Chapter 9

Using Self-Organizing Maps to Visualize, Filter and Cluster Multidimensional Bio-Omics Data

Ji Zhang and Hai Fang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51702

1. Introduction

In the face of ever-growing of biological data at the genome scale (denoted as omics data) [1,2], investigators of virtually every aspect of biological research are shifting their attention to massive information extracted from omics data. The 'omics' refers to a complete set of biomolecules, such as DNAs, RNAs, proteins and other molecular entities. Omics data are produced by high-throughput technologies. At first, these technologies were known as cDNA microarray [3] and oligonucleotide chips [4]. Then, they were diversely evolved into ChIP-on-Chip [5] and ChIP-Sequencing [6,7], two-dimensional gel electrophoresis and mass spectrometry [8] and high-throughput two-hybrid screening [9]. Recently, they are highlighted by next-generation sequencing technologies such as DNA-seq [10] and RNA-seq [11]. Because of these technological advances, biological information can be quantified in parallel and on a genome scale, but at a much-reduced cost. Nearly, omics data cover every aspect of biological information and thus secure the studies being carried out from a genome-wise perspective. To name but a few examples, they can be used (i) to catalog the whole genome within a living organism (genomics), (ii) to monitor the gene expression at RNA level (transcriptomics) or at protein level (proteomics), (iii) to study the protein-protein interactions (interactomics) and transcription factor-DNA binding patterns (regularomics), and (iv) to characterize DNA or histone modifications exerting on the chromosomes (epigenomics). These multi-layer omics data not just constitute a global overview of molecular constituents, but also provide an opportunity for studying biological mechanisms. In contrast to conventional reductionism focusing on individual biomolecules, omics approaches allow the study of emergent behaviors of biological systems. This conceptual advance has led to the advent of systems biology [12], an interdisciplinary research field with the ultimate goal of *in silico* modeling of biological systems.

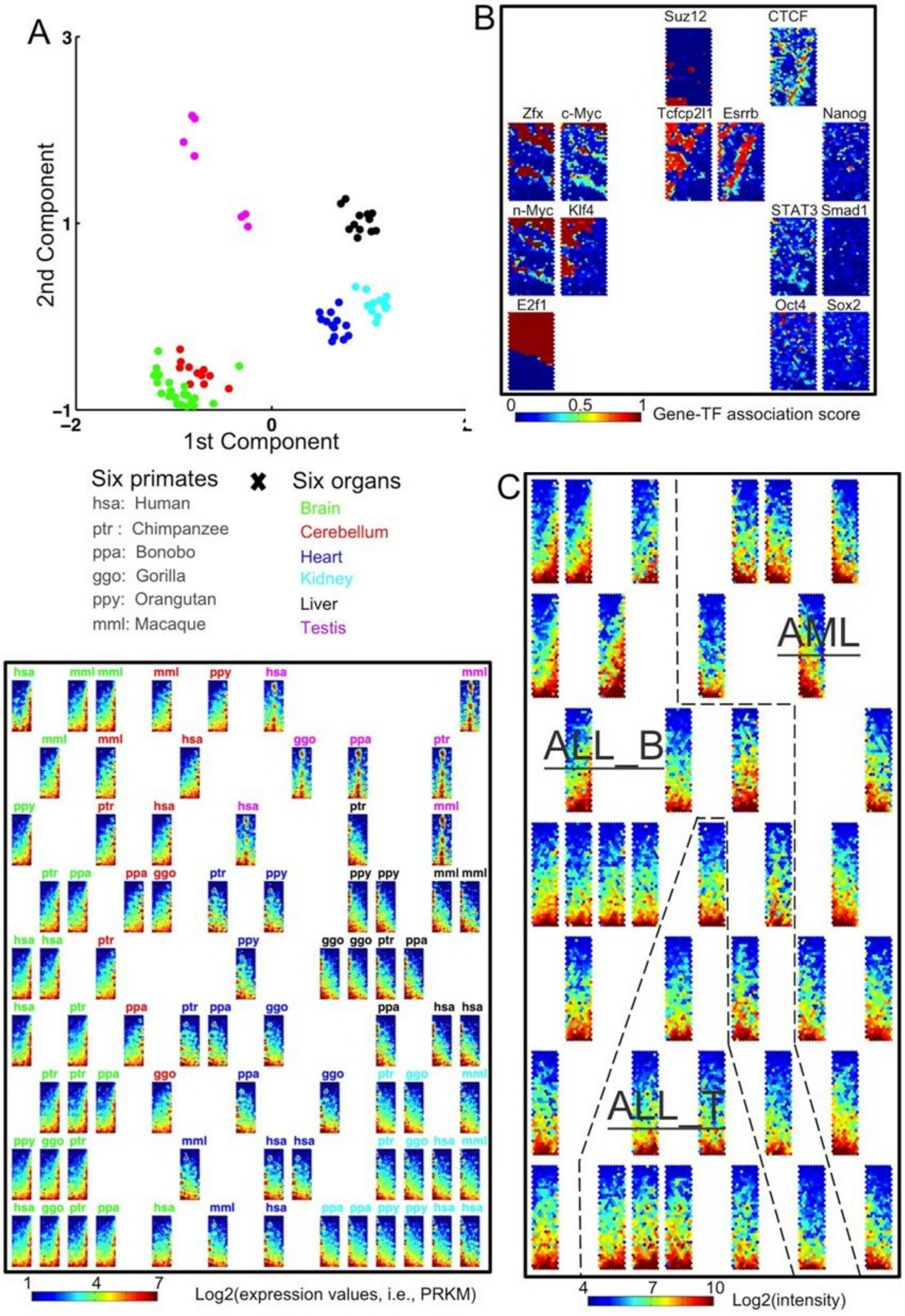

Figure 1. Reanalysis of three different sets of omics data by the reorganized CPPs. **(A)** Transcriptome evolution in mammalian organs. Sammon mapping onto the first two components is displayed in the top panel. Each dot corresponds to one of 36 samples, color-encoded based on their organ origins for the better visualization. The reorganized CPPs are shown in the bottom panel. Each component plane illustrates the sample-specific transcriptome map and is placed within a two-dimensional rectangular lattice (framed in black). Within each component plane, genes with the same or similar expression patterns are mapped to the same or nearby map nodes. When zooming out to look at between-planes/samples relationships, samples with the similar expression profiles are placed closer to each other. The title above each plane is texted in abbreviation and marked in color. The meanings of these abbreviations and colors are described in the middle panel. **(B)** Regularome of multiple transcription factors in embryonic stem cells. The reor-

ganized CPPs not only display regularome of each of 14 transcription factors, but also reveal their relationships by geometric closeness within the two-dimensional rectangular lattice. **(C)** Transcriptome profiling in cancer classification. The transcriptome similarities and distinctions among 38 leukemia samples are visualized by the reorganized CPPs. The dotted lines are used to intuitively indicate the boundary between the AML-ALL separation, and within the ALL, the boundary between its two subtypes (i.e., the ALL_B and ALL_T). Since each sample class occupies distinctive regions within the two-dimensional rectangular lattice, the sample labels are texted uniformly as indicated. AML: acute myeloid leukemia; ALL: acute lymphoblastic leukemia; ALL_B: B-cell ALL; ALL_T: T-cell ALL.

Today, all areas of biological science are confronted with ever-increasing amounts of omics data whereas interpretations of the data appear to lag far behind the rate of data accumulation [13]. It is largely due to a lack of understanding the complexity of the data, and is also partially explained by algorithms being applied inappropriately. For example, transcriptome data are tabulated as gene expression matrix, measuring expression levels of genes against experimental samples. Two factors limit the power of many conventional multivariate statistical methods. First, gene expression matrix contains data with low signal-to-noise ratio and missing values as well. Second, such matrix usually involves tens of thousands of genes but a much smaller number of samples, known as ‘small sample sizes relative to huge gene volumes’. To overcome the limitations of conventional algorithms, bringing human intelligence into the data processing represents a crucial factor for the discovery of *bona fide* relationships between genes or samples, in which visual control is indispensible. Interestingly, early pioneered efforts on transcriptome data mining were primarily focused on data organization and visualization [14,15].

Visual inspection represents a crucial aspect in omics data mining, providing many potential benefits. However, such potential benefits are largely limited by using conventional algorithms such as hierarchical and K-mean clustering. Instead, we use the vector space model to conceptually express omics data. This model allows biological molecules (e.g., genes) to be automatically organized into data clouds in the virtual reality environment based on their numerical values across all samples tested. Take transcriptome data as an example, wherein each gene activity pattern (e.g., gene expression pattern) across N related samples could be referred to as a data point in an N-dimensional hyperspace. Tens of thousands of such data points would therefore form data clouds in the space. Accordingly, the methods used for the gene clustering or the projection/visualization of output results should respect the ‘natural’ structure of input expression matrix, that is, to preserve the shape and density (collectively called ‘topological structure’) of the data. Notably, the more similar activity the genes exhibit, the closer geometric space they occupy. Exploring geometric relationships in a topology-preserving manner provides a natural basis for discovering biologically meaningful knowledge. Such topological preservation is of particular significance at the exploratory phase of omics data mining since *a priori* knowledge of the data structure is usually unknown.

The self-organizing map (SOM), as a learning algorithm [16], appears to be suitable for topology-preserving analysis of multi-dimensional data. In an interactive manner, the SOM summarizes the input data by vector quantization (VQ) and simultaneously carries out topological preserving projection by vector projection (VP). More importantly, optimization of neighborhood kernels may control the extent to which the VP influences the VQ. For the sake of human-centric visualization, this algorithm usually produces a regular two-dimensional hexagonal grid of map nodes. Each map node is associated with a prototype vector in the high-dimensional space, collectively forming the codebook matrix. In terms of gene activity matrix (such as gene expression matrix) as input, the SOM produces a map, wherein (i) genes with the same or similar activity patterns (i.e., gene activity vectors) are mapped to the same or nearby map nodes, (ii) the density of genes mapped to this two-dimensional map follows the data density in the high-dimensional space. When all map nodes are color-encoded according to values in each component of prototype vectors, the resulting component map (or called 'component plane' due to a regular shape of the map [17]) can be used as a sample-specific presentation of gene activities. Based on this scheme, we have applied a method of component plane presentations (CPPs) to visualize microarray data analysis [18]. In essence, the CPPs take advantage of the visual benefits of the ordered SOM map to illustrate the codebook matrix in a sample-specific fashion.

In addition, we here aim to formally introduce a SOM-centric analytical pipeline, an extension to our previously proposed approaches [19], for in-depth mining of biological information. At the core is the plasticity of SOM neighborhood kernels in preserving local versus global topology of the input data to a varied degree. The remainder of this chapter is organized as follows. First, we will introduce the reorganized CPPs, originally called 'component plane reorganization' for correlation hunting [20], and illustrate the visual benefits in characterizing omics data of various types. Then, we will give a timeline review of the SOM in the context of its past applications in omics data. After that, we will focus on our proposed pipeline. Through a representative real-world case of transcriptome changes during early human organogenesis, we will provide a tutorial overview of how this pipeline can be used for the simultaneous visualizations of genes and samples, topology-preserving gene selection and clustering, and the temporal expression-active sub-network detection. Finally, we will conclude this chapter along with the future directions of the pipeline for the further developments.

2. The reorganized CPPs and the potential benefits for visualizing omics data of various types

As demonstrated in many applications [21-28], the CPPs enable straightforward and widespread use. They allow users to interpret omics data in a sample-specific fashion but without loss of information on tens of thousands of genes (still visible but being clustered and orderly organized). Very often users tend to mistake CPPs as microarray chips. It suggests the importance of sample-specific visualization of omics data from the biologists' point of view. Instead of the correction, we can further interpret the CPPs as a related set of microarray chips, in which probes (representing genes to be measured) are artificially reconfigured according to their patterns. Such metaphor might increase the circulation of the CPPs and thus the SOM within the omics community. Another way for increasing the circulation is to further improve the CPPs by adding new functionalities.

Since the SOM algorithm is robust to the missing data and rare outliers, the codebook matrix is not just an approximation to the input matrix, but can be more useful than previously thought. For instance, the codebook matrix can be further used to explore relationships between samples. It is an equivalent of reorganizing component planes by placing similar component planes closer to each other. Such reorganization can be realized by using a new SOM map (usually a rectangular lattice on a two-dimensional map) to train component plane vectors (i.e., column-wide vectors of output codebook matrix). To ensure the unique placement, each component plane mapped to this rectangular lattice can be determined in an order from the best matched to the next compromised one. Comparing to the ordinary ones, the CPPs being reorganized in such a way are rich in the information revealed; genes and samples can be simultaneously visualized in a single display. The organized CPPs are easier to interpret, especially when the number of samples (i.e., component planes) is relatively large and the relationships between samples are unclear. To give a sense of such visual benefits, we provide three examples involving omics data generated by different high-throughput technologies (Figure 1).

2.1. Transcriptome evolution in mammalian organs

Comparative study of different organs or of different species can be a useful approach for the insights into transcriptome features underlying phenotypic changes [29]. To demonstrate the power of our analysis tools in this regard, we first selected a recently published dataset comprising 13,277 one-to-one orthologous genes across six primates, each with six organs [30]. The data were generated by the RNA-seq technology, and expression levels were quantified by reads per kilobase of exon model per million mapped reads (RPKM). These reads were normalized across species/tissues on the basis of rank-conserved genes, followed by logarithm transformation. The higher the normalized and transformed RPKM indicates the higher expression levels. We projected samples onto two-dimensional space by Sammon mapping [31]. As showed in the top panel of Figure 1A, samples are grouped together according to the tissue origins except for the neural tissues (i.e. brain and cerebellum), which is slightly better than the originally published results using principle component analysis. When the reorganized CPPs were used instead, more informative relationships were revealed just by visual inspection (the bottom panel of Figure 1A). First, each component plane provides a sample-specific transcriptome map (rather than a dot). Second, samples are better separated even for the neural tissues. Last but not the least, there is much room left to label the samples; the species origins can be titled/colored above each plane. To conclude, the reorganized CPPs permit the direct comparisons of cross-species transcriptome evolution within the same tissue and cross-tissue trancriptome changes within the same species as well.

2.2. Regularome of multiple transcription factors in embryonic stem cells

Characterizing transcription factor (TF) binding sits from a genome-wise scale is the key to the understanding of pluripotency and reprogramming [32]. Also, such an approach has been widely used in various biological investigations. To further illustrate the visual benefits of using the SOM and the reorganized CPPs, we chose a second dataset generated by the ChIP-seq technology, which contained binding sites of 14 TFs at the promoter regions of 17,442 genes in mouse [33]. TF-gene association scores were calculated to estimate the strength of binding, with higher scores implying higher chance of a gene (in rows) being targeted by a TF (in columns). As shown in Figure 1B, the visual inspection of the reorganized CPPs suggests several features associated with this multiple TF regularome dataset: (i) binding profiles of five TFs (i.e., Nanog, Sox2, Oct4, Smad1 and STAT3) are similar both in the number and strength of target genes, being exclusively located into the bottom-right corner; (ii) another four TFs (i.e., n-Myc, c-Myc, Klf4 and Zfx) share much more common binding profiles than the rest, and are placed together; (iii) when examining regularome of two TFs (i.e., E2f1 and Suz12), their component planes are far apart, which is consistent with the observation that their binding profiles are mutually exclusive. Unlike the original publication, the reorganized CPPs spotlight these prominent features under a single informative display.

2.3. Transcriptome profiling in cancer classification

Cancer classification based on transcriptome profiling is one of the most popular applications [34]. For this regard, we chose a third dataset generated by oligonucleotide chip, consisting of 5,000 genes expressed at 38 leukemia samples [35]. These samples include 11 acute myeloid leukemia (AML) and 27 acute lymphoblastic leukemia (ALL) that can be further sub-typed into 19 B-cell ALL (ALL_B) and 8 T-cell ALL (ALL_T). This dataset is typically used as classification benchmark to evaluate the performance of the methods being tested. Here we used it for the reorganized CPPs to visualize three known classes and their boundaries. Figure 1C intuitively displays the AML-ALL distinction, each occupying its own landscape (AML on the right and ALL at the left). Within the ALL-occupied landscape, the partition between ALL-B and ALL-T can also be observed despite the fact that this benchmark dataset contains sample outliers (probably due to incorrect diagnosis of ALL samples). Since the cancer is a highly heterogeneous population with ambiguous boundary for the subpopulations/subtypes, the information provided by visualized data both in genes and samples is fairly important for the cancer classification and the identification of subtype-specific molecular signatures as well.

3. Timeline of the SOM-based applications in omics data mining

The SOM, originally proposed by Kohonen [36], is a special instance of artificial neural networks (ANNs) as an competitive learning algorithm inspired by the cortex of human brain.

Unlike other ANNs, a unique feature of SOM is that it can use neighborhood kernels to preserve and also control the topological structure of high-dimensional input data [16]. For this reason, the SOM has become a valuable tool and primary choice for visualizing and characterizing a relatively massive amount of data. Announced by Kohonen in the WSOM 2011 conference, there are already over 10,000 scientific papers published using SOM. The major contributions to this huge publication list come from its broad applications in engineering, economics and biomedicine [37,38].

Literature surveys of the bibliography suggest the existence of three periods, which can be used to summarize the past developments and applications of the SOM in multidimensional omics data. Namely, they are the opening, maturing, and turning periods along the timeline ahead. The opening period last from the year 1999 to 2001, in which the SOM was widely introduced into the field of genomics research. It attracted a great deal of interest by its superiority. Compared to other existing methods such as hierarchical clustering [14] at that time, it was scalable to large datasets, and was robust to noise and outliers. Also, two factors could explain the sudden popularity. At very end of the last century, there was a great need to develop effective tools for the extraction of the inherent biological information from explosive gene expression data. Another factor is that, although mathematically hard to understand, the computational implementation of the SOM algorithm was just available for the practical use, together with user-friendly documentations regarding data pre-processing, training and post-processing [17,39-41]. The following years (2002-2004) could be considered as the maturing period. During this period, biologists realized that it could be misleading without knowing the context of omics data. Accordingly, special attentions were given to visual potentials of the SOM when analyzing omics data. Also, numerous attempts were made to solve the problems associated with the algorithm itself, such as the requirements of pre-defined cluster numbers and the doubts on stability of clusters obtained. From the year 2005 on, the fewer advances have been achieved in gene expression data applications, although several combinations with other methods have also been reported. It can be explained by the shift from emphasis on the numeric gene expression data to the nonnumeric sequenced genomic data. This shift discourages the direct application of the SOM, and several variants of the SOM were instead tried. For these reasons, we call the third as the turning period. In the rest of this section, we will give a fair review of these three periods by focusing on successful applications and innovative improvements in the context of omics data mining.

3.1. Opening period by emerging gene expression data analysis

The SOM was first applied to interpret gene expression data of hematopoietic differentiation [15]. In the same year, several applications in other biological systems were also reported. These included the use of the SOM to analyze and visualize yeast gene expression data during diauxic shift [42], to process the developmental gene expression data during metamorphosis in Drosophila [43], and to discover and predict cancer classes based on gene expression data [35]. Thereafter, the exploratory nature of the SOM for the use was exploited in the context of gene expression data analysis [44,45]. In addition to expression data, the SOM was also proved as a powerful tool to characterize horizontally transferred genes by looking at the codon usage patterns of bacterial genomic data [46,47].

3.2. Maturing period for algorithm optimizations and improvements

Visual advantages of the SOM were systematically demonstrated in revealing relationships among genes of known functional classes [48], classifying tissues of different origins [49] and tumor origins [50], and both [51]. In particular, component plane-based visualizations were much appreciated [18,51-53]. As illustrated in the previous section, our experience of using reordered CPPs started with microarray data analysis. Such sample-specific presentations are intuitive to biologists, because it is straightforward to interpret biological significances of genes (being clustered) with respect to each sample [18]. Another major improvement during this period was the development of SOM variants, as highlighted by adaptive double SOM [54] and hierarchical dynamic SOM [55], to address the issue of how to identify unknown/consistent cluster number. The adaptive double SOM adapts its free parameters during the training process to find consistent cluster number, while hierarchical dynamic SOM uses growing SOM to hierarchically improve the clustering process. To account for the random initial conditions and to assess clustering stability, a generic strategy called resampling-based consensus clustering was also proposed to represent the consensus over multiple runs of the SOM algorithm with random restart [56]. Unfortunately, performance evaluations showed that consensus clustering for the SOM produced slightly worse results than that for the hierarchical clustering, and both were overtaken by another method based on nonnegative matrix factorization [57]. Using the SOM for the biological sequence analysis were also attempted, including the nonvectorial SOM algorithm for the clustering and visualization of a large protein sequences [58], the partitioning of similar protein sequences for the subsequent conserved local motif prediction [59], hidden genome signature visualization [60] and gene prediction [61].

3.3. Turning period with the emphasis on the nonnumeric data and the combination of the SOM with other methods

One of active attempts to analyze the DNA sequences was TF binding site identification [62] and sequence motif discovery [63], both using sequence motif representations as input vectors. Such DNA motif identifications were recently improved by using a heterogeneous node model [64]. Several variants of the SOM were reported to analyze microbial metagenomes for clustering and visualizing taxonomic groups. With the DNA oligonucleotide frequencies as input, emergent SOM was used for the increase in the projection resolution [65,66], growing SOM was used for speed improvements [67,68], and the main parameters of the SOM were studied for the accuracy [69]. Using other representations of genomic sequences was also reported in the hyperbolic SOM [70]. TaxSOM implementing the growing SOM and batch-learning SOM [71,72] was recently made available for the ease of use [73]. In terms of gene expression data, combinations with other methods were actively studied. The SOM was used as a data-filtering to improve classification performance of the support vector machine [74]. Multi-level SOM of SOM was proposed to determine the cluster number [75]. Minimum spanning tree and ensemble resampling were also employed to post-process the SOM for automatic clustering [76]. The combination of the SOM and the singular value decomposition (SVD) was suggested by us for topology-preserving gene selection [19].

4. A SOM-centric pipeline and its tutorial for the in-depth mining of transcriptome changes during early human organogenesis

The aforementioned three examples clearly show that the SOM with the reorganized CPPs enables straightforward and widespread use in a variety of omics data. From previous applications, a lesson can be learned that the popularity of the SOM during the opening period is not merely driven by the explosive gene expression data, but is also attributable to the availability of algorithm implementation and tutorial documentations. Accordingly, we attempt to develop a SOM-centric pipeline for maximizing its beneficial potentials in visualizing, selecting and clustering multidimensional omics data. Briefly, the implementation of pipeline starts with the preparation of data, in the form of gene activity matrix, to record biological activities of a large number of genes (rows) against related samples (columns). It is always advisable to pre-process raw data, such as normalization by rows and/or columns, and logarithmic transformation to approximate normal distribution. After that, it is highly recommended to simultaneously visualize genes and samples by the reorganized CPPs; these dual visualizations aim to effectively characterize data structure and to visually monitor data quality. Hybrid SOM-SVD is applied for topology-preserving gene selection, while the distance matrix-based clustering of the SOM (a special type of a SOM-based two-phase gene clustering) is used for topology-preserving gene clustering. The obtained genes clusters can facilitate many aspects of biological interpretations by applying enrichment analysis to examine whether clustered genes share functional, regulatory, or phenotypic characteristics. Also, the dominant patterns revealed by SOM-SVD can facilitate the graph mining tools for detecting temporal expression-active subnetworks. To demonstrate these multifaceted functionalities of this SOM-centric pipeline, we provide a tutorial overview of in-depth mining transcriptome changes during early human organogenesis, together with the necessary details of the underlying algorithms and the biological explanations.

Prior to the tutorial, it is necessary to clarify the technical issues with respect to the SOM used here. In terms of the SOM topology, the map size is heuristically determined based on the input training data, as suggested by the MATLAB SOM toolbox [77]. During the SOM training, the map is linearly initialized along two greatest eigenvectors of the input data. Then, map nodes compete to win the input data, followed by updating the winner node and its topological neighbors. This iterative training is implemented using the batch algorithm and contains two phases: rough phase and fine-tuning phase. To increase the reproducibility of the trained map, we purposely prolong the fine-turning phase until the successive fine-tunings reach a steady state; the quality of the SOM map (i.e., average quantization error and topographic error) does not change any more. Among various parameters associated with the SOM training, the neighborhood kernel is the most important one because it dictates the final topology of the trained map. In addition to the commonly used Gaussian function (see Equation 1), others, such as Epanechikov function (see Equation 2), Gut-gaussian function and Bubble function, can also be chosen depending on the tasks [77]. From the mathematical definitions as well as the practical comparisons using the same test of data, we have observed that Epanechikov neighborhood kernel puts more emphasis on local topological relationships than the other threes, suitable for the use in gene selection. On the other extreme, the Gaussian neighborhood kernel preserves global topology relationships to the most extent, and thus is ideal for the use in global gene clustering and visualization. As demonstrated below, the dual strengths of the SOM in preserving both local and global topological properties (via choosing different neighborhood kernels) can optimize the data processing from multi-aspects.

$$h_{ci}(t)=\max\left\{0,\ 1-\frac{\|\vec{r}_c-\vec{r}_i\|^2}{\sigma^2(t)}\right\} \tag{1}$$

$$h_{ci}(t)=\exp\left(-\frac{\|\vec{r}_c-\vec{r}_i\|^2}{2\sigma^2(t)}\right) \tag{2}$$

where the positive integer $\sigma(t)$ defines the width of the kernel at training time t, and $\vec{r}_c$ and $\vec{r}_i$ are respectively the location vectors of the winner node c and a node i on the two-dimensional SOM map grid.

4.1. Simultaneous visualizations of genes and samples

In our previous work [27], we have analyzed transcriptome data during early human organogenesis (hORG), which involves human embryos at six consecutive stages (Carnegie stages 9-14, S9-S14) with three replicates for each. Here, we use it for pipeline tutorials and for demonstrations on further improvements. After normalization and pre-filtering, the gene expression matrix contains expression values of 5,441 genes (in rows) × 18 samples (in columns; six developmental stages S9-S14 in triplicate R1-R3 for each) (available at the supplemental Table 1 in the original publication [27]). To account for variance stabilization and to focus on the relative expression across the samples, we further pre-process this matrix by base-2 logarithm transformation and the row-wise centering. From the hORG matrix, the gene expression vectors are input for the SOM training with the Epanechikov neighborhood kernel and the grid of 360 (30 × 12) hexagonal nodes. Each column of SOM codebook matrix corresponds to one component plane. The column-wise component plane vectors are then used to train a new SOM with the Gaussian neighborhood kernel (see [2]) and the grid of 40 (5 × 8) rectangular nodes. The placement of a component plane is determined in a sequential rank from the best-matching node (BMN) to the second BMN and so on (using Pearson correlation coefficients as the similarity metric). The above process repeats until all the component planes find the non-overlapping location in the rectangular lattice. As shown in Figure 2B, the reorganized CPPs enhance the visual convenience by placing component planes in a biologically meaningful manner. The relative geometric distance intuitively illustrates the correlations within the three replicates and across the six developmental stages. Remarkably, such simultaneous visualizations of genes and samples reveal developmental trajectory in the transcriptome landscape of early human organogenesis.

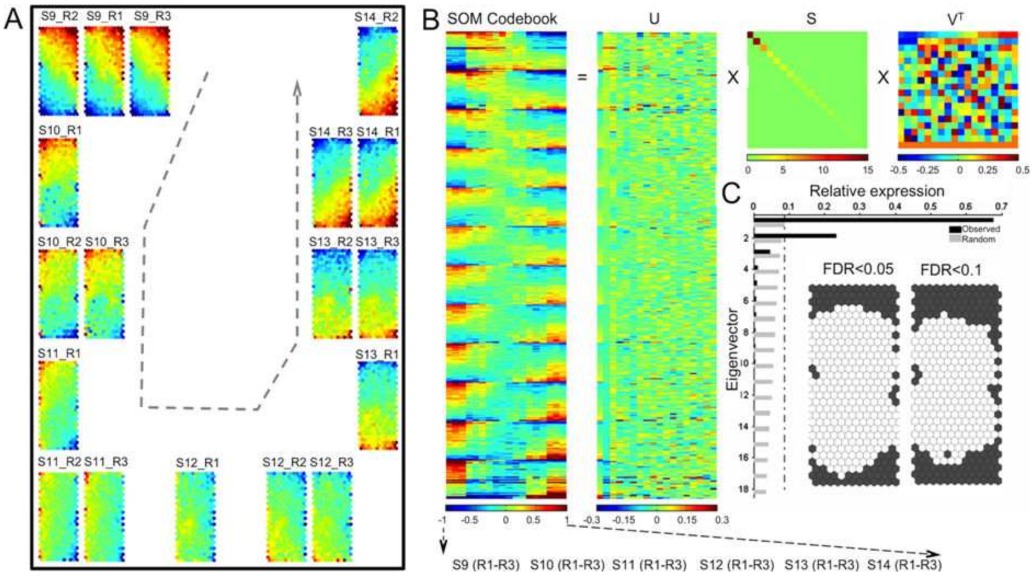

Figure 2. A tutorial on the simultaneous visualizations of genes and samples by the reorganized CPPs, and topology-preserving gene selection through the SOM-SVD. (A) The reorganized CPPs of transcriptome changes during early human organogenesis. Each component plane illustrates a sample-specific transcriptome map. Sample similarities and differences are also illustrated by the extent to which component planes are geometrically related to each other. Owing to simultaneous visualizations of genes and samples, the dotted line can be intuitively drawn to denote the developmental trajectory. (B) Decomposition of the SOM codebook matrix by SVD. This codebook matrix is linearly decomposed into three matrices of U, S and V^T. Values of eigensamples (columns of U), eigenexpressions (on-diagonal entries of S) and eigenvectors (rows of V^T) are color-encoded as indicated by bar underneath. (C) SOM node selection by false discovery rate (FDR) to account for multiple hypothesis tests. Bars on the left illustrate the relative contribution (in relative to the overall variation) of observed eigenvectors (filled in black) and randomized eigenvectors (filled in gray) from a randomization. The dominant eigenvectors are selected if their observed relative expression is larger than the maximum of random relative eigenexpression (as indicated by the vertical dotted line). On the right displays the SOM grid map with nodes being selected (in heavy gray) or not (in white) under the threshold of FDR as indicated.

4.2. Topology-preserving gene selection

In our recent work [19], we have developed hybrid SOM-SVD for topology-preserving selection of genes that show statistically significant changes in expression. Unlike conventional arbitrary or manual gene selection procedures, this approach permits the entire gene selection process to be realized automatically and on the basis of statistical inference. Through comparisons with other methods, this approach has demonstrated to be more effective in selecting cell cycle genes with a characteristic period. Also, the gene selection by hybrid SOM-SVD can facilitate the downstream clustering analysis, as direct application of the clustering method on unselected data may distort the topology of global clustering [19].

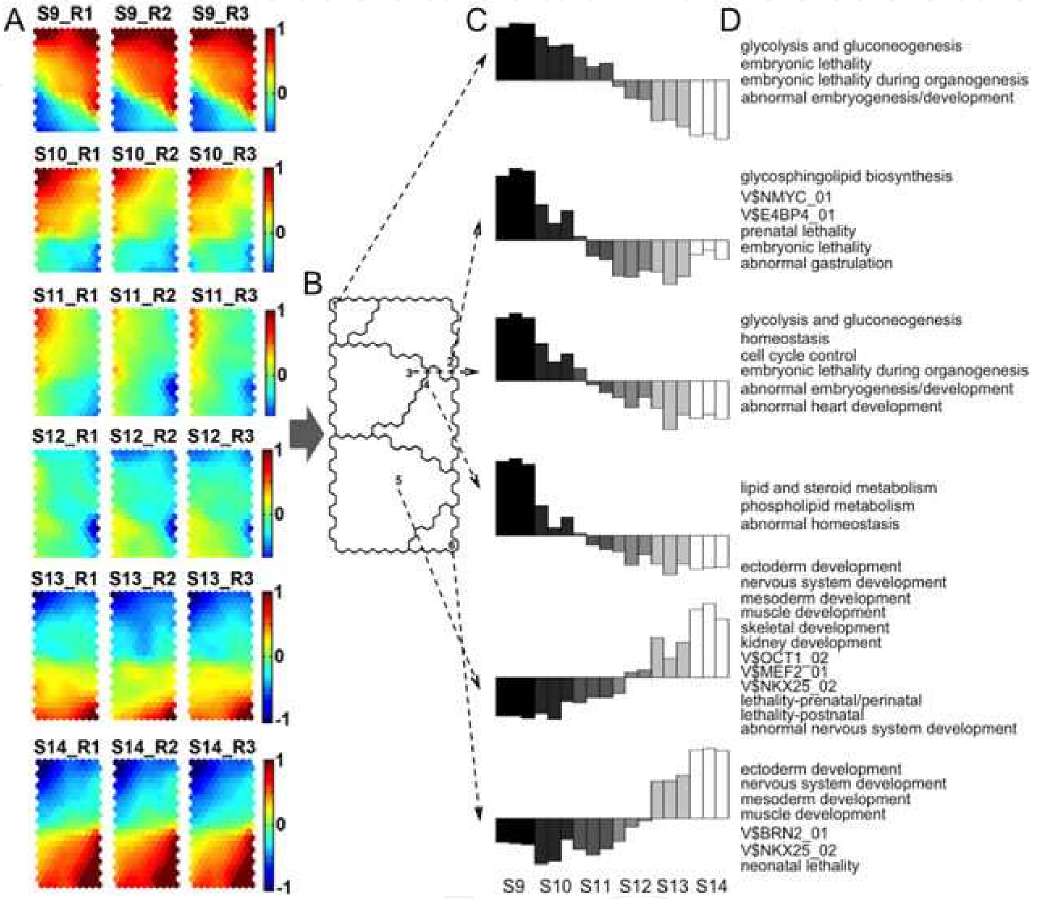

Figure 3. A tutorial on topology-preserving gene clustering by the distance matrix-based clustering of the SOM. (A) The CPPs of the SOM outputs using the input of the gene expression matrix selected by SOM-SVD. (B) Ideogram illustration of six gene clusters on a SOM grid map. The cluster index is marked in the seed node. From each seed node, the corresponding cluster is obtained through a region growing procedure. (C) Bar-graph display of SOM outputs in seed nodes. (D) Significant functional, regulatory and phenotypic features associated with gene clusters.

The hORG tabulated gene expression matrix (5,441 genes × 18 samples) is first subjected to non-linear transformation using the SOM algorithm with the Epanechikov neighbourhood kernel and the grid of 360 (30 × 12) hexagonal nodes. The resultant codebook matrix (i.e., 360 nodes in rows × 18 samples in columns) serves as an intermediate format for pattern recognition by SVD (Figure 2B). It is sequentially followed by two dominant eigenvector selection, SVD subspace projection and distance statistic construction, significant node assessment using the false discovery rate (FDR) procedure for multiple hypothesis tests, and finally the selection of significant nodes and their genes as defined by the BMN (Figure 2C).

A total of 2,148 genes are selected under an FDR cutoff of 0.1. The selected gene expression matrix (2,148 genes × 18 samples) forms the characteristic matrix, which can be used for further clustering analysis. Notably, the motivations behind the combination of the SOM with the SVD are: (i) the separation of features and artifacts by the SOM training with the Epanechikov neighbourhood kernel, (ii) the pattern recognition of features and artifacts by SVD decomposition, and (iii) the statistical selection of features by the FDR.

4.3. Topology-preserving gene clustering

Gene clustering in a topology-preserving manner is implemented using a SOM-based two-phase clustering algorithm that takes into account SOM neighborhoods. In the first phase, the gene expression vectors (preferably from gene expression matrix selected by SOM-SVD) are trained by SOM with the Gaussian neighbourhood kernel to better preserve the topology of the data. In the second phase, the resultant SOM map is divided into a set of clusters using a region growing procedure. By calculating the SOM distance matrix from U-matrix (i.e., distances between each map node and its neighbors) [78], this procedure starts with local minima of distance matrix as seeds, followed by the assignment of the remaining nodes to their corresponding clusters [79]. Like other hierarchical agglomerative or k-means partitive algorithms used at the second phase [40], this distance matrix-based algorithm can reduce the complexity of the clustering task from tens of thousands of genes to the hundreds of nodes in the SOM map. Unlike others, this distance matrix-based clustering of the SOM enables more reliable estimates of gene clusters in a topology-preserving manner. In our previous work [19], we have shown that, for the same data as input, using k-means clustering at the second phase could not result in topology-preserving gene clusters. Also, we have demonstrated the preferential use of the SOM-SVD gene selection ahead of the topology-preserving gene clustering. Otherwise, it would distort the topology of global clustering when directly applying on the unselected data.

Therefore, the gene expression matrix of 2,148 genes × 18 samples, as selected by the SOM-SVD, is used as input for the SOM-based two-phase gene clustering. Specifically, the input data is first trained using the SOM with 220 (22 × 10) nodes and Gaussian neighborhood kernel, and the SOM codebook matrix is displayed by CPPs in Figure 3A. The trained map is then divided using the region growing procedure. As showed in Figure 3B, the map nodes at the second phase of the gene clustering are continuously organized into six clusters according to neighborhood relationships and without any pre-knowledge of data structure. Since the seed nodes are identified as local minima (i.e., cluster centres), the pattern seen in a seed node can be viewed as the average expression pattern of genes mapped to that seed. More loosely, it can also be approximated as the overall pattern in the gene cluster obtained from the seed. As show in Figure 3C, seeds in clusters 1-4 display gradually decreasing expression patterns, while those for clusters 5-6 have gradual increasing pattern in expression. More importantly, gene clusters facilitate the downstream biological interpretations based on the paradigm of ‘coexpression-cofunction-coregulation’. Such interpretations are coupled with external biological annotations such as Gene Ontology (GO) [80], conserved TF binding sites (in the form of positional weighted matrix) from the UCSC Genome Browser database [81] and mammalian phenotype ontology [82]. Using these diverse annotations, enrichment analysis is conducted to identify functional, regulatory and phenotypic features that are shared by genes being clustered together. Figure 4D lists shared features associated with each gene cluster. Genes in clusters 1-3 are functionally related to cellular metabolism and homeostasis, are possibly regulated by survival-related transcription factors, and are largely linked to embryonic lethality and abnormal embryogenesis. By contrast, genes in cluster 5-6 are functionally involved in the establishment of organ morphogenesis, are regulated by organogenesis-specific TFs, and are primarily linked to postnatal lethality and diverse organ/system defects.

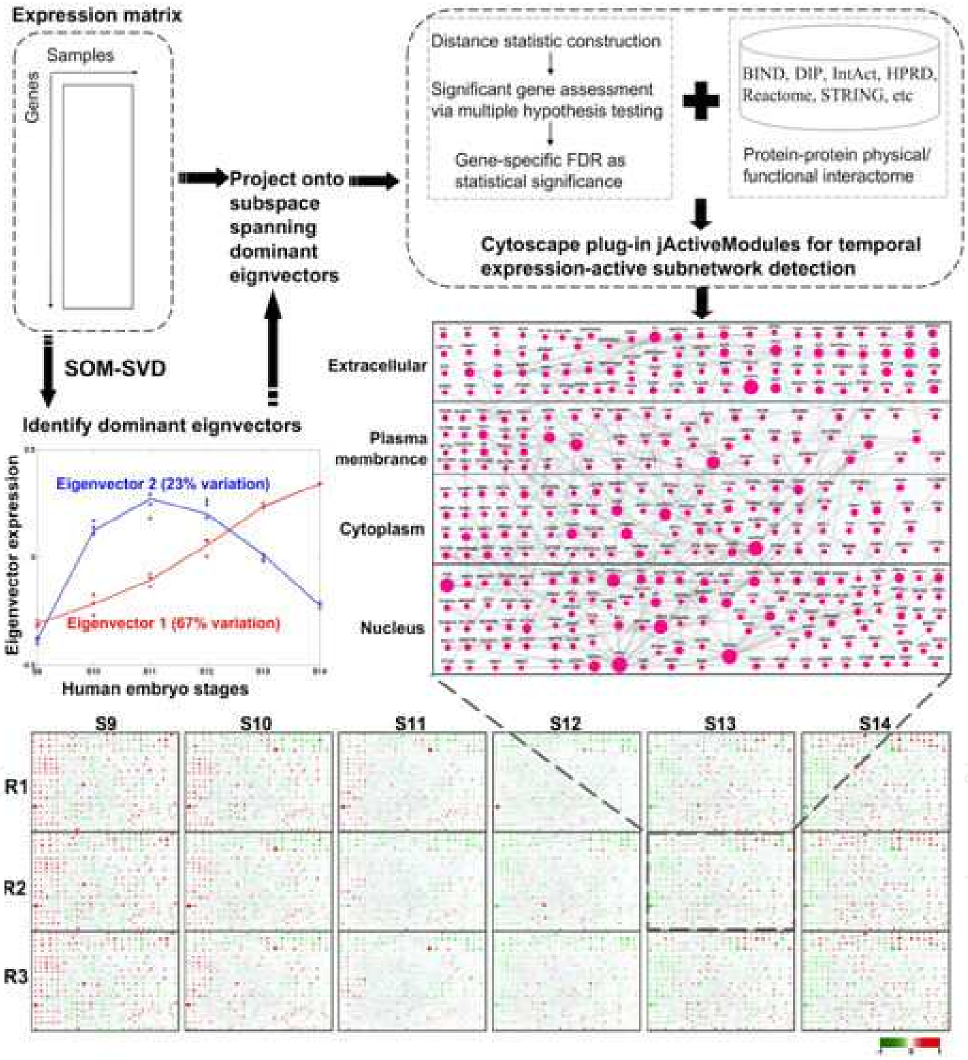
Expression matrix
Samples
Genes
Project onto subspace spanning dominant eignvectors
SOM-SVD
Identify dominant eignvectors
Eigenvector expression
Eigenvector 2 (23% variation)
Eigenvector 1 (67% variation)
Human embryo stages
Distance statistic construction
Significant gene assessment via multiple hypothesis testing
Gene-specific FDR as statistical significance
BIND, DIP, IntAct, HPRD, Reactome, STRING, etc
Protein-protein physical/ functional interactome
Cytoscape plug-in jActiveModules for temporal expression-active subnetwork detection
Extracellular
Plasma membrance
Cytoplasm
Nucleus
S9
S10
S11
S12
S13
S14
R1
R2
R3

Figure 4. A tutorial on temporal expression-active subnetwork detection by jActiveModules. The Cytoscape plug-in jActiveModules, as a subgraph-searching tool, requires the input of both a user-predefined network being searched against and a gene-specific metric to measure the significance of expression change (top-right corner). For the network to be input, the existing protein physical interaction databases such as BIND, DIP, IntAct, HPRD, Reactome can be compiled together, which can be further complemented by the functional interactions from the database like STRING to improve the network coverage. For the temporal change measure, the dominant eigenvectors identified by SOM-

SVD analysis can be used (top-left corner). As suggested here, it consists of three steps, including gene projection onto the subspace spanning dominant eigenvectors, distance statistic construction, significant gene assessment through multiple hypothesis tests for FDR calculation. The gene-specific FDR is then used as the significance of expression change. With both data as input, jActiveModules uses the simulated annealing to detect expression-active subnetworks containing genes with expression patterns highly similar to dominant eigenvectors as identified by SOM-SVD analysis. The middle-right panel displays the detected temporal expression-active subnetwork, the layout of which is reconfigured according to subcellular localization. By overlaying gene expression data from each of 18 samples (i.e., three replicates R1-R3 in rows × six stages S9-S14 in columns) onto the subnetwork, each plane (such as S13_R2 as highlighted in dot lines) illustrates sample-specific subnetwork with genes/nodes color-encoded based on their expression values as indicated underneath (bottom panel). Similar to the CPPs, such plane visualization permits the monitoring of the subnetwork expression changes, indicative of this subnetwork activity being dynamically changed during early human organogenesis.

4.4. Temporal expression-active subnetwork detection

A temporal expression-active subnetwork is the connected region of an interactome/ network, constrained by that this subnetwork should contain genes that show significant changes in expression over a biological process. Such active subnetworks can bring the value of omics data into the higher level. Biologically, genes do not act alone but are interconnected into cohesive networks. Methodologically, the integration of two or more sources of omics data can increase the chance of identifying biologically meaningful knowledge than either data source. Temporal expression-active subnetworks can be viewed as the integration of the context-independent interactome (static, unionizing all possible interactions) and the context-specific transcriptome (dynamic, involving only genes being expressed under the conditions). The Cytoscape plug-in jActiveModules [83] is one of algorithms that have been successfully used for identifying expression-active subnetworks. In addition to a user-predefined network, it also requires the input of a gene-specific metric to measure the significance of expression change. This method is effective for the transcriptome data obtained from the 'case-control' experimental design because the significance of expression change can be evaluated by testing the differences. In a time-series setting, however, this method can be problematic. Although any two-successive expression change can result in the corresponding expression-active subnetworks, these subnetworks may not overlap at all and ignore the temporal dependency. It is appealing to identify subnetworks that are cohesively active across the whole time series. For the use of jActiveModules in this purpose, we propose to calculate a gene-specific FDR as a measure of significance in temporal expression. The basic idea is to weigh genes according to their similarity with dominant eigenvectors (as identified by SOM-SVD). Similar to the calibration strategy, genes with expression pattern similar to the dominant eigenvector expression are up-weighed; otherwise down-weighed.

Schematic flowchart in Figure 4 illustrates a temporal expression-active subnetwork during early human organogenesis. Brief explanations can be found in the legend. Here, we only detail the steps of how to calculate the gene-specific FDR from gene expression matrix (denoted as M with G genes × N samples) and the L dominant eigenvectors (e.g., the first 2 dominant eigenvectors identified by SOM-SVD analysis in Figure 2). Let $\vec{x}$ be gene expression vector, and $\Re^{L}$ be SVD subspace spanning by the L dominant eigenvectors. We project $\vec{x}$onto$\Re^{L}$, obtaining projection vector$\vec{q} \in \Re^{L}$. In$\Re^{L}$, we compute the Euclidian distance (distance statistic, DS) of projection vector $\vec{q}$ away from the coordinate-wise zero point. The

DS measures similarity between gene expression and the dominant eigenvector expression, with the larger value indicating the higher similarity. When comparing multiple hypothesis tests simultaneously, we assess statistical significance of gene-specific *DS* by a method of FDR, described as follows. For the matrix *M*, we first use the above procedure to obtain a list of *DS*, being ranked as$DS_{r1} \leq DS_{r2} \leq \cdots \leq DS_{rG}$. Then, obtain $b = 1, \ldots, B$ randomized matrix M^{b}, which is generated by randomly permuting matrix *M* in both row and column directions. Analogously, compute projection values of randomized gene expression vector $\vec{x}^{b}$ on the chosen *L* dominant eigenvectors to obtain projection vector and calculate the distance statistic DS^{b}, and rank the distances:$DS_{r1}^{b} \leq DS_{r2}^{b} \leq \cdots \leq DS_{rG}^{b}$. Finally, assess statistical significance in terms of FDR for each gene. For the ri^{th} gene as ordered, compute the number of genes called significant ($rG - ri + 1$), and the median number of genes falsely called significant by calculating the median number of genes among each of the *B* sets of reference data, whose DS_{rj}^{b}satisfy:$DS_{rj}^{b} \geq DS_{ri}$, $j = 1, \ldots, G$. Thus, FDR for the ri^{th} ordered gene is quantized as the median number of falsely called genes divided by the number of genes called significant.

5. Conclusion

A great number of advances in the SOM have been made during the past decades. The applications in the omics data mining are largely driven by the persuasive gene expression data, as well as by the availability of the user-friendly tools. The ongoing applications are to analyze the nonnumeric genomic sequenced data, probably combined with other existing methods. In principle, the same SOM procedures could also be applied to the nonnumeric sequenced data, if these sequenced data could be numerically transformed in an appropriate way (such as regularome data illustrated in Figure 1B). We envisage that these massive omics data, whether be quantified numerically or not, offer an unprecedented opportunity for the next-wave applications of the SOM. It requires the better appreciation of its dual strengths in preserving both local and global topological properties through adjusting neighborhood functions. To guide towards this direction, we have extended our previous approach into a SOM-centric pipeline, and through a real-world transcriptome data, have demonstrated its practical usefulness in achieving multifaceted functionalities. Below, we discuss future directions for further improvements.

Owing to the advantage in simultaneously displaying genes and samples, the reorganized CPPs have been demonstrated powerful for use in a variety of omics data (Figure 1). As an improvement to the ordinary CPPs, geometric location within a rectangular lattice has been utilized to reveal natural relationships between samples. At the current state, the ambiguous boundary is identified by visual inspection (Figure 1C). In the future, an automatic procedure is needed to avoid any subjective intervention from human. Another issue regarding the reorganized CPPs is limited space left for displaying component planes, especially when hundreds of samples are involved. One of the possible solutions is to use the tree-like structure [84]. The tree-structured is a natural way to link together component planes that have been clustered into different groups. Each node of the tree is a set of component planes visualized by the reorganized CPPs. Further efforts in this direction can increase the value of the reorganized CPPs in transcriptome profiling-based cancer classifications.

Another promising direction is to improve the stability of the gene clusters obtained by SOM-based two-phase clustering algorithm. The obtained clusters not only depend on random variations in the data, which has been reduced through the SOM-SVD gene selection (Figure 2), but also the stochastic nature of the SOM algorithm. As a result, distance matrix from U-matrix would differ from multiple runs, which will affect the determination of the seed nodes (i.e., local minima of distance matrix; Figure 3). The strategies like consensus clustering [56] could be used for the improvements.

The use of the SOM in network-level interpretations of omics data is poorly attempted in the literature. We have showed such possibility of aiding in temporal expression-active subnetwork detections (Figure 4). However, the SOM here only plays an indirect role. It has been reported to be used in the social network mining [85]. Much more work remains to be done so that the SOM could be directly applied to the intereactome data. Since the networked data are primarily represented as an adjacent matrix, the SOM of the matrix data (rather than the vectors) seems to be possible too [86].

Author details

Ji Zhang[1,2*] and Hai Fang[1,2]

*Address all correspondence to: jizhang@sibs.ac.cn

1 State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology and Sino-French Center for Life Science and Genomics, Rui-Jin Hospital affiliated to Shanghai Jiao Tong University School of Medicine, China

2 Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China

References

[1] Ledford, H. (2010). Big science: The cancer genome challenge. *Nature*, 464(7291), 972-974.

[2] Toft, C., & Andersson, S. G. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet*.

[3] Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467-470.

[4] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., & Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13), 1675-1680.

[5] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., & Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290(5500), 2306-2309.

[6] Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., Brodsky, A. S., Keeton, E. K., Fertuck, K. C., Hall, G. F., Wang, Q., Bekiranov, S., Sementchenko, V., Fox, E. A., Silver, P. A., Gingeras, T. R., Liu, X. S., & Brown, M. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nat Genet*, 38(11), 1289-1297.

[7] Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science; , 316(5830), 1497-1502.

[8] Domon, B., & Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science*, 312(5771), 212-217.

[9] Walhout, A. J., & Vidal, M. (2001). High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, 24(3), 297-306.

[10] Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol*, 26(10), 1135-1145.

[11] Wang, Z., Gerstein, M., & Snyder-Seq, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1), 57-63.

[12] Hood, L., Heath, J. R., Phelps, M. E., & Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696), 640-643.

[13] Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, 13(1), 36-46.

[14] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25), 14863-14868.

[15] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., & Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6), 2907-2912.

[16] Kohonen, T. (2001). Organizing Maps. Third, extended edition Springer

[17] Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2), 111-126.

[18] Xiao, L., Wang, K., Teng, Y., & Zhang, J. (2003). Component plane presentation integrated self-organizing map for microarray data analysis. *FEBS Lett*.

[19] Fang, H., Du, Y., Xia, L., Li, J., Zhang, J., & Wang, K. A. (2011). A topology-preserving selection and clustering approach to multidimensional biological data. *OMICS*.

[20] Vesanto, J., & Ahola, J. Hunting for Correlations in Data Using the Self-Organizing Map. In Proc. of International ICSC Congress on Computational Intelligence Methods and Applications (CIMA'99), Rochester, New York, USA, June 22-25

[21] Xu, K., Guidez, F., Glasow, A., Chung, D., Petrie, K., Stegmaier, K., Wang, K. K., Zhang, J., Jing, Y., Zelent, A., & Waxman, S. (2005). Benzodithiophenes potentiate differentiation of acute promyelocytic leukemia cells by lowering the threshold for ligand-mediated corepressor/coactivator exchange with retinoic acid receptor alpha and enhancing changes in all-trans-retinoic acid-regulated gene expression. *Cancer Res*, 65(17), 7856-7865.

[22] Zheng, P. Z., Wang, K. K., Zhang, Q. Y., Huang, Q. H., Du, Y. Z., Zhang, Q. H., Xiao, D. K., Shen, S. H., Imbeaud, S., Eveno, E., Zhao, C. J., Chen, Y. L., Fan, H. Y., Waxman, S., Auffray, C., Jin, G., Chen, S. J., Chen, Z., & Zhang, J. (2005). Systems analysis of transcriptome and proteome in retinoic acid/arsenic trioxide-induced cell differentiation/apoptosis of promyelocytic leukemia. *Proc Natl Acad Sci U S A*, 102(21), 7653-7658.

[23] Du, Y., Wang, K., Fang, H., Li, J., Xiao, D., Zheng, P., Chen, Y., Fan, H., Pan, X., Zhao, C., Zhang, Q., Imbeaud, S., Graudens, E., Eveno, E., Auffray, C., Chen, S., Chen, Z., & Zhang, J. (2006). Coordination of intrinsic, extrinsic, and endoplasmic reticulum-mediated apoptosis by imatinib mesylate combined with arsenic trioxide in chronic myeloid leukemia. *Blood*, 107(4), 1582-1590.

[24] Fang, H., Wang, K., & Zhang, J. (2008). Transcriptome and proteome analyses of drug interactions with natural products. *Curr Drug Metab*, 9(10), 1038-1048.

[25] Wang, K., Fang, H., Xiao, D., Zhu, X., He, M., Pan, X., Shi, J., Zhang, H., Jia, X., Du, Y., & Zhang, J. (2009). Converting redox signaling to apoptotic activities by stress-responsive regulators HSF1 and NRF2 in fenretinide treated cancer cells. *PloS one*, .

[26] Bi, Y. F., Liu, R. X., Ye, L., Fang, H., Li, X. Y., Wang, W. Q., Zhang, J., Wang, K. K., Jiang, L., Su, T. W., Chen, Z. Y., & Ning, G. (2009). Gene expression profiles of thymic neuroendocrine tumors (carcinoids) with ectopic ACTH syndrome reveal novel molecular mechanism. *Endocr Relat Cancer*, 16(4), 1273-1282.

[27] Fang, H., Yang, Y., Li, C., Fu, S., Yang, Z., Jin, G., Wang, K., Zhang, J., & Jin, Y. (2010). Transcriptome analysis of early organogenesis in human embryos. *Dev Cell*, 19(1), 174-184.

[28] Wu, K., Dong, D., Fang, H., Levillain, F., Jin, W., Mei, J., Gicquel, B., Du, Y., Wang, K., Gao, Q., Neyrolles, O., & Zhang, J. (2012). An Interferon-Related Signature in the

Transcriptional Core Response of Human Macrophages to Mycobacterium tuberculosis Infection. *PloS one*, e38367.

[29] Khaitovich, P., Enard, W., Lachmann, M., & Paabo, S. (2006). Evolution of primate gene expression. *Nat Rev Genet*, 7(9), 693-702.

[30] Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grutzner, F., Bergmann, S., Nielsen, R., Paabo, S., & Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369), 343-348.

[31] Sammon, J. W. (1969). A Nonlinear Mapping for Data Structure Analysis. *IEEE Trans. Comput.*, 18(5), 401-409.

[32] Plath, K., & Lowry, W. E. (2011). Progress in understanding reprogramming to the induced pluripotent state. *Nat Rev Genet*, 12(4), 253-265.

[33] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L., & Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6), 1106-1117.

[34] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J,., Jr, Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., & Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature; , 403(6769), 503-511.

[35] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531-537.

[36] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), 59-69.

[37] Oja, M., Kaski, S., & Kohonen, T. (2002). Bibliography of Self-Organizing Map (SOM) Papers : 1998-2001 Addendum. *Neural Networks*, 3(1), 1-156.

[38] Po, M., Honkela, T., & Kohonen, T. (2009). Bibliography of self-organizing map (som) papers: 2002-2005 addendum. *TKK Reports in Information and Computer Science, Helsinki University of Technology, Report TKK-ICS-R23*.

[39] Juha, V., Johan, H., Esa, A., & Juha, P. (1999). Self-Organizing Map in Matlab: the SOM Toolbox.

[40] Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Trans Neural Netw*, 11(3), 586-600.

[41] Siponen, M., Vesanto, J., Simula, O., & Vasara, P. An approach to automated interpretation of SOM. In Advances in Self-Organizing Maps: Springer: (2001). , 2001, 89-94.

[42] Toronen, P., Kolehmainen, M., Wong, G., & Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Lett*, 451(2), 142-146.

[43] White, K. P., Rifkin, S. A., Hurban, P., & Hogness, D. S. (1999). Microarray analysis of Drosophila development during metamorphosis. *Science*, 286(5447), 2179-2184.

[44] Kaski, S. (2001). SOM-Based Exploratory Analysis of Gene Expression Data. N, Yin H, Allinson L, and Slack J. London: Springer , 2001124-131.

[45] Torkkola, K., Gardner, R. M., Kaysser-Kranich, T., & Ma, C. (2001). Self-organizing maps in mining gene expression data. *Inf. Sci.*

[46] Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H., & Ikemura, T. (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*.

[47] Wang, H. C., Badger, J., Kearney, P., & Li, M. (2001). Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol Biol Evol*, 18(5), 792-800.

[48] Nikkila, J., Törönen, P., Kaski, S., Venna, J., Castrén, E., & Wong, G. (2002). Analysis and visualization of gene expression data using self-organizing maps. *Neural Netw.*

[49] Covell, D. G., Wallqvist, A., Rabow, A. A., & Thanki, N. (2003). Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data. *Mol Cancer Ther*, 2(3), 317-332.

[50] Buckhaults, P., Zhang, Z., Chen, Y. C., Wang, T. L., St, Croix. B., Saha, S., Bardelli, A., Morin, P. J., Polyak, K., Hruban, R. H., Velculescu, V. E., & Shih, Ie. M. (2003). Identifying tumor origin using a gene expression-based classification map. *Cancer Res*, 63(14), 4144-4149.

[51] Wang, J., Delabie, J., Aasheim, H., Smeland, E., & Myklebost, O. (2002). Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics*.

[52] Sultan, M., Wigle, D. A., Cumbaa, C. A., Maziarz, M., Glasgow, J., Tsao, M. S., & Jurisica, I. (2002). Binary tree-structured vector quantization approach to clustering and visualizing microarray data. Bioinformatics (Oxford, England) Suppl 1S , 111-119.

[53] Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, Pi., Kallioniemi, A., Wolf, M., Ruiz, J., Mousses, S., & Kallioniemi-P, O. (2003). Analysis and Visualization of Gene

Expression Microarray Data in Human Cancer Using Self-Organizing Maps. Mach. Learn.

[54] Ressom, H., Wang, D., & Natarajan, P. (2003). Clustering gene expression data using adaptive double self-organizing map. *Physiol Genomics*, 14(1), 35-46.

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

[56] Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52(1), 91-118.

[57] Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12), 4164-4169.

[58] Kohonen, T., & Somervuo, P. (2002). How to make large self-organizing maps for nonvectorial data. *Neural Netw*.

[59] Yang, Z. R., & Chou, K. C. (2003). Mining biological data using self-organizing map. *J Chem Inf Comput Sci*, 43(6), 1748-1753.

[60] Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., & Ikemura, T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res*, 13(4), 693-702.

[61] Mahony, S., McInerney, J. O., Smith, T. J., & Golden, A. (2004). Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models. *BMC Bioinformatics*.

[62] Mahony, S., Hendrix, D., Golden, A., Smith, T. J., & Rokhsar, D. S. (2005). Transcription factor binding site identification using the self-organizing map. *Bioinformatics (Oxford, England)*, 21(9), 1807-1814.

[63] Liu, D., Xiong, X., Das, Gupta. B., & Zhang, H. (2006). Motif discoveries in unaligned molecular sequences using self-organizing neural networks. *IEEE Trans Neural Netw*, 17(4), 919-928.

[64] Lee, N. K., & Wang, D. (2011). SOMEA: self-organizing map based extraction algorithm for DNA motif identification with heterogeneous model. BMC Bioinformatics Suppl 1S16.

[65] Ultsch, A, & Orchen, F. (2005). ESOM-Maps: tools for clustering, visualization,and classification with Emergent SOM.

[66] Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome biology R85.*

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

[68] Chan, C. K., Hsu, A. L., & Tang, S. L. (2008). Halgamuge SK.Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. *J Biomed Biotechnol*.

[69] Gatherer, D. (2007). Genome signatures, self-organizing maps and higher order phylogenies: a parametric analysis. *Evol Bioinform Online*, 3211-236.

[70] Martin, C., Diaz, N. N., Ontrup, J., & Nattkemper, T. W. (2008). Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. *Bioinformatics (Oxford, England)*, 24(14), 1568-1574.

[71] Abe, T., Sugawara, H., Kanaya, S., Kinouchi, M., & Ikemura, T. (2006). Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. *Gene*, 36527-34.

[72] Abe, T., Hamano, Y., Kanaya, S., Wada, K., & Ikemura, T. (2009). A Large-Scale Genomics Studies Conducted with Batch-Learning SOM Utilizing High-Performance Supercomputers.

[73] Bio-Inspired Systems: Computational and Ambient Intelligence. (2009). 5517829-836.

[74] Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, B. M., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., Amann, R., & Glockner, F. O. (2011). Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *ISME J*, 5(5), 918-928.

[75] Wu, W., Liu, X., Xu, M., Peng, J. R., & Setiono, R. A. (2005). A hybrid SOM-SVM approach for the zebrafish gene expression analysis. *Genomics Proteomics Bioinformatics*, 3(2), 84-93.

[76] Ghouila, A., Yahia, S. B., Malouche, D., Jmel, H., Laouini, D., Guerfali, F. Z., & Abdelhak, S. (2009). Application of Multi-SOM clustering approach to macrophage gene expression analysis. *Infect Genet Evol*, 9(3), 328-336.

[77] Newman, A. M., & Cooper, J. B. (2010). AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC Bioinformatics*.

[78] Vesanto, J. (2000). SOM Toolbox for Matlab 5: Helsinki University of Technology. ;.

[79] Vellido, A., Lisboa, P. J. G., & Meehan, K. (1999). Segmentation of the on-line shopping market using neural networks. *Expert Systems with Applications*, 17(4), 303-314.

[80] Vesanto, J., & Sulkava, M. (2002). Distance matrix based clustering of the Self-Organizing Map. *Artificial Neural Networks- Icann*, 2415951-956.

[81] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver,

L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet; , 25(1), 25-29.

[82] Dreszer, T. R., Karolchik, D., Zweig, A. S., Hinrichs, A. S., Raney, B. J., Kuhn, R. M., Meyer, L. R., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Pohl, A., Malladi, V. S., Li, C. H., Learned, K., Kirkup, V., Hsu, F., Harte, R. A., Guruvadoo, L., Goldman, M., Giardine, B. M., Fujita, P. A., Diekhans, M., Cline, M. S., Clawson, H., Barber, G. P., Haussler, D., & James, Kent. W. (2012). The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res (Database , 40(D918-923), 918-923.

[83] Smith, C.L., & Eppig, J.T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med*, 1(3), 390-399.

[84] Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics (Oxford, England) Suppl 1S , 233-240.

[85] Barreto, S. M. A., & Pérez-Uribe, A. (2007). Improving the correlation hunting in a large quantity of SOM component planes: classification of agro-ecological variables related with productivity in the sugar cane culture. In Proceedings of the 17th international conference on Artificial neural networks.

[86] Boulet, R., Jouve, B., Rossi, F., & Villa, N. (2008). Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomput*.

[87] Seo, S., & Obermayer, K. (2004). Self-organizing maps and clustering methods for matrix data. *Neural Netw*.