We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

4,100
Open access books available

116,000
International authors and editors

120M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
1. Introduction

The technology of high throughput screening is nowadays widely available in life science especially in the fields of molecular diagnosis and drug discovery. This is due to the establishment of microarray procedure, which deals with huge number of genes at one time. Cluster analysis is usually performed on the results of DNA microarray experiments. However, the routine procedure of data mining dealing with the huge number of signal information obtained from microarray is not fixed yet. We can find various way of clustering in hierarchical and non-hierarchical methods, which are applied for the analyses. The most popular ones appear non-hierarchical clustering such as k-means (MacQueen, 1965), partitioning around medoids (Kaufman and Rousseeuw, 1990) and cluster affinity search technique (Ben-Dor et al., 1999).

We have employed spherical self organizing map (sSOM), which is also a non-hierarchical clustering, to cluster genes by gene expression profiles of cells and tissues (Tuoya et al., 2008). Analyzing various types of carcinoma cells and normal tissues, we could find interesting cell surface molecules, which should serve as the molecular markers. This procedure, which we are demonstrating, is rather new to the data analyses of gene clustering from the gene expression profiles obtained from DNA microarray technique. Flexible arrangement of the data obtained allows us to cluster cells and tissues as well as genes to find definitely fantastic direction of further advancement of study.

Furthermore, we applied sSOM to classify bioactive chemical compounds by their mechanism of action (MOA), which should enable us virtual screening in silico (Reddy et al., 2007). In recent years, many intriguing methods for virtual screening have been developed in this field (Gasteiger et al., 2003; Melville et al., 2009). Especially, ligand-based method is suitable for selecting drug candidates from enormous compounds library because it simply requires computational resources, which are less expensive. Although SOM has partly been used as a tool of ligand-based methods to classify compounds by their properties in chemoinformatics, spherical SOM has not been used in chemoinformatics to the best of our knowledge (Brüstle et al., 2002; Schneider & Nettekoven, 2003; Schneider & Schneider, 2003; Wang et al., 2005; Kaiser et al., 2007; Renner et al., 2007; Li & Gramatica, 2010). We propose here the extended application of sSOM to classify bioactive compounds by their MOA together with their structural information. In the future this procedure should
be extremely helpful in the field of drug discovery as well as those of molecular biology and oncology.
This chapter is dedicated to introduce our concept of the application of sSOM procedure.

2. Materials, methods and tools

2.1 Cell lines and cell culture
Human breast cancer derived cell lines Hs-578T, MCF-7, MDA-MB-134, MDA-MB-231, SK-BR-3, T-47D and ZR-75-1 were obtained from American Type Culture Collection (ATCC, VA). Hs-578T cells were cultured in DMEM containing 10 % fetal bovine serum (FBS), 10 µg/mL insulin and 2 mM L-glutamine. MCF-7 cells were cultured in MEM containing 10 % FBS, 10 µg/mL insulin and 2 mM L-glutamine. MDA-MB-134 cells were cultured in Leibovitz-15 containing 10 % FBS, 2 mM L-glutamine buffered with 10 mM HEPES. MDA-MB-231 cells were cultured in DMEM containing 10 % FBS and 2 mM L-glutamine. SK-BR-3 cells were cultured in RPMI 1640 containing 20 % FBS and 2 mM L-glutamine. T-47D cells cultured in RPMI 1640 containing 10 % FBS, 2 mM L-glutamine, 10 µg/mL insulin and 30 ng/mL EGF. ZR-75-1 cells were cultured in RPMI 1640 containing 10 % FBS. All cells were maintained at 37°C in a humidified 5 % CO2 atmosphere except MDA-MB-134 cells, which were maintained in 100 % air.

2.2 Preparation of total RNA and cDNA synthesis
Total RNA was extracted from the cells used in this study. Cells were harvested at a confluence of 80% for preparation using RNeasy Mini kits (Qiagen), following the manufacturer’s instructions. Total RNA from human normal breast and mouse normal tissues was purchased from Stratagene (CA). RNA integrity and purity were assessed by OD260/280 measurements and by the ratio of 28S and 18S rRNA with Experion system (BioRad Labs, VA). The total RNA was further treated with DNase and purified. The integrity of template RNA was assessed by OD260/280 measurements. Twenty micrograms of total RNA was used to synthesize cDNA in the presence of aminoalkyl-dUTP. To monitor the efficiency of cDNA synthesis and hybridization control RNAs were added in the reaction as describe previously (Tuoya et al., 2008; Abou-Sharieha et al. 2009). Cy3-labeled cDNAs were prepared by indirect labeling method adapted from the Brown Web site (http://cmgm.stanford.edu/pbrown/protocols).

2.3 Microarray analysis
We originally proposed DNA microarray, which focused cell membrane-bound proteins to identify cell surface marker specific to the cells or tissues of interest (Tuoya et al., 2008; Abou-Sharieha et al. 2009). Two different microarrays were designed to contain 1,795 oligonucleotide probes corresponding to human genes and 1,405 corresponding to mouse genes, respectively. These genes were limited to those coding membrane-bound proteins so as to cover cell surface proteins. To avoid the effect of alternative splicing, the coding sequence for the membrane-bound region or GPI-anchor modified region was focused to design the oligonucleotide probes. The probes were conjugated on the slide glass coated with diamond-like carbon as described previously (Tuoya et al., 2005).

The Cy3-labeled cDNA synthesized above was hybridized to the cell surface marker DNA microarray in 5x SSC/0.5 % SDS solution at 55 °C for 15 h. After washing, arrays were scanned on a FLA8000 scanner (Fuji Film, Japan). Intensity for each spot of the array was
captured by GenePix® Pro5.1 image analysis software (Axon Instrument). The fluorescent intensity of each spot referred as "relative fluorescent intensity (RFI)", which represented the expression level of each gene. Gene expression levels were compared to one another by RFI value to identify differentially expressed genes.

2.4 Data filtering in breast cancer cell
In order to eliminate genes that did not change significantly between cancer cell lines and normal tissue, each gene was given a score \( S \) by a formula:

\[
S = N - C - V_c
\]

where \( N \), \( C \) and \( V_c \) denote the expression level of the gene in normal breast, the average of the expression levels of the gene in the seven cancer cell lines and the standard deviation of the gene expression level in the seven cancer cell lines, respectively. Genes were eliminated from further consideration when \( S < 0 \) or \( S = 0 \), since only the genes with a score greater than a threshold (i.e., zero) are deemed potentially significant (Tuoya et al., 2008).

2.5 ssSOM analysis of gene expression
The expression levels of each gene were normalized among the breast cancer cell lines and normal breast tissue and among mouse normal tissues. First, the maximal RFI value of each gene was taken as 1, the minimum RFI was taken as 0 and other RFI values were linearly calculated into the values between 0 and 1. Secondly the average expression levels of each gene were calculated and each average was divided by the maximal average value. The resultant values were further multiplied to each normalized value calculated above. The normalized data were clustered and displayed by ssSOM software Cluster Blossom (Ver. 1.0.2, SOM Japan Co-Ltd., http://www.somj.com/). The training of Cluster Blossom were performed 50 times. Other parameters were automatically set by the software. Then the dendrograms were drawn from the final map after training by group average method with a glyph value 1.0.

2.6 Datasets for chemicals
The dataset analyzed in this study was taken from the previous report, in which 131 compounds were classified by the self organizing map with screening data against the 60 human cancer cell lines as input vectors (van Osdol et al., 1994). All these compounds structure data were downloaded from NCI databases by using Enhanced NCI Database Browser (http://129.43.27.140/ncidb2/). The names of compounds analyzed in this study are listed in Table 2 with NSC Nos. and MOA.

2.7 Descriptors of chemicals
All downloaded structures were submitted to the chemical descriptor calculation software, CDK Descriptor Calculator GUI (ver. 1.0.5; http://rguha.net/code/java/cdkdesc.html) to calculate 283 theoretical descriptors, including molecular descriptors, bond descriptors and atom descriptors (Steinbeck et al., 2003).

2.8 Descriptor scaling and selection
All above calculated descriptors were normalized by each row that they have mean 0 and variance 1 by the function of "normalize" in the "som package" of statistical software R
2.9 ssOM analysis of chemicals
Clustering were performed with the software Cluster Blossom (Ver. 1.0.3, SOM Japan Co-Ltd.). The trained ssOM was developed using the dataset above mentioned as input vectors. The same training parameters of Cluster Blossom were used as described above. Similarly, the dendrogram was drawn from the trained map as described above. The accuracy of clustering \( A \) was calculated as following.

\[
A = \frac{N_{tG_i}}{N_{G_i}} \times 100
\]

where \( N_{tG_i} \) is the number of compounds correctly assigned in cluster \( G_i \) and \( N_{G_i} \) is the number of compounds assigned in cluster \( G_i \) where \( i \) depicts the number of cluster.

3. Results and discussion
3.1 ssOM clustering of human breast cancer cell lines
We performed DNA microarray gene expression analysis in order to screen genes commonly and specifically expressed in the seven cell lines derived from breast cancer when compared to normal breast. As the result of data filtering, 840 genes were found to suffice the criteria described in “2.4 Data filtering in breast cancer cell”. The expression levels of these genes were then normalized and clustered by ssOM. The gene expression profiles were visualized on the sphere surface map and the dendrogram indicating themselves classified by the origin of the cells (Fig. 1). It is interesting to note that Hs-578T and MDA-MB-231 cells, which are derived from basal-like breast cancer known to have poor prognosis, are clustered in the same group (Ray et al., 2010). T-47D, ZR-75-1, MCF-7 and MDA-MB-134 cells, which are derived from luminal breast cancer, are well known to have good prognosis. Since SK-BR-3 cells are Her2 positive, which is an efficient target for the cancer therapy, and derived from breast cancer of medium level of prognosis. Thus, the gene expression profiles were successfully visualized by the ssOM clustering, suggesting the clusters of prognosis. From the patterns, cells derived from luminal breast cancer appear to be clustered into three groups of “close to normal”, “medium” and “poorer”. Namely, it might be possible to diagnose SK-BR-3 cells as “close to normal” while MDA-MB-134 as poorer than the other luminal derived cells.

In order to find genes highly expressed in all the seven cancer cell lines, ssOM was performed with an assumptive gene inserted into the dataset of the 840 genes. The assumptive gene stood for an ideal point IP, which was supposed to be expressed in all the breast cancer cell lines analyzed in this study but not in the normal breast tissue, so that the genes clustered close to IP should be potential diagnostic markers of breast cancer. In the result of ssOM clustering, IP was mapped in the red part of the pattern in all the seven cancer cell lines (Fig. 1) but blue in normal breast tissue. Since this mapped position of IP is consistent with the assumption, the genes close to IP should be selected as candidates of cancer-specific genes on the ssOM. Each spot on the surface of ssOM contains a group of clustered genes (Fig. 2). The spots mapped close to IP are shown in Fig. 2 and the candidate genes clustered in each spot are listed in Table 1. It is noteworthy that ErbB3 and ROBO2 have been nominated as potential diagnostic
markers here and some reports are found describing their relationships with breast cancers (Lemoine et al., 1992; Gasparini, 1994; Quinn et al. 1994 Travis et al., 1996; Naidu et al., 1998; Fogel et al., 1999; Holbro et al., 2003; Barnes et al., 2005; Schabath et al., 2006; Shiau et al. 2008). MUC1 is also known as a diagnostic marker in various cancers including breast cancers (Singh et al., 2008). Considering the results that contain these potential candidates, the other genes listed in Table 1 could be a potential candidate for the diagnostic marker of breast cancers still unknown.

Fig. 1. The gene expression profiles analyzed by sSOM for cancer derived cell lines and normal breast. The normalized data set was clustered and visualized by Cluster Blossom. Each position of genes is fixed on the global surface. The colors indicate the expression level for each gene. Red, high; yellow, slightly high; white, median; light blue, slightly low; deep blue, low. See text for the names of cell lines and diagnostic levels. The alignment of cells is the result of sSOM clustering, which was drawn by dendrogram.

| Gene No. | GenBank Accession No. | Gene Name |
|----------|-----------------------|-----------|
| 1586     | NM_032038             | spinster-like protein |
| 1423     | NM_016372             | seven transmembrane domain orphan receptor |
| 1784     | AH006947              | vitelliform macular dystrophy protein 2 |
| 1777     | M29366                | v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (ErbB3) |
| 1682     | NM_012471             | transient receptor potential cation channel, subfamily C, member 5 |
| 734      | NM_002099             | glycophorin A (includes MN blood group) |
| 1399     | AF040991              | roundabout, axon guidance receptor, homolog 2 (ROBO2) |
| 163      | NM_001188             | BCL2-antagonist/killer 1 |
| 247      | NM_001218             | carbonic anhydrase XII |
| 1699     | NM_003271             | transmembrane 4 superfamily member 7 |
| 241      | NM_022131             | calsyntenin 2 |
| 1015     | NM_002456             | mucin 1, transmembrane (MUC1) |

Table 1. Candidate genes for the potential diagnostic marker for breast cancer as picked up from genes commonly expressed in all the cancer derived cells studied in this paper.
3.2 sSOM clustering of mouse normal tissues

In this section, gene expression profiles of normal tissues in mouse were clustered by sSOM. Since the breast cancer cell lines were successfully clustered, we expected normal tissues should be also clustered with the features of each tissue. Clustering of brain, colon, heart, kidney, liver, lung, muscle, small intestine, spleen, stomach, testis and thymus was performed and the resultant gene expression profiles were aligned on the anatomical sketch of mouse body (Fig. 3). The relationship between each tissue was shown in a global map obtained by sSOM (Fig. 4). In this map, each distance between the nodes was not adjusted to a sphere surface (glyph =0) but reflected the distance when calculated by SOM (glyph =1.0) resulting in a meteoritic form of map.

The alignment of gene expression profiles around the body sketch reveals some similarities between the tissues. The similarity of the profiles between colon and intestine appears consistent. The similarity of profiles between spleen and thymus also sounds reasonable because of the deep relationship of these tissues with immunological system. The similarity is also found in heart, liver and lung. Although it is difficult to explain their close relationship from the embryonic development of tissues in mouse, it might be important to try to make viewpoints shared in these three tissues but not in other tissues as suggested by the sSOM clustering. Further application of sSOM on the gene expression profiles comparing with normal tissues and diseased tissues would lead to a challenging opportunity to find novel diagnostic markers in the future.
Fig. 3. Gene expression profiles of mouse normal tissues clustered by sSOM. The normalized data set was clustered and visualized by Cluster Blossom. Each position of genes is fixed on the global surface. See Fig. 1 for the colors indicating the expression level for each gene. Views of clustered global map from front side (A) and back side (B).
3.3 sSOM clustering of bioactive chemicals

The bioactive compounds previously screened for anti-cancer reagents were evaluated for clustering in this study. The compounds were clustered by sSOM. The dendrogram was drawn based on the trained map by group average method to obtain 9 clusters of compounds, which were colored by their clusters on the surface of global map (Fig. 5). The compounds in the dataset are summarized in Table 2 with their assigned MOA and the cluster groups. Table 3 shows the clustering results of compounds. The accuracy of clustering was overall 86.2%, ranging from 60 to 100% in each cluster. The alkylating agents, AC, A7, and A1, are misclassified relatively in higher frequency than other agents. It is interesting to note the anti-DNA agents, DI, DP, and DR, and the inhibitors of nucleotide synthesis, RI, RO, and R, are clustered into the same group. This might be the result due to the character of these agents associating with the enzymes associated with nucleotide metabolisms.

In this study, 16 compounds (ID 5, 17, 29, 33, 44, 46, 52, 64, 81, 86, 88, 92, 93, 96, 99) were misclassified. These results suggest that they might have another activity other than those experimentally defined because small organic compounds frequently exhibit
polypharmacology. In fact, trimetrexate (ID 88) and DUP785 (ID 96) might have topoisomerase inhibiting activity because both of them have resemble planar heteroaromatic ring, which is the feature of topoisomerase inhibitor. Additionally, mitzolamide (ID 89) has also heteroaromatic ring implying DNA interacting ability. The chemical structures of these three compounds are shown in Fig. 6. Currently, exploring new targets and activity of already approved drugs is fascinating strategy to develop novel therapeutic drugs with less risks of the clinical trial (Keiser et al., 2009). Although further investigation is needed, sSOM would be a comprehensive and useful tool to classify the compounds and to find novel activities in themselves.

| ID | NSC No. | Drug Name          | MOA | cluster |
|----|---------|--------------------|-----|---------|
| 1  | NSC740  | Methotrexate       | RF  | G1      |
| 2  | NSC750  | Busulfan           | A7  | G6      |
| 3  | NSC752  | Thioguanine        | DI  | G7      |
| 4  | NSC755  | Thiopurine         | DI  | G7      |
| 5  | NSC757  | Colchicine         | TU  | G9      |
| 6  | NSC762  | Mechlorethamine    | A7  | G6      |
| 7  | NSC1895 | Guanazole          | DR  | G7      |
| 8  | NSC3088 | Chlorambucil       | A7  | G6      |
| 9  | NSC6396 | Thiotepa           | A7  | G7      |
| 10 | NSC8806 | Melphalan          | A7  | G6      |
| 11 | NSC9706 | Triethylenemelamine| A7  | G7      |
| 12 | NSC19893| Fluprouracil       | A7  | G7      |
| 13 | NSC25154| Pipobroman         | A7  | G6      |
| 14 | NSC26980| Mitomycin          | A2  | G7      |
| 15 | NSC27640| Fluorouridine      | DP  | G7      |
| 16 | NSC32065| Hydroxyurea        | DR  | G7      |
| 17 | NSC33410| Colchicine derivative| TU  | G9      |
| 18 | NSC34462| Uracil mustard     | A7  | G6      |
| 19 | NSC49842| Vinblastine sulfate| TU  | G3      |
| 20 | NSC51143| Pyrazoloimidazole  | DR  | G7      |
| 21 | NSC56410| Porfiromycin       | A2  | G7      |
| 22 | NSC63878| Cytarabine         | DP  | G7      |
| 23 | NSC67574| Vincristine sulfate| TU  | G3      |
| 24 | NSC71261| beta-2-Deoxythioguanosine| DI  | G7  |
| 25 | NSC71851| alpha-2-Deoxythioguanosine | DI | G7  |
| 26 | NSC73754| Fluorodopan        | A7  | G6      |
| 27 | NSC79037| Lomustine          | AC  | G6      |
| 28 | NSC82151| Daunorubicin       | T2  | G2      |
| No. | NSC Code | Compound Name                              | Source | Stage |
|-----|----------|--------------------------------------------|--------|-------|
| 29  | NSC83265 | Trityl cysteine                            | TU     | G9    |
| 30  | NSC94600 | Camptothecin                               | T1     | G9    |
| 31  | NSC95382 | Camptothecin derivative                    | T1     | G9    |
| 32  | NSC95441 | Semustine                                  | AC     | G6    |
| 33  | NSC95466 | PCNU                                        | AC     | G7    |
| 34  | NSC95678 | 3-Hydroxypicolinaldehyde thiosemicarbazone | DR     | G7    |
| 35  | NSC100880| Camptothecin derivative                    | T1     | G9    |
| 36  | NSC102627| Yoshio-864                                 | A7     | G6    |
| 37  | NSC102816| Azacytidine                                | RO     | G7    |
| 38  | NSC107124| Camptothecin derivative                    | T1     | G9    |
| 39  | NSC107392| 5-Hydroxypicolinaldehyde thiosemicarbazone | DR     | G7    |
| 40  | NSC118994| Inosine glycodialdehyde                    | DR     | G7    |
| 41  | NSC122819| Teniposide                                 | T2     | G2    |
| 42  | NSC123127| Doxorubicin                                | T2     | G2    |
| 43  | NSC125973| Paclitaxel derivative                      | TU     | G3    |
| 44  | NSC126771| Dichloroallyl lawsone                      | RO     | G6    |
| 45  | NSC127716| 5-Aza-2'-deoxyctydine                      | DI     | G7    |
| 46  | NSC132313| Dianhydrogalactitol                        | A7     | G7    |
| 47  | NSC132483| Aminopterin                                | RF     | G1    |
| 48  | NSC134033| Aminopterine derivative                    | RF     | G1    |
| 49  | NSC135758| Piperazinedione                            | A7     | G6    |
| 50  | NSC139105| Baker's soluble antifolate                 | RF     | G5    |
| 51  | NSC141540| Etoposide                                  | T2     | G2    |
| 52  | NSC142982| Hycanthone                                 | A1     | G9    |
| 53  | NSC143095| Pyrazofurin                                | RO     | G7    |
| 54  | NSC145668| Cycloctydine                               | DP     | G7    |
| 55  | NSC14895 | Ftorafur                                   | R      | G7    |
| 56  | NSC153353| L-Alanosine                                | RO     | G7    |
| 57  | NSC153858| Maytansine                                 | TU     | G5    |
| 58  | NSC163501| Acivicin                                    | RI     | G7    |
| 59  | NSC164011| Zorubicin                                  | T2     | G2    |
| 60  | NSC167780| Asaley                                     | A7     | G5    |
| 61  | NSC172112| Spiromustine                               | A7     | G6    |
| 62  | NSC174121| Methotrexate derivative                    | RF     | G1    |
| 63  | NSC176323| Camptothecin derivative                    | T1     | G9    |
| 64  | NSC178248| Chlorozotocin                              | AC     | G7    |
|    | NSC           | Chemical Name                                           | Tissue | Growth Zone |
|----|---------------|---------------------------------------------------------|--------|-------------|
| 65 | NSC182986     | Diaziridinylbenzoquinone                               | A7     | G7          |
| 66 | NSC184692     | Aminopterin derivative                                 | RF     | G1          |
| 67 | NSC224131     | N-(phosphonoacetyl-L-aspartic acid, tetrasodium salt)    | RO     | G7          |
| 68 | NSC249910     | Camptothecin derivative                                | T1     | G9          |
| 69 | NSC249992     | Amsacrine                                               | T2     | G9          |
| 70 | NSC264880     | 5,6-Dihydro-5-azacytidine                              | RO     | G7          |
| 71 | NSC267469     | Deoxydoxorubicin                                        | T2     | G2          |
| 72 | NSC268242     | N,N-Dibenzyldaurubicin                                  | T2     | G2          |
| 73 | NSC269148     | Menogaril                                               | T2     | G2          |
| 74 | NSC295500     | Camptothecin derivative                                 | T1     | G9          |
| 75 | NSC295501     | Camptothecin derivative                                 | T1     | G9          |
| 76 | NSC296934     | Teroxirone                                              | A7     | G7          |
| 77 | NSC301739     | Mitoxantrone                                            | T2     | G2          |
| 78 | NSC303812     | Aphidicolin glycinate                                   | DP     | G4          |
| 79 | NSC308847     | Amonafide                                               | T2     | G9          |
| 80 | NSC329680     | Hepesulfam                                              | A7     | G6          |
| 81 | NSC330500     | Geldanamycin                                            | DP     | G9          |
| 82 | NSC332598     | Rhizoxin                                                | TU     | G5          |
| 83 | NSC337766     | Bisantrene                                              | T2     | G9          |
| 84 | NSC338947     | Clomesone                                               | AC     | G6          |
| 85 | NSC344007     | Piperazine alkylator                                     | A7     | G6          |
| 86 | NSC348948     | Cyclodisone                                             | AC     | G7          |
| 87 | NSC349174     | Oxanthrazole                                            | T2     | G2          |
| 88 | NSC352122     | Trimetrexate                                            | RF     | G9          |
| 89 | NSC353451     | Mitozolamide                                            | AC     | G7          |
| 90 | NSC354646     | Morpholino adriamycin                                   | T2     | G2          |
| 91 | NSC355644     | Anthrapyrazole derivative                               | T2     | G2          |
| 92 | NSC357704     | Cyanomorpholinodoxorubicin                             | A1     | G2          |
| 93 | NSC361792     | Thiocolchicine                                          | TU     | G9          |
| 94 | NSC364830     | Camptothecin derivative                                 | T1     | G9          |
| 95 | NSC366140     | Pyrazoloacridine                                        | T2     | G9          |
| 96 | NSC368390     | DUP785 (brequinar)                                     | RO     | G9          |
| 97 | NSC374028     | Camptothecin derivative                                 | T1     | G9          |
| 98 | NSC376128     | Dolastatin 10                                           | TU     | G5          |
| 99 | NSC406042     | Allocolchicine                                          | TU     | G9          |
|100 | NSC409962     | Carmustine                                              | AC     | G6          |
Table 2. Compounds in the dataset and the result of clustering. Abbreviations in MOA are as following. DNA alkylating agents: A2, alkylating at N-2 position of guanine; AC, alkyl transferase-dependent cross-linkers; A7, alkylating at N-7 position of guanine; AI, DNA intercalators. Anti-DNA agents: DI, incorporated; DP, polymerase inhibitors; DR, ribonuclease reductase inhibitors. Nucleotide synthesis inhibitors: RF, antifolates; RI, irreversible inhibitors; RO, anti other precursors; R, unknown locus of inhibition. Topoisomerase inhibitor: T1, topoisomerase I inhibitors; T2, topoisomerase II inhibitors. Tubulin-active antimitotic agents: TU

| Compound ID | Name | MOA | Cluster | Accuracy (%) |
|-------------|------|------|---------|--------------|
| NSC603071   | Camptothecin derivative | T1 | G9 | 100 |
| NSC606172   | Camptothecin derivative | T1 | G9 | 92 |
| NSC606173   | Camptothecin derivative | T1 | G9 | 100 |
| NSC606497   | Camptothecin derivative | T1 | G9 | 60 |
| NSC606499   | Camptothecin derivative | T1 | G9 | 94 |
| NSC606985   | Camptothecin derivative | T1 | G9 | 82 |
| NSC608832   | Paclitaxel derivative | TU | G3 | 100 |
| NSC610456   | Camptothecin derivative | T1 | G9 | 75 |
| NSC610457   | Camptothecin derivative | T1 | G9 | 92 |
| NSC610458   | Camptothecin derivative | T1 | G9 | 100 |
| NSC610459   | Camptothecin derivative | T1 | G9 | 60 |
| NSC618939   | Camptothecin derivative | T1 | G9 | 94 |
| NSC623017   | an. Antifol II | RF | G1 | 100 |
| NSC629971   | Camptothecin derivative | T1 | G9 | 75 |
| NSC633713   | an. Antifol II | RF | G1 | 100 |
| NSC643833   | Camptothecin derivative | T1 | G9 | 75 |

Table 3. Summary of clustering compounds in this study. See Table 2 for the abbreviations for MOA. Accuracy was calculated as described in “2.9 sSOM analysis of Chemicals”
Fig. 5. Projections of clustered bioactive compounds by sSOM. Cluster colors: G1, blue; G2, yellow; G3, gray; G4, green; G5, pink; G6, cyan; G7, yellowish green; G8, white; G9, red. Numbers at the nodes indicate the ID of compounds. Two views on a single global map are shown from the opposite directions.

Fig. 6. Chemical structures of trimetrexate (ID 88), brequinar (ID 96), and mitozolamide (ID 89).

4. Conclusions

In order to characterize cells and tissues, gene expression profiling is one of the most popular procedures nowadays. Through these procedures, identification of cell surface markers specific to some cells or tissues is a key for diagnosing and molecular targeting. DNA microarray is a high-throughput technology believed to be a powerful tool to find genes differentially expressed in the cells or tissues. Although it can provide critically important and useful information even from one experiment, the amount of data is usually too large to be handled. Therefore, highly sophisticated software is expected to support to transform the multidimensional datasets into simple dimensions or glyphs. For example, visual cues such as shape and color, which make it comprehensive for researchers to...
recognize and analyze the patterns hidden in the datasets. Here we successfully demonstrated surface marker analyses using our DNA microarray coupled with novel sSOM clustering procedure. The cell surface markers, which are common and specific to cancer derived cells, are proposed in this study and further assessment is now underway. Here we have also examined sSOM for the classification of chemical compounds. sSOM successfully clustered 116 anti-cancer agents into 9 groups by their MOA using simple chemical descriptors as inputs. So we are now trying to apply this procedure to larger dataset for virtual screening.

Thus, we conclude sSOM is a powerful tool for data mining, knowledge discovery and visualization of multi-dimensional data.

5. References

Abou-Sharieha, S; Sugii, Y; Tuoya; Yu, D; Chen, L; Tokutaka, H & Seno, M. (2009) Identification of TM9SF2 as a candidate of the cell surface marker common to breast carcinoma cells. Clin. Oncol. Cancer Res. 6(1), 1-9, ISSN: 1674-5361.

Barnes, N.L.; Khavari, S.; Boland, G.P.; Cramer, A.; Knox, W.F. & Bundred, N.J. (2005) Absence of HER4 expression predicts recurrence of ductal carcinoma in situ of the breast. Clin. Cancer Res. 11(6), 2163-2168, ISSN: 1078-0432.

Ben-Dor, A.; Shamir, R. & Yahkini, Z. (1999) Clustering gene expression patterns. Journal of Computational Biology, 6(3/4) 281-297, ISSN: 1557-8666.

Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T & Clark T. (2002) Descriptors, physical properties, and drug-likeness. J. Med. Chem. 45(16), 3345-3355, ISSN: 0022-2623.

Fogel, M.; Friederichs, J.; Zeller,Y.; Husar, M.; Smirnov, A.; Roitman, L.; Altevogt, P. & Sthoeger, Z.M. (1999) CD24 is a marker for human breast carcinoma. Cancer Lett. 143(1), 87-94, ISSN: 0304-3835.

Gasparini, G.; Gullick, W.J.; Maluta, S.; Dalla Palma, P.; Caffo, O.; Leonardi, E.; Boracchi, P.; Pozza, F.; Lemoine, N.R. & Bevilacqua, P. (1994) C-erbB-3 and c-erbB-2 protein expression in node-negative breast carcinoma—an immunocytochemical study. Eur J Cancer, 30A(1), 16-22, ISSN: 0014-2964.

Gasteiger, J.; Teckentrup, A.; Terfloth, L. & Spycher, S. (2003) Neural networks as data mining tools in drug design, J. Phys. Org. Chem. 16 (4), 232-245, ISSN: 1099-1395.

Holbro, T.; Givneni, G.; Hynes, N.E. (2003) The ErbB receptors and their role in cancer progression. Exp Cell Res. 284(1), 99-110, ISSN: 0014-4827.

Kaiser, D.; Terfloth, L.; Kopp, S.; Schulz, J.; Laet, de R.; Chiba, P.; Ecker F. G. & Gasteiger J. (2007) Self-organizing maps for identification of new inhibitors of P-glycoprotein. J. Med. Chem. 50(7), 1698-1702, ISSN: 0022-2623.

Kaufman, L. & Rousseeuw, P.J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, ISBN: 978-0-471-73578-6, New York.

Keiser, M.J.; Setola, V.; Irwin, J.J.; Laggner, C.; Abbas, A.L.; Hufeisen, S.J.; Jensen, N.H.; Kuijer, M.B.; Matos, R.C.; Tran, T.B.; Whaley, R.; Glennon, R.A.; Hert, J.; Thomas, K.L.; Edwards, D.D.; Shoichet, B.K. & Roth, B.L. (2009) Predicting new molecular targets for known drugs. Nature 462(7270), 175-181, ISSN 0028-0836.

Lemoine, N.R.; Barnes, D.M.; Hollywood, D.P.; Hughes, C.M.; Smith, P.; Dublin, E.; Prigent, S.A.; Gullick, W.J.; Hurst HC. (1992) Expression of the ERBB3 gene product in breast cancer. Br J Cancer 66(6), 1116-1121, ISSN 0007-0920.
Clustering Genes, Tissues, Cells and Bioactive Chemicals by Sphere SOM

Li J. & Gramatica P. (2010) Classification and virtual screening of androgen receptor antagonists. *J. Chem. Inf. Model.* 50(5), 861-874, ISSN: 1549-9596.

MacQueen, J. (1965) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281-297.

Melville, J. L.; Burke, E. K.; & Hirst, J. D. (2009) Machine learning in virtual screening. *Combinatorial Chemistry and High Throughput Screening*, 12(4), 332-343, ISSN 1386-2073.

Naidu, R.; Yadav, M.; Nair, S. & Kutty, M.K.(1998) Expression of c-erbB3 protein in primary breast carcinomas. *Br J Cancer*. 78(10), 1385-1390, ISSN 0007-0920.

Quinn, C.M.; Ostrowski, J.L.; Lane, S.A.; Loney, D.P.; Teasdale, J., Benson, F.A. (1994) C-erbB-3 protein expression in human breast cancer: comparison with other tumor variables and survival. *Histopathology*, 25(3), 247-252, ISSN 0309-0167.

R Development Core Team (2007) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ray, P.S.; Wang, J.; Qu, Y.; Sim, M.S., Shamonki, J.; Bagaria, S.P.; Ye, X.; Liu, B.; Elashoff, D.; Hoon, D.S.; Walter, M.A.; Martens, J.W.; Richardson, A.L.; Giuliano, A.E.& Cui, X. (2010) FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. *Cancer Res.*, 70(10), 3870-3786, ISSN 0008-5472.

Reddy S. A.; Pati P. S.; Kumar P. P.; Pradeep N. H. & Sastry N. G. (2007) Virtual screening in drug discovery – a computational perspective. *Current Protein And Peptide Science*, 8(4), 329-351, ISSN 1389-2037.

Renner, S.; Hechenberger, M.; Noeske, T.; Böcker, A.; Jatzke, C.; Schmuker, M.; Parsons, G. C.; Wel, T. & Schneider G. (2007) Searching for Drug Scaffolds with 3D Pharmacophores and Neural Network Ensembles. *Angew. Chem. Int. Ed.* 46(28), 5336-5339, ISSN: 1521-3773.

Schabath, H.; Runz, S.; Joumaa, S. & Altevogt, P. (2006) CD24 affects CXCR4 function in pre-B lymphocytes and breast carcinoma cells. *J Cell Sci.*, 119(P12), 314-325, ISSN: 0021-9533.

Schneider, G. & Nettekoven, M. (2003) Ligand-Based Combinatorial Design of Selective Purinergic Receptor (A2a) Antagonists Using Self-Organizing Maps. *J. Comb. Chem.* 5(3), 233-237, ISSN: 1520-4766.

Schneider, P. & Schneider, G. (2003) Collection of Bioactive Reference Compounds for Focused Library Design. *QSAR Comb. Sci.* 22(7), 713-718, ISSN: 1868-1751.

Shiau, C.E.; Lwigale, P.Y.; Das, R.M.; Wilson, S.A.& Bronner-Fraser M. (2008) Robo2-Slit1 dependent cell-cell interactions mediate assembly of the trigeminal ganglion. *Nat. Neurosci.*, 11(3), 269-276, ISSN: 1097-6256.

Singh, A.P.; Senapati, S.; Ponnusamy, M.P.; Jain, M.; Lele, S.M.; Davis, J.S.; Remmenga, S.& Batra SK. (2008) Clinical potential of mucins in diagnosis, prognosis, and therapy of ovarian cancer. *Lancet Oncol.* 9(11), 1076-1085, ISSN: 1470-2045.

Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E. & Willighagen, E. (2003) The chemistry development kit (CDK): An open source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* 43(2), 493-500, ISSN: 1549-9596.
Travis, A.; Pinder, S.E.; Robertson, J.F.; Bell, J.A.; Wencyk, P.; Gullick, W.J.; Nicholson, R.I.; Poller, D.N.; Blamey, R.W.; Elston, C.W. & Ellis, I.O. (1996) C-erbB-3 in human breast carcinoma: expression and relation to prognosis and established prognostic indicators. *Br J Cancer* 74(2), 229-233, ISSN 0007-0920.

Tuoya; Hirayama, K; Nagaoka, T; Yu, D; Fukuda, T; Tada, H; Yamada, H. & Seno M. (2005) Identification of cell surface marker candidates on SV-T2 cells using DNA microarray on DLC-coated glass. *Biochem. Biophys. Res. Commun.* 334(1), 263-268, ISSN: 0006-291X.

Tuoya; Sugii, Y.; Satoh, H.; Yu, D.; Matsuura, Y.; Tokutaka, H. & Seno, M. (2008) Spherical self-organizing map as a helpful tool to identify category-specific cell surface markers. *Biochem Biophys Res Commun.* 376(2), 414-418, ISSN: 0006-291X.

van Osdol, W.W.; Myers, T.G.; Paull, K.D.; Kohn, K.W. & Weinstein, J.N. (1994) Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* 86(24), 1853-1859, ISSN 0027-8874.

Wang, Y.; Li, Y.; Yang, S. & Yang, L. (2005) Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* 45(3), 750-757, ISSN: 1549-9596.
Kohonen Self Organizing Maps (SOM) has found application in practical all fields, especially those which tend to handle high dimensional data. SOM can be used for the clustering of genes in the medical field, the study of multi-media and web based contents and in the transportation industry, just to name a few. Apart from the aforementioned areas this book also covers the study of complex data found in meteorological and remotely sensed images acquired using satellite sensing. Data management and envelopment analysis has also been covered. The application of SOM in mechanical and manufacturing engineering forms another important area of this book. The final section of this book, addresses the design and application of novel variants of SOM algorithms.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yuh Sugii, Takayuki Kudoh, Takayuki Otani, Masashi Ikeda, Heizo Tokutaka and Masaharu Seno (2011). Clustering Genes, Tissues, Cells and Bioactive Chemicals by Sphere SOM, Self Organizing Maps - Applications and Novel Algorithm Design, Dr Josphat Igadwa Mwasiagi (Ed.), ISBN: 978-953-307-546-4, InTech, Available from: http://www.intechopen.com/books/self-organizing-maps-applications-and-novel-algorithm-design/clustering-genes-tissues-cells-and-bioactive-chemicals-by-sphere-som
