Exploratory Analysis of High-Throughput Metabolomic Data

ARTICLE *in* METABOLOMICS · DECEMBER 2013

Impact Factor: 3.86 · DOI: 10.1007/s11306-013-0545-6

Jairus B Bowne
University of Melbourne
6 PUBLICATIONS **152** CITATIONS

Arthur L Hsu
University of Melbourne
44 PUBLICATIONS **612** CITATIONS

Ute Roessner
University of Melbourne
117 PUBLICATIONS **6,531** CITATIONS

Saman K Halgamuge
University of Melbourne
238 PUBLICATIONS **4,115** CITATIONS

Metabolomics (2013) 9:1311–1320
DOI 10.1007/s11306-013-0545-6

ORIGINAL ARTICLE

Exploratory analysis of high-throughput metabolomic data

Chalini D. Wijetunge · Zhaoping Li ·
Isaam Saeed · Jairus Bowne · Arthur L. Hsu ·
Ute Roessner · Antony Bacic · Saman K. Halgamuge

Received: 17 March 2013 / Accepted: 8 May 2013 / Published online: 21 May 2013

Abstract In order to make sense of the sheer volume of metabolomic data that can be generated using current technology, robust data analysis tools are essential. We propose the use of the growing self-organizing map (GSOM) algorithm and by doing so demonstrate that a deeper analysis of metabolomics data is possible in comparison to the widely used batch-learning self-organizing map, hierarchical cluster analysis and partitioning around medoids algorithms on simulated and real-world time-course metabolomic datasets. We then applied GSOM to a recently published dataset representing metabolome response patterns of three wheat cultivars subject to a field simulated cyclic drought stress. This novel and information rich analysis provided by the proposed GSOM framework can be easily extended to other high-throughput metabolomics studies.

Keywords Metabolomics data analysis · Growing self-organising map · Unsupervised learning

1 Introduction

High-throughput metabolomics aims to comprehensively identify and quantify small-molecules (metabolites) in a given biological sample (Wishart 2008). The data generated from a metabolomics experiment complements genomic, transcriptomic and proteomic studies, and has profound advantages in bridging the gap between genotype and phenotype (Fiehn 2002). The significant advances in high-throughput technologies have produced an avalanche of data in biology. Unfortunately, traditional statistical methods seem insufficient to handle such data and have many limitations, especially for high dimensional and huge volume datasets. These two obstacles often hinder the analysis and interpretation of metabolomics data. Therefore, in order to extract global information from biological experiments subjected to high-throughput analyses, a

Electronic supplementary material The online version of this article (doi:10.1007/s11306-013-0545-6) contains supplementary material, which is available to authorized users.

Chalini D. Wijetunge and Zhaoping Li are equal first authors.

C. D. Wijetunge · Z. Li · I. Saeed · A. L. Hsu ·
S. K. Halgamuge (✉)
Optimisation and Pattern Recognition Group, Department of Mechanical Engineering, Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
e-mail: saman@unimelb.edu.au

J. Bowne · U. Roessner
Australian Centre for Plant Functional Genomics, School of Botany, The University of Melbourne, Parkville, VIC 3010, Australia

J. Bowne · U. Roessner · A. Bacic
Metabolomics Australia, School of Botany, The University of Melbourne, Parkville, VIC 3010, Australia

A. L. Hsu
Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia

A. Bacic
ARC Centre of Excellence in Plant Cell Walls, School of Botany, The University of Melbourne, Parkville, VIC 3010, Australia

multidisciplinary approach is required to facilitate the integration of biological questions into mathematical and computational challenges. In this regard, multivariate data analysis techniques—such as supervised classifiers and, more frequently, unsupervised clustering algorithms—are typically applied in favor of a purely statistical approach. Within the unsupervised paradigm, self-organizing maps (SOMs), and its various extensions (Abe et al. 2002; Alahakoon et al. 2000; Weber et al. 2011), have been widely applied.

In particular, batch-learning self-organizing map (BL-SOM) has become a commonly used method for analyzing metabolomics data. Hirai et al. (2004, 2005) have used BL-SOM to analyse integrated metabolomics and transcriptomics datasets to investigate plant responses to abiotic stress. Kim et al. (2007b) used the BL-SOM to analyse time-course metabolomic data to better understand the salt-stress response mechanisms of *Arabidopsis thaliana*. Similarly, Ohta et al. (2007) used BL-SOM to investigate metabolite accumulation patterns of *A. thaliana* under a series treatment by herbicidal enzyme inhibitors. Sawada et al. (2009) used BL-SOM to cluster a combined dataset of a measured metabolomic dataset and a model metabolomic dataset representing three plant families to search for family-specific metabolites.

Despite its common use in metabolomics studies, the rationale for selecting BL-SOM as a data analysis tool has not been clearly justified—stimulated perhaps by its popularity in a wide variety of other "omics" studies (Abe et al. 2006a, b, 2009; Hayashi et al. 2005; Kanaya et al. 2001; Uchiyama et al. 2004)—nor has it been compared to alternative clustering algorithms. As such, it is necessary that the performance of BL-SOM be compared against other algorithms to investigate whether metabolomic data can be analyzed more effectively. In this study we focus our attention to the growing self-organizing map (GSOM) algorithm. To the best of our knowledge, this is the first such attempt to benchmark methods for metabolomic data analysis and also the first attempt to apply GSOM to metabolomic data.

GSOM has also recently gained popularity in the analysis of biological data generated from high-throughput experiments, and at the same time has also demonstrated improved performance and clustering capability (such as topology preservation) over standard, fixed-size SOMs such as BL-SOM. Notably, Hsu et al. (2003a) combined the dynamic SOM tree with GSOM to perform class discovery and marker gene identification in micro-array data. Moreover, Chan et al. (2008a) improved the performance of clustering nucleotide frequency vectors for binning metagenomic data. The same authors also proposed a semi-supervised extension to GSOM (Chan et al. 2008b). These studies demonstrated that GSOM performed better than conventional SOM-based algorithms in clustering various datasets.

Considering the flexibility of the GSOM algorithm and its successful applications on metagenomic and transcriptomic data, we compare the clustering performance of GSOM with 3 other widely-used clustering algorithms, including: hierarchical cluster analysis (HCA); partitioning around medoids (PAM); and the BL-SOM algorithm. Since there are no widely accepted benchmarks for metabolome time-course data, we propose the use of simulated metabolomic datasets (of controlled complexity), in addition to a real-world metabolomic dataset (Kim et al. 2007a) to evaluate the performance of the algorithms under investigation. The best performing algorithm is then applied to a new, recently published real-world dataset (Bowne et al. 2012), with the results validated by complementary analyses of the data under investigation.

2 Methods

2.1 Data sets

2.1.1 *Simulated benchmark data*

We constructed 64 simulated datasets each of which represents a time-course metabolomic dataset. Based on our previous analyses of real-world time-course metabolomic data, seven different and frequently occurring concentration patterns of metabolites were selected to simulate these datasets (Supplementary Fig. 1). The 64 datasets were created by varying the sample size from 10 to 40 and the number of time points from 5 to 20, with 4 datasets generated at each setting to mimic the effect of replicates. Random noise was also added to each dataset to simulate the effects of imperfect data.

2.1.2 *Real-world benchmark: Arabidopsis thaliana under NaCl stress*

This benchmark dataset is based on a study originally published by Kim et al. (2007a, b). The data was generated from an experiment which first imposed NaCl on *A. thaliana* cells and then used gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS) to obtain time-course metabolomic data. Samples from both control group and salt-stress group were harvested after 0.5, 1, 2, 4, 12, 24, 48 and 72 h. The dataset contains 47 metabolites in which 25 metabolites have been emphasized in their analysis. The published data is given as fold changes of the salt-stress group over the corresponding control group. This dataset is similar in design to the wheat metabolome dataset which we aim to investigate using appropriate unsupervised methods, and also it provides a sound base with which to benchmark different analysis methods.

2.1.3 *Metabolomic time-course data of wheat cultivars under drought stress*

Three wheat cultivars—Kukri (drought-susceptible), RAC875 (drought-tolerant) and Excalibur (drought-tolerant)—were subjected to cyclic drought stress by controlling the amount of water they receive (Bowne et al. 2012). All three cultivars were reported to show different growth performances under intermittent periods of rainfall in South Australia. For the purpose of investigating drought tolerance mechanisms among these three cultivars, an experiment which simulates the intermittent rainfall during the growing season was conducted by Izanloo et al. (2008). High-throughput metabolite profiling of these cultivars through GC-MS was performed on samples of both control and drought-stress groups, resulting in the identification of 86 metabolites across all three cultivars (Bowne et al. 2012). Supplementary Fig. 2 illustrates the drought treatment on the drought-stress group, where water was initially withdrawn gradually to reduce soil moisture levels. Upon reaching wilting point, the plants were re-watered to field capacity. For the purposes of this study, fold changes of the drought-stress group data over the corresponding control group data were analyzed.

2.2 Clustering algorithms

2.2.1 *Growing self-organizing map*

The (GSOM) (Alahakoon et al. 2000) is a useful extension to the conventional SOM. The algorithm initially produces a feature map with a small number of nodes and generates new nodes when necessary to better conform with the topology of the input data whereas both the SOM and BL-SOM algorithms require definition of the map size prior to execution, and remain fixed during the learning process. The GSOM can be initialised with only four nodes in the rectangular topology, or seven nodes in the hexagonal topology. The latter was selected in this study for better performance (Hsu et al. 2003a). Input vectors are then sequentially processed by the algorithm (similar to SOM). The size of the feature map is controlled by growth threshold (*GT*) which governs the growth of new nodes. The *GT* (Chan and Halgamuge 2009) is defined as:

$$GT = -D \times \ln(SF), \tag{1}$$

where *D* represents the dimensionality of the input vectors; and *SF* represents the user-defined spread factor, given $SF \in (0, 1]$. A value of $GT \approx 0$ represents minimum growth and a value of 1 represents maximum growth. The common parameters of GSOM that were used in this study (applicable to all data sets) can be summarised as follows: a hexagonal map topology, a Euclidean similarity measure, and a fixed initial value of 0.5 for all weight vectors (see Chan et al. 2008b). Clusters were identified based on a threshold of the distance between a node and its neighbors. GSOM used in this work is available at: http://140.109.29.21/s-gsom/.

2.2.2 *Batch learning self-organizing map*

This algorithm, a modified version of SOM, was originally proposed by Abe et al. (1999) for gene classification. The primary difference between BL-SOM and the conventional SOM is the learning process, where BL-SOM compares an entire batch of input data with all weight vectors of a feature map before updating weight vectors (Kohonen 2001), whereas the conventional SOM processes one input vector at a time (so called sequential learning)—batch learning can efficiently reduce algorithm runtime. BL-SOM initializes weight vectors using PCA instead of random values, which yields more reproducible feature maps compared to sequential learning SOMs (Kanaya et al. 2001; Yano et al. 2006). The weight vector W_{ij} of a BL-SOM feature map in the *ij*-th lattice is initialized using:

$$W_{ij} = X_{av} + 5\sigma_1 \frac{b_1\left(i - \frac{I}{2}\right) + b_2\left(j - \frac{J}{2}\right)}{I}, \tag{2}$$

where *I* represents the first dimension of the weight vector defined by $5\sigma_1$ and *J* represents the second dimension of the weight vector defined by the nearest integer greater than $\left(\frac{\sigma_2}{\sigma_1}\right) \times I$; σ_1 and σ_2 represent the standard deviation of the first and second principal components; X_{av} is the average vector at each node; and b_1 and b_2 represent the eigenvectors for the first and second principal components, respectively. The only user-defined parameter during initialization is the map width (the height is inferred). BL-SOM used in this work is available at: http://prime.psc.riken.jp/?action=blsom_index.

2.2.3 *Hierarchical cluster analysis*

Hierarchical cluster analysis (HCA) clusters a dataset into different groups by creating a hierarchical dendrogram representing the relationships of data points with respect to a predefined distance measure. Two procedures exist for constructing the dendrogram: agglomerative and divisive methods. The agglomerative method initially considers each observation as a separate cluster and then iteratively merges the two closest clusters into one larger cluster until a singleton cluster is formed containing all data points. The divisive method proceeds in the opposite direction, i.e. the initial cluster contains all data points, and is recursively split into two. HCA has become an important algorithm in metabolomics data analysis (Sawada et al. 2009).

2.2.4 Partitioning around medoids

Partitioning around medoids (PAM) algorithm requires the number of clusters to find as an *a priori* input to the algorithm, while the initial set of cluster centres are set randomly. Each data point is then associated to the closest cluster selected based on a valid distance metric such as Euclidean distance, Manhattan distance or Minkowski distance. After each data point is assigned to a cluster, all cluster centres are updated. These two steps are repeated until all cluster centres become stable and the algorithm converges. Although PAM is similar to the standard K-means algorithm, it is more robust than K-means due to its ability to incorporate dissimilarity matrices other than those based on Euclidean distances.

2.2.5 Cluster selection

In regards to GSOM, cluster selection is automatic, and (if desired) can also be inferred after a map is produced. For the remaining algorithms, they require the number of clusters *a priori*, which is their drawback in the context of automated exploratory approaches. Incorrect specification of the number of clusters can lead to incorrect interpretations of the data. However, there are existing and well established methods to estimate a plausible number of clusters based on the properties of the data under investigation. In this study the silhouette width—which measures how accurately each observation in the dataset is assigned to a cluster—will be used. Values for the Silhouette width range between [−1,1], and should be maximized to obtain an *optimal* clustering result.

3 Results

The data sets used in this study were normalised prior to clustering (see Supplementary Information 1).

3.1 Benchmark results

3.1.1 Simulated benchmark data

The three clustering algorithms, GSOM, HCA and PAM were applied to all simulated datasets and their respective clustering performances were evaluated using Adjusted Rand Index (ARI) (Supplementary Information 2).

We conducted two types of benchmark analyses: in the first case we supplied the clustering algorithms with the correct number of clusters to examine their performance in the optimal scenario; and in the second, more realistic case, we used a metric to estimate the number of clusters and allowed the clustering algorithm to select the most ideal cluster solution based on the data at hand.

When the exact numbers of clusters are provided as input to HCA and PAM, their performance on the simulated datasets are comparable to that of the GSOM algorithm (Supplementary Fig. 3 and Supplementary Table 1). However, this is not an unbiased comparison since GSOM does not require such knowledge about the number of clusters in advance. Moreover, when dealing with most real-world metabolomic datasets, we cannot expect to have such *a priori* knowledge about the groupings in an exploratory setting. Therefore, in each case, the number of clusters was estimated using the Silhouette width. When the estimated numbers of clusters were given instead of the exact numbers of clusters, the performances of the PAM and HCA algorithms were dramatically decreased (Supplementary Table 2). The graphs in Fig. 1 clearly illustrate the significant improvement in performance of the GSOM when compared to HCA and PAM algorithms for each sample size. GSOM not only clusters the datasets accurately but also visualizes the relationships between clusters thus providing some additional information to biologists.

3.1.2 Real-world benchmark: Arabidopsis thaliana under NaCl stress

Neither HCA nor PAM algorithm was able to produce the set of clusters suggested by Kim et al. (2007a), instead both algorithms were only able to identify two abstract clusters. Both BL-SOM and GSOM were able to group metabolites which have similar response patterns over a series of predefined time intervals, but the extent to which these groups accurately represent the underlying biology of the samples can differ. In this regard, we evaluated the clustering performance by first examining the error within the groups (measured by Euclidean distance) as well as the Pearson correlation within the groups. The groups produced by each method were then interpreted based on the metabolites identified in each group.

The ideal set of clusters that Kim et al. (2007a) suggested through their analysis, was obtained by GSOM (G1, G2 and G3 in Fig. 2) whereas BL-SOM generated the same set of clusters reported in the original study (B1, B2, B3 and B4 in Supplementary Fig. 4). Figure 3 illustrates that the average correlation of metabolites within each group clustered by GSOM is higher than that of the BL-SOM groups, while residual errors within the GSOM groups are lower than those of BL-SOM. We observed that GSOM provided a 23.36 ± 9.64 % improvement in the correlations within groups of metabolomic profiles and 14.11 ± 8.45 % reduction in the residual errors within groups. These results demonstrate the significant

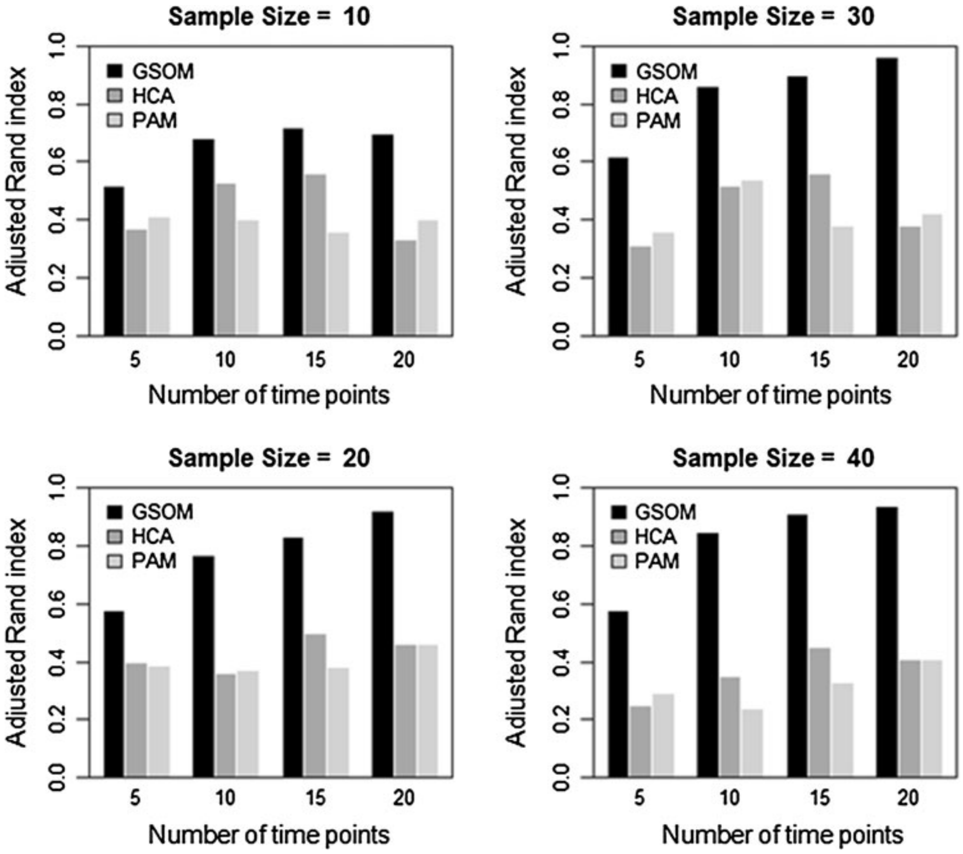

Fig. 1 Comparison of the clustering performance of *GSOM*, *HCA* and *PAM* algorithms on the 64 simulated datasets based on the adjusted rand index when the *estimated* numbers of clusters were provided as input to *HCA* and *PAM*

improvement in performance of GSOM when compared with BL-SOM.

Given these benchmark results, we then focused our comparison on a subset of metabolites from each BL-SOM cluster (Table 1), which have been further analyzed in the original study due to their significance. Metabolites clustered in each of the BL-SOM and GSOM groups are summarized in Table 1. GSOM groups G1 and G3 are exactly similar to the BL-SOM groups B1 and B4 respectively. Kim et al. (2007a) noted that metabolites from Group B2 and Group B3 should be combined (even though they were identified by BL-SOM as distinct groups) as their patterns were similar. Interestingly, the metabolites in GSOM Group G2 are the same as the combination of BL-SOM Groups B2 and B3, and therefore GSOM provides a better clustering representation.

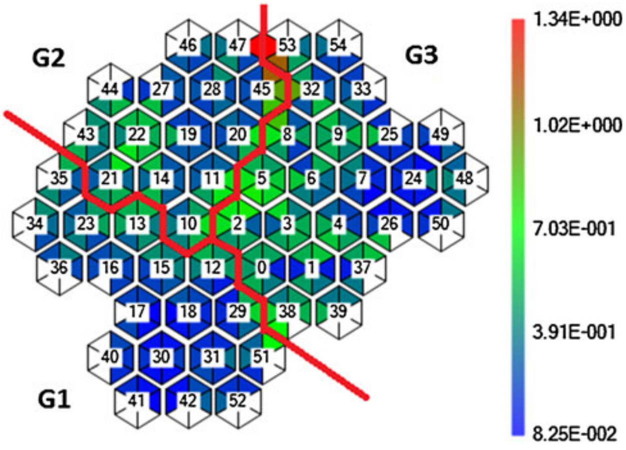

Fig. 2 Application of GSOM on the real-world benchmark (*Arabidopsis thaliana* under NaCl stress), using *SF* = 0.4 (*SF*—Spread Factor); *G1*– *G3* represent Group *G1* to Group *G3*

The improved performance of GSOM may be due to two of its main advantages: (1) the growing size of the map circumvents the need to specify a predefined map structure, which may better preserve the inherent structure of the data (Chan et al. 2008a); and (2) the GSOM spread factor can flexibly control the resolution of clustering without having to define a possibly unrepresentative map size *a priori*. Even though BL-SOM can also achieve different map resolutions by specifying various predefined map sizes, it still lacks flexibility in clustering compared to GSOM, which is considered as an important property in data analysis (Chan et al. 2008a). Notably, we do acknowledge that the computational cost of GSOM is comparable to BL-SOM for the data sets considered—this however does not include the time required to estimate the map sizes for BL-SOM to suit the data at hand.

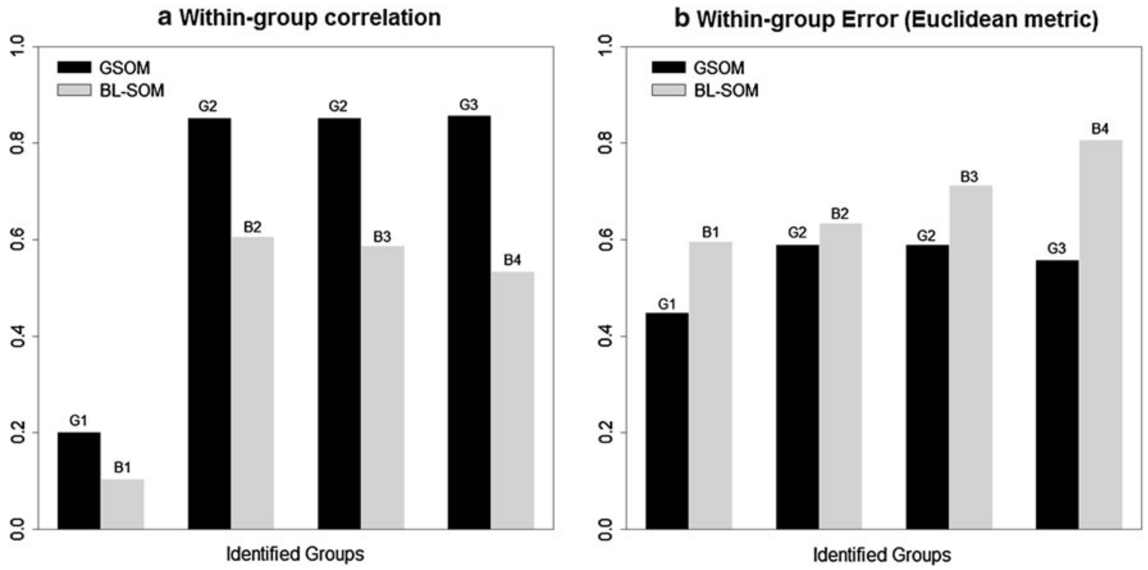

Fig. 3 Comparison of all the *GSOM* and *BL-SOM* generated groups (for the real-world benchmark, *Arabidopsis thaliana* under NaCl stress) based on **a** pearson correlation and **b** residual errors within groups

Table 1 Comparison of clustering results produced by BL-SOM and GSOM using the subset of metabolites emphasised by Kim et al. (2007a)

BL-SOM Group B1	GSOM Group G1	BL-SOM Group B2	GSOM Group G2	BL-SOM Group B4	GSOM Group G3
Glutamate	Glutamate	Gluconate	Gluconate	Lactate	Lactate
Glycine	Glycine	Cysteine	Cysteine	Sucrose	Sucrose
Folate	Folate	Ethanolamine	Ethanolamine	Fumarate	Fumarate
Formate	Formate			Malate	Malate
Serine	Serine	*Group B3*		Shikimate	Shikimate
Cadaverine	Cadaverine	Citrate	Citrate	D-fructose	D-fructose
Proline	Proline	Lysine	Lysine	Succinate	Succinate
n-Butylamine	*n*-Butylamine	Tyrosine	Tyrosine	Pyruvate	Pyruvate
		Tryptophan	Tryptophan		
		Phenylalanine	Phenylalanine		
		Methylation	Methylation		

Methylation is the ratio of S-adenosyl-L-methionine to S-adenosyl-L-homocysteine and is used for the evaluation of endogenous methylation activity

3.1.3 *Application of GSOM to a new real-world metabolomic dataset: wheat cultivars under drought stress*

We then applied GSOM to a recently published wheat metabolomic data set. The results obtained agreed with previous complementary analyses of the data, and in cases where additional information was uncovered, an interpretation of the results is presented and found to be consistent with the biology of the sample.

The wheat metabolomic dataset used in this study contains three types of metabolites namely sugars, amino acids and organic acids. GSOM was applied to these three types of metabolites separately in order to uncover any hidden patterns among metabolites in each category. The application of GSOM to the amino acid metabolite data produced four clusters (AA1, AA2, AA3 and AA4 in Supplementary Table 3). Supplementary Table 4 provides an overview of the distribution of amino acids of the three cultivars suggesting that there are overlaps.

Figure 4 shows a Venn diagram of the corresponding amino acid metabolites, showing the metabolite patterns shared between each of the three cultivars and their association with the four pattern groups. It shows that the number of amino acid metabolites shared between drought-tolerant RAC875 and Excalibur only (13 metabolites in Intersection 3) is significantly higher than Intersections 1 and 2. Interestingly, all the metabolites in Intersection 3

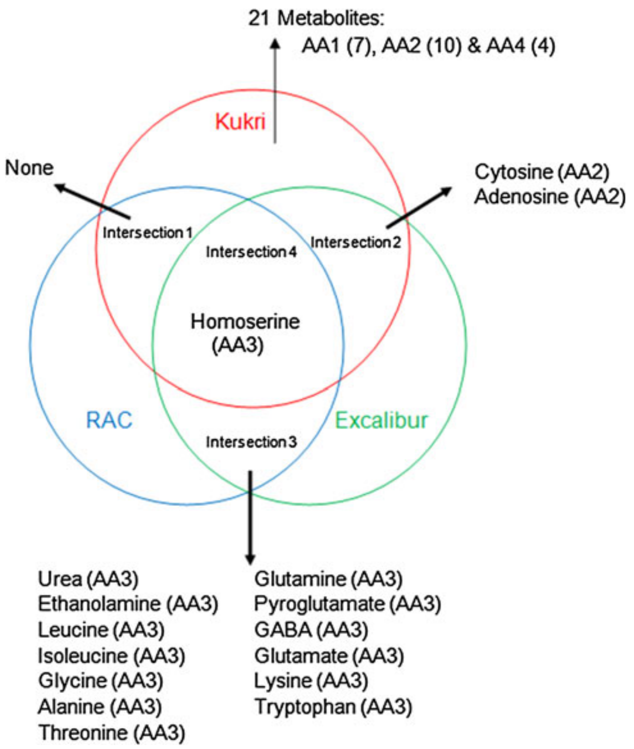

Fig. 4 Application of GSOM to metabolomic time-course data of wheat cultivars under drought stress—summary of corresponding amino acid metabolites that share the same patterns among different cultivars. The three circles represent the metabolites of Kukri, RAC875 and Excalibur. Intersections *1–4* represent corresponding metabolites shared by Kukri and RAC875, Kukri and Excalibur, RAC875 and Excalibur, Kukri and RAC875 and Excalibur in the same pattern, respectively. *AA1–AA4* represent the patterns which such metabolites belong to

follow the pattern AA3. Moreover, a significant number of amino acid metabolites (21 out of 24) in drought susceptible Kukri are unique and the majority among them follow the patterns AA1 and AA2.

Based on these results, it is reasonable to assume that amino acid metabolites that follow patterns AA1 and AA2 affect the drought susceptibility of the cultivar, given that Kukri is known to be drought susceptible. Similarly, it can be assumed that pattern AA3 which is unique to RAC875 and Excalibur (drought-tolerant cultivars) potentially includes amino acid metabolites specific to drought tolerant mechanisms. However, more experiments are required to support this observation.

We then focused our attention on the trends of amino acids in each cultivar over the time-course. In the first drought cycle (see Supplementary Fig. 5), fold changes of amino acids in RAC875 and Excalibur show a negative correlation in AA3 with decreasing water content, while Kukri shows positive and then negative correlations in AA2. When comparing the first two severe drought conditions (i.e. when the relative water content is approximately 40 % at time points C and D), fold changes of Kukri amino acids show no significant change in the second drought condition with respect to the first drought condition. Contrastingly, fold changes of RAC875 and Excalibur amino acids show an increase in the second drought condition with respect to the first drought condition. Since a significant number of RAC875 and Excalibur amino acids follow this trend, it suggests that the amino acid metabolites affect the drought tolerance mechanisms of cultivars strongly. Bowne et al. (2012) also observed this significant change in the levels of amino acids of the three cultivars over the duration of drought stress.

After analyzing the amino acids in the wheat metabolomic dataset, we then applied the GSOM algorithm to the organic acid metabolites and obtained four clusters (OA1, OA2, OA3 and OA4 in Supplementary Table 5). Supplementary Table 6 summarizes the distribution of organic acids of each of the three cultivars in each pattern group. According to Fig. 5, the number of organic acid metabolites shared between RAC875 and Excalibur (Intersection 3) is not significantly different from Intersection 1 or Intersection 2. These results may indicate that RAC875 and Excalibur possess different mechanisms of drought resistance even though they share some similar metabolite patterns. This was also suggested by Izanloo et al. (2008) from physiological measurements of the three cultivars.

Finally, with the application of GSOM to the sugar metabolites, 6 clusters were obtained. Supplementary Tables 7, 8 and Supplementary Fig. 6 summarize the obtained clustering results. However, these results do not directly reveal a similarity in between the two drought tolerant cultivars or a clear difference among the drought susceptible and tolerant cultivars.

4 Discussion

4.1 Performance of GSOM in metabolomic data analysis

The selection of an unsupervised approach to exploratory pattern discovery in metabolomics data is motivated by the often limited *a priori* knowledge of the biological sample under investigation. The approach used in this study offers unbiased clues into such biological mechanisms, allowing the data itself to inform interpretation of the experimental results. In this regard, the advantages of GSOM can be summarized as follows (in the context of the wheat metabolome):

Firstly, since the abiotic stress imposed on the wheat cultivars is dynamic and non-linear (cyclic drought stress) the metabolite responses to this environmental stress are

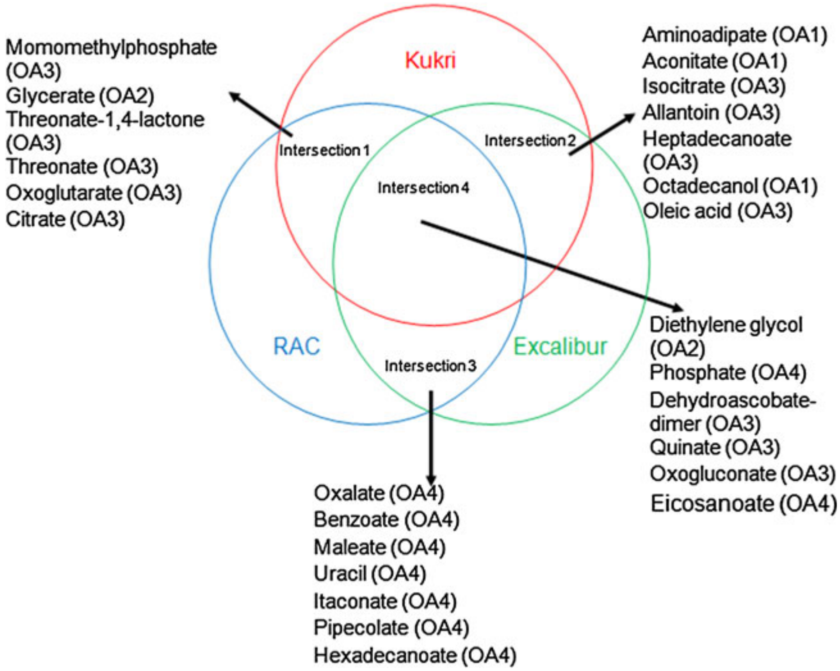

Fig. 5 Application of GSOM to metabolomic time-course data of wheat cultivars under drought stress—summary of corresponding organic acid metabolites that share the same patterns among different cultivars. The three circles represent the metabolites of Kukri, RAC875 and Excalibur. Intersections *1–4* represent corresponding metabolites shared by Kukri and RAC875, Kukri and Excalibur, RAC875 and Excalibur, Kukri and RAC875 and Excalibur in the same pattern, respectively. *OA1–OA4* represent the patterns which such metabolites belong to

less likely to be linear. Therefore, simple linear analysis or basic statistics may not be sufficient to capture such variations. The GSOM algorithm, on the other hand, reduces the dimensionality of the data non-linearly, thus preserving the topology of the possibly non-linear relationships of interest. Moreover, GSOM uses a dynamic feature map model rather than a predefined map structure, and uses the spread factor to control the resolution of clustering, which contributes to topology preservation and facilitates analysis from different levels of data abstraction (Alahakoon et al. 2000). These features enable the GSOM algorithm to outweigh traditional metabolomic data analysis tools such as principal component analysis (PCA) and as this study shows HCA, PAM algorithm and the competing SOM-based method BL-SOM algorithm.

Secondly, in terms of data size, the wheat metabolomics dataset contains a large volume of relative metabolite concentrations and is of high dimensionality (3 cultivars, 86 identified metabolites for each cultivar, and each metabolite observed at 5 time points). The GSOM algorithm can effectively and efficiently project such high-dimensional data onto a representative two dimensional feature map, enabling the visualization of cluster structure as well as the similarity of metabolite response patterns.

The application of GSOM to larger, higher dimensional metabolomics data sets can be made possible by two of its key characteristics: topology preservation and scalable computational speed. The GSOM algorithm like its predecessor (SOM) is inherently designed to handle high-dimensional data, and with a modified Growth Factor, the map topology has been shown to be preserved with increasing data dimensionality (Chan and Halgamuge 2009) as shown in several other applications with large data sets from metagenomics (Weber et al. 2011) and microarray (Hsu et al. 2003a) studies. Overall, its flexibility in dynamically generating a map largely contributes to its topology preservation capability for high dimensional data (a trait that is crucial to obtaining a meaningful clustering of the data, and for subsequent extraction of relevant patterns that arise in the data) as described in Hsu and Halgamuge (2003b). In terms of increasing the computational speed, there are several implementations and extensions of GSOM that have been designed for this purpose. For example, our previous work has explored a hardware specific implementation of GSOM (Preethichandra et al. 2002), as well as a modified GSOM algorithm tailored for parallel computation (Zhai et al. 2006).

Clustering patterns in data is a common approach in time-course microarray and metabolomics data analysis. Even though the observed performance of GSOM is higher than BL-SOM, HCA and PAM algorithms, there are certain limitations of GSOM that can be addressed to further improve clustering performance. For example, the current GSOM algorithm will produce different feature maps depending on the order of input vectors, even though the topology of the data will still be preserved. In order to overcome this limitation, the GSOM algorithm was executed several times on each dataset. Clustering results of multiple GSOM runs are not reported in this paper because the observed differences were negligible. In addition, the current GSOM algorithm incorporates only Euclidean and Manhattan distances in clustering. For complex and multidimensional datasets neither of them may be the best choice. The Minkowski matrix with an optimized variable, which can vary the distance measure accordingly, has proven to be a better alternative to improve the accuracy of the classifier especially when dealing with multidimensional data (Halgamuge 1997). Therefore, we propose to modify the current GSOM algorithm by incorporating the Minkowski distance matrix, which can replace both Euclidean and Manhattan distance matrices, in order to improve its clustering performance further.

Although the two standard clustering algorithms—HCA and PAM perform poorly when analyzing complex metabolomics datasets with no prior knowledge about the groupings, they perform well when the exact number of clusters in the dataset are known, which can be true for some applications. For example, these algorithms may reveal promising results in a biomarker study which aims at finding two groups comprising diseased and healthy individuals.

5 Concluding remarks

Through rigorous benchmarking and validation on simulated and real-world time-course metabolomic data, we have proposed the use of GSOM as an effective means to extract novel patterns in metabolite response profiles. While the benchmark results on simulated data showed that GSOM was the best performing algorithm, the subsequent real-world benchmark further revealed that GSOM was able to group similar metabolite response patterns with lower residual error and higher correlation than BL-SOM. These results indicated that the output of GSOM is biologically more informative than BL-SOM. The results of GSOM on a recently published wheat metabolome dataset were validated by complementary analyses of the sample, where GSOM was successfully able to uncover metabolite response patterns with respect to drought stress, revealing potentially different adaptation mechanisms of the three wheat cultivars in the dataset. The benchmark procedure proposed in this study paves the way for quantitatively evaluating new tools for metabolomic data analysis, ensuring that the resulting observations are indeed better able to capture the underlying biology of the sample. The proposed GSOM framework can also be easily extended to other high-throughput metabolomics studies.

Acknowledgments The wheat metabolome analysis was supported by the Australian Centre for Plant Functional Genomics, funded by grants from the Australian Research Council and the Grains Research and Development Corporation, the South Australian Government, the University of Melbourne, the University of Adelaide and the University of Queensland and the Victorian Government through the Victorian Centre for Plant Functional Genomics. Work on the development of near-unsupervised learning algorithms was supported by the Australian Research Council (Grant number: DP1096296).

Conflict of interest None declared

References

Abe, T., Kanaya, S., & Ikemura, T. (2009). Batch-learning self-organizing map for predicting functions of poorly-characterized proteins massively accumulated. In: Principe, J., Miikkulainen, R. (eds). *Lecture notes in computer science*, pp. 1–9.

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., & Ikemura, T. (2002). A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: Self-organizing map of oligonucleotide frequency. *Genome Informatics, 13*, 12–20.

Abe, T., Kanaya, S., Kinouchi, M., et al. (1999). Gene classification method based on batch-learning SOM. In: Asai, K., Miyano, S., Takagi, T. (eds) Genome Informatics Series No. 10. Tokyo, Universal Academy Press, pp. 314–315.

Abe, T., Sugawara, H., Kanaya, S., & Ikemura, T. (2006a). A novel bioinformatics tool for phylogenetic classification of genomic sequence fragments derived from mixed genomes of environmental uncultured microbes. *Polar Biosci., 20*, 103–112.

Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., & Ikemura, T. (2006b). Self-organizing map (som) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. *Gene, 365*, 27–34.

Alahakoon, D., Halgamuge, S. K., & Srinivasan, B. (2000). Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks, 11*(3), 601–614.

Bowne, J. B., Erwin, T. A., Juttner, J., et al. (2012). Drought responses of leaf tissues from wheat cultivars of differing drought tolerance at the metabolite level. *Molecular Plant, 5*(2), 418–429.

Chan, C.K.K., & Halgamuge, S.K. (2009). A new generalized growth threshold for dynamic som for comparing average mutual information and oligonucleotide frequency as a species signature. *International Journal of Bio-Science and Bio-Technology, 1*(1), 1–10.

Chan, C. K. K., Hsu, A. L., Halgamuge, S. K., & Tang, S. L. (2008a). Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics, 9*, 215.

Chan, C. K. K., Hsu, A. L., Tang, S. L., & Halgamuge, S. K. (2008b). Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. *Journal of Biomedicine Biotechnology, 2008*, 1–10.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology, 48*, 155–171.

Halgamuge, S. K. (1997). Self-evolving neural networks for rule-based data processing. *IEEE Transactions on Signal Processing, 45*(11), 2766–2773.

Hayashi, H., Abe, T., Sakamoto, M., et al. (2005). Direct cloning of genes encoding novel xylanases from human gut. *Canadian Journal of Microbiology, 51*(3), 251–259.

Hirai, M. Y., Klein, M., Fujikawa, Y., et al. (2005). Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *The Journal of Biological Chemistry, 280*(27), 25590–25595.

Hirai, M. Y., Yano, M., Goodenowe, D. B., et al. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in arabidopsis thaliana. *Proceedings National Academy Sciences of the USA, 101*(27), 10205–10210.

Hsu, A. L., & Halgamuge, S. K. (2003b). Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation. *International Journal of Approximate Reasoning, 32*, 259–279.

Hsu, A. L., Tang, S. L., & Halgamuge, S. K. (2003a). An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data. *Bioinformatics, 19*(16), 2131–2140.

Izanloo, A., Condon, A. G., Langridge, P., Tester, M., & Schnurbusch, T. (2008). Different mechanisms of adaptation to cyclic water stress in two south australian bread wheat cultivars. *Journal of Experimental Botany, 59*(12), 3327–3346.

Kanaya, S., Kinouchi, M., Abe, T., et al. (2001). Analysis of codon usage diversity of bacterial genes with a self-organizing map (som): Characterization of horizontally transferred genes with emphasis on the e. coli O157 genome. *Gene, 276*, 89–99.

Kim, J. K., Bamba, T., Harada, K., Fukusaki, E., & Kobayashi, A. (2007a). Time-course metabolic profiling in *Arabidopsis thaliana* cell cultures after salt stress treatment. *Journal of Experimental Botany, 58*(3), 415–424.

Kim, J. K., Cho, M. R., Baek, H. J., et al. (2007b). Analysis of metabolite profile data using batch-learning self-organizing maps. *Journal of Plant Biology, 50*(4), 517–521.

Kohonen, T. (2001). Self-organizing maps. In: Huang, T., Kohonen, T., Schroeder, M. (eds). *Self-organizing maps*, third, extended 3rd edn, Berlin, Springer.

Ohta, D., Shibata, D., & Kanaya, S. (2007). Metabolic profiling using fourier-transform ion-cyclotron-resonance mass spectrometry. *Analytical and Bioanalytical Chemistry, 389*, 1469–1475.

Preethichandra, D. M. G., Hsu, A., Alahakoon, D., & Halgamuge, S. K. (2002) . A modified dynamic self-organizing map algorithm for efficient hardware implementation. *Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery*.

Sawada, Y., Akiyama, K., Sakata, A., et al. (2009). Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant and Cell Physiology, 50*(1), 37–47.

Uchiyama, T., Abe, T., Ikemura, T., & Watanabe, K. (2004). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nature Biotechnology, 23*, 88–93.

Weber, M., Teeling, H., Huang, S., et al. (2011). Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *The ISME Journal, 5*, 918–928.

Wishart, D. S. (2008). Metabolomics: Applications to food science and nutrition research. *Trends in Food Science & Technology, 19*, 482–493.

Yano, M., Kanaya, S., Altaf-Ul-Aminb, M., Kurokawa, K., Hirai, M. Y., & Kazuki, S. (2006). Integrated data mining of transcriptome and metabolome based on BL-SOM. *Journal of Computer Aided Chemistry, 7*, 125–136.

Zhai, Y. Z., Hsu, A., & Halgamuge, S. K. (2006) . Scalable dynamic self-organising maps for mining massive textual data. *Neural Information Processing*, Vol. 4234 of *Lecture Notes in Computer Science*. Berlin, Springer, pp. 260–267.