Spherical SOM with arbitrary number of neurons and measure of suitability

Hirokazu Nishio, Md. Altaf-Ul-Amin, Ken Kurokawa, Kotaro Minato, Shigehiko Kanaya
Nara Institute of Science and Technology
Nara-ken Japan
hiroka-n@is.naist.jp

Abstract - *Spherical SOM, where neurons are arranged on spherical surface, is suitable for clustering objects based on the trends such as the direction of vectors in multidimensional space, for example, for classification of genes based on expression profiles. It is difficult to arrange arbitrary number of neurons uniformly on spherical surface. To overcome this difficulty, we propose spiral arrangement of neurons on spherical surface, which allows arrangement of arbitrary number of neurons and call it as SphSOM-SPIRAL. Statistically, the spiral arrangement has rather higher uniformity compare to icosahedron subdivision arrangement which is generally used in Spherical SOM. We 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 that the Spherical SOM is more suitable than the Plane SOM.*

Key words - Spherical SOM, arbitrary number of neurons, microarray, uniformity, suitability

1 Introduction

Self-organizing Maps (SOM) are clustering methods that can map high-dimensional data to low-dimensional representation space [1]. In the present study we consider two types of SOMs, one of them uses planar representation space called ‘ Plane SOM ’ and the other uses spherical space called ‘ Spherical SOM. ’ Ritter first proposed Spherical SOM [2], whose neurons are arranged on spherical surface by subdividing an icosahedron recursively, and it is referred to as SphSOM-ICOSA. When we compare genes based on expression profiles, normalization of profile vectors of gene expression to unity in length has advantage that genes are compared by the trends of expression profiles and are classified based on similarity of expression regulation in cells [3]. The normalized vectors are distributed on the surface of an S-dimensional hypersphere, where S is the number of measurements for genes. Spherical SOM is adequate for clustering the normalized data. In the present paper, we propose a method for arranging arbitrary number of neurons uniformly on Spherical SOM called SphSOM-SPIRAL, compare the uniformity of neuron arrangement on spherical surface between SphSOM-SPIRAL and SphSOM-ICOSA, and show that SphSOM-SPIRAL makes it possible to arrange neurons more uniformly on spherical surface in comparison to SphSOM-ICOSA. We also propose a measure to select suitable neuron arrangements, that is, whether spherical or Plane SOM is suitable for classification of high dimensional data, and demonstrate that Spherical SOM is more suitable for normalized gene expression profile vectors than Plane SOM.

2 Method

2.1 Plane SOM and Spherical SOM

We use the term “Plane SOM” for a type of SOM whose neurons are arranged on plane surface [1], and “Spherical SOM” for a type of SOM whose neurons are arranged on spherical surface [2].

2.1.1 Arrangement of neurons

In Spherical SOM proposed by Ritter, neurons are arranged by subdividing an icosahedron recursively [2]. Example of a triangle of an icosahedron and its three recursive divisions are shown in Figure 1(a). We 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, we can’t always arrange arbitrary number of neurons. To solve this problem, we proposed a method to arrange arbitrary number of neurons by dividing a helix which goes around a sphere of unity radius (Fig. 1(c)) into pieces of identical length [4]. While the arrangement has high uniformity, this method need numerical integration to calculate the length of helix. In the present paper we use the concept of generalized spiral points. The generalized spiral points are explicitly defined sets of points [5]; where the set of points $\{(\theta_k, \phi_k) | 0 \leq k < N\}$ is determined by the Eq. 1.

$$
\begin{array}{rcll}
h_k & := & \dfrac{2k}{N-1} - 1, & (0 \leq k < N) \\
\phi_k & := & \arccos(h_k), & (0 \leq k < N) \\
\theta_0 & := & \theta_{N-1} := 0, & \\
\theta_k & := & \left(\theta_{k-1} + \dfrac{C}{\sqrt{N(1-h_k^2)}}\right). & (1 \leq k < N-1)
\end{array}
\tag{1}
$$

Here, C is a constant and was chosen as $C = 3.6$ so that the distance between successive point will be approximately the same. Spherical coordinates θ and ϕ can be transformed to orthogonal coordinates by Eq. 2. (See Figure 1(b).)

$$x = \cos\theta \sin\phi, y = \sin\theta \sin\phi, z = \cos\phi \tag{2}$$

Using the generalized spiral points, we can arrange neurons of arbitrary number on the spherical surface. We use the term “SphSOM-SPIRAL” to refer to a type of SOM whose neurons are arranged by this method. The set of neighborhood neurons of a neuron η corresponding to a particular radius r is the set of neurons that are within the Euclidean distance r from η. (See Figure 1(e).)

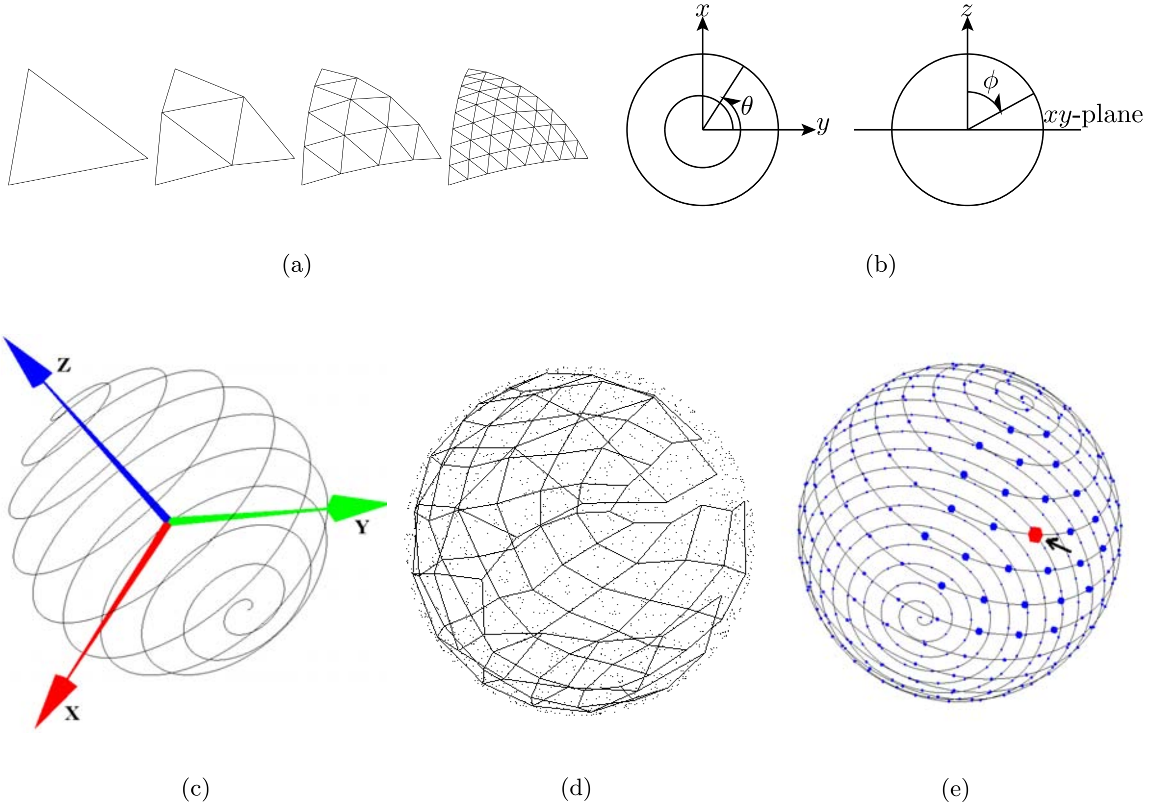


Figure 1: (a) Recursive subdivision of a triangle of an icosahedron (b) Relation between axis x, y, z, θ, ϕ (c) A helix on a sphere (d) Distortion when Plane SOM learns the data on the spherical surface (e)28 neurons are within radius 0.5 from the indicated neuron η i.e. f(η, 0.5)=28.

2.2 *UNTIDINESS*, the measure of uniformity

"Border effects" is one of the important disadvantages for Plane SOM. The problems occur because the number of neighborhood neurons of a neuron near a border is different from that of a neuron near the center. In the case of Spherical SOM, even though it has no border, there is a problem how to arrange arbitrary number of neurons uniformly. In the present study, 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 where the number of neighborhood neurons is the same for each of the neurons of the map. The concept of uniformity is represented by Eq. 3, where $f(\eta, r)$ is the numbers of neurons within radius r from neuron η (See Figure 1(e) as an example) and $V(r)$ is the variance of $f(\eta, r)$ for all neurons.

$$V(r) = \frac{n\sum_\eta f(\eta, r)^2 - \left(\sum_\eta f(\eta, r)\right)^2}{n^2} \tag{3}$$

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. An example for *ICOSA*$_3$

and *SPIRAL*$_3$ are shown in Figure 2(a). In order to allow any value for radius r, we propose the measure of uniformity *UNTIDINESS* by Eq. 4.

$$\mathit{UNTIDINESS} := \int_{\theta=0}^{\theta=\pi/2} V\left(2\sin\left(\theta/2\right)\right) \mathrm{d}\theta \tag{4}$$

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

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

$$\mathbf{w_i} := R(x_i\mathbf{e_1} + y_i\mathbf{e_2} + z_i\mathbf{e_3}). \tag{5}$$

Here R is the radius of hypersphere.

2.4 N-measure, the measure of suitability

In this paper we compare two different type of SOM: Plane SOM and Spherical SOM. When there are 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 we can't observe high dimensional distribution directly. So we propose a measure of suitability. When a map 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. This linearity can be a basis to determine whether spherical or planar arrangement of neurons is suitable for a specific data set. This is carried out by plotting two distances between all pairs of the neurons defined in representative space such as spherical or plane surface and in the original space, respectively, which are called d-d plot. In Fig. 1(d), for example, neurons on opposing corner of 2-dimensional plane arrangement are close in 3-dimensional space. So the linear relation is broken in d-d plot because of the distortion occurred from projecting high dimensional data of original space to representation space. 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. 6.)

$$\text{minimize} \sum_i (y_i - \hat{y_i})^2 \tag{6}$$

The coefficient of determination R^2 in regression analysis is defined by Eq. 7.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y})^2} \tag{7}$$

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

$$\bar{y} = \frac{1}{N}\sum_i y_i \tag{8}$$

R^2 means how the given data fit to the given model. It should be quantified how d-d plot satisfy the linear model $\hat{y_i} = kx_i$. Here, k for minimal square difference is represented in Eq. 9.

$$k = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \tag{9}$$

This is applied to d-d plot as follows. For mapping Φ and set on its domain Ω, $X := \{\|\mathbf{u}-\mathbf{v}\| | \mathbf{u} \in \Omega, \mathbf{v} \in \Omega\}$ and $Y := \{\|\Phi(\mathbf{u}) - \Phi(\mathbf{v})\| | \mathbf{u} \in \Omega, \mathbf{v} \in \Omega\}$. The N-measure $\mathcal{N}$ can be defined by combining these equations,

$$\mathcal{N} := \frac{NS_{xy}^2 - S_{xx}S_y^2}{NS_{xx}S_{yy} - S_{xx}S_y^2} \tag{10}$$

Here $S_y := \sum_i y_i$, $S_{xx} := \sum_i x_i^2$, $S_{xy} := \sum_i x_i y_i$ and $S_{yy} := \sum_i y_i^2$.
$\mathcal{N}$ takes value between -1 and 1. The larger the value of $\mathcal{N}$, the higher the linearity.

3 Result and Discussion

3.1 Uniformity

We use the term *SPIRAL*$_N$ to refer to spiral arrangement of a particular number of neurons that is equal to the number of neurons in *ICOSA*$_N$. The relations between the radius r and the variance $V(r)$ for *ICOSA*$_3$ and *SPIRAL*$_3$ are shown in Figure 2(a) and for *ICOSA*$_4$ and *SPIRAL*$_4$ are shown in Figure 2(b) as an example. It shows $V(r)$ in the case of helix is mostly lower than that in the case of an icosahedron.

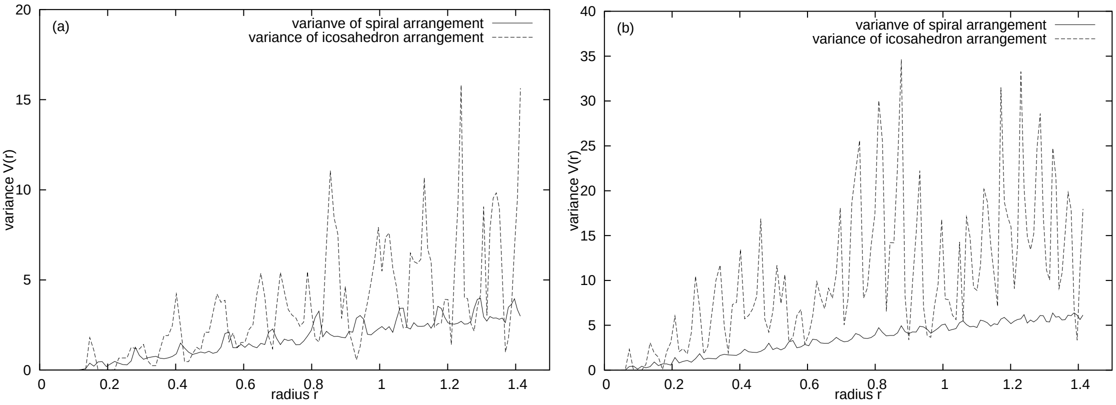


Figure 2: relation between the radius and the variance (a) *SPIRAL*$_3$ and *ICOSA*$_3$ (b) *SPIRAL*$_4$ and *ICOSA*$_4$

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

N	# of neurons	*UNTIDINESS* of $ICOSA_N$	*UNTIDINESS* of $SPIRAL_N$
0	12	0.0000	0.2611
1	42	0.4253	0.4270
2	162	2.8707	0.9057
3	642	3.9360	1.8011
4	2562	10.4419	3.4898
5	10242	82.1324	6.8744

Table 1 lists the *UNTIDINESS* of $ICOSA_N$ and $SPIRAL_N$ for N in range $0 \leq N \leq 5$. $ICOSA_0$ is completely uniform because it is an icosahedron itself. $ICOSA_1$ has smaller *UNTIDINESS* than $SPIRAL_1$. Since *UNTIDINESS* of $ICOSA_N$ increases quickly, it is larger than that of $SPIRAL_N$ when N is two or more. Figure 3 shows the relation between the number of neurons and *UNTIDINESS*. By conventional method we 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, we can arrange any number of neurons by proposed 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 conventional 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. To determine whether plane or Spherical SOM is suitable for gene expression profile vectors, we consider SphSOM-SPIRAL instead of SphSOM-ICOSA.

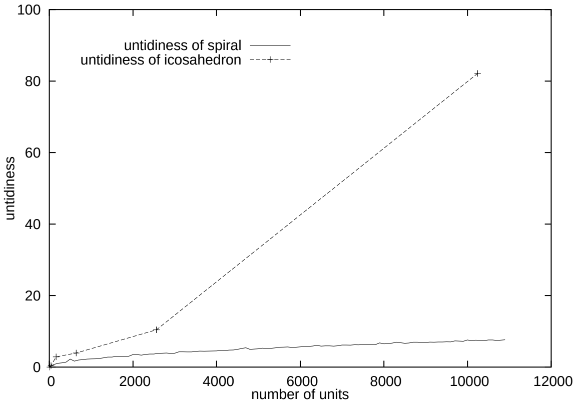


Figure 3: Relation between *UNTIDINESS* and the number of neurons

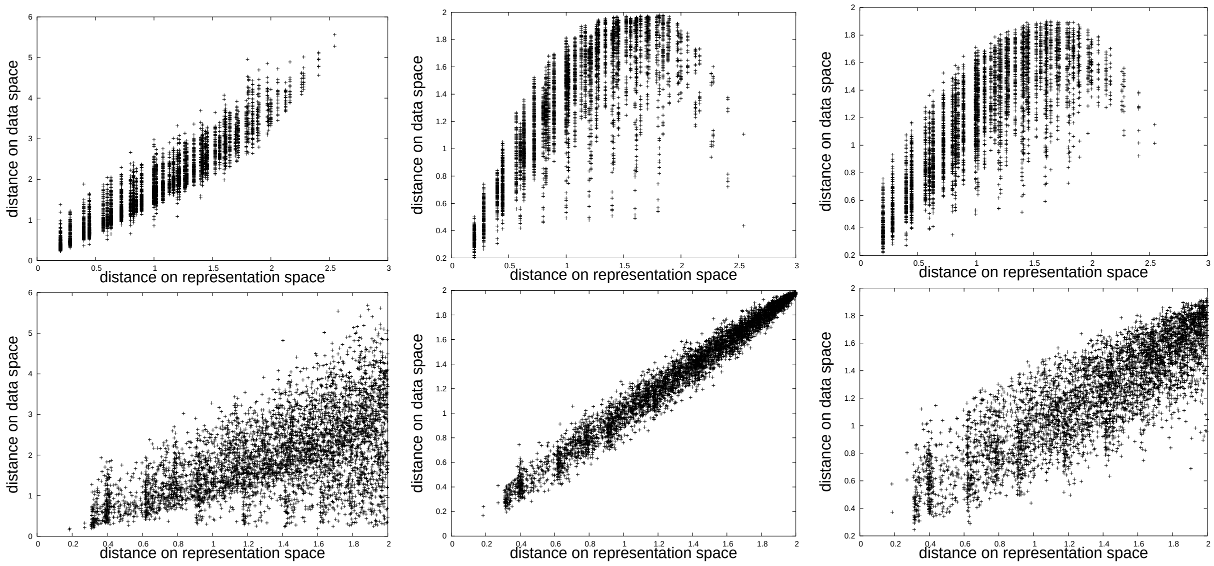
3.2 Suitability

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

- Randomly distributed 1800 points on a plane surface consisting of two axes. Variance of each axis is unity.

- 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.
- Actual 8-dimensional normalized gene expression profiles of *Bacillus subtilis* from time series experiments in LB medium.

Figure 4 shows the d-d plots for those data with Plane SOM (upper) and Spherical SOM (lower). In case of plane surface data (Fig. 4(a)), 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 2) is greater than that of Spherical SOM (0.3212). On the other hand, in case of spherical surface data (Fig. 4(b)), 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(c)), 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. We suggest that Spherical SOM may be suitable for the analysis of normalized gene expression data.



(a) data on plane surface

(b) data on spherical surface

(c) normalized gene expression data

Figure 4: d-d plot(upper:Plane SOM, lower:Spherical SOM). Horizontal axis means distance on representation space and vertical axis means distance on data space

Table 2: N-measure

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

4 Conclusion

Conventional Spherical SOM that uses icosahedron subdivision arrangement, can not arrange arbitrary number of neurons on spherical surface. Therefore we propose a SOM with spiral arrangement of neurons on spherical surface (SphSOM-SPIRAL). We show that SphSOM-SPIRAL is better than 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. To determine whether plane or Spherical SOM is suitable for gene expression profile vectors, we proposed a measure called N-measure. In the present paper, we observed that Spherical SOM is more suitable than Plane SOM for classification of genes based on normalized gene expression profile vectors for a particular set of microarray time-series data of *B. subtilis.*

References

[1] Teuvo Kohonen (1990), The self-organizing map, *Proc. of IEEE* **vol. 78 No. 9** p. 1464-1480.

[2] Helge Ritter (1999), Self-Organizing Maps on non-euclidean Spaces, *Kohonen Maps* p. 97-108.

[3] H. Nishio, Md. Altaf-Ul-Amin, T. Sato, K. Wada, Y. Wada, K. Minato, K. Kobayashi, N. Ogasawara, S. Kanaya (2003), Visualization of Gene Classification Based on Expression Profile Using BL-SOM, *Proc. of WSOM'03* p. 101-106.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.