QSAR models can be used for discovery of potentially new active ingredients based on structural similarities to known compounds.

Another computational approach that can be used in plant-based drug discovery is virtual screening. Using large libraries of plant-based chemical metabolites available, structure-based virtual screening mine for potential drugs [88, 90]. Petersen *et al.* [88] used such an approach to identify naturally occurring substances that can be used as agonist for PPAR gamma. Their approach involved creating a virtual pharmacophore model from 13 PPAR gamma partial agonists and then virtually screening it for plant derived natural products from the Chinese Natural Product Database (CNPD). The virtual screening revealed that an oleoresin (Oleanonic acid) from *Pistacia lentiscus* was a potential suitable naturally occurring ligand. Effect of oleanonic acid on PPAR gamma activity was subsequently experimentally examined in cell lines transiently transfected with luciferase reporter under the transcriptional control of a fusion of Gal4 DNA binding domain and human PPAR gamma ligand binding domain. Treatment with oleanonic acid resulted in transcriptional activation of PPAR gamma in a dose-dependent manner. However, applicability of virtual screening-based approach does have some limitations arising from vast phytochemical space, higher number of conformations of receptor arising as a result of rotatable bonds and difficulties in calculation of binding affinities [91].

Bioinformatics approaches toward evaluation of patterns among phytochemical data shows the immense potential for identification of new drug candidates. In particular, the application of machine-learning techniques for analysis of computational models and classification of phytochemicals can be used to narrow down the search for new sources. In addition, bioinformatics approach in context of network systems biology for facilitating drug discovery shows promise. Studying the pharmacological aspects in a network perspective may provide essential tools for designing drug discovery strategies. A few examples of the application of this approach the field of plant-based medicine was provided in 'Approaches to Optimize Natural Product Usage and Remedies' section. Arrell *et al.* [92] review opportunities and applicability of network systems biology in general context of drug discovery. Taken together, *in silico* methodologies may be an important approach to reduce the cost and time involved by conventional screening strategies.

AUTHENTICITY OF PLANT MATERIALS (DNA BARCODING)

There are numerous plants that may be of medicinal importance based on their characteristic bioactive secondary metabolites. However, the efficacy of drugs derived from plant sources depends on the reliable identification of correct source plants. There have been documented instances where the usage of incorrect plant material resulted in poisoning. For example, in Europe several cases of renal failure were reported due to incorrect use of the poisonous plant *Aristolochia fangchi* [93]. Another instance was due the incorrect use of Japanese star anise (*Illicium anisatum*) in an herbal tea mixture, which resulted in nausea and vomiting after consumption [94]. To avoid these kinds of incidents, a more reliable method than physical characteristics of plants is required. DNA-based markers have shown to be an efficient and reliable means of correctly identifying species. DNA barcodes are built on a standardized short sequence of DNA from a small region of an organism's genome that can differentiate the species from others in the same kingdom [95, 96]. For animals, the chosen sequence is from the mitochondrial gene cytochrome c oxidase subunit 1 (CO1). There has been previous controversy about which marker should be used for plants [97]. After initial studies showing the efficacy of a matK-based barcode for angiosperms (which constitute the majority of land plants) [98], it and rbcL were chosen in combination to be used as the plant barcode [99].

The utility of a DNA barcoding system to serve as a standard way for species identification will depend on public interfaces that can match DNA barcode sequences from an unidentified species using comprehensive libraries of plant barcodes. A generalized platform for barcode-based classification is the Barcode of Life Database (BOLD) [100]. Different classification algorithms can be leveraged from the Barcode of Life Data Portal [101]. Within the specific context of medicinal plants, the Medicinal Materials DNA Barcode Database (MMDBD) [102] integrates available information on medicinal plants along with their DNA barcodes and provides bioinformatics tools for searching and sequence

comparison. A significant bioinformatics challenge posed by the plant barcode is that the classification methods need to accommodate multi (two) barcodes. Kress and Erickson have reviewed the needs for designing appropriate algorithms for searching barcode database [97].

SAFETY EVALUATION AND TOXICOLOGY STUDIES RELATED TO MEDICINAL PLANTS

Study of the toxicological profile of a given drug is an important step in drug discovery process. Adverse drug reaction can pose a severe threat that can lead to significant complications and perhaps even death. It is thus imperative to have a clear understanding of potential hazards associated with new drugs. As a part of regulatory decisions, appropriate risk assessments studies are required for a potential therapeutic [103]. As such, a large number of drugs fail to reach the market due to potential risk. For pharmaceutical chemical drugs, there are several stages involved in preclinical toxicology testing [104]: (i) acute studies; (ii) repeated dose studies; (iii) genetic toxicity studies; (iv) reproductive toxicity studies; (v) carcinogenicity studies; and (vi) toxicokinetic studies. Even though plant sources may be widely used, there is a lack of comprehensive toxicology data related to plant materials. Toxicology studies are an important part of safety assessment in developing and standardizing drugs from plant materials [105, 106]. Very few phytochemicals have been tested for their associated toxicity. Toxicological studies and safety evaluations can involve significant expenses. For example, carcinogenic toxicity in rodent models can cost approximately $3 million for each standardized study [107]. Computational approaches can offer a cost-effective method to eliminate unsafe products at initial stages of safety assessment [108]. The QSAR has been shown to be an *in silico* approach for toxicity assessment. Valerio *et al.* [107] have shown how QSAR-based tools (LMA [109] and MC4PC [110]) can be used in the modeling and prediction of rodent carcinogenicity for phytochemicals. They tested the predictive performance of two QSAR-based tools approved by FDA on a set of phytochemicals with known carcinogenic potential and a set of synthetic chemicals. MC4PC displayed predictive performance for predicting noncarcinogens. However, the overall performance for predicting carcinogens was poor for both MC4PC and LMA. Statistical learning methods can also be applied in for toxicology testing. These methods do not put restrictions on the features of structures or types of molecules that can be used for predicting the toxicity of a given chemical [111–113]. Statistical learning methods focus on general structural and physicochemical properties rather than structural and chemical types. Although they have not been reported for use in testing the toxicology of phytochemicals, statistical-learning methods have been successfully applied for several chemical compounds. However, the performance of these methods can be practically limited by the quality of molecular descriptors, diversity of training and testing data and the efficiency of statistical learning algorithm. Li *et al.* [114] have evaluated a number of such methods for predicting genotoxicity of chemical compounds [including Probabilistic Neural Network (PNN), k-Nearest Neighbor (kNN), Support Vector Machine (SVM) and Decision Tree (DT)]. In their study, Li *et al.* used agents with known genotoxicity. They divided the compounds randomly into five subsets of equal size. Four subsets were used as training set and the fifth as testing set. Evaluation by an independent validation set was carried out by dividing the compounds into training, testing and independent validation sets. Their results show comparable accuracies for SVM, kNN and PNN with SVM giving highest accuracies (77.8% for positive and 92.7% for negative genotoxic agents).

A related field that also aims at improving the screening of chemicals for toxicity by being quicker and cheaper is toxicogenomics. Toxicogenomics is the integration of genomics and toxicology for investigating the molecular mechanisms associated with the expression of toxicity. Toxicogenomics further aims to derive molecular expression patterns that can predict toxicity. The importance of toxicogenomics in the field of TCM and the methodology currently used has been reviewed by Youns *et al.* [115]. This review also highlights the importance of bioinformatics for predicting molecular actions of different classes of toxicants and for refining pathways to distinguish the effect of different agents that represent a range of toxic effects.

CONCLUSION

Plants can be a valuable source of pharmacologically important compounds. However, plant-specific research has been hampered by a number of resource

challenges. Conventional, mostly manual bioscreening strategies for identifying and studying medicinal plants have not been sufficient to keep pace with current pharmaceutical needs. Bioinformatics approaches may provide an essential set of tools for designing efficient and targeted searches for plant-based remedies. This review highlighted the different aspects associated with medicinal plant research where bioinformatics strategies could be employed to attain significant progress. The combination of bioinformatics strategies may enable a new era of plant-based drug discovery.

Key Points

- Plants have been used as a source of many drugs used in contemporary medicine and may still be an important source for new drugs.
- Data associated with medicinal plants are scattered and generally unlinked, which limits the potential to identify new sources for drugs.
- Bioinformatics approaches can facilitate identification of potential plant sources for future therapeutics.

FUNDING

This work was funded in part by National Institutes of Health, National Library of Medicine (grant R01LM009725).

References

1. Cowan MM. Plant products as antimicrobial agents. *Clin Microbiol Rev* 1999;**12**:564–82.
2. Patwardhan B, Warude D, Pushpangadan P, *et al*. Ayurveda and traditional Chinese medicine: a comparative overview. *eCAM* 2005;**2**:465–73.
3. Fabricant DS, Farnsworth NR. The value of plants used in traditional medicine for drug discovery. *Environ Health Perspect* 2001;**109(Suppl 1)**:69–75.
4. Clarkson C, Maharaj VJ, Crouch NR, *et al*. In vitro antiplasmodial activity of medicinal plants native to or naturalised in South Africa. *J Ethnopharmacol* 2004;**92**:177–91.
5. Strohl WR. The role of natural products in a modern drug discovery program. *Drug Discov Today* 2000;**5**:39–41.
6. Miller J. The discovery of medicines from plants: a current biological perspective1. *Econ Botany* 2011;**65**:396–407.
7. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ* 2003;**22**:151–85.
8. Jorgensen WL. The many roles of computation in drug discovery. *Science* 2004;**303**:1813–8.
9. Saito K, Matsuda F. Metabolomics for functional genomics, systems biology, and biotechnology. *Annu Rev Plant Biol* 2010;**61**:463–89.
10. Manniche L. *An Ancient Egyptian Herbal*. Austin, Texas: University of Texas Press, 1989.
11. Oppenheim AL. Mesopotamian medicine. *Bull Hist Med* 1962;**36**:97–108.
12. Korin YD, Brooks DG, Brown S, *et al*. Effects of prostratin on T-cell activation and human immunodeficiency virus latency. *J Virol* 2002;**76**:8118–23.
13. Cox PA. Saving the ethnopharmacological heritage of Samoa. *J Ethnopharmacol* 1993;**38**:181–8.
14. Buenz EJ, Johnson HE, Beekman EM, *et al*. Bioprospecting Rumphius's Ambonese herbal: Volume I. *J Ethnopharmacol* 2005;**96**:57–70.
15. Wang JF, Li ZR, Cai CZ, *et al*. Assessment of approximate string matching in a biomedical text retrieval problem. *Comput Biol Med* 2005;**35**:717–24.
16. Buenz EJ, Bauer BA, Johnson HE, *et al*. Searching historical herbal texts for potential new drugs. *BMJ* 2006;**333**:1314–5.
17. Sarkar IN. Biodiversity informatics: the emergence of a field. *BMC Bioinformatics* 2009;**10(Suppl 14)**:S1.
18. Wagner JC, Rogers JE, Baud RH, *et al*. Natural language generation of surgical procedures. *Stud Health Technol Informatics* 1998;**52**(Pt 1):591–5.
19. Cao C, Wang H, Sui Y. Knowledge modeling and acquisition of traditional Chinese herbal drugs and formulae from text. *Artificial Intell Med*. 2004;**32**:3–13.
20. Samwald M, Dumontier M, Zhao J, *et al*. Integrating findings of traditional medicine with modern pharmaceutical research: the potential role of linked open data. *Chinese Med* 2010;**5**:43.
21. Chen H, Mao Y, Zheng X, *et al*. Towards semantic e-science for traditional Chinese medicine. *BMC Bioinformatics* 2007;**8(Suppl 3)**:S6.
22. Sayers EW, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2010;**38**:D5–16.
23. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
24. Afendi FM, Okada T, Yamazaki M, *et al*. KNApSAcK Family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol* 2012;**53**:e1.
25. Loub WD, Farnsworth NR, Soejarto DD, *et al*. NAPRALERT: computer handling of natural product research data. *J Chem Inform Comput Sci* 1985;**25**:99–103.
26. Wang JF, Zhou H, Han LY, *et al*. Traditional Chinese medicine information database. *Clin Pharmacol Therap* 2005;**78**:92–3.
27. Ji ZL, Zhou H, Wang JF, *et al*. Traditional Chinese medicine information database. *J Ethnopharmacol* 2006;**103**:501.
28. Chen X, Zhou H, Liu YB, *et al*. Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *Br J Pharmacol* 2006;**149**:1092–03.
29. Ye H, Ye L, Kang H, *et al*. HIT: linking herbal active ingredients to targets. *Nucleic Acids Res* 2011;**39**:D1055–9.
30. Liu X, Zhu F, Ma X, *et al*. The Therapeutic Target Database: an internet resource for the primary targets of approved, clinical trial and experimental drugs. *Exp Opin Therapeutic Targets* 2011;**15**:903–12.

31. Bernstein FC, Koetzle TF, Williams GJ, *et al*. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch Biochem Biophys* 1978;**185**:584–91.
32. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. *Database J Biol Databases Curation* 2011;**2011**:bar009.
33. Punta M, Coggill PC, Eberhardt RY, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2012;**40**: D290–301.
34. Gaikwad J, Khanna V, Vemulpad S, *et al*. CMKb: a web-based prototype for integrating Australian Aboriginal customary medicinal plant knowledge. *BMC Bioinformatics* 2008;**9(Suppl 12)**:S25.
35. Masoudi-Nejad A, Goto S, Jauregui R, *et al*. EGENES: transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG. *Plant Physiol* 2007;**144**:857–66.
36. Li Y, Luo HM, Sun C, *et al*. EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. *BMC Genomics* 2010;**11**:268.
37. Chen S, Luo H, Li Y, *et al*. 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in Panax ginseng. *Plant Cell Rep* 2011;**30**:1593–601.
38. Luo H, Li Y, Sun C, *et al*. Comparison of 454-ESTs from Huperzia serrata and Phlegmariurus carinatus reveals putative genes involved in lycopodium alkaloid biosynthesis and developmental regulation. *BMC Plant Biol* 2010;**10**:209.
39. Yuan D, Tu L, Zhang X. Generation, annotation and analysis of first large-scale expressed sequence tags from developing fiber of Gossypium barbadense L. *PloS One* 2011; **6**:e22758.
40. Kavitha K, Venkataraman G, Parida A. An oxidative and salinity stress induced peroxisomal ascorbate peroxidase from Avicennia marina: molecular and functional characterization. *PPB/Societe Francaise De Physiologie Vegetale* 2008;**46**: 794–804.
41. Cabral A, Stassen JH, Seidl MF, *et al*. Identification of Hyaloperonospora arabidopsidis transcript sequences expressed during infection reveals isolate-specific effectors. *PloS One* 2011;**6**:e19328.
42. Xu B, Huang L, Cui G, *et al*. [Functional genomics of Salvia militiorrhiza IV–analysis of ethylene responsive element binding protein gene]. *Zhongguo Zhong Yao Za Zhi = Zhongguo Zhongyao Zazhi = China Journal Of Chinese Materia Medica* 2009;**34**:2564–6.
43. Legrand S, Valot N, Nicole F, *et al*. One-step identification of conserved miRNAs, their targets, potential transcription factors and effector genes of complete secondary metabolism pathways after 454 pyrosequencing of calyx cDNAs from the Labiate Salvia sclarea L. *Gene* 2010;**450**:55–62.
44. Joshi RK, Kar B, Nayak S. Exploiting EST databases for the mining and characterization of short sequence repeat (SSR) markers in Catharanthus roseus L. *Bioinformation* 2011;**5**: 378–81.
45. Victoria FC, da Maia LC, de Oliveira AC. In silico comparative analysis of SSR markers in plants. *BMC Plant Biol* 2011;**11**:15.
46. Zeng S, Xiao G, Guo J, *et al*. Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, Epimedium sagittatum (Sieb. Et Zucc.) Maxim. *BMC Genomics* 2010;**11**:94.
47. Wang CM, Liu P, Yi C, *et al*. A first generation microsatellite- and SNP-based linkage map of Jatropha. *PloS One* 2011;**6**:e236–32.
48. Diaz A, Fergany M, Formisano G, *et al*. A consensus linkage map for molecular markers and Quantitative Trait Loci associated with economically important traits in melon (Cucumis melo L.). *BMC Plant Biol* 2011;**11**:111.
49. Korotkova N, Borsch T, Quandt D, *et al*. What does it take to resolve relationships and to identify species with molecular markers? An example from the epiphytic Rhipsalideae (Cactaceae). *Am J Botany* 2011;**98**:1549–72.
50. Venuprasad R, Bool ME, Quiatchon L, *et al*. A QTL for rice grain yield in aerobic environments with large effects in three genetic backgrounds, TAG. Theoretical and applied genetics. *Theoretische Und Angewandte Genetik* 2011;**124**(2): 323–332.
51. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 2005;**23**:48–55.
52. da Maia LC, Palmieri DA, de Souza VQ, *et al*. SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *Int J Plant Genomics* 2008;**2008**:412696.
53. Wise RP, Caldo RA, Hong L, *et al*. BarleyBase/PLEXdb. *Methods Mol Biol* 2007;**406**:347–63.
54. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* 2006;**411**:352–69.
55. Rocca-Serra P, Brazma A, Parkinson H, *et al*. ArrayExpress: a public database of gene expression data at EBI. *Comptes Rendus Biologies* 2003;**326**:1075–8.
56. Mochida K, Uehara-Yamaguchi Y, Yoshida T, *et al*. Global landscape of a co-expressed gene network in barley and its application to gene discovery in Triticeae crops. *Plant Cell Physiol* 2011;**52**:785–803.
57. Kanehisa M, Goto S, Furumichi M, *et al*. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;**38**:D355–60.
58. Kanehisa M. Representation and analysis of molecular networks involving diseases and drugs. *Genome Inform Int Conf Genome Inform* 2009;**23**:212–3.
59. Moriya Y, Shigemizu D, Hattori M, *et al*. PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res* 2010;**38**:W138–43.
60. Saito K, Hirai MY, Yonekura-Sakakibara K. Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'. *Trends Plant Sci* 2008;**13**:36–43.
61. Hirai MY, Yano M, Goodenowe DB, *et al*. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc Natl Acad Sci USA* 2004;**101**:10205–10.
62. Mercke P, Kappers IF, Verstappen FW, *et al*. Combined transcript and metabolite analysis reveals genes involved in spider mite induced volatile formation in cucumber plants. *Plant Physiol* 2004;**135**:2012–24.
63. Rischer H, Oresic M, Seppanen-Laakso T, *et al*. Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in Catharanthus roseus cells. *Proc Natl Acad Sci USA* 2006;**103**:5614–9.
64. Chan K. Progress in traditional Chinese medicine. *Trends Pharmacol Sci* 1995;**16**:182–7.

65. Fischedick JT, Hazekamp A, Erkelens T, *et al*. Metabolic fingerprinting of Cannabis sativa L., cannabinoids and terpenoids for chemotaxonomic and drug standardization purposes. *Phytochemistry* 2010;**71**:2058–73.
66. Yu F, Kong L, Zou H, *et al*. Progress on the screening and analysis of bioactive compounds in traditional chinese medicines by biological fingerprinting analysis. *Combinatorial Chem High Throughput Screening* 2010;**13**:855–68.
67. Zhao XP, Fan XH, Yu J, *et al*. [A method for predicting activity of traditional Chinese medicine based on quantitative composition-activity relationship of neural network model]. *Zhongguo Zhong Yao Za Zhi = Zhongguo Zhongyao Zazhi = China Journal Of Chinese Materia Medica* 2004;**29**: 1082–5.
68. Brown AC, Fraser TR. On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *J Anat Physiol* 1868; **2**:224–42.
69. Wang Y, Wang X, Cheng Y. A computational approach to botanical drug design by modeling quantitative composition-activity relationship. *Chem Biol Drug Design* 2006;**68**:166–72.
70. Nayak SK, Patra PK, Padhi P, Panda A. Optimization of herbal drugs using soft computing approach. *Int J Logic Computat* 2010;**1**:34–9.
71. Dudek G, Grzywna ZJ, Willcox ML. Classification of antituberculosis herbs for remedial purposes by using fuzzy sets. *Bio Sys* 2008;**94**:285–9.
72. Li S, Zhang B, Zhang N. Network target for screening synergistic drug combinations with application to traditional Chinese medicine. *BMC Sys Biol* 2011;**5(Suppl 1)**:S10.
73. Li S, Zhang B, Jiang D, *et al*. Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae. *BMC Bioinformatics* 2010;**11(Suppl 11)**:S6.
74. Li S. [Network target: a starting point for traditional Chinese medicine network pharmacology]. *Zhongguo Zhong Yao Za Zhi = Zhongguo Zhongyao Zazhi = China Journal Of Chinese Materia Medica* 2011;**36**:2017–20.
75. Zhao J, Jiang P, Zhang W. Molecular networks for the study of TCM pharmacology. *Brief Bioinformatics* 2010;**11**: 417–30.
76. Kang TH, Moon E, Hong BN, *et al*. Diosgenin from dioscorea nipponica ameliorates diabetic neuropathy by inducing nerve growth factor. *Biol Pharm Bull* 2011;**34**: 1493–8.
77. Miner J, Hoffhines A. The discovery of aspirin's antithrombotic effects. Texas Heart Institute J / from the Texas Heart Institute of St. Luke's Episcopal Hospital, Texas Children's Hospital 2007;34:179–86.
78. Prinzing A, Durka W, Klotz S, *et al*. The niche of higher plants: evidence for phylogenetic conservatism. *Proc Biol Sci Roy Soc* 2001;**268**:2383–9.
79. Paton AJ, Springate D, Suddee S, *et al*. Phylogeny and evolution of basils and allies (Ocimeae, Labiatae) based on three plastid DNA regions. *Mol Phylogenet Evol* 2004;**31**: 277–99.
80. Ronsted N, Savolainen V, Molgaard P, Jager AK. Phylogenetic selection of Narcissus species for drug discovery. *Biochem Sys Ecol* 2008;**36**:417–22.
81. Saslis-Lagoudakis CH, Klitgaard BB, Forest F, *et al*. The use of phylogeny to interpret cross-cultural patterns in plant use and guide medicinal plant discovery: an example from pterocarpus (leguminosae). *PloS One* 2011;**6**: e22275.
82. Larsen MM, Adsersen A, Davis AP, *et al*. Using a phylogenetic approach to selection of target plants in drug discovery of acetylcholinesterase inhibiting alkaloids in Amaryllidaceae tribe Galantheae. *Biochem Sys Ecol* 2005;**38**:1026–34.
83. Fraser L, Mulholand A, Fraser DD. Classification of limonoids and protolimonoids using neural networks. *Phytochem Anal* 1997;**8**:301–11.
84. Xue CX, Zhang XY, Liu MC, *et al*. Study of probabilistic neural networks to classify the active compounds in medicinal plants. *J Pharm Biomed Anal* 2005;**38**:497–507.
85. Gasteiger J, Zupan E. Neural networks in Chemistry. *Angew Chem Int Ed Engl* 1993;**32**:503–27.
86. Schwabe T, Ferreira MJP, Alvarenga SAV, Emerenciano VP. Neural networks for secondary metabolite prediction in Artemisia Genus (Asteraceae), Internet Electron. *J Mol Des* 2004;**4**:9–16.
87. Zhou X, Li Y, Chen X. Computational identification of bioactive natural products by structure activity relationship. *J Mol Graphic Modelling* 2010;**29**:38–45.
88. Petersen RK, Christensen KB, Assimopoulou AN, *et al*. Pharmacophore-driven identification of PPARgamma agonists from natural sources. *J Comp Aided Mol Design* 2011;**25**:107–16.
89. Yadav DK, Meena A, Srivastava A, *et al*. Development of QSAR model for immunomodulatory activity of natural coumarinolignoids. *Drug Design Dev Ther* 2010;**4**:173–86.
90. Sakkiah S, Thangapandian S, Lee KW. Ligand based virtual screening and molecular docking studies to identify the critical chemical features of potent cathepsin D inhibitors. *Chem Biol Drug Design* 2012;**80**:64–79.
91. Shoichet BK. Virtual screening of chemical libraries. *Nature* 2004;**432**:862–5.
92. Arrell DK, Terzic A. Network systems biology for drug discovery. *Clin Pharmacol Therap* 2010;**88**:120–5.
93. Vanherweghem JL, Depierreux M, Tielemans C, *et al*. Rapidly progressive interstitial renal fibrosis in young women: association with slimming regimen including Chinese herbs. *Lancet* 1993;**341**:387–91.
94. Johanns ES, van der Kolk LE, van Gemert HM, *et al*. [An epidemic of epileptic seizures after consumption of herbal tea]. *Nederlands Tijdschrift Voor Geneeskunde* 2002;**146**: 813–6.
95. Hebert PD, Ratnasingham S, deWaard JR. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci Roy Soc* 2003; **270(Suppl 1)**:S96–9.
96. Miller SE. DNA barcoding and the renaissance of taxonomy. *Proc Natl Acad Sci USA* 2007;**104**:4775–6.
97. Kress WJ, Erickson DL. DNA barcodes: genes, genomics, and bioinformatics. *Proc Natl Acad Sci USA* 2008;**105**: 2761–2.
98. Lahaye R, van der Bank M, Bogarin D, *et al*. DNA barcoding the floras of biodiversity hotspots. *Proc Natl Acad Sci USA* 2008;**105**:2923–8.
99. CBOL Plant Working Group. A DNA barcode for land plants. *Proc Natl Acad Sci USA* 2009;**106**:12794–7.

100. Ratnasingham S, Hebert PD. Bold: the barcode of life data system (http://www.barcodinglife.org). *Mol Ecol Notes* 2007;**7**:355–64.

101. Sarkar IN, Trizna M. The Barcode of Life Data Portal: bridging the biodiversity informatics divide for DNA barcoding. *PloS One* 2011;**6**:e14689.

102. Lou SK, Wong KL, Li M, *et al*. An integrated web medicinal materials DNA database: MMDBD (Medicinal Materials DNA Barcode Database). *BMC Genomics* 2010; **11**:402.

103. US Department of Health and Human Services Food and Drug Administration *'Guidance for Industry: Q8(R2) Pharmaceutical Development'* 2009. Rockville, MD.

104. US Department of Health and Human Services Food and Drug Administration *'Guidance for Industry: M3 Nonclinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals'* 1997. Rockville, MD.

105. Wu KM, Ghantous H, Birnkrant DB. Current regulatory toxicology perspectives on the development of herbal medicines to prescription drug products in the United States. *Food Chem Toxicol Int J published for the Br Industrial Biol Res Ass* 2008;**46**:2606–10.

106. Wu KM, Farrelly J, Birnkrant D, *et al*. Regulatory toxicology perspectives on the development of botanical drug products in the United States. *Am J Therapeut* 2004;**11**: 213–7.

107. Valerio LGJr, Arvidson KB, Busta E, *et al*. Testing computational toxicology models with phytochemicals. *Mol Nutr Food Res* 2010;**54**:186–94.

108. Rusyn I, Daston GP. Computational toxicology: realizing the promise of the toxicity testing in the 21st century. *Environ Health Perspect* 2010;**118**:1047–50.

109. Yang C, Hasselgren CH, Boyer S, *et al*. Understanding genetic toxicity through data mining: the process of building knowledge by integrating multiple genetic toxicity databases. *Toxicol Mechanisms Methods* 2008;**18**:277–95.

110. Saiakhov RD, Klopman G. MultiCASE expert systems and the REACH initiative. *Toxicol Mechanisms Methods* 2008;**18**: 159–75.

111. Mosier PD, Jurs PC, Custer LL, *et al*. Predicting the genotoxicity of thiophene derivatives from molecular structure. *Chem Res Toxicol* 2003;**16**:721–32.

112. He L, Jurs PC, Custer LL, *et al*. Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers. *Chem Res Toxicol* 2003;**16**:1567–80.

113. Mattioni BE, Kauffman GW, Jurs PC, *et al*. Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble. *J Chem Inform Comput Sci* 2003;**43**:949–63.

114. Li H, Ung CY, Yap CW, *et al*. Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem Res Toxicol* 2005;**18**:1071–80.

115. Youns M, Hoheisel JD, Efferth T. Toxicogenomics for the prediction of toxicity related to herbs from traditional Chinese medicine. *Planta Medica* 2010;**76**:2019–25.