Using sequence DNA chips data to Mining and Diagnosing Cancer Patients

Zakaria Suliman Zubi[1], Marim Aboajela Emsaed[2]
[1]Sirte University, Faculty of Science, Computer Science Department
Sirte, P.O Box 727, Libya,
{zszubi@yahoo.com}
[2]Alfateh University, Faculty of Science, Computer Science Department, Tripoli, Libya, P,O Box 13210,
{meemee_02@yahoo.com}

Abstract: Deoxyribonucleic acid (DNA) micro-arrays present a powerful means of observing thousands of gene terms levels at the same time. They consist of high dimensional datasets, which challenge conventional clustering methods. The data's high dimensionality calls for Self Organizing Maps (SOMs) to cluster DNA micro-array data. The DNA micro-array dataset are stored in huge biological databases for several purposes [1]. The proposed methods are based on the idea of selecting a gene subset to distinguish all classes, it will be more effective to solve a multi-class problem, and we will propose a genetic programming (GP) based approach to analyze multi-class micro-array datasets. This biological dataset will be derived from multiple biological databases. The procedure responsible for extracting datasets called DNA-Aggregator. We will design a biological aggregator, which aggregates various datasets via DNA micro-array community-developed ontology based upon the concept of semantic Web for integrating and exchanging biological data. Our aggregator is composed of modules that retrieve the data from various biological databases. It will also enable queries by other applications to recognize the genes. The genes will be categorized in groups based on a classification method, which collects similar expression patterns. Using a clustering method such as k-mean is required either to discover the groups of similar objects from the biological database to characterize the underlying data distribution.

Key-Words: DNA micro-array, Data Mining, Sequence Mining, Biological Database, Genetic Programming, Clustering, Classification, K-means.

1 Introduction

Data mining techniques used to make predictions and typically using only recent static data. Sequence mining is a special case of structured data mining and concerned with finding statistically relevant patterns between data examples where the values delivered in a sequence. These values delivered and then stored in huge collections of data; Examples of such collections include transaction databases were the DNA sequence databases and web site usage logs are available. The availability of these collections has produced great interest in the problem of extracting useful knowledge from the data. However, these data is a sequential data in nature requires a technique for discovering sequential patterns; this technique could be sequence-mining technique. The principle of sequence mining is to discover useful sequential knowledge. Knowledge obtains the form of insight into the structure of the data, "since it is structure that makes things predictable, and it is predictability that can be exploited" [11].

DNA (gene) is an extraordinary chip data with thousands of attributes which represents the gene expression values [14]. Cancers caused through gene mutations and other types of chromosomal or molecular abnormalities. The rare hereditary cancer predisposition syndromes have given much interest in recent years, because genes found that account for the marked preponderance of exacting neoplasm's in such families. Individuals with hereditary cancer predisposition display germline mutations in such genes in their constitutional DNA. The frequent sporadic cancers, i.e. cancers in individuals with a

negative family history for cancer, carry somatic gene mutations acquired at mitosis. Genes caught up with cancers are mainly those involved in normal homeostasis of cellular proliferation, differentiation and death. Cancer growth usually requires that some different gene mutations accumulate in a cell of origin and in its sub clones during colonial evolution of malignant growth. Gene mutations in cancers invariably lead to alterations of gene expression patterns with respect to normal cellular counterparts, including the mutated genes themselves and their downstream targets [9]. The selection of genes involves some sort of 'educated guessing'. For example, a gene might be attractive because it is known to be involved in cellular differentiation, or because it is placed in a genomic area which is targeted by chromosomal aberrations in a precise tumour type [9]. As the Human Genome Project now locates thousands of genes and their sequences, a wealth of genetic information has become available for a probable diagnostic use. Many molecular methods may be too cumbersome to evaluate all related molecular markers in a tumour biopsy. New techniques may help to overcome this limitation indicated in [9] called Genetic programming (GP).

Genetic programming (GP) based is an essential method for both feature selection and generating simple models based on a few genes demonstrated on cancer data. An alternative in the analysis of complex multi-class micro-array datasets is the development of micro-array technology, it is possible to diagnose and classify some particular cancers directly based on DNA micro-array datasets. Genetic programming (GP) has been widely applied with classification problems because it can discover underlying data relationships. GP is a promising solution for the discovery of potentially important gene by generating comprehensible rules for classification. GP based methods have been successfully applied to analyze two-class micro-array since traditional GP is represented in a tree form, which can produce a 'yes/no' answer for a classification problem.

In this proposal, we will focus on the analysis of multi-class micro-array datasets, whereas, dealing with a multi-class problem.

1.1 Sequence mining definition

Sequence mining is concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus Time series mining is closely related, but usually considered a different activity. Sequence mining is a special case of structured data mining [18].

There are two different kinds of sequence mining: string mining and itemset mining. String mining is widely used in biology, to examine gene and protein sequences, and is primarily concerned with sequences with a single member at each position. There exist a variety of prominent algorithms to perform alignment of a query sequence with those existing in databases. The kind of alignment could either involve matching a query with one subject e.g. BLAST or matching multiple query sets with each other e.g. ClustalW. Itemset mining is used more often in marketing and CRM applications, and is concerned with multiple-symbols at each position. Itemset mining is also a popular approach to text mining [20, 21]

The purpose of sequence mining is to discover useful sequential knowledge. Knowledge takes the form of insight into the structure of the data. In particular, much of this data is sequential in nature, so there is a need for techniques for exploring sequential patterns - what might be called sequence mining. Since it is structure that makes things predictable, and it is predictability that can be exploited. [19].

Frequent Sequence Mining is used to discover a set of patterns shared among objects which have between them a specific order.

1.2 Data Sequence

We are given a database D of sequences called data-sequences. Each data-sequence consists of the list of transactions, ordered by increasing transaction-time. A transaction has the following fields: sequence-id, transaction-id, transaction-time, and the items present in the transaction. We assume that the set of items $I = \{i_1, i_2,...i_m\}$, is the set of literals

that can be sorted in lexicographical order. The items in the transaction are sorted in lexicographical order [17].

An itemset i is a non-empty set of items, denoted by (i_1, i_2,...i_m), where *ij* is an item. The support of an itemset is defined as the fraction of the total transactions that contain this itemset. An itemset is said to frequent if its support is above a certain user-specified minimum threshold. Given a database of D of transactions, the problem of mining for association is to find the all frequent itemsets among all transactions.

A sequence is an ordered list of itemsets, denoted by < S_1, S_2 ...S_n >, where S_i is an itemset. The support of a sequence is defined as the fraction of total data-sequences that contain this sequence. A sequence is said to be frequent if its support is above a certain user-specified minimum threshold. Given a database D of data-sequences, the problem of mining for sequential patterns is to find the all frequent sequences among all data-sequences. Each such frequent sequence represents a sequential pattern. It is important to note that from now on the term sequential is adjective of pattern, while term serial is adjective of algorithm [16].

Application domain where Frequent Sequence Mining may be used is Web click log analysis in Information Retrieval systems, in which case the system performance may be refined by analyzing the sequence of interactions that the user exposed while searching or browsing for a specific information. This kind of usage becomes specially clear when we consider the huge amount of data obtained by industrial search engines in the form of query logs [15].

In biology Frequent Sequence Mining may be used to extract information hidden in DNA sequences. Sequence databases in biology are often huge and complex due to variations from genetic mutations and evolution. For example, Frequent Sequence Mining can be used to extract patterns which may be determinant to the development of genetic conditions [20,21].

2 Structure of the DNA

DNA is a one-dimensional fragment, made of two paired strands, coiled around each other as a double helix and held together by hydrogen bonds that connect a linear sequence of complementary pairs of bases. There are four types of bases, referred as C, G, A, T; the bound pairs are G–C (three hydrogen bonds) and A–T (two). DNA inhabits in the nucleus of (eukaryotic) cells.

A gene is a part of DNA, which includes the formula for the chemical composition of one exacting protein. The genome holds the collection of all the genes that code for the entire proteins that an organism wants and produces. A gene expressed in a cell when the protein it codes for a truly synthesized. The human genome has between 20 000 and 30 000 genes.

Moreover, each cell contains the whole genome, dissimilar genes expressed in different cell types.

Transcription and translation: Synthesis of proteins placed at the ribosomes, huge complexes that inhabit in the cytoplasm; the information determined on the DNA relocated from the nucleus to the ribosomes by a molecule called messenger RNA (mRNA). When a gene expressed, specific copies of the information written on one of the DNA strands made, in the form of linear mRNA molecules that leave the nucleus and diffuse through the cytoplasm. The process of copying DNA onto mRNA called transcription.

Consequently, a ribosome reads the message from the mRNA and translates it into a sequence of amino acids, added one at a time, comprising the corresponding Protein. When many copies of a certain protein needed, the cell generates many copies of corresponding mRNA molecules, which are “read” by several ribosome's' [4].

The information one can achieve from sequencing a gene:

1. The sequence of the protein it encodes;
2. Learning the function of the gene;
3. Seeking for the presence of mutations;
4. Comparing the gene sequence with the protein it encodes in different animal species;
5. Studying the evolution of genes [8].

2.1 DNA Micro-arrays overview

The DNA micro-arrays produced by placing small drops of liquid include genes on a glass microscope slide, and allowing the spots to dry. Each spot of liquid contains numerous copies of a single gene and the characteristics of each spot's of gene are known in figure 1, shows how the placing drops of liquid in a precise grid pattern, or array.

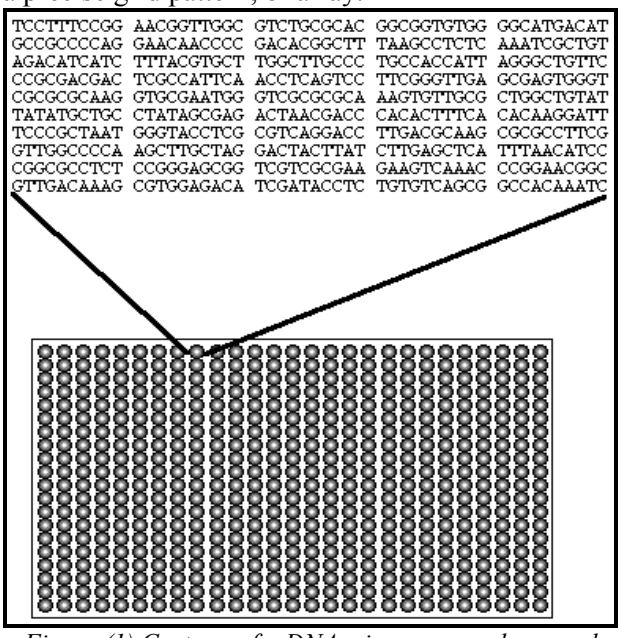

Figure (1) Cartoon of a DNA micro-array where each circle represents a spot on the micro-array and each spot contains a different DNA sequence. One such a sequence is shown in the above the micro-array.

There are many ways to used DNA micro-arrays, but the most ordinary ways viewed in figures 3 and 4. In figure 3, we see two populations of yeast grown under dissimilar conditions. The essential conception is that a genome must react to the environment in which it has placed in and therefore regulate its genes consequently. DNA micro-arrays tolerate us to measure the gene movement of every gene in an organism's genome. The mRNA is isolated from each population and each population of mRNA converted into colored cDNA usually in red and green.

Once the two populations of cDNA's produced, they will be mixed and incubated with the DNA micro-array and unbound cDNA is washed off, figure 3 shows the incubate process. The spotted DNA on the micro-array has been denatured so it is single stranded. The quantity of the colored cDNA that ties to its complementary single-stranded DNA is relative to the activity of the gene. If a gene A produced 10 mRNAs and gene B produced two mRNAs, then we would expect spot A to be 5 times brighter than spot B. The DNA micro-array scanned to discover the two colours of cDNA and then the green and the red images will be stored. Software merges the two colours and spots bound by both colours of cDNA appear yellow shown in figures 2 and 3.

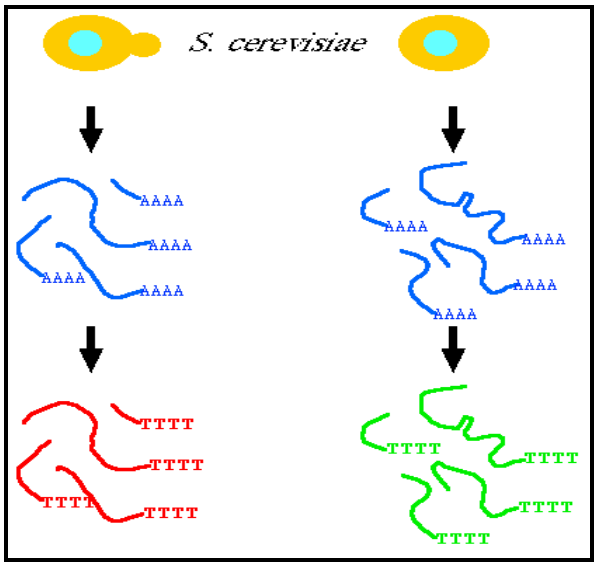

Figure (2) Shows the method for producing labeled cDNA from two populations of cells. The colored cDNAs will be used to probe the DNA micro-array to determine which genes were activated in each growth condition.

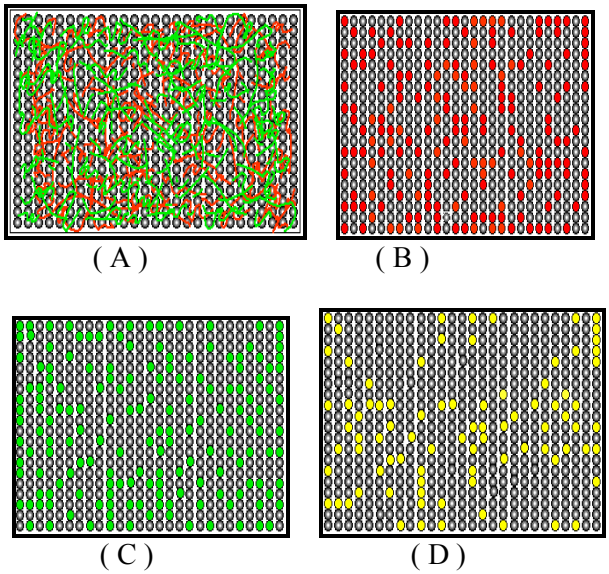

Figure (3) Colored cDNA is incubated with the micro-array (panel A) and complementary sequences are allowed to bind. Unbound cDNA washed off and then the micro-array scanned for red cDNA (panel B) and green cDNA (panel C). Spots bound by both colors of cDNA appear yellow on the computer screen used to visualize the micro-array data.

We indicate some real data in figure 4 using an

application program to analyze the data. The first gene shows orange because it was transcribed under both growth conditions but more strongly so in the red growth condition. The middle gene appears pure yellow because it was equally transcribed in both growth conditions. The third gene appears lime yellow because it was transcribed too in both growth conditions, but more powerfully so in the green growth condition. This type of comparison lets us to make comparative measurements of gene activity between two growth conditions. If the red condition is experimental and the green is control, then gene one was encouraged by the experimental conditions, gene two was unaffected but gene three was reserved by the experimental growth condition.

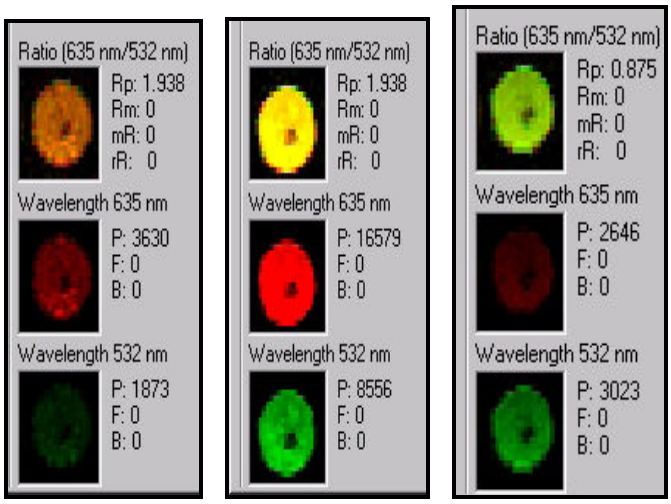

Figure (4) illustrates the real micro-array data for three genes.

Conversion of Raw Data to Numerical Data

The computer application program measures the amount of red and a green light produce from each spot and generates a number for each spot in the table. Also included in the table are the spots location (first three columns), the light strength for each channel (red and green). The ratio is the last product from the procedure with which the remaining attempt of analysis. Characteristically, we want to see how cells approve over time so a time course experiment achieved where mRNA samples collected every two hours from zero to ten hours in alike previous table denoted as table x.

Table x shows all the data but the genes are listed in alphabetical order. It is very hard to deduce these data, so cluster genes, which mean genes with similar patterns of ratios, located next to each other in a table y. Large tables such as these (yeast has about 6000 genes) are difficult to read and understand. Therefore, a new color-coded are required [2].

Analyses of cancers with DNA micro-arrays

Most established molecular diagnostic techniques are insufficient to permit the comprehensive screening of a tumour biopsy sample for all possible types of genetic markers.

Hitherto, standard molecular diagnostics such as the Southern blot have depended on tagged precise DNA probes complementary to the sequences of interest in a sample. PCR based on the exact annealing of nucleic acid sequences (known as primers), to the left and to the right of a gene sequence of interest, which thus enable specific amplification of a defined stretch of DNA or RNA. Technological advances have now permitted these standard molecular detection methods to be miniaturised. DNA micro-arrays are also known as 'chips', biochips, or gene arrays, not to be confused with the tissue arrays. DNA micro-arrays typically consist of rows and rows of oligonucleotide sequence strands, or cDNA sequences immobilised and lined up in dots on a silicon chip or glass slide shown in figer5. Arrays can accommodate up to 20 000 precise sequences on a single chip, either chosen randomly, or deliberately 'biased' to characterize collections of genes typically expressed in a cell type of interest, for example, lymphoid B-cells. With further advances of the technology, it is likely that single chips will contain comprehensive human cDNA .

The major application of microchips falls into three categories:

1. **Gene expression profiling,** while RNA is extracted from tumour samples and hybridised to the micro-array to assess concurrently and in a single experiment the term of thousands of genes within the sample.

2. **Genotyping,** Genomic DNA from an individual tested for hundreds or thousands of genetic markers [notably single nucleotide polymorphisms (SNPs) or 'snips', or micro-satellite markers] in a single hybridisation. This will yield a genetic fingerprint, which in turn may be linked to the risk of developing single gene disorders or particular common complex

diseases.

3. **DNA sequencing,** Sequence variations of specific genes can be monitored in a test DNA sample, thereby greatly increasing the scope for precise molecular diagnosis in single gene disorders or complex genetic diseases.

In cancers, the diagnostic material regularly consists of RNA samples extracted from tumours of interest, which are branded for hybridisation on chips to study large-scale gene expression profiles. The use of RNA implies that freshly frozen intact tumour tissue must be used.

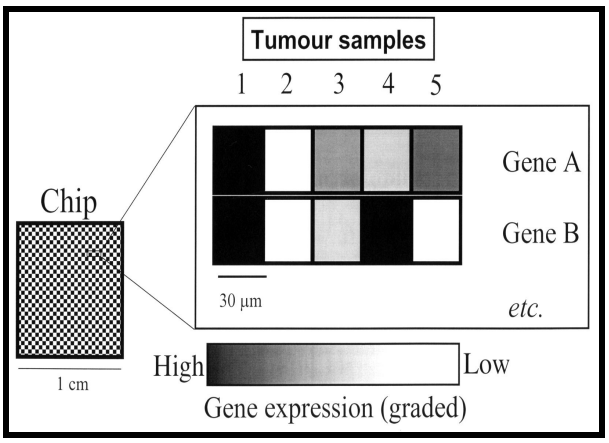

Figure (5) Represents sequences immobilised and lined up in dots on a silicon chip or glass.

The principle of cDNA micro-array gene expression analysis in tumours depends on Schematic representation of a DNA microchip. The chip consists of a silicones or glass surface (an area of about 1 cm2) where up to 10 cDNAs or 250 oligonucleotide gene sequences are schemed in an orderly fashion. In the two-colour hybridisation scheme regularly employed, labelled RNA from tumour samples 1–4 is hybridised to the chip concurrently and in direct competition with labelled RNA from defined control samples. Different fluorescent dyes used for sample and control RNA (shown here in black and white). The relative difference in gene expression between tumour and normal cells can quantified during image analysis of the chip, and assessed as comparative amounts of the two different fluorochrome signals arising from each defined quadrant of the chip. Tumours 1 and 2 show identical terms patterns of genes A and B, and may thus grouped together by virtue of a clustered gene expression profile. In contrast, tumours 3–5 carry distinct molecular signatures with respect to the expression of these two genes [9].

Many protocols foresee a comparative and competitive hybridisation of tumour RNA samples on the chip against normal or reference RNA characterized with different colours. A more conventional molecular diagnostic method, are the enrichment of tumour cells is important. This can be realized by 'virtual dissection' of chip data in a computer application program where gene groups may be clustered and filtered out of which are known to be derived from normal cells or from inflammatory infiltrates in a tumour. Laser capture micro dissection to isolate tumour cells mechanically may, however, still be compulsory, although more laborious.

The 'Lymphochip' is a cDNA micro-array collecting genes preferentially expressed in lymphoid cells. Lymphoid malignancies considered with this chip demonstrate an orderly picture of gene expression patterns, reflecting both B- or T-cell lineage characteristics, stage of maturation of lymphoid cells and proliferation signatures. Diffuse large B-cell lymphoma (a clinically heterogeneous group of lymphomas despite morphological similarity) can be split into subtypes show signs of gene expression profiles either typical of germinal centre B-cells, or activated B-cells, perhaps with implications for prognosis [9].

DNA Sequencing Process:

1. Mapping
   - *Identity set of clones that span region of genome to sequence.*
2. Library Creation
   - *Make sets of smaller clones from mapped clones.*
3. Template Preparation
   - *Purify DNA from smaller clones*
   - *Set up and perform Sequencing chemistries*
4. Gel Electrophoresis
   - *Determine sequences from smaller clones*
5. Pre-finishing and Finishing
   - *Specialty techniques to produce high quality sequences*
6. Data Editing/ Annotation
   - Quality assurance
   - Verification
   - Biological annotation
   - Submission to public database [16].

Applications of DNA micro-arrays or 'chips' in oncology

- Global understanding of abnormal gene expression contributing to malignancy, i.e. snapshots of genes either up or down regulated in tumours.
- Molecular classification of neoplasm's by gene expression signatures, forecasting the tissue of origin of a tumour in the context of multiple cancer classes.
- Classification of novel molecular-based subclasses in the tumours with clinical relevance.
- Discovery of new prognostic or predictive indicators and biomarkers of therapeutic response;
- Identification and validation of new molecular targets for drug development;
- Prediction of drug side effects during preclinical development and toxicology studies;
- Identification of genes conferring drug resistance;
- Prediction or selection of patients most likely to benefit from, or suffer from particular side effects of drugs (pharmacogenomics) [9].

3 The objectives of the paper

Our paper proposal aims to achieve the following:

1. Use DNA micro-array technology in comparing gene expression profiling between tumour cells or tissues and corresponding normal cells or tissues in humans, and for classification, prediction of prognoses.
2. Extracting useful knowledge from DNA micro-array dataset that may help patients discover there early sickness "cancer" before it is too late.
3. Appling methods in gene appearance data analysis using gene selection and an appropriate GP.

4 Sequence Mining in DNA chips data

Task of Mining DNA Chip data it focuses on the use of DNA micro-array to prediction and diagnosis of cancer, so that it expectedly helps us to exactly predict and diagnose cancer. To precisely classify cancer we have to select genes related to cancer, and the resultant impact it might have on routine clinical practice. We will use some common methods to help us produce classification rules.

4.1 Biological Dataset

Biological dataset is a data or measurements collected from biological sources, which is stored or exchanged in a digital form. Biological dataset is regularly stored in files or databases. Examples of biological data are DNA base-pair sequences, and population data used in ecology. Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses. They contain information from research areas including genomics, proteomics and micro-array gene expression. Information enclosed in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

There are a number of micro-array datasets from published cancer gene expression, including leukemia cancer dataset, colon cancer dataset, lymphoma dataset, breast cancer dataset, NCI60 dataset, and ovarian cancer dataset. Among them three datasets will be used in this proposal work.

The first and third datasets will involve samples from two variants of the same disease but the second dataset absorb tumour and normal samples of the same tissue.

Leukemia cancer dataset

Leukemia dataset consists of 72 samples: 25 samples of Acute Myeloid Leukemia (AML) and 47 samples of Acute Lymphoblastic Leukemia (ALL). The source of the gene expression measurements was taken form 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples measured using high-density oligonucleotide micro-arrays. The 38 out of 72

samples used as training data and the remaining used as test data in this proposal work. Each sample contains 7129 gene expression levels.

Colon cancer dataset

Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. Even though original data consists of 6000 gene expression levels, 4000 out of 6000 removed based on the confidence in the measured expression levels. The 40 of 62 samples are colon cancer samples and the remaining are normal samples. Each sample was taken from tumors and normal healthy parts of the colons of the same patients and measured using high-density oligonucleotide arrays. The 31 out of 62 samples used as training dataset in this proposal work.

Lymphoma cancer dataset

B cell diffuse large cell lymphoma (B-DLCL) is a heterogeneous group of tumors, based on important variations in morphology, clinical presentation, and response to treatment. Gene expression profiling has revealed two distinct tumor subtypes of B-DLCL: germinal centre B cell-like DLCL and activated B cell-like DLCL. Lymphoma dataset consists of 24 samples of GC B-like and 23 samples of activated B-like. The 22 out of 47 samples used as a training data and the remaining used as test data as well in this proposal.

Analysis of results

The below table shows the IDs of genes overlapped by Pearsons correlation coefficient, cosine coefficient, Euclidean distance in each dataset. Among these genes there are some genes overlapped by other feature selection methods. For example, gene 2288 of Leukemia has been third-ranked in information gain. The number of overlapped genes of leukemia dataset is 17. The number of overlapped genes of colon dataset is 9. The number of overlapped genes of lymphoma dataset is 19. These overlapped genes are very informative. In particular, Zyxin, gene 4847 of leukemia, has reported as informative, but there are no genes appeared regularly in every method.

Table 1 The IDs of genes overlapped by Pearsons correlation coefficient, cosine coefficient, and Euclidean.

Leukemia	461	1249	1745	1834	2020
	2043	2242	2288	3258	3320
	4196	4847	5039	6200	6201
	6373	6803			
Colon	187	619	704	767	1060
	1208	1546	1771	1772	
Lymphoma	36	75	76	77	86
	86	678	680	1636	1637
	2225	2243	2263	2412	2417
	2467	3890	3893	3934	

Table 1, The IDs of genes overlapped

In figure 6, we illustrate the expression level of genes chosen by Pearson's correlation coefficient method in Leukemia dataset. The 1~27 samples are ALL and 28~38 samples are AML. The differences of brightness between AML and ALL correspond to those genes chosen by Pearson's correlation coefficient method divide samples into AML and ALL.

The results of recognition rate on the tested data are as shown in tables 2,3,4. Column is the list of feature selection methods: Pearson's correlation coefficient (PC), Spearman's correlation coefficient (SC), Euclidean distance (ED), cosine coefficient (CC), information gain (IG), mutual information (MI), and signal to noise ratio (SN). KNNPearson and MLP seem to produce the best recognition rate among the classifiers on the average. KNNPearson is better than KNNcosine. SVM is poorer than any other classifiers.

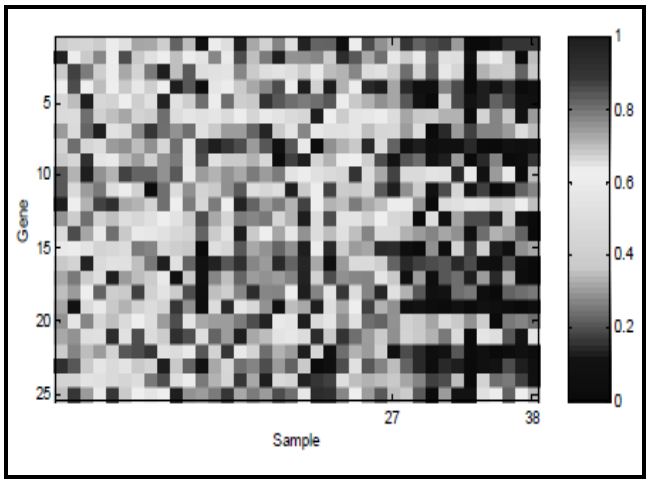

Figure 6 Expression level of genes chosen by rPearson in Leukemia dataset

In table (2) we indicates the recognition rate with features and classifiers in percentages (%) in Leukemia dataset as follow:

	MLP	SASOM	SVM		KNN	
			Linear	RBF	Cosine	Pearson
PC	97.1	76.5	79.4	79.4	97.1	94.1
SC	82.4	61.8	58.8	58.8	76.5	82.4
ED	91.2	73.5	70.6	70.6	85.3	82.4
CC	94.1	88.2	85.3	85.3	91.2	94.1
IG	97.1	91.2	97.1	97.1	94.1	97.1
MI	58.8	58.8	58.8	58.8	73.5	73.5
SN	76.5	67.7	58.8	58.8	73.5	73.5
Mean	85.3	74.0	72.7	72.7	84.5	85.3

Table 2, The recognition rate with features and classifiers in Leukemia.

Table (3) shows the recognition rate with features and classifiers in percentages format (%) in Colon dataset in the below table:

	MLP	SASOM	SVM		KNN	
			Linear	RBF	Cosine	Pearson
PC	74.2	74.2	64.5	64.5	71.0	77.4
SC	58.1	45.2	64.5	64.5	61.3	67.7
ED	67.8	67.6	64.5	64.5	83.9	83.9
CC	83.9	64.5	64.5	64.5	80.7	80.7
IG	71.0	71.0	71.0	71.0	74.2	80.7
MI	71.0	71.0	71.0	71.0	74.2	80.7
SN	64.5	45.2	64.5	64.5	64.5	71.0
Mean	70.1	62.7	66.4	66.4	72.7	77.4

Table 3, The recognition rate with features and classifiers in Colon.

In table (4) the recognition rate with features and classifiers represented in percentages form (%) in Lymphoma dataset are also indicated in the following table as well [12]:

	MLP	SASOM	SVM		KNN	
			Linear	RBF	Cosine	Pearson
PC	64.0	48.0	56.0	60.0	60.0	76.0
SC	60.0	68.0	44.0	44.0	60.0	60.0
ED	56.0	52.0	56.0	56.0	56.0	68.0
CC	68.0	52.0	56.0	56.0	60.0	72.0
IG	92.0	84.0	92.0	92.0	92.0	92.0
MI	72.0	64.0	64.0	64.0	80.0	64.0
SN	76.0	76.0	72.0	76.0	76.0	80.0
Mean	69.7	63.4	62.9	63.4	69.1	73.1

Table 4, the recognition rate with features and classifiers in Lymphoma.

5 Methods and Models

5.1 Ensemble genetic programming

Genetic programming first proposed by Koza (Koza, 1992) with the purpose of automatically generates a program that could solve a given problem. It was initially similar to the genetic algorithm in many ways, but it was different in representation [3]. An individual was represented as a tree composing of functions and terminal symbols. Various functions and terminal symbols developed for the target application, and classification was one of the goals of genetic programming.

In another word, GP is a branch of genetic algorithm (GA), and the main difference between GP and GA is the structure of individuals: GA has string-structured individuals, while GP's individuals are trees. Due to the structure, GP can produce classification rules by formulating important features.

Generally, the terminal set consists of features and constants, and the function set consists of arithmetical or logical functions. The leaf nodes and non-leaf nodes of trees are chosen from the terminal and function sets, correspondingly. Let F be the set of functions, and T be the set of terminals. The trees built in the evolution process are the set of all possible compositions of functions and terminals selected from F and T. When used for the classification task, a tree evolved from a training dataset and validated against an independent test dataset.

5.2 The structure of individuals

An ensemble system has established to be more

accurate and robust than an excellent single classifier in many areas. As the output of an group is based on all trees in the ensemble, when a tree fail to make a distinction a 'hard' sample, other trees in the ensemble still have a ability to correct it. Then the final ensemble can generate a correct output. So instead of concerning a tree to a two-class problem, an ensemble of k trees is installed in this study. For an n-class micro-array dataset, n ensembles are necessary to solve the relevant two-class problems.

Based on this thought, a new individual structure for GP will be proposed, as illustrated in figure 7. In the scheme indicated below in figure 7, an individual is a multi-class classifier and can deal with a multi-class problem directly. In an entity, there are n ensemble systems, which named as sub-ensemble (SE) systems for clarifying their roles. As an individual is composed of n× k trees in all, the size of SEs should not be too large for building a capable and compact classifier. [3].

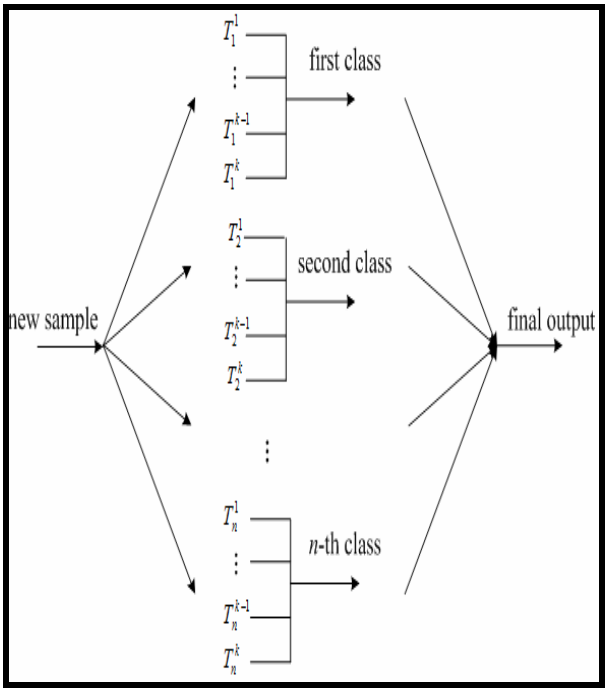

Figure 7, The new individual structure for the GP.

5.3 The initialization of GP

In this process, an equal number of trees initialized for each depth between two and the preliminary maximum tree depth value. For each depth level measured, half of the trees receive non-leaf nodes from F until trees are fully-grown; the other half allowed receiving nodes from both F and T randomly excepting for the root node, which generates a set of a seriously unbalanced trees. This method will results in balanced and unbalanced trees with several different depths. Bloat means that trees keep growing without the corresponding improvements in fitness. It creates the results complex without any advantage, the dynamic maximum tree depth technique used to control it here. When using this technique, there are two important parameters: strict depth limit and dynamic maximum tree depth limit.

The tree depth originally set to be no deeper than the dynamic depth limit. If a tree does not surpass the dynamic maximum depth, it participates in the current population. When a tree is deeper than the dynamic maximum depth but does not exceed the strict maximum depth, it is evaluated by 10-fold CV. In case of nominating the tree as a best tree for the relevant two-class problem currently, the dynamic maximum depth is increased and the new tree allowed joining the population; otherwise, the new tree rejected. If a tree is deeper than the strict maximum depth, it is rejected and one of its parents enters the population. Once the dynamic maximum depth is increased, it would never be lowered again. The original maximum dynamic depth limit is set to be small to force the GP to look for simpler trees firstly before accepting complex solutions. An effective SE consists of accurate and diverse trees. Explicitly, the trees employed in a SE should be of high classification accuracy and avoid making coincident errors.

In this way, the fused outputs can be more accurate than that of the best tree. No gains will be reaching when fusing trees producing the same outputs. Many different diversity measures proposed based on different theories, and most of them based on the difference among the classifier outputs. Here we define a new diversity measure based on the difference in the feature subsets among trees, named as diversity in features (DF) [3].

5.4 The genetic operators

The crossover and mutation operators should be operated on trees instead of individuals or SEs. Ahead of applying the operators, a selection procedure is essential to pick up trees firstly. Comprehensively, to select the trees for the i-th two-class problem, the first step is to select an individual from the current population in likelihood proportional to the individual's fitness value. The next step is to select a tree in a prospect proportional to the tree's CV accuracy rate in the i-th SE of the

selected individual. Both steps based on the theory of roulette, which allows all trees in the current population getting a chance. Two/one trees will selected as parent(s) for crossover/mutation.

After the selection process, the standard crossover and mutation operators set up. In the crossover procedure, two random nodes are chosen from both parents, and then the particular branches are swapped to create two offspring. The mutation operator randomly chooses a node from a parent and replaces it with a new randomly generated sub-tree. In our GP, the lately generated sub-tree only contains the features barred in the parent tree. In this way, more features will evaluate in the evolution process. As the task of the GP is to discover potentially important genes in a huge number of candidates, the mutation operator is significant to improve the exploration in the great search space. For each offspring, k trees will generate in each SE using crossover and mutation operators. However, due to the random mechanism, the lately produced SEs in new offspring may not always be diverse enough. Therefore, the Heuristic Algorithm I is also functional to adjust the tree assignment for the new individuals, just as in the initialization process [3].

5.5 The validation of classifier

A classifier validated in a self-governing test set. The validation of an individual's performance usually takes three steps. Initially, all trees in an individual will check a new sample. The trees classify a sample via the process described above. Secondly, the trees' outputs combined to form the particular SEs' outputs based on the weighted majority vote. At this time, the outputs of SEs indicated by +1/-1. Ultimately, the particular SEs' outputs are fused to generate the results for an individual. If only a SE returns +1, the sample will be assigned to the respective class. The conflicting situation occurs, when, the covering scores of the conflicted SEs compared, and only the output of the SE with the highest score will be chosen as a final decision. Many feature subsets produced in parallel during evolution, a set of globally optimal or at least near optimal trees obtained for each two-class problem. High diversity in SEs can easily achieve, so the GP is promising in achieving predictive classifiers [3].

5.6 Models used

DNA program kit is a commonly known application program it contains three different DNA sub models, each of which illustrates an important characteristic of each DNA sub models. These models indicted as follows:

K'Nex model is a very easy model to use and provides a good opportunity to demonstrate the helical structure of DNA. The K'Nex model has four different colours of connectors that used to represent the four-nucleotide bases, but it is difficult to see the base pairing.

Pop Bead model this model is an excellent example of the base pairing and ladder structure of DNA.

Molecular model illustrates the fine chemical structure of DNA and the molecular interactions involved in forming the alpha helix.

6 Data Classification and Clustering

A variety of methods of data reduction and classification formulated to identify groups of genes that show similar expression patterns.

To present the results of such classification, it is helpful to have an intuitive visual representation. This often accomplished by drawing dendrograms and/or colour-coded representations of similarly expressed genes [10].

6.1 Clustering Workloads

Clustering is the process of discovering the groups of similar objects from a database to characterize the underlying data distribution. It has wide applications pattern recognition, and spatial data analysis these spatial data could be biological datasets [6].

Hierarchical Clustering Algorithms: Most gene-clustering algorithms are hierarchical. These techniques derived from algorithms used to construct phylogenetic trees; the most-similar genes clustered first, while those with more-diverse profiles are subsequently included in a stepwise hierarchy of increasing diversity. This means that, in the first clustering step, the single most-similar expression profiles correlated to form nodes, the most similar of which linked further in the second clustering step, etc, until all nodes finally linked and the complete hierarchical tree of proximities (dendrogram) obtained. Starting from the second clustering step and higher, each node may consist of two or more objects.

The distances between nodes must be recomputed at each step. This will be done, for example, by computing the distance between the nodes as the average distance between its objects, as in the average linkage procedure, or as the distance between two of its closest objects, as in the kth nearest-neighbour linkage procedure (KNN). Other options include distances computed between the centres of mass of clusters or their modifications. In most cases, however, average linkage procedure is considered acceptable.

These different linkage choices made to recompense for potential problems with hierarchical clustering. Namely, as clusters grow in size, at higher levels of hierarchy, the expression vector that represents the cluster may no longer be representative of any of the genes in the cluster. Thus, actual expression patterns of the genes themselves become less relevant on higher levels of hierarchy. If a gene assigned to as "wrong" cluster, this error cannot be corrected later under hierarchical clustering [10].

Non-hierarchical Clustering Algorithms

1. **K-means Clustering**
   Sometimes, when a priori knowledge exists about the number of clusters that should be achieve, one can use non-hierarchical K-means clustering to partition the data. In this procedure, one first indicates the number of clusters (K), and then randomly allocates expression vectors to them. Distances between clusters recomputed, expression vectors reassigned to the nearest cluster, and the procedure iterated until the point reached when no new assignments made. The K-means clustering process simply partitions expression data into K groups and does not generate a dendrogram, while one can constructed later by a hierarchical process [18,19].

2. **Self-organizing Maps**
   An additional frequently used non-hierarchical process is self-organizing maps (SOMs), a neural network-based method for clustering. In this algorithm, one also indicates in advance the number of clusters, selected typically as the nodes of a grid. The nodes mapped into K-dimensional space, primarily at random, and then iteratively adjusted. During each iteration, a data point is randomly chosen and the node is moved towards that point by the amount proportional to its proximity, so that more distant nodes are moved to the least amount. In this way, neighbouring points in the primary geometry mapped to close by points in the data space. This procedure regularly iterated ten thousands of times. SOMs are mainly helpful for examining data analysis, in order to expose the global patterns in the data. [10]

Biomedical Applications Of Classification

The use of classification processes in micro-array testing has been mainly productive in cancer research because cancers are complex, mutagenic diseases with a natural control group for the analysis non-cancerous tissue [11].

We present a purpose of the feed forward neural network (SLFN) trained by the singular value decomposition (SVD) approach for DNA micro-array classification.

Experimental results show that the SVD trained feed forward neural network is simple in both training procedure and network structure; it has low computational complexity and can generate better performance with compact network architecture [7].

7 Some Facts about DNA micro-array

One of the benefits of DNA micro-array technology is that it can evaluate simultaneously the relative expression of thousands of genes by using small amounts of materials, as long as gene signatures for particular disease situations. Additionally, the measures can easily be automated. Furthermore, the ability of measurement of gene expression by DNA micro-array is huge.

On the other hand, the main difficulty of DNA micro-array technology is that it only estimates gene expression at a transcriptional, but not translational, level, as posttranscriptional modifications (such as phosphorylation) often play important roles in the regulation of protein functions. Besides, DNA micro-array technology is still not established enough for decision-making based on the micro-array data sometimes when the data is incomplete [17].

8 Conclusion

In this paper, we proposed a GP based approach

to deal with the gene selection and classification tasks for multi-class micro-array datasets. The multi-class problem will divide it into multiple two-class problems, and a set of sub-ensemble systems deployed to deal with respective two-class problems. The procedure responsible for extracting datasets called DNA-Aggregator. We will design a biological aggregator, which aggregates various datasets via DNA micro-array community-developed ontology based upon the concept of semantic Web for integrating and exchanging biological data. Then by fusing these ensembles, an individual built to deal with a multi-class problem directly. Trees constructed with different genes; important genes selected as important references for clinic diagnosis or cancer development. For each dataset, the biological significance of the selected genes validated from a biological database.

Finally, we hope our GP based method could present useful alternatives in the analysis of complex multi-class micro-array datasets.

References:

[[1] Alfred Ultsch, David Kämpf. " Knowledge Discovery in DNA Microarray Data of Cancer Patients with Emergent Self Organizing Maps" . ESANN'2004 proceedings - European Symposium on Artificial Neural Networks Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 501-506.page 1

[2] A. Malcolm Campbell and Laurie J. Heyer" DNA Microarrays: Background, Interactive Databases, and Hands-on Data Analysis" .page 5

[3] De-Shuang Huang, Kun-Hong Liu and Chun-Gui Xu , Associate Editor: Prof. David Rocke "A Genetic Programming Based Approach to the Classification of Multiclass Microarray Datasets . page 2

[4] Eytan Domany"Analysis of DNA-chip and antigen-chip data: studies of cancer, stem cells and autoimmune diseases". S0010-4655(05)00138-4/FLA AID:2871.page 2

[5] Girish kumar jha "ARTIFICIAL NEURAL NETWORKS" PUSA,New Delhi-110 012 .page 3

[6] Joseph Zambreno Berkin O¨zıs.ıkyılmaz Gokhan Memik Alok Choudhary"Performance Characterization of Data Mining Applications using MineBench". Jayaprakash Pisharath .Architecture Performance and Projections Group .Intel Corporation .Santa Clara, CA 95054 . jayaprakash.pisharath@intel.com.

[7] Hieu Trung Huynh, Jung-Ja Kim and Yonggwan Won" Classification Study on DNA Microarray with Feedforward Neural Network Trained by Singular Value Decomposition" International Journal of Bio- Science and Bio- Technology Vol. 1, No. 1, December, 2009.page 1

[8] Mandana Sassanfar and Graham Walker "DNA Microarray Technology. What is it and how is it useful?" Copyright 2003 MIT Dept of Biology . page 3

[9] 2. M. F. Fey" The impact of chip technology on cancer medicine". DOI: 10.1093/annonc/mdf647.

[10] N.M. SVRAKIC, O. NESIC, M.R.K. DASU, D. HERNDON, AND J.R. PEREZ-POLO"Statistical Approach to DNA Chip Analysis". Page 5

[11] Philip Hingston" Using Finite State Automata for Sequence Mining" .page 1

[12] Sung-Bae Cho and Hong-Hee Won" Machine Learning in DNA Microarray Analysis for Cancer Classification".page 6

[13] Timothy James Stich, Dr. Julie K. Spoerre & Dr. Tomás Velasco" The Application of Artificial Neural Networks to Monitoring and Control of an Induction Hardening Process" The Official Electronic Publication of the National Association of Industrial Technology • www.nait.org © 2000 .page 3

[14] W. B. Langdon and B. F. Buxton" Genetic Programming for Mining DNA Chip data from Cancer Patients" Computer Science, University College, Gower Street, London, WC1E 6BT, UK, fW.Langdon, B.Buxtong@cs.ucl.ac.uk .http://www.cs.ucl.ac.uk/sta_/W.Langdon, /sta_/B.Buxton .page 1

[15]W. van der Aalst, T. Weijters, and L. Maruster.Work°ow mining: Discovering process models from event logs. IEEE Trans. Knowl. Data Eng., 16(9):1128{1142, 2004.

[16] Yunqing YU, Kazuaki MURAKAMI "Reconfigurable Neural Network Using DAP/DNA" 6-1 Kasuga-koen, Kasuga, Fukuoka 816-8580, Japan .page 1

[17] M. J. Zaki. Sequence mining in categorical domains: Incorporating constraints. In Proceedings of CIKM'00.

[18] M. J. Zaki. Spade: An efficient algorithm for mining frequent
sequences. Machine Learning, 42(1/2):31–60, 2001.

[19] Zimback L, Mori ES, de Morase ML, Rosa DD, Furtado EL, Marino, CL, Wilken CF, Velini ED, Guerrini AI, Maia IG, and Camargo LE (2004), "Data mining of Eucalyptus ESTs involved in the mechanism non hormonal growth genes," International Plant & Animal Genomes XII Conference, San Diego, CA, January 10-14, 2004.

[20]http://www.ornl.gov/sci/techresources/Human_ Genome/graphics/DNASeq. Process.pdf . 'Page header: DNA Sequencing Process'. Login clock 11:03pm. Date 16-2-2010. Page 1

[21] http://worldscibooks.com/etextbook/6712/6712_ chap01.pdf. page header 'CHAPTER 1 DNA Microarray Technology'. Login clock 09:47pm.date 21-2-2010. Page 8.