NAIST-IS-DD

Doctoral Dissertation

Spherical SOM for non-Euclidean metric space

Hirokazu Nishio

December 8, 2005

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Science.

Thesis Committee:

Professor Shigehiko Kanaya	(Supervisor)
Professor Kotaro Minato	(Co-supervisor)
Professor Kunihiro Chihara	(Co-supervisor)
Associate Professor Ken Kurokawa	(Co-supervisor)

Spherical SOM for non-Euclidean metric space *

Hirokazu Nishio

Abstract

Self-Organizing Maps (SOM) is a method for unsupervised learning based on a network of artificial neurons. It shows us intrinsic structural features of the inaccessible high dimensional data space. Spherical SOM, where neurons are arranged on spherical surface, is suitable for visualization based on the correlation between data vectors and the angle between data vectors in multidimensional space for example expression profiles of genes. It is difficult to arrange arbitrary number of neurons uniformly on spherical surface. To overcome this difficulty, I propose spiral arrangement of neurons on spherical surface, which allows arrangement of arbitrary number of neurons. I also propose a measure to access the suitability of any SOM to any data set, and confirm for a particular set of gene expression data and randomly generated high dimensional data that the Spherical SOM is more suitable than the original Plane SOM.

Keywords:

Self organizing maps, Spherical SOM, arbitrary number of neurons, uniformity, suitability

*Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD, December 8, 2005.

List of Publications

Journal Papers

1. Hirokazu Nishio, Md. Altaf-Ul-Amin, Ken Kurokawa, Shigehiko Kanaya, "Spherical SOM and arrangement of neurons using helix on sphere", *IPSJ Transactions on Mathematical Modeling and its Applications*, Vol. 14 (in press)

2. Hirokazu Nishio, Md. Altaf-Ul-Amin, Ken Kurokawa, Kotaro Minato, Shigehiko Kanaya, "Spherical SOM with arbitrary number of neurons and measure of suitability", *IEEE TRANSACTIONS ON NEURAL NETWORKS* (submitted)

International Conferences (Reviewed)

1. Hirokazu Nishio, Md. Altaf-Ul-Amin, Tetsuo Sato, Ken-nosuke Wada, Yoshiko Wada, Kotaro Minato, Kazuo Kobayashi, Naotake Ogasawara, Shigehiko Kanaya, "Visualization of gene classification based on expression profile using BL-SOM", *Proc. of WSOM03*, p. 101-106.

2. Hirokazu Nishio, Ken-nosuke Wada, Yoshiko Wada, Md. Altaf-Ul-Amin, Shigehiko Kanaya, "Suitability of Spherical SOM for Gene Expression Analysis", *Proc. of RECOMB2004*, p.79-80.

3. Hirokazu Nishio, Md. Altaf-Ul-Amin, Yukiko Nakamura, Takashi Abe, Makoto Kinouchi, Toshimichi Ikemura, Kazuo Kobayashi, Naotake Ogasawara, Shigehiko Kanaya, "Gene classification Based on Expression

Profile using BL-SOM: Suitability assessment of multivariate gene expression data to Spherical and Plain SOM by N-measure" *Proc. of SCI2004*, vol.8. p.189-192.

4. Hirokazu Nishio, Md. Ultaf-Ul Amin, Ken Kurokawa, Shigehiko Kanaya, "SphSOM: A Toolkit for Visualizw Intrinsic structure in high dimensional data space" *Proc. of APBC* (in press).

Contents

List of Figures

List of Tables

1 Introduction

Self-organizing Maps (SOMs) are clustering methods that can map high-dimensional data to low-dimensional representation space [3]. They are applied to broad range of fields, including pattern recognition, natural language processing, signal processing, data mining, brain theory, analysis of genome sequence and gene expression. SOM is a dimension reduction method similar to principal component analysis (PCA). Though PCA makes a linear mapping from high-dimensional data space to low-dimensional space, SOM makes a non-linear mapping. Because of this difference SOM has an ability to visualize high-dimensional adjacent relation. (see Fig. 1[11])

In the present study I consider two types of SOMs, one of them uses planar representation space called ' Plane SOM ' and the other uses spherical space called 'Spherical SOM.'When I compare genes based on expression profiles, I normalize profile vectors of gene expression to unity in length as a preprocess. It has advantage that genes are compared by the trends of expression profiles and are classified based on similarity of expression regulation in cells [18]. This 'nomalization metric' is almost same to popular 'cosine metric' using dot product.

$$d_c(\mathbf{u},\mathbf{v}) \equiv 1 - \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = 1 - \cos\theta \tag{1}$$

$$d_n(\mathbf{u},\mathbf{v}) \equiv \left|\frac{\mathbf{u}}{|\mathbf{u}|} - \frac{\mathbf{v}}{|\mathbf{v}|}\right| = 2\sin\frac{\theta}{2} \tag{2}$$

Here, θ is an angle between vector $\mathbf{u}$ and $\mathbf{v}$.

$$\begin{aligned} d_n(\mathbf{u},\mathbf{v})^2 &= \left(2\sin\frac{\theta}{2}\right)^2 && (3)\\ &= 4\sin^2\frac{\theta}{2} && (4)\\ &= 4\left(\frac{1-\cos\theta}{2}\right) && (5)\\ &= 2\left(1-\cos\theta\right) && (6)\\ &= 2d_c(\mathbf{u},\mathbf{v}) && (7) \end{aligned}$$

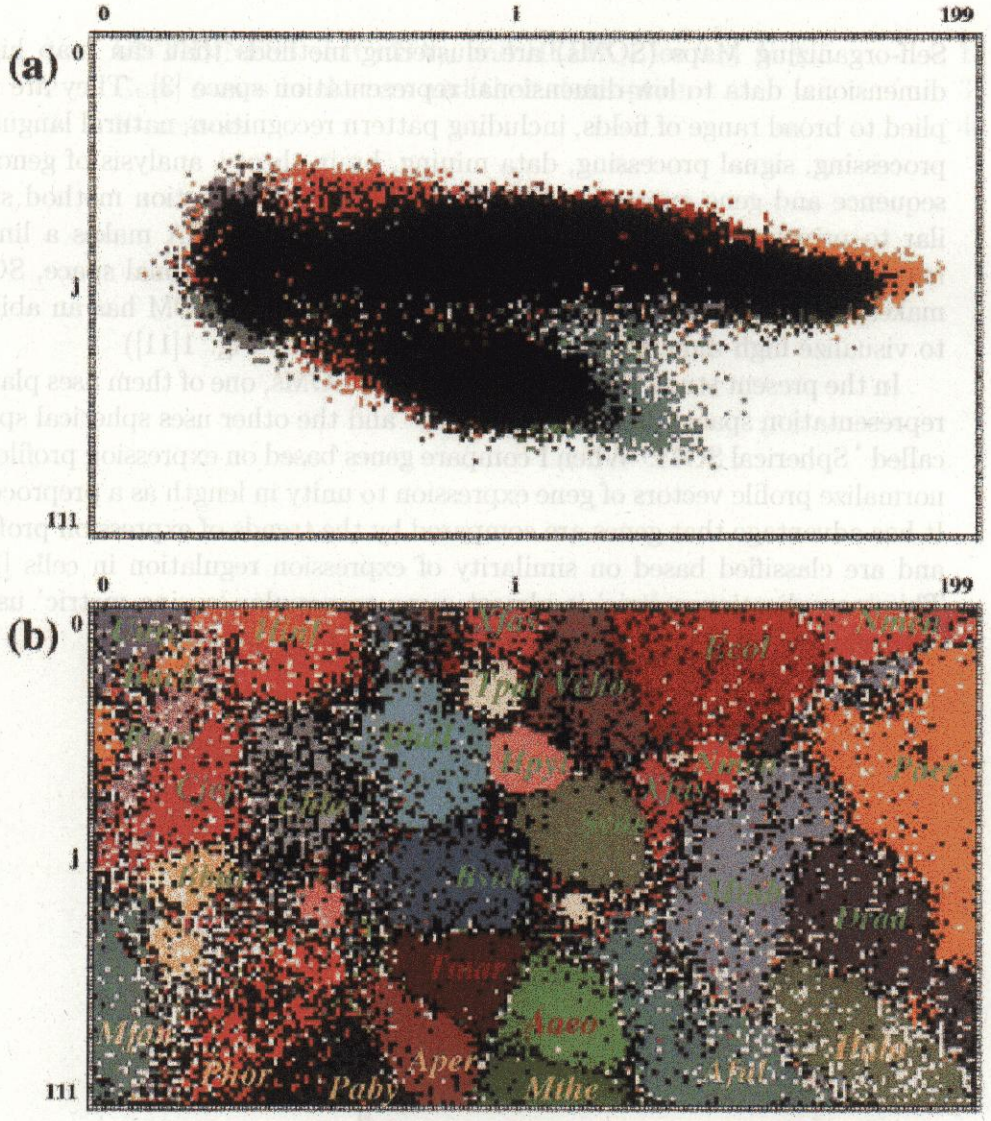


Figure 1: Visualization of codon usage by (a) PCA and (b) SOM

A. aeolicus (abbreviated as Aaeo), A. fulgidus (Aful), A. pernix (Aper), B. subtilis (Bsub), B. halodurans (Bhal), B. burgdorferi (Bbur), Buchnera sp. (Buch), C. jejuni (Cjej), C. trachomatis and C. pneumonia (Chla), D. radiodurans (Drad), E. coli (Ecol), H. influenzae (Hinf), Halobactgerium sp. (Halo), H. pylori (Hpyl), M. jannaschii (Mjan), M. thermoautotrophicum (Mthe), M. tuberculosis (Mtub), N. meningitidis (Nmen), P. aeruginosa (Paer), P. abyssi (Paby), P. horikoshii (Phor), R. prowazekii (Rpro), Synechocystis sp. (Syne), T. maritima (Tmar), T. pallidum (Tpal), U. urealyticum (Uure), V. cholerae (Vcho), and X. fastidiosa (Xfas). Archaea, eubacteria and two thermophilic bacteria are denoted by yellow, green, and blue letters.

Thus, $2d_c(\mathbf{u},\mathbf{v}) = d_n(\mathbf{u},\mathbf{v})^2$. Pearson product-moment correlation coefficient $C(\mathbf{x},\mathbf{y})$ is same to $1 - d_c(\mathbf{u},\mathbf{v})$ if we denote $\{x_i - \bar{x}\}$ by $\mathbf{u}$ and $\{y_i - \bar{y}\}$ by $\mathbf{v}$.

$$
\begin{aligned}
C(\mathbf{x},\mathbf{y}) &\equiv \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} && (8)\\
&= \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} && (9)\\
&= \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{u}||\mathbf{v}|} && (10)\\
&= 1 - d_c(\mathbf{u},\mathbf{v}) && (11)
\end{aligned}
$$

This 'normalization metric' has advantage that I can use many speed-up methods supposing Euclidean metric. Spherical SOM is adequate for clustering the normalized data.

2 Arrangement of arbitrary number of neurons

2.1 Introduction

Self-organization maps (SOM) are a kind of neural network, and are applied to many fields for understanding the distribution in high dimensional space. In many applications, its neurons are arranged on plain surface and there are borders surrounding the neuron arrangement. It causes a problem which is called 'Border effect' because the number of neighborhood neurons of a neuron near a border is different from that of a neuron near the center[3]. A sphere has no border. Though this problem can be solved by arranging neurons uniformly on the surface of a sphere[5], a disadvantage of the conventional method is that it cannot arrange neurons of arbitrary number. To solve this problem, I proposed a method to arrange arbitrary number of neurons by dividing a helix which goes around a sphere of unity radius (Fig. 3) into pieces of identical length [7]. While the arrangement has high uniformity, this method need numerical integration to calculate the length of helix. In the present paper I use the concept of generalized spiral points.

2.2 Method

2.2.1 Icosahedron recursive subdivision

In Spherical SOM proposed by Ritter, neurons are arranged by subdividing an icosahedron recursively [5]. Example of a triangle of an icosahedron and its three recursive divisions are shown in Figure 2.2.1. I refer to this arrangement of neurons as ICOSA_N, where N is the number of recursive subdivision. The ICOSA_N has $2 + 10 \cdot 4^N$ neurons (12, 42, 162, 642, 2562, 10242,...) Since it increases exponentially, I can't always arrange arbitrary number of neurons.

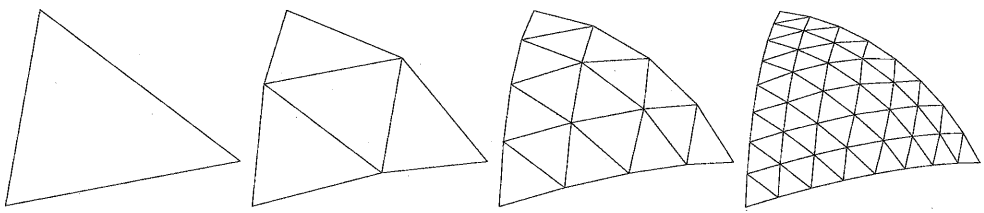


Figure 2: Recursive division of a triangle of an icosahedron

2.2.2 Division of helix on sphere

First, I consider a helix which goes around a sphere of unity radius. (See Figure 3.) The helix is determined by the equation below.

$$\theta = 2k\phi (0 \leq \phi \leq \pi)$$

In the above equation, k is the number of turns in the helix and spherical coordinates θ and ϕ can be transformed to orthogonal coordinates using the following equations. (See Figure 4.)

$$x = \cos\theta \sin\phi, y = \sin\theta \sin\phi, z = \cos\phi \tag{12}$$

There is no restriction for the value of k. In this paper I set k as follow

$$k \equiv \sqrt{N}$$

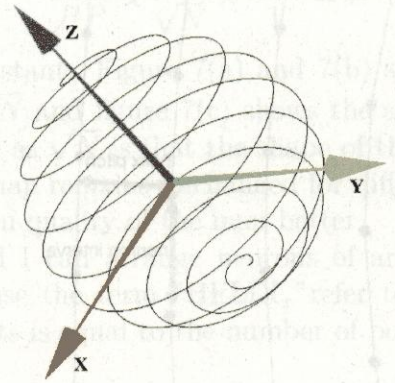


Figure 3: A helix on a sphere

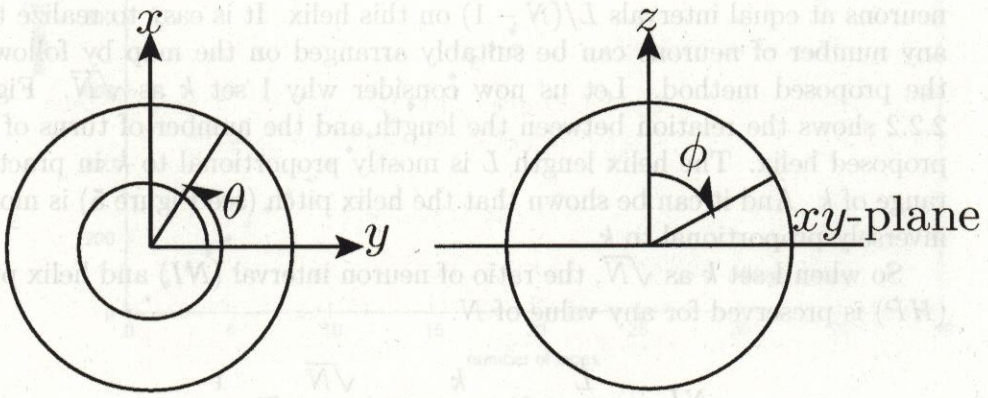


Figure 4: Relation between axis x, y, z, θ and ϕ

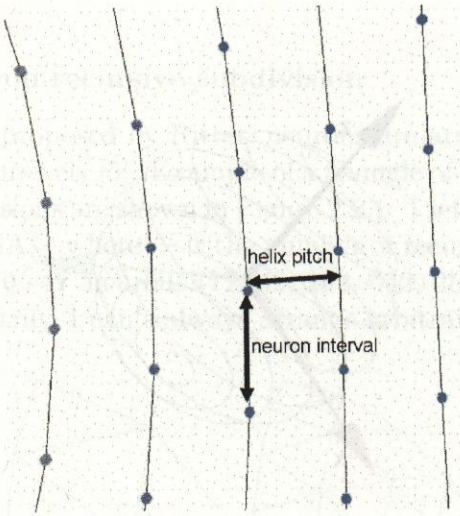


Figure 5: Helix pitch and neuron interval

where N is the number of neurons. The reason of this will be discussed later. Second, I calculate the helix length L. Since it is difficult to calculate it analytically, I calculate it by numerical integration. Finally, I arrange neurons at equal intervals $L/(N-1)$ on this helix. It is easy to realize that any number of neurons can be suitably arranged on the map by following the proposed method. Let us now consider why I set k as $\sqrt{N}$. Figure 2.2.2 shows the relation between the length and the number of turns of the proposed helix. The helix length L is mostly proportional to k in practical range of k. And it can be shown that the helix pitch (see Figure 5) is mostly inversely proportional to k.

So when I set k as $\sqrt{N}$, the ratio of neuron interval (NI) and helix pitch (HP) is preserved for any value of N.

$$NI = \frac{L}{N-1} \propto \frac{k}{N-1} = \frac{\sqrt{N}}{N-1} \approx \frac{1}{\sqrt{N}}$$

and,

$$HP \propto \frac{1}{k} = \frac{1}{\sqrt{N}}$$

therefore,

$$\frac{NI}{HP} \propto \frac{1}{\sqrt{N}} \times \frac{\sqrt{N}}{1} = 1$$

i.e. $\frac{NI}{HP}$ is almost constant. Figure 7(a) and 7(b) shows the arrangements of neurons for $k = \sqrt{N}$ and figure 7(c) shows the same for $k \neq \sqrt{N}$. The advantage of setting k as $\sqrt{N}$ is that the shape of the dots representing the neurons on the final map remains unchanged for different values of N. This makes the visualization quality of the map better.

Using this method I can arrange neurons of arbitrary number on the spherical surface. I use the term "HELIX$_x$"refer to refer an arrangement which number of points is equal to the number of points in ICOSA$_x$.

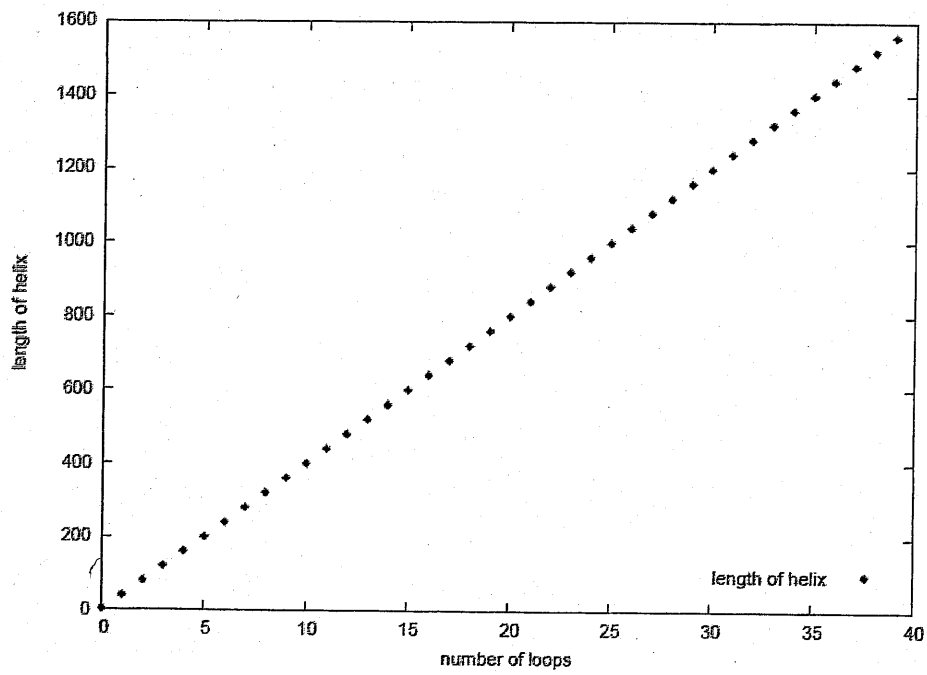
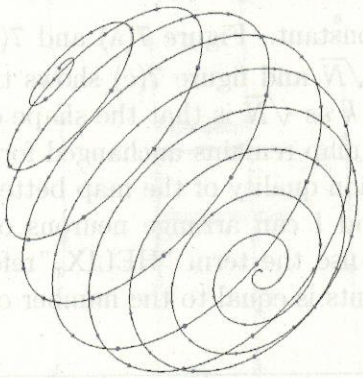
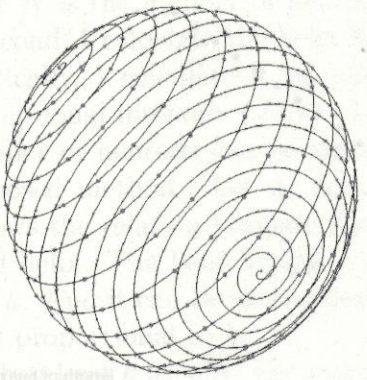


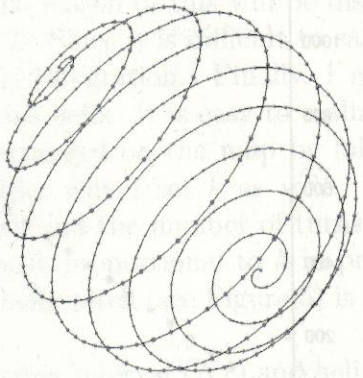
Figure 6: Relation between number of loops k and the helix length L



(a) Arrangement when $N = 100$ and $k = 10$



(b) Arrangement when $N = 400$ and $k = 20$



(c) Arrangement when $N = 400$ but $k = 10$

Figure 7: Relation between arrangement and parameters (N: number of neurons, k: number of loops)

2.2.3 Generalized spiral points

The generalized spiral points are explicitly defined sets of points proposed by E. A. Rakhmanov et al. [6]; where the set of points $\{(\theta_k, \phi_k)|0 \leq k < N\}$ is determined by the Eq. 13.

$$\begin{aligned} h_k &\equiv \frac{2k}{N-1} - 1, && (0 \leq k < N) \\ \phi_k &\equiv \arccos(h_k), && (0 \leq k < N) \\ \theta_0 &\equiv \theta_{N-1} \equiv 0, && \\ \theta_k &\equiv \left(\theta_{k-1} + \frac{C}{\sqrt{N(1-h_k^2)}}\right). && (1 \leq k < N-1) \end{aligned} \tag{13}$$

Here, C is a constant and was chosen as $C = 3.6$ so that the distance between successive point will be approximately the same. The value of h_k corresponds to $-x$ or $\cos(\phi_k)$, and the value of $\sqrt{1-h_k^2}$ correspons to $\sin(\phi_k)$. Spherical coordinates θ and ϕ can be transformed to orthogonal coordinates by Eq. 12. (See Figure 4.)

Using the generalized spiral points I can arrange neurons of arbitrary number on the spherical surface. I use the term "SPIRAL$_x$"refer to refer an arrangement which number of points is equal to the number of points in ICOSA$_x$.

2.2.4 Uniformity of neuron arrangement

In this section I will compare ICOSA_x with HELIX_x in the context of uniformity of the arrangements of the neurons. I use HELIX_N to mean the arrangement by proposed method which has the same number of neurons as ICOSA_N has. Uniformity means the equality of the number of neighborhood neurons in the context of all the neurons in the map. The map is completely uniform in the case when the number of neighborhood neurons is the same for each of the neurons of the map. Let $f(\eta, r)$ is the numbers of neurons which are within radius r from neuron η. (See Figure 8.)

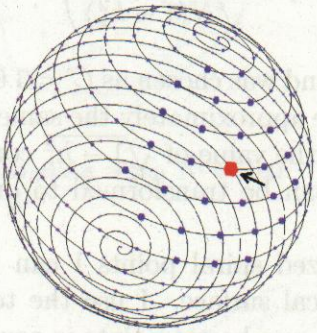


Figure 8: 28 neurons are within radius 0.5 from the indicated neuron η i.e. f(η, 0.5)=28.

And let $V(r)$ is the variance of $f(\eta, r)$ for all neurons i.e.

$$V(r) = \sum_{\eta} \frac{(f(\eta, r) - \overline{f(\eta, r)})^2}{n}$$

The variance $V(r)$ can only be a real positive number or zero. When the number of neighborhood neurons $f(\eta, r)$ is the same for all neuron η, $V(r)$ is zero. The relations between the radius r and the variance $V(r)$ are shown in Figure 10(a). It shows $V(r)$ in the case of helix is mostly lower than that in the case of icosahedron. I propose UNTIDINESS as a measure of uniformity and define UNTIDINESS as follows:

$$\mathrm{UNTIDINESS} \equiv \int_{\theta=0}^{\theta=\pi/2} V\left(2\sin\left(\theta/2\right)\right) \mathrm{d}\theta$$

Here $2\sin(\theta/2)$ is the length of chord for center angle θ.

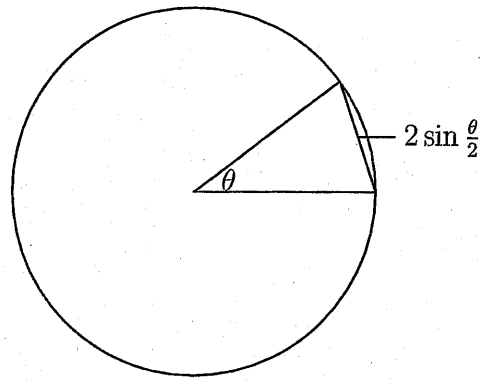


Figure 9: center angle θ and length of chord $2\sin\frac{\theta}{2}$

UNTIDINESS can only be a real positive number or zero. The more uniform is the neuron arrangement the less is the value of untidiness. When the arrangement is completely uniform UNTIDINESS is zero.

2.3 Result and discussion

I calculate the value of $V(r)$ in range $0 \leq \theta \leq \frac{\pi}{2}$ i.e. $0 \leq r \leq \sqrt{2}$ for (a) ICOSA_3 and HELIX_3, (b) ICOSA_4 and HELIX_4 and (c) ICOSA_5 and HELIX_5. The relations between the radius r and the variance $V(r)$ are shown in Figure 10(a). It shows $V(r)$ in the case of helix is mostly lower than that in the case of icosahedron.

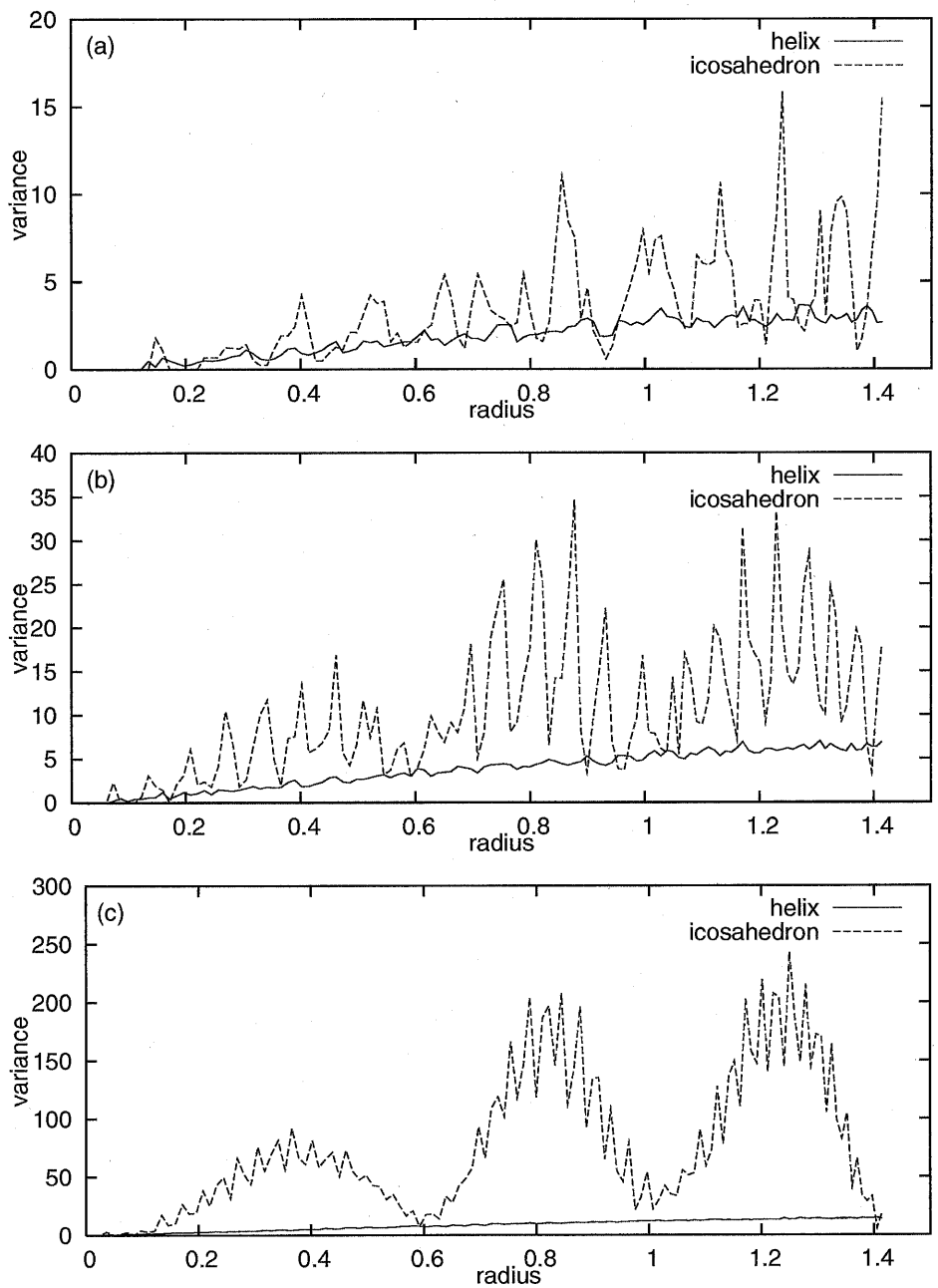


Figure 10: Relation between the radius and the variance. (a) $ICOSA_3$ and $HELIX_3$ (b) $ICOSA_4$ and $HELIX_4$ (c) $ICOSA_5$ and $HELIX_5$

The following table 1 is the UNTIDINESS of ICOSA_N and HELIX_N for N in range $0 \leq N \leq 5$.

Table 1: UNTIDINESS of ICOSA_N and HELIX_N for $0 \leq N \leq 5$

N	UNTIDINESS of ICOSA_N	UNTIDINESS of HELIX_N
0	0.0000	0.3611
1	0.4253	0.4539
2	2.8707	1.3165
3	3.9360	2.3874
4	10.4419	4.4689
5	82.1324	9.5022

ICOSA_0 is completely uniform because it is icosahedron itself. And ICOSA_1 has smaller UNTIDINESS than HELIX_1. Since UNTIDINESS of ICOSA_N increases quickly, it is larger than that of HELIX_N when N is two or more. Figure 11 shows the relation between the number of neurons and UNTIDINESS. By ICOSA method I can't arrange arbitrary number of neurons. So the UNTIDINESS for the permissible number of neurons are shown by crosses on the dotted line. On the other hand, I can arrange any number of neurons by HELIX method and the solid line shows the relation between the number of neurons and untidiness. The UNTIDINESS in case of the proposed method is much lower compared to the ICOSA method. The result shows that the proposed method has two advantages, (i) ability to select an arbitrary number of neurons which is not possible in the conventional icosahedron method, and (ii) improved uniformity of the arrangement of neurons. The arrangement by the proposed approach is more uniform than the icosahedron approach when the number of neurons are 43 or more. Thus the proposed approach relaxes the usability and improves the usefulness of the Spherical SOM. It is considered that the regularity of arrangement may be important in view of visibility, and conventional method might be more regular. The variance of ICOSA_N is less than that of HELIX_N on some particular radii. It means that both regularity and uniformity are satisfied when I use those radii only. The SOM which use conventional method can avoid border effect by using those radii for learning process.

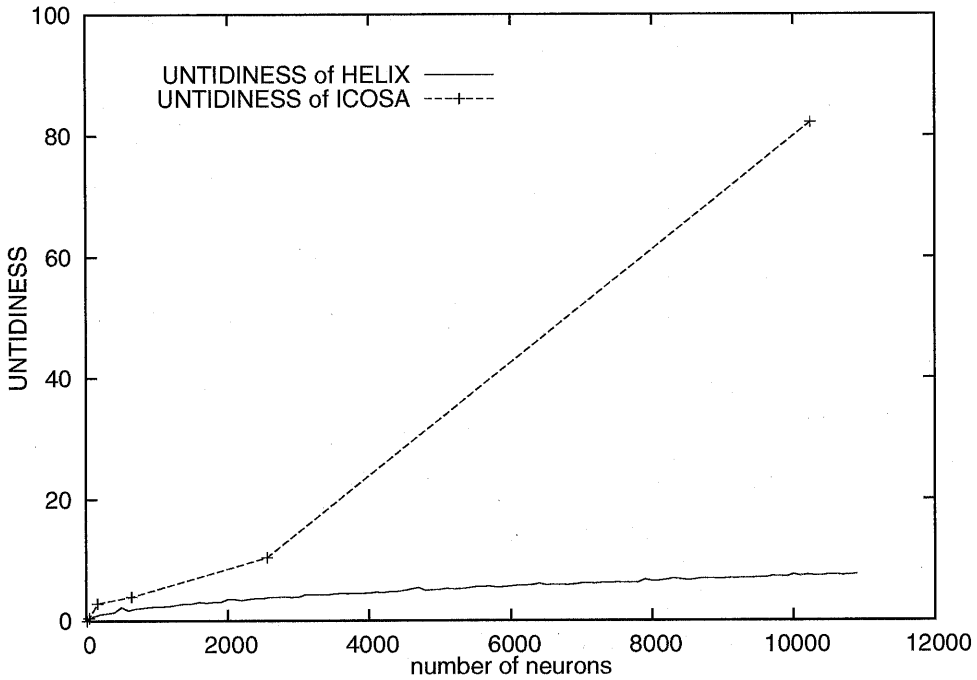


Figure 11: Relation between UNTIDINESS and the number of neurons (ICOSA and HELIX)

Figure 12 shows the UNTIDINESS of HELIX and SPIRAL. The UNTIDINESS in case of the SPIRAL method is much lower compared to the HELIX method in most case. The result means the arrangement by the SPIRAL approach is more uniform than the HELIX approach.

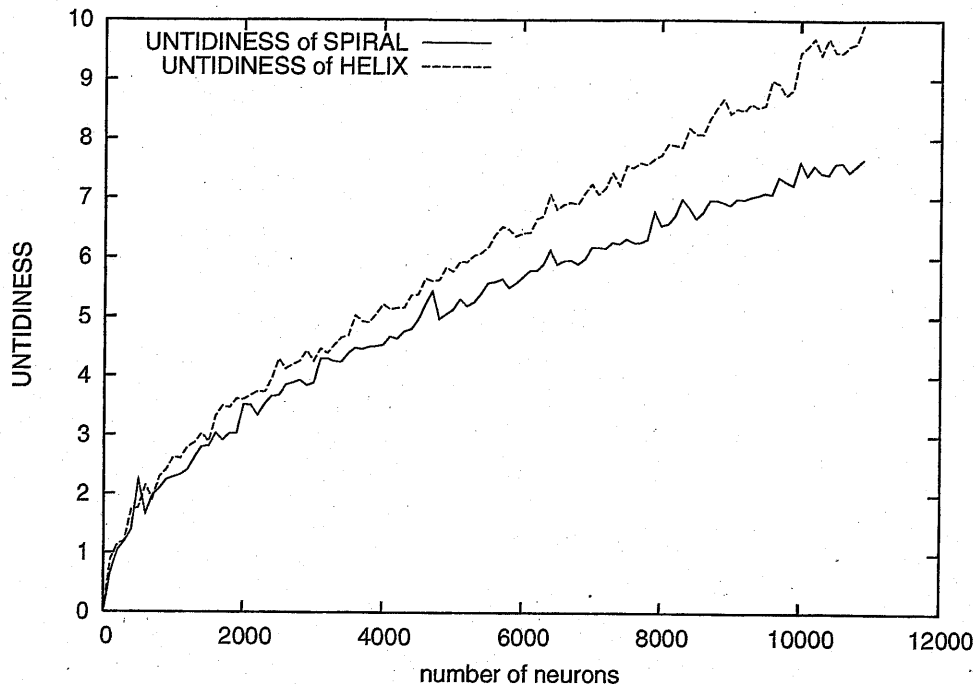


Figure 12: Relation between UNTIDINESS and the number of neurons (HELIX and SPIRAL)

2.4 Conclusion

In this paper I propose a new approach for the arrangement of neurons concerning spherical SOM. This approach enables us to suitably arrange an arbitrary number of neurons which is not possible in conventional icosahedron method. Another advantage of the proposed approach is the uniformity of the arrangement. The arrangement by the proposed approach is more uniform than the icosahedron approach when the number of neuron are 43 or more. Thus the proposed approach relaxes the usability and improves the usefulness of the spherical SOM.

3 Spherical SOM

3.1 Introduction

SOM is a dimension reduction method and a clustering method. This remarking feature comes from that a neuron has two vectors; position and weight. In this section I talk about how to initialize weight vectors and how a SOM learns input data.

3.2 Method

3.2.1 Initialization of weight vectors with PCA

Initialization of weight vectors with random values is not preferred because of the reproducibility for maps. In this work, the weight vectors of Spherical SOM are initialized on the basis of principal component analysis (PCA) for original data. Initially, principal axes $\mathbf{e_1}$, $\mathbf{e_2}$ and $\mathbf{e_3}$ with the three largest variance are estimated using the original data. Then, the weight vector $\mathbf{w_i}$ is initialized as

$$\mathbf{w_i} \equiv R(x_i\mathbf{e_1} + y_i\mathbf{e_2} + z_i\mathbf{e_3}). \tag{14}$$

Here R is the radius of hypersphere. For the value R I use the average length of input vectors projected to 3-dimensional space using PCA.

$$R \equiv \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\mathbf{v_i}\cdot\mathbf{e_1})^2 + (\mathbf{v_i}\cdot\mathbf{e_2})^2 + (\mathbf{v_i}\cdot\mathbf{e_3})^2} \tag{15}$$

Here $\mathbf{v_i}$ is the i-th input data vector.

The result of SOM when $R = 1$ is not much different. It seems that the value of R doesn't exert a strong influence on the result of the SOM.

3.2.2 Learning process and neighborhood function

Each neuron in SOM has two vector, an weight vector and a position vector. Let S be the dimension of input data vectors. An weight vector is a S-dimensional vector and a position vector is more low-dimensional, usually one, two or three dimensional vector. The learning process of SOM for a teacher vector is realized by the following algorithm.

1. Select BMU (Best Match Unit), or "winner neuron", the neuron η_i which has the nearest weight vector $\mathbf{w_i}$ to the teacher vector $\mathbf{t}$.

$$\mathbf{w_{win}}, \quad \text{such that} \quad d_d(\mathbf{t}, \mathbf{w_{win}}) = \min_i(d_d(\mathbf{t}, \mathbf{w_i})) \tag{16}$$

Here d_d is a distance function in data space. I use the normalization metric d_n in order to take account to the correlation.

$$d_n(\mathbf{u}, \mathbf{v}) \equiv \left| \frac{\mathbf{u}}{|\mathbf{u}|} - \frac{\mathbf{v}}{|\mathbf{v}|} \right| \tag{17}$$

2. Update weight vectors according to following equation

$$\mathbf{w_i} \leftarrow \mathbf{w_i} + \alpha(t)h(t)(\mathbf{t} - \mathbf{w_i}) \tag{18}$$

where h is the neighborhood function and α is the learning rate function. Both function depends on time.

The neighborhood function can be of type:

- "crisp":

$$h(t) = \begin{cases} 1 & (d_r(\mathbf{p_i}, \mathbf{p_{win}}) \leq r(t)) \\ 0 & (d_r(\mathbf{p_i}, \mathbf{p_{win}}) > r(t)) \end{cases} \tag{19}$$

- "fuzzy":

$$h(t) = \exp\left(\frac{-d_r(\mathbf{p_i}, \mathbf{p_{win}})}{2r(t)^2}\right) \tag{20}$$

Here d_r is a distance function in representation space. In original Plane SOM it is the box metric d_b

$$d_b(\mathbf{u}, \mathbf{v}) \equiv \max(\mathbf{u}_1 - \mathbf{v}_1, \mathbf{u}_2 - \mathbf{v}_2) \tag{21}$$

but it is not suitable to Spherical SOM. I use Euclidean metric for the distance function in representation space.

$$d_e(\mathbf{u}, \mathbf{v}) \equiv \sqrt{\sum_i (\mathbf{u}_i - \mathbf{v}_i)^2} \tag{22}$$

I use a crisp type of neighborhood function because it saves the amount of calculation.

3.2.3 Batch learning

The original SOM has a problem that the result depends on the order of input data. It is solved by using Batch Learning[10, 11].

1. Initialize "weight correction vectors" $\mathbf{u_i}$ with zero vector.

$$\mathbf{u_i} \leftarrow \mathbf{0} \tag{23}$$

2. Initialize "number of correction" n_i with zero.

$$n_i \leftarrow 0 \tag{24}$$

3. For each input data $\mathbf{t}$ do:

   (a) Select the neuron η_i which has the nearest weight vector $\mathbf{w_i}$ to the teacher vector $\mathbf{t}$.

   (b) Update "weight correction vectors" and "number of correction" according to following equation

$$\mathbf{u_i} \leftarrow \mathbf{u_i} + \alpha(t)h(t)(\mathbf{t} - \mathbf{w_i}) \tag{25}$$

$$n_i \leftarrow n_i + 1 \qquad \text{if } h(t) > 0 \tag{26}$$

4. Update weight vectors according to the rule:

$$\mathbf{w_i} \leftarrow \mathbf{w_i} + \mathbf{u_i}/n_i \tag{27}$$

Batch Learning SOM (BL-SOM) which proposed by Abe and others is accompanied by initialization of weight vector using PCA. I adopt both batch learning and PCA based initialization.

3.2.4 SphSOM package

I made 'SphSOM package', a toolkit for visualization of the high dimensional data space using proposed method to arrange arbitrary number of neurons. SphSOM package contains two components: (i) SphSOM is visualization toolkits based on spherical SOM, customizable with Python script, that is, without any development environments and (ii) SphSOMViz is pure Java applet which allows publishing your 3-dimensional visualization result on a web site. This software are available on http://www.nishiohirokazu.org/sphsom/.

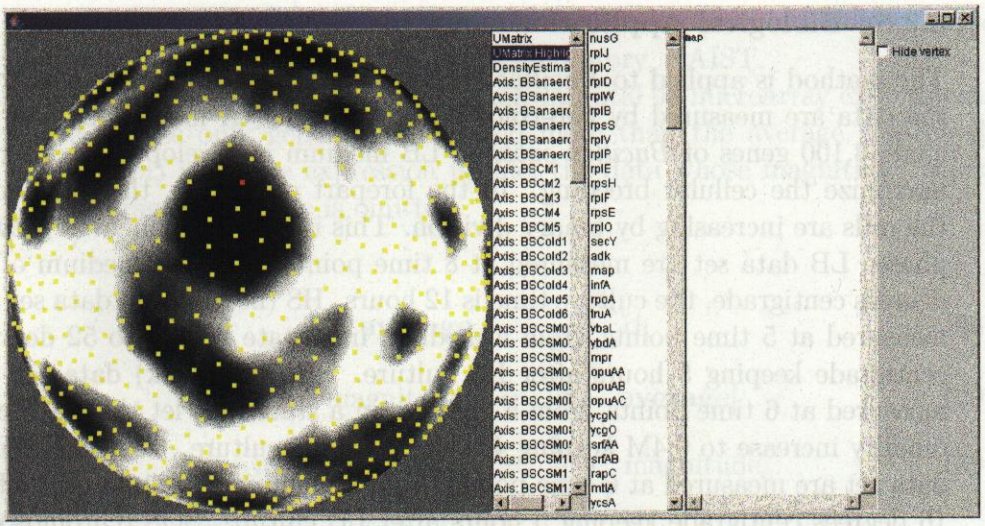


Figure 13: SphSOMViz screenshot 1

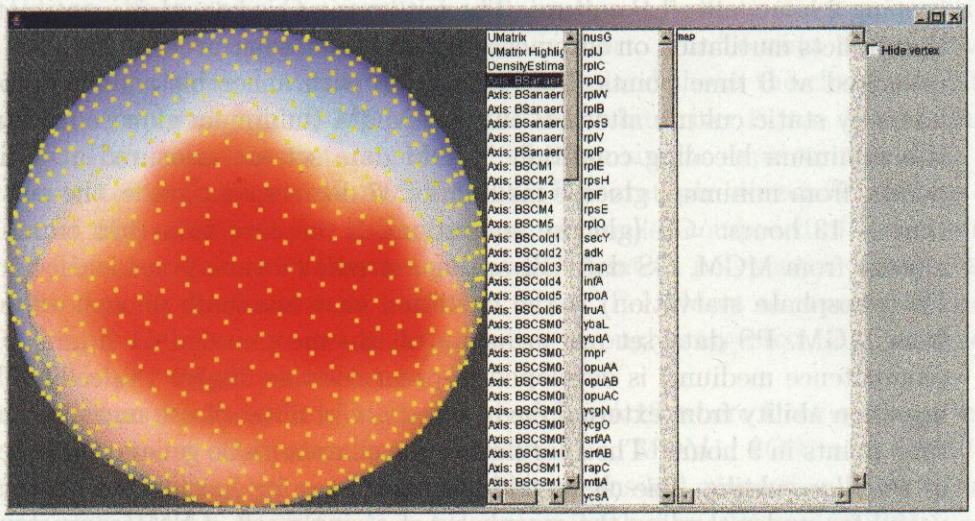


Figure 14: SphSOMViz screenshot 2

3.2.5 Biological Application

The method is applied to *Bacillus subtilis* microarray time series data. All the data are measured by two-color fluorescence cDNA microarray and include 3,100 genes of *Bacillus subtilis*. LB medium is developed in order to maximize the cellular breeding. In the forepart of culture, the number of the cells are increasing by binary division. This is called exponential growth phase. LB data set are measured at 8 time points, from LB medium of 37 degrees centigrade, the culture time is 12 hours. HS (heat shock) data set are measured at 5 time points, from a medium in a state heated to 52 degrees centigrade keeping 5 hours after LB culture. SS (salt shock) data set are measured at 6 time points, from a medium in a state that let the NaCl cardinality increase to 0.4M keeping 4 hours after LB culture. CS (cold shock) data set are measured at 6 time points, from a medium in a state cool off to 16 degrees centigrade keeping 5 hours after LB culture. SOS (chromosome damege by mitomycin C) data set are measured at 6 time points, from a medium in a state that let the mitomycin C cardinality increase to 1 $\mu g/ml$ keeping 2 hours after LB culture. The mitomycin C is known as a antibiotics that inflicts mutilation on chromosomes. AG (anaerobic growth) data set are measured at 6 time points, from a medium with anaerobic condition for 2 hours by static culture after LB culture. MGM (minimum glucose medium) is a minimum bleeding condition. MGM data set are measured at 8 time points, from minimum glucose medium of 37 degrees centigrade, the culture time is 13 hours. GS (glucose starvation) is the condition that eliminate glucose from MGM. GS data set are measured at 5 time poins in 10 hours. PS (phosphate starvation) is the condition that eliminate phosphoric acid from MGM. PS data set are measured at 6 time poins in 11 hours. CM (competence medium) is developed in order to make higher the cell's DNA ingestion ability from external environment. CM data set are measured at 5 time points in 9 hours. The sporulation is a characteristic generation process of *Bacillus subtilis*. The characterization of genes for sporulation process is useful for understanding the morphosis of a single cell. CSM (competence-sporulation medium) is a kind of medium which promotes sporulation. CSM data set are measured at 13 time points in 15 hours. DSM (Difco sporulation medium) is another kind of medium which promotes sporulation. DSM data set are measured at 19 time points in 12 hours. DGG (Difco glucose glutamine) is a medium that glucose and glutamine are added to DSM medium in order to inhibit spoluration. another kind of medium which promotes

sporulation. DGG data set are measured at 6 time points in 9 hours. These data were offered from the Ogasawara laboratory, NAIST.

I examined gene expression profiles from total 98 microarray experiments above. I use only genes that changed more than the average. Given $\mathbf{d_i}$ $(1 \leq i \leq n)$ is a gene expression profile, the data whose magnitude_i is less than averageMagnitude is omitted.

$$\mathbf{average} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d_i} \tag{28}$$

$$\text{magnitude}_i = |\mathbf{d_i} - \mathbf{average}| \tag{29}$$

$$\text{averageMagnitude} = \frac{1}{n}\sum_{i=1}^{n}\text{magnitude}_i \tag{30}$$

3.3 Result and discussion

Figures 15-18 show the result of visualization. A black color means the distance between neurons in data space is short. The neurons have a character to gather in data cluster. So the dark island shows the cluster in data space. Two dark islands separated by broad white river is two distant clusters. Those clusters contain genes which have similar expression profiles. I illustrated it in Figure 19 and Table 2. For example cluster A contains a lot of ribosomal protein. Ribosomes are organelles composed of rRNA and ribosomal proteins and they translate mRNA into a polypeptide chain. The infA gene is the translation initiation factor IF-1. Most of genes in cluster A and B are related to translation, so it is strongly suggested that the other genes also related to translation.

Figures 20-23 is the same position of figure 19 where I colored by a expression value of a paticular experiment LB1, MGM1, SOS1 and CSM01. Red color means high expression and blue means low expression. Most of expression of genes in cluster A and B are similar but in LB1 and CSM01 I can observe the difference of expression. This suggests that there is a difference of some functions between those genes.

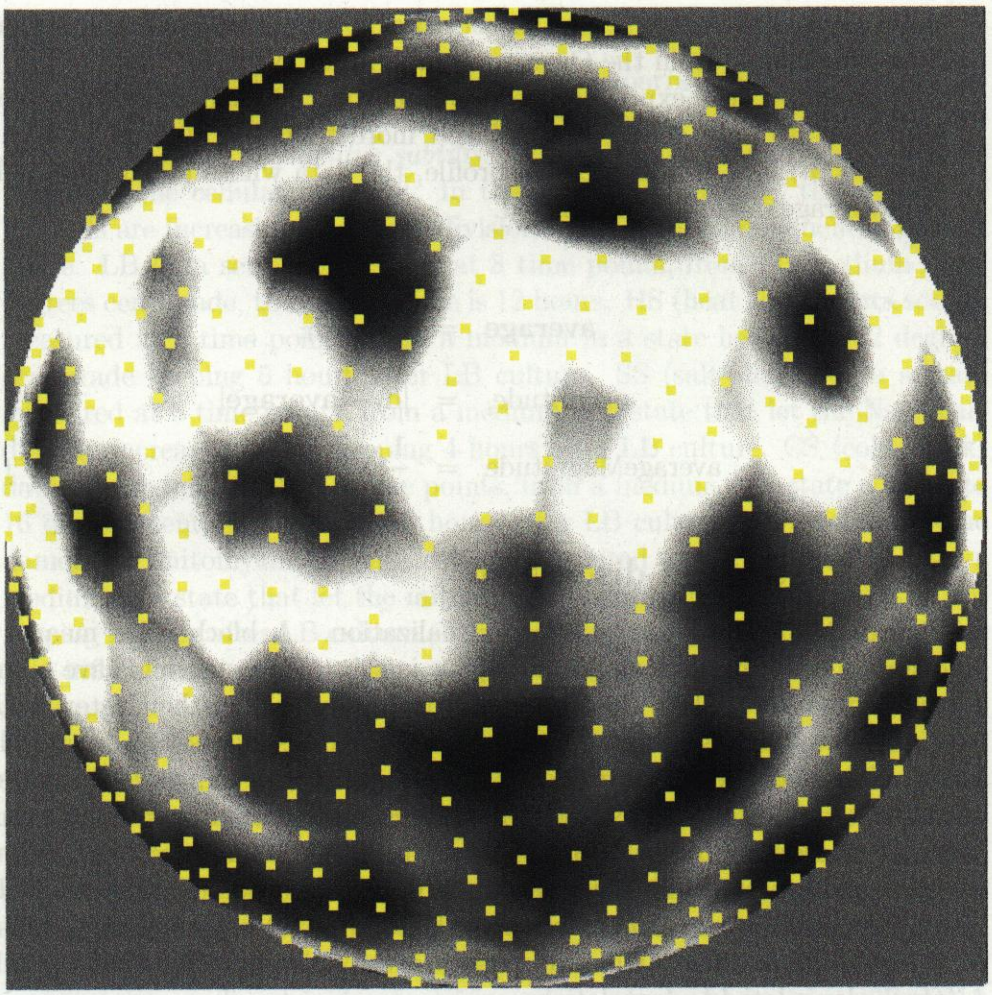


Figure 15: Result of spherical SOM 1

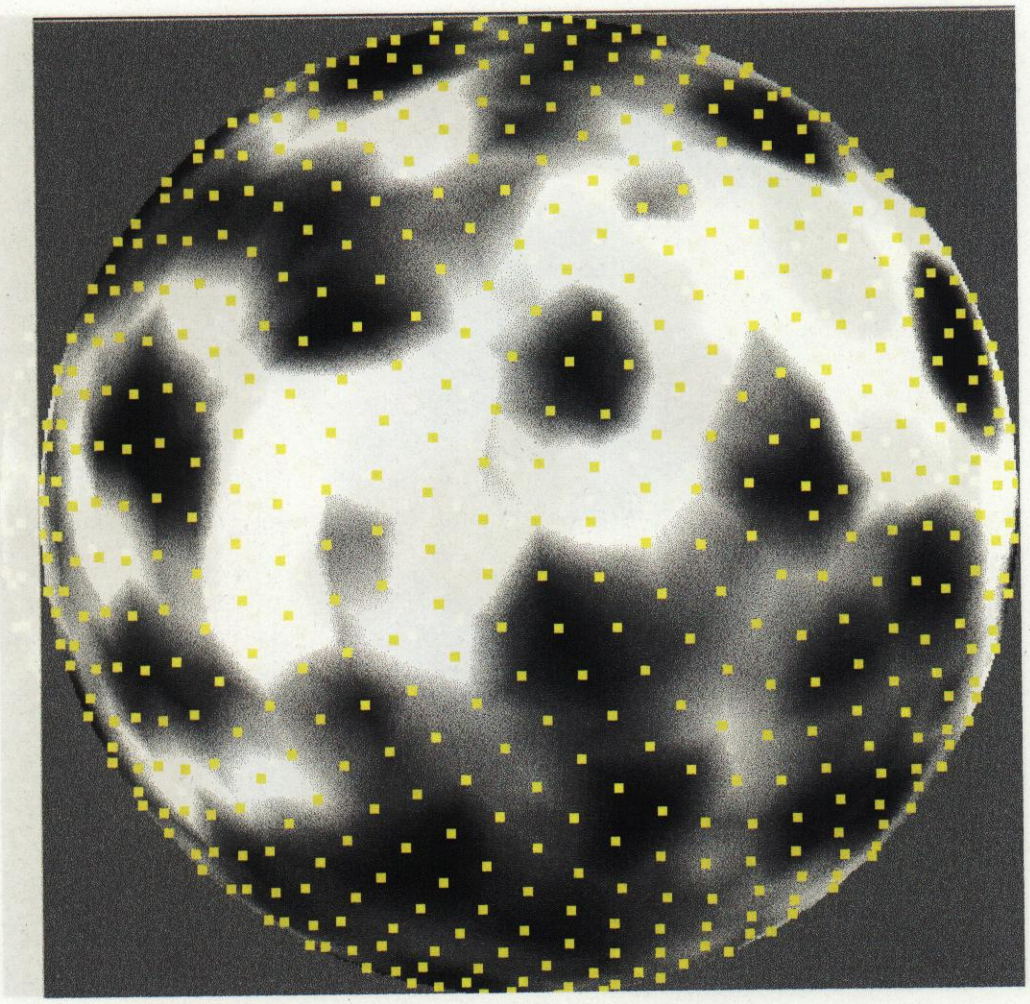


Figure 16: Result of spherical SOM 2

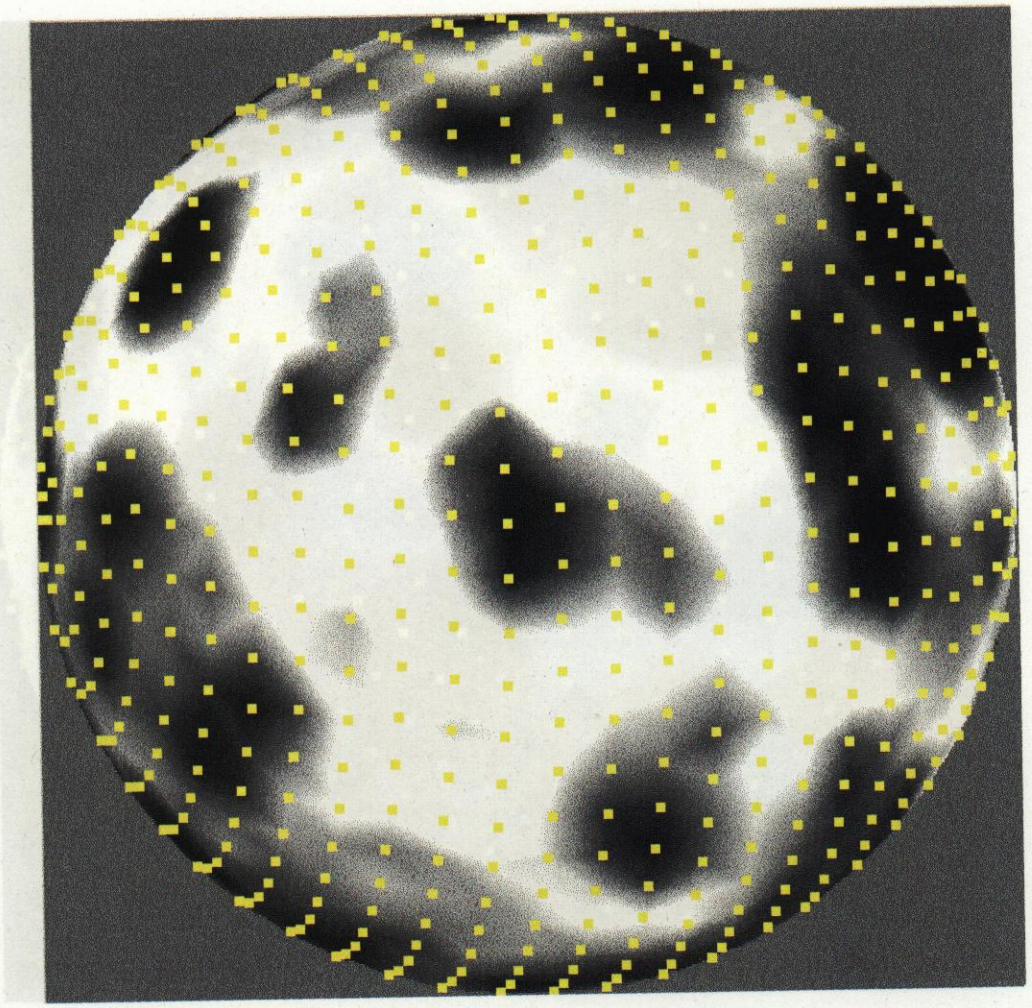


Figure 17: Result of spherical SOM 3

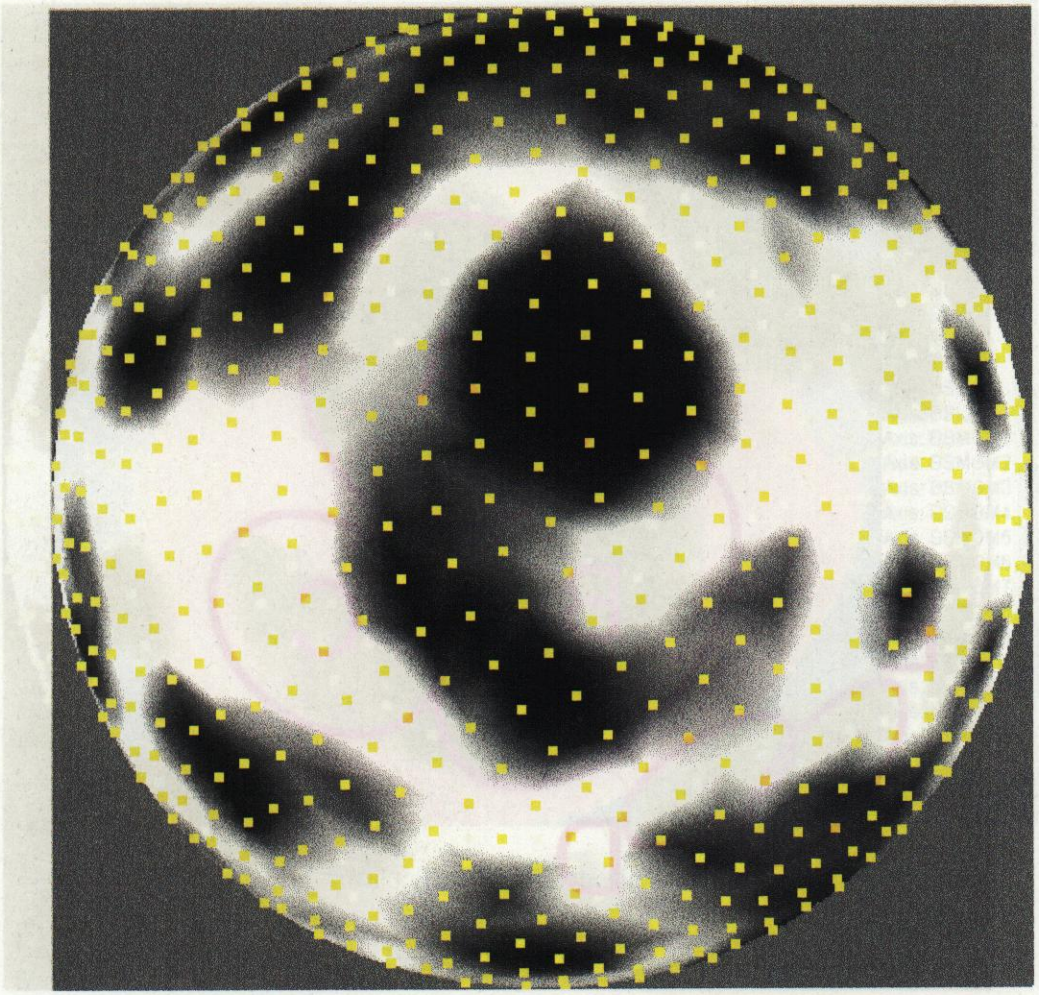


Figure 18: Result of spherical SOM 4

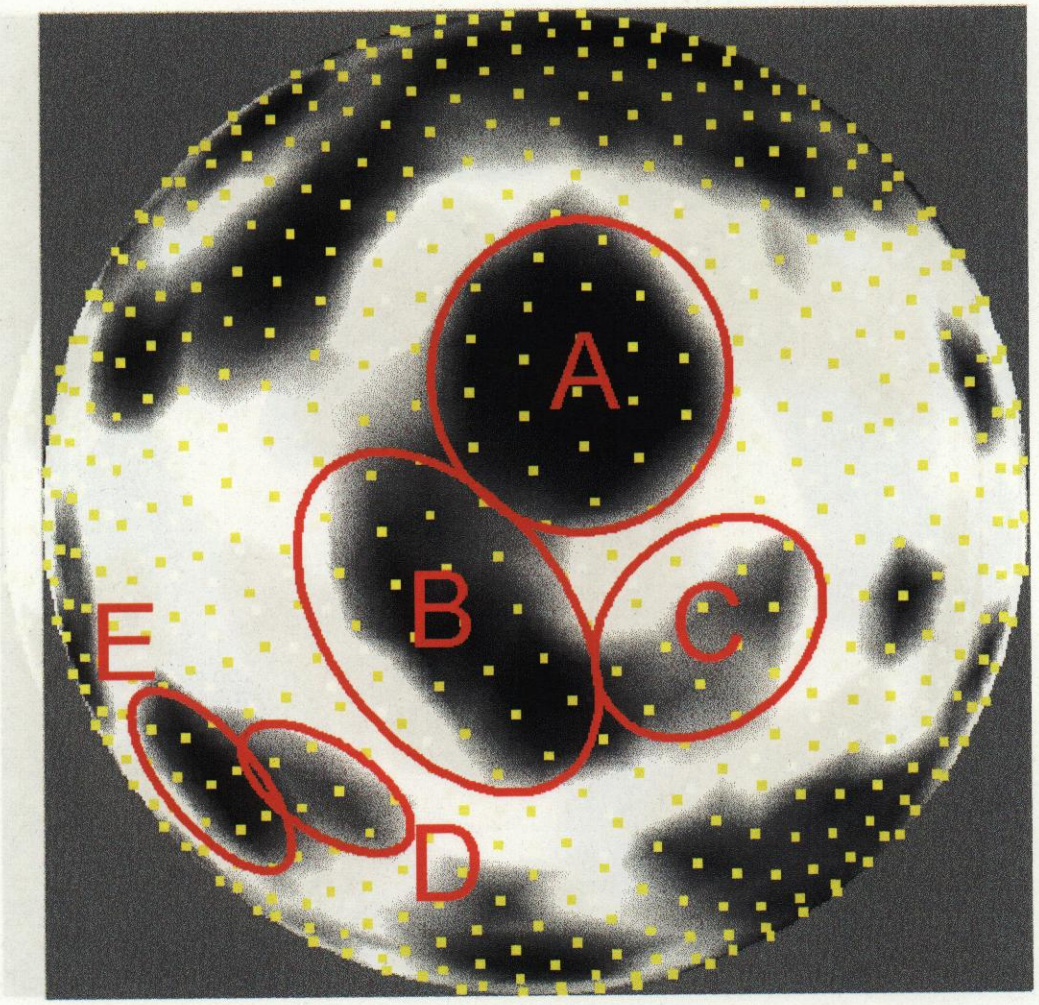


Figure 19: Result of spherical SOM: some clusters

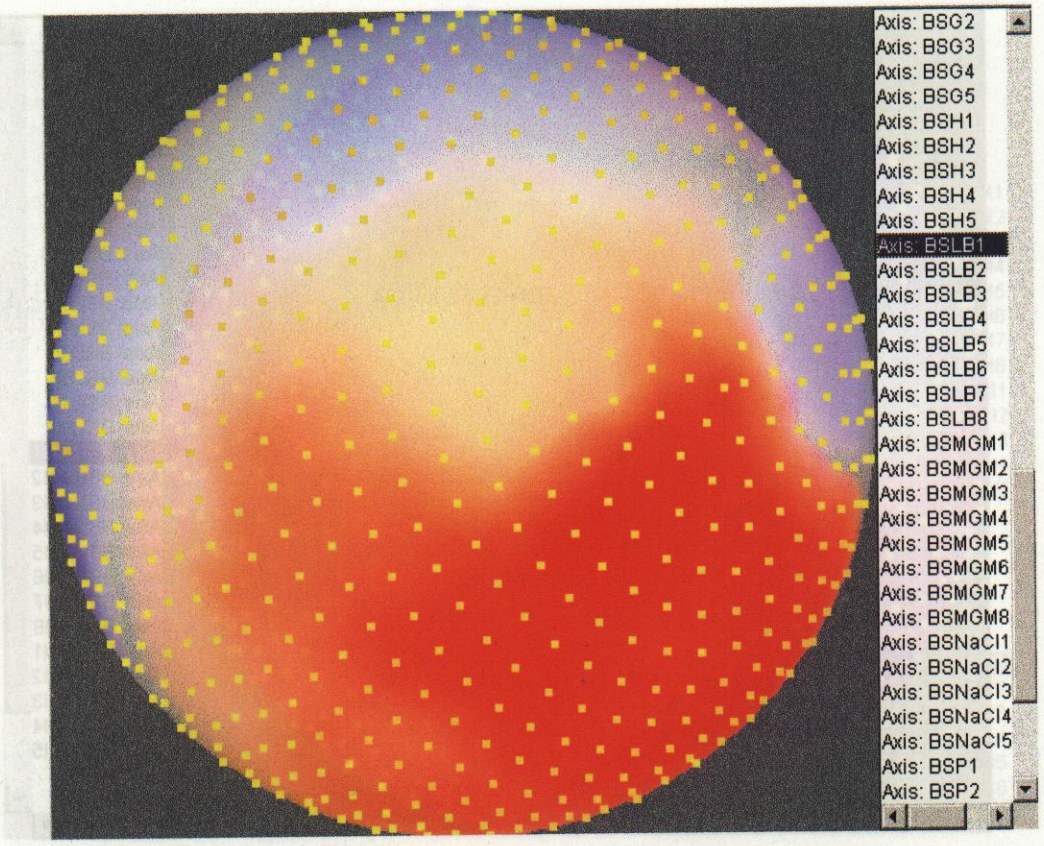


Figure 20: Result of spherical SOM: LB1

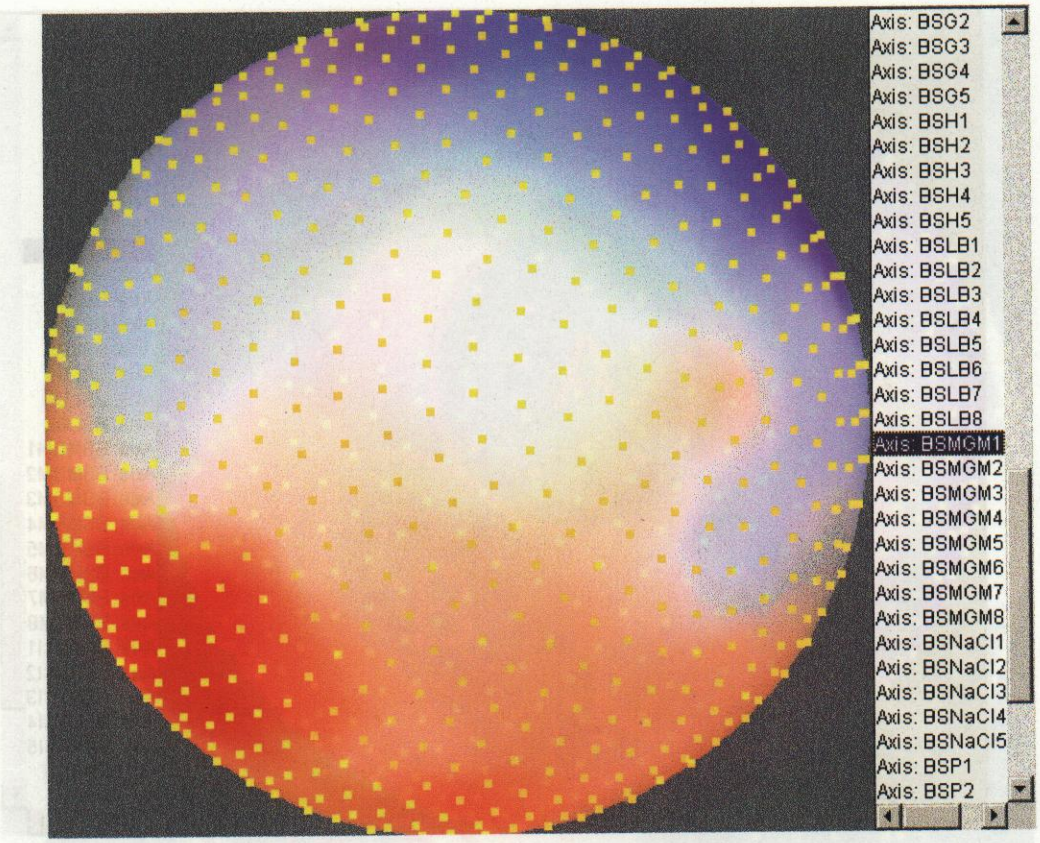


Figure 21: Result of spherical SOM: MGM1

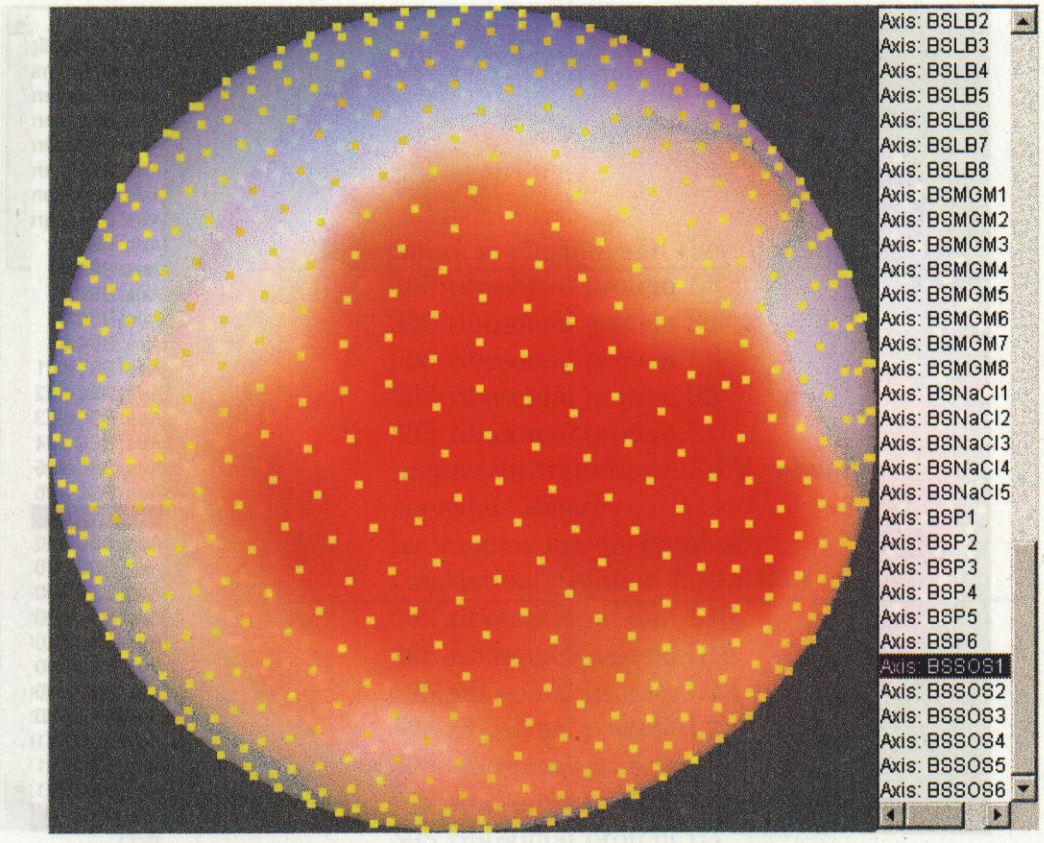


Figure 22: Result of spherical SOM: SOS1

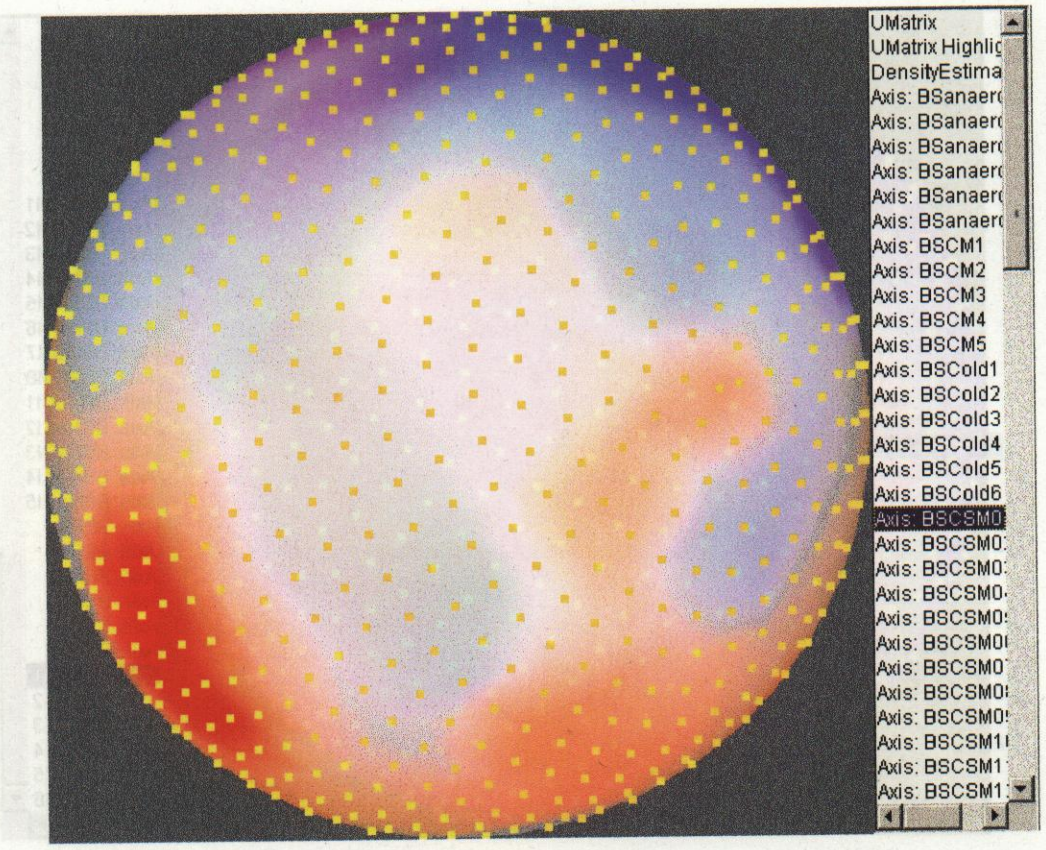


Figure 23: Result of spherical SOM: CSM01

3.4 Conclusion

I develop software to visualize high dimensional data using spherical SOM and publish it. I examined actual gene expression profiles from 98 microarray experiments. The result shows that Spherical SOM can visualize 98 dimensional expression profiles.

Table 2: Gene names in clusters and their description

A	rplE	50S ribosomal protein L5
	rpsH	30S ribosomal protein S8
	rplF	50S ribosomal protein L6
	rpsE, spcA	30S ribosomal protein S5
	rplO	50S ribosomal protein L15
	secY	preprotein translocase SecY subunit
	adk	adenylate kinase
	map	methionine aminopeptidase
	infA	translation initiation factor IF-1
B	rplJ	50S ribosomal protein L10
	rplP	50S ribosomal protein L16
	rplW	50S ribosomal protein L23
	rplB	50S ribosomal protein L2
	rpsS	30S ribosomal protein S19
	rplV	50S ribosomal protein L22
	rplC	50S ribosomal protein L3
	rplD	50S ribosomal protein L4
C	truA, ybaH	pseudouridylate synthase I
	yrrM	SAM-dependent methyltransferases
D	pyrF	orotidine 5'-phosphate decarboxylase
E	purH, purJ, purHJ	phosphoribosylaminoimidazole carboxy formyl formyltransferase inosine-monophosphate cyclohydrolase
	purL	phosphoribosylformylglycinamidine synthase II
	purN	phosphoribosylglycinamide formyltransferase

4 Suitability assessment

4.1 Introduction

When there are some feature maps of a particular data set from different type of SOM, it is not evident which map is the most suitable for a given data set because I can't observe high dimensional distribution directly. So I propose a measure of suitability, N-measure.

4.2 Method

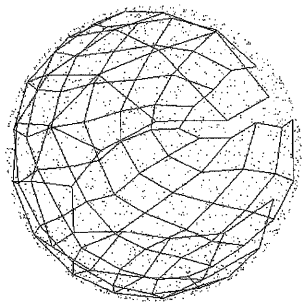
4.2.1 Linearity of distance-distance plot

A SOM is a non-linear mapping in essentials. When a mapping is generated by a learning process without distortion, the distances between neurons in spherical surface should be linearly related with those between them in the original space. The linearity means the quality of distance preservation, which is the main idea of dimension reduction methods, for example Multi-Dimensional Scaling.[9] This linearity can be a basis to determine whether spherical or planar arrangement of neurons is suitable for a specific data set. For any mapping Φ, I can observe the linearity using the d-d plot [8]. The d-d plot is a scatter plot two distances between all pairs of the elements.

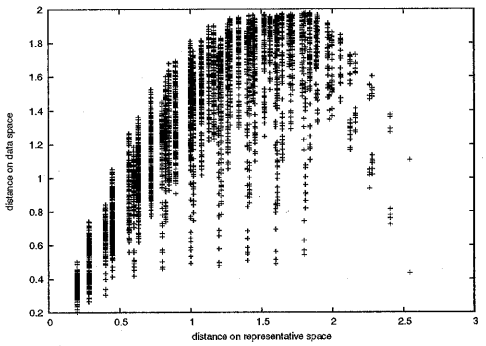
$$x_{uv} \equiv \|\Phi(\mathbf{u}) - \Phi(\mathbf{v})\| \tag{31}$$

$$y_{uv} \equiv \|\mathbf{u} - \mathbf{v}\| \tag{32}$$

Here $\mathbf{u}$ and $\mathbf{v}$ are data in the domain of the mapping Φ. In this paper, the mapping Φ is plane and sphrical SOM and its domain is the data space. For example (Fig. 24(a)), neurons on opposing corner of 2-dimensional plane arrangement are close in 3-dimensional space. So the linear relation is d-d plot (Fig. 24(b)) is broken because of the distortion occurred from projecting high dimensional data of original space to representation space.



(a)



(b)

The linearity is measured by correlation coefficient in the regression analysis fashion. In regression analysis the sum of the squares of differences between actual value y and the value $\hat{y}$ estimated with a given model $\hat{y}_i = f(x_i)$ is minimized (see Eq. 33.)

$$\text{minimize} \sum_i (y_i - \hat{y}_i)^2 \tag{33}$$

In this case we'd like to quantify the linearity of d-d plot, $\hat{y}_i = kx_i$. Here, k for minimal square difference is represented in Eq. 34.

$$k = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \tag{34}$$

The coefficient of determination R^2 in regression analysis is defined by Eq. 35.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{35}$$

Here $\bar{y}$ is the avarage of y_i;

$$\bar{y} = \frac{1}{N} \sum_i y_i \tag{36}$$

R^2 means how the given data fit to the given model. So the value shows how d-d plot satisfy the linear model $f(x) = kx$.

The N-measure $\mathcal{N}$ can be defined by combining these equations,

$$\mathcal{N} \equiv 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{37}$$

$$= 1 - \frac{\sum_i (y_i^2 - 2y_i\hat{y}_i + \hat{y}_i{}^2)}{\sum_i (y_i^2 - 2y_i\bar{y} + \bar{y}^2)} \tag{38}$$

$$= 1 - \frac{\sum_i (y_i^2) - 2\sum_i (y_i\hat{y}_i) + \sum_i (\hat{y}_i{}^2)}{\sum_i (y_i^2) - 2\sum_i (y_i\bar{y}) + \sum_i (\bar{y}^2)} \tag{39}$$

$$\tag{40}$$

To elide annoying sigmas I use notations as $S_y \equiv \sum_i y_i$, $S_{xx} \equiv \sum_i x_i^2$, $S_{xy} \equiv \sum_i x_i y_i$ and $S_{yy} \equiv \sum_i y_i^2$.

$$= 1 - \frac{S_{yy} - 2\sum_i (y_i\hat{y}_i) + \sum_i (\hat{y}_i{}^2)}{S_{yy} - 2\sum_i (y_i\bar{y}) + \sum_i (\bar{y}^2)} \tag{41}$$

$$= 1 - \frac{S_{yy} - 2k\sum_i (x_i y_i) + k^2\sum_i (x_i^2)}{S_{yy} - 2\bar{y}\sum_i (y_i) + N\bar{y}^2} \tag{42}$$

$$= 1 - \frac{S_{yy} - 2kS_{xy} + k^2 S_{xx}}{S_{yy} - 2S_y^2/N + S_y^2/N} \tag{43}$$

$$= 1 - \frac{S_{yy} - 2(S_{xy}/S_{xx})S_{xy} + (S_{xy}/S_{xx})^2 S_{xx}}{S_{yy} - 2S_y^2/N + S_y^2/N} \tag{44}$$

$$= 1 - \frac{S_{yy} - 2S_{xy}^2/S_{xx} + S_{xy}^2/S_{xx}}{S_{yy} - 2S_y^2/N + S_y^2/N} \tag{45}$$

$$= 1 - \frac{S_{yy} - S_{xy}^2/S_{xx}}{S_{yy} - S_y^2/N} \tag{46}$$

$$= 1 - \frac{NS_{xx}S_{yy} - NS_{xy}^2}{NS_{xx}S_{yy} - S_{xx}S_y^2} \tag{47}$$

$$= \frac{NS_{xy}^2 - S_{xx}S_y^2}{NS_{xx}S_{yy} - S_{xx}S_y^2} \tag{48}$$

$$\tag{49}$$

$\mathcal{N}$ takes value less than 1. The larger the value of $\mathcal{N}$, the higher the linearity.

4.2.2 Data set

In order to demonstrate the validity of N-measure, I examine three data sets as follows:

- D1: Randomly distributed 1800 points on a plane surface consisting of two axes. Variance of each axis is unity.
- D2: Randomly distributed points on a spherical surface consisting of three axes. Initially three dimensional 1800 vectors are randomly generated. Here, variance of each axis is unity. Then all the vectors are normalized to unity in length.
- D3: Actual 8-dimensional normalized gene expression profiles of *Bacillus subtilis* from time series experiments in LB medium.
- $D4_i$: Randomly distributed 1000 data points on a i-dimensional space. $(3 \leq i \leq 32)$
- $D5_i$: Randomly distributed 40 clusters on a i-dimensional space. Each cluster has 25 data points.
- $D6_i$: Randomly distributed 40 clusters on a i-dimensional space. Each cluster has 25 data points and is 10^i times denser than $D5_i$.
- $D7_i$: Randomly distributed 40 clusters on a i-dimensional space. Each cluster has 25 data points and is 10^i times denser than $D6_i$.

Those data were normalized to unity in length. D4 has high uniformity but D5, D6 and D7 don't.

4.3 Result and discussion

Figures 24-26 show the d-d plots for those data with Plane SOM (upper) and Spherical SOM (lower). In case of plane surface data (Fig. 4.3), a linear relationship is observed in the d-d plot of Plane SOM. The N-measure of Plane SOM (0.8901: see underlines in Table 3) is greater than that of Spherical SOM (0.3212). On the other hand, in case of spherical surface data (Fig. 4.3), a linear relationship is observed in the d-d plot of Spherical SOM. The N-measure of Plane SOM (0.3782) is less than that of Spherical SOM (0.9764). The N-measure increases when a SOM can represent a set of

high dimensional data without much distortion in the context of distances between all pairs in the original space. Thus, suitability of SOM to a given data set can be estimated by N-measure. In case of actual 8-dimensional gene expression data (Fig. 4.3), the N-measure of Plane SOM (0.1682) is less than that of Spherical SOM (0.6044). It shows Spherical SOM is suitable for the analysis of the data. I suggest that Spherical SOM may be suitable for the analysis of normalized gene expression data.

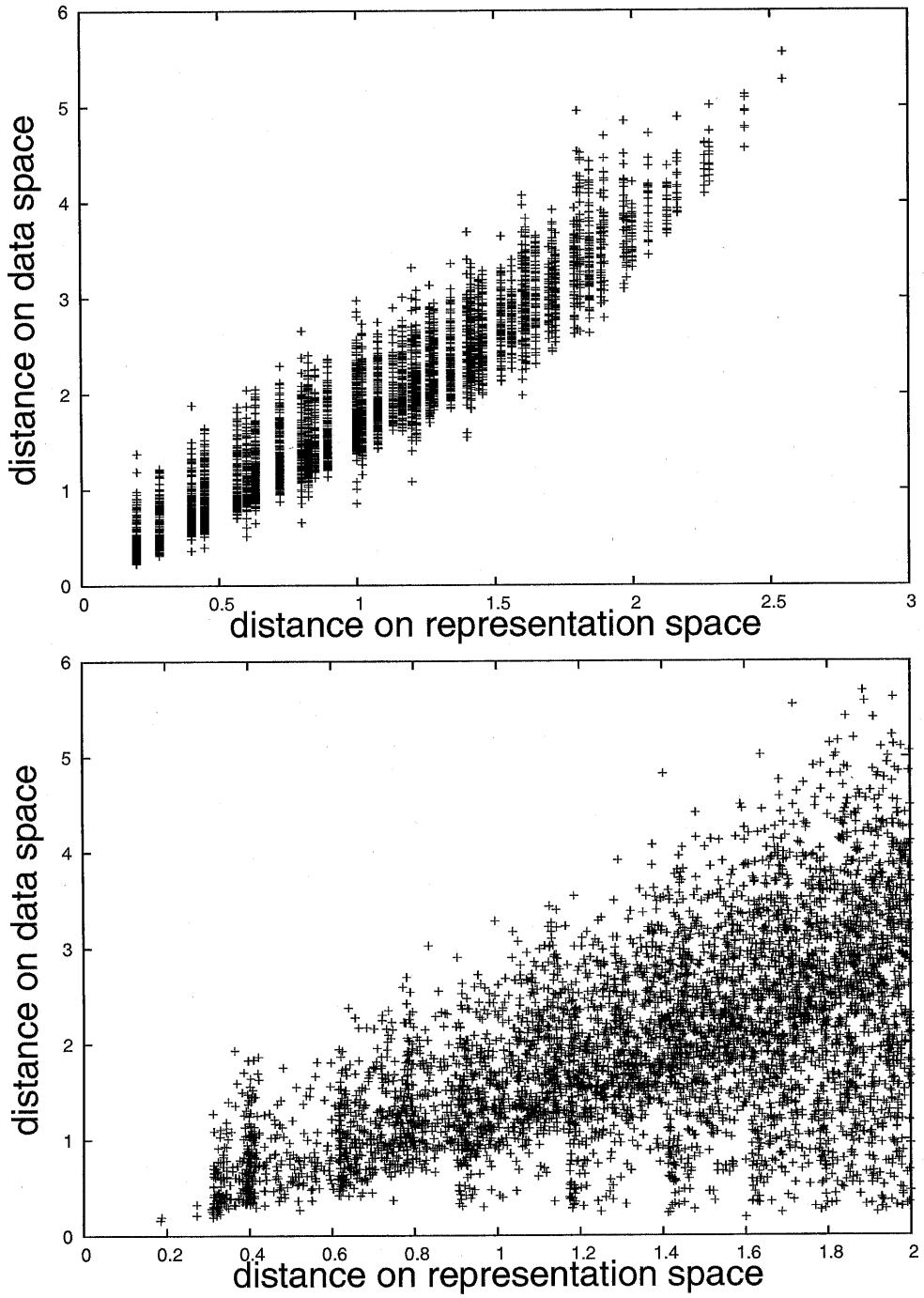


Figure 24: d-d plot from D1(upper:Plane SOM, lower:Spherical SOM)

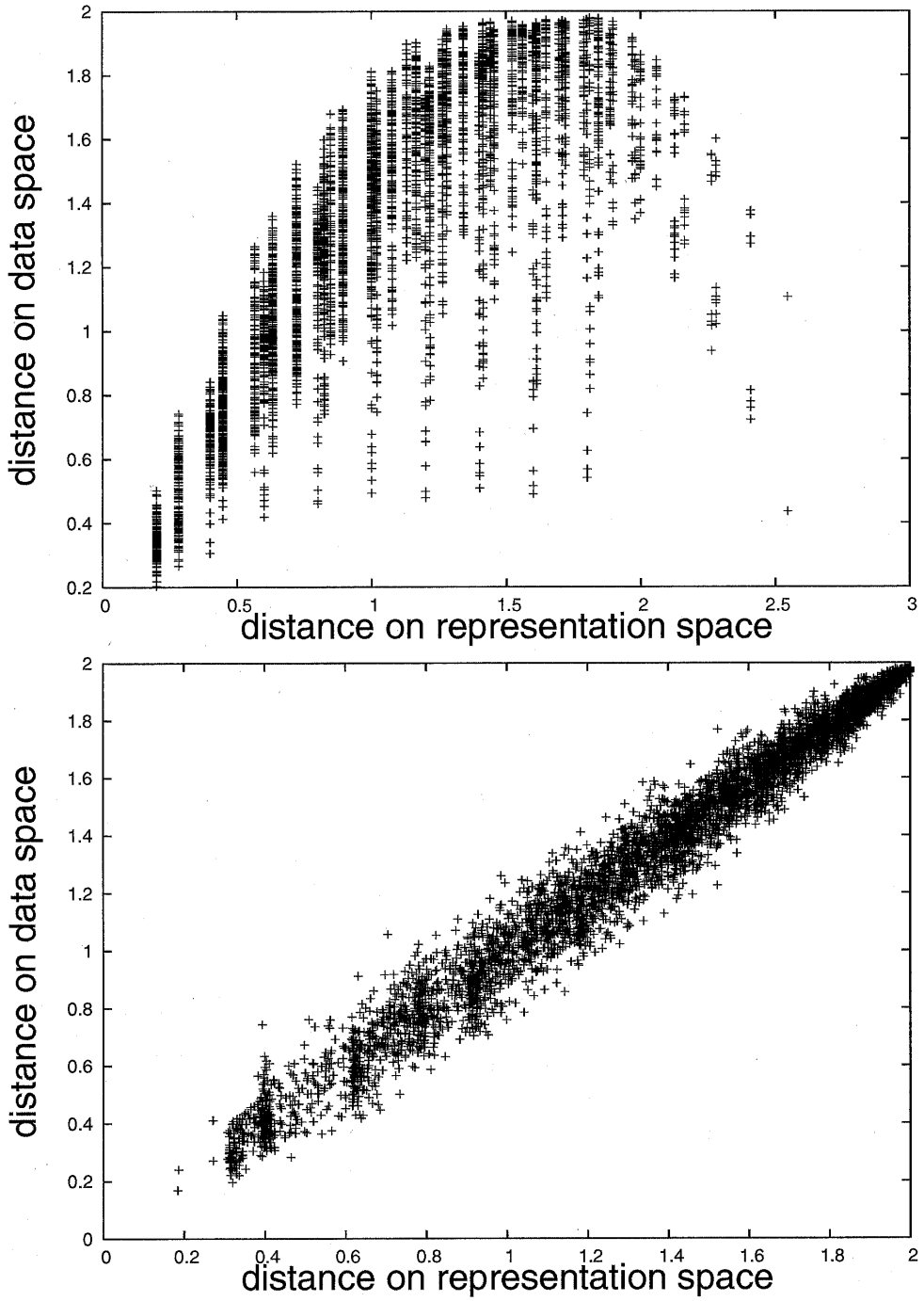


Figure 25: d-d plot from D2(upper:Plane SOM, lower:Spherical SOM)

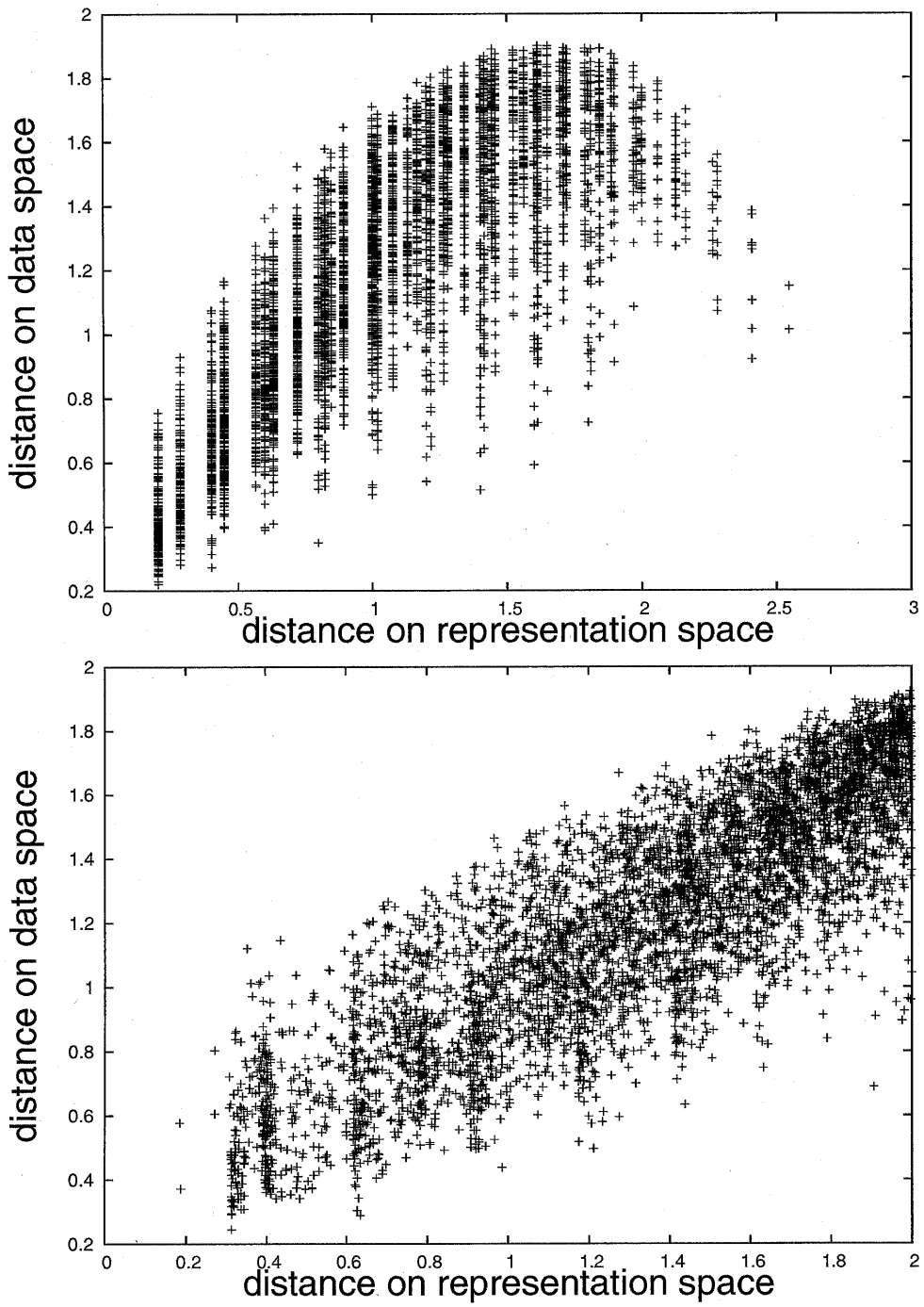


Figure 26: d-d plot from D3(upper:Plane SOM, lower:Spherical SOM)

Table 3: N-measure

Data set	Plane SOM	Spherical SOM
D1: data on plane surface	0.8901	0.3212
D2: data on spherical surface	0.3782	0.9764
D3: data of normalized gene expression	0.1682	0.6044

Figure 27 shows the relation between the dimension i and the average N-measure for the data $D4_i$ of Plane SOM and Spherical SOM from 100 trials, and Figure 28, 29 and 30 show the relation between the dimension i and the average N-measure for the data $D5_i$, $D6_i$ and $D7_i$ of Plane SOM and Spherical SOM from 10 trials. Clusters in D6 or D7 are separated by broad gaps. Borders of representation space of Plane SOM fit to some gaps. It is the reason why the N-measure of Plane SOM rises. On the other hand, the values of N-measure of Spherical SOMs for the data D4 and D5 are apparently greater than those of Plane SOMs. It suggest Spherical SOM is effective to visualize the intrinsic structure of a high dimensional data taking account to the correlation.

4.4 Conclusion

To determine whether Plane or Spherical SOM is suitable for gene expression profile vectors, I proposed a linearity measure of mapping called N-measure as a suitability measure of SOMs. And then I observed that Spherical SOM is more suitable than Plane SOM to visualize the a particular 8 dimensional gene expression data taking account to the correlation. It suggests that Spherical SOM may be suitable for the analysis of high dimensional gene expression data. Generally Spherical SOM is effective to visualize the intrinsic structure of a high dimensional data taking account to the correlation. In the case of the data structure has many broad gaps, Plane SOM is more suitable than Spherical SOM. In that case the data is easily separated to clusters by some gaps, and Spherical SOM is suitable to visualize the intrinsic structure of each cluster.

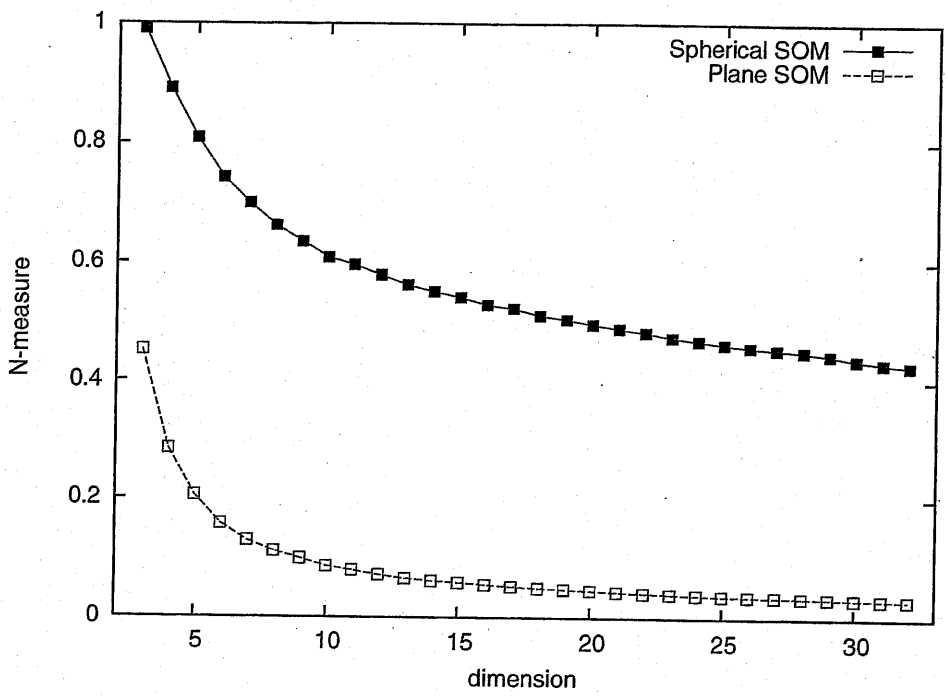


Figure 27: Relation between the dimension of data D4 and its N-measure

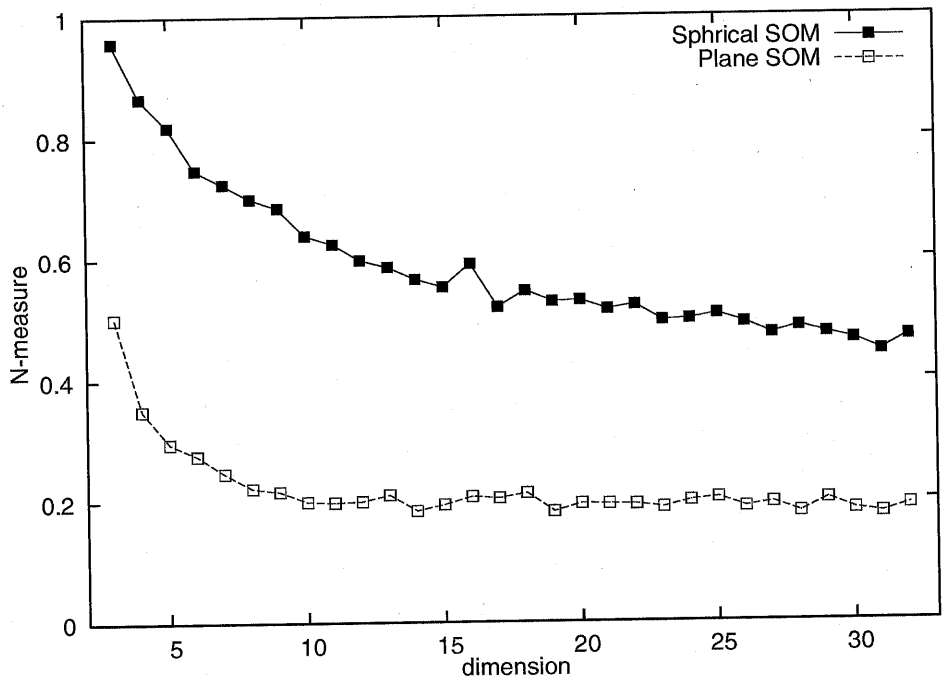


Figure 28: Relation between the dimension of data D5 and its N-measure

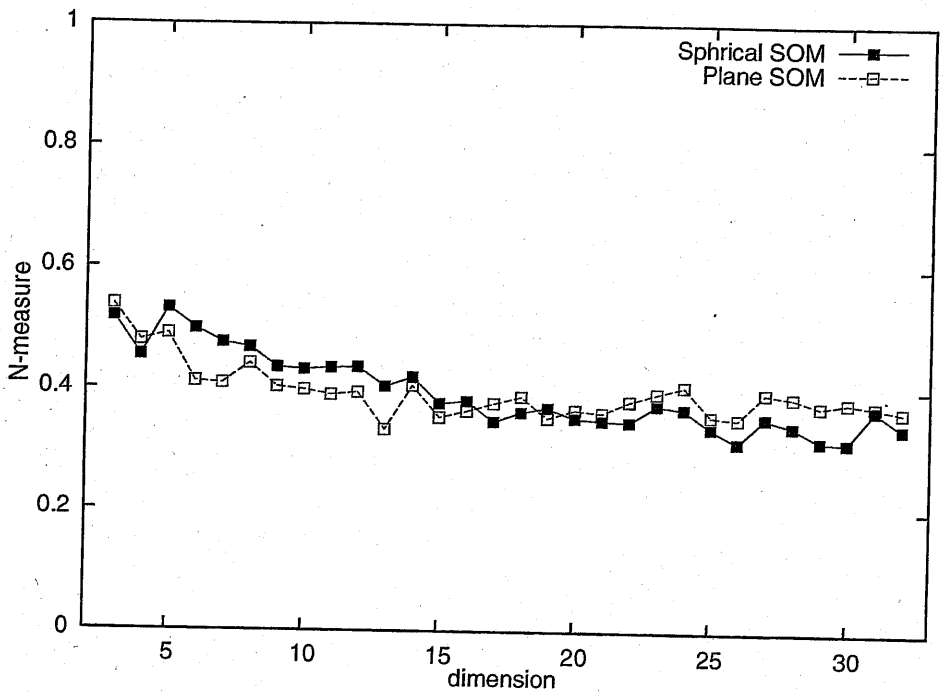


Figure 29: Relation between the dimension of data D6 and its N-measure

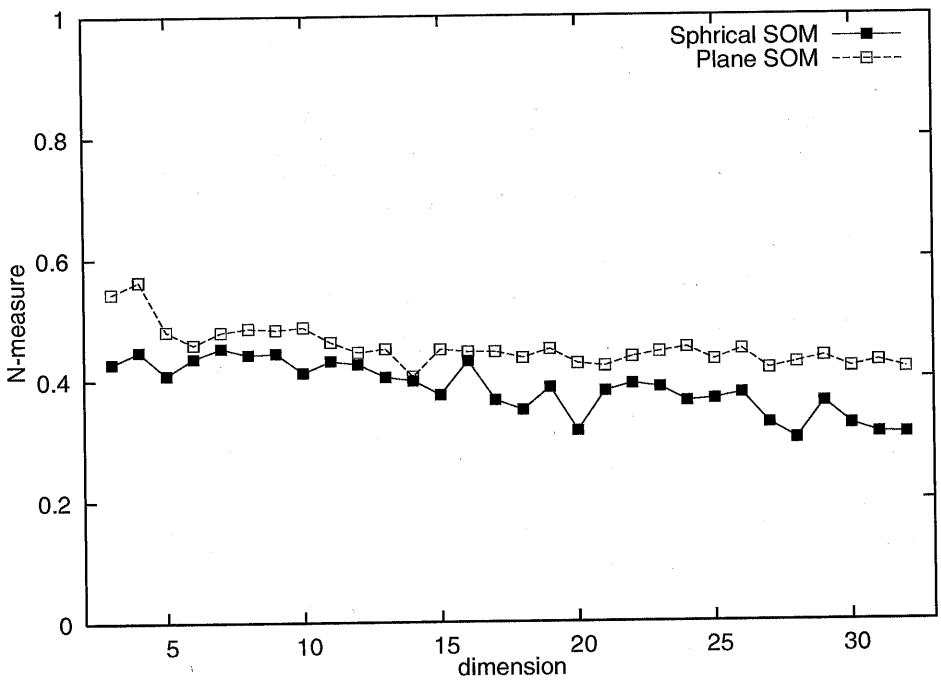


Figure 30: Relation between the dimension of data D7 and its N-measure

5 Conclusion

The conventional Spherical SOM, which uses icosahedron subdivision arrangement, can not arrange arbitrary number of neurons on spherical surface. Therefore I propose a novel SOM which uses spiral arrangement of neurons on spherical surface. I show the novel SOM is better than the icosahedron based SOM in terms of flexibility of number of neurons and uniformity of neuron arrangements. Thus the proposed approach relaxes the usability and improves the usefulness of the Spherical SOM. I developed a software using proposed method to arrange arbitrary number of neurons. I applied the software to actual biological data from microarray experiments and confirmed utility of spherical SOM. I proposed a linearity measure of mapping called N-measure as a suitability measure of SOMs, observed that Spherical SOM is more suitable than Plane SOM to visualize the high dimensional data following a normal distribution taking account to the correlation. When data are extremely concentrated in a limited part, superiority of spherical SOM is lost.

6 Acknowledgements

I am thankful to Professor Shigehiko Kanaya, Kotaro Minato, Kunihiro Chihara and Associate Professor Ken Kurokawa for their worthful guidances and this thesis' review. I am thankful to Assistant Professor Md. Altaf-Ul-Amin for precious guidances. I am thankful to Professor Naotake Ogasawara and members of Ogasawara Lab. for the expression profiles of *Bacillus subtilis*.

References

[1] D. J. Willshaw, C. Von Der Malsburg. How patterned neural connections can be set up by self-organization. Proc. Roy. Soc. London B, vol. 194, pp. 431-445, 1976.

[2] T. Kohonen. Self-organized formation of topologically correct feature maps. Biol. Cybern., vol. 43, pp. 59-69, 1982.

[3] Teuvo Kohonen. Self-Organizing Maps. Springer Series in Information Sciences. Springer, second edition, 1997.

[4] Juha Vesanto, Mika Sulkava, Jaakko Hollmén. On the Decomposition of the Self-Organizing Map Distortion Measure.

[5] Helge Ritter. Self-Organizing Maps on non-euclidean Spaces. Kohonen Maps, pp. 97-108, ed. E. Oja and S. Kaski, 1999

[6] E. A. Rakhmanov, E. B. Saff, Y. M. Zhou (1994), Minimal Discrete Energy on the Sphere, *Mathematical Research Letters* **vol. 1** p. 647-662.

[7] H. Nishio, Md. Altaf-Ul-Amin, K. Kurokawa, S. Kanaya (2004), Spherical SOM and arrangement of neurons using helix on sphere, *ISPJ Symposium Series (in press)* p. 113-120.

[8] H. Nishio, K. Wada, Y. Wada, Md. Altaf-Ul-Amin, S. Kanaya (2004), Suitability of Spherical SOM for Gene Expression Analysis, *Proc. of RECOMB 2004* p. 79-80.

[9] Duch W and Naud A, Simplexes, Multi-Dimensional Scaling and Self-Organized Mapping. Proceedings of the 8th joint EPS-APS International Conference on Physics Computing '96, pp. 367-370, 17?21.9.1996, Krakow, Poland.

[10] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, T. Ikemura (2003), Informatics for unvailing hidden genome signatures, *Genome Res.* p. 693-702.

[11] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, T. Ikemura (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* p. 89-99.

[12] H. Nishio, Md. Altaf-Ul-Amin, Y. Nakamura, K. Kurokawa, Y. Shinbo, T. Abe, M. Kinouchi, T. Ikemura, K. Kobayashi, N. Ogasawara, S. Kanaya (2004), Gene classification Based on Expression Profile using BL-SOM: Suitability assessment of multivariate gene expression data to Spherical and Plain SOM by N-measure, *Proc. of SCI 2004* **vol. 8** p. 189-192.

[13] Janne Nikkilä, Petri Törönen, Samuel Kaski, Jarkko Venna, Eero Castrén, Garry Wong. Analysis and visualization of gene expression data using Self-Organizing Maps. *Neural Networks.*

[14] Daniel Pullwitt. Integrating contextual information to enhance SOM-based text document clustering. *Neural Networks.*

[15] Mu-Chun Su, Hsiao-Te Chang. Fast Self-Organizing Feature Map Algorithm. *IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 11, NO. 3, MAY 2000 721.*

[16] K. Marzouki and T. Yamakawa . Novel Learning Algorithm Aiming at Generating a Unique Units Distribution in Standard SOM. Being submitted to ICANNGA 2005, 21st -23rd March, Coimbra Portugal. Paper Accepted on November 15th.

[17] D. Deng, and N. Kasabov, ESOM: An algorithm to evolve self-organizing maps from on-line data streams, Proc. of IJCNN 2000, (IEEE Press, 2000), VI:3-8.

[18] Hirokazu Nishio, Md. Altaf-Ul-Amin, Tetsuo Sato, Ken-nosuke Wada, Yoshiko Wada, Kotaro Minato, Kazuo Kobayashi, Naotake Ogasawara, Shigehiko Kanaya. Visualization of Gene Classification Based on Expression Profile Using BL-SOM.

[19] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest. Introduction to Algorithms. MIT Press.

[20] Significant of Digital Gene Expression Profiles (SAGE) Audic,S., and Claverie, J. M. (1997) The significance of digital gene expression profiles, Genome Res., 7, 986-995.

[21] von Kampen, A., van Schaik, B., pauws, E., Michiels, E., Ruijter, J., Caron, H., Versteeg, R., Heisterkamp, S., Leunissen, J., Baas, F., and van der Mee, M. (2000) USAGE: A web-based approach towards the analysis of SAGE data. Bioinformatics, 16, 899-905.

[22] Lash, A., Tolstoshev, C., Wagner, L., Schuler, G.., Strausberg, R., Riggins G., and Altschul, S. (2000) SAGE map: A public gene expression resource. Genome Res., 10, 1051-1060.

[23] Man, M. Z., Wang, X., and Wang, Y. (2000) POWER_SAGE: Comparing statistical tests for SAGE experiments. Bioinformatics, 16, 953-959.

[24] Velculescu, V., Zhang, L., Vogelstein, B., and Kinzler, K. (1995) Serial analysis of gene expression. Sci., 270, 484-487.

[25] Yamamoto, M., Wakatsuki, T., Hada, A., and Ryo, A. (2001) Use of serial analysis of gene expression (SAGE) technology. J. Immunol. Methods, 250, 45-66.

[26] Differential display Matz, M., Usman, N., Shagin, D., Bogdanova, E., and Lukyanov, S. (1997) Ordered defferential display: A simple method for systematic comparison of gene expression profiles. Nucleic Acids Res., 25, 2541-2542.

[27] GeneChip Baugh, L. R., Hill, A. A., Brown, E. L., and Hunter, C. P. (2001) Quantitative analysis of mRNA amplification by in vitro transcription. Nucleic Acids Res., 29, E29.

[28] Lipshutz, R., Fodor, S., Gingeras, T., and Lockhart, D. (1999) High density synthetic oligonucleotide arrays. Nature Genetics Chipping Forecast, 21, 20-24.

[29] Lockhart, D. J., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang C., Kobayashi, M., Horton, H., and Brown, E. (1996) Expression monitoring by hybridization to high-density oligonulcleotide arrays. Nature Biotech. 14, 1675-1680.

[30] Wodicka, L., Dong, H., Mittmann, M., Ho, M., and Lockhart, D. (1997) Genome-wide expression monitoring in Saccharomyces cerevisiae. Nature Biotech., 15, 1359-1367.

[31] Bao, P. Frutos, A., Greef, C., Lahiri, J., Muller, U., Peterson, T., Warden, L., and Xie, X., (2002) High-sensitivity detection of DNA hybridization on microarrays using resonance light scattering., Anal. Chem., 74, 1792-1797.

[32] Bowtell, D. (1999) Options available ? from start to finish ? for obtaining expression data by microarray. Nature Genet. (supplement), 21, 25-32.

[33] Brown P., and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays, Nature Genetics, 21 (Suppl. 1), 33-37.

[34] Schena, M., Shalon, D., Davis, R., Brown, P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Sci., 270, 467-470.

[35] Pennie W. (2002) Custom cDNA microarrays: technologies and applications. Toxicol., 181/182, 551-554.

[36] Xiang, C., and Chen, Y. (2000) cDNA microarray technology and its application., Biotechnol. Adv., 18, 35-46.

[37] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M. Downing, J., Caligiuri, M., Bloomfield, C., Lander, E., (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Sci., 286, 531-537.

[38] Nishio, H., Amin, M., Sato, T., Wada, K., Wada, Y., Minato, K., Kobayashi, K., Ogasawara, N., and Kanaya, S. (2003) Visualization of gene classification based on expression profile using BL-SOM., Proc. of WSOM03, 101-106.

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

[40] Toronen, P., Kolehmainer, M., Wong, G., and Castr?n E., (1999) Analysis of gene expression data using self-organizing maps., FEBS Letters, 451, 142-146.

[41] Alter, O., Brown, P.O., Botstein, D., (2000) Singular value decomposition for genome-wide expression data processing and modeling., Proc. Natl. Acad. Sci. USA., 97, 10101-10106.

[42] Horn, D., and Axel, I., (2003) Novel clustering algorithm for microarray expression data in a truncated SVD space., Bioinformatics, 19, 1110-1115.

[43] Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Bostein, D., Brown, P., (2000) ' Gene shaving ' as a method for identifying distinct sets of genes with similar expression patterns., Genome Biol., 1, 0003.1-0003.21.

[44] Lee, S., Batzoglou, S., (2003) Application of independent component analysis to microarrays. Genome Biol., 2003,4, R76.

[45] Martoglio, A., Miskin, J. Smith, S. MacKay, D., (2002) A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer., 18, 1617-1624.

[46] Chang, H., Chi, J., Dudoit, S., Bondre, C., van de Rijn, M., Botstein, D., and Brown P., (2002) Diversity, topographic differentiation, and positional memory in human fibroblasts., Proc. Natl. Acad. Sci. USA., 99, 12877-12882.

[47] Eisen, M., Spellman, P., Brown, P., and Botstein, D., (1998) Cluster analysis and display of genome-wide expression patterns., Proc. Natl. Acad. Sci. USA., 95, 14863-14868. Heyer, L. J., Kruglyak, S., and Yooseph, S., (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res., 9, 1106-1115.

[48] Pan, W., Lin, J., Le, and C.T. (2002) Model-based clustering analysis of microarray gene-expression data., Genome Biol., 3, 0009.1-0009.8.

[49] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L., (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics, 17, 977-987.

[50] Getz, G., Levine, E., and Domany, E., (2000) Coupled two-way clustering analysis of gene microarray data., 97, 12079-12084.

[51] Lazzeroni, L., and Owen, A., (2002) Plaid models for gene expression data, Statistica Sinica, 12, 61-85.

[52] Sultan, M., Wigle, D.A., Cumbaa, C.A., Maziarz, M., Glasgow, J., Tsao, M.S., and Jurisica, I. (2002) Binary tree-structured vector quantization approach to clustering and visualizing microarray data., Bioinformatics, 18, S111-S119.

[53] Qin, J., Lewis, D., Noble W.S., (2003) Kernel hierarchical gene clustering from microarray expression data., Bioinformatics, 19, 2097-2104.

[54] Chen, G. Jaradat, S., Banerjee, N., (2002) Evaluation and comparison of clustering algorithms in anglyzing ES cell gene expression data., Statistica Sinica, 12, 241-262.

[55] McShane, L., Radmacher, M., Freidlin, B., Yu, R., Li, M., and Simon, R., (2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data., Bioinformatics, 18, 1462-1469.

[56] Ishida, N., Hayashi, K., Hoshijima, M., Ogawa, T., Koga, S., Miyatake, Y., Kumegawa, M., Kimura, T., Takeya, T., (2002) Large scale gene expression analysis of osteoclastogenesis in vitro and elucidation of NFAT2 as a key regulator., J. Biol. Chem., 277, 41147-41156.