Integrative data mining in functional genomics of Brassica napus and Arabidopsis thaliana

Pan, Youlian; Tchagang, Alain; Bérubé, Hugo; Phan, Sieu; Shearer, Heather; Liu, Ziying; Fobert, Pierre; Famili, Fazel

IEA/AIE'10 Proceedings of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, 3, pp. 92-101, 2010-06-01

Integrative data mining in functional genomics of *Brassica napus* and *Arabidopsis thaliana*

Youlian Pan[1], Alain Tchagang[1], Hugo Bérubé[1], Sieu Phan[1], Heather Shearer[2], Ziying Liu[1], Pierre Fobert[2], Fazel Famili[1]

[1] Knowledge Discovery Group, Institute for Information Technology, NRC, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada; [2] Seed Systems, Plant Biotechnology Institute, NRC, 110 Gymnasium Place, Saskatoon, SK S7N 0W9
{youlian.pan, alain.tchagang, hugo.berube, sieu.phan, heather.shearer, ziying.liu, pierre.fobert, fazel.famili}@nrc-cnrc.gc.ca

Abstract. Vast amount of data in various forms have been accumulated through many years of functional genomic research throughout the world. It is a challenge to discover and disseminate knowledge hidden in these data. Many computational methods have been developed to solve this problem. Taking analysis of the microarray data as an example, we spent the past decade developing various data mining strategies and software tools. It appears still insufficient to cover all sources of data. In this paper, we summarize our experiences in mining microarray data by using two plant species, *Brassica napus* and *Arabidopsis thaliana*, as examples. We present several success stories and also a few lessons learnt. The domain problems that we dealt with were the transcriptional regulation in seed development and during defense responses against pathogen infection.

Keywords: Integrative data mining, microarray, transcription regulation, seed development, plant defense.

1 Introduction

Knowledge discovery from various sources, such as biological experiments and clinical or field trial information, is a complex and challenging task. This requires an in-depth understanding of the domain and development of appropriate strategies for data preprocessing and subsequent analysis. High throughput determination of gene expression profiles has been prevalent in the past decades, particularly with the advent of microarray technology. This has motivated researchers to utilize tools, techniques, and algorithms developed through many years of data mining and knowledge discovery research, to search for useful patterns in the gene expression data. This is exemplified by the abundance of computerized data analysis tools that have become available to perform clustering, pattern recognition, and motif identification in gene's promoters. One of the greatest challenges is to understand how the expression pattern of thousands of genes in a living organism is regulated and related to one another. Two examples of these are: (*i*) the discovery of relationships between genes and their

expression profiles over a time-series, such as genes' progressive responses to drug treatment over time or stages during embryonic development, and (*ii*) genes' responses at one discrete time point to various treatments, to knock-out or knock-down of certain transcription factors.

Generally, no single data analysis method is able to be successfully applied to all different datasets. Often, data mining researchers have to select methods or develop a new algorithm based on a particular dataset. Microarray gene expression data is subject to multiple sources of noise [1]. To cope with such instability in the data, many normalization techniques have been developed, but these techniques can only ease rather than solve the problems completely. As a consequence, the confidence in knowledge derived from the data by a single analysis tool is dependent on the extent of noise and bias. One of the important questions in data mining is how to understand the scope and minimize the impact of such noise and bias within the data.

Our research team is currently working on knowledge discovery from plant genomes, specifically *Brassica napus*, the canola oil producer, and *Arabidopsis thaliana*, a small model plant, with regard to seed development and defense mechanism against pathogen infection, respectively. The data were produced by using various microarray platforms. Thus the data have various degrees of complexity. We have used several integrative approaches to mine these data. This paper is to present some successful stories and lessons learnt from our data mining investigations.

Both *Brassica napus* and *Arabidopsis thaliana* belong to the Brassicaceae family. *Brassica napus* (rapeseed) currently contributes over $11B in economic activity with the canola industry being responsible for over 214,000 jobs in Canada. Canola oil has high content of healthy fatty acids, such as oleic acid, linoleic acid, and α-linolenic acid [2], and contains only a trace amount of erucic acid, which may adversely affect heart tissue [3]. Therefore, canola oil is prized as healthy oil by consumers. Our research problem is to identify genes with the potential to improve key aspects of *Brassica* oilseed and canola productivity by increasing total oil production, seed yield and seedling vigour. Our role in this research is to help biologists at Plant Biotechnology Institute, NRC to identify these genes and their behaviors under various conditions, at various seed developmental stages and in different tissues.

Our research problem in *A. thaliana* is to identity genes responsible to pathogen infection and alteration of their expression profiles during systemic acquired resistance (SAR). Our role is to discover the transcriptional regulatory relationship between these genes and their upstream regulators through knock-out of certain transcription factors that are key regulators with regard to plant's SAR.

In the following sections, we present the two biological problems as examples of our current research. We first describe the problems and our solutions. Then, we present result highlighting the benefit of integrative approaches. This is followed by a discussion and a conclusion.

2 The biological problems and solutions

2.1 Endosperm of *Brassica napus*

This problem was to identify highly expressed genes and understand the mechanism behind the changes of gene expression in the endosperm during embryogenesis of *B. napus* seeds. The stages of embryogenesis considered in this study were defined according to the shapes of the imbedded embryos: globular, heart, and cotyledon. The microarray experiment was done in dual channel array representing two different developmental stages, i.e. heart vs. globular, cotyledon vs. globular, or cotyledon vs. heart. The experiment was performed with two biological replicates; each had four technical repeats with dye swaps. Paralleled with the microarray data, there were also EST (expressed sequence tag) data. Details are available in [4].

Our approach was first to identify a group of significantly and differentially expressed genes. In this step, it is critical to refer to the domain questions so that not to exclude genes which are necessary in answering the questions and also not to introduce much noise in subsequent data analysis steps. There are two aspects in the domain question: *1*) to find significantly expressed genes *a*) in some stages but not necessarily in other stages, and *b*) across all stages; *2*) to group differentially expressed genes based on their patterns of variations [5]. In this study, our main point was to highlight the importance of considering both ratio data and intensity data from each channel at the same time.

2.2 Defense response in *Arabidopsis thaliana*

This problem was to identify the effect of key transcription regulators in plant defense responses against pathogen infection, using data generated by microarray. The microarray analysis addressed two key variables: the effect of salicylic acid (SA), a key elicitor of pathogen-induced SAR in plants, and the effect of mutating the NPR1(Non-expresser of Pathogenesis Related gene 1) gene and TGA family genes. The establishment of SAR, an inducible defense response that leads to broad spectrum of systemic resistance, requires an endogenous increase in SA levels [6]. However, the exogenous application of low concentrations of SA, as used in this study, can also trigger a SAR response. In *Arabidopsis*, the NPR1 gene is essential for SA-mediated SAR [7]. Currently there is no evidence to suggest that NPR1 binds DNA directly to regulate transcription, but indirectly regulates the expression of genes involved in SAR through its interaction with TGA family of bZIP transcription factors [8-12]. Seven (TGA1-TGA7) of the ten TGA factors in *Arabidopsis* have been characterized to interact with NPR1 [13-15]. These seven TGAs can be divided into three groups based on sequence homology [16]. Group I consists of TGA1 and TGA4; Group II TGA2, TGA5 and TGA6; and group III TGA3 and TGA7. In this research, we used four genotypes: Columbia wild type plant and three sets of mutants (*npr1*, knock off of group I TGA factors, *tga1 tga4*, and knock off of group II TGA factors, *tga2 tga5*

tga6). Small amount of SA was sprayed to each plant to mimic pathogen infection that induced a series changes in expression of genes involved in SAR. Samples were taken 0, 1, and 8 hours after the plants subjected to SA application.

This research was performed in two phases. The first phase consists of five biological experiments and SA was applied on two genotypes, wild type and the mutant *npr1*. We used an approach that consists of several iterations of integration of three components: (*i*) clustering (unsupervised learning), (*ii*) pattern recognition (supervised learning), and (*iii*) identification of transcription factor binding sites. Briefly, a group of informative genes were identified from the entire dataset through pattern recognition and compared to interesting clusters generated by K-Means. Interesting motifs in the upstream promoter region were identified for each gene and compared with other genes in the same cluster. A combination of results of informative genes, gene expression profiles and motif information constituted a representative gene for each interesting cluster. These representative genes were used as seeds for subsequent re-clustering of the data through K-Means to determine more refined clusters [17].

In the second phase, SA was applied to all four genotypes. The knowledge discovery was done by using integration of an expanded version of frequent itemset mining approach [18] and an order preserving three-dimensional-clustering approach [19]. The order preserving clustering approach is a combination of order preserving pattern feature [20] with clustering. Before applying the expanded version of frequent itemset mining algorithm, the gene expression matrix was first discretized into three distinct values (-1, 0, 1) representing down-regulation, no significant difference, and up-regulation, respectively, based on a predefined threshold value, and relative to a baseline, which is the wild type in our study. All the interesting associations and association rules between the transcription factors and their target genes were then identified [18]. A wiring diagram was inferred to describe gene regulatory network during pathogen-induced SAR in *A. thaliana* specifying sets of genes that were differentially expressed as a result of one, two or all three mutant sets (*npr1*, *tga1 tga4*, and *tga2 tga5 tga6*) (Fig. 1).

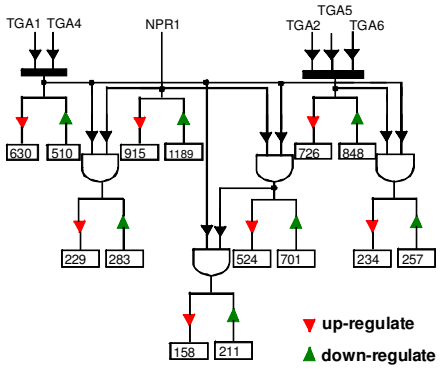

Fig. 1. The wiring diagram of 8th hour after SA treatment on *Arabidopsis thaliana*.

The order preserving 3D clustering approach discretizes a gene expression profile by ranking expression value at all time points based on a predefined threshold value disregarding up or down regulation [19]. The number of discretization values is $\leq T$ depending on the threshold and the variability of the expression profile; where T is the number of time points. This approach identifies similarities and differences in terms of gene expression profiles between the wild type and the mutant sets. In other words, it identifies groups of genes that have the same sequential variation patterns unique in one genotype plants or others and may be the same between two, three or across all four genotypes [19].

3 Results

3.1 Endosperm of *Brassica napus*

The detailed result of this research has been recently published in [4, 5]. Here we highlight relationship of two key transcription factors that might shine some light in cascading of transcription regulation of seed development and fatty acid metabolism. Leafy cotyledon1 (LEC1) is a key regulator of fatty acid biosynthesis in *Arabidopsis*. In the LEC1-overexpressing transgenic plants, over 58% of known enzyme-coding genes involved in the plastidial fatty acid synthetic pathway are up-regulated; levels of major fatty acid species (e.g. oleic acid, linoleic acid, and α-linolenic acid) and lipids were substantially increased [21]. The function of LEC1 is partially dependent on WRINKLED1 (WRI1) and other two transcription factors (ABI3 and FUS3) in the regulation of fatty acid biosynthesis. Over-expression of WRI1 up-regulates a set of genes involved in fatty acid (FA) synthesis in plastids [22]. In our work, the expression profiles of LEC1 and WRI1 are identical based solely on the log ratio values (Fig. 2A), which lead us to a conclusion that the LEC1 and WRI1 are co-regulating the FA metabolism in *B napus*. While looking into the expression intensity of the two TFs, we found that they were about one order of magnitude different from

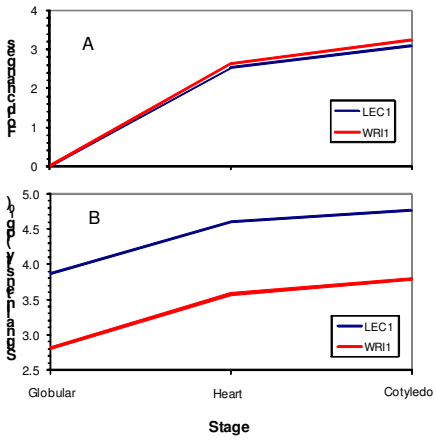

Fig. 2. Gene expression of the two transcription factors, LEC1 and WRI1.

each other (Fig. 2B). This result was consistent with the number of ESTs found for each TF [5]. When the expression of LEC1 was moderate at the globular stage, the expression of WRI1 was very low. When expression of LEC1 became more significant at later stages, expression of WRI1 increased significantly. This observation allows us to deduce that high expression of LEC1 probably enhances the expression of WRI1 in *B napus*, whether directly or indirectly through another transcription factor. A recent study revealed that LEC2 directly regulated WRI1 [23]. However, over-expression of LEC1 does not directly affect the level of LEC2 [21]. Therefore, we can conclude that both LEC1 and LEC2 are the upstream regulator to WRI1 (Fig. 3). This result is similar to what is found in *Arabidopsis* [21]. However, this cascading relationship between the two transcription factors would not be possibly revealed without considering the signal intensity data.

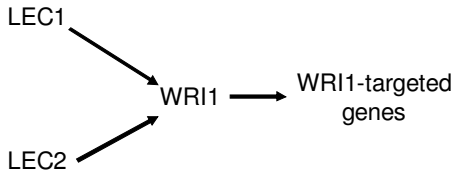

Fig. 3. Schematic transcriptional relationship between transcription factors

3.2 Defense response in *Arabidopsis thaliana*

The details of phase one work have been published in [17]. Here we provide a highlight of this phase. Through several iterations of clustering, we were able to identify and confirm 24 genes that were differentially expressed, 12 up-regulated and 12 down-regulated, in mutant *npr1* as compared to the wild type following SA treatment. Using the pattern recognition approach, we were able to identify 15 highly informative genes, the majority (8) of which were in the down-regulated cluster described above and highly enriched with ASF-1 motif (TGACG [24]) and W box (TTGAC [25]) in their promoters. The TGA factors binds to ASF-1 motif and WRKY transcription factors, which are dependent on NPR1 [26], bind to W box. This is consistent with the fact that NPR1 indirectly regulates the target genes of both TGA factors and WRKY factors.

In the second phase, through the frequent itemset mining approach [18], we were able to identify genes that were regulated by one set of transcription factors alone, and those collectively regulated by two or all three sets of transcription factors. For example, the wiring diagram in Fig. 1 shows that 8 hours after application of SA, 158 and 211 genes are up and down regulated, respectively, by the combined function of NPR1 and all five TGA factors in this study. Similarly, 76 (=234-158) and 46 (=257-211) genes are exclusively (excluding the effect of NPR1) up and down regulated, respectively, by the combined function of all five TGA factors. But 320 (=915-524-229+158) and 416 (=1189-701-283+211) genes are uniquely (excluding the effect of TGA factors) up and down regulated by NPR1, respectively.

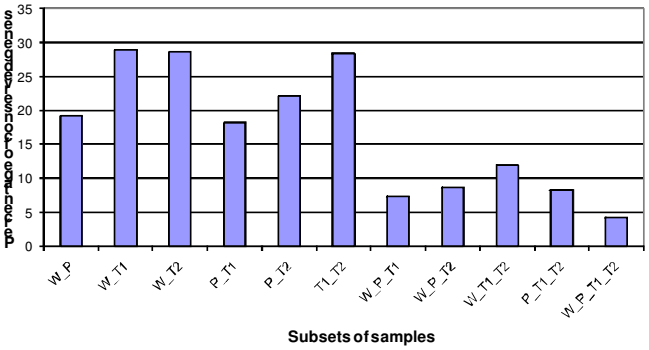

Fig. 4. Summary result of the order preserving clustering approach. W = wild type, P = *npr1*, T1 = *tga1*×*tga4,* T2 = *tga2*×*tga5*×*tga6.* W_P_T1_T2 = genes whose expression profile conserved across all genotypes, i.e. independent of any mutations.

Through the order preserving clustering approach [19], we were able to identify genes that are not affected by one or more mutations (Fig. 4) across the time series. For example, 4.32% genes are independent of any genotypes in this study (W_P_T1_T2); 12.02% genes are independent of mutation of TGA factors (W_T1_T2). From the difference between these two numbers, we were able to derive 7.70% are affected by mutant *npr1*, i.e. regulated by NPR1.

After integration of these two sets of analyses, we will be able to pin-point certain group of genes that are regulated by one transcription factor or co-regulated by more transcription factors at one specific time point, or across the entire time series, therefore infer the dynamic regulatory behavior during SAR in *A. thaliana*. More post processing of this phase of work is in progress. Additionally, we are applying the same integrative approach to seed developments and its association with fatty acid metabolism in *B. napus*. Preliminary results indicate this approach is promising and has broad application.

4 Result integration – knowledge base

Through many years of experimentation, vast amount of data have been accumulated and currently available at Plant Biotechnology Institute, NRC. Currently, we are developing a knowledge base named BRISKA (*Brassica* Seed Knowledge Application) [27] that integrates knowledge discovered through our many years of data mining processes with publically available knowledge in literature and public databases, such as GO [28], KEGG [29], TRASFAC [30], etc. The objective of this knowledge base is to support subsequent integrated reasoning in new knowledge discovery processes. Our ultimate goal is to build a robust virtual seed system through incremental learning.

Currently, a prototype of the knowledge base has been developed. It contains various tools and results of analysis from both public and private sources. BRISKA

integrates microarray data alongside sequence-based data, such as ESTs and promoter sequences, and provides data analysis results generated by using various tools. For example, sequence similarity has been done on all sequence-based data to allow linkage from EST, to contigs and genes while also proposing possible orthologs for any gene of interest in related species. The schema used in BRISKA is based on the Chado [31] model and, through its ontology-driven design, supports complex representation of biological knowledge such as clusters of co-expressed genes, gene regulatory network modules, and expression plots. Public microarray experiment data are acquired from Gene Expression Omnibus (GEO) and has been selected by the relevance to our actual research interests. Genes, ESTs, and contigs are from the TAIR database [32] while binding site and transcription factor information are mostly from PlantTFDB [33]. Annotation information such as gene ontology, KEGG identifiers has also been added. Private data are mostly acquired from the Plant Biotechnology Institute and consist mainly of EST and microarray expression data.

An interactive web-based interface along with visualization tools have been developed to provide intuitive access to the knowledge base. An analysis explorer tool grants user access to their data, analyses results along with the protocol used to generate the results. For example, gene expression analysis results are provided in an interactive spreadsheet; gene networks analyses can be visualized via BRISKA's viewer, which were built by our team as an extension of the Guess [34] application to provide multiple functionalities such as network search, connection-depth search, and network manipulation.

5 Discussion and conclusion

In this paper, we described several integrative approaches applied in mining microarray gene expression data of *B. napus* and *A. thaliana*. These methods have been incorporated in various software tools developed in-house or obtained from publically available resources. It is important first to investigate the data structure, their associated features and attributes, the noise content and the associated strategy to minimize its impact, and the domain problems to be addressed. Once this basic information is known, one can then look into the tools needed to address these sets of problems.

In the *B. napus* endosperm work, we have investigated both the signal intensity and the differential expression between different stages of embryogenesis. The conventional approach of analyzing the dual channel microarray data by looking into only the ratio data between the channels would miss the cascading relationship between LEC1 and WRI1. This research also alerts us that it is important to identify the characteristics of data. For example, the ratio data in the dual array experiment is derived from two signal intensities. In this case, the signal intensities are the primary data, while the ratio is derived data. Two ratios of the same value do not imply that the primary data, from which they are derived, are the same or similar. There could be a big difference between two sets of primary data (Fig. 2). Yet, some ratio data in a specific region of their parental primary data could be misleading [5]. This fact has been neglected in many of the dual channel array data analysis performed earlier.

In the first phase of *Arabidopsis* work, we integrated the unsupervised and supervised learning approaches with sequence motif search and identified a group of genes that were closely related with the transcription factor NPR1. Microarray data usually contains much noise [1]. A cross checking of results by various methods are necessary to ensure quality of the results. Our final results from the three sets of analyses support each other and increase our confidence. In the second phase of this work, we used the frequent itemset mining approach to identify the effect of mutations on gene expression at a given time point; we used the order preserving clustering approach to mine the sequential ranking pattern in the expression profile. The former identifies the static nature and the later illustrates the functionally dynamics nature of the gene expression data matrix.

Each step of knowledge discovery enriches our knowledge base, as we see in our *Brassica* Seed Knowledge Base that we are developing. The discovered knowledge will incrementally enhance the subsequent knowledge discovery process. For such reason, many bioinformatics knowledge bases, such as gene ontology [28], KEGG pathway [29], etc. became available. Our objective is to structure all forms of discovered and validated knowledge in order to provide means to augment our capability in subsequent knowledge discovery.

Through worldwide genomic research effort of the past decade, large amount of data in various forms have been accumulated. It is a challenge to integrate information in many different forms (e.g. ESTs, SNPs, small RNAs, microarray, protein-protein interactions and metabolomics data) and from various platforms (for example the microarray data could be from Affymetrix, Agilent chips, or in house chips) and to extract knowledge from this vast information pool. Our group has developed various data mining strategies, algorithms and tools based on our expertise in data mining and machine learning in order to effectively discover new knowledge from this vast information pool. This knowledge will be validated through literature search, local and public knowledge bases search, and follow-up wet lab experimentation. Finally, the knowledge can be presented as gene-gene associations, gene-metabolites associations, metabolic pathways, gene networks and various forms of predictive or descriptive models. These forms of knowledge presentation will be facilitated and interconnected through our BRISKA knowledge base. Our ultimate goal is to provide biologists an integrative and interactive environment to visualize seed development and fatty acid metabolism of *B. napus* and related species and to further conduct experiments in perturbing/modifying genetic parameters in a virtual world to improve canola oil production, seed yield and seedling vigour.

Acknowledgments. The data for *Brassica* endosperm work were from Jitao Zou's lab at Plant Biotechnology Institute, NRC. This work is co-funded by Genomics and Health Initiative, Plant Biotechnology Institute, and Institute for Information Technology, National Research Council Canada. This is publication NRCXXXXX of the National Research Council.

References

1. Churchill, G.A.: Fundamentals of experimental design for cDNA microarrays. Nat. Genet. 32, 490--495 (2002)
2. Canola Council of Canada: Comparison of Dietary Fats Chart. Retrieved on Nov 25, (2009)
3. Food Standards Australia New Zealand: Erucic acid in food: a toxicological review and risk assessment. Technical report series No. 21; ISBN 0 642 34526 0, ISSN 1448-3017 (2003)
4. Huang, Y., Chen, L., Wang, L., Phan, S., Liu, Z., Vijayan, K., Wan, L., Ross, A., Datla, R., Pan, Y., Zou, J.: Probing endosperm gene expression landscape in Brassica napus. BMC Genomics 10, 256 (2009)
5. Pan, Y., Zou, J., Huang, Y., Liu, Z., Phan, S., Famili, F.A.: Goal driven analysis of cDNA microarray data. In: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2009), pp.186-192, IEEE Press, New York (2009)
6. Ryals, J.A., Neuenschwander, U.H., Willits, M.G., Molina, A., Steiner H.Y., Hunt, M.D.: Systemic acquired resistance. Plant Cell 8, 1809-1819 (1996)
7. Delaney, T.P., Friedrich, L., Ryals, J.A.: *Arabidopsis* signal transduction mutant defective in chemically and biologically induced disease resistance. Proc. Natl. Acad .Sci. 92, 6602-6606 (1995)
8. Zhang, Y., Fan, W., Kinkema, M., Li, X., Dong, X.: Interaction of NPR1 with basic leucine zipper protein transcription factors that bind sequences required for salicylic acid induction of the PR-1 gene. Proc. Natl. Acad. Sci. 96, 6523-6528 (1999)
9. Després, C., DeLong, C., Glaze, S., Liu, E., Fobert, P.R.: The *Arabidopsis* NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. Plant Cell 12, 279-290 (2000)
10. Kinkema, K., Fan, W., Dong, X.: Nuclear localization of NPR1 is required for activation of PR gene expression. Plant Cell 12, 2339-2350 (2000)
11. Subramaniam, R., Desveaux, D., Spickler, C., Michnick, S.W., Brisson, N.: Direct visualization of protein interactions in plant cells. Nat. Biotech. 19, 769-772 (2001)
12. Johnson, C., Boden, E., Arias, J.: Salicylic acid and NPR1 induce the recruitment of trans-activating TGA factors to a defense gene promoter in *Arabidopsis*. Plant Cell 15, 1846-1858 (2003)
13. Kesarwani, M., Yoo, J., Dong, X,: Genetic Interactions of TGA transcription factors in the regulation of pathogenesis- related genes and disease resistance in *Arabidopsis thaliana*, Plant Physiol. 44, 336-346 (2007)
14. Jakoby, M., Weisshaar, B., Droge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., Parcy, F.: bZIP transcription factors in *Arabidopsis*. Trends Plant Sci. 7, 106-111 (2002)
15. Dong, X.: NPR1, all things considered. Curr. Opin. Plant Biol. 7, 547-552 (2004)
16. Xiang, C., Miao, Z., Lam, E.: DNA-binding properties, genomic organization and expression pattern of TGA6, a new member of the TGA family of bZIP transcription factors in *Arabidopsis thaliana*. Plant Mol. Biol. 34, 403-415 (1997)
17. Pan, Y., Pylatuik, J.D., Ouyang, J., Famili, A., Fobert, P.R.: Discovery of functional genes for systemic acquired resistance in *Arabidopsis thaliana* through integrated data mining. J. Bioinfo. Comput. Biol. 2, 639-655 (2004)
18. Tchagang, A.B., Shearer, H., Phan, S., Bérubé, H., Famili, F.A., Fobert, P., and Pan, Y.: Towards a temporal modeling of the genetic network controlling systemic acquired resistance in *Arabidopsis thaliana*. *Under review*: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2010)
19. Tchagang, A.B., Phan, S., Famili, F.A., and Pan, Y.: OPTricluster: The Order Preserving Triclustering Algorithm. Technical Report, Knowledge Discovery Group, Institute for Information Technology, National Research Council Canada, 2008.

20. Phan, S., Famili, F., Tang, Z., Pan, Y., Liu, Z., Ouyang, J., Lenferink, A., O'Connor, M.: A novel pattern based clustering methodology for time-series microarray data. Intern. J. Comput. Math. 84, 585-597 (2007)
21. Mu, J., Tan, H., Zheng, Q., Fu, F., Liang, Y., Zhang, J., Yang, X., Wang, T., Chong, K., Wang, X.-J., Zuo, J.: LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in *Arabidopsis*. Plant Physiology 148, 1042-1054, (2008)
22. Maeo, K., Tokuda, T., Ayame, A., Mitsui, N., Kawai, T., Tsukagoshi, H., Ishiguro, S., Nakamura, K.: An AP2-type transcription factor, WRINKLED1, of *Arabidopsis thaliana* binds to the AW-box sequence conserved among proximal upstream regions of genes involved in fatty acid synthesis. Plant J. 60, 476-487 (2009)
23. Baud, S., Mendoza, M.S., To, A., Harscoet, E., Lepiniec, L., Dubreucq, B.: WRINKLED1 specifies the regulatory action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in *Arabidopsis*. Plant J. 50, 825–838 (2007)
24. Lebel, E., Heifetz, P., Thorne, L., Uknes, S., Ryals, J., Ward, E.: Functional analysis of regulatory sequences controlling PR-1 gene expression in *Arabidopsis*. Plant J. 16, 223-233, (1998)
25. Eulgem, T., Rushton, P.J., Robatzek, S., Somssich, I.E.: The WRKY super-family of plant transcription factors. Trends Plant Sci. 5, 199-205 (2000)
26. Yu, D., Chen, C., Chen, Z.: Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression. Plant Cell 13, 1527-1539 (2001)
27. Bérubé, H., Tchagang, A., Wang, Y., Liu, Z., Phan, S., Famili, F., Pan Y.: BRISKA: brassica seed knowledge application. Poster at 17th International Conference on Intelligent Systems in Molecular Biology, Stockholm, (2009).
28. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25-29 (2000)
29. Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., Kanehisa, M.: KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. 36, W423–W426 (2008)
30. Matys, V., Kel-Margoulis, O.V., Fricke, E., et al.: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34, D108-110 (2006)
31. Mungall, C.J., Emmert, D.B., The FlyBase Consortium: A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 23, i337-i346(2007)
32. The Arabidopsis Information Resource (TAIR). [http://www.arabidopsis.org/]
33. Guo, A.Y., Chen, X., Gao, G., Zhang, H., Zhu, Q.H., Liu, X.C., Zhong, Y.F., Gu, X., He, K., Luo, J.: PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res. 36, D966-D969 (2008)
34. Adar, E.: GUESS: a language and interface for graph exploration. In: CHI 2006, ACM (2006)