International Conference on Information Technology & Multimedia (ICIMU'05). November 22-24. Cyberjaya, Malaysia.

BIOINFORMATICS: A TOOL FOR BIO-KNOWLEDGE GENERATION

Safaai Deris, Hany Taher Alashwal, Mohd Saberi Mohamad, Yeo Lee Chin, Muhammad Razib Othman, Yuslina Zakaria, Suhaila Zainudin, Nazar Mustapha Zaki, Saad Othman Abdalla, and Satya Nanda Arjunan
e-mail: safaai@fsksm.utm.my

Artificial Intelligence and Bioinformatics Laboratory, Department of Software Engineering, Faculty of Computer Science and Information System, Univerisiti Teknologi Malaysia.

Abstract

With the knowledge-based economy (KBE) emerging as a new paradigm, it becomes important that economic actors acquire new knowledge and technology. This will stimulate innovation and contribute further growth and development. In the knowledge-based economy, it is not tangible capital but intangible and innovations that determine international competitiveness, productivity growth and worker's status in the labour market. This intangible capital and innovation are supplied by knowledge-based society (KBS). The knowledge are acquired by society through research and innovation and disseminated through lifelong teaching and learning. After digital and ICT revolution comes biotechnology revolution which again change the way of life of society. With this revolution the knowledge from biotechnology and life sciences become the dominant component of KBS. With abundant of data generated by experiments using new automated equipment make bioinformatics a major tool to transform data into information and knowledge. This paper describes features of KBS and it role in KBE and how it is related to the roles of bioinformatics as a tool for generating knowledge in KBS. The paper also presents the current results of our research including the techniques for predicting protein structure and function and system for life sciences which is the basic knowledge for designing and applying new bio-products and organism.

1. Introduction

As we are approaching the final decade of the 20th century, two great changes began transforming economies and ways of life around the world. The first one is the globalization - as economies everywhere became increasingly interdependent, a global economy was being born. The second was the technological revolution – the coming of the internet and of new information and communication technologies. Social and economic ecosystem is a dynamic process. We have observed that the society has passed through at least 3 different era of ecosystem in the past several decades; from industrial society to information society and now entering the knowledge-based society (KBS).

KBS is a society where social system including life styles and business methods of all economic subjects such government, companies and individuals are reorganized through embracing applications of IT-related technologies. Transition from information-based to knowledge-based society has brought about changes in business and education in terms of the way information is processed and problems resolved. In KBS, knowledge plays a central role in economic and social development.

Knowledge-based economy (KBE) can only be achieved by KBS. Digital and knowledge is a catalyst to the formation of a knowledge-based socio-economy. Elements of KBE (World Bank KAM project, 2002) can be summarized as follows:

- Creating an appropriate economic incentive and industrial regime that encourage the widespread and efficient use local and global knowledge in all sector of economy, that foster entrepreneurship and that permit and support the economic and social transformation.

- Creating a society of skilled, flexible and creative people with opportunities for quality education and life-long learning available to all.

- Building a dynamic information infrastructure and a competitive and innovative information services and tools available to all sectors of society – include high end ICT and other element of information rich society such as radio, TV, etc.

- Creating an efficient innovation system comprising firms, science parks, incubators, research centers, universities, think tanks and other organization that tap into and contribute to the growing stock of global knowledge, adapt to local needs and use to create new product services and ways of doing business.

KBE is characterized as follows:

- KBE is a service economy
- KBE is an information society
- KBE deals with the production and the use of knowledge
- KBE is an innovation society

There are several issues need to be address in moving toward KBS. First, how to motivate human resources to continuously create an apply new knowledge; Second, how to maintain the organization that provide thoughtful perspective on application and third, how to address questions of basic rights of access to knowledge

A successful modern economy is founded on a strong scientific base that has the ability to convert scientific research and knowledge into products and services which bring social and economic benefits. The future of the country is to make major investment in research and development in ICT and Biotechnology.

2. KBS around the world

Countries around the world work very hard to become more competitive by promoting KBS. In 2000, the EU has set a few strategic goals for the next decade: to become the most competitive and dynamic KBE in the world, capable of sustaining economic growth with more and better jobs and greater social cohesion. This new strategic goal has been accepted as an inspiring example by many countries in the world. Among those countries are Finland and Korea which persistently attempting to provide required infrastructure and environment for growth. Other countries are Canada, Ireland, India, Thailand, Indonesia, and Malaysia which are working toward more competitive KBE by creating KBS. Among other efforts toward KBS include:

- business and citizens must have access to an inexpensive, world-class communication infrastructure and a wire range of services;
- every citizen must be equipped with the skills needed to live and work in new KBS

The government of Korea is one example of aggressively pursuing toward becoming KBS from the beginning through implementation of several mega projects as follows:

- National Basic Information System (1987 – 1996)
- Reorganisation of MOC into MIC (1994)
- Informatization Promotion Act (1995)
- Cyber Korea 21 (1999)

In addition to above projects, there are other initiatives such as “Trans-Eurasia Network” and privatization of Korean Telecom. As a result, the digital economy represented by e-commerce and information technology will take up a greater part in Korean economy. IT industry contribution to gross domestic products increase to 10.7% in 1999 and export industry reached 58% of trade surplus in the same year.

In EU, the sixth framework programme of the European Community for Research and Technological Development (RTD) activities is another example of serious effort to alleviate economic competitiveness against US and Japan where over €11 billions have been allocated for a period of 2002 – 2006. Three of seven priority areas under this program are related to ICT, biotechnology and knowledge-based society. The three areas are genome and biotechnology for health, Information society technologies and citizen and governance in an open European knowledge-based society.

Other countries such as Finland, Ireland, India, Canada, and Indonesia, etc have their own programs leveraging their strengths, resources preparing their society for knowledge-based society. Malaysia also is aggressively promoting KBS through MSC and Bio-valley initiatives.

3. Roles of research and innovation in KBS

Society in the 21st century calls for active development of new knowledge focusing on science and technology and the application of such knowledge to society. In economy, competition will intensify in which the winner will be the one who obtain knowledge and information fastest. Science and technology and society become more closer. In this way, society of 21st century will make a transition to a KBS in which all activities of society, such as industry and public lifestyles will develop rapidly based on knowledge. In order to realize the above vision, the roles innovation, research, teaching and learning, and entrepreneurship become increasingly importance. The policy, system and method for the above processes need to be accordingly reviewed and revised to suit with the new ecosystem of KBS. Each plays its important role. Research is responsible for acquiring new knowledge required by society. Innovation is related to discovery and invention. Discovery represents a completely new addition to the existing body of knowledge. Invention is a creation of something that did not exist before but based on existing knowledge. Innovation is based on some combination of discovery and invention. It is based on a combination of existing knowledge and a new way of applying that knowledge or it is the application of existing knowledge to different situations or problem.

Teaching and learning is the process of dissemination of knowledge to all layer of society for economic or social purposes. Entrepreneurship is the utilization or application of knowledge for economic and business purposes. The transition to KBS can be facilitated by having efficient innovation system that comprise of clusters of universities, research centers, incubators, science parks, public institutions and business institution or firms. The entities in this cluster are linked together based on the value chain of the products. These clusters must be supported by suitable socio-political and economic environment, policies, procedures, legal, and infrastructure.

4. What are biosciences and bioinformatics

Bioscience is the study (science) of life. It is concerned with the characteristics and behaviors of organisms, how species and individuals come into existence, and the interactions they have with each other and with the environment. Biosciences provide the basis for applied fields such as agriculture (fisheries, forestry, food sciences, animal sciences), biotechnology, medical sciences, and health sciences. Bioinformatics is research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data including those to acquire, store, organize, analyze or visualize such data. The related discipline of computational biology is the development of and application of data analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavior and social sciences.

We can summarize the entire field of bioinformatics with three perspectives. The first perspective is the cell. The central dogma of molecular biology is that DNA is transcribed into RNA and translated into protein. The focus of molecular biology has been on individual genes, messenger RNA transcripts and protein. The second perspective focuses on individual organisms. Each organism changes across different stages of development and across different region of the body. The third perspective is the tree of life represents the largest scale of biosciences. There are many millions of species alive today and the can be grouped into the three major branches of bacteria, archaea and eukaryotes.

In general terms, bioinformatics is the application of analytical theory and practical tools of mathematics and computer science to provide the computational management of all kinds of biological information.

5. Bio-data and Bio-database

The explosion of interest in bioinformatics has been driven by the emergence of experimental techniques that generate data in a high throughput fashion. Examples of these techniques are DNA sequencing, mass spectrometry, and micro-array expression analysis, X-ray crystallography, and electron microscopy. Bioinformatics depend on the availability of large data sets that are too complex to allow manual analysis.

The availability of high-throughput experimental techniques mentioned above generates large volume of data of different scale. Analysis of these data require input from many different disciplines such as computational scientists, statisticians, and computer scientists, to analyze and interpret and transform into a knowledge to be used by users.

Biology has always been a haven for microscopes, test tubes, and Petri dishes, but this conventional picture of the field is expanding rapidly. Sophisticated techniques adapted from physics, chemistry, and engineering enable scientists to use computers and robots to separate molecules in solution, read genetic information, reveal the three-dimensional shapes of natural molecules like proteins, and take pictures of the brain in action. All of these techniques generate large amounts of data, and biology is changing fast into a science of information management.

Today's biomedical researcher routinely generates an amount of data that would fill multiple compact discs, each containing billions of bytes of data. (A byte is approximately the amount of information contained in an individual letter of type on this page.) There is no way to manage these data by hand. What researchers need are computer programs and other tools to evaluate, combine, and visualize these data. In some cases, these tools will greatly benefit from the awesome strength of supercomputers or the combined power of many smaller machines in a coordinated way but, in other cases, these tools will be used on modern personal computers and workstations.

This new field of bioinformatics born out of the need to process extremely large volume of data generated by high throughput equipments. These data can be categorize as follows; genome sequence data, gene expression micro-array data, protein sequence data, protein structural data, and pathway data. These data have been organized and store in the following databases; literature databases, microarray databases, nucleotide databases, protein databases, proteomic databases, structure databases, pathway databases. These databases have been made available to access by public over internet. Today, there several major sites that provide access to these databases and other related databases. The following are the examples of the major sites provide access to public.

Table 1: Major sites hosting major bio-database.

Resource	Description	URL
DNA database of Japan (DDBJ)	Associated with the Center for Information Biology	http://www.ddbj.nig.ac.jp/
European Bioinformatics Institute (EBI)	Maintain the EMBL database	http://www.ebi.ac.uk/
National Center for Biotechnology Information (NCBI)	Maintain GenBank	http://www.ncbi.nlm.nih.gov
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Database of pathway, genes, ligand, brite	http://www.genome.jp/kegg/
Encyclopedia of Metabolic Pathway (MetaCyc)	Maintain database of metabolic pathway from more than 300 organisms	http://metacyc.org/
Gene Ontology	Provide control vocabulary of gene and gene product	http://www.geneontology.org/
PubMed	Archive of literature database from Life Sciences Journal	http://www.pubmedcentral.nih.gov/

Today, there are several hundreds, even thousand sites of different sizes storing wide range of data related to gene, genome, protein, structures, and pathways.

6. Bioinformatics and Bio-knowledge Generation

Bioinformatics field provide computational techniques and tools for analysis of data from the above databases or data direct from high-throughput experiments. Computational tool and techniques from Bioinformatics will be used to process and to transform basic data such DNA macromolecules, nucleotide sequence, or DNA microarray expression data into a series of representation from as simple as classes, clusters, relation, and moving into more complexes representation such networks, rules, reaction and into more complexes representation such as pathway, function and even process and finally in the form of cell and even organ (Figures 1 and 2).

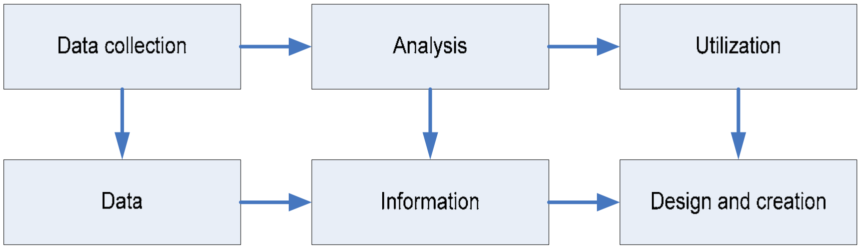


Figure 1: Data analysis and utilization**.**

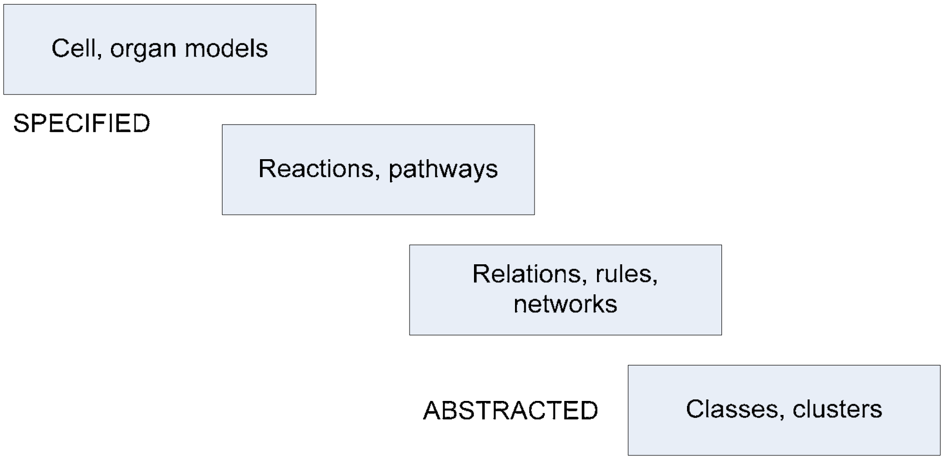


Figure 2: Knowledge creation stages.

In order to transform raw basic data from life sciences field to higher level information and finally to knowledge involve wide range of computational domain and artificial intelligence as shown in Figure 3.

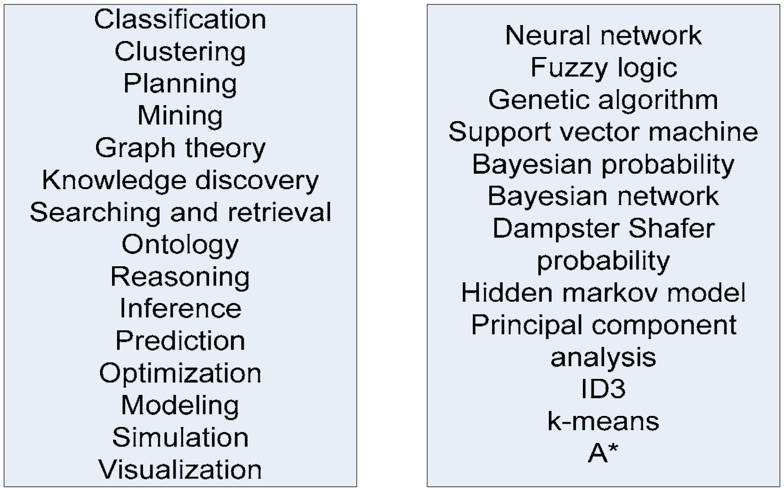


Figure 3: Examples of domain and the associated techniques for data analysis.

The above techniques may be applied independently or in the form of hybrid and combination of several above techniques (Figure 4). Life sciences data from high throughput experiment normally huge and extremely large in size. The nature of data from life sciences is very different from numerical data from engineering design problem. The data from life sciences mainly in the form of strings or sequences and the size of sequence is huge. Therefore bioinformatics exercise involve compute intensive and data intensive computation. As a result, in most cases bioinformatics computation requires high performance computing facilities utilizing cluster or grid computing.

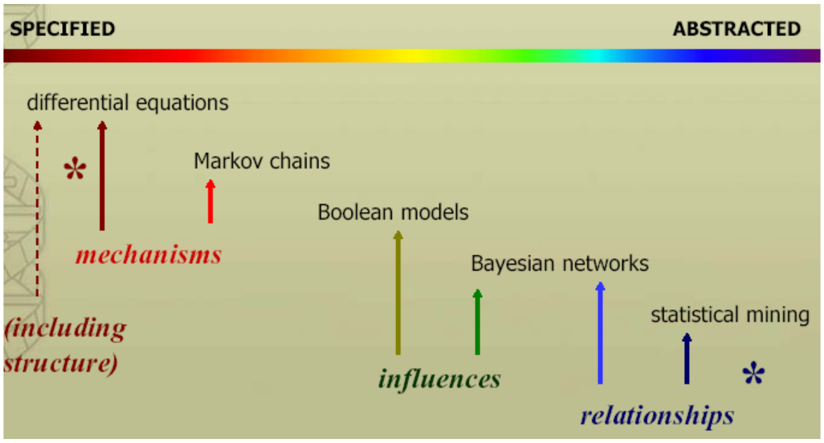


Figure 4: Techniques for knowledge creation.

The highest level of study of living organism is the modeling and simulation of cell, organ and even whole human body. Examples of these project are e-cell (http://www.e-cell.org/) and digital human http://www.sci.utah.edu/ncrr/software/biopse.html, http://www.fas.org). Through modeling and simulation we expect we can learn more about the function and structure of the living organism.

7. Bio-knowledge and Knowledge-Based Society

Our society has been transformed from agrarian society to industrial society and then to information society and now into knowledge-based society. This transformation is a result of industrial revolution, and the ICT revolution, and now at the beginning of biotechnology revolution. Every step of this revolution changes the way of live of human being (Figure 5).

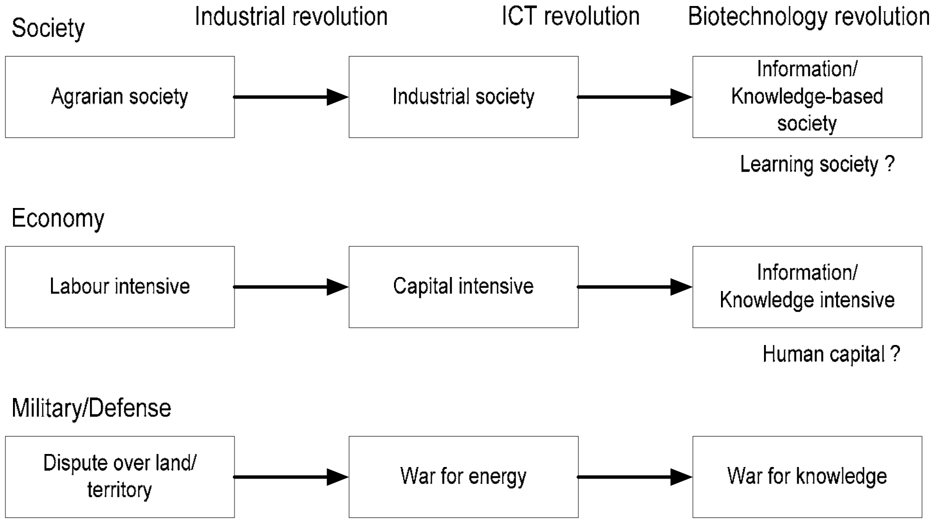


Figure 5: Technology revolutions.

In recent years we noticed that new revolution has started, i.e. biotechnology revolution. Biotechnology field has direct relationship because bioinformatics provides tools and methods to support and accelerate the experiment and development by in-silico instead of in-vivo. By this way scientist save time and effort to conduct experiment and testing. Biotechnology and bioinformatics have direct implication on basic need of human being, i.e. the food production and medicine and health care. Without these two basic need human been can be in very difficult situation and even have use as a weapon to suppress others. The knowledge generated by bioinformatics can be use to save human been in many ways as follows:

- Medical and healthcare applications
- Drug design
- Agriculture and food
- Defense and security
- Bio-based industrial product

Currently site http://www.biobase.de/pages/who/biobase.html has been providing the information and knowledge related to diseases, important and organism in systematic manner. This information is accessible to anybody connect to internet. In future there will be many more organisms added to the database.

The main purpose of bioinformatics, bio-medical scientist and biologist is to understand the structure and function of organisms include human been. So far, we just understand less than 10 % of function and structure of our body. If we understand more, more can be done in making the human been live in harmony in this world.

8. Bio-informatics research at AIBIL, UTM

In the past 7 years, our research focus mainly on understanding the structures, functions and biological systems using genome sequence, protein sequence, gene expression microarray data and path way. By understanding the structure, function and system will lead us to understanding of diseases and illness and also the design of new organism and new drugs. This knowledge hopefully will benefit human being for better and comfortable life.

8.1. Protein Structure Prediction

The most common structure prediction is for their secondary structure from genome sequence. We started our research in this area using combination neural network and Dempster-Shafer probability and then further improve by using information theory.

8.1.1. Secondary Structure Prediction using Neural Network and Dampster Shafer Probability

Recently, Denoeux proposed a novel neural network classifier based on the Dempster-Shafer theory (DBNN). Several of his preliminary experiments in some typical problems demonstrated that the classifier has an excellent performance when compared to other statistical and machine learning approaches. As a result, this research extends the initial work by examining its potential improvements and applicability in a new real world task such as the protein secondary structure prediction. The classifier permitted rigorous experiments to be conducted in two other benchmark problems with disparate dimensions to determine the classifier's inherent attributes and drawbacks.

The experiments showed that although the classifier performed better than some of the best methods such as Support Vector Machines and Kernel Fisher Discriminants in the small dimensional problem (dimension size = 9), its performance deteriorated significantly in the higher dimensional problem (dimension size = 60). This presented a substantial challenge because the secondary structure prediction exhibits high dimensionality as well. An improved version of the classifier was designed by introducing Multilayer Perceptrons to replace the distance measure of the classifier, which appeared to be impaired in high dimensions. Figure 5 illustrates the figure for our proposed approach (αMLPDS).

Experimental results of the secondary structure prediction have shown in Table 1 and demonstrated that the new classifier (αMLPDS) performed better than the original one (DBNN). Moreover, at the level of sequence-to-structure prediction, its performance was comparable to the PHD (Profile network from Heidelberg) method, which is one of the best secondary structure prediction schemes.

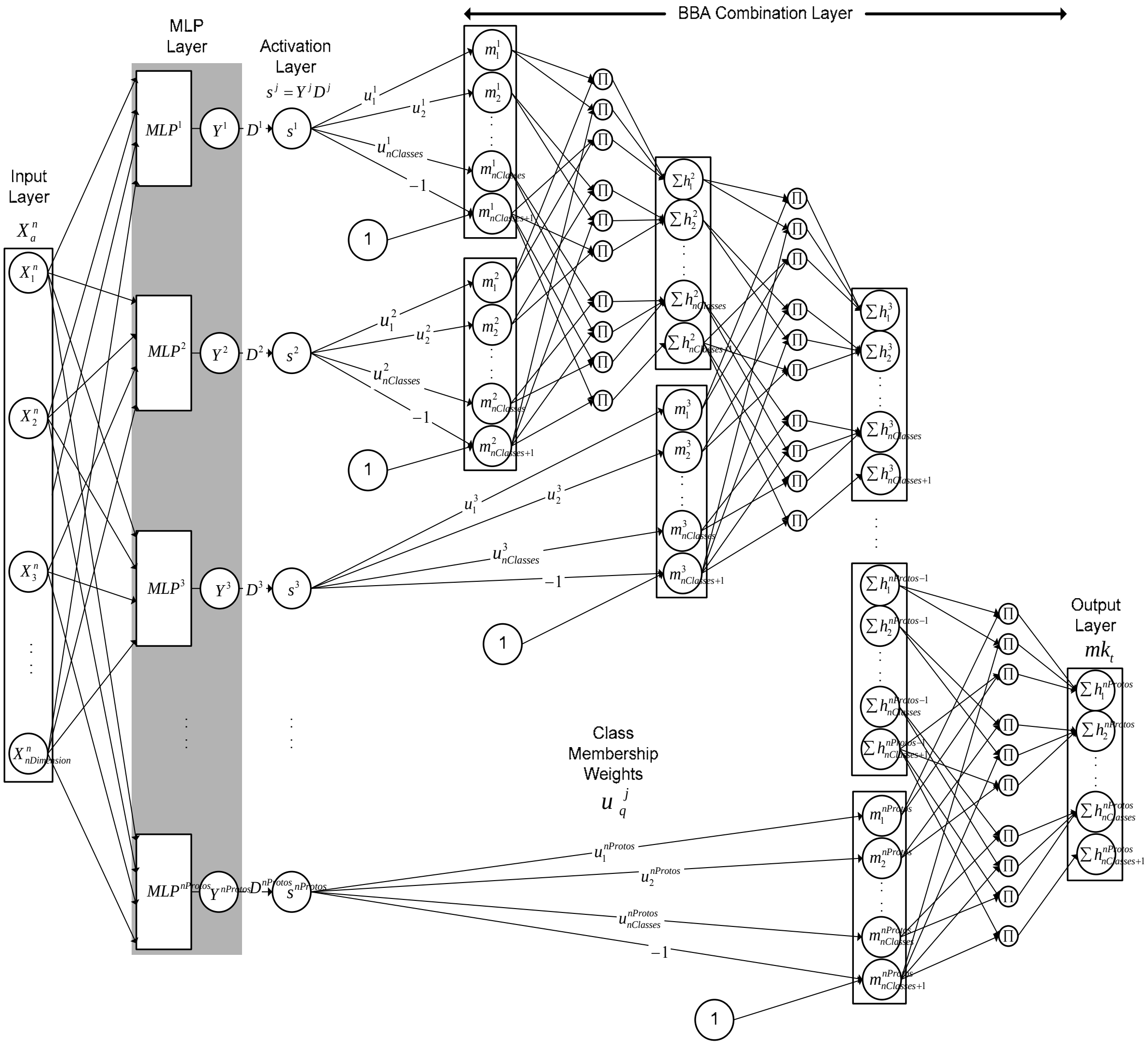


Figure 5: The architecture of the hybrid system (MLP + DBNN).

Table 1: αMLPDS seven-fold cross validation results of the RS126 and SSPRO data sets.

Data Set	Classifier	Train Accuracy Q_3 %	Test Accuracy Q_3 %	Convergence Time (hours)
SSpro	DBNN	61.88	61.86	58.9
	αMLPDS	62.82	62.91	149.5
RS126	DBNN	61.51	61.85	25.5
	αMLPDS	62.73	62.39	38.6
	PHD		62.10	

8.1.2. Secondary Structure Prediction using Neural Network and Information Theory

Protein secondary structure prediction is a fundamental step in determining the 3D structure of a protein. In this paper, a new method for predicting protein secondary structure from amino acid sequences has been proposed and implemented. Cuff and Barton 513 protein data set is used in training and testing the prediction methods under the same hardware, platforms, and environments. The newly developed method utilizes the knowledge of the GOR-V information theory and the power of the neural networks to classify a novel protein sequence in one of its three secondary structure classes (helices, strands, and coils). The newly developed method (NN-GORV-I) is improved further by applying a filtering mechanism to the searched database and, hence, named NN-GORV-II.

Considering the nature of the composition of protein secondary structure, it is worth mentioning that prediction accuracy of about 50% is worse than random guess, since the coils composition of most databases is about 50% of the whole database (Baldi et al., 2000). Figure 6 represents a histogram that elucidates the performance (Q3) of the seven prediction methods from the 50% level and above. Figure 1 shows clearly that NN-I and GOR-IV methods predicted most of the 480 proteins at prediction levels near the 55% to 65% levels, while NNGORV-II, NN-GORV-I, PROF, NNII, and GOR-V methods predicted many of the 480 proteins around the 85% to 90% levels. The new methods, NNGORV-II and NN-GORV-I, can be observed to predict many proteins at the 95% to 100% levels, compared to other methods. This analysis of Figure 3 suggests that generally, NN-I and GOR-IV methods are less performance predictors; NN-GORV-II and NN-GORV-I are high performance predictors; and the remaining three methods performed between these two levels.

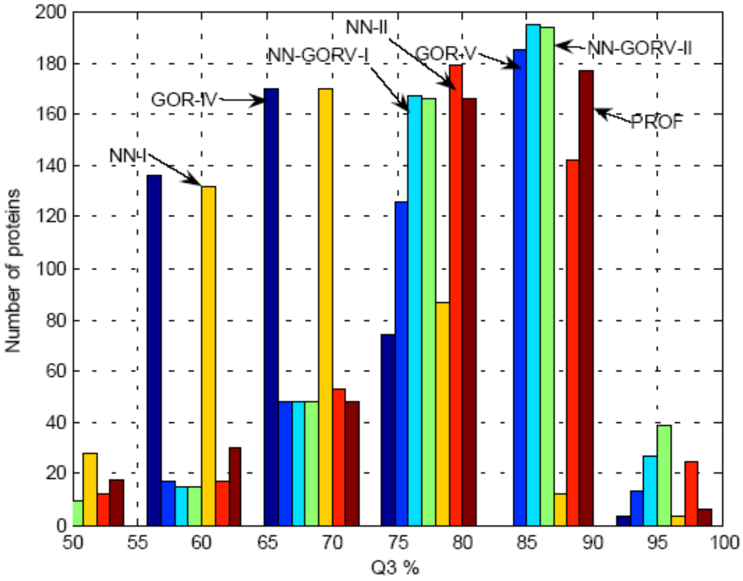


Figure 6. Histogram showing the Q3 performance of the seven prediction methods

The experimental results indicate that our prediction methods performed significantly better in terms of accuracy and quality compared to other prediction methods studied in this work. The NNGORV-II outperformed the GOR-V methods by 8.7% in Q3 accuracy and the neural networks method (NN-II) by 6.9%. Also, the SOV measure and the Mathews Correlation Coefficients (MCC) showed that NN-GORV-II significantly outperformed the other prediction methods. Our future work involves further refinement of our approach and conducting additional experimental validation.

8.1.3. Protein homology detection

The amount of information being churned out by the field of biology has jumped manifold and now requires the extensive use of computer techniques for the management of this information. Protein classification is a principal information management and retrieval task. In this research, we presented, applied and analyzed effective learning methods for detecting remote protein homology. We first propose a method uses a discriminative framework and in particular Support Vector Machine (SVM) derived from a generative statistical model.

We presented Hidden Markov Model combining scores (HMMcs) approach. Six scoring algorithms are combined as a way of extracting features from a protein sequence. We evaluated the proposed method by applying it to assign protein sequence to a predefined family. Experimental results validate the effectiveness of the proposed method.

Moreover, we applied a novel approach for automated protein homology detection based on a special kernel that computes the similarity between protein sequences in a feature space generated by all subsequences of bounded length. Figure 7 shows the new kernel in general framework for detecting remote protein homology which we call as string kernel in conjunction with SVM. This kernel is tested on SCOP database.

Figure 8 shows the overall performance of the various methods. This figure show that SVM-SK approach delivers comparable performance in detecting protein homology. The method outperformed all the generative based methods and it’s comparable to SVM-Fisher method which is among the most accurate. Moreover, SVM-SK is proved to outperforming the SVM-Fisher method in classifying some of the families.

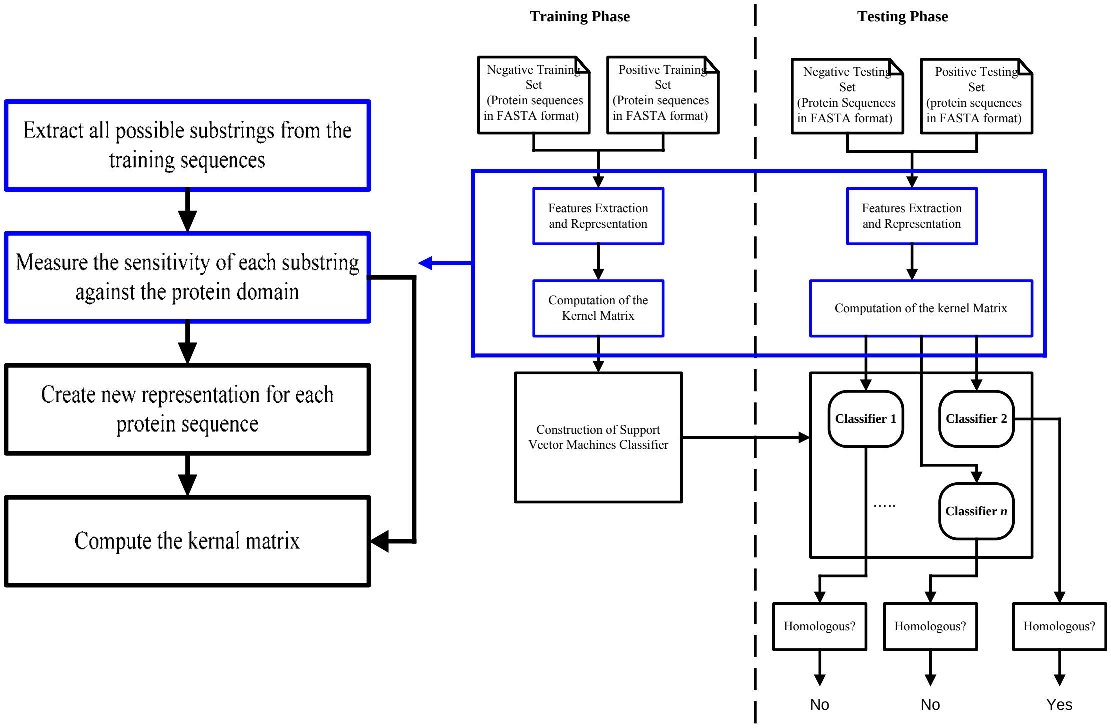

Training Phase
Testing Phase
Negative Training Set (Protein sequences in FASTA format)
Positive Training Set (Protein sequences in FASTA format)
Negative Testing Set (Protein Sequences in FASTA format)
Positive Testing Set (protein sequences in FASTA format)
Extract all possible substrings from the training sequences
Measure the sensitivity of each substring against the protein domain
Create new representation for each protein sequence
Compute the kernal matrix
Features Extraction and Representation
Computation of the Kernel Matrix
Features Extraction and Representation
Computation of the kernel Matrix
Construction of Support Vector Machines Classifier
Classifier 1
Classifier 2
Classifier n
.....
Homologous?
Homologous?
Homologous?
No
No
Yes

Figure 7: Kernel String method (SVM-SK) in general framework for detecting remote protein homology.

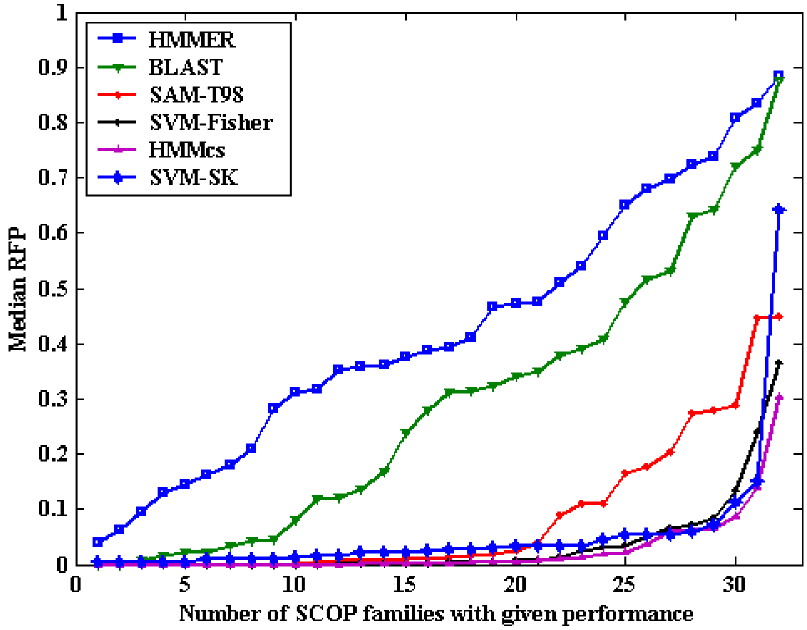


Figure 8: Overall performance in terms of rate of false positive (RFP) of protein remote homology detection methods on the 33 test families.

8.2. Protein Function Prediction

8.2.1. Protein-protein interaction

Since proteins collaborate or interact with one another for a common purpose, it is possible to infer and deduce functions of a protein through the functions of its interaction partners. However, the interactions data that have been identified by high-throughput technologies are known to yield many false positives. Therefore, it is proposed in this research to computationally predict protein-protein interactions (PPI) using Support Vector Machines. Then combine both the experimental and the computationally predicted protein-protein interactions data to construct a reliable dataset that can be used for the prediction of proteins functions prediction. Bayesian Network will be used to develop the protein functions prediction system.

One possibility to computationally predict interacting proteins is by correlating experimental data on interaction partners with computable or manually annotated features of protein sequences using machine learning approaches, such as support vector machines (SVM) and data mining techniques, such as association rule mining. The most common sequence feature used for this purpose is the protein domains structure. The motivation for this choice is that molecular interactions are typically mediated by a great variety of interaction domains. It is thus logical to assume that the patterns of domain occurrence in interacting proteins provide useful information for training PPI prediction methods.

Another sequence feature that has been used to predict PPI in-silico is the hydrophobicity properties of the amino acid residues. Therefore, in this research we proposed a better and more realistic method to construct the negative interaction set. Then we compared the using of domain structure and hydrophobicity properties as the protein features for the learning system.

The main results of our experiments are summarized in Table 1. When only domain structure was considered as the protein feature without information on domain appearance score, the cross-validation accuracy and ROC score were respectively 80.848% and 0.8759. When domain scores were included the cross-validation accuracy and ROC score were decreased to 76.397% and 0.8190 respectively. This result indicates that it is not important to include the domains score information to the feature representation of the protein pairs. It is informative enough to consider only the existence of domains structure in the protein pairs. It is important here to note that the performance of the prediction algorithm is far better than an absolute random approach which has ROC score of 0.5. This indicates that the difference between interacting and non-interacting protein pairs can be learned from the available data.

In the case of hydrophobicity dataset, the cross-validation and ROC sore were respectively 80.677% and 0.7914. We can see from these results that both domain dataset and hydrophobicity dataset have similar cross-validation accuracy of about 80%.

However, ROC score indicates that domain structure is noticeably better than hydrophobic properties (Figure 4). Another aspect is the running time for both features. Clearly, when domain structure used, the data set is much smaller than the data set for the hydrophobic properties. Consequently, the running time required for domain structure training data is much less than the running time required for the hydrophobic training data as shown in Table2.

Table 2. The overall performance of SVM for predicting PPI using domain and hydrophobicity features.

Experiment Feature	Accuracy	ROC score	Running time
Domain	79.4372 %	0.8480	34 seconds
Domain Scores	76.397 %	0.8190	38 seconds
Hydrophobicity	78.6214 %	0.8159	20,571 seconds (5.7 hours)
Hydrophobicity Scales	79.1375 %	0.7716	34,602 seconds (9.6 hours)

8.2.2. Gene Network Inference

Almost all cells in an organism contain the same genes. To understand how an organism functions at the molecular level, we must understand which genes are expressed, when are they expressed and what are the expression levels. Genes do not act alone. A gene must interact with other genes in order to function. Genes interact via gene expression which is a cellular process by which genetic information flows from genes to mRNA to proteins. The complex interaction between genes can be represented in a gene network. In a gene network, some genes will be over expressed whereas others will be under expressed. Using microarray technology, we can capture gene expression data for tens of thousands of genes simultaneously. Figure 9 is a representation of the steps involved in obtaining microarray gene expression data.

Techniques such as Bayesian network (BN), Differential Equation and others have been used to infer gene network from the gene expression data. From the literature, BN is currently actively used to infer gene network because it is able to handle the stochastic aspects and noisy measurements of gene expression data. . A Bayesian network is made up of two parts; a Directed Acyclic Graph (DAG) and a Conditional Probability Distribution (CPD). In a DAG, a node represents a random variable (gene expression) and an edge represents the influence between the nodes. An example of a Bayesian network is in Figure 10.

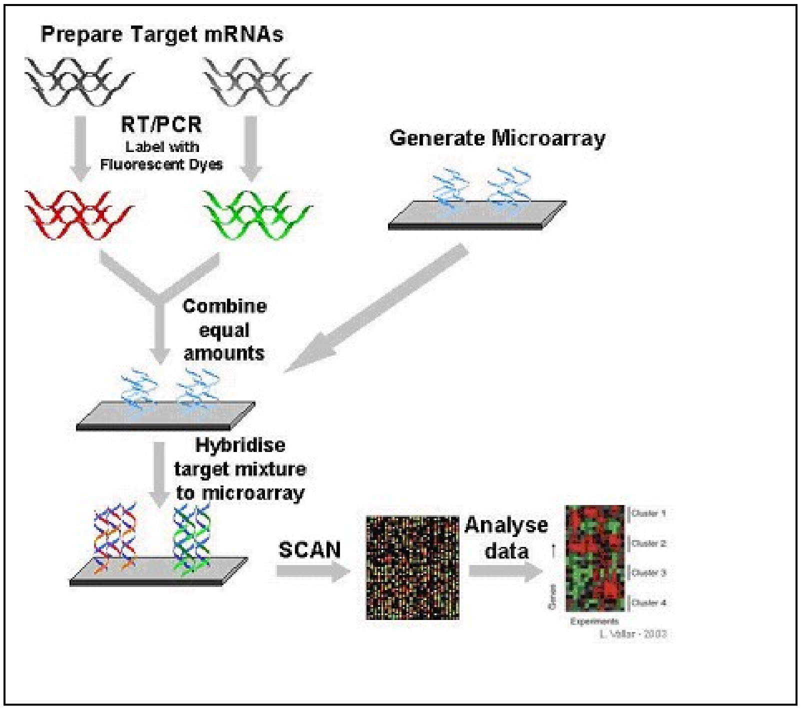
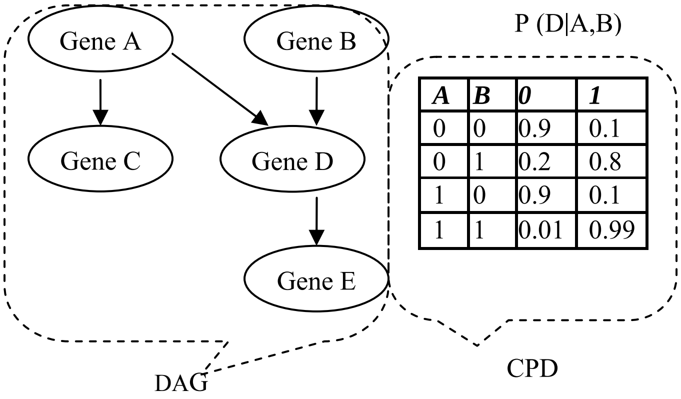


Figure 9: Steps in Obtaining Microarray Gene Expression Data



A	*B*	*0*	*1*
0	0	0.9	0.1
0	1	0.2	0.8
1	0	0.9	0.1
1	1	0.01	0.99

Figure 10: Bayesian Network consisting of DAG and CPD

The CPD represents the set of parameters for the DAG. In Figure 2, the CPD for gene D is shown. The CPD contains parameters for the probabilities of gene D given the probabilities of its parents, gene A and gene B. The process of inferring gene network using Bayesian Network is as follows, given a set of expression data D in the form of independent values for X, learning techniques for Bayesian Networks allow one to induce the network, that best matches D. The technique relies on a matching score to evaluate those networks with respect to the data and searches for the network with the optimal score.

Based on the literature, Bayesian network is the most used technique. Using Bayesian network alone is not sufficient to produce a reliable gene network. Combining BN with other techniques will produce a better network in less time. Using multiple sources of data may also do the same. The trend of research has also focused in unravelling small sized networks from a handful of genes and not focusing on unravelling the complete network from thousands of genes due to time needed to model the network. The time to process increases exponentially as the number of genes increases. Currently we are in the process of improving various aspect of the technique and the final outcome of this research is as shown in Figure 11.

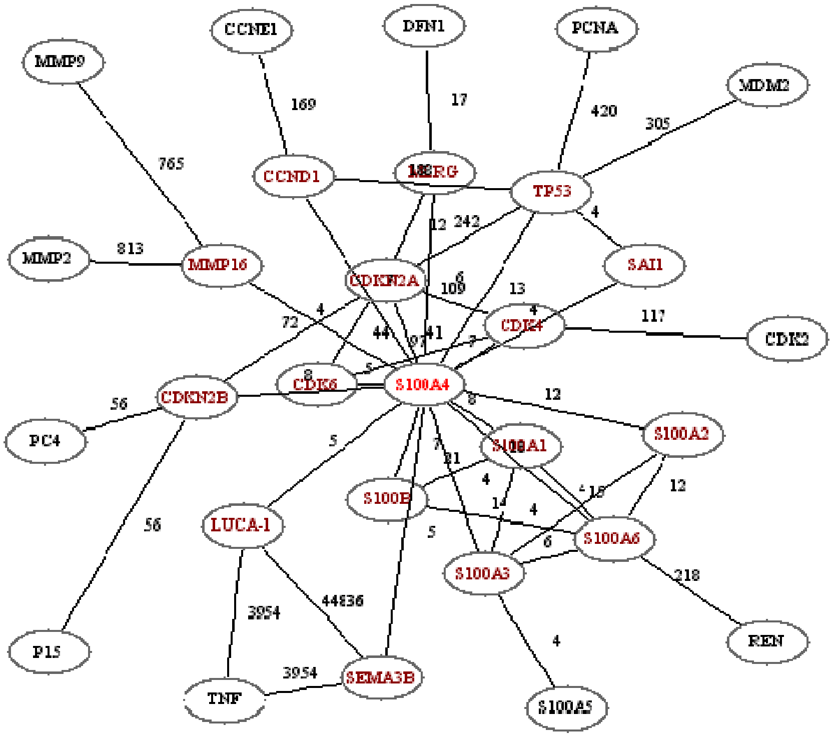


Figure 11: Gene network from Bayesian Network

8.2.3. Gene classification using hybrid GA

Advancement in gene expression technology offers the ability to measure the expression levels of thousand of genes in parallel. Gene expression data is expected to significantly aid in the development of efficient cancer diagnosis and classification platforms. Key issues that need to be addressed under such circumstances are the efficient selection of a small subset of genes that might profoundly contribute to disease identification from the thousand of genes measured on microarrays that are inherently noisy. This research deals with finding a small subset of informative genes from gene expression data which maximizes the classification accuracy.

This research proposed a hybrid of Genetic Algorithm and Support Vector Machine classifier (New-GASVM) for selecting an optimal small subset of informative genes and classifying the optimal subset. The GA will select the subsets of features and then the SVM classifier evaluates the subsets during a classification process. The result of the classification is used for the fitness value of GA. The key steps in the hybrid GA and SVM classifier are shown in Figure 12.

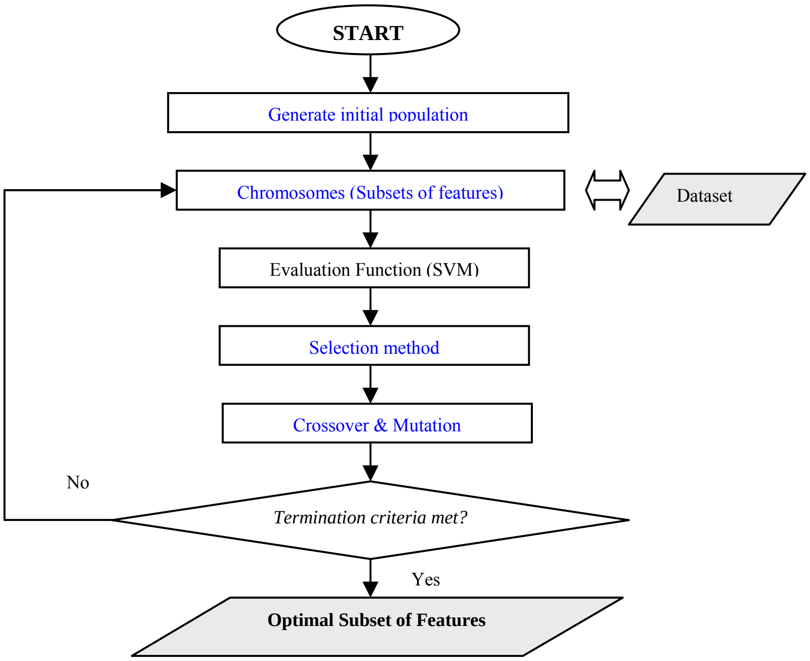

START
Generate initial population
Chromosomes (Subsets of features)
Dataset
Evaluation Function (SVM)
Selection method
Crossover & Mutation
No
Termination criteria met?
Yes
Optimal Subset of Features

Figure 12: Flowchart of hybrid GA and SVM classifier.

Two benchmark data sets, namely leukemia cancer and colon cancer were used to evaluate the useful of the approach for small and high dimension data. Table 3 displays the results of benchmarking of New-GASVM and the best results of latest Leukemia Cancer dataset outcome. Based on the LOOCV and the accuracy test, it was noted that New-GASVM performances were equal to current methods produced by Xu et al., Cho and Won, and Nguyen and Rocke. The results in Table 2 shows that the New-GASVM achieved 98.3871% accuracy and performed better than previous methods by using 30 selected genes.

Table 3. Benchmark of New-GASVM performances and current best of previous methods on Leukemia Cancer dataset.

Method	Author (Year)	Number of Selected Genes	Accuracy (%)	
			LOOCV	Test
New-GASVM		40	100	97.0588
GASVM		3568	94.7368	85.2941
SVM		7129	94.7368	85.2941
ART-NN	Xu et al., 2002	10	100	97.0588
LD	Nguyen and Rocke, 2003	50	100	97.0588
MN	Su et al., 2002	10	100	90.0
SVM	Furey et al., 2000	25	100	94.1177
	Mukherjee, 2001	49	100	100
GAWV	Liu et al., 2001	29	94.7368	88.2353
WV	Golub et al., 1999	50	94.7368	85.2941
	Slonim et al., 2000	50	94.7368	85.2941

Note:
Methods in boldface were experimented in this research. Best results shown in shaded cells.
GASVM : Hybrid GA with SVM
New-GASVM : Hybrid GA with SVM (proposed algorithm)
SVM : Support Vector Machine
WV : Weight Voting
GAWV : Hybrid GA with Weight Voting
ART-NN : Adaptive Resonance Theory Neural Network
LD : Logistic Discriminant
MN : Modular Neural Network

Table 4. Benchmark of New-GASVM performances and current best of previous methods on Colon Cancer dataset.

Method	Author (Year)	Number of Selected Genes	LOOCV Accuracy (%)
New-GASVM		30	98.3871
GASVM		1009	90.3226
SVM		2000	85.4839
JCFO	Krishnapuram et al., 2004	26	96.7742
LD	Nguyen and Rocke, 2003	50	93.5484
GAWV	Liu et al., 2001	30	91.9355
ART-NN	Xu et al., 2002	50	90.3226
SVM	Furey et al., 2000	1000	90.3226
	Ben-Dor et al., 2000	2000	77.4194
NBB	Ben-Dor et al., 2000	2000	80.6452
AB	Ben-Dor et al., 2000	2000	72.5807

Note:
Methods in boldface were experimented in this research. Best results shown in shaded cells.
GASVM : Hybrid GA with SVM
New-GASVM : Hybrid GA with SVM (proposed method)
GAWV : Hybrid GA with Weight Voting
ART-NN : Adaptive Resonance Theory Neural Network
LD : Logistic Discriminant
SVM : Support Vector Machine
NBB : Nearest Neighbor
AB : Ada-Boost
JCFO : Joint Classifier and Feature Optimization

The results of the gene expression classification demonstrated that our proposed method (New-GASVM) performed better than the original (GASVM) and the previous methods. The informative genes from the experiment results proved to be biologically plausible when compared with the biological results produced from biologist and computer scientist researches.

8.2.4. Gene classification using model-based reasoning

Recent introduction of microarray technology allows researchers to monitor thousands of gene expression levels in a microarray experiment. Classification of tissue samples into tumour or normal is one of the applications of microarray technology. When classifying tissue samples, gene selection plays an important role. In this research, some existing gene selection techniques are studied and better gene selection techniques are proposed and developed.

Figure 13 illustrates the flowchart for our proposed approach, namely model-based clustering procedure. Given the data, the chosen model and number of clusters, *c*, agglomerative model-based clustering will initialize the clusters. Bayesian Information Criterion (BIC) is computed and the highest BIC among different number of clusters and models is saved. After obtaining the clusters, cluster quality is assigned to each cluster. The final result is the number of clusters, *c*, the best model, the estimated parameters which fits the data best and the cluster quality for these parameters.

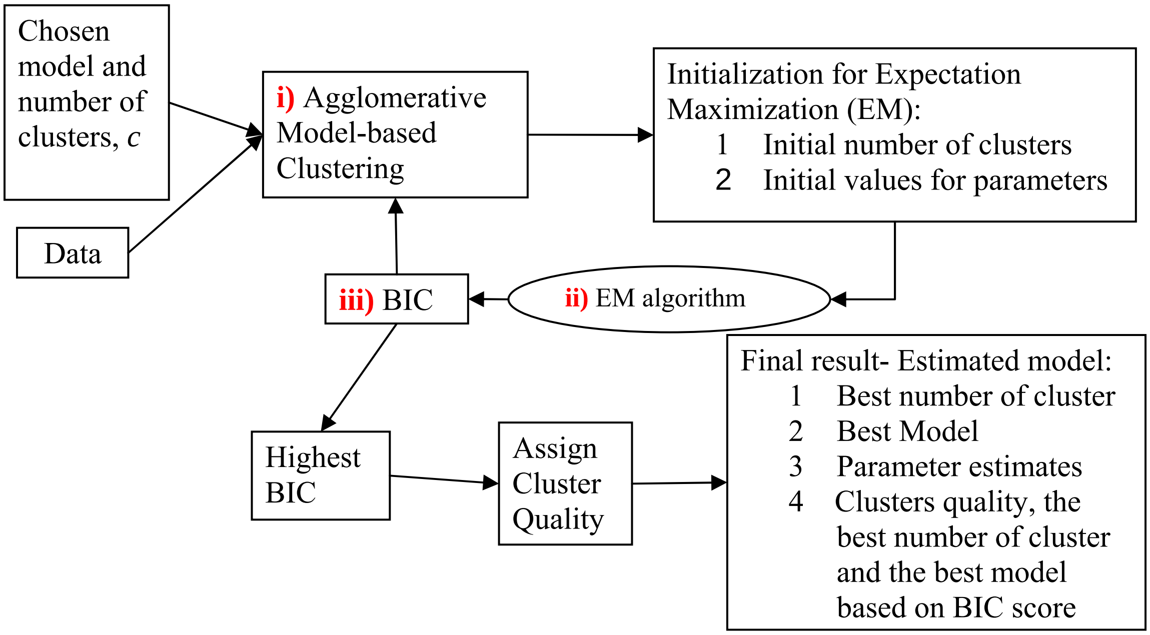

Chosen model and number of clusters, c
Data
i) Agglomerative Model-based Clustering
Initialization for Expectation Maximization (EM):
1 Initial number of clusters
2 Initial values for parameters
iii) BIC
ii) EM algorithm
Highest BIC
Assign Cluster Quality
Final result- Estimated model:
1 Best number of cluster
2 Best Model
3 Parameter estimates
4 Clusters quality, the best number of cluster and the best model based on BIC score

Figure 13: Flowchart illustrates the model-based clustering procedure.

Various *k*-means clustering algorithms and model-based clustering algorithms are proposed to group the genes. Support Vector Machine (SVM) and *k*-nearest Neighbour (*k*-nn) are used for the classification purposes. Receiver operating characteristic (ROC) score is used to analyze the results. Colon data with 2000 genes and 62 tissue samples is used for the testing. Figure 14 shows the classification performance for the results from previous techniques and best combination of the third gene selection technique (using model-based clustering).

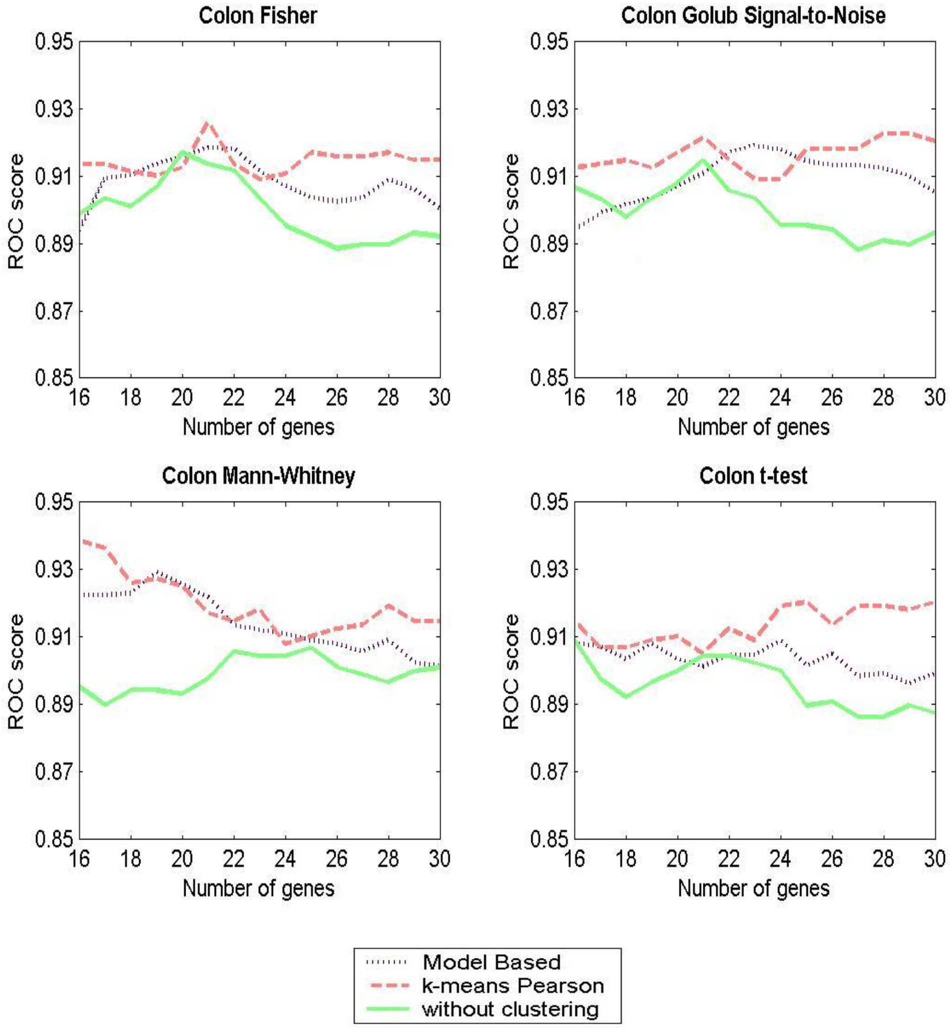


Figure 14: Classification performance for the results from previous techniques and best combination of the gene selection using model-based clustering algorithm.

Figure 14 shows that the classification performance for the proposed technique is higher compared to the previous techniques where no clustering algorithm is involved. This shows that the clustering algorithm helps in selecting more informative genes. Highest ROC score recorded from the experiments achieved 0.95, corresponding to five misclassifications. This should be of significant value for diagnostic purposes as well as for guiding further exploration of the underlying biology.

8.3. Automatic Clustering of Gene Ontology by Genetic Algorithm

Gene Ontology has been used widely by many researchers for biological data mining and information retrieval, integration of biological databases, finding genes, and incorporates knowledge in the Gene Ontology for gene clustering. However, the increase in sizes of the Gene Ontology has causes problems in maintaining and processing them. One way to obtain their accessibility is by clustering them into fragmented groups. Clustering the Gene Ontology is a difficult combinatorial problem and can be modeled as a graph partitioning problem. Additionally, deciding number *k* of clusters to use is not easily perceived and is a hard algorithmic problem. An approach for solving the automatic clustering of the Gene Ontology is proposed by incorporating cohesion-and-coupling metric into a hybrid algorithm consisting of a genetic algorithm and split-and-merge algorithm.

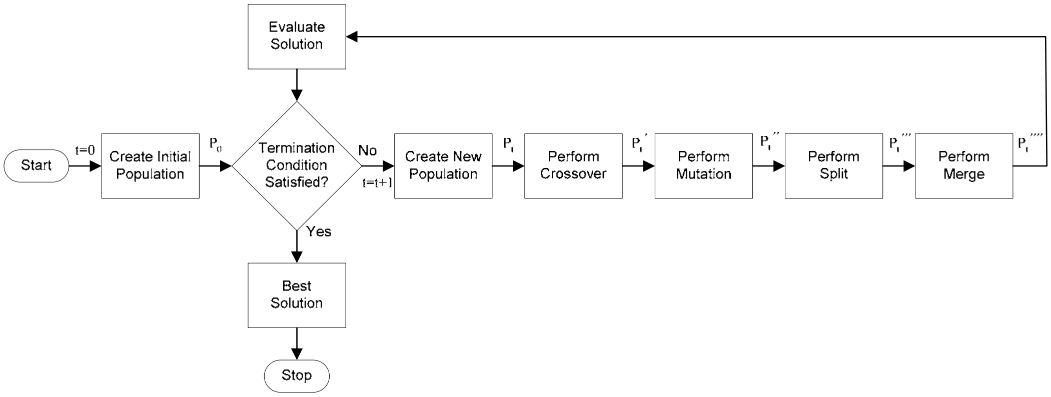


Figure 14: The overall methodology.

Function for merging clusters

```
Algorithm Merge (x, $k_{min}$);
Input: x (a chromosome), $k_{min}$ (a minimum number of clusters)
Output: x′ (a  modified chromosome)
begin
  n := x.NoOfClusters();
  if n $\neq$ 1 then
    for p := 1 to n do
      for q := p + 1 to n do
        if x.NoOfClusters() > $k_{min}$ and x.Coupling(p, q) $\neq$ 0 then
          $x_{merge}$ := x;
          for r := 1 to $x_{merge}$.Length() do
            if $x_{merge}$.Gene(r) = q then $x_{merge}$.Gene(r) := p; end-if
          end-for
          if $x_{merge}$.QOC($C_p$) > x.QOC($C_p$, $C_q$) and $x_{merge}$.DependencyIndex($C_p$) < $I_{max}$
            then x := $x_{merge}$; end-if
        end-if
      end-for
    end-for
  end-if
End
```

Figure 15: Merge and split algorithm for graph balancing.

Computational Result:

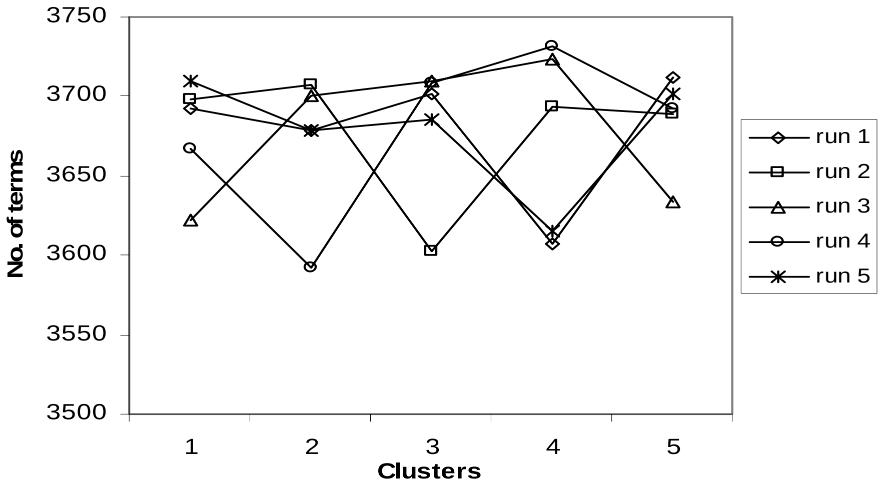


Figure 16: Clustering utilization of 5 runs

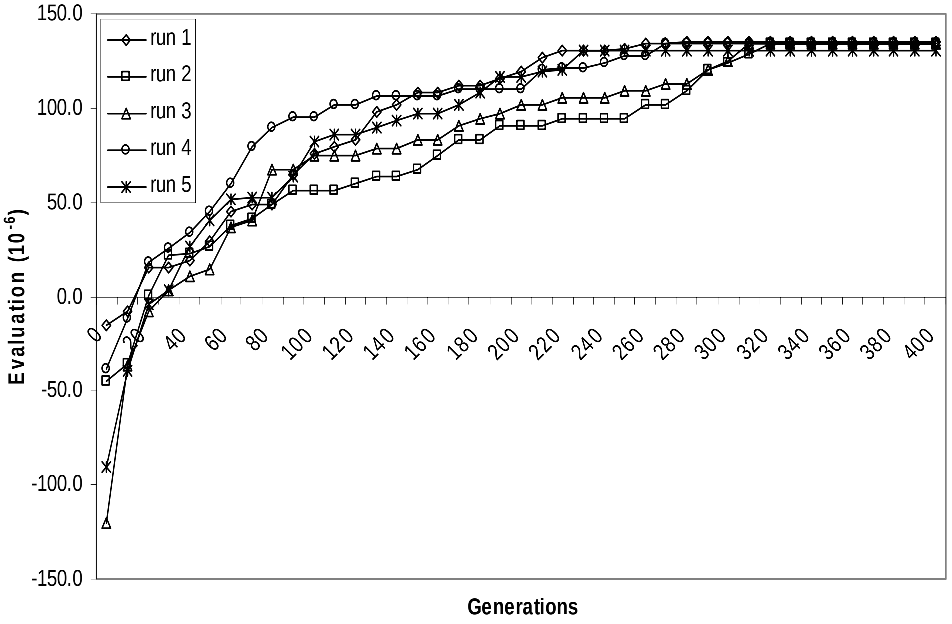


Figure 17: Evolution of 5 runs

8.4. Pathway prediction

The genome database of *Saccharomyces cerevisiae* from GenBank as well as MetaCyc pathway reference databases have been used for predicting metabolic pathways. MetaCyc is a Pathway/Genome Database (PGDB) which describes the genes, proteins, and metabolic reactions and pathways interrelated with each others. The annotated genome database contains of 1909 genes and 5 set of chromosome data.

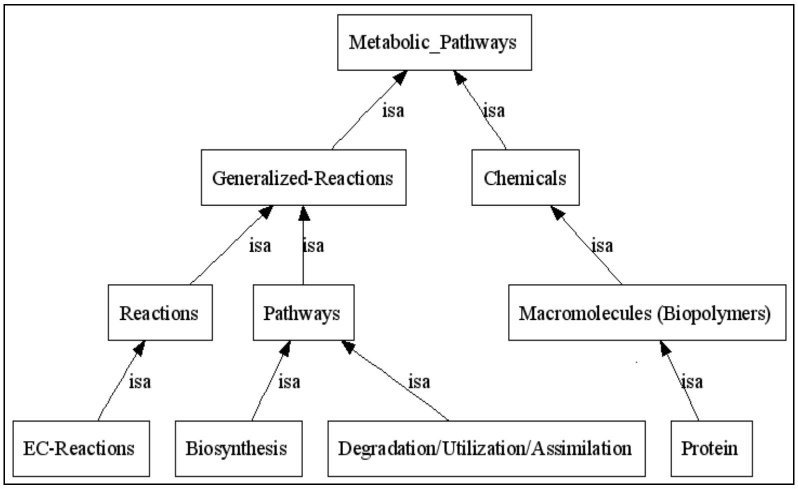


Figure 18: Extended hierarchy for metabolic pathways ontology

Based on this experiment, it did predict the presence of the pathways that are expected to exist from available pathway databases. Table 5 shows the pathways and biological information that can be predicted from *Saccharomyces cerevisiae* annotated genome data.

Table 5: Pathways predicted from S.cerevisiae data

Pathways	41
Enzymatic Reactions	228
Transport Reactions	1
Protein Complexes	0
Enzymes	69
Transporters	1
Compounds	249

This result indicates that ontology and problem solving method are capable to provide the meaningful data and also able to solve the problems mention above. The integration of ontology and problem solving method as a possible solution for predicting metabolic pathway which states that representing knowledge for solving this problem is strongly affected by the nature of the problem and the inference strategy to be applied to the problem.

9. Conclusion

The technology revolution has changed the society from agrarian society to industrial society to information society and to knowledge-based society. In the KBS, role of research becoming more important and at the same time changed the paradigm of education, teaching and learning. Research and innovation play major and important roles in acquiring the knowledge in KBS. Digital and biotechnology revolution has created new a field called bioinformatics which is responsible for generating accumulating life sciences knowledge for KBS. Since biotechnology may change many aspects of future society, acquiring this knowledge become inevitable. Bioinformatics research covering several areas such as structure and function prediction and modeling biological process and system are a few initial steps for moving toward KBS and KBE.