Component plane presentation integrated self-organizing map for microarray data analysis

Li Xiao[a,1], Kankan Wang[b,1], Yue Teng[c], Ji Zhang[a,c,*]

[a]*Center for Human Molecular Genetics, MMI, University of Nebraska Medical Center, Omaha, NE, USA*
[b]*State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai, PR China*
[c]*Department of Pathology and Microbiology, University of Nebraska Medical Center, 985454 Nebraska Medical Center, Omaha, NE 68198, USA*

Received 9 December 2002; revised 16 January 2003; accepted 30 January 2003

First published online 19 February 2003

Edited by Gianni Cesareni

Abstract We describe a powerful approach, component plane presentation integrated self-organizing map (SOM), for the analysis of microarray data. This approach allows the display of multi-dimensional SOM outputs of microarray data in multiple sample specific presentations, providing distinct advantages in visual inspection of biological significances of genes clustered in each map unit with respect to each RNA sample. Beneficial potentials of the approach are highlighted by processing microarray data from yeast cells as well as human breast malignancies.

Key words: Microarray data display; Component plane; Self-organizing map

1. Introduction

With the completion of human genome sequencing being rapidly approached, functional genomics is becoming extremely prominent in the field of biology. This is represented by the emergence of DNA microarray technology [1]. The typical microarray methodology is where inserts from tens of thousands of cDNA clones (i.e. probes) are robotically arrayed onto a glass slide and subsequently probed with two differentially labeled pools of RNA (i.e. targets). Typically, the RNA sample is labeled with a fluorescence conjugated nucleotide, such as Cy-3 dUTP or Cy-5 dUTP, and these targets are selected to show a contrast between two states of mRNA activity, such as a normal vs. diseased cells/tissue, wild-type vs. transgenic animal/organism or a general control vs. a series of study samples. The slide is then excited by appropriate wavelength laser beams to generate two 16 bit TIF images. The pixel number of each spot is proportional to the number of fluorescent molecules and hence permits the quantification of the number of target molecules which have hybridized with the spotted cDNA. The differences in intensities of signal at each of the wavelengths reflect the proportion of molecules from the two different target sources that have hybridized to the same cDNA probe.

Since the amount of data generated by each microarray experiment is substantial, potentially equivalent to that obtained through tens of thousands of individual nucleotide hybridization experiments done in the manner of traditional molecular biology (i.e. Northern blots), it is extremely challenging to convert such a massive amount of data into meaningful biological networks. Current efforts toward this direction have primarily focused on clustering and visualization. A commonly applied method is hierarchical clustering, which is an unsupervised clustering algorithm primarily based on the similarity measure between individuals (genes or samples) using a pairwise average linkage clustering [2]. Through the pairwise comparison, this algorithm eventually clusters individuals into a phylogenetic structure, visualized as a tree view. A major drawback of the algorithm is the phylogenetic structure, which is more suitable to true phylogenetic situations such as an evolution process rather than multiple molecular networks in cells. Accordingly, this application may lead to error clustering, particularly with respect to large and complex microarray data. The recently introduced self-organizing map (SOM) [3,4], an artificial intelligent algorithm based on unsupervised learning, appears to be particularly appealing in this regard. This algorithm configures output vectors into a topological presentation of the original multi-dimensional data, producing a SOM in which individuals with similar features are mapped to the same map unit or nearby units. This creates a smooth transition of related individuals to unrelated individuals over the entire map. In addition, this ordered map provides a convenient platform for various visual inspections of large numerical data sets. Although SOM has been utilized by several groups for gene clustering analysis [5–8], many of its beneficial potentials, particularly those of visual inspections, have not yet been explored, which may have led to the underutilization of this powerful data mining tool for microarray data analyses. We therefore describe the use of component plane presentations [9,10], an important visualization tool of SOM, to display SOM outputs of microarray data. Benefits of this approach to microarray analysis are highlighted by processing different microarray data sets including microarray data from single cell organism systems such as yeast [11], and from more complex breast cancer samples [12].

*Corresponding author. Fax: (1)-402-559 4001.
E-mail address: jzhang@unmc.edu (J. Zhang).

[1] These authors contributed equally to this study.

Abbreviations: SOM, self-organizing map

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00156-X

2. Materials and methods

For SOM and its visualizations, we utilized a SOM toolbox built in the Matlab 5 computation environment [13] (http://www.cis.hut.fi/projects/somtoolbox/). Hierarchical clustering was performed using tree view (http://genome-www5.stanford.edu/MicroArray/SMD/restech.html). The yeast diauxic shift data (http://cmgm.stanford.edu/pbrown/explore/) were scaled by logarithm with base 2 before being analyzed by SOM. The breast cancer data (http://www.rii.com/publications/default.htm) contained numerical values of 24 479 genes across 98 tumor samples. These data were firstly filtered to eliminate genes with unreadable values and problematic samples. The remaining expression values of 23 606 genes across 96 tumor samples were then centered before being processed by SOM. The yeast data were processed using 256 (16×16 grids) neurons and the breast cancer data were processed using 400 (20×20 grids) neurons. Each of these neurons is represented by a multi-dimensional (seven and 96 respectively) hexagonal prototype vector. The number of dimensions of the prototype vector is equal to the dimensions of input vectors. The number of input vectors, however, is equal to the number of inputs (the number of genes). The neurons are connected to adjacent neurons by neighborhood relationship, which dictates the topology or structure of the map. The prototype vectors are initiated with random numbers and trained iteratively. Each actual input vector is compared with each prototype vector on the mapping grid based on:

$$\| \vec{x} - \vec{m}_{\mathrm{c}} \| = \min_{i} \{ \| \vec{x} - \vec{m}_{\mathrm{i}} \| \}$$

where $\vec{x}$ stands for input vector and $\vec{m}_{\mathrm{c}}$ for output vector. The best-matching unit (BMU) is defined as the smallest distance between prototype and input vectors. Simultaneously, the topological neighbors around the BMU are stretched towards the training input vector so as to have them updated as denoted by: $\vec{m}_{\mathrm{i}}(t+1) = \vec{m}_{\mathrm{i}}(t) + \alpha(t)[\vec{x}(t) - \vec{m}_{\mathrm{i}}(t)]$ [3]. After iterative trainings, SOM is eventually formed in the format that inputs with similar features are mapped to the same map unit or nearby neighboring units, creating a smooth transition of related individuals over the entire map. Different visualizations, including component plane presentations, presented in this paper were also performed using the SOM Toolbox as mentioned above.

3. Results

To test our approach, we first selected a yeast data set, containing expression values of 6400 genes from RNA samples collected at seven time points before, during and after the yeast diauxic shift [11]. This data set was also previously utilized by other groups for testing their data mining tools, including hierarchical clustering and SOM approaches [2,6]. As shown in Fig. 1A, the SOM outputs were firstly visualized by a bar graphic display. This visualization is similar to previously published methods [5–8], providing a global view of gene clustering, particularly with respect to expression patterns of clustered genes. The number of genes mapped to individual map units varied from 5 to 89 and the bar chart displayed in each hexagonal unit represented the average expression pattern of genes mapped in the unit. It can be seen that the map has been organized in such a way that related patterns are placed in nearby map units, producing a smooth transition of patterns over the entire map. Therefore, a gene cluster can also be recognized from genes represented by closely related neighboring map units in addition to its core unit. Obviously, genes mapped to edge and particularly corner areas appear to be mostly regulated during the diauxic shift, while genes in a large area near the central part of the map appear to be less regulated, as also suggested previously [6]. Inserts in the right panel of Fig. 1A detail patterns of the four corner map units. Named genes mapped to these units are listed in Table 1.

To further reveal features other than expression patterns of clustered genes, we introduced a powerful visualization approach known as component plane presentation [9,10]. This approach allows the illustration of SOM outputs in multiple, vector component (sample) specific presentations. Each of these presentations illustrates values of a single vector component in all map units. For instance, the first presentation (R1) in Fig. 1B shows the SOM values of all map units at the time point of 9 h and last one (R7) shows the SOM values of all map units at 21 h during the diauxic shift [11]. Interestingly, each of these presentations also appears as a sample specific, genome-wide transcriptional display, in which all up-regulated units (hexagons in red), down-regulated units (hexagons in blue), and moderately transcribed units (hexagons in green and yellow) are well delineated. It is straightforward to determine functional significances of genes regulated at each time point during the diauxic shift. By comparing these presentations, we can also learn many additional features. For instance, these presentations are sequentially correlated with each other, depicting the process of metabolic change from fermentation to respiration at the genome-wide scale. The sequential inactivation of genes mapped to two upper corners suggests that the functional group represented by genes on the left is more sensitive to the depletion of glucose than the one on the right, although both of them are suppressed toward the end of the diauxic shift. The sequential activation of genes mapped to the two bottom corners even gives us a more vivid picture of the process leading to ethanol consumption in the yeast cells. Genes in the bottom left corner are particularly activated at the end of the shift, indicating that these genes are specifically associated with ethanol metabolism, whereas the progressively increased expression level of genes in the right corner suggests that these genes are associated not only with ethanol metabolism but also with glucose consumption. This is confirmed by known genes mapped to these corner units (Table 1). It is clearly shown that genes represented by the upper two corner units (C1 and C16) are mostly cell growth and protein synthesis related. Particularly, genes grouped in C16 are almost exclusively ribosome encoding genes, whereas genes in the bottom left corner are specifically involved in ethanol metabolism, including the glyoxylate cycle. Genes in the bottom right corner are involved in glucose metabolism, including the tricarboxylic acid cycle, in addition to some stress activated heat shock protein and cytochrome *c* related genes. Of course, glucose pathways and the tricarboxylic acid cycle are also utilized during ethanol metabolism. As compared with previous analyses of this set of data [6,11], our results appear to be more targeted, complete and meaningful (Table 1).

To validate that our approach is also applicable to larger and more complex microarray data, we selected a recently published breast cancer data set [12], containing readable values of 23 606 genes from RNA samples of 96 individual tumors. Cancer, a heterogeneous population, with the same stage of disease can have markedly different treatment responses and overall outcome. Accordingly, microarray based investigation may help to identify previously unrecognized and clinically significant subtypes of tumor and thus develop more sophisticated clinical protocols [12–14]. Since the previous analysis of this set of data was primarily based on hierarchical clustering, we speculated that the application of our approach would generate additional information. SOM was

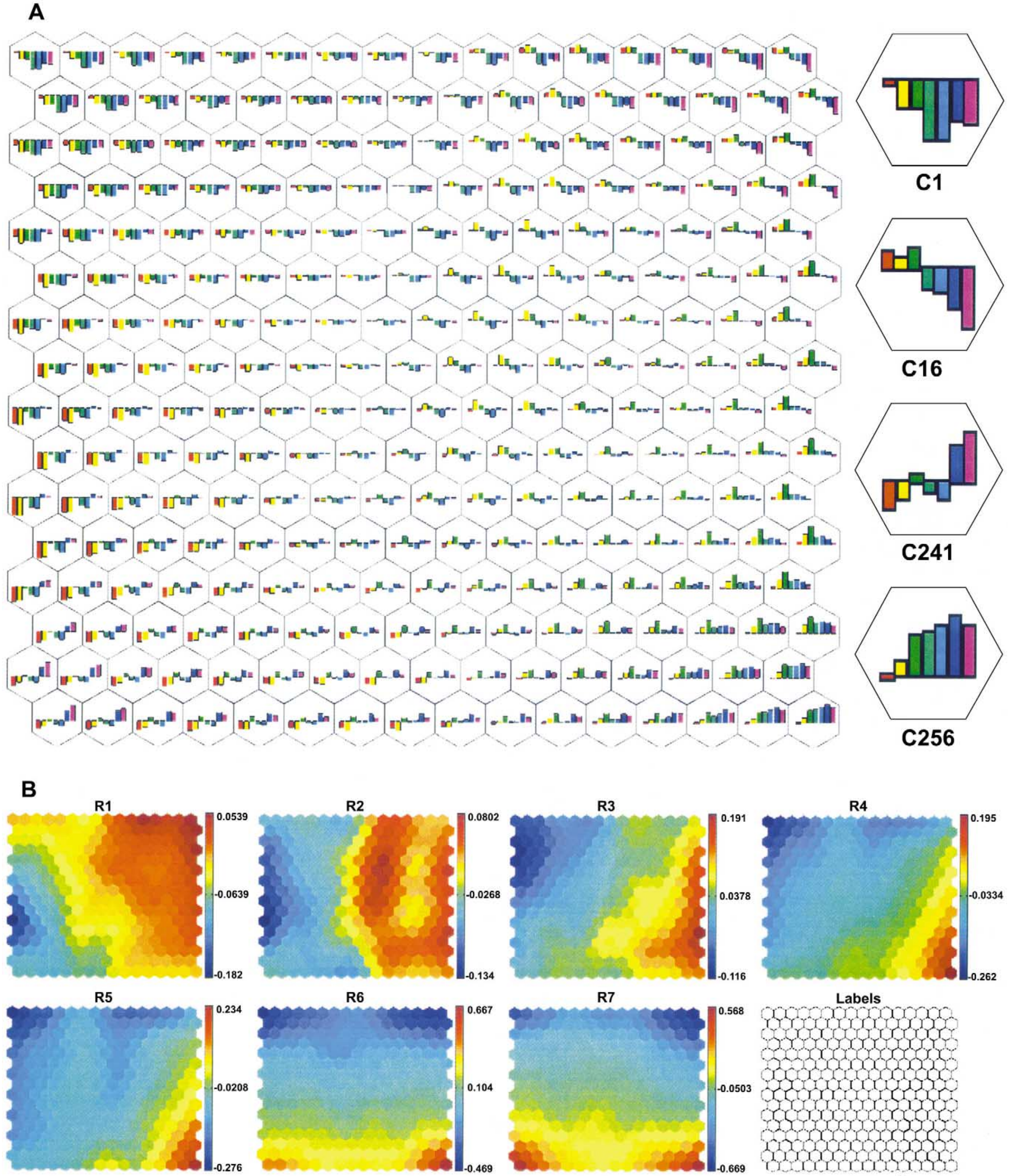

Fig. 1. SOM outputs of the yeast diauxic shift data. A: A bar graphic display illustrating expression patterns of clustered genes in all map units. The bar chart in each unit represents the average expression values of genes mapped to the unit across all seven samples. Inserts on the right panel detail four corner map units: upper left (C1), upper right (C16), bottom left (C241) and bottom right (C256) respectively. B: Component plane presentations (R1 to R7) depicting genome-wide transcriptional changes from fermentation to respiration during the diauxic shift. Color coding index stands for the expression values of genes. The brighter the color, the higher the value. The color index of each display is established on the basis of all the values of a single component plane. All these presentations are linked by position: in each display, the hexagon in a certain position corresponds to the same map unit. The label display shows positions of each unit on the map.

Table 1
Genes clustered in the four corner units

	ORF	Name	Description
C1	YBR247C	ENP1	*N*-glucosylation protein
C1	YCR053w	THR4	threonine synthase (*o-p*-homoserine *p*-lyase)
C1	YCR065w	HCM1	transcription factor
C1	YDL037c		putative glucan 1,4-α-glucosidase
C1	YDL182w		putative homocitrate synthase or isopropylmalate synthase
C1	YDL213c		putative RNA binding protein
C1	YDR144C	MKC7	aspartyl protease of the periplasmic space
C1	YEL026w		homology to high mobility group-like protein Nhp2p
C1	YEL046c	GLY1	putative aminotransferase
C1	YGR034W	RPL33B	ribosomal protein
C1	YIL053W	GPP1	DL-glycerol phosphatase
C1	YIL069C	RP50B	ribosomal protein S24.e
C1	YJL189W	RPL46	ribosomal protein L39.e
C1	YJL194W	CDC6	cell division control protein
C1	YKL009W		similarity to Rpl10p and *Sulfolobus solfataricus* ribosomal protein L10
C1	YKL181W	PRPS1	ribose-phosphate pyrophosphokinase
C1	YKL191W	DPH2	diphtheria toxin resistance protein
C1	YLL008w	DRS1	RNA helicase of the DEAD box family
C1	YLL012W		similarity to triacylglycerol lipases
C1	YLR009W		similarity to ribosomal protein L24.e.B
C1	YLR056w	ERG2	C-5 sterol desaturase
C1	YLR129w	DIP2	Dom34p interacting protein
C1	YLR179C		similarity to Tfs1p and Nsp1p
C1	YLR214W	FRE1	ferric (and cupric) reductase
C1	YLR355C	ILV5	ketol-acid reducto-isomerase
C1	YMR058W	FET3	cell surface ferroxidase for high affinity ferrous iron uptake
C1	YMR108W	ILV2	acetolactate synthase
C1	YMR239C	RNT1	double-stranded ribonuclease
C1	YMR290C		putative helicase
C1	YNL002C	RLP7	ribosomal protein L7.e
C1	YNL111C	CYB5	cytochrome b_5
C1	YNL132W		homology to *A. ambisexualis* antheridiol steroid receptor
C1	YNL141W		similarity to adenosine deaminase
C1	YNL216W	RAP1	DNA binding protein with repressor and activator activity
C1	YNL256W		similarity to bacterial dihydropteroate synthase
C1	YNL327W	EGT2	involved in cell separation in G1
C1	YOR116C	RPO31	DNA directed RNA polymerase III 160K chain
C1	YOR361C	PRT1	eIF3 initiation factor complex subunit
C1	YPL226W		putative translocation elongation factor eEF-3
C16	YAL012W	CYS3	cystathionine γ-lyase
C16	**YAL038W**	**PYK1**	**pyruvate kinase**
C16	YAR073W	FUN63	homology to Pur5p
C16	YBR181C	RPS101	ribosomal protein S6.e
C16	YBR189W	SUP46	ribosomal protein S9.e.B
C16	YBR191W	URP1	ribosomal protein L21.e
C16	YDL136w		putative ribosomal protein
C16	YDL208W	NHP2	putative ribosomal protein
C16	YDR418W	RPL15A	ribosomal protein
C16	YEL054c	RPL15A	ribosomal protein L12.e.a
C16	**YGL103W**	**CYH2**	**ribosomal protein**
C16	YGR148C	RPL30B	ribosomal protein
C16	YGR214W	NAB1A	40S ribosomal protein p40 homolog A
C16	YHL015W	URP2	ribosomal protein
C16	YHL033C	RPL4A	60S ribosomal protein L7A-1
C16	**YHR203C**	**RPS7A**	**ribosomal protein S4.e**
C16	YHR216W	PUR5	IMP dehydrogenase
C16	**YIL018W**	**RPL5A**	**ribosomal protein L8.e**
C16	**YIL052C**		**ribosomal protein L34.e**
C16	**YIL133C**	**RP22**	**ribosomal protein**
C16	**YJL136C**	**RPS25B**	**ribosomal protein S21.e**
C16	YJL177W	RPL20B	ribosomal protein L17.e.c10
C16	**YJL190C**	**RPS24A**	**ribosomal protein S15a.e**
C16	YJR123W	RPS5	ribosomal protein S5.e
C16	YJR145C	RPS7B	ribosomal protein S4.e.c10
C16	**YKR057W**	**RPS25A**	**ribosomal protein S21.e**
C16	**YKR059W**	**TIF1**	**translation initiation factor 4A**
C16	YLL045c	RPL4B	ribosomal protein L7a.e.B
C16	YLR048w	NAB1B	40S ribosomal protein p40 homolog b
C16	**YLR175W**	**CBF5**	**centromere/microtubule binding protein**
C16	YLR249W	YEF3	translation elongation factor eF-3
C16	**YLR325C**		**putative ribosomal protein L38**
C16	YLR340W	RPLA0	acidic ribosomal protein L10.e

Table 1 (*Continued*).

	ORF	Name	Description
C16	YLR432W		homology to IMP dehydrogenases, Pur5p and YM9958.06c
C16	**YML063W**	**RP10B**	**ribosomal protein**
C16	**YNL069C**	**RP23**	**ribosomal protein**
C16	YNL096C		homology to ribosomal protein S7
C16	YNL301C	RP28B	ribosomal protein L18.e
C16	YNR053C		homology to human breast tumor associated autoantigen
C16	YOL120C	RP28A	ribosomal protein
C16	**YOL121C**	**RPS16B**	**ribosomal protein S19.e**
C16	YOL040C	RPS21	ribosomal protein
C16	YOR063W	TCM1	ribosomal protein L3.e
C16	YOR310C		homology to SIK1 protein
C16	YOR312C	RPL18B	ribosomal protein
C16	YPL131W	RPL1	ribosomal protein L5.e
C16	YPL220W	SSM1A	ribosomal protein
C241	**YAL054C**	**ACS1**	**acetyl-CoA synthetase**
C241	YBR067C	TIP1	cold and heat shock induced protein of the Srp1/Tip1p family
C241	**YBR117C**	**TKL2**	**transketolase 2**
C241	YBR298C	MAL3T	maltose permease
C241	YCL025C		putative amino acid transport protein
C241	**YCR005c**	**CIT2**	**peroxisomal citrate (si)-synthase**
C241	YDL085w		putative NADH dehydrogenase (ubiquinone)
C241	YDL223c		putative microtubule binding protein
C241	**YEL012w**	**UBC8**	**ubiquitin conjugating enzyme**
C241	YER024w		similarity to carnitine *O*-acetyltransferase Yat1p
C241	**YER065c**	**ICL1**	**isocitrate lyase**
C241	**YHR096C**	**HXT5**	**putative hexose transporter**
C241	YJL045W		homology to succinate dehydrogenase flavoprotein
C241	**YJL137C**	**GLG2**	**self-glucosylating initiator of glycogen synthesis**
C241	YJR048W	CYC1	cytochrome *c* isoform 1
C241	YJR095W	ACR1	regulator of acetyl-CoA synthetase activity
C241	YKL093W	MBR1	required for optimal growth on glycerol
C241	**YKR097W**	**PPC1**	**phosphoenolpyruvate carboxykinase**
C241	YLR164W		putative succinate dehydrogenase
C241	**YLR174W**	**IDP2**	**cytoplasmic isocitrate dehydrogenase (NADP$^+$)**
C241	**YLR377C**	**FBP1**	**fructose-1,6-bisphosphatase**
C241	YML042W	CAT2	carnitine *O*-acetyltransferase
C241	YML054C	CYB2	lactate dehydrogenase cytochrome b_2
C241	YML120C	NDI1	NADH-ubiquinone-6 oxidoreductase
C241	YNL009W		homology to isocitrate dehydrogenase
C241	**YNL117W**	**MLS1**	**malate synthase 1**
C241	YOR100C		homology to mitochondrial carrier protein YMC1
C241	YPL135W		homology to nitrogen fixation protein (nifU)
C241	YPL262W	FUM1	fumarate hydratase
C241	YPR184W		similarity to glycogen debranching enzyme (4-α-glucanotransferase)
C256	YBL015W	ACH1	acetyl-CoA hydrolase
C256	YBL045C	COR1	ubiquinol–cytochrome *c* reductase 44K core protein
C256	YBL064C		similarity to thiol specific antioxidant enzyme
C256	**YBR072W**	**HSP26**	**heat shock protein**
C256	YBR139W		homology to carboxypeptidase
C256	YBR169C	SSE2	heat shock protein of the HSP70 family
C256	YCL035C		homology to glutaredoxin
C256	**YCR021c**	**HSP30**	**heat shock protein**
C256	YDL022w	GPD1	glycerol-3-phosphate dehydrogenase (NAD$^+$) precursor
C256	**YDR001C**	**NTH1**	**neutral trehalase (α,α-trehalase)**
C256	YDR077W	SED1	abundant cell surface glycoprotein
C256	**YDR171W**	**HSP42**	**heat shock protein with similarity to Hsp26p**
C256	**YDR178W**	**SDH4**	**succinate dehydrogenase membrane anchor subunit for sdh2p**
C256	YDR258C	HSP78	mitochondrial heat shock protein of clpb ATP-dependent proteases
C256	YDR342C	HXT7	high affinity hexose transporter
C256	YDR343C	HXT6	high affinity hexose transporter
C256	YDR513W	TTR1	glutaredoxin
C256	YDR529C	QCR7	ubiquinol–cytochrome *c* reductase subunit 7
C256	**YEL011w**	**GLC3**	**1,4-glucan branching enzyme (glycogen branching enzyme)**
C256	YEL024w	RIP1	ubiquinol–cytochrome *c* reductase iron–sulfur protein precursor
C256	YER053c		homology to mitochondrial phosphate carrier protein
C256	**YFL014W**	**HSP12**	**heat shock protein**
C256	**YFR015C**	**GSY1**	**UDP glucose–starch glucosyltransferase 1**
C256	YFR033C	QCR6	ubiquinol–cytochrome *c* reductase 17K protein
C256	YGL187C	COX4	cytochrome *c* oxidase chain IV
C256	YGL191W	COX13	cytochrome *c* oxidase chain VIa
C256	YGR008C	STF2	ATPase stabilizing factor
C256	YGR043C		putative transaldolase
C256	**YGR088W**	**CTT1**	**cytosolic catalase T**

Table 1 (*Continued*).

	ORF	Name	Description
C256	YGR244C		putative β-succinyl CoA synthetase
C256	YHR051W	COX6	cytochrome *c* oxidase subunit VI
C256	YIL111W	COX5B	cytochrome *c* oxidase chain Vb
C256	**YIL125W**	**KGD1**	**α-ketoglutarate dehydrogenase**
C256	YIL136W	OM45	protein of the outer mitochondrial membrane
C256	YIR039C		similarity to Yap3p
C256	YJR073C	PEM2	methylene fatty acyl phospholipid synthase
C256	YJR096W		similarity to *Leishmania* reductase
C256	YKL026C		homology to glutathione peroxidase
C256	**YKL085W**	**MDH1**	**mitochondrial malate dehydrogenase precursor**
C256	YKL103C	LAP4	vacuolar aminopeptidase yscl precursor
C256	YKL109W	HAP4	transcriptional activator
C256	**YKL141W**	**SDH3**	**cytochrome b_{560} subunit of respiratory complex II**
C256	**YKL148C**	**SDH1**	**succinate dehydrogenase flavoprotein precursor**
C256	YKL217W	JEN1	carboxylic acid transporter protein
C256	YLL026w	HSP104	heat shock protein
C256	**YLL041c**	**SDH2**	**succinate dehydrogenase iron–sulfur protein subunit**
C256	YLR178C	TFS1	cdc25-dependent nutrient and ammonia response cell cycle regulator
C256	**YLR258W**	**GSY2**	**UDP-glucose–starch glucosyltransferase, isoform 2**
C256	**YLR304C**	**ACO1**	**aconitate hydratase**
C256	YLR327C		homology to STF2 protein
C256	YLR345W		similarity to 6-phosphofructo-2-kinase
C256	YLR356W		similarity to SCM4 protein
C256	YML100W	TSL1	α,α-trehalose phosphate synthase (UDP forming) subunit
C256	YMR081C	MBR1	with Nam7p/Upf1p in suppression of mitochondrial splicing defect
C256	**YMR105C**	**PGM2**	**phosphoglucomutase, major enzyme**
C256	YMR110C		putative aldehyde dehydrogenase
C256	YMR170C	ALD2	mitochondrial aldehyde dehydrogenase 2 (NAD$^+$)
C256	YMR250W		putative glutamate decarboxylase
C256	YMR297W	PRC1	carboxypeptidase Y (CPY) (YSCY), serine-type protease
C256	YNL015W	PBI2	proteinase B inhibitor 2
C256	YNL134C		homology to *Cochliobolus carbonum toxD* gene
C256	YNL160W	YGP1	secreted glycoprotein
C256	YNL173C		pheromone response G protein
C256	YNL274C		similarity to dehydrogenases
C256	**YNR001C**	**CIT1**	**citrate (si)-synthase**
C256	YOR065W	CYT1	cytochrome c_1
C256	YOR178C	GAC1	regulatory subunit for protein Ser/Thr phosphatase Glc7p
C256	YOR374W		putative aldehyde dehydrogenase
C256	YPL154C	PEP4	aspartyl protease
C256	YPR149W	NCE2	involved in non-classical protein export pathway

Genes in bold are those previously reported [6,11].

then performed using 400 (20×20 grids) hexagonal prototype vectors and the component outputs were further organized by hierarchical clustering. In the dendrogram shown in Fig. 2A, the length and the subdivision of the branches depict the relationship between tumors, where the shorter the branch the more similarity there is between tumors. It is clearly shown that these 96 samples are separated into two distinct categories, representing 32 and 64 tumors respectively. These two categories are equivalent to the previously described ER negative and ER positive groups [12]. Seventeen out of 18 tumor samples with germ line BRCA1 mutations are found in the first category and 62 out of 64 ER positive cases are found in the second category. Obviously, samples in the first category are all grade 3 tumors and most of them are lymphocytic infiltrate. In addition, subdivisions of the tree branches in our dendrogram appear to form distinct clusters in each of the two categories, suggesting that tumor samples in both categories can be further divided into subcategories based on the SOM outputs of the 23 606 genes. This observation is fully elucidated in the following component plane presentations shown in Fig. 2B. Each presentation illustrates transcriptional changes of a specific tumor sample at the genome-wide scale, exhibiting characteristic patterns. By comparing patterns in identical positions between presentations, we can recognize tumors potentially belonging to the same clinically significant type, i.e. with the same or similar transcriptional mechanisms. In addition, we can simultaneously recognize genes potentially important for the type of tumor. For instance, tumor samples classified into the first subcategory (i.e. S44, S92, S89, S75, S86, S65, S100, S50, S20, S98, S83 and S97) of the ER negative category display similar patterns across many identical positions, particularly with respect to the mostly up-regulated units in the top left corner and down-regulated units in the bottom right corner. This further indicates that these tumors may belong to the same pathologically significant subtype, sharing the same or similar molecular mechanisms underlying the genesis of the tumor, while commonly regulated genes, particularly those mapped to the top left and bottom right corner units, may represent potentially important genes whose regulation is strongly associated with the type of tumor. Considering the heterogeneity of tumor tissue used in microarray hybridization and individual variation, it is logical to expect that some of the patterns, particularly those occurring sporadically, may actually symbolize noise interruptions. Therefore, it is critical to compare patterns between presentations and hence identify commonly

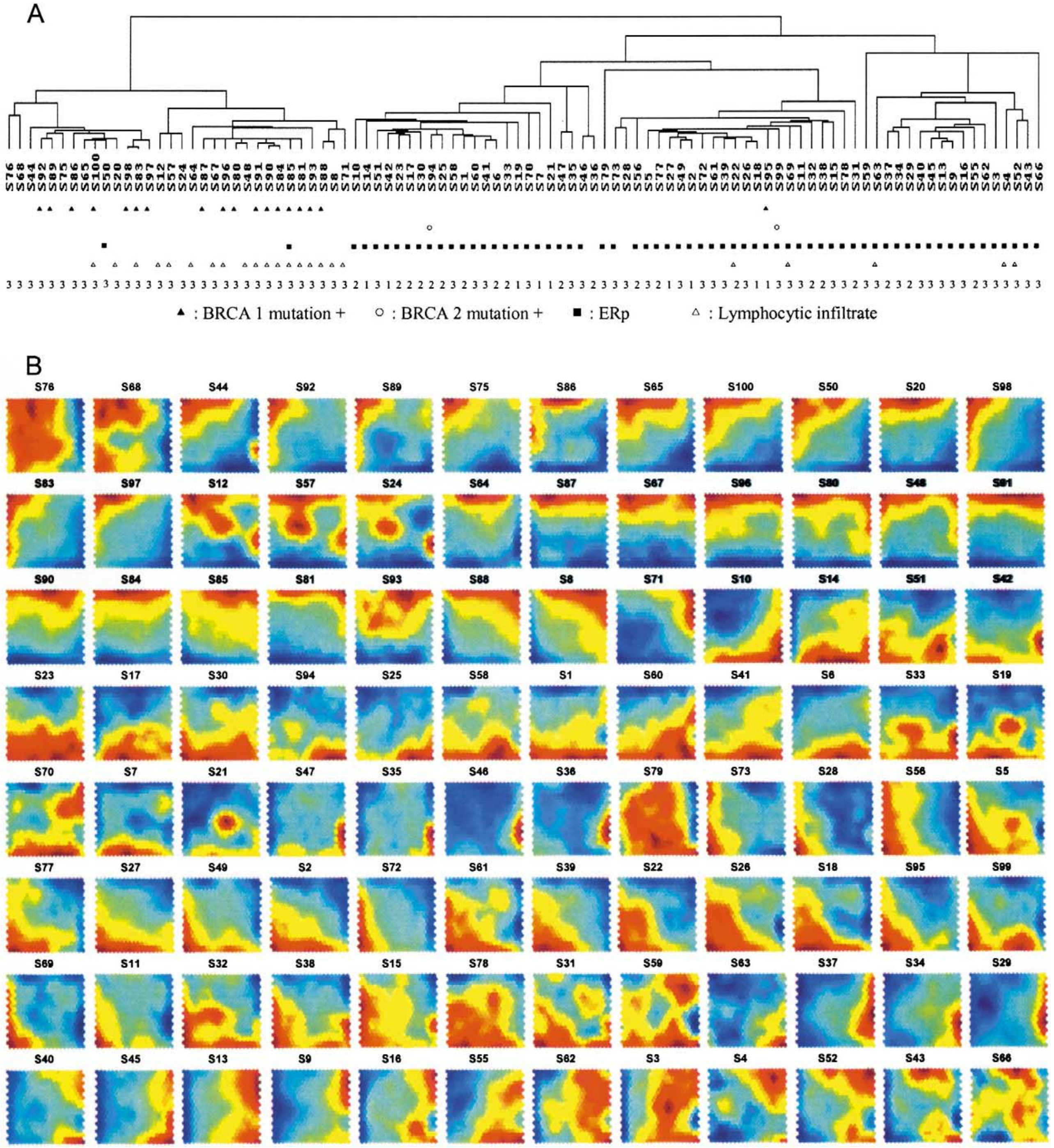

Fig. 2. Component plane presentation based classification of human breast malignancies. A: Hierarchical clustering of component plane presentations of 96 tumor samples. Each presentation is represented by a specific case number under each tree branch. Molecular and clinical descriptions of the tumors are marked using different symbols as indicated underneath. B: Reordered component plane presentations of 96 tumors, permitting direct visual comparison of transcriptional similarities between tumors at the genome-wide scale. As described above, color coding index stands for the expression values of genes. The brighter the color, the higher the value. Color coding scales and label display were omitted.

or recurrently regulated map units so as to find pathologically important genes specifically for the tumor type. It is obvious that breast cancer is a highly heterogeneous population with multiple subtypes of tumors. Therefore, additional efforts are needed to establish the detailed tumor classification and identify subtype specific signature molecules, which may eventually help us to improve current diagnostic, prognostic and therapeutic protocols in this malignancy.

4. Discussion

The component plane presentation integrated SOM is a powerful approach for the analysis of microarray data. This has been demonstrated by analyzing microarray data from yeast cells as well as human breast tumors. In particular, component plane presentations permit in-depth visualization of vector component variables that contribute to SOM and thus allow the display of multi-dimensional SOM outputs of microarray data in multiple, sample specific presentations, in which all the regulated genes are well delineated. These presentations therefore facilitate the visual inspection of functional significances of genes mapped to each unit with respect to each sample, providing distinct advantages for us to understand microarray data. As demonstrated in the processing of the yeast diauxic shift data, each presentation illustrates genome-wide transcriptional changes of a specific stage and hence the sequentially correlated presentations depict the entire process of metabolic changes from fermentation to respiration at the transcriptional level. Beneficial potentials of the component plane presentation integrated SOM approach are further demonstrated in the processing of the breast cancer data. Breast cancer, like many other human malignancies, is a highly heterogeneous population with multiple tumor types. Accordingly, the microarray data generated from expressions of 23 606 genes across 96 tumor samples is not only large but also complex. Considering the heterogeneity of tumor tissue and individual variation, the complexity of the tumor classification data can go even beyond our imagination. Using hierarchy based methods to categorize genes or samples on the basis of such data is apparently not adequate and probably problematic as well, whereas our approach appears to be particularly appealing in this regard. It permits the direct visualization of transcriptional changes of each tumor sample at the genome-wide scale, providing unequivocal advantages for us to inspect the tumor classification data in detail. For instance, by comparing different presentations, we can recognize tumors potentially belonging to the same clinically significant subtype, i.e. with similar transcriptional changes in identical positions. By targeting commonly or recurrently regulated units within the type of tumors, we can identify potentially relevant genes. It is straightforward and easy to interpret. It is obvious that the approach described in this paper has distinct advantages over commonly applied hierarchy based methods for the analysis of microarray data, particularly with respect to visual advantages provided by the approach. We believe the approach is applicable to any kind of microarray/genechip data. We also believe the potential impact of the approach on gene expression based investigations will be substantial.

Acknowledgements: We thank Drs. Kirk Beisel and Wing C. Chan for their helpful comments and suggestions on the manuscript. We also thank Mr. Yu Liu for his artistic support.

References

[1] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Science 270, 467–470.

[2] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14868.

[3] Kohonen, T. (1995) Self-organizing Maps, Springer Series in Information Sciences, Vol. 30, Springer, Berlin.

[4] Kohonen, T., Oja, E., Simula, O., Visa, A. and Kangas, J. (1996) Proc. IEEE 84, 1358–1384.

[5] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Proc. Natl. Acad. Sci. USA 96, 2907–2912.

[6] Toronen, P., Kolehmainen, M., Wong, G. and Castren, E. (1999) FEBS Lett. 451, 142–146.

[7] Chen, J.J., Peck, K., Hong, T.M., Yang, S.C., Sher, Y.P., Shih, J.Y., Wu, R., Cheng, J.L., Roffler, S.R., Wu, C.W. and Yang, P.C. (2001) Cancer Res. 61, 5223–5230.

[8] White, K.P., Rifkin, S.A., Hurban, P. and Hogness, D.S. (1999) Science 286, 2179–2184.

[9] Vesanto, J. (1999) Intell. Data Anal. 3, 111–126.

[10] Vesanto, J. (2000) in: Neural Network Tool for Data Mining: SOM Toolbox. Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems, Oulun yliopistopaino, Oulu, pp. 184–196.

[11] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Science 278, 680–686.

[12] van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Nature 415, 530–536.

[13] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S. and Golub, T.R. (2001) Proc. Natl. Acad. Sci. USA 98, 15149–15154.

[14] Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, L., Glatfelter, A., Pollock, P., Gillanders, E., Leja, D., Dietrich, K., Berens, M.C., Alberts, D., Sondak, V., Hayward, H. and Trent, J. (2000) Nature 406, 536–540.