Annals of Genetics and Genetic Disorders

Review Article
Published: 10 Aug, 2018

TAGPT: A Web Server for Prediction of Trait Associated Genes using Gene Expression Data

Dwijesh Chandra Mishra, Sanjeev Kumar, Lal SB, Arijit Saha, Chaturvedi KK, Neeraj Budhlakoti and Anil Rai*

Department of Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, India

Abstract

Background: Transcriptomics plays a vital role in functional annotation of genes. This is the well proven fact that genes related to a particular trait can be predicted through modelling and analysis of transcriptomic data. Massive gene expression data related to different useful traits are available in public repositories. However, this valuable information are underutilized due to the lack of appropriate methodologies and efficient tools for identification of related genes and transcription factors. In order to properly utilize this genomic information an efficient algorithm and a sophisticated computational tool based on efficient algorithm is of urgent need.

Results: In this study, an efficient algorithm has been developed by utilizing non-linear penalized regression technique Kernelized Least Absolute Shrinkage and Selection Operator (LASSO) for predicting genes related to useful traits using gene expression data. Also, a user friendly Web based tool (TAGPT, Trait Associated Genes Prediction Tool) has been developed based on this algorithm. The efficiency of TAGPT over existing tools has been demonstrated by predicting the trait specific genes using five sets of gene expression data from various experiments related to gene expression. It has been observed that the developed algorithm behind TAGPT is accurate and fast while predicting the specific genes related to a particular trait by analysing the predicted genes.

Conclusion: TAGPT is user friendly Web server which provides the list of predicted genes related to a specific trait with their expression value along with their evaluation measures (sensitivity, specificity, classification accuracy etc.). This web server can be accessed by using any web browsers like Internet Explorer, Google Chrome, Mozilla Firefox etc. TAGPT is freely available at http://cabgrid.res.in/tagpt.

Keywords: Biotic and abiotic stress; Microarray; Gene expression; Algorithm; Kernelized LASSO; Web server

OPEN ACCESS

Correspondence:
Dwijesh Chandra Mishra, Department of Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, India,
E-mail: dwijesh.mishra@icar.gov.in

Received Date: *20 May 2018*
Accepted Date: *03 Aug 2018*
Published Date: *10 Aug 2018*

Citation:
Mishra DC, Kumar S, Lal SB, Saha A, Chaturvedi KK, Budhlakoti N, et al. TAGPT: A Web Server for Prediction of Trait Associated Genes using Gene Expression Data. Ann Genet Genet Disord. 2018; 1(1): 1003.

Abbreviation

TAGPT- Trait Associated Genes Prediction Tool; FAO- Food and Agriculture Organization; LASSO- Least Absolute Shrinkage and Selection Operator; GEO- Gene Expression Omnibus; SVM- Support Vector Machine; LOOCV- Leave One Out Cross Validation; GUI- Graphical User Interface; JSP- Java Server Pages; HTML- Hypertext Mark-up Language; RWUI- R Web User Interface

Background

Food security is a prime goal of sustainable agriculture (FAO). Gene transfer technology is being used to introduce many important agricultural traits like higher yield, resistance to abiotic or biotic stress, nutritional quality of food etc. into a wide range of food crops [1]. Crops or animals are engineered for useful traits to improve the quantity and quality of the food [2]. Identification of genes related to these useful traits has been a major task in biological research. These genes can be identified either by conducting experiments in lab (*in-vivo* approach) or by using some computational tool on already available genomic data (*in-silico* approach) or following hybrid approach i.e. designing and conducting lab experiments followed by computational analysis. Compare to *in-vivo* approach, *in-silico* approaches are time, cost and resource efficient [3].

As per current experimental trends in *in-vivo* approach, transcriptomics plays an important role in identifying genes related to a specific trait [4-6]. Microarrays and RNA-Seq are the two main technologies used to profile the expression level of genes. These technologies are widely used to detect the expression level of genes specific to a particular condition. A massive microarray datasets

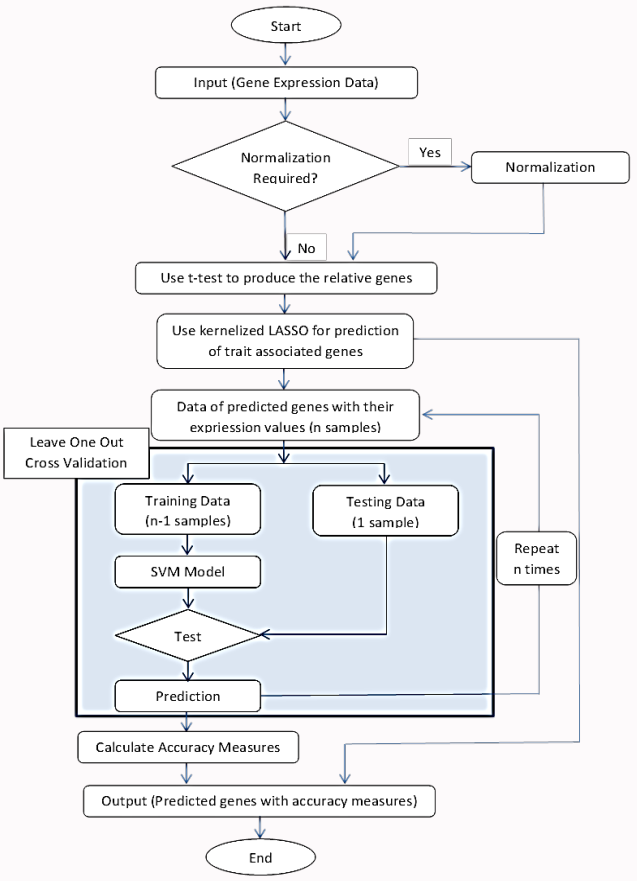

Figure 1: Schematic workflow of the proposed algorithm**.** Figure is showing algorithmic flowchart of proposed algorithm for prediction of trait specific genes using gene expression data.

generated through conducting trait related experiments are available in public repositories. These datasets can be used to detect trait specific genes. Therefore, for proper utilization of this huge dataset for *in-silico* identification of genes related to a specific trait, an efficient algorithm along with Web tool was required. Further, data from gene expression experiments i.e. both microarray and transcriptomics can also be analysed for identification of genes related to specific trait using this tool. This Web tool will not only helps biological scientists in accurately and efficiently identifications of genes related to specific trait but also provide easy and flexible Graphical User Interface (GUI) with high end advanced computational techniques applied to the data at the backend. The problem of identifying genes related to a particular trait can be framed as a feature selection problem, where, genes involved in microarray/transcriptomic data are considered as features and the selected key genes are indicative of a specific trait. Many feature selection techniques are proposed to select important genes responsible for a particular trait [7-22]. Among these techniques, penalized regression like Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net etc. are most popular, as these techniques deals with small sample size (n) and large number of features (p) i.e. n<p problem [23,24]. But being linear in nature, these techniques are unable to handle the problem of non-linear input-output relationships. Therefore, non-linear kernelized LASSO has been proposed to resolve this issue by using kernelized-penalized regression technique for feature selection and non-linear dependency of response with predictors [24]. By using this concept, an algorithm for gene classification has been proposed in this study. Furthermore, there is lack of readily available, user friendly tools based on this algorithm. Therefore, it is highly imperative to develop a tool based on this efficient algorithm for the purpose of prediction of trait specific genes to assist the biological scientists for accurately and efficiently identification of genes related to specific trait. This Web tool namely TAGPT (Trait Associated Genes Prediction Tool) has been developed for identification of trait specific genes using gene expression data i.e. microarray/ transcriptomic data. TAGPT employs an algorithm by integrating the kernelized LASSO to address two of the most important statistical issues *i.e.* (n<p) and non-linear dependency of predictors to the response. The efficacy of the proposed TAGPT has been assessed using Leave One Out Cross Validation (LOOCV) with Support Vector Machine (SVM) on publically available microarray/ transcriptomic data related to various traits.

Implementation of Algorithm

Gene expression data are having multiple sources of variation which affects measured gene expression levels. Normalization is a process which attempts to remove such variation and ensure correct data for subsequent analysis for identification of differentially expressed genes. The process of normalization of the data can also be applied to transcriptomic data set. After this data normalization data generated from both these technologies i.e. microarray experimentation and RNA-seq can be treated together. Therefore, the proposed algorithm begins with a normalization step followed by t-test for initial selection of differentially expressed genes [25,26]. Now, prediction of highly expressed genes has been done using a latest feature selection technique i.e. Kernelized LASSO. It was reported in the literature, that non-linear kernelized LASSO is able to reduce number of predictors less than the number of observations with appropriate degree of accuracy [24]. Classification power of the predicted highly expressed genes was assessed by classifying the labelled samples (control vs trait) into their respective classes. For this Support Vector Machine (SVM) has been used for the purpose of classification, as it is known to be best binary linear/non-linear classifier [27,28]. Classification accuracy has been assessed by using cross validation technique [29,30]. The schematic work flow of the proposed algorithm used in the TAGPT is given in the Figure1.

Data processing stage

Quantile based method of data normalization has been used in this algorithm. "Limma" package of Bioconductor (http://www.bioconductor.org/), an open-source tool for bioinformatics based on R statistical programming language, has been used for the same [25].

Preliminary selection of genes

Gene expression data contains much fewer data points as compare to the genes observed in experiments. Usually, tens of thousands of genes are observed with very less number of data points. Therefore, it would be of high complexity to use Kernelized LASSO directly for valid interpretation. Hence, a t-test with corrected p-values was employed to filter out non-significant differentially expressed genes for a specific trait and control condition [26].

Prediction of trait specific genes

The filtered out significant differentially expressed genes in the previous step were used for subsequent analysis. Here, non-linear penalized regression technique called Kernelized LASSO was used on the filtered data for final selection or identification of the trait specific responsive genes [24]. In order to select genes through Kernelized LASSO technique, the observed data on significant differentially

expressed genes is denoted by

$X = [x_1, \ldots, x_n] \in R^{d \times n}$ and different trait level of output data as $Y = [y_1, \ldots, y_n]^T \in R^n$,

where, n is the total number of observation and d is the number of significant differential expressed genes from the expression data obtained through t-test. Here, Y has taken as binary variable i.e. related to specific trait and control.

Now Kernelized LASSO is written as

$$\hat{\alpha} = \text{argmin} \frac{1}{2} \left\| \bar{R} - \sum_{k=1}^{d} \alpha_k \bar{Z}^{(k)} \right\|_2 + \delta \|\alpha\|_1$$

where, $\alpha = [\alpha_1, \ldots, \alpha_d]^T$ is a regression coefficient vector, α_k denotes the regression coefficient of the k-th feature and $\delta > 0$ is the regularization parameter.

$\bar{Z}^{(k)} = \Gamma Z^{(k)} \Gamma$ and $\bar{R} = \Gamma R \Gamma$ are centered kernel functions for input and response variable respectively. Here, $\Gamma = I_n - \frac{1}{n} 1_n 1_n^T$ is the centering matrix,

A universal reproducing kernel i.e. Gausian kernel is used for input **x** in the present algorithm

$$Z(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^2}{2\sigma_x^2}\right),$$

where, $\sigma_x = \text{median}\left(\{|x_i - x_j|\}_{i,j=1}^n\right)$

and delta kernel for response **y**,

$$R(\mathbf{y}, \mathbf{y}') = \begin{cases} \frac{1}{n_y} \text{ if } \mathbf{y} = \mathbf{y}' \\ 0 \text{ otherwise} \end{cases}$$

where, n_y is the number of samples in class **y.**

Support Vector Machine (SVM) and Cross validation

Predictive power of the **s**elected trait specific genes were assessed through SVM as it constructs a hyper plane which separates two groups of labelled samples with a maximum margin. Here, the selected trait specific genes were used as predictor variables and trait levels (trait and control) as response variable in training SVM [27]. CRAN package: e1071 version 1.6-2 of R software has been used for training SVM [28]. Classification accuracy of the trained SVM was further tested by using Leave One Out Cross Validation (LOOCV) Technique [29,30].

Software Development

The software for Trait Associated Gene Prediction Tool (TAGPT) has been developed and implemented on Web to provide easy and flexible GUI based tool to potential users. Programming has been done using HTML, R, JSP and Java programming languages [31]. User interface has been developed by using R Web User Interface (RWUI) too

along with Net Beans development environment as a platform for development of this tool [32,33]. The developed Web based tool is deployed on Apache tomcat server and can be accessed at www.cabgrid.res.in/tagpt [34,35].

Architecture and Design of the Software

The Web application has been developed using client-server architecture. A Web browser is the first tier (presentation), an engine using dynamic web content technology (JSP) is the middle tier (Web Application), and Shell, R language is the third tier (System Application). Architecture of the software is represented in Figure 2. The schematically design showing different modules developed for this tool has been shown in the Figure 3. It can be seen that TAGPT has four main modules i.e. data management, data analysis, report generation and help.

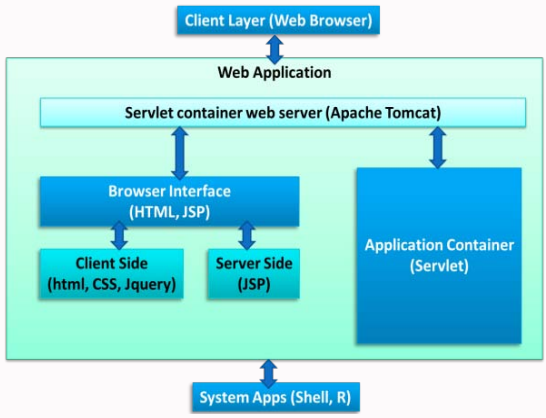

Figure 2: Architecture of the TAGPT. Figure represents architecture used in development of the Trait Associated Genes Prediction Tool (TAGPT).

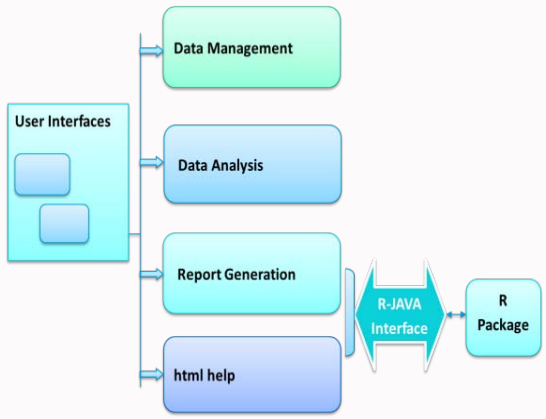

Figure 3: Design of the TAGPT. Figure represents the design of the TAGPT showing various modules i.e. user profile management, data analysis, report generation and help of the web server.

Software Features

This tool provides login facility to the users (Figure 4). It requires three types of input parameters i.e. gene expression matrix in .csv format, response variable or class type also in .csv format and p-value on the basis of which initial selection of the genes is to be done (Figure 4). Tool provides four types of output in tabular form i.e. first table is list of differentially expressed genes with all statistical values such as Log Fold Change (logFC) value, t-value, p-value, adjusted p-values, B-values etc., second table contains differentially expressed genes with their expression values, third table consists of predicted genes with their expression values and the fourth table is confusion matrix along with other evaluation parameters (Figure 4).

Results and Discussion

The performance of proposed algorithm implemented as TAGPT Web based tool for prediction of genes related to a specific trait was demonstrated using various microarray data. Five case studies has been considered for this purpose. Five microarray data sets of related to various traits were downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) with accession IDs GSE10670, GSE37940, GSE5185, GSE31885 and GSE32642. These dataset contain gene expression data related to drought stress, cold stress, ethanol tolerance, biosynthesis

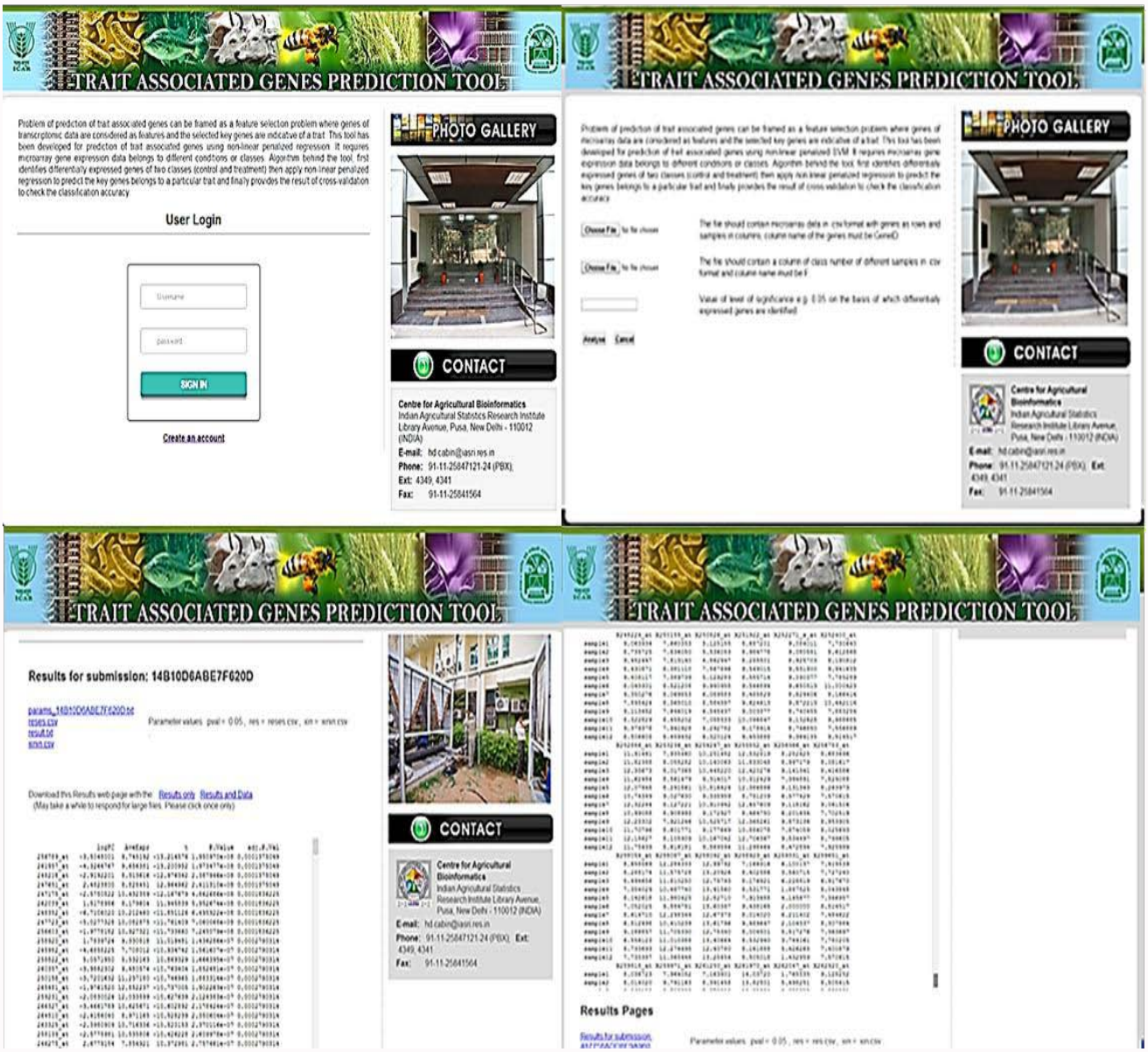

Figure 4: Features of TAGPT. Figure represents various pages (Login page, data input and analysis page, results page) of the web server.

of waxy substance and microbe associated molecular pattern (Table1). As per the algorithm, at the first step t-test was employed to filter out unlikely genes involved in controlling specific traits. In this preliminary filtration of unlikely trait associated genes through t-test, the p-value was assigned as 0.05, resulting in subset of 3615, 6749, 836, 1763 and 2882 filtered genes out of total 22810, 57194, 10928, 37478 and 66659 genes (Table 2) from different experiments considered above. Thereafter, Kernelized LASSO was applied to expression data of these filtered genes, which resulted in 27, 40, 15, 17 and 51genes respectively (Table 2). Hence, it can be seen than this algorithm reduced the number of genes related to a specific trait up to a great extent. Again, to test predictive power of these selected genes, SVM used as a classifier on expression data based on final selected genes to classify into two class i.e. trait vs control. Performance of the classification by SVM was tested using Leave One Out Cross Validation (LOOCV). Result of the analysis was presented in the form of classification accuracy, sensitivity and specificity (Table 2). Results indicate that finally predicted genes has high predictive power and able to classify the samples with high accuracy.

Conclusion

In this paper, a novel algorithm for prediction of genes related to a particular trait based on gene expressions data has been developed. Also, a user friendly Web tool for Trait Associated Genes Prediction (TAGPT) using proposed algorithm has been developed. Performance of the developed algorithm has been demonstrated on a real data set from microarray experiments. These case studies show that the algorithm behind TAGPT is accurate and fast while predicting the trait specific genes by analysing the gene expression data. This developed Web solution would be highly useful for biological scientists to use this highly sophisticated computational tool through easy and flexible GUI without going through its mathematical complexity and get the results quickly saving their precious time and energy.

Authors' Contribution

Overall study was planned by DCM and he worked in development

Table 1: Summary of the data used in analysis.

Accession No.	Organism	Data Type	Trait	No. of Samples		No. of Replicates	
				Treatment	Control	Biological	Technical
GSE10670	*Arabidopsis thaliana*	Microarray (Affymatrix)	Drought Stress	6	6	2	3
GSE37940	*Oryza sativa japonica*	Microarray (Affymatrix)	Cold Stress	15	3	5	3
GSE5185	*Sacharomyces cerevisiae*	Microarray (Affymatrix)	Ethanol Tolerance	6	6	2	3
GSE31885	*Arabidopsis thaliana*	Microarray (Affymatrix)	Biosynthesis of Waxy Substance	4	4	1	4
GSE32642	*Glycine max*	Microarray (Affymatrix)	Microbe Associated Molecular Pattern	12	12	4	3

Table 2: Summary of the predicted genes using TAGPT.

Accession No.	Total no. of genes in study	No. of genes in primary selection	No. of finally predicted genes	Evaluation measures of predicted genes		
				Sensitivity	Specificity	Accuracy
GSE10670	22810	3615	27	1	0.83	0.92
GSE37940	57194	6749	40	0.67	1	0.94
GSE5185	10928	836	15	1	1	1
GSE31885	37478	1763	17	1	1	1
GSE32642	66659	2882	51	1	1	1

of most of the modules of the software. SK helped in planning of the study and writing of codes. SBL and KKC designed and developed the software. AS worked in development of the user interface of the software. NB performed the computation. AR helped in planning of the study and provided statistical commentary. All the authors contributed towards the writing of the manuscript. All authors read and approved the manuscript.

Acknowledgement

This work was done under in-house project entitled 'Algorithm for Gene Classification based on Gene Expression Data' at Indian Agricultural Statistics Research Institute, New Delhi, India.

References

1. Acquaah G. Principles of plant genetics and breeding. Blackwell, Oxford, UK. 2007.
2. Wang W, Vinocur B, Altman A. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta. 2003;218(1):1-14.
3. Zhu Mengjin, Zhao Shuhong. Candidate Gene Identification Approach: Progress and Challenges. Int J Biol Sci. 2007;3(7):420-7.
4. Chana Yagil, Norbert Hubner, Jan Monti, Herbert Schulz, Marina Sapojnikov, Friedrich CL, et al. Identification of Hypertension-Related Genes Through an Integrated Genomic-Transcriptomic Approach. Circ Res. 2005;96(6):617-25.
5. Wan Y, Underwood C, Toole G, Skeggs P, Zhu T, Leverington M, et al. A novel transcriptomic approach to identify candidate genes for grain quality traits in wheat. Plant Biotechnol J. 2009;7(5):401-10.
6. Sebastián Aguilar Pierlé, Michael J Dark, Dani Dahmen, Guy H Palmer, Kelly A Brayton. Comparative genomics and transcriptomics of trait-gene association. BMC Genomics. 2012;13:669.
7. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics. 2001;17(6):509-19.
8. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 2005;3(2):185-205.
9. Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D. Prediction of Drought-Resistant Genes in Arabidopsis thaliana Using SVM-RFE. PLoS ONE. 2011;6(7):e21750.
10. Dash S, Patra B. Study of Classification Accuracy of Microarray Data for Cancer Classification using Hybrid, Wrapper and Filter Feature Selection Method. BIOCOMP. 2012.
11. Ding Y, Wilkins D. Improving the performance of SVM-RFE to select genes in microarray data. BMC Bioinformatics. 2006;7:S12.
12. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46(1-3):389-422.
13. Huang DQ, Wu WR, Abrams SR, Cutler AJ. The relationship of drought related gene expression in Arabidopsis thaliana to hormonal and environmental factors. J Exp Bot. 2008;59(11):2991-3007.
14. Jornsten R, Yu B. 'Simultaneous gene clustering and subset selection for sample classification via MDL.' Bioinformatics. 2003;19(9):1100-9.
15. Kankainen M, Brader G, To¨ro¨nen P, Palva ET, Holm L. Identifying functional gene sets from hierarchically clustered expression data: map of abiotic stress regulated genes in Arabidopsis thaliana. Nucleic Acids Res. 2006;34(18):e124.
16. Li J, Gao Z. Random forests: an important feature genes selection method of tumor. Acta Biophys Sin. 2009;25:51-6.
17. Sumathi A, Santhosh Kumar S, Sakthivel NK. Development of an Efficient Data Mining Classifier with Microarray Data Set for Gene Selection and Classification. Journal of Theoretical and Applied Information Technology. 2012;35(2):208-14.
18. Van't Veer L, Dai H, van de Vijver M, He Y, Hart A, Mao M, et al. 'Gene expression profiling predicts clinical outcomes of breast cancer'. Nature. 2002;415(6871):530-6.
19. Wang JX, Zhang F, Wang Y, Fu Y, Xu D, et al. Salt tolerance genes selection in Oryza Sativa using SVMRFE based on Microarray. 3rd International Conference on Bioinformatics and Computational Biology (BICoB). New Orleans. 2011;30-5.
20. Yousef M, Ketany M, Manevitz L, Showe LC, Showe MK. Classification and biomarker identification using gene network modules and support vector machines. BMC Bioinformatics. 2009;10:337.
21. Zhang W, Ruan J, Ho TH, You Y, Yu T, Quatrano RS. Cis-regulatory element based targeted gene finding: genome-wide identification of

abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana. Bioinformatics. 2005;21(14):3074-81.

22. Zhou X, Tuck DP. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Bioinformatics. 2007;23(9):1106-14.

23. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B. 1996;58(1):267-88.

24. Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M. High-dimensional feature selection by feature-wise kernelized lasso. Neural Comput. 2014;26(1):185-207.

25. Smyth GK. Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Edited by: Gentleman R, Carey V, Dudoit S, R Irizarry WH. New York: Springer. 2005;397-420.

26. Y Benjamini, Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B (Methodological). 1995;57(1):289-300.

27. Vapnik, Vladimir N. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc. 1995.

28. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-24. 2010.

29. Verweij PJ, Van Houwelingen HC. 'Cross-validation in survival analysis'. Stat Med. 1993;12(24):2305-14.

30. Arlot Sylvain, Celisse Alain. A survey of cross validation procedures for model selection. Statistical Surveys. 2010;4:40-79.

31. Guay R. Protect web application control flow, 2003.

32. Newton R, Wernisch L, Rwui A. Web Application to Create User Friendly Web Interfaces for R Scripts, R News. 2007;7(2):32-5.

33. Böck, Heiko. The Definitive Guide to Net Beans Platform 7 (First ed.). Apress. 2011;592. ISBN 978-1-4302-4101-0.

34. Apache Struts. http://struts.apache.org/

35. Apache Tomcat. http://tomcat.apache.org/