Proceedings of International Conference on Rural Information and Communication Technology 2009

Institut Teknologi Bandung
17-18 June 2009

General Chair's Message

This conference addresses rural information and communication technology (R-ICT) issues with digital culture objectives. We believe that everybody must have a chance to benefit from digital technology. The gap between the haves and the have nots to benefit from digital technology has been known as digital divide. There are approximately 3.6 billion people without benefits from ICT. We need to find solutions to allow digital inclusions for everybody.

On key issue is broadband infrastructure for rural areas. This requires innovations at many levels, from technology, products, applications, operations, and investments. In particular, we are looking at rural next generation network (R-NGN), next generation computing (NCG), next generation application platforms, as well as cultural contents. This conference gathers researchers to report their results pertaining those issues. But more importantly, this conference should consolidate research community to actively dedicate research efforts toward solving digital divide issues.

This conference is inspired by the idea that ICT empowers its users to be productive in creating values despite of their shortcomings. Without clear concept of how the users benefit from ICT, rural ICT will not be sustainable. We strongly believe, in additions to the usual education, health, trade, farming, and information applications, it is the cultural potentials that will eventually make rural ICT sustainable. It is the unleashing of the creative excitement embedded in each culture that will drive the value creation by rural communities. That is the creative entrepreneurial knowledge society that we are visioning.

This year's conference in particular is held within The Golden Anniversary of the Institute Technology Bandung. In conjunction to this conference, ITB holds 4 more international conferences with a common theme "Energy and The Environment: Reinvention for Developing Countries". All these conferences focuses on the issues related to the developing countries. In addition to the technical sessions within these conferences we also hold Millennium Research Agenda (MRA) meeting.

We hope that through these conferences and MRA meeting, we could be able to join our focus in issues of helping the bottom pyramid of the society.

Bandung, 15 June 2009

Dr. Armein Z. R. Langi

Chairman

r-ICT 2009 Committee

General Chair

Armein Z.R. Langi

Conference Chair

Mervin T. Hutabarat

Scientific Committee

Leo P. Ligthart (TU Delft)
Adang S. Ahmad (ITB)
Andriyan B. Suksmono (ITB)
Witold Kinsner (Univ. Manitoba, Canada)
Chun Tung Chou (UNSW, Australia)
Yo-Sung Ho (KIST, Korea)
Dharmendra Sharma (Canberra Univ., Australia)
Isao Nakajima (Tokai University/ITU)
Dwi H. Widyantoro (ITB)
Adi Indrayanto (ITB)
Soegijardjo Soegijoko (ITB)
Tati L. R. Mengko (ITB)
Ary Setijadi P., (ITB)
Bobby Nazif (Univ. Indonesia)
Suhono H. Supangkat (ITB)
Dadang Gunawan (Univ. Indonesia)
Budi Rahardjo (ITB)
Hendrawan (ITB)
Onno W. Purbo
Bambang Riyanto (ITB)
Joko Siswanto (ITB)
Suhardi (ITB)
Ken Ferens (Univ. Manitoba, Canada)
Santika Wachyudin P. (ITB)
Sarwono Sutikno (ITB)
Hira Laksmiwati Zoro (ITB)
Iyas Munawar (ITB)
Kuspriyano (ITB)
Ngapuli Irmea Sinisuka (ITB)

Table of Contents

APPLICATIONS

E-GOVERNMENT

E-HEALTH

E-LEARNING

POLICY STUDY ON RURAL COMMUNICATION

SOSIOECONOMIC ASPECTS OF R-ICT

INFORMATION & SYSTEM SECURITY

LANGUAGE PROCESSING

SIGNAL & IMAGE PROCESSING

Analysis of Gen Expression and Microarray Data on Pattern Recognition Techniques: Review

Ermatita[1] , Sri Hartati[2]
[1]*Faculty of Computer Science, University of Sriwijaya*
Jalan Raya Jakarta-Inderalaya Km.32 Sum-Sel
[2]*Faculty of Mathematics and Natural Sciences, University of Gadjah Mada*
Bulak Sumur Sekip Utara Yogyakarta
ermatitaz@yahoo.com, tati_ugm@yahoo.com

***Abstract*—Analysis of genes expression can be done with the investigation of a particular microarray data for the description of a gen. This is done to identify what genes that were active in the human body, if wearing a particular treatment. This activity is useful to predict the occurrence of a disease or reaction to medications. This information is an important consideration for medical experts to determine the incidence of disease mechanisms, and determine which therapy is right for the patient. Various studies and research on the microarray data and genes expression has been conducted to identify the genes is. Genes is identified with the introduction of patterns (pattern recognition). Introduction of various technical patterns (pattern recognition) analysis on the expression of genes has been performed. Introduction of techniques such as the patterns in data mining techniques. PCA, k-means Clustering (Partitioning), Bayesian belief Networks (BBN), Hierarchical Clustering, mixture models and EM, Gene Shaving, Hidden Markov Model, GA/KNN, Boolean Network, Novel algorithm. Each has experienced the development of techniques by the researchers in the field, to produce a good performance techniques.**

***Keywords*—genes expression, microarray data, pattern recognition, PCA, k-means Clustering (Partitioning), Bayesian belief Networks (BBN), Hierarchical Clustering, mixture models and EM, Gene Shaving, Hidden Markov Model, GA / KNN, Boolean Network, Novel algorithm.**

I. INTRODUCTION

Science and technology develop rapidly at this time. Many of the development of science and technology has helped people in various aspects of life. In the field of information technology and biological science has developed bioinformatics. Bioinformatics learn that science is the application of the computational technique to manage and analyze biological information. This field include the application of the methods of mathematics, statistics, and informatics to solve biological problems, especially with the use sekuens DNA and amino acids as well as information relating to it. Sample topics include the main areas of this database to manage biological information, sequence alignment, to predict the structure prediction of protein structure and RNA secondary structure, filogenetik analysis, and expression analysis of genes. Next (Nugroho, 2003) [18] states Bioinformatics is a cross-disciplinary science of information technology and biological technology, to address complex problems in the field of biology. Bioinformatics development based on human needs to analyze these data quantity increasing rapidly. Acceleration of the availability of biological data is not released from the harmonious cooperation of information technology and advances in the field of biotechnology [17].

One of the activities in which bioinformatics do is in many investigations of human genes and microarray data. Complete mapping of all the genes possessed by humans have been conducted scientists joined in the field of molecular biology in the call to the Human Genome Project (HGP). Disclosure of data on human genes can recognize all biochemical processes that occur in the human body, the effect on the nature-nature.

Various studies and research on the microarray data and genes expression has been conducted to identify the genes is. Gene is identified with the introduction of patterns (pattern recognition). The introduction is a discipline of learning how to classify the object to some class or category and identify preference of data. This subject is also called the pattern recognition. Pattern Recognition at this time has been learned in the developed and the methods. Activities in this pattern recognition is to map the data in a particular concept that was defined previously [17]. Methods in pattern recognition has been developed. Pattern recognition (The pattern) in the data mining methodology includes: classification, klastering, graphical modeling, etc. Various research has been conducted to analyze microarray data or genes, and different methodologies have also been developed and in use for analize microarray data and gene expression. For example, PCA, k-means Clustering (Partitioning), Bayesian belief Networks (BBN), Hierarchical Clustering, mixture models and EM, Gene Shaving, Hidden Markov Model, GA / KNN, Boolean Network, novel algorithm and many technical analysis genes expression of other microarray data.

II. GEN EXPRESSION AND DNA MICROARRAY

Genom is the genetic material, which is a set of gen-set of genes from a complete organism [19]. Are gen is a DNA Sequence that reduce a protein that has a specific function in a cell is preparing the body, further valavar states. "A DNA molecule is a double-stranded polymer structured in the form of a double-Helix. A gene is a segment of protein coding in the chromosomal DNA that directs the synthesis of a protein. While the cells is a primary and basic unit of work in the body of living creatures [19].

DNA prepared by the four basic molecular units in which a Nucleotides. Each nucleotide consists of a phosphate group, a deoxyribose carbohydrate (sugar), and one of the four nitrogen base called adenine (A), guanine (G), cytosine (C), and thymine (T) (picture 1). Two of the e chain DNA closely related to the hydrogen bonds between nitrogen bases (base-pairs). Base pairing occurs only between G and C, or between A and T [14].

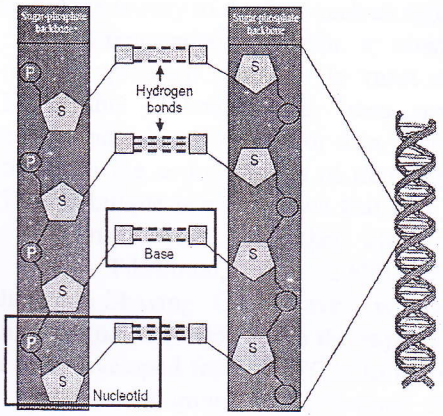

Figure 1. DNA molecule is a double-stranded, double Helix polymer.

Protein in the cell to work efficiently and harmoniously in touch with one another, to prepare an integrated organization in the body. Protein does not hold all available (diekspresi) in the cell but will be available when the protein is needed and will be immediately removed when no longer needed. In other words, the organization of the cells that will provide the protein in the type, amount of time and the right / pas. There is a shift or change, good intention, or when the amount of protein in the provision of a significant, akan can cause illness or aberration.

Provision of protein in the cell body of individuals that this is in an expression with the genes. Protein pattern of providing each individual naturally not the same, it can be said that no two individuals that have a similarity in the pattern of provision of these proteins. However, the differences that still functions in the limit kenormalan overall. Differences in the provision of this protein the cause of the difference between the certainty of the one with the other. Differences are fenotif will seen for example in the form of physical, intellect, emotions, in particular the ability (talent) to all stimulus sensitivity, congenital disease, sensitive against all the influence. Attributes attached to the personal can be tracked from the cell in the provision of protein (genotype). The progress of science in bioinformatics that supported by computer allows the system to be able to analyze the genetic identity that have in each person that will know the nature fenotifnya.

Protein pattern of providing each individual can be in the analysis. Various research has been done to analyze the expression patterns of a large number of genes in human have. Many Bantu tool to help analyze the expression of genes has been developed. With this technology can help people to identify all the inherent nature of someone. This technology can also help people in making the diagnosis, monitor and predict a disease.

III. PROBLEMS IN GEN EXPRESSION AND MICROARRAY DATA

In human genes, the problems that arise is how to understand the structure of genes and the expression, and how genes function in the human body. In this paper will focus on how the techniques in the expression of genes can help in analyzing the expression of genes and microarray data.

IV. GEN EXPRESSION ANALYSIS

Analysis of expression of genes can be done with the investigation of a particular microarray data for the description of one genes, to identify what genes-genes that were active in the human body if wearing a particular treatment. This is useful to predict the occurrence of a disease or reaction to medications .. This information is an important consideration for medical experts to determine the incidence of disease mechanisms, and determine which therapy is right for the patient. Micorarray in the process of analysis can be simply described as follows. First diisolasi from the mRNA sample is returned first in the form of DNA using a reverse transcription reaction. Then through the process of hybridization, only the DNA that are complementary akan berikatan with the DNA on the chip. DNA that has been labeled different colors will show a unique pattern. By using image processing technology (image processing), this pattern is transferred to the next in a numeric expression to be processed with various methods of pattern recognition [17]. Various techniques introduction patterns (pattern recognition) analysis on the expression of genes has been performed [19].

Burge.C, 1997 to do research on the structure of genes with general introduce probabilistic model. By using the computer program GENSCAN to identify genes. Principal Component Analysis (PCA), one technique used in pattern recognition. PCA-based principal components (PCs)., Have been able to reduce the dimensionality

problem space without the more common conditions and information. Karhunen-loe `ve expansion, Mallat (1999) in [19] in research on pattern recognition have been using principal component analysis (PCA), as well as singular value decomposition (SVD) by Anderson (1984) in [19] in statistics is also PCA. Eigengenes used to expand the (model) problem space, which is more accurate. However, the lower the sense of a eigengene, the more noise it. Required balance between the need for maximal expansion of the problem space and the need for noise reduction.

PCA can reduce the problem of dimensional space without losing information in general. This method is easy to understand each eigengene as the main expression vector represents a cluster of expression data (expression pattern). In most studies involving PCA, this technique has been used to find patterns or "modes" in the expression data with the purpose of this link to the transcriptional regulatory action. An eigengene (PC) can be emerge of a major pattern in the data set (expression data). This technique is easy to think of each as eigengene expression vector representation mean a cluster of expression data (expression pattern). In most studies involving PCA, this technique has been used to menemukanpola Modes in expression or data with intent to connect this with the action Modes of transcriptional regulators. This technique is a technique that both when in use in combinasi with the other classification techniques. Hastie, Tibshirani, et al (2000) [25] has developed the Gene Shaving. Gene shaving is a popular approach Statistical pattern detection in the expression of genes in the data developed from the PCA. In the Hastie eksperiment two patient groups were defined from a hirarchical clustering tree grown from a 380 gene cluster. As a predictor, the grouping was just significant in the low IPI group only, at the 0:05 level.

Thus 25/36 = 69% of patients are classified the same way by both groupings. The patients grouping of Alizadeh et al. was based on a cluster of 380 genes, chosen for Their large variation over the samples, the cluster of 234 genes has 38 indicated sign of the gene is to be flipped before averaging); by an asterisk indicates a gene that also falls in the 380 gene cluster from Alizadeh et al, and five of the 234 genes has 38 genes also appear in the unsupervised clusters found earlier, in the second of the three clusters.

PCA has developed many techniques to combine with others to produce the optimal analysis. Bayesian belief Networks (BBN): Bayesian belief Networks (BBN), one technique that has been in use to researchers in pattern recognition. Gifford, et al have used this approach to distinguish between two competing models for galactose regulation. Friedman, et al BBNs use to analyze the genome-wide expression data in the work to identify significant interactions between genes in a variety of metabolic and Regulatory pathway19]. Baldi, et al [2] develop a Bayesian probabilistic framework for microarray data analysis using BBNs to model "value log-normal distribution of expression dependence, parameterized with the way the variances before the appropriate normal distribution to obtain the hierarchical point estimates of both the meter and make a hyperparameter uniform expression of the expression of different genes for each combine with local and empirical different backgrounds with different genes based on proximity. Techniques BBN has developed according to the needs of each analysis, by combination techniques with others, this is done to get optimal results in the analysis of genes expression.

K-means Clustering (Partitioning): K-means Clustering (partition): k-means approach is a disruptive klastering groups. Data (genes or experiments) were divided into groups that have the same expression pattern. k is the number of groups. Number k is the input value is given to the algorithm. The k-mean clustering algorithm is a process of three langkahyaitu: The first step, a random algorithm to provide training data to a group k. In the second stage, the distance mean inter-and intra-class distances are calculated. The average distance between classes (δc) each cluster is calculated with the mean vector (μc) for each cluster, and the average distance between the vectors (data) from the cluster and the average vector. vector is called the average expression vector. In this formula, it is assumed Euclidian distance measurement, and averaging to calculate the arithmetic average. The advantages of k-means algorithm is simple and can be used in the different problem. K-means clustering algorithm designed specifically to evaluate gene spots (on the array images) is done by Bozinov [19]. This technique is based on clustering pixels of a target area in the front and the back of the clusters. Results from the analysis of real gene spots indicate that the performance is very good approach to data analysis methods gen.

Hierarchical Clustering: Hierarchical Clustering is one of the techniques in the analysis of genes and expression microarray data. There are many hierarchical clustering algorithms that can be in the application to the analysis of microarray data. Includes one-linkage clustering, complete-linkage clustering, averagelinkage clustering, weighted pair-group average, and in pair-averaging .. Hierarchical clustering algorithms usually generate a number of the same genes for all combinations of genes. Hierarchical clustering algorithms usually generate gene similarity score for all gene combinations, which put the score in the matrix, genes that have a high similarity value akan join in, and then proceed to combine the pair have little similarity. In the process of clustering, after the calculation of similarity score, the pair identified above related figures-diagonal matrix. At this process node in a hierarchy is created to pair with the highest number, the gene expressed two genes from the profilers average, and merging the elements of weight matrix based on a number of elements in it. elements that are weighted by the

number of elements they contain. For n genes, this process is repeated n-1 times until one element (which contains all the genes) remains.

Wen et al. 26 using clustering and data-mining techniques to analyze data for large scale expression of genes. In the report shows how to integrate the results in the can by using various distance metrics can reveal the differences but the patterns in the data .. Eisen et al.27 also create an elegant demonstration of the ability of hierarchical clustering in the analysis of microarray data. Mixture models: mixture models are a divide and conquer approach to the statistical model. Mixture modeling comes from the fact that not all variables in measuring can. Some of the variables that play a role sometimes described as the behavior of latent or hidden variables. Unconditional varieties are generally used for density estimation, and the one used for Regression condition and Classification problems. Mixture models using a positive convex combination of distribution and the measurement of latent variables to build a model. The use of latent variables that are intended to accommodate the natural system, such as biological and medical can not measure all the variables involved. So that the latent variables that need It is still hidden.

In general there are two types of mixture models, the conditional and unconditional. Unconditional type generally used for density estimation. This condition is used for the problem. Mixture model is used more in the clustering process.

Expectation-maximization (EM) is a two-step iterative process that maximizes the log-likelihood of a mixture model. This valuation of the address can be trusted from the lack of hierarchical clustering in the main the question is sometimes important to consider the data in the study of biological microarray data. Mixture models and EM have been doing clustering in microarray expression data analysis.

McLachlan and colleagues, using a mixture model and EM in developing a software package called EMMIX-Gene. EMMIX Gene-cluster is used to record the microarray expression data from the network examples colon and leukemia. For both data sets, relevant subsets of genes that appear klastering the biological significance of the network in the selection. Cluster-cluster is consistent with the results of the examination or network priori biological knowledge [19].

This method is developed to accommodate data with the variables that are not measurable, with the maximum log-likelihood of the mixture model, this method can overcome the kekurang on hierarchical clustering microarray data in the study.

Support Vector Machine (SVM): Support Vector Machine (SVM) Byun (2003) [24] and Tsuda (2000) [23] have used the Support Vector Machine (SVM) developed by Boser, Guyon, Vapnik, and was first presented in 1992 at the Annual Workshop on Computational Learning Theory in the research. Basic concept of SVM is actually a combination of the harmonic theories of computing have been dozens of previous years, such as margin hyperplane (widower & Hart 1973, Cover 1965, Vapnik 1964, etc..), The kernel was introduced by Aronszajn 1950, and also with the concepts of supporting the other. SVM to solve the problem with the expression vector mapping genes from expression space into higher-dimensional feature, where distance is measured using a mathematical function known as a kernel function, and data may be in for two to the restrictions in the class room linier map features to the restriction non linier expression of genes in space. SVM can be considered as a nonlinear Separation technique. Each experiment generated data points by DNA microarray hybridization represents a comparison of the level of expression of genes under two experimental conditions the differences [4]. SVM uses hyperplanes as separators between the positive and negative points in the feature space.

Select SVM margin hyperplane that provides maximum surface points between positive and negative. After memimisahkan hyperplanes for the points classification function that involves the dot product between points in space. SVM will begin with a set of genes that have a function: for example, genes coding for ribosomal proteins or genes coding for components of the proteasome. In separate groups of genes that are not known to be functional members of a specified class. Both sets of genes in combination form a set of training examples labeled as positive if their genes are in functional class and labeled negative if there is no known functional class. Set of training examples can be easily collected from the literature and data sources. Using this training set of SVM learning akan determine between members and non members of a given functional class based on expression data. After studying the expression-class features, SVM could recognize new genes as members or not members of the class based on expression data.

Results from the Brown experiment with n genes on a single chip, is a series of n expression-level ratio. Results from this eksperiment indicates that some functional classes of genes-genes can be recognized by using SVMs trained on DNA microarray expression data from a comparison of SVMs with four non-SVM methods, show that SVMs provide the best performance [4]. Various studies have been conducted in the SVM's ability to analyze expression of genes. SVM is one of the methods used for the analysis of expression of genes that high dimension. Use of SVM in high dimensions, will not cause negative effects that occur because the curse of dimensionality, from studies with the combination of SVM and development with other techniques shows a good performance compared with other techniques patern recognition in the expression of genes for analysis.

Hidden Markov Models: Hidden Markov Models have been widely learned in the ability for recognition and processing. Hidden Markov models: Hidden Markov

models have been studied extensively to speak and recognition processing. Haussler et. and Krogh et al.36. al.37 is among the pioneers in the use of biological data in the Well has been applied to various problems in biological data mining, such as gene finding and sequence alignment.

Haussler et. al. and Krogh et. al. one of the researchers who use data mining Hmm in biology. HMMs are state-based models, is defined based on a priori knowledge of the biological system into the model. HMMs on data mining applications in biology, among others, find the genes and sequence alignment. Sonnhammer create Pfam, a database of protein Hmm engines to use the provision of a tool to align the various sekuens. Implementation is another Sequence alignment and modeling (SAM) made by Karplus is another example of the application in the sequence alignment Hmm. Hmm, using the binding nucleotides, or groups of nucleotides, in DNA binding sites may be independent [19].

Kulp, D, et all do research for the recognition of human genes using hidden markov model that developed into the Genie gene finding is a model [12].

Genetic algorithm (GA) and k-Nearest neighbor (KNN) (GA/KNN) is another approach in identifying the genes expressed by Li, et al. This approach combines the Genetic algorithm (GA) and k-Nearest neighbor (KNN) to identify genes that can jointly discriminate between different classes of samples. In GA coordinates dots arranged in the problem space as a sequence, as the withdrawal sequence gen. The process of searching the maximum and minimum can be a mutation in the sequence that reaches the new coordinates. In the new coordinates every function in the evaluation, the new points in the set to be more optimal from the past. Points in the new store as the extrema (minimum or maximum) new. Is used in many applications bebagai sequencing. For example, Person has been using GA for DNA fragment assembly. Zhang and Wong GA to implement multiple molecular sequence alignment. Most of the way different types of GA mutation in the sequence, and explore the problem space in a different pattern.

GA/KNN is a supervised stochastic pattern recognition method which is able to select a subset of predictive genes with a set of data that have a data set of large noisy data for sample Classification.

In research Liu, (2002) analyzed the primary sequences using Numerical Characterization and Similarity which defines the scheme provides a logical order of DNA sequences in the primary period of classification based on the nucleic acid. Using logic sequences they produce a set of 4 × 6 matrices to represent DNA primary sequences, which are based on the calculation of all (0,1) triplets in the logic sequences. Using the condensed representation of primary DNA primary sequences and the eigenvalues of the Symmetric real matrix a comparison is made between the primary sequences for exon-1 of human β-globin and seven other species. This method is the development of the matrix method to determine the invariant as new descriptors for DNA sequences [2]. This method is a good method to use for DNA Sequence Analysis to analyze the expression of genes [14].

Kapushesky (2004) have created a tool to help identify and ekspresii genes microarray and other genomic functions, namely Expression profiler. Expression profiler (EP, http://www.ebi.ac.uk/expressionprofiler) is a web-based platform for microarray gene expression and other functional genomics-related data analysis. Web-based design of Expression profiler support data sharing and analysis in a collaborative environment that is guaranteed. Development tool integrated with the microarray gene expression database. EP: NG is an open-source project that gives hope for the distribution and extension of the scientific community [3]. Tools developed in this very helpful in identifying the microarray and the expression of genes, to genes from manganalisis conditions in order to provide optimal results from the analysis of genes [11].

Cho, (2003) tried to explore some classification techniques and feature selection. In percobaannya adopted seven feature selection methods and classification of the four techniques used in datamining and pattern recognition in classifying cancer precisely. Feature selection methods including Pearson's and Spearman's correlation coefficients, Euclidean distance, patcosine coefficient, information gain, mutual information and signal to noise ratio. Classification techniques and multi-layer perceptron (MLP), k-nearest neighbour (KNN), support vector machine (SVM) and structure adaptive self-organizing map (Som) in machine learning. In this experiment also combine several classifiers with majority voting to enhance the performance of classification. Machine learning is defined for a DNA microarray select discriminative genes associated with the Classification of gene expression data, trains a new classifier using the data using the learned classifier.

Gene expression data calculated from the DNA microarray. We predict two-phase system with the feature selection and patternclassification stages. Feature selection can be kemukakan as gene selection, to get a list of genes that may be informative for the prediction with the Statistical, informatioputusan to categorize the pattern of genes located at the input.

This method is very simple and there are many methods to combination classification techniques in the field of machine learning and data mining. Applications with more sophisticated methods from the same datasets to strengthen the results obtained in order to become better. This research has developed an optimal combination of feature-classifier to produce the best performance in classification. Information obtained, Pearson's correlation coefficient is the top feature selection methods, and MLP and KNN is best classifiers [7].

Jaeger (2006) in the research entitled "Selecting

normalization genes for small Diagnostic microarrays" results obtained from the difference in normalizing large microarrays and small Diagnostic microarrays. In this research to include the proposed additional normalization genes on the small Diagnostic microarrays and propose two strategic select for each genome wide microarray studies. The first is a data driven selection of univariate normalization gen-gen. Multivariate and the second is based on the discovery of Diagnostic balanced signature. Then do the second comparison method for standard normalization protocols known from extending widely in the microarray system. Here recognize that this method can be said to provide better GAMI with the ability to express a variable-length motifs and allow GAMI to menhubungkan difference motives in CTFR Data [6]. Method developed is very well applied to the microarray for r not too big not [10].

Congdon (2008) conducted a research entitled "An Evaluation of Information Content as a Metric for the Inference of Putative Conserved Noncoding Regions in DNA sequences Using a Genetic Algorithms Approach." On this research he memperkenakan MC and IC metrics and some of the variation than to use it as a Fitness functions in GAMI. GAMI method is for a set of nucleotide sequence patterns that appear at least once in each sequence. Motif representation is the standard consensus motif: an N-mer base that consists of A, C, G, and T. For example, if we find 8mers, possible motives identified include akan CATGCAAT, TAGGAACT, ACTTACGT, and so forth. As a function of initial fitness, used a metric called the "match count" (MC). To evaluate patterns of MC is given, each sequence of consecutive sought to find the best match in the sequence. Match between Forward and reverse-Complement considered in each sequence. Match maximizes the best of a number of bases is according to the motif in all sequences. But there is one or more of one the most appropriate patterns and nucleotide sequences. Maximum score value of bases matched in each sequence is the motif of the sequence. Score for each motif in the sequence data is the overall score of the motif. MC good performance in this experiment. With the addition to the motif in the matrix that is used to calculate the IC, can increase the IC as a fitness function for GAMI (Congdon) [8].

Yin in research to introduce an exploratory framework for learning patterns of conditional coexpression in the expression data gen. This approach can also estimate how the content of information disseminated by a set of M nodes in a network. (where each node is associated to an expression profile) change the situation on a set of L conditioning variables (in the case of simple direpresentakian a separate set of expression profiles).

VCD algorithm that identifies possible coexpression groups in the data is based on the scope of identification for vector variation gen. Algorithm include the vector variation tendencies have any similarity, that is the same direction will be the same group. Center coverage will always be in the area of space vector space vectors Centroids of a representative group of vectors, an algorithms are proposed, should meet two constraints:

1. Must include the number of minimized.
2. The number of combination of two elements must b minimized.

This method can investigate biological Annotations fo coexpression patterns. This experiment using the Gen Ontology (GO). Saccharomyces Genome Databas (SGD). This tool is designed to search for significan shared GO terms of genes and groups provide ways t identify characteristics of genes that may hav similarities.

This method is nonparametric and is based on th concept of Statistical coinformation, which is not th same as conventional techniques based on correlation which does not limit the scope to the linear conditiona dependency patterns.dengan applying exploratory metho has been able to learn the statistical. significant pattern of conditional.

Also introduced in the research ability is a nove method of detecting linear, the nonlinear patterns o conditional coexpression in the measurement o expression of genes [3] Furthermore Yin (2008) i research, propose a novel scheme, which is a variatio diesain-based Coexpression Detection (VCD) algorithm to analyze the trend of expression based on the diference [21].

Chin, et al (2008) conducted research entitled "DNA Motif Representation with Nucleotide Dependency" introduces a new motif representation called scored Position Specific Pattern (SPSP), which is a generalization of the string representation matriks and that in considering the use dependent occurrences of neighboring nucleotides. In this algorithm starts with a good set of string patterns, based on local search, find several areas related to the optimal SPSP binding sites. Algorithm has two main stages. First, the seed serves as a searching, to find the set of motif length-l string binding in the input sequence. The second phase begins with each length-l string R as a seed SPSP representation and combine some of R's position to the binding sites in the form of SPSP representation with p-value is smaller. Repetition is done until the p-value not dapaat reduced. This research has the SPSP Finder in C + +. SPSP Finder is used to find motifs in biological data between simulation and real [6].

Chalco, (2008) in research to introduce a method based on a modified Gabor-wavelet transform (MGWT) to identify the protein region it. Several methods of DNA it is based on the model independent of specific patterns of nucleotide in the region it .. This result indicates that the source of the identification error generated by the previous method is the fixed working scale, this new method not only avoid the source of the error but also help create a tool that provides a detailed exploration of the nucleotide occurrence. This method allows use on

multiple scales, analysis of coding regions between the small scales with small and larger coding regions with large scales. The advantages of this method is ketangguhannya for different scale on the analysis of DNA sequences [6].

Using logic Sequence we produce a set of matrik × 4 to 6 represents the main DNA Sequence, which base on all the calculate (0,1) triplets in the logic sekuens. Using the condensed representation of primary DNA primary sequences of eigenvalues and corresponding Symmetric real matrix a comparison made between Sequence primer for exon 1 of human-globin seven other species. This study suggests that shrinkage method to bring information from the relevant genes goup. Found that the shrinkage method consistently works well in a variety of scenario [20].

Elo (2008) Introducing a procedure to increase reproducibility-optimization, which allows selection of a statistical ranking of the genes directly from the appropriate data in comparison with existing ranking methods, reproducibility-optimized statistic shows good performance consistently in the different simulation conditions and the Affymetrix spike-in data set. Feasibility of the novel is set in a statistical practical research setting using data from a microarray study in-house cDNA in asthma-related changes in expression of genes. This result suggests that facilitate the procedure of selection of the appropriate statistical test to data set without relying on a priori assumptions and allow the discovery of ienterpretasi be biased. General reproducibility-optimization procedure is not limited to the range of other applications that better [9].

Liang, Fuhrman and Somogyi (PSB98, 18-29, ·1998) have described the algorithm for inferring a genetic network architectures from the state transition table that is connected to the time series pattern of expression of genes. Using the Boolean network model, they generate komputasional experiment, and suggested that a small part of the transition state (INPUT / OUTPUT) command pair is suficient to infer the correct original Boolean network. This study provides verification of the mathematics observasinya. They make the computational experiment that in order to ekspose constant factor involved in the notation O (log n). The results show that the Boolean network with a size of 100000 can be identified with the algorithm from about 100 INPUT / OUTPUT, if the maximum indegree pair in the limit 2. algortma virtue of this is also a conceptual simple that can be extended to the next realistic network models [1].

The motivation of this activity is to identify methods for the accuracy of the DR menganalisi genes with higher dimensions and study of protein expression. In this study comparing the three schemes DR nonlinier (Isomap, Locally Linear Embedding and Laplacian Eigenmaps and three DR scheme linier (PCA, Linear discriminant Analysis, and Multidimensional scaling) with the purpose of determining reduced subspace repesentation, where the individual class of objects is more easily can be distinguished. In a statistical significant improvement in the quality of the right womb cancer 10 datasets using three non linier DR scheme in the top three results of the DR technique linier scheme has been in the observation [13].

No.	Researcher	Method	Algorithm	Achievements
1	Akusu T. and Miyano, S	Identification	Boolean network	development of faster algorithms development of faster algorithms
2	Baldi,P, and Long,A.D and Friedman	Identification	Bayesian probabilistic framework with BBN	dependent normal distributions for point estimates
3	Boscolo, R, Liao, J. C. and Roychow-dhury, V. P	detection	Information Theoretic Exploratory Method	conditional coexpression in the measurement of expression genes
4	Brown, M. P. S. et all		Knowledge-based analysis	
5	Bozinov	clustering	K-means clustering algorithm	Development of the paitern analysis of genes and microarray data
6	Burge,C and Karlin,S		general probabilistic model	
7	Chin, F. and Leung, H. C.M.	Analisis	Scored Position Specific Pattern (SPSP)	can not guarantee search optimal motif in SPSP representation
8	Cho, S. and Won, H.	classification	Machine Learning	feature-classifier to produce the best performance on the classification and clustering
9	Congdon, C. B	identification	Genetic Algorithms Approach	Development of other methods with the fitness function
10	Elo, L. L, Lahesmaa, R.and Aittokallio, T	selection	Reproducibilit y-Optimized Test Statistic	facilitate the procedure of selection of the appropriate statistical test to data set without relying on a priori assumptions and allow the discovery of ienterpretasi be biased
11	Hastie, Tibshirani, et al dan Mallat	classification	Gene Shaving dan Principal Component Analysis (PCA)	Development of the PCA with other techniques
12	Jaeger, J. and Spang, R	identification	Selecting normalization genes	Development of the PCA with other techniques
13	Kapushesky , M	identification	Expression Profiler	Development of software that better
14	Yin, Z. and Chiang, J		Variation-based Coexpression Detection algorithm	analyze the trend of expression based on the differences and similarities

V. DISCUSSION

Various techniques in analyzing the expression of genes and microarray data was done in a lot of research. Research that has been done is the development or testing of a particular method or technique in the field bioinformatics. This is intended to produce an analysis and prediction of gen in the micrroarray data and can provide stimulus for the best match and the conditions and circumstances of a particular man.

The development of techniques in the analysis of genes expression and micrroarray data this is done in order to meet the needs of their data to the gene expression and micrroarray data. All techniques that have been done show that each of the existing techniques have been developed and through advanced research conducted by scientists can be used in the genes expression to analyze the micrroarray data . Each technique has advantages and disadvantages of each. And every technique that has been developed, the ability and reliability in the analysis also depends on the condition of the data to be in the analysis.

ACKNOWLEDGEMENTS

The work described in this paper was partly supported methode in the Classification of data mining that will be in application in the next research.

REREFENCES

[1] T.Akusu dan S.Miyano, "Identification of Genetic Networks From A Small Number of Gene Expression Patterns Under The Boolean Network Model",

[2] P. Baldi and A.D, Long, 2001, "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes",bioinformaticsVol. 17 no. 6 2001 Pages 509–519

[3] R. Boscolo J. C. Liao and V. P. Roychowdhury, 2008, "An Information Theoretic Exploratory Method for Learning Patterns of Conditional Gene Coexpression from Microarray", Data IEEE/ACM Transactions on Computational Biology and Bionformatics, VOL. 5, NO. 1, January-March 2008,15-24

[4] M. P. S Brown, et all, "Knowledge-based analysis of microarray gene expression data by using support vector machines", pp: 262–267 u PNAS u January 4, 2000 u vol. 97 u no. 1

[5] C. Burge and S. Karlin "Prediction of Complete Gene Structures in Human Genomic DNA", JMB J.Mol Bol,1997,268. 78-94 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.3107

[6] F.Chin and H. C.M. Leung, 2008, "DNA Motif Representation with Nucleotide Dependency", IEEE/ACM Transactions on Computational Biology and Bionformatics, VOL. 5, NO. 1, January-March 2008,110-119

[7] S.Cho and H.Won,. , 2003, "Machine Learning in DNA Microarray Analysis for Cancer Classification", This paper appeared at First Asia-Pacific Bioinformatics Conference, Adelaide, Australia. Conferences in Research and Practice in Information Technology, Vol. 19. Yi-Ping Phoebe Chen, Ed.,2003

[8] C. B Congdon, 2008, "An Evaluation of Information Content as a Metric for the Inference of Putative Conserved Noncoding Regions in DNA Sequences Using a Genetic Algorithms Approach", IEEE/ACM Transactions on Computational Biology and Bionformatics, VOL. 5, NO. 1, January-March 2008,1-14

[9] L. L, Elo, R.Lahesmaa and , T. Aittokallio, 2008, " Reproducibility-Optimized Test Statistic for Ranking Genes in Microarray Studies", IEEE/ACM Transactions on Computational Biology and Bionformatics,, VOL. 5, NO. 3, July-Septe4mber 2008, 423-431

[10] J.Jaeger and R. Spang, 2006, "Selecting normalization genes for small diagnostic microarrays", BMC Bioinformatics 2006, 7:388 , doi:10.1186/1471-2105-7-388

[11] M.Kapushesky, "Expression Profiler: next generation—an online platform for analysis of microarray data", Nucleic Acids Research, 2004, Vol. 32, Web Server issue, W465–W470 available on DOI: 10.1093/nar/gkh

[12] D,Kulp, D.Haussler, et all, "A Generalized Hidden markov Model for the recognition of human Genes in DNA", http:// citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.6842

[13] Lee, C. Rodriguez, and A. Madabhushi, , 2008, "Investigating the Efficacy of Nonlinear Dimensionality Reduction Schemes in Classifying Gene and Protein Expression Studies", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 5, No. 3, July-September 2008, 368-384

[14] Y. Liu,. "The Numerical Characterization and Similarity Analysis of DNA Primary Sequences", Internet Electronic Journal of Molecular Design 2002, , Volume 1, Number 12, Pages 675–684, December 2002

[15] J.P Mena-Chalco,. et all, 2008, "Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform", IEEE/ACM Transactions on Computational Biology and Bionformatics, VOL. 5, NO. 1, January-March 2008,198-207

[16] E Meiyanto, ,2001, "Farmakogenomik: sebuah paradigma Baru Dalam Sistem Pengobatan",Poster Forum-Zoa Biotek 2001 /available on http://www. std.ryu.titech.ac.jp/ ~indonesia/zoa/paper/html/posteredymeyanto.html, akses 4 Desember 2008

[17] A.S Nugroho, A.B. Witarto, dan D.Handoko, 2008 ," Support Vector Machine: Teori dan aplikasinya dalam Bioinformatika, http://ilmukomputer.com, Desember 2008 (in Indonesian)

[18] A.S Nugroho, "Bioinformatika dan Pattern Recognition", http;//asnugroho.net/papers/ikcbioneinf.pdf

[19] F. Valafar, 2002 "Pattern Recognition Techniques in Microarray Data Analysis: A Survey", Special issue of Annals of New York Academy of Sciences, Techniques in Bioinformatics and Medical Informatics. (980) 41-64, December 2002.

[20] P. Wei, and W. Pan, , 2008, "Incorporating Gene Functions into Regression Analysis of DNA-Protein Binding Data and Gene Expression Data to Construct Transcriptional Networks", IEEE/ACM Transactions on Computational Biology and Bionformatics, VOL. 5, NO. 3, July-September 2008 , 401-415

[21] Z. Yin and J Chiang, "Novel Algorithm for Coexpression Detection in Time-Varying Microarray Data Sets", IEEE/ACM Transactions on Computational Biology and Bionformatics, VOL. 5, NO. 1, January-March 2008,120-135

[22] McLachlan, D. Peel, K.E. Basford, P. Adams, The Emmix Software For The Fitting Of Mixtures Of Normal And t-Components, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.1530

[23] K. Tsuda , 1999, "Support Vector Classifier with Asymmetric Kernel Functions ",ESANN'1999 proceedings - European Symposium on Artificial Neural Networks Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 183-188. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.3123

[24] H. E. Byun and K., 2003, "Cheverst Supporting proactive "intelligent" behaviour: the problem of uncertainty" http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.2704

[25] T. Hastie, R.Tibshirani, et all, 2000, Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns (2000) http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.719

[26] H. Mallat, , Dr. Tech Esko Nuutila EMIL: A Toolkit for Program Analysis and Modification (1999) http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.4324