· Reviews ·

Application of Bioinformatics and Systems Biology in Medicinal Plant Studies

DENG You-ping[1,2 *], AI Jun-mei[3], XIAO Pei-gen[4]

1. SpecPro, Vicksburg, MS 39180, USA
2. Department of Biological Sciences, University of Southern Mississippi, Hattiesburg, MS 39406, USA
3. School of Computing, University of Southern Mississippi, Hattiesburg, MS 39406, USA
4. Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing 100193, China

Abstract: One important purpose to investigate medicinal plants is to understand genes and enzymes that govern the biological metabolic process to produce bioactive compounds. Genome wide high throughput technologies such as genomics, transcriptomics, proteomics and metabolomics can help reach that goal. Such technologies can produce a vast amount of data which desperately need bioinformatics and systems biology to process, manage, distribute and understand these data. By dealing with the "omics" data, bioinformatics and systems biology can also help improve the quality of traditional medicinal materials, develop new approaches for the classification and authentication of medicinal plants, identify new active compounds, and cultivate medicinal plant species that tolerate harsh environmental conditions. In this review, the application of bioinformatics and systems biology in medicinal plants is briefly introduced.

Key words: bioinformatics; functional genomics; medicinal plants; metabolomics; systems biology
DOI: 10.3969/j.issn.1674-6384.2010.03.003

Introduction

Bioinformatics is the application and development of computational tools for biological sciences. The main tasks of bioinformatics are the management and analyses of biological data. Over the past few decades, rapid developments of high-throughput technologies such as genomics, proteomics, and metabolomics have generated a huge amount of data in molecular biology. To fully understand biological processes, bioinformatics has increasingly become an essential tool for many biological areas (Kann, 2009). Because each "omics" data is not isolated, systems biology aims to understand biology as a system and tries to integrate different "omics" data sources for the analysis of networks, regulation, and understanding of how the system works from a whole system point of view (Ray, Chong, and Gough, 2002). The publications of the completed *Arabidopsis thaliana* genome sequence (Aubourg *et al*, 2000) and sequence for rice genome (Goff *et al*, 2002) as well as some agricultural organisms such as maize (Schnable *et al*, 2009) open a new era for plant and agricultural sciences. Bioinformatics and systems biology play a critical role in processing, distributing and interpreting this novel and hard to be handled data. Chinese herbal medicine is an important medicinal resource for the world, but the genome information of medicinal plants is far behind model organisms and other economic plants. In order to identify functional genes and enzymes that control bioactive compound production of medical plants, improve quality of traditional medicinal materials, develop new methods for the classification and authentication, and cultivate medicinal plants species with pathogen and abiotic stress and other hard environmental resistances, and save endangered traditional medicinal species, more genomics, proteomics and metablomics information needs to be produced. As next generation sequencing technology has dramatically reduced the cost of genome sequencing, bioinformatics

* Corresponding author: Deng YP Address: SpecPro, on site Collaborator of US Army Environmental Research Development Center, 3909 Halls Ferry Road, Vicksburg, MS 39180, USA Tel: +601634 5229 Fax: +601634 3838 E-mail: Youping.deng@usace.army.mil
Received: January 28, 2010; Revised: April 12, 2010; Accepted: April 14, 2010

and systems biology will be required more in the field (Xu, 2003). In this review, the potential applications of transcirptomics, metabolomics, proteomics, genomics, epigenomics, and systems biology in medicinal plants are described. The review discusses the importance of information management of medicinal plants with the focus more on medicinal plants themselves other than bioactive compounds with prediction of targets and their medicinal effects using bioinformatics.

Transcriptomics

Transcriptomics is to analyze the information of a whole transcriptome of an organism. A transcriptome is the set of all RNA molecules, including mRNA, tRNA, rRNA, and non-coding RNA produced in one or group of cells. Generally, transcriptomics indicates the global analysis of gene expression profile at mRNA level. Genome-wide gene expression technologies mainly include cDNA-AFLP, SAGE, DNA microarray (or gene chip), and oligo-microarray. Currently the most popular transcriptomics method is oligo-microarray due to its whole genome coverage advantage. Microarray can be used to measure gene expression changes of medicinal plants. Materials used for microarray could be different organs for the same organism, the same organ at different development stages, or even cultured plant cells. Others include different geographic locations, natural growth environments, or cultivation conditions. Unfortunately, there are only a few publications using high throughput gene expression profiles to study gene expression changes in medicinal plants. cDNA microarray was utilized to investigate the gene expression profiles of hairy root of *Salvia miltiorrhiza* Bunge at different stages (Cui *et al*, 2007). The ultimate goals for a gene expression study are to identify genes that are responsible for regulating active medicinal compounds, anti-pathogen infection, or adaption to hard environment. A transcriptome analysis approach was applied to the isolation of trichome-specific genes from a medicinal plant *Cistus creticus* subsp. *creticus* (L.) Heywood (Falara *et al*, 2008). Ethylene responsive element binding protein genes for *S. militiorrhiza* was analyzed using functional transcriptomics (Xu *et al*, 2009). DNA microarray can also be used for authentication of medicinal plants. However, it usually doesn't use whole genome information, instead, 5S ribosomal RNA genes (Carles *et al*, 2005) or 18SRNA genes (Zhu, Fushimi, and Komatsu, 2008) across close related species are used.

To design a DNA microarray, the best way is to use the whole genome sequence information for a specific organism. Since almost no medicinal plants whole genome has been sequenced, an alternative way to accomplish this feat is to get the whole transcriptome information through generating expression sequence tags (ESTs). We see that more ESTs from several medicinal plants such as *Panax quinquefolius* L. (Wu *et al*, 2010; Chen *et al*, 2008), *Huperzia serrata* (Thunb. ex Murray) Trev. (Luo *et al*, 2009), *P. notoginseng* (Burk.) F. H. Chen. (He, Zhu, and Zhang, 2008), *Rehmannia glutinosa* Libosch. (Sun *et al*, 2010), and *Catharanthus roseus* (L.) G. Don (Shukla *et al*, 2006) have been produced. My team has built an automatic system for large scale EST sequence retrieval, assembly, function and pathway analysis, and all processed data were put into ESTMD, EST model database (Pirooznia and Deng, 2007; Deng *et al*, 2006a). The system has been successfully used for the analysis of plant EST sequences (Thara *et al*, 2004) and animal EST sequences (Pirooznia *et al*, 2007; Pirooznia *et al*, 2009; Pozhitkov *et al*, 2009; Deng *et al*, 2006b; Pirooznia *et al*, 2009; Pozhitkov *et al*, 2009; Boyko *et al*, 2006).

Since the traditional sequencing technologies such as Sanger Sequencing usually generate less than ten thousands ESTs for a non-model organism, cDNA microarrays were widely designed for the gene expression study. Our collaborators have successfully used the strategy for microarray design with a wide range of purposes of such gene expression studies (Gust *et al*, 2009; Gong *et al*, 2008a; Gong *et al*, 2007; Majji *et al*, 2009a; Majji *et al*, 2009b; Milev-Milovanovic *et al*, 2009; Pondugula *et al*, 2006). Recently developed next generation sequencing technologies, such as illumine, 454, and solid approaches have revolutionized the EST generation. My team has analyzed millions of EST sequences of Northern bobwhite (*Colinus virginianus* L.) (Rawat *et al*, unpublished data) and red earthworm from the 454 platform. Assembled unique sequences are used to design oligo-microarrays (Gong *et al*, 2008b). This new strategy enables us to design an array with almost the whole genome transcriptome.

Metabolomics

Metabolomics is to survey the information of the

whole metabolome of an organism. Metabonomics is more specifically defined as "the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification" (Nicholson, Lindon, and Holmes, 1999). Metabolome is the complete set of small-molecule metabolites including metabolic intermediates, hormones and other signaling molecules, and secondary metabolites that exist in a biological sample, such as a single organism (Oliver *et al*, 1998). Because metabolomics actually focuses on real physiological and biochemical processes, it becomes increasingly more important in biomedical researches. About 7900 metabolites are contained at the Human Metabolome Database (HMDB is available at: www. hmdb. ca) and based on the analysis of the current scientific literature the database is still far from complete. In contrast, much more is known in plants. More than 50 000 metabolites have been characterized from the plant kingdom. Metabolomics should be extremely important for medicinal plants, because one of our objectives to study medicinal plants is to find active compounds, which, in fact, are metabolites. Therefore, the fast development of metabolomics technology provides us a valuable opportunity to advance the studies of medicinal plants. One advantage of conducting metabolomics is that genomics information is not needed. Metabolomics can be used to compare metabolite quantitative changes in medicinal materials between different organisms, ages, origins, organs, developmental stages, environmental cultivation and culture conditions, and processing methods. It can help us understand the metabolic pathways for the production of these bioactive compounds generate metabolic fingerprinting of medicinal plants for the authentication and quality control, classify medicinal plants, and establish a quantitative version of chemotaxonomic analysis to advance our knowledge of the evolutional relationship of medicinal plants. Moreover, the technology can help us quickly and efficiently recognize the composition and quantity of a medical plant, find new potential compounds (Wang *et al*, 2007), identify existing compounds, and confirm and generate new knowledge of the pharmacological and toxic effects of the plant.

Metabolic fingerprinting techniques have been successfully employed to evaluate the quality of herbal material and phytopharmaceutical with MS and NMR methods (Van der *et al*, 2009). For instance, metabolomics profiling was used to differentiate 12 Cannabis sativa cultivars (Choi *et al*, 2004b; Choi *et al*, 2004a) and determine the quantity of ginkgolic acids from Ginkgo leaves and products in six commercial Ginkgo products (Choi *et al*, 2004a). Metabolic profiling of *Angelica acutiloba* (Sieb. & Zucc.) Kitag. roots, utilizing gas chromatography-time-of-flight-mass spectrometry, could be used accurately to assess the quality of the medicinal materials based on cultivation area and cultivar *via* multivariate pattern recognition (Tianniam *et al*, 2008). Comparative metabolomics strategy coupled with cell- and gene-based assays were used for species classification and anti-inflammatory bioactivity validation of medicinal Echinacea species, i.e. *Echinacea purpurea* (L.) Moench, *E. pallida* Nutt., and *E. angustifolia* DC. (Hou *et al*, 2009). More metabolic profiling have been conducted in the medicinal plant barrel medic (*Medicago truncatula*) which is also a model organism for legume biology (Schliemann *et al*, 2008; Farag *et al*, 2008; Barsch *et al*, 2006b; Barsch *et al*, 2006a).

Popular methods used to perform metabolomic experiments are mass spectrometry (MS) and nuclear magnetic resonance (NMR). My team has used metabolomic profiling to study type 2 diabetes. We focused more on lipid compounds and tried to measure lipid species (nearly 400) simultaneously (Welti *et al*, 2007), named lipidomics. Using human blood plasma samples, we compared the lipid profiles among type 2 diabetic (n = 62), pre-diabetic (n = 23) and healthy subjects (n = 35). Ten to 13 lipid species were significantly changed (1.2–3 fold) in pre-diabetic and diabetic subjects compared with controls. Our data indicate that lipidomic technology is a valuable high-throughput method that can potentially be used to identify novel biomarkers for diagnosis and treatment of type 2 diabetes (Mao *et al*, 2009). An automated electrospray ionization-tandem mass spectrometry (ESI-MS/MS) approach was used for the lipidomics technology in our study. Fig. 1 illustrates the ion current chromatography of PA, PC and PE of human plasma. The same technology is being used in the studies of breast cancer and prostate cancer with some interesting preliminary data being achieved (Deng, unpublished data). The foundation of metabolomics

technology is already developed in China. For instance, qualitative compound composition analyses have been extensively studied for a broad purposes in China (Li *et al*, 2008; Li *et al*, 1996b; Li *et al*, 1996a; Li *et al*, 1995; Wu *et al*, 1998; Yu *et al*, 1999). We need to set up a quantitative metabolomics platform and bioinformatics certainly plays an important role to make it functional.

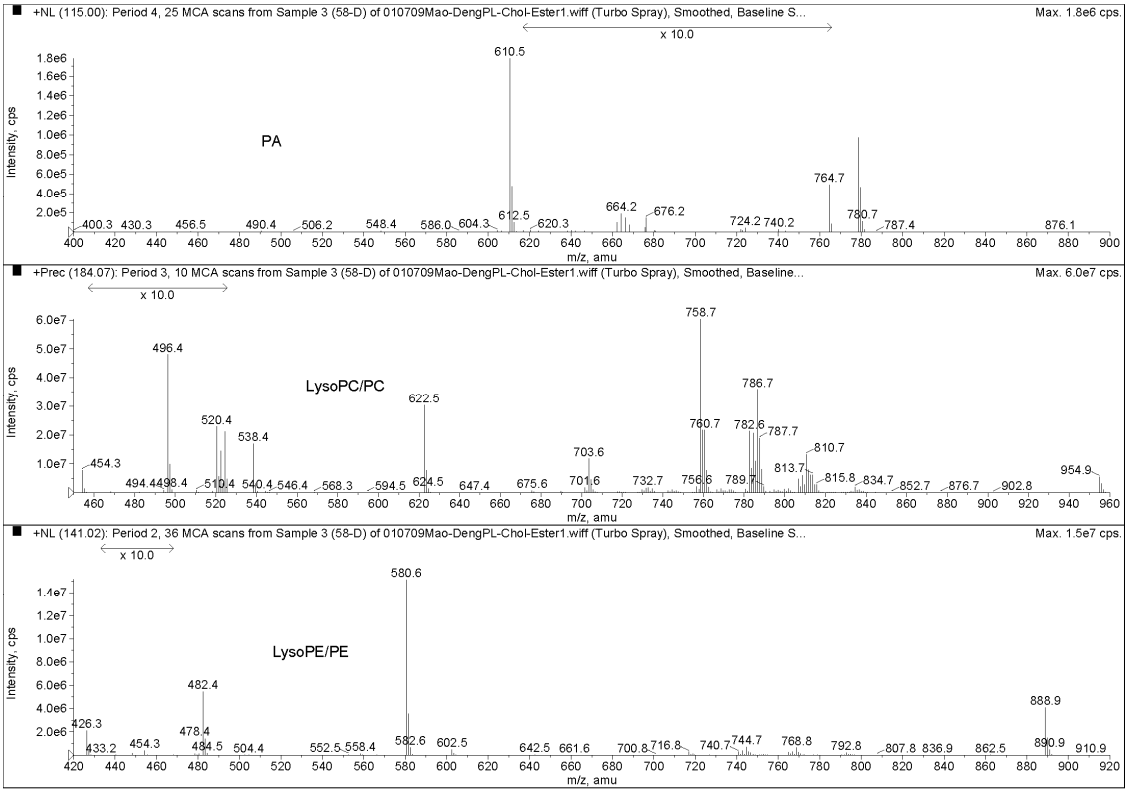

Fig. 1 The ion current chromatogram of ESI-MS/MS analysis of phospholipids PA, PC, and PE of human plasma sample

Proteomics

Proteomics is the information of a whole proteome, which refers to the entire complement of proteins including the modulation made to a particular set of proteins within an organism or system. This will alter over time or under different environmental conditions that a cell or organism undergoes (Wilkins *et al*, 1996). Besides dealing with protein expression changes, proteomics also covers the studies of protein structure and function, protein and protein interaction, and protein post-translational modifications that majorly include phosphorylation and ubiquitination. Proteins can be subjected to methylation, acetylation, glycosylation, oxidation, and nitrosylation (Blackstock and Weir, 1999). Published proteomic data of medicinal plants is most obtained from medicago. Most reports focus on the protein expression and phosphorylation changes at various conditions (Castillejo *et al*, 2009; Dam *et al*, 2009; Aloui *et al*, 2009; Van Noorden *et al*, 2007; Prayitno *et al*, 2006). There is no report of proteomic observations on Chinese medicinal plants, and the major reason is the lack annotation of protein and gene sequence information of Chinese medicinal plants. The current situation of proteomics studies indicates that it is urgent for us to generate more genome and proteome information of medicinal plants. My team has worked on some data analysis of proteomics data such as comparison of different feature selection and classification methods for MALDI-MS data (Liu *et al*, 2009b). Based on whole proteome information, we found some new drosophila chitinases (Zhu *et al*, 2004).

Genomics and epigenomics

Genomics is the study of an organism's whole genome. Genome refers to all of the DNA sequences in an organism. So far, there are very few genomes of medicinal plants that have been fully sequenced. The draft genome of the transgenetic tropical fruit tree papaya (*Carica papaya* Linnaeus), which also possesses medicinal value, was sequenced (Ming *et al*, 2008). The genome sequencing of *M. truncatula* is close to being

completed. However, there are a high volume of publications related to identification and authentication of medicinal plants at the DNA level. DNA-based techniques that do not require whole genome information like PCR, RFLP, AFLP, RAPD, and sequencing are employed to resolve ambiguities in plant identification and discrimination (Titanji, Ngwa, and Ngemenya, 2007; Chen *et al*, 2010; Song *et al*, 2009; Yu, 2010). Since there are more than 5000 medicinal plants and many of them are endangered species, it is essential to select one or more model medicinal plants to sequence their whole genomes which will profoundly enhance the research of medicinal plants.

Epigenomics is the whole genome level study of epigenetic elements. Epigenetic mechanisms, including DNA methylation and modifications to histone proteins, regulate high-order DNA structure and gene expression. Small RNA studies could also be part of epigenomics. Epigenomics is important for us to understand the mechanisms of gene expression changes of medicinal plants. Our team has also begun to apply epigenomics related technologies to environmental and human disease studies.

"Omics" data analyses

"Omics" data, usually high-dimensional data, requires common statistical and machine learning methods. Here, we use transcriptomics data as examples to talk about "omics" data analyses. After a microarray design is finished, the next critical steps are experimental design, data quality control and normalization. Our team has developed a web application tool for two-channel microarray interwoven loop design which any biologist can easily use for analysis (Pirooznia *et al*, 2008a). We also used a simulated strategy and developed a hybrid normalization method for microarray normalization (Pirooznia and Deng, 2007). We developed a distribution-free convolution model for background correction of oligonucleotide microarray data as well (Chen *et al*, 2009). The original purpose of a DNA microarray experiment is to identify differential genes across different conditions. Next, a clustering algorithm could be applied in order to cluster genes that have similar behaviors. The genes in the same cluster should have similar functions. For instance, we could employ clustering algorithms to identify a group of functionally similar genes that are involved in the same biological process or pathway that affects an active compound's production in a medicinal plant. Clustering algorithms can be divided into unsupervised and supervised algorithms. Unsupervised algorithms mainly include hierarchical K-means and self organizing map methods. K-means is one of the most popular methods used in gene expression data analyses due to its high computational performance. However, it is well known that K-means might converge to a local optimum, and its result is subject to the initialization process, which randomly generates the initial clustering. In other words, different runs of K-means on the same input data might produce different solutions.

Genetic K-means Algorithm (GKA) is an algorithm which hybridizes a genetic algorithm with the K-means algorithm. This hybrid approach combines the robust nature of the genetic algorithm with the high performance of the K-means algorithm. As a result, GKA will always converge to the global optimum faster than other genetic algorithms. We have successfully developed a faster version of GKA, i.e. FGKA which runs 20 times faster than GKA. We further developed an extension to FGKA, *Incremental Genetic K-means Algorithm* (*IGKA*) that inherits all the advantages of FGKA including the convergence to the global optimum. This addition outperforms FGKA when the mutation probability is small (Lu *et al*, 2004). We also developed two new algorithms: markov chain correlation based clustering algorithm (Deng, Chokalingam, and Zhang, 2005a) and particle swarm optimization K-means algorithm (Deng *et al*, 2006c) for clustering gene expression data. Both algorithms perform better than the existing K-means algorithm. We also developed an automatic method for the determination of cluster number in clustering microarray data based on dynamic validity index (Shen *et al*, 2005). We have proposed a graph theory method to accurately characterize co-regulated genes for Arabidopsis microarray data (Rawat, Seifert, and Deng, 2008). In terms of gene functional analysis, we have successfully predicted the function of a yeast gene YJL103C based on gene expression clustering (Deng *et al*, 2005b). Recently we created a better web tool to search Gene Ontology more efficiently (Pirooznia *et al*, 2008b).

More advanced microarray and other high throughput data analysis allow the use of classification

supervised machine learning algorithms to identify predictive biomarkers. The strategy has been successfully used for the diagnosis and prognosis of various diseases such as cancer (Cooper, 2001). An advantageous application of classification algorithms in medicinal plants could be authentication and quality control, as well as characterization, of the right species that could aid in the evolutional studies of medicinal plants. Actually, all the information related to a medicinal material such as DNA markers, gene expression data, chemical compound composition and quantity, morphology and so on could be used as training data for running classification algorithms. The goal here is to identify a list of efficient markers to precisely distinguish different medicinal materials. Before building a classifier, feature selection is important. There are many feature selection methods including support vector machine recursive feature elimination (SVM-RFE), chisquare, infogain, gainratio, relief, wrapper, and CSF. Generally used classification algorithms include decision tree J48, random forest (RF), Naivy Bayes (NB), simple logistic (SL), RBF Neural Nets, MLP Neural Nets and support vector machines (SVMs). We have compared different feature selection and classification algorithms based on multiple types of microarray data using training and test cross-validation to judge the performances. We found SVMs usually outperforms other classification methods (Pirooznia *et al*, 2008c). It is necessary to build user-friendly software with graphical a user interface (GUI) to classify high-throughput data for general biologists. So far, almost all classification algorithms need experienced statisticians to perform the task. Since SVM is a very good algorithm for microarray data classification, many users may use it for data analysis by creating a GUI. We (Pirooznia and Deng, 2006) have developed a user-friendly, java GUI application allowing users to perform SVM training, classification and prediction. We demonstrated that our software can accurately classify genes into functional categories based upon expression data from DNA microarray experiments. We have also developed Parallel Multicategory Support Vector Machines (PMC-SVM) for classifying microarray data (Zhang *et al*, 2006). Recently, we proposed a new gene feature selection method called Recursive Feature Addition (RFA), which combines supervised learning and statistical similarity measures, that outperforms other popular feature selection methods including the Support Vector Machine Recursive Feature Elimination (SVMRFE) (Liu *et al*, 2009a).

Databases

Database is an important aspect of bioinformatics. It is critical to develop web based searchable databases related to medicinal plants. Current medicinal plant databases are mainly medicinal resource databases which usually include taxonomy, biogeography, cultivation conditions, medical organs, medicinal functions, and biological active compounds of medicinal plant species as well as images of individual species. These databases could be generated based on a country or district's need. For example, CMKb is a web-based relational prototype database for integrating Australian Aboriginal customary medicinal plant knowledge. Some databases only contain the medicinal plants (Jarayaman, 2000) that can treat a specific diseases such as antifertility (Ghosh and Chattopadhyay, 2000), asthma (Kasirajan *et al*, 2007), and diabetes (Arulrayan *et al*, 2007). China has developed many nationwide and regional medicinal plant resource databases, some of which are prescription or active and inactive compounds, chemical and targets. One disadvantage is that most of these resource databases are only written in Chinese and the knowledge cannot be shared with the world.

Databases can also be developed based on genetic and genomics information. For instance, there is a lot of DNA finger printing information available that could be stored in a web accessible and searchable database. DNA, mRNA level, and gene function for a group of medicinal plants or even a single plant species could be used to develop medicinal plant genome databases. Functional genomics, proteomics and metabolomics data can also be managed to build functional "omic" databases for medicinal plants. For example, if more metabolites are available, we could develop a metabolite database of medicinal plants. My team has participated in the development of BeetleBase, an online Tribloium genome database (Wang *et al*, 2007). We also developed EST database (Deng *et al*, 2005b) and RiboaptDB, a comprehensive online searchable database of ribozyme and aptamers (Thodima, Pirooznia, and Deng, 2006).

Integrative systems biology

Integrative systems biology involves the integration

of genomics, transcriptomics, proteomics, and metabolomics information to create a whole system network view of a biological entity using bioinformatics (Saito K). Integrating transcirptomics and metabolomics data can help us to predict gene function particularly for genes involved in complicated pathways that can produce bioactive constituents. Through integrated metabolite and transcript profiling, a biosynthetic mechanism for hispidol in *Medicago truncate*, cell cultures were characterized (Farag *et al*, 2009). We have done an integrative metal-analysis for lung cancer marker gene identification (Jiang *et al*, 2004) using different microarray data resources. Gene shaving (GS) methods based on Random Forests (RF) and Fisher's Linear Discrimination (FLD) were applied separately to the joint data set for cancer gene selection. The two methods discovered 13 and 10 marker genes (5 in common) for RF and FLD, respectively, with expression patterns differentiating diseased from normal samples. We also found that patterns of 36 genes were significantly correlated with patient survival ($P < 0.05$). One of the most challenging tasks in systems biology in the post-genomic era is to reconstruct the transcriptional regulatory networks using reverse engineering algorithms. This is also the major task of systems biology. By collaborating with computer scientists, we have recently developed an ensemble learning approach to reverse-engineering transcriptional regulatory networks from time-series gene expression data. Our method starts with building an ensemble of decision trees for each microarray data to capture the association between the expression levels of yeast genes and the binding of transcription factors to gene promoter regions as determined by integrating with another type of omics data, a chromatin immunoprecipitation microarray (ChIP-chip) data. Cross-validation experiments show that the method is more accurate and reliable than the naive decision tree algorithm and several other ensemble learning methods (Ruan *et al*, 2009). We also designed a slice pattern model to reconstruct gene regulatory network using time-series data (Wang *et al*, 2009).

Conclusion

This article demonstrated the importance of bioinformatics and systems biology in medicinal plant research. Functional genomics, proteomics, metabolomics and integrated systems biology have the ability to advance the studies of modern medicinal plants and so help us to understand how bioactive compounds are produced and improving the quality of medicinal materials. Understanding of medicinal evolutional relationship and new methods for the authentication and quality control of the medicinal plants may also be benefited. However, the generation of "omics" data for medicinal plants is still in the beginning stages compared with other biomedical and agricultural areas, leading to an urgent need to develop more genomic resources of medicinal plants. We expect that there will be a rapid development of functional genomic approaches for studying traditional and herbal medicine in the future.

Acknowledgments

We thank Mitchell S. Wilbanks from US Army Environmental Research Development Center for reading the manuscript and giving the suggestions.

References

Aloui A, Recorbet G, Gollotte A, Robert F, Valot B, Gianinazzi-Pearson V, schi-Smiti S, Dumas-Gaudot E, 2009. On the mechanisms of cadmium stress alleviation in *Medicago truncatula* by arbuscular mycorrhizal symbiosis: A root proteomic study. *Proteomics* 9: 420-433.

Arulrayan N, Rangasamy S, James E, Pitchai D, 2007. A database for medicinal plants used in the treatment of diabetes and its secondary complications. *Bioinformation* 2: 22-23.

Aubourg S, Boudet N, Kreis M, Lecharny A, 2000. In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. *Plant Mol Biol* 42: 603-613.

Barsch A, Carvalho HG, Cullimore JV, Niehaus K, 2006a. GC-MS based metabolite profiling implies three interdependent ways of ammonium assimilation in *Medicago truncatula* root nodules. *J Biotechnol* 127: 79-83.

Barsch A, Tellstrom V, Patschkowski T, Kuster H, Niehaus K, 2006b. Metabolite profiles of nodulated alfalfa plants indicate that distinct stages of nodule organogenesis are accompanied by global physiological adaptations. *Mol Plant Microbe Interact* 19: 998-1013.

Blackstock WP, Weir MP, 1999. Proteomics: Quantitative and physical mapping of cellular proteins. *Trends Biotechnol* 17: 121-127.

Boyko EV, Smith CM, Thara VK, Bruno JM, Deng Y, Starkey SR, Klaahsen DL, 2006. Molecular basis of plant gene expression during aphid invasion: Wheat Pto- and Pti-like sequences are involved in interactions between wheat and Russian wheat aphid (Homoptera: Aphididae). *J Econ Entomol* 99: 1430-1445.

Carles M, Cheung MK, Moganti S, Dong TT, Tsim KW, Ip NY, Sucher NJ, 2005. A DNA microarray for the authentication of toxic traditional Chinese medicinal plants. *Planta Med* 71: 580-584.

Castillejo MA, Maldonado AM, Dumas-Gaudot E, Fernandez-Aparicio M, Susin R, Diego R, Jorrin JV, 2009. Differential expression proteomics to investigate responses and resistance to *Orobanche crenata* in *Medicago truncatula*. *BMC Genomics* 10: 294.

Chen SL, Yao H, Han JP, Liu C, Song JY, Shi LC, Zhu YJ, Ma XY, Gao T, Pang XH, Luo K, Li Y, Li XW, Jia XC, Lin YL, Leon C, 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 5: e8613.

Chen SL, Sun YQ, Song JY, Li Y, Li CJ, Hu SN, Li XW, Yao H, Zhang XW, 2008. Analysis of expressed sequence tags (EST) from *Panax quinquefolium* root. *Acta Pharm Sin* 43: 657-663.

Chen Z, McGee M, Liu Q, Kong M, Deng Y, Scheuermann RH, 2009. A distribution-free convolution model for background correction of oligonucleotide microarray data. *BMC Genomics* 10(S1): 19.

Choi YH, Choi HK, Peltenburg-Looman AM, Lefeber AW, Verpoorte R, 2004a. Quantitative analysis of ginkgolic acids from Ginkgo leaves and products using 1H-NMR. *Phytochem Anal* 15: 325-330.

Choi YH, Kim HK, Hazekamp A, Erkelens C, Lefeber AW, Verpoorte R, 2004b. Metabolomic differentiation of Cannabis sativa cultivars using ^{1}H-NMR spectroscopy and principal component analysis. *J Nat Prod* 67: 953-957.

Cooper CS, 2001. Applications of microarray technology in breast cancer research. *Breast Cancer Res* 3: 158-175.

Cui GH, Huang LQ, Qiu DY, Yuan Y, Fu GF, 2007. Functional genomics studies of *Salvia miltiorrhiza* II--gene expression profiling of different stage of hairy root. *China J Chin Mater Med* 32: 1267-1272.

Dam S, Laursen BS, Ornfelt JH, Jochimsen B, Staerfeldt HH, Friis C, Nielsen K, Goffard N, Besenbacher S, Krusell L, Sato S, Tabata S, Thogersen IB, Enghild JJ, Stougaard J, 2009. The proteome of seed development in the model legume *Lotus japonicus*. *Plant Physiol* 149: 1325-1340.

Deng YP, Chokalingam V, Zhang CY, 2005. Markov chain correlation based clustering of gene expression data. *ITCC* 2005: 750-755.

Deng YP, Dong Y, Brown SJ, Zhang CY, 2006a. An integrated web-based model for management, analysis and retrieval of EST biological information. *LNCS* 3842: 931-938.

Deng YP, Dong Y, Thodima V, Clem RJ, Passarelli AL, 2006b. Analysis and functional annotation of expressed sequence tags from the fall armyworm *Spodoptera frugiperda*. *BMC Genomics* 7: 264.

Deng Y, He T, Wu Y, Vanka P, Yang G, Huang Y, Yao H, Brown SJ, 2005b. Computationally analyzing the possible biological function of YJL103C--an ORF potentially involved in the regulation of energy process in yeast. *Int J Mol Med* 15: 123-127.

Deng YP, Kayarat D, Elasri MO, Brown SJ, 2006c. Microarray data clustering using particle swarm optimization K-means algorithm. *Patt Recognit Mach Intellig* 1: 114-119.

Falara V, Fotopoulos V, Margaritis T, Anastasaki T, Pateraki I, Bosabalidis AM, Kafetzopoulos D, Demetzos C, Pichersky E, Kanellis AK, 2008. Transcriptome analysis approaches for the isolation of trichome-specific genes from the medicinal plant *Cistus creticus* subsp. *creticus*. *Plant Mol Biol* 68: 633-651.

Farag MA, Deavours BE, de FA, Naoumkina M, Dixon RA, Sumner LW, 2009. Integrated metabolite and transcript profiling identify a biosynthetic mechanism for hispidol in *Medicago truncatula* cell cultures. *Plant Physiol* 151: 1096-1113.

Farag MA, Huhman DV, Dixon RA, Sumner LW, 2008. Metabolomics reveals novel pathways and differential mechanistic and elicitor-specific responses in phenylpropanoid and isoflavonoid biosynthesis in *Medicago truncatula* cell cultures. *Plant Physiol* 146: 387-402.

Ghosh S, Chattopadhyay S, 2000. Development of a computer system in search of antifertility drug from indigenous plants. *Acta Pol Pharm* 57: 459-463.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong JP, Miguel T, Paszkowski U, Zhang SP, Colbert M, Sun WL, Chen LL, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S, 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92-100.

Gong P, Guan X, Inouye LS, Deng YP, Pirooznia M, Perkins EJ, 2008a. Transcriptomic analysis of RDX and TNT interactive sublethal effects in the earthworm *Eisenia fetida*. *BMC Genomics* 9 (S1): 15.

Gong P, Guan X, Inouye LS, Pirooznia M, Indest KJ, Athow RS, Deng YP, Perkins EJ, 2007. Toxicogenomic analysis provides new insights into molecular mechanisms of the sublethal toxicity of 2,4,6-trinitrotoluene in *Eisenia fetida*. *Environ Sci Technol* 41: 8195-8202.

Gong P, Pirooznia M, Guan X, Vera JC, Liang C, Deng YP, Zhang CY, Perkins EJ, 2008b. A strategy for high throughput microarray oligo probe design using 454 and sanger sequencing data of *Eisenia Fetida* expressed sequence tags. The 2008 World Congress in Computer Science Computer Engineering and Applied Computing; 2008 July 14-17 USA, Biocomp; Las Vegas, Nevada, USA, 1040-1043.

Gust KA, Pirooznia M, Quinn MJ, Jr, Johnson MS, Escalon L, Indest KJ, Guan X, Clarke J, Deng YP, Gong P, Perkins EJ, 2009. Neurotoxicogenomic investigations to assess mechanisms of action of the munitions constituents RDX and 2,6-DNT in Northern bobwhite (*Colinus virginianus*). *Toxicol Sci* 110: 168-180.

He FM, Zhu YP, Zhang YZ, 2008. Identification and characterization of differentially expressed genes involved in pharmacological activities of roots of *Panax notoginseng* during plant growth. *Plant Cell Rep* 27: 923-930.

Hou CC, Chen CH, Yang NS, Chen YP, Lo CP, Wang SY, Tien YJ, Tsai PW, Shyur LF, 2009. Comparative metabolomics approach coupled with cell- and gene-based assays for species classification and anti-inflammatory bioactivity validation of Echinacea plants. *J Nutr Biochem* [Epub ahead of print].

Jarayaman KS, 2000. India pushes ahead with plant database. *Nature* 405: 267.

Jiang HY, Deng YP, Chen HS, Tao L, Sha QY, Chen J, Tsai CJ, Zhang SL, 2004. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 5: 81.

Kann MG, 2009. Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. *Brief Bioinform* 11: 96-110.

Kasirajan B, Maruthamuthu R, Gopalakrishnan V, Arumugam K, Asirvatham H, Murali V, Mohandass R, Bhaskar A, 2007. A database for medicinal plants used in treatment of asthma. *Bioinformation* 2: 105-106.

Li MH, Chen JM, Peng YP, Wu Q, Xiao PG, 2008. Investigation of Danshen and related medicinal plants in China. *J Ethnopharmacol* 120: 419-426.

Li WK, Pan JQ, Lu MJ, Xiao PG, Zhang RY, 1996a. Anhydroicaritin 3-*O*-rhamnosyl(1→2) rhamnoside from epimedium koreanum and a reappraisal of other rhamnosyl (1→2, 1→3 and 1→4) rhamnoside structures. *Phytochemistry* 42: 213-216.

Li WK, Xiao PG, Tu GZ, Ma LB, Zhang RY, 1995. Flavonol glycosides from *Epimedium koreanum*. *Phytochemistry* 38: 263-265.

Li WK, Zhang RY, Xiao PG, 1996b. Flavonoids from *Epimedium wanshanense*. *Phytochemistry* 43: 527-530.

Liu QZ, Sung AH, Chen ZX, Liu JZ, Huang XD, Deng YP, 2009a. Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *PLoS One* 4: e8250.

Liu QZ, Sung AH, Qiao MY, Chen ZX, Yang JY, Yang MQ, Huang XD, Deng YP, 2009b. Comparison of feature selection and classification for MALDI-MS data. *BMC Genomics* 10(S1): S3.

Lu Y, Lu SY, Fotouhi F, Deng YP, Brown SJ, 2004. Incremental genetic K-means algorithm and its application in gene expression data analysis. *BMC Bioinformatics* 5: 172.

Luo HM, Sun C, Li Y, Wu Q, Song JY, Wang DL, Jia XC, Li RT, Chen SL, 2009. Analysis of expressed sequence tags from the *Huperzia serrata* leaf for gene discovery in the areas of secondary metabolite biosynthesis and development regulation. *Physiol Plantarum* 139: 1-12.

Majji S, Thodima V, Arnizaut A, Deng Y, May W, Sittman D, Waldbieser GC, Hanson L, Cuchens MA, Bengten E, Chinchar VG, 2009a. Expression profiles of cloned channel catfish (*Ictalurus punctatus*) lymphoid cell lines and mixed lymphocyte cultures. *Dev Comp Immunol* 33: 224-234.

Majji S, Thodima V, Sample R, Whitley D, Deng Y, Mao J, Chinchar VG, 2009b. Transcriptome analysis of Frog virus 3, the type species of the genus *Ranavirus*, family Iridoviridae. *Virology* 391: 293-303.

Mao JH, Shen-Wu M, Roth M, Welti R, Deng YP, 2009. Lipidomic profiles reveal the potential useful plasma biomarkers associated with type II diabetes in black and white patients. *Biocomp* 2009: 774-779.

Milev-Milovanovic I, Majji S, Thodima V, Deng YP, Hanson L, Arnizaut A, Waldbieser G, Chinchar VG, 2009. Identification and expression analyses of poly [I:C]-stimulated genes in channel catfish (*Ictalurus punctatus*). *Fish Shellfish Immunol* 26: 811-820.

Ming R, Hou SB, Feng Y, Yu QY, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen CX, Qian WB, Shen JG, Du P, Eustice M, Tong E, Tang HB, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang JM, Wang JP, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang XY, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei HR, Irikura B, Paidi M, Jiang N, Zhang WL, Presting G, Windsor A, Navajas-Perez R, Torres MJ, Feltus FA, Porter B, Li YJ, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang JM, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M, 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452: 991-996.

Nicholson JK, Lindon JC, Holmes E, 1999. 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29: 1181-1189.

Oliver SG, Winson MK, Kell DB, Baganz F, 1998. Systematic functional analysis of the yeast genome. *Trends Biotechnol* 16: 373-378.

Pirooznia M, Crawford DL, Vanwye JD, Gibson R, Zhang CY, Deng YP, McLean T, 2009. Genomic resources for *Karenia brevis*. The 2009 World Congress in Computer Science Computer Engineering and Applied Computing; 2009 July 13-16 USA, Biocomp; Las Vegas, Nevada, USA, 228-234.

Pirooznia M, Deng YP. 2006. SVM Classifier - a comprehensive java interface for support vector machine classification of microarray data. *BMC Bioinformatics* 7(S4): S25.

Pirooznia M, Deng YP, 2007. Efficiency of hybrid normalization of microarray gene expression: A simulation study. *AINA Workshops* 1: 739-744.

Pirooznia M, Gong P, Guan X, Inouye LS, Yang K, Perkins EJ, Deng YP, 2007. Cloning, analysis and functional annotation of expressed sequence tags from the Earthworm *Eisenia fetida*. *BMC Bioinformatics* 8(S7): S7.

Pirooznia M, Gong P, Yang JY, Yang MQ, Perkins EJ, Deng YP, 2008a. ILOOP - a web application for two-channel microarray interwoven loop design. *BMC Genomics* 9(S2): S11.

Pirooznia M, Habib T, Perkins EJ, Deng YP, 2008b. GOfetcher: A database with complex searching facility for gene ontology. *Bioinformatics* 24: 2561-2563.

Pirooznia M, Yang JY, Yang MQ, Deng YP, 2008c. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* 9(S1): S13.

Pondugula SR, Raveendran NN, Ergonul Z, Deng YP, Chen J, Sanneman JD, Palmer LG, Marcus DC, 2006. Glucocorticoid regulation of genes in the amiloride-sensitive sodium transport pathway by semicircular canal duct epithelium of neonatal rat. *Physiol Genomics* 24: 114-123.

Pozhitkov A, Pirooznia M, Zhang CY, Gong P, Perkins EJ, Deng YP, Brouwer M, 2009. Generation and analysis of expressed sequence tags from the Sheepshead minnow (*Cyprinodon variegatus*). The 2009 World Congress in Computer Science Computer Engineering and Applied Computing; 2009 July 13-16 USA, Biocomp; Las Vegas, Nevada, USA, 174-180.

Prayitno J, Imin N, Rolfe BG, Mathesius U, 2006. Identification of ethylene-mediated protein changes during nodulation in *Medicago truncatula* using proteome analysis. *J Proteome Res* 5: 3084-3095.

Rawat A, Deng YP, 2008. Novel implementation of conditional co-regulation by graph theory to derive co-expressed genes from microarray data. *BMC Bioinformatics* 9(S9): S7.

Ray LB, Chong LD, Gough NR, 2002. Computational biology. *Sci STKE* 2002: eg10.

Ruan JH, Deng YP, Perkins EJ, Zhang WX, 2009. An ensemble learning approach to reverse-engineering transcriptional regulatory networks from time-series gene expression data. *BMC Genomics* 10(S1): S8.

Schliemann W, Ammer C, Strack D, 2008. Metabolite profiling of mycorrhizal roots of *Medicago truncatula*. *Phytochemistry* 69: 112-146.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen WZ, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin JK, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan CZ, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren LY, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van BP, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han YJ, Lee H, Li PH, Lisch DR, Liu SZ, Liu ZJ, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang LX, Yu Y, Zhang LF, Zhou SG, Zhu QH, Bennetzen JL, Dawe RK, Jiang JM, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK, 2009. The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326: 1112-1115.

Shen JD, Chang SI, Lee ES, Deng YP, Brown SJ, 2005. Determination of cluster number in clustering microarray data. *Appl Math Comput* 169: 1172-1185.

Shukla AK, Shasany AK, Gupta MM, Khanuja SP, 2006. Transcriptome analysis in *Catharanthus roseus* leaves and roots for comparative terpenoid indole alkaloid profiles. *J Exp Bot* 57: 3921-3932.

Song JY, Yao H, Li Y, Li XW, Lin YL, Liu C, Han JP, Xie CX, Chen SL, 2009. Authentication of the family Polygonaceae in *Chinese pharmacopoeia* by DNA barcoding technique. *J Ethnopharmacol* 124: 434-439.

Sun P, Guo YH, Qi JJ, Zhou LL, Li XE, 2010. Isolation and expression analysis of tuberous root development related genes in *Rehmannia glutinosa*. *Mol Biol Rep* 37: 1069-1079.

Thara VK, Seilaniantz AR, Deng YP, Dong YH, Yang YN, Tang XY, Zhou JM, 2004. Tobacco genes induced by the bacterial effector protein AvrPto. *Mol Plant Microbe Interact* 17: 1139-1145.

Thodima V, Pirooznia M, Deng Y, 2006. RiboaptDB: A comprehensive database of ribozymes and aptamers. *BMC Bioinformatics* 7(S2): S6.

Tianniam S, Tarachiwin L, Bamba T, Kobayashi A, Fukusaki E, 2008. Metabolic profiling of *Angelica acutiloba* roots utilizing gas chromatography-time-of-flight-mass spectrometry for quality assessment based on cultivation area and cultivar via multivariate pattern recognition. *J Biosci Bioeng* 105: 655-659.

Titanji VP, Ngwa AA, Ngemenya M, 2007. Applications of biotechnology techniques to the study of medicinal plants. *Afr J Med Med Sci* 36(S): 23-29.

Van der KF, Maltese F, Choi YH, Kim HK, Verpoorte R, 2009. Quality control of herbal material and phytopharmaceuticals with MS and NMR based metabolic fingerprinting. *Planta Med* 75: 763-775.

Van Noorden GE, Kerim T, Goffard N, Wiblin R, Pellerone FI, Rolfe BG, Mathesius U, 2007. Overlap of proteome changes in *Medicago truncatula* in response to auxin and *Sinorhizobium meliloti*. *Plant Physiol* 144: 1115-1131.

Wang LJ, Wang SZ, Li YH, Paradesi MS, Brown SJ, 2007. BeetleBase: The model organism database for *Tribolium castaneum*. *Nucleic Acids Res* 35: D476-D479.

Wang JN, Gu SP, Tan RX, Wang JW, 2007. Methology for quick finding of leading compounds based on botanical metabonomic concept of determination for mixture by ^{1}H-NMR. *Chin Tradit Herb Drugs* 38(6): 812-814.

Wang YD, Wang GH, Yang B, Tao HJ, Yang JY, Deng YP, Liu YL, 2009. Reconstruct gene regulatory network using slice pattern model. *BMC Genomics* 10(S1): S2.

Welti R, Roth MR, Deng YP, Shah J, Wang XM, 2007. Lipidomics: ESI-MS/MS-based profiling to determine the function of genes involved in metabolism of complex lipids. In: Nikolau BJ, Wurtele ES, editors. *Concepts Plant Metabolomics*. Kluwer Academic Publishers, New York, 87-92.

Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser DF, 1996. From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Nature Biotechnol* 14: 61-65.

Wu Q, Song JY, Sun YQ, Suo FM, Li CJ, Luo HM, Liu Y, Li Y, Zhang XW, Yao H, Li XW, Hu SN, Sun C, 2010. Transcript profiles of *Panax quinquefolius* from flower, leaf and root bring new insights into genes related to ginsenosides biosynthesis and transcriptional regulation. *Physiol Plantarum* 138: 134-149.

Wu QL, Wang SP, Du LJ, Yang JS, Xiao PG, 1998. Xanthones from *Hypericum japonicum* and *H. henryi*. *Phytochemistry* 49: 1395-1402.

Xu B, Huang L, Cui G, Mao Y, Zhang H, 2009. Functional genomics of *Salvia militiorrhiza* IV--analysis of ethylene responsive element binding protein gene. *China J Chin Mater Med* 34: 2564-2566.

Xu ZN, 2003. Relationship between development of bioinformatics and study of Chinese herbal medicine. *Chin Tradit Herb Drugs* 34(6): 481-486.

Yu HH, Yang ZL, Yang X, Tan ZF, Shu X, Liu RN, 2010. Advances in studies on ITS sequence of medicinal plants germplasm resources. *Chin Tradit Herb Drugs* 41(3): 491-496.

Yu SC, Wu QL, Wang LW, Yang JS, Xiao PG, 1999. Flavonoid glycosides from *Thalictrum przewalskii*. *J Asian Nat Prod Res* 1: 301-306.

Zhang CY, Li P, Rajendran A, Deng YP, Chen DQ, 2006. Parallelization of multicategory support vector machines (PMC-SVM) for classifying microarray data. *BMC Bioinformatics* 7(S4): S15.

Zhu QS, Deng YP, Vanka P, Brown SJ, Muthukrishnan S, Kramer KJ, 2004. Computational identification of novel chitinase-like proteins in the *Drosophila melanogaster* genome. *Bioinformatics* 20: 161-169.

Zhu S, Fushimi H, Komatsu K, 2008. Development of a DNA microarray for authentication of ginseng drugs based on 18S rRNA gene sequence. *J Agric Food Chem* 56: 3953-3959.