Joint analyses of transcriptomic and metabolomic data to probe ryegrass-endophyte symbiosis

M. CAO, L. JOHNSON, R. JOHNSON, A. KOULMAN, G.A. LANE and S. RASMUSSEN
AgResearch Grasslands, Private Bag 11008, Palmerston North, New Zealand
Mingshu.Cao@agresearch.co.nz

Abstract

Fungal endophytes (*Neotyphodium lolii*) in perennial ryegrass (*Lolium perenne*) produce a range of bioactive alkaloids which are implicated in both toxicity to grazing animals and resistance to insects. The understanding of regulatory and biochemical mechanisms of the symbiosis will provide clues for the genetic manipulation of beneficial alkaloid production. This paper presents approaches to analyse data from high-throughput microarray experiments and targeted metabolomic analyses. Combined with bioinformatics analyses, potential genes were found associated with the accumulation of alkaloids and other metabolites. The advantages and limitations of our approach to address the molecular mechanisms of the symbiosis will be discussed.

Keywords: *Lolium perenne*, *Neotyphodium lolii*, metabolomics, microarray

Introduction

Fungal endophytes (*Neotyphodium lolii*) in perennial ryegrass (*Lolium perenne*) produce a range of bioactive alkaloids which are implicated in toxicity to grazing animals but also in resistance to insects. Better understanding of the regulatory and biochemical mechanisms of the symbiosis will provide clues for the genetic manipulation of beneficial alkaloid production. High-throughput technologies in functional genomics can provide comprehensive information on a biological system. However, the integration of data from heterogeneous sources poses challenges for the effective formation of hypotheses because of the complexity of data collected from the high-throughput technologies, such as microarray and mass spectrometry etc. We report in this paper data reduction and integration methods to gain insights into the complex biological system of the symbiosis of ryegrass and its fungal endophyte from the perspective of changes in gene expression and metabolite concentration.

Methods

Samples

Twenty-four perennial ryegrass samples were examined in this study comprising three tissue types (immature leaves, blades and mature leaves or sheaths) of both endophyte LP19 (*N. lolii*) infected (E+) and endophyte free (E-) isogenic ryegrass lines. Four replicates were used for each tissue type of E+ and E-.

The source of microarray and metabolomics data

In brief, about 15 000 ESTs (expressed sequence tags) were generated from six suppressive subtractive hybridisation (SSH) libraries and other sources (see Johnson *et al*. 2006 for details of microarray data generation). Targeted metabolite analyses of the samples were conducted by a range of chemical analyses such as GCMS, LC-PDA and LC-fluorescence etc. In total, 70 targeted measurements of sugars, amino acids and alkaloids were included in the analysis. The large scale metabolomic screening on the same set of samples based on direct infusion MSMS will be reported elsewhere.

Data processing

A balanced incomplete block design including dye-swaps was used for the original microarray experiments to make direct treatment comparisons robust. For the purpose of data integration with metabolite data, the original microarray data ($\log_2$ ratio) were converted to $\log_2$ intensity values for each treatment. This data transformation generates the intensity of gene expression of each treatment by comparison with the overall mean effect (David Baird, personal communication). The same approach was also applied to the metabolite data. For the comparison with expression values of ESTs, the concentration of each estimated metabolite was normalised against the average of the observations in all the samples, using $\log_2$ ($\mathbf{x}$(i) / mean ($\mathbf{x}$)), where vector $\mathbf{x}$ is the concentration measurements of each metabolite, and $\mathbf{x}$(i) is the concentration of each individual treatment with i: 1 ~ 24.

The missing values were treated as follows. For metabolite data, trace values (tr) and lower concentrations beyond the detection (nd) were replaced by 1/5 and 1/10 of the mean of all other valid observations, respectively. Metabolites which were consistently below the detection limit in all the samples, such as elaidic acid, were removed from the data set. Fifty-eight metabolites finally remained in the data set. For microarray data, the missing

Figure 1 Outline of the data integration strategy for the gene expression data and metabolite data. There are wide range of algorithms for normalisation, feature selection or generation, clustering to classification and computational validation.

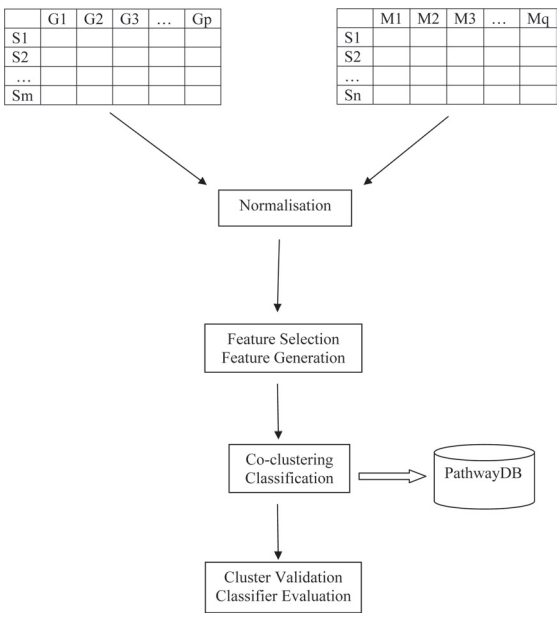
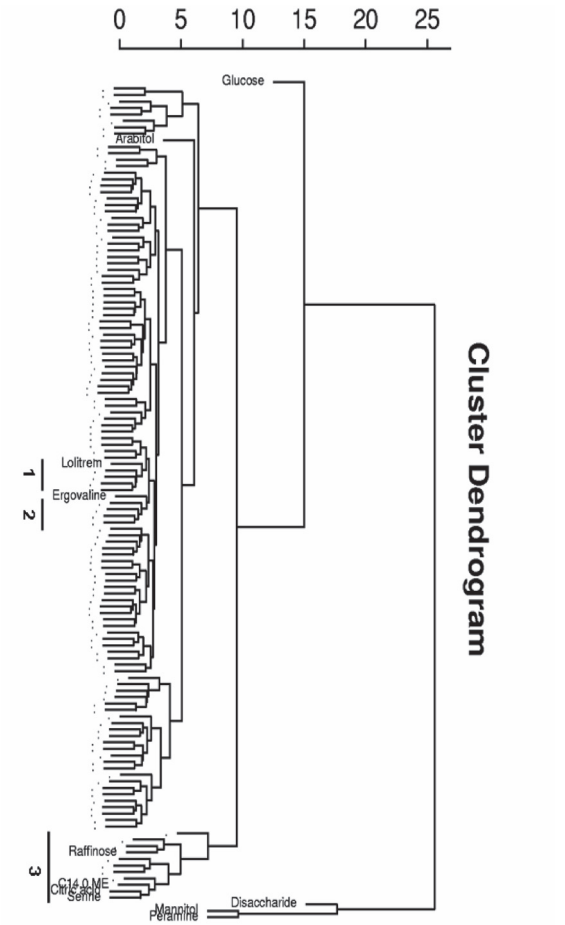


Figure 2 Co-clustering of selected ESTs and metabolites based on Euclidean distance and complete linkage.



values in the transformed data were replaced by 0, which means there are no differences between each treatment and the overall background. The transformed microarray data were then passed through a filter and ESTs with an absolute $\log_2$ value of each EST less than 1.0 (2-fold changes) across all the treatments were removed. Three thousand and sixty ESTs were then left for the following analysis.

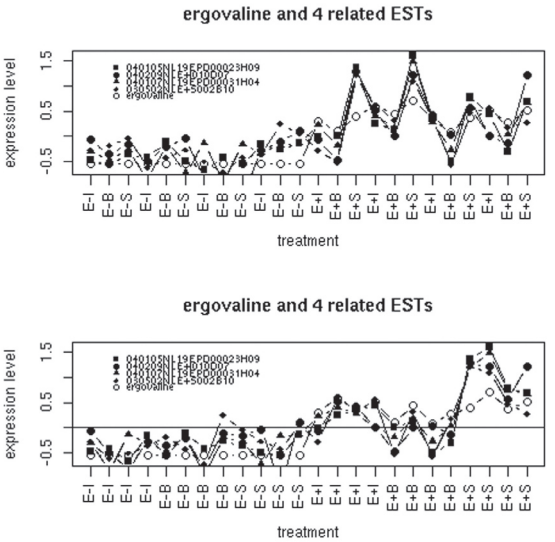
The general process of data analysis followed the strategy described in Fig. 1. After data normalisation, feature selection (or variable reduction) approaches were used to reduce the dimensionality of both microarray and metabolomics data. Various approaches were used for the feature selection, including t-test, principal component analysis (PCA) and Random Forest (RF) algorithm (Breiman 2001). Salient ESTs and metabolites were selected in relation to differential effects of endophyte infection (E+ vs E-). Co-clustering (co-occurrence) analysis of two sources of data was employed to reveal coherent relationships between ESTs and metabolites.

Results

The salient ESTs and metabolites relevant to the symbiosis

Feature selection approaches were conducted for the microarray

Figure 3 The pattern of gene expression and ergovaline accumulation (from cluster 3) in the E+ and E- tissues. Four ESTs are highly regulated in the ryegrass sheath (sorted by the tissue type in the lower plotting) along with ergovaline.



and metabolite data. Based on the simple t-test, 14 metabolites were selected which were statistically significantly different between E+ and E- ($P<0.05$). The PCA analysis showed that the 1st and 2nd principal components (PC) explain the tissue effect with 36.6% of variance by PC1 and 29.5% by PC2. The metabolic difference between E+ and E- was captured by PC3 (11.8%). The absolute loading values of the PC3 were sorted and the top 14 metabolites were selected and compared with the list from t-test. There were 11 metabolites (arabitol, mannitol, citric acid, disaccharide, raffinose, ergovaline, lolitrem B, peramine, C14.0 methyl ester, glucose, serine) found in common by the two analyses and used for the joint analysis with the microarray data.

Since a large number of ESTs was used in the microarray experiments, the cut-off of P-value must be adjusted to control for false positives when conducting statistical testing for significant genes. Based on the t-test with adjusted P-value (Benjamin & Hochberg's false discovery rate), 77 ESTs were selected with adjusted P-value ($P<0.075$). The cut-off of adjusted P-value means the expected proportion of false discoveries is 7.5%. Only seven ESTs were left with adjusted P value ($P<0.05$). The problems with the t-test approach are 1) the variance estimation for small number of samples is unreliable; 2) t-test treats each gene independently by ignoring the interaction between genes. Different feature selection approaches may be necessary for searching for target genes. The Random Forest algorithm was used to identify the important ESTs which differentiate between E+ and E- samples. The variables (ESTs) were sorted according to their importance as measured by the Gini index and the first 77 most highly ranked ESTs were selected for comparison with the ESTs selected by t-test with adjusted P-value ($P<0.075$). RF retains interactions among ESTs while avoiding over-fitting (9.7% error rate of sample classification here). Since there is no standard for setting up a threshold for cutting off the list, arbitrary

decisions were made here to use the union of the two lists by t-test and RF. Hence, in total 119 ESTs were retained for the joint analysis assuming that the interactions between ESTs might be maintained.

A joint analysis of selected ESTs and metabolites

There are generally two categories of analysis that may help to reveal a coherent relationship between gene expression and metabolite (see Cao *et al*. 2006 for a discussion in a broader sense). Clustering analyses are based on distance metrics, and the aim of classification is to find a classifier (model) which best differentiates the classes (Fig. 1). Simple clustering analysis was used in this study. Based on the Euclidean distance and complete linkage, 119 EST and 11 metabolites were co-clustered and the dendrogram is shown in Fig. 2.

Biological interpretation of the clusters is not a simple task no matter what kind of algorithm is used. All the 119 ESTs are likely to be relevant to the symbiosis due to their differentiated expression in E+ and E- ryegrass. Although a number of ESTs found here have a known role in the symbiosis, such as proteinase, chitinase and genes related to lolitrem biosynthesis (Johnson *et al*. 2006), most ESTs in the list have unknown functions, but might be novel components involved in the symbiosis. A close examination of the co-clusters in Fig. 2 shows that four ESTs share a sub-cluster (cluster 1) with lolitrem B. These four ESTs were highly up-regulated in the mature tissue (ryegrass sheath) of endophyte-infected ryegrass, which is also where lolitrem accumulates (Lane *et al*. 2007). Bioinformatics analysis has indicated that two genes have unknown annotations and two others are ascorbate peroxidise and DMAT gene, respectively. DMAT (dimethylallyl tryptophan synthase) is a gene involved in the lolitrem biosynthesis (Young *et al*. 2005). Cluster 2 (Fig. 2) comprises a sub-cluster of ergovaline related ESTs from the clustering analysis. The plot of this cluster also shows these four ESTs are highly expressed in the mature tissue of endophyte-infected ryegrass along with ergovaline (Fig. 3). These four ESTs have unclear annotations or their relation to symbiosis is unknown, for example, rice vhs2 (domain) protein. Cluster 3 (Fig. 2) is a cluster of the primary metabolites (serine, citric acid and C14.0 methyl ester and raffinose) related to the symbiosis. The known annotations of those seven ESTs in the cluster include rubisco, s-adenosyl-l-methionine synthetase, chlorophyll a/b binding protein, ribosomal RNA large subunit etc. all of which are likely related to the production of primary metabolites.

Discussion

The changes in metabolic phenotype may be the outcome of genes, but the relationship between gene expression and metabolite concentration change is inherently indirect. In the simplest case, it is mediated by the levels of protein (enzyme). However, combined analysis of metabolomic and microarray data, even with the absence of proteomic information, has been used to identify gene function and regulatory networks in several organisms (Phelps *et al.* 2002: Colebatch *et al.* 2004), to find key metabolic pathways relevant to nutritional stress in Arabidopsis (Hirai *et al.* 2004), and to identify target genes which are then confirmed by Northern blot (Verdonk *et al.* 2003). Our initial joint investigations of gene expression and metabolite accumulation has identified several ESTs which show distinctive expression patterns which coincide with the pattern of accumulation of alkaloids and other metabolites in the mature tissues of endophyte-infected ryegrass. We hope that increasing genomic information of the ryegrass and endophyte with effective data integration methods will improve our understanding of this complex system.

There are many uncertainties in high-throughput functional genomic data generation and analysis (Malo *et al.* 2006). The integration of data from different sources helps reinforce *bona fide* observations and reduce false negatives (Urbanczyk-Wochniak *et al.* 2003; Hwang *et al.* 2005). Other information, such as biochemical pathways and expert knowledge are indispensable parts of data interpretation. Questions to be examined in this kind of co-clustering analysis include: are the co-clustered genes and metabolites co-located in the same metabolic pathway or under the same regulatory mechanism? What novel metrics are the best to describe the coherent relationships between genes and metabolites with or without information about proteins?

High-throughput technologies in functional genomics produce comprehensive information about a biological system but on the other hand generate noisy data. Laboratory verification of candidate genes from microarray analysis is considered as a required standard. However, the list of candidates could be quite long for such complex biological systems. Computational biology may be needed to model the mode of interaction for providing better valid targets. From the practical perspective, the selected ESTs could be used to screen for genetic markers, such as EST-SSR (EST derived simple sequence repeats) (Varshey *et al.* 2005). The putative function of ESTs could then prove useful for marker-assisted selection in breeding for ryegrass. This approach could serve as a bridge from high-throughput functional genomics to traditional plant improvement.

ACKNOWLEDGEMENTS

We thank Zaneta Park-Ng for being involved in the processing of microarray data; David Baird for advice on the microarray data transformation; Alan McCulloch for EST assembly and blasting multiple gene databases; Karl Fraser and Brian Tapper for targeted analysis of metabolites data.

REFERENCES

Breiman, L. 2001. Random Forests. *Machine Learning 45*: 5-32.

Cao, M.; Koulman, A.; Pacheco, D.; Lane, G.A.; Rasmussen, S. 2006. Systems Biology and Metabolomics: Experimental and Computational Challenges. *Proceedings of the New Zealand Society of Animal Production 66*: 213-218.

Colebatch, G.; Desbrosses, G.; Ott, T.; Krusell, L.; Montanari, O.; Kloska, S.; Kopka, J.; Udvardi, M.K. 2004. Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *Plant Journal. 39(4)*: 487-512.

Hirai, M.Y.; Yano, M.; Goodenowe, D.B.; Kanaya, S.; Kimura, T.; Awazuhara, M.; Arita, M.; Fujiwara, T.; Saito, K. 2004. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stress in *Arabiodopsis thaliana*. *Proceedings of the National Academy of Sciences, USA. 101(27)*: 10205-10210.

Hwang, D.; Rust, A.G.; Ramsey, S.; Smith, J.J.; Leslie, D.M.; Weston, A.D.; de Atauri, P.; Aitchison, J.D.; Hood, L.; Siegel, A.F.; Bolouri, H. 2005. A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences, USA. 102(48)*: 17296-17301.

Johnson, R.D.; Bassett, S.; Cao, M.; Christensen, M.; Gaborit,

C.; Johnson, L.; Koulman, A.; Rasmussen, S.; Voisey, C.; Bryan, G. 2006. A multidisciplinary approach to dissect the molecular basis of the *Neotyphodium lolii*/ryegrass symbiosis. pp. 107-114. *In*: Advances in Pasture Plant Breeding. Ed. Mercer, C.F. Grassland Research and Practice Series No. 12. New Zealand Grassland Association.

Lane, G.A.; Cao, M.: Johnson, L.J.; Koulman, A. Popay, A.J.; Rasmussen, S.; Tapper, B.A. 2007 Anti-herbivore factors of grass endophytes: new prospects from metabolomics. pp 307 *In*: Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses, Eds. Popay, A.J.; Thom, E.R. Grassland Research and Practice Series No. 13. New Zealand Grassland Association.

Phelps, T.J.; Palumbo, A.V.; Beliaev, A.S. 2002. Metabolomics and microarrays for improved understanding of phenotypic characteristics controlled by both genomics and environmental constraints. *Current Opinion in Biotechnology. 13*: 20-24.

Malo, N.; Hanley, J.A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. 2006. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology 24(2)*: 167-175.

Urbanczyk-Wochniak, E.; Luedemann, A.; Kopka, J.; Selbig, J.; Roessner-Yunali, U.; Willmitzer, L.; Ferenie, A.R. 2003. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports 4(10)*:989-993.

Varshney, R.K.; Graner, A.; Sorrells, M.E. 2005. Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology 23(1)*: 48-55.

Verdonk, J.C.; Ric de Vos, C.H.; Verhoeven, H.A.; Haring, M.A.; van Tunen, A.J.; Schuurink, R.C. 2003. Regulation of floral scent production in petunia revealed by targeted metabolomics. *Phytochemistry 62*: 997-1008.

Young, C.A.; Bryant, M.K.; Christensen, M.J.; Tapper, B.A.; Bryan, G.T.; Scott, B. 2005. Molecular cloning and genetic analysis of a symbiosis-expressed gene cluster for lolitrem biosynthesis from a mutualistic endophyte of perennial ryegrass. *Molecular Genetic Genomics 274*: 13-29.