DISCOVERING GENOMIC PATTERNS USING FUZZY SELF-ORGANISING MAPS

by

CHUN HO YI

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science (Mathematics)

FEBRUARY 2009

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Yahya Abu Hasan from the School of Mathematics, Universiti Sains Malaysia, for his guidance during the course of my studies. It is with his advice and wisdom that I am able to complete this thesis.

I would also like to thank Dr. James Ashman, then a visiting lecturer from the School of Pharmaceutical Sciences, Universiti Sains Malaysia, for his advice and wisdom on microarrays, genetics and biochemistry in general. I thank him for helping me expand my knowledge on genetics and biochemistry, and indirectly, helping me greatly with my research. My thanks go out to Dr. Zainudin as well for his guidance in statistical methods, of which has helped me in my work.

I would also like to thank my family and friends, all of who have supported me for my choice to take my studies further. To that special person in my life who has been patient with me throughout the course of my studies. Last, but not least, to my parents especially, who have continued to support me throughout my studies regardless of the difficulties I faced.

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION

CHAPTER 2 – BASICS OF GENETICS AND MICROARRAYS

CHAPTER 3 – SELF-ORGANISING MAPS (SOM)

LIST OF TABLES

LIST OF FIGURES

MENEMUI POLA GENOM MENGGUNAKAN PETA PENYUSUNAN SENDIRI KABUR

ABSTRAK

Teknologi barisan-mikro telah membolehkan pengumpulan beribu-ribu data genetik pada masa yang sama. Masalah yang dihadapi oleh data barisan-mikro adalah bilangan ciri yang banyak tetapi bilangan sampel yang sedikit. Pengerumunan menjadi pilihan yang praktikal dalam menganalisa data begini kerana untuk mengkaji setiap satu data adalah tak praktis. Walapun terdapat berbilang jenis kaedah pengerumunan seperti "k-means" dan pengerumunan berpangkat, namun ini tidak mencukupi apabila menglibatkan set data yang sangat besar. Peta menyusun sendiri (SOM) adalah satu konsep yang diperkenalkan oleh Teuvo Kohonen yang membolehkan pengerumunan yang tidak diawasi dan penggambaran data. Penggambaran ini dilakukan dengan pemancaran data tersebut ke atas satu jaringan peta 2-dimensi sambil mengekalkan hubungan antara unit peta yang mewakili data tersebut. Lagipun, penggunaan satah komponen dalam SOM membolehkan analisis setiap ciri dan perhubungan antara ciri. Secara keseluruhan, pengerumunan dan penggambaran ini membolehkan analisis set data secara keseluruhan. Dalam genetik, pengerumunan keras mungkin tidak bersesuaian kerana berbilang gen boleh mempunyai berbilang sifat, dan ini bermakna berbilang kelompok dalam pengerumunan. Ini pula membawa kepada topik untuk menggabungkan peraturan kabur, dimana satu ciri tidak terhad kepada satu sifat. Oleh itu, satu penyelesaian yang munasabah adalah untuk mengacukkan "c-means" kabur dengan SOM. Walaupun SOM kabur (FSOM) telah diselidiki, namun ia memerlukan banyak parameter sebelum pemprosesan. Kerja ini mencadangkan satu FSOM terubah yang memerlukan parameter awal yang kurang, dan mengujinya ke atas suatu set data tumor otak dan set data kanser payudara. Hasilnya menunjukan bahawa prestasi FSOM terubah ini adalah lebih baik daripada SOM asal, dan hasil analisis pengerumunan adalah serupa dengan penyelidikan awal.

DISCOVERING GENOMIC PATTERNS USING FUZZY SELF-ORGANISING MAPS

ABSTRACT

Microarray technology has allowed for the collection of thousands of genetic data simultaneously. The problem with microarray data lies in the fact that while there are thousands of attributes, the number of samples is few. Clustering becomes a practical option in analysing such data as it is impractical to study every data individually. Although there are numerous clustering methods such as k-means and hierarchical clustering, it is insufficient when it comes to a massive scale dataset. Self-organising maps (SOM) is a concept introduced by Teuvo Kohonen that allowed for an unsupervised clustering and subsequent visualisation of data. This visualisation is by means of projecting the data onto a 2-dimensional map grid while retaining relations among map units that represent the data. Furthermore, the use of component planes in SOM allows for the analysis of individual attributes and correlation between attributes. On the whole, this clustering and visualisation allows for a "big picture" analysis of a massive dataset. In genetics, hard clustering may not be entirely suitable as multiple genes can have multiple traits, and thus, in clustering, can belong to multiple clusters. This brings up the subject of incorporating fuzzy rules, where one attribute is not restricted to have only one characteristic. Therefore, a feasible solution would be a hybridisation of fuzzy c-means and SOM. While fuzzy SOM (FSOM) has been explored, it requires many parameters prior to processing. This work proposes a modified FSOM that requires very few initial parameters, and was tested on a brain tumour dataset and a breast cancer dataset. The results showed that the modified FSOM performed better than the ordinary SOM, and the results of the cluster analysis matched a prior work.

CHAPTER 1

1.1 Introduction

Microarray technology has hastened oncological research by leaps and bounds with the ability to measure the expression of thousands of genes in a single slide or chip. While microarray experiments are costly and time consuming, this has opened up many possibilities in oncological research and provided for the advancement in finding the much sought after cure for cancer. This breakthrough, however, has its drawbacks. The production of a single massive dataset containing thousands of genes for a single sample has brought up many statistical issues. One sample with thousands of attributes is hardly "ideal" statistically. Typically, the type of data used for statistical analysis consists of a few attributes with many samples. While a microarray research involves many more samples, it still does not match the number of genes. A single experiment may use about 100 samples, but most will have genes numbering in the thousands, and this figure may well extend to tens of thousands of genes. It would be a major task for a microarray experiment to be conducted on 1000 samples, much less a few thousand samples. Thus, we are left with a major statistical problem of analysing such massive dataset.

Self-organising maps (SOM) is a concept introduced by Teuvo Kohonen (Kohonen 1995). It has the characteristic of, as its name goes, organising the data on its own with every iteration of its algorithm, based on the concept of neighbouring association. SOM has the advantage of clustering and projecting a large dataset onto a simple 2-dimensional projection, allowing for a visual analysis as opposed to a numerical one. The interpretation of SOM is very subjective due to its visual nature, which makes it ideal for a "big picture" analysis before any detailed analysis.

Conclusions and interpretations are drawn from individual knowledge and expertise based on the visualisations. This makes SOM a tool that can be used by anyone in any field.

The nature of SOM has made microarray data analysis much simpler. The ability to visualise massive datasets onto a 2-dimensional projection has allowed for thousands of incomprehensible numbers to become meaningful and patterns of correlation to become clear. Toronen et al (1999), Tamayo et al (1999), Torkkola et al (2001), Xiao et al (2003), Hautaniemi et al (2003), Garrigues et al (2004) are some works that have explored the use of SOM in microarray data mining involving breast cancer, prostrate cancer, yeast and macrophages. In those works, SOM has proven to be efficient in visualising the massive datasets containing thousands of genes. Similarly, variations of SOM have also been explored and used in the analysis of microarray data, such as in Sultan et al (2002) and Wu et al (2005), proposing hybrids of SOM, respectively with partitive k-means and multidimensional scaling. Here, a variation of SOM incorporating fuzzy set rules is explored.

Fuzzy rules are non-discrete rules that involve assigning membership of attributes to more than just one characteristic. Like its name, it allows for an attribute to be "fuzzy", having some of each characteristic (at varying degrees), rather than being confined to a single one. In the case of microarray data analysis, traditional clustering methods restrict one gene to one cluster. This can be very limiting, especially in the field of genetics. A single gene produces a single type of protein, but multiple proteins are involved in various bodily functions. Proteins are the primary components of enzymes, most hormones and many cellular components

(Keeton 1972). By incorporating fuzzy rules into SOM, a different kind of self-organisation is expected where the organisation of data is no longer confined to rigid assignments. This concept of incorporating fuzzy rules into SOM has been explored as well, such as in Bezdek et al (1992), Vuorimaa (1994) and Hu et al (2004), but little is seen of its use on massive datasets.

Hu et al (2004) presented a fuzzy SOM (FSOM) that was used for learning activity patterns. For testing, the FSOM algorithm was implemented and used on a microarray dataset. The result was a very homogenous SOM, making any interpretation impossible. In this work, a modified version of fuzzy SOM is presented and tested on a large scale microarray dataset. This work is motivated by the need for a fuzzy algorithm to require as few parameters and assumptions as possible. Fuzzy c-means clustering require some parameters prior to processing, such as the number of centres and "fuzziness factor" (or fuzzy variable). Other fuzzy SOM algorithms would fix these parameters (such as in Bezdek et al 1992, Abonyi et al 2002). The fuzzy SOM proposed by Hu et al (2004) did not require such parameters, but have omitted the neighbourhood function from the SOM algorithm. The neighbourhood function forms the core of SOM, and is vital in its self-organising characteristic. Furthermore, the proposed FSOM does not work well with massive datasets such as microarray datasets. Thus, a modified fuzzy SOM (MFSOM) algorithm is proposed here that requires few parameters and retains the characteristics of SOM while incorporating fuzzy rules, namely membership functions, into SOM.

There are 2 different microarray datasets that are used here, and they are a brain tumour dataset and a breast cancer dataset. The publicly available brain tumour dataset was trimmed to a very small number of genes and processed with the

MFSOM as a test to see how well the proposed algorithm works on a small scale dataset. Subsequently, the proposed algorithm will be tested on a much larger scale dataset, which is the breast cancer dataset. This dataset was obtained from the supplementary of the paper by Sotiriou et al (2003). This work does not seek to confirm or verify previous works, but only to be used as comparison for testing out the proposed algorithm. Therefore, the biological implications and conclusions are not explored here.

1.2 Methodology

Prior to using the MFSOM algorithm, both datasets first undergo a preparation stage. All raw datasets require pre-processing before they can be used by any software package. In the first step of data preparation, both datasets are treated to the same process of removing unnecessary data and text are from the data files, keeping only the necessary gene labels and sample names. For the subsequent parts, the brain tumour and breast cancer datasets are subjected to different preparation methods.

In this work, the data preparation was done using Mathematica 5.0 by Wolfram Research and the implementation of SOM and the fuzzy variations are done in Matlab 6.5 with the SOM Toolbox for Matlab. The SOM Toolbox for Matlab was developed by Vesanto et al (2000) from the Helsinki University of Technology. Modifications to the program code were made to implement algorithm changes, namely the implementation of the FSOM and the MFSOM algorithms.

The brain tumour dataset, comprising of 7070 genes with 69 samples and 5 classes, was first subjected to thresholding, i.e. forcing the data range of the entire dataset to [20, 16000]. Next, the numbers of genes are trimmed by filtering out those with low fold difference. Subsequently, the T-Value for the remaining genes based upon their classes is calculated, and the top 30 genes with the highest T-Value for each class are selected. Finally, the dataset, now comprising of 145 genes, is edited to a specific format for use by the SOM Toolbox for Matlab.

The breast cancer dataset was prepared differently. Comprising of 7650 genes with 99 samples, the number of genes were trimmed by using the standard deviation, a simple measure of variability. Genes with high variability were kept while removing those with low variability. The result was a dataset of 2081 genes. From the patient information table, 4 more variables were added to the dataset. The final dataset contained 2085 attributes with 99 samples. The last step, like the brain tumour dataset, was to edit the dataset to a specific format for use by the SOM Toolbox.

Both datasets are then processed using the normal SOM batch algorithm, the FSOM as proposed by Hu et al (2004) and the MFSOM as proposed in this work. For both datasets, the quantisation and topographic error were used as the measure of quality of the algorithm. Quantisation error refers to the mapping precision of the algorithm while topographic error refers to how well the topology of the input data has been preserved during training and how smooth the map is.

1.3 Contribution

This work has contributed an improved fuzzy SOM algorithm that works on massive datasets such as microarray datasets. Also, this improved fuzzy SOM algorithm produced lower mapping error on the massive datasets as compared to the original SOM. Furthermore, the clusters produced in the projections are more defined and apparent, allowing for easier visual analysis.

1.4 Chapter Summary

Chapter 1 introduces the reader to the works of this thesis. It introduces a summary of the scope and work of this thesis.

Chapter 2 describes in brief the workings of genetics and microarrays. Specifically, the process of protein synthesis is explained. This in turn illustrates the way microarrays work, and how it is related to the study of genetics.

Chapter 3 explores SOM as a whole, including how it works to self-organise data and its visualisations. The function of the SOM visualisations and their use is also explained.

Chapter 4 describes the steps taken to prepare the data for analysis. The data preparation steps for both the brain tumour and breast cancer datasets are described in detail.

Chapter 5 explains in detail the work of this thesis, mainly the proposed MFSOM algorithm.

Chapter 6 discusses the results of the proposed algorithm, as well as its performance and the visualisations obtained. The visualisations of the datasets are also presented.

Chapter 7 concludes this thesis with a summary of the results and discussion, as well as proposed future research based upon this thesis.

CHAPTER 2

2.1 Basics of Genetics

In order to understand microarrays, one must first understand the concept of genetics and protein synthesis at the cellular level. One must first know that DNA (deoxyribonucleic acid) is where the information for protein production is stored, and proteins are basically polypeptide chains consisting of a string of amino acids. DNA is a double-stranded double helix (refer to figure 2.1), consisting of only four types of nucleic acid. These nucleic acids are arranged in pairs, thus giving it its double-stranded characteristic. The nucleic acids are adenine, thymine, cytosine and guanine. The pairings of these nucleic acids within the DNA strand are very specific. Adenine is paired only with thymine, and cytosine is paired only with guanine. Thus, a DNA strand consists of multiple repetitions of only these 2 complementary pairings (called base pairs), making its sequence a very important factor in protein synthesis as it acts as a template for protein synthesis.

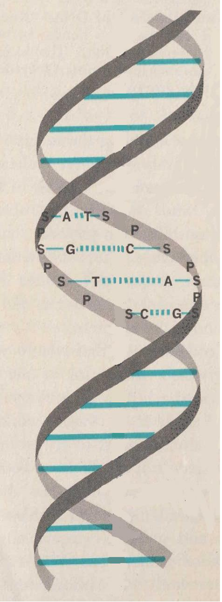

Figure 2.1: The Watson-Crick model of the DNA structure as known today. Note the complementary pairings of the nucleic acids labelled as A, T, C and G. Illustration from Keeton (1972)

Proteins are made of amino acids (20 different types of amino acids in all), and these amino acids are marked by a specific sequence on the DNA. The actual process of protein synthesis begins with ribonucleic acid (RNA). The DNA is first "copied" by creating a messenger RNA (mRNA) from one side of the DNA strand. The mRNA also consists of the same nucleic acids, but having uracil in place of thymine. As it copies only from one strand of the DNA, this makes the mRNA single stranded. The mRNA contains genetic information that is derived directly from a specific part of the DNA, which serves as a template for protein production. Specifically, every sequence of 3 consecutive nucleic acids on the mRNA codes for a very specific amino acid. This is an important fact that will become apparent in the later stage of protein synthesis. As the DNA is found only in the nucleus of a cell, and protein synthesis can only happen in the main body of a cell outside of this nucleus (called the cytoplasm), the role of the mRNA is to carry the information out of the nucleus and into the cytoplasm.

Having been transferred out of the nucleus and into the cytoplasm, one of many ribosomal RNA (rRNA) binds to a specific end of the mRNA (the specific end is determined by the sequence of nucleic acids). The rRNA here acts as the "factory" that produces the protein based on the sequence of the mRNA. An rRNA is very unspecific, and can bind to different strands of mRNA for producing varying types of proteins.

All transfer RNA (tRNA) found in the cell cytoplasm each carries a single amino acid, whose type is determined by a part of the of the tRNA structure that consist of 3 very specific nucleic acids that is complementary (the base pairing) to the sequence as found on the mRNA. These 3-length nucleic acids are called anticodons. The complementary sequence of an anticodon that is found on the mRNA is called a codon. Figure 2.2 shows the structure of a tRNA and its anticodon region.

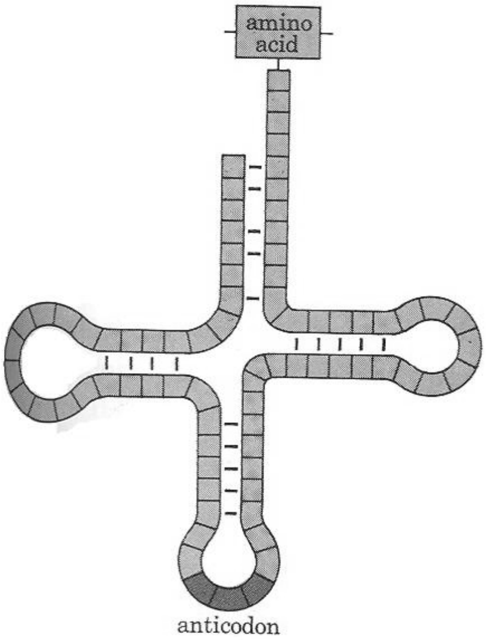

Figure 2.2: The cloverleaf structure of a tRNA with the anticodon labelled. Illustration from Keeton (1972)

As the rRNA moves along the mRNA, it "reads" the codons found on the mRNA. For every codon it "reads", a tRNA with the appropriate anticodon will bind to the rRNA and "release" the amino acid it was carrying. As the rRNA moves on to the next codon, another tRNA (with the appropriate anticodon) will bind to the rRNA and "release" its amino acid which will be attached to the previous amino acid by the rRNA. This process continues until the end of the mRNA has been reached. At the same time, a string of amino acids is built by the rRNA. After reaching the end of the mRNA, the rRNA will detach itself and release the built protein chain. Thus, a protein has been synthesised. Figure 2.3 summarises the protein synthesis process.

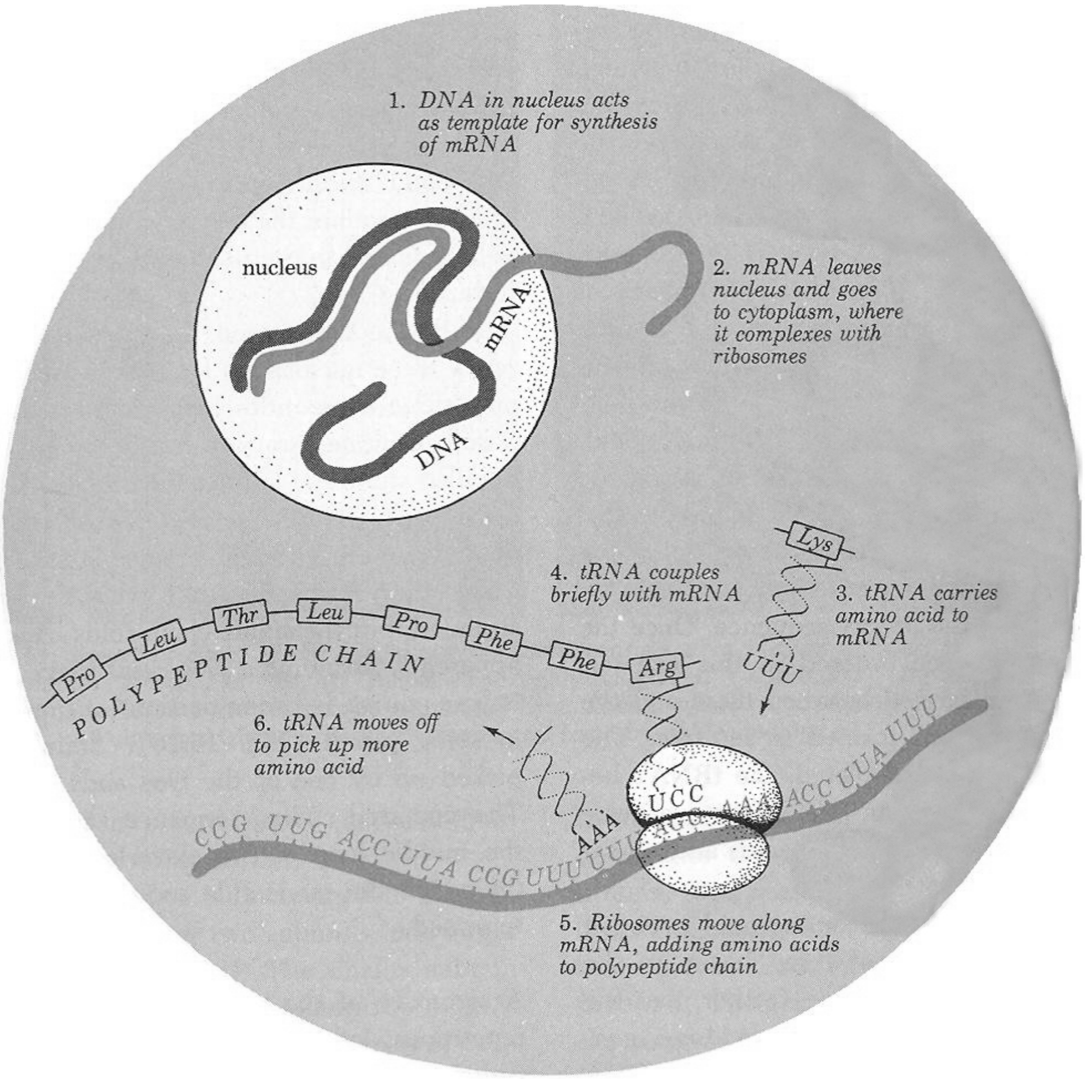
1. DNA in nucleus acts as template for synthesis of mRNA
nucleus
mRNA
DNA
2. mRNA leaves nucleus and goes to cytoplasm, where it complexes with ribosomes
Lys
4. tRNA couples briefly with mRNA
3. tRNA carries amino acid to mRNA
UUU
Pro
Leu
Thr
Leu
Pro
Phe
Phe
Arg
POLYPEPTIDE CHAIN
6. tRNA moves off to pick up more amino acid
AAA
UCC
CCG UUG ACC UUA CCG UUU UUU AGG AAA ACC UUA UUU
5. Ribosomes move along mRNA, adding amino acids to polypeptide chain

Figure 2.3: Protein synthesis process. Illustration from Keeton (1972)

2.2 Microarrays

A microarray is a collection of microscopic DNA spots attached to a solid surface, typically silicon chips, to form an array. Its purpose is to measure the expression levels of many genes simultaneously. Gene expression is a 2 stage process of transcribing a gene's DNA sequence into protein. The first stage is the transcription stage, whereby mRNA is produced from the DNA template. The second stage is the translation stage, where the mRNA is then used to produce a polypeptide, or protein. Essentially, gene expression would mean how much a particular gene was “used” for protein synthesis. A high expression level of a particular gene would mean that a lot of mRNA was produced from that particular gene sequence. This in turn would logically imply that a lot of the associated protein was produced as well (Brown 1998). This mRNA is then spliced to remove the non-essential segments to create a purified mRNA, whereby complementary DNA (cDNA) is then derived from the purified mRNA. Complementary DNA here refers to the DNA strand made from an mRNA. A single strand is first derived from the sequence provided in the mRNA, and subsequently paired with other nucleic acids to form the complete double strand. These cDNAs are subsequently used for microarray analysis. The use of cDNA in genomic experiments is common because cDNAs are copies of genes without any excessive and non-coding segments (such as exons and introns). They are also double-stranded, and are practical clones to a DNA strand. More importantly, as opposed to mRNA, cDNA can be litigated into a cloning vector for replication for practical use in experiments and can be chemically dyed for labelling and subsequent scanning (van Kampen et al 2003, Brown 1998).

As living organisms contain thousands, if not millions, of genes, measuring the expression levels of each and every gene individually in a laboratory would be time consuming, tedious and costly. Although there have been improvements and advancement in measuring gene expressions at a single gene level, such as reverse transcriptase polymerase chain reaction (RT-PCR) and TaqMan (van Kampen et al 2003), genome-wide research is more feasible as it is more effective to measure thousands of genes from a single sample in one experiment. Genome-wide experiments were made possible with technology such as serial analysis of gene expression and microarray, and further advanced with genomic databases accessible online.

There are many types of microarrays, and Piatetsky-Shapiro et al (2003a) describes some examples of microarrays that include short oligonucleotide arrays (such as the GeneChip by Affymetrix), cDNA or spotted arrays (microarray concept by Patrick O. Brown at Stanford, Schena et al, 1995), long oligonucleotide arrays (by Agilent Inkjet) and fibre-optic arrays. Gabig et al (2001) described several microarray techniques in greater detail. The spotted arrays, also known as fragment based DNA printing, features an array of spotted polymerase chain reaction (PCR) products. Arrays of prefabricated oligonucleotides can be made with microelectrodes or gel pads. The method of *in situ* synthesis of oligonucleotides can be done with photolitography (as in the case of Affymetrix GeneChips) or inkjet technology (method by Agilent).

In a microarray experiment, DNA samples are fixed upon a glass slide, with each sample positioned on a known location on the array. The target and reference samples are then labelled with red and green fluorescent dyes respectively, and are then hybridised on the slide. The intensities of mRNA hybridising at each site is then measured using a fluorescent microscope and the image analysed by the microarray reader. Figure 2.4 illustrates this process. This results in a few thousand numbers, each measuring the expression levels of each gene in the experimental sample relative to the reference sample. Positive values mean that there is a higher expression value in the target sample with respect to the reference sample, and vice versa for negative. This process is repeated for every sample and gene, and will result in a gene expression matrix (Aas, 2001). Further detailed microarray processes and information can be found explained in Gabig et al (2001), Nakanishi et al (2001) and Bednar (2000).

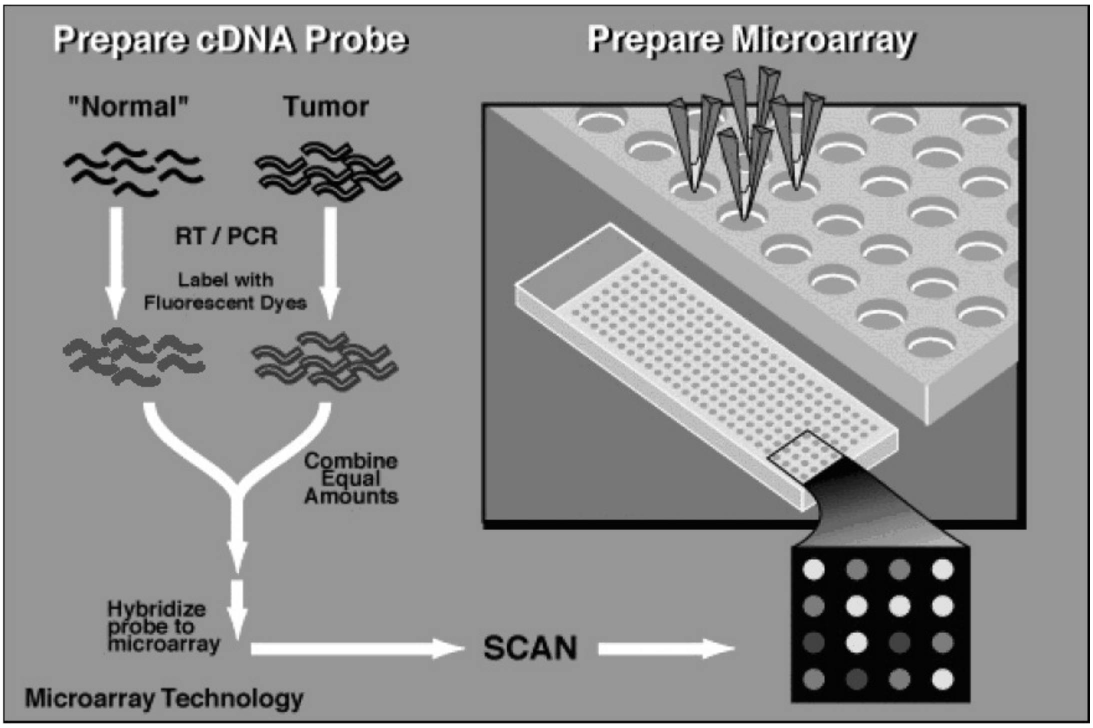

Figure 2.4: The microarray process. Illustration from The National Human Genome Research Institute, National Institutes of Health website

Measuring gene expression using microarrays is relevant to many areas of biology and medicine, such as studying treatments, disease and developmental stages. Microarrays have made possible the collection of molecular information into datasets to represent many biological functions, and thus have expanded the uses of microarray in the field of medicine, of which includes DNA microarray, tissue microarray, protein microarray, plant microarray and so forth. With the rise in microarray technology, genetic databases have also been established with some being open to public access. Some public-access databases include Array Express, Genetic Expression Omnibus, Stanford Microarray Database, University of North Carolina (UNC) Microarray Database, Medical University of South Carolina (MUSC) DNA Microarray Database and so forth. The genes studied in this project involve the microarray for brain tumour genes and breast cancer.

CHAPTER 3

3.1 Self-Organising Maps (SOM)

The idea of SOM was first introduced by Teuvo Kohonen in 1982 (Kohonen 1982a, Kohonen 1982b) and proceeded to expand to a full clustering algorithm in Kohonen (1995), which was then further expanded theoretically and explored by many scholars, such as in M. Cottrell et al (1998).

SOM's rise in popularity stems from its ease of use and efficient clustering algorithm as well as its ability to project the multi-dimensional data onto a 2-dimensional plane. It has been proven to be superior to hierarchical clustering methods by Nikkila et al (2002) due to its more efficient visualisation method as compared the hierarchical clustering that will result in an unwieldy dendrogram of thousands of nodes, and perhaps in hundreds of treatments. SOM, however, can visualise the entire data in a plane that eases analysis and interpretation, along with the individual component planes to further discover relationships among clusters. Furthermore, Toronen et al (1999) demonstrated that SOM is a computationally faster and more reliable method than Sammon's mapping, especially in the case of gene expression data.

The popularity of SOM, especially in the field of microarray data mining, continues to expand with research in the genomic field. This includes yeast (Torkkola et al 2001), breast tumours, leukemia, prostate cancers, esophageal cancers, budding yeast and system tumours. Gene families with similar gene expression patterns were easily identified using SOM, in addition to rapidly finding similar patterns when comparing their results with a previous work (Nikkila et al, 2002). This was due in part to SOM's ability to learn unsupervised, and subsequently visualise the high dimensional data onto a lower 2-dimensional plane.

SOM has its share of novel hybridisations and variations as well. As with all other forms of clustering, variations were explored to suit current research, such as in Yin (2002), Sultan et al (2002), Martin-Merino et al (2004), Wu et al (2005), and Kusumoto et al (2006). Yin (2002) sought to improve the visualisation and topology of SOM by controlling the resolution of the map. The variation of SOM used by Sultan et al (2002) incorporates SOM with binary tree-structured vector quantisation, which combines SOM and partitive k-means clustering. This resembles SOM visualising a dataset, and subsequently performing k-means clustering upon the many component planes generated during SOM training. Another variation, used by Manuel Martin-Merino et al (2004) was an asymmetric version of SOM, which was found to produce excellent results and was the best visualisation method among other techniques examined in the paper, such as multidimensional scaling and Sammon mapping. Wu et al (2005) proposed a hybridised SOM and multidimensional scaling method. Kusumoto et al (2006), on the other hand, proposed a variation of SOM that is without the neighbourhood function that was used to define SOM's self-organising ablilities.

However, many have used SOM on its own as it proves to be a powerful enough visualisation tool, such as Nikkila et al (2002) and Xiao et al (2003). For Xiao et al (2003), the ability for SOM to generate component planes was fully utilised, allowing for visual inspection of all biological significance of the genes clustered by SOM. Huang et al (2003) on the other hand, used SOM as a pre-processing tool. SOM was used to cluster and visualise a yeast cell cycle to extract patterns within the clusters trained by SOM, and subsequently a three-layer artificial neural network was used to find relationships among the clusters as found by SOM.

To this day, SOM continues to be used as an effective data mining and clustering tool in various fields, such as in de Almeida (2007), Adiletta (2008), Hu et al (2008) Zhang et al (2009), and Das et al (2009). One of the strengths of SOM is that is requires very few parameters, more so parameters that are specific to individual datasets. Unlike other clustering methods such as k-means, fuzzy c-means and hierarchical, SOM does not require various parameters such as the number of clusters or generations prior to processing. This makes SOM a robust clustering algorithm that can be applied to a varied range of datasets and situations as well as being hybridised and modified for use in specific cases.

3.2 Basic concept of SOM

The basic idea of SOM is that it takes high-dimensional data, organises the data (hence the self-organising feature) and projects it onto a lower-dimensional grid (usually 2-dimensional). The projection is the output in the form of a U-Matrix (unified distance matrix) where the relative distances of the data have been preserved during the course of the training and thus, the visualisation represents the distribution of the map units. Interpretation is further aided by the use of component planes, which are an extension of SOM, where every attribute (genes in the case of microarray data) is visualised on its own projection.

The first step in SOM is defining an output space, which is usually a 2-dimensional map grid that consists of map units, and every map unit is then associated with a reference vector. Together, all of the reference vectors form the codebook of the map. These reference vectors are initialised before training. The iterative training begins with the selection of a best-matching unit (BMU). The associated reference vectors of the BMU are then updated based on the updating function of SOM. During the updating of the BMU, the topological neighbours are updated as well, as determined by the neighbourhood function. Figure 3.1 illustrates this manner of updating the weight vectors, with the solid lines representing the grid before updating, and the dashed lines after.

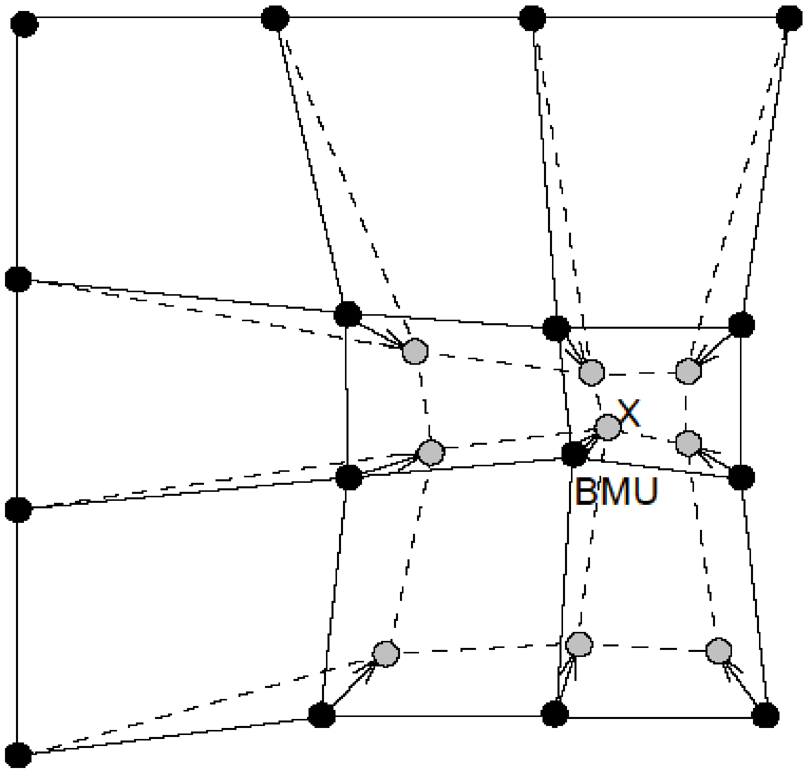

Figure 3.1: Map units being "pulled" towards the BMU. Black dots represent data points "before" and grey dots represent data points "after". BMU represented by X (Figure from Vesanto et al, 1999)

From figure 3.1, it can be seen that the map units nearest to the BMU are pulled closer, while those further away are less affected, and those even further away are unchanged. The map units affected are said to be "in the neighbourhood" of the

BMU, which is defined by the neighbourhood function used during the updating of the vectors. As the BMU randomly changes with each iteration, the “pulling” effect of the updating process will continue until map units have been ordered on the grid with other closely correlated map units while map units with less or no correlation are segregated even further, i.e. self-organising the data. The idea is that through this training process, map units with correlations will be pulled together, while those with little or no correlation will remain distant from each other. During the training phase, the structure and number of map units remain unchanged so as to represent the data in its entirety. Rather, the map (output) “folds and stretches” during the training to an approximation of the density of the data (input), which is where the neighbourhood function plays its role. This updating process repeats for a fixed training length or until the reference vectors in the codebook has changed less than a predetermined error level ε. The result of the clustering will be projected onto a U-Matrix that will allow for visual cluster analysis.

The resulting processed dataset is then subjected to error measures, namely quantisation error (mapping precision) and topographic error (topological preservation). Quantisation error measures the accuracy of the mapping, and how well the map units match the input data. It is essentially the mean of the difference in the distance of the input vector with the corresponding BMUs. Topographic error measures how much of the topology of the input data has been preserved during training, as well as whether twists have been formed on the map. For both of these error measurements, a lower value indicates a better trained map.

3.3 Visualisations of SOM

To illustrate a visualisation of SOM, a publicly available leukaemia microarray dataset was used for the purpose of providing an example. The test dataset was cleaned and subsequently processed using normal batch SOM. The U-matrix and component planes were then visualised. For the component planes, only 10 were visualised for the sake of this example.

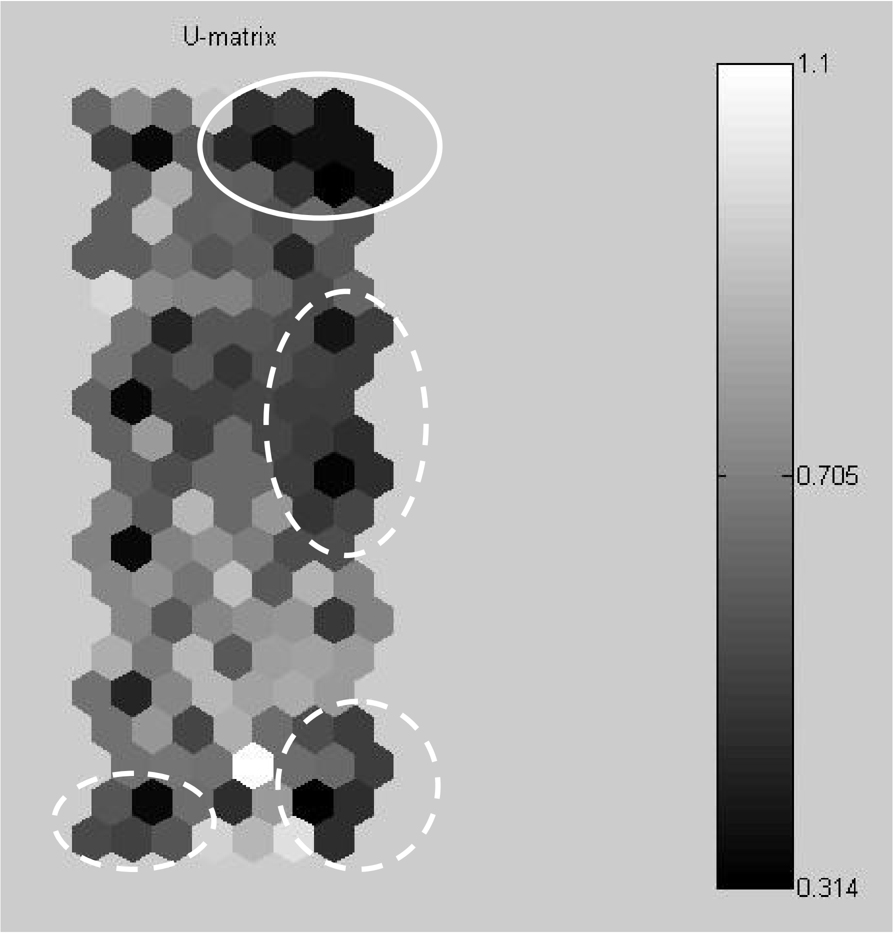

Figure 3.2: Sample SOM U-matrix

It must first be noted that the interpretation of a U-matrix is very subjective. The U-matrix is the projection of the codebook, with map units representing the reference vectors and their location on the map with the colour representing the distance with the neighbouring map units. As observed in figure 3.2, from the colour bar on the side, darker hues signify a shorter distance, while lighter hues signify a larger distance. Thus, an area consisting of several dark-coloured map units will mean that they are close to each other, and are thus collectively a cluster. Conversely, light coloured areas signify map units that are far from each other, and thus have little to no correlation with each other.

As visualisations are subjective and distance scales are different with every U-matrix, rules to determine clusters in this work will be based on colour hue as the colour scale is consistent with every U-matrix. For a cluster to be considered as definite, the colour hue would have to be within the bottom third of the colour bar as seen on the side of the U-matrix. The circled areas show visible clusters, indicated by the low distance between several neighbouring map units. However, as it is a visualisation, its interpretation is entirely subjective. The solid circle can be considered to be a definite cluster, while the dashed circles indicate possible clusters. An individual may deem the possible clusters as not having a close enough distance between map units, while it may be sufficiently close to another individual.

Additionally, component planes complement the SOM visualisation of data. Figure 3.3 shows a sample component planes illustration. Every attribute (genes in the case of microarrays) is mapped out on planes similar to the U-Matrix, but rather than representing the distances between map units, the component planes represent the attribute values and are linked to the U-Matrix by the relative location on the plane. Furthermore, correlations between attributes can be determined by the use of the component planes. The outlook and distribution of relative values indicate how closely related the various attributes are. The more similar the outlook, the more correlated the attributes.

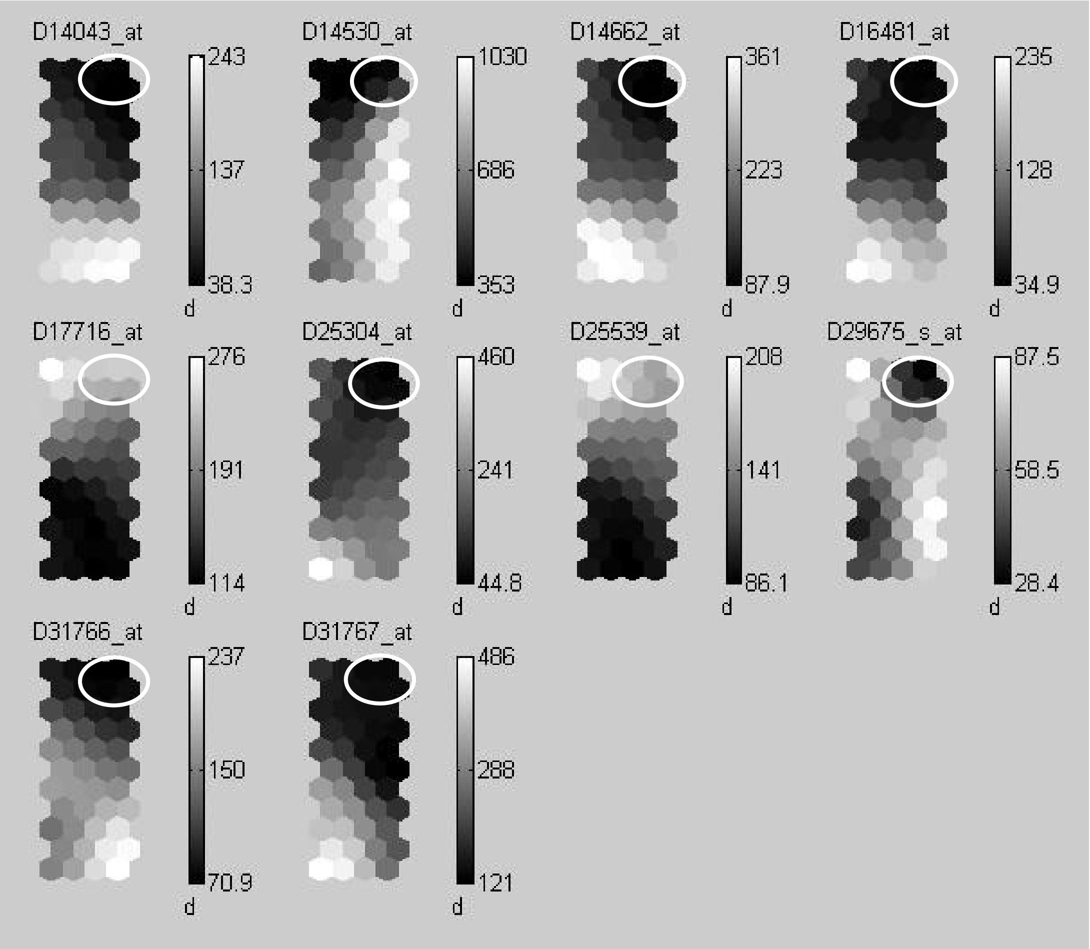

Figure 3.3: Sample SOM component planes

In the component planes, each projection represents a single attribute, which are genes in the case of microarray data. The colour scale here does not represent distance between map units, as in the case of the U-matrix. Rather, the colour scale, and subsequently the associated values, represents actual attribute values. In the same example illustrated above, the first component plane, labelled "D14043_at", is minimally expressed for the cluster shown in the U-matrix. This is observation is made by noting the location of the cluster in the U-matrix (solid white circled area in figure 3.2) and comparing the relative location on every component plane. The genes D16481_at and D25304_at have similar expression values for the same cluster. The genes D14662_at, D31766_at and D31767_at, however, have a similar dark area at the same cluster location, but do not have similar expression values as

D16481_at and D25304_at. These genes, however, are correlated in the sense that they are all expressed at their lowest respective levels for the cluster. Precise absolute expression levels (attribute values) cannot be determined from here, but an estimate based on the colour scale be made.

Visual inspection of the component planes may produce another observation. As mentioned, component planes with similar outlooks are closely correlated to each other. The genes labelled D14043_at, D14662_at and 16481_at have very similar looking component planes, with only mild differences. It is the same for the genes D17716_at and D25539_at. It can be concluded that these 2 groups of genes have some correlation with their respective group members. In the case of genetics, this could mean that the 2 genes are expressed at the same time, although at different levels. The extent of the correlation is, again, subjective to every individual, but the principle remains the same.

The role of the component planes have shown to be very important in analysis using SOM. While it does not allow for determining precise values, it is invaluable when visualising a massive dataset to clearly see a "big picture". Together with the U-matrix projection, clusters and the behaviour of the associating attributes can be easily observed visually.

CHAPTER 4

4.1 Data Preparation

Data preparation is an important part in microarray data mining as data mining software needs to be able to read the data and subsequently interpret it. In its raw state, microarray data can be quite incomprehensible as different microarray equipment will provide different output formats, notably home grown microarray experiments by some labs. Thus, raw datasets have to be formatted for use in analysis. Although there are standardised microarray formats such as Minimum Information About Microarray Experiment (MIAME) and other formats as established at the Microarray and Gene Expression Data Society (MGED), these standards apply only to the laboratory microarray experiments and the necessary supplementary information of outputs, and not the analysis of such data.

Data preparation is a multi phase process. The methods used for preparing data varies with different datasets, applications and tools as each uses different algorithms and have different requirements. Piatetsky-Shapiro et al (2003b) described the most important steps necessary in data preparation for the purpose of data analysis, regardless of algorithm and application. Among them are removing unnecessary fields, thresholding, formatting and feature selection. It is sometimes known as filtering, or in the case of microarray, gene reduction. But all its variations, such as gene filtering, feature reduction and data trimming, all mean the same thing. Normalisation is another step in the preparation process, but the inclusion of this step is largely dependent on the dataset. Large datasets such as microarray usually require normalisation. Subsequently, as each data analysis tool,