Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest

Shaurya Jauhari and S.A.M. Rizvi

Abstract—An understanding towards genetics and epigenetics is essential to cope up with the paradigm shift which is underway. Personalized medicine and gene therapy will confluence the days to come. This review highlights traditional approaches as well as current advancements in the analysis of the gene expression data from cancer perspective. Due to improvements in biometric instrumentation and automation, it has become easier to collect a lot of experimental data in molecular biology. Analysis of such data is extremely important as it leads to knowledge discovery that can be validated by experiments. Previously, the diagnosis of complex genetic diseases has conventionally been done based on the non-molecular characteristics like kind of tumor tissue, pathological characteristics, and clinical phase. The microarray data can be well accounted for high dimensional space and noise. Same were the reasons for ineffective and imprecise results. Several machine learning and data mining techniques are presently applied for identifying cancer using gene expression data. While differences in efficiency do exist, none of the well-established approaches is uniformly superior to others. The quality of algorithm is important, but is not in itself a guarantee of the quality of a specific data analysis.

Index Terms—Association rules, cancer, classification, clustering, data mining, gene expression data, gene therapy, epigenetics, next generation sequencing, clinicopathology

1 Introduction

Cancer is a major cause of all the natural mortalities and morbidities throughout the world. Nearly 13 percent of deaths caused are due to cancer [1]. It is a disease getting constantly challenged by many eminent and premier researchers. Although, some advancements have been reported for its clinical prevention and cure and there has been a noticeable decline in the lives' lost [49], but they are not quite adequate [50]. The lack of affordable treatment and early detection is the crux of this hostile situation. It is becoming hard for the perennial biomedical scientists and researchers to exorcize this daemon.

The growth in a body is observed when the division and multiplication of cells takes place. When the appropriate division levels have been achieved, the process is deactivated. In an unusual scenario however, cells continue to replicate and form lumps in the body, although it commences with a paltry entity. Cancer is an abnormal and uncontrollable growth of cells in the body that turn malignant. This is not to be confused with tumors. Even a tumor is an abnormal growth of cells, but it can be classified as (non-cancerous) benign and malignant, the latter one causing cancer. It is noteworthy that all cancers are tumors, but the reverse is not true. Cancer can develop in almost any organ or tissue, such as the lung, colon, breast, skin, bones, or nerve tissue. Various types of cancer have been identified namely, breast cancer, colon cancer, lung cancer, brain cancer, cervical cancer, kidney cancer, liver cancer, leukemia, Hodgkin's lymphoma, non-Hodgkin's lymphoma, ovarian cancer, skin cancer, thyroid cancer, uterine cancer, and testicular cancer [2]. Cancer causes quick dissemination of cells and a cancer type can fortify and extend to another one if not treated appropriately.

There are many causes of cancers, including:

- Benzene and other chemicals.
- Drinking excess alcohol.
- Environmental toxins, such as certain poisonous mushrooms and a type of poison that can grow on peanut plants (aflatoxins).
- Excessive sunlight exposure.
- Genetic problems.
- Obesity.
- Radiation.
- Viruses.

However, the cause of many cancers remains unknown. Apart from internal (genetic) causes, there are certain environmental and external factors too that participate in cancer formation within an organism, viz. environmental toxins, adulterated food intake, air pollution, and irregular lifestyle; (share as depicted in the Fig. 1). These can be categorized under *epigenetics*. Epigenetics is an unignorable issue to be addressed by the biomedical community.

Symptoms of cancer depend on the type and location of the cancer. For example, lung cancer can cause coughing, heavy breathing, chest pain, etc. Colon cancer often causes diarrhoea, constipation, dysentery, and blood in the stool. Some cancers may not have any symptoms at all. In certain cancers, such as pancreatic cancer, symptoms often do not start until the disease has reached an advanced stage. Cancer progression can be aggressive or benign. That corresponds to the suitable treatment required for ailing the cancer [2].

• *The authors are with the Department of Computer Science, Jamia Millia Islamia, Jamia Nagar, Okhla, New Delhi 110025, India. E-mail: shaurya126906@s.jmi.ac.in, samsam_rizvi@yahoo.com.*

Manuscript received 11 Nov. 2013; revised 6 Feb. 2014; accepted 10 Mar. 2014. Date of publication 17 Mar. 2014; date of current version 5 June 2014. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2014.2312002

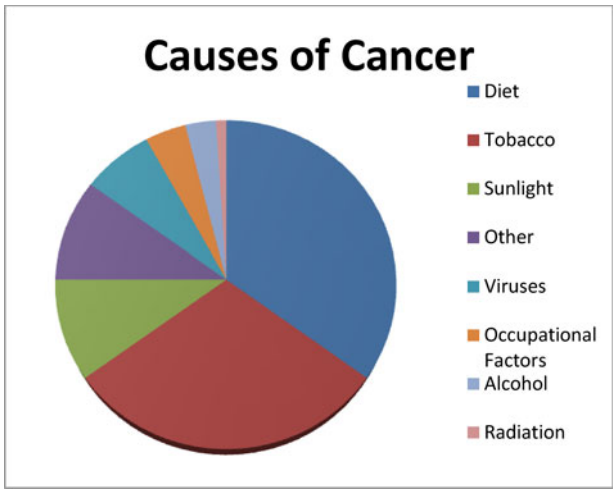

Fig. 1. A pie chart depicting various factors contributing to the initiation and growth of the disease.

There exists a *self-correcting mechanism* in our body. In diseases like cancer, the misarrangements in the underlying genetic coding aggravates to an extent, wherein this mechanism fails to rectify it. This scenario is vital to identify the malignancy state and decipher the inter-gene collaboration.

2 Mathematics in Nature: An Intuitive Certitude

Mathematics is known to be an indispensable part all sciences. It forms the edifice of all existences being a formidable aegis that "holds" all parts together. One can have certain propensity towards it and more when going through John A. Adam's texts [28], [29]. Docile is the symmetry found in vegetation, anatomy of living creatures, shapes of heavenly bodies: planets being spherical and much so their orbits, to name a few. Mathematical modeling of any natural phenomena and its inherent dogma at core, which ensures it compositional as well as physical traits can be extremely expedient towards developing an understanding. A mathematical model is a feat if it fits the known data and makes accurate predictions for the future, as rendered in the Fig. 2 [28]. A snowball is defined to grow in size and attain an almost intermediary shape between a circle and a sphere as it rolls through ice. However, external factors like intensity of sunlight, heat produced due to friction/resistance, dymanic surface area of the ball, etc., are the variates that contribute to the problem, profoundly (Adam, 2006). The author also registers that all mathematical models are flawed to some extent owing to the inappropriate presumptions made during their construction.

The aesthetic of all natural and physical phenomena is brought to life once the mathematical undergird is realized.

"*Mathematics is to nature as Sherlock Holmes is to evidence*."

With a compendium of suggestions to consult, it renders highly probabilistic scenario that the gene expression data won't be an exception. Studies have shown that mathematical and statistical models can be built around it and they are seminal to identify biomarkers. The only apprehension is the uncertainity attached to it which is subjected to validation to establish the baseline parameters. All the models drawn till date, are rarely "euphoric", and although get us close to finding targets, but still are far off.

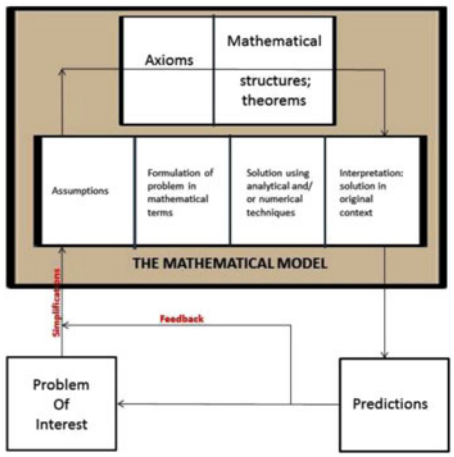

Fig. 2. Flowchart of mathematical modeling [28].

3 Epigenetics

Literally speaking, "epi" stands for "on top of" and epigenetics is on top of genetics. Governing factors for the gene expression and protein building, explicit to the inherent DNA code, delineates epigenetics. The environment and our lifestyle can significantly direct our genetic behavior and even that of our kids. The multicellular organisms have optimally identical underlying code, yet they have incongruent phenotypes. This extraneous information that renders this disparity and uniqueness is held outside the DNA, in an extracellular memory, if you will.

There is also a debate on the non-coding parts of the RNA that may be causative to epigenetic traits. These are non- functional to protein production and DNA coding is not held here.

3.1 Why Your DNA Isn't Your Destiny

The more our understanding towards the human genome is concreted, the vastly it is realized that genes and DNA aren't really orchestrating it all [30]. An epigenome is conceived to be a filter of biochemical reactions that controls gene expression. It can hold as much clues to the gene regulation and transcription factor mechanism, as otherwise. By unearthing epigenome, scientists can move closer towards precise medication and better understanding disease contrivance.

By mapping the epigenome and linking it with genetic and hereditary traits, scientists think that they can unveil the mysteries of many genetic disorders that get transferred through ages. Also, they can elucidate clues about the fact that why two twins cannot be the same. Epigenome changes with the environment and deeply associated with disease and development. Ehrlich's research in 1983 linked human cancer with epigenome, but then it was greatly flouted. In a current viewpoint, he admits that though in an embryonic stage, epigenomics has a long way to go.

John Cloud's article in the famous Time Magazine [31], (dated January 18, 2010) reports that it was a sacrilegious idea. No matter what we do to our health or lifestyle, our

DNA composition remains the same. In 1986, a pioneering research published by *Lancet* revealed that a malnourished pregnant woman is likely to give birth to a child that will have high probability to inhibit cardiovascular disorder during adulthood. This was also alleged to start before pregnancy. Was something else affecting the profile of the new born? Isn't this nebulous to envisage that after the kids are born, the DNA machinery is to start afresh? Although the child may adapt to the new environment other than his/her parents and that could change the epigenome, but also lots of his genes will partly or fully hold hereditary information passed from the parents. This establishes that DNA and epigenome both collectively oversee an individuals genotype and phenotype.

Kaati et al. [33] illustrate through their findings in Overkalix, Sweden that during slow growth period of father and grandfather respectively, child will exhibit low cardiovascular disorder tendency owing to low food availability and high diabetic ailment due to high food resource. This was seminal in understanding hereditary linkages to inherent disease transitions and genetic mutations leading to disorders. In 2004, the Food and Drug Administration (FDA), US, approved an epigenetic drug, Azacitidine, which could repress the expression of genes causative of myelodysplastic syndromes (MDS). So, with due regards to Charles Darwin, epigeneticists can now model and improvise on the existing DNA lines and decipher ways to cure cancer, alzheimer's, and other likewise complex maladies, and also contain their progression and probability of occurrence down the hierarchy.

It must be mentioned here that epigenetics does not change the underlying genetic code but alters the mechanism and functionality by adding certain extraneous features to it. It should also be noted that changes due to epigenetics can be temporary or permanent, reasons still sought. The epigenome is not only vital to one's own life but also lies underneath is the idea how the coming generation will survive.

3.2 DNA Methylation: An Outlook from Gene Expression and Cancer

As previously known, epigenetics sits on top of our DNA and modifies the genetic code "superficially" to result in a distinguished functionality; the least of which can be explained by DNA Methylation.

DNA Methylation is the process of addition of a methyl group to the gene. A methyl group is fairly significant to organic chemistry and consists of one carbon atom bonded with three hydrogen atoms (CH_3). DNA Methylation can modify the gene expression, diminishing it or making it gaudier.

Das and Singal [41] portray DNA methylation as an epigenetic event that highly correlates to the regulation of gene expression. As one facet, DNA methylation exhibits direct interception with the binding sites of particular transcription factors to their promoters. Also, they are involved with the direct binding of specific transcriptors to the methylated DNA.

From the cancer perspective, malignant cells as opposed to their normal counterparts show exaggerated disturbances in their DNA [41]. This trait is a lucrative distinctive measure to study the methylation pattern-change behavior. Hypomethylation is another characteristic of the solid tumor types as cervical and prostate cancers.

In an independent study conducted by Duke University's oncologist, Randy Jirtle et al., it was revealed that on regulation of agouti gene in mice, yellow coats and propensity for obesity and diabetes were rendered, when continuously expressed. The B vitamins acted as methyl donors causing methyl group to attach with agouti genes more readily, thereby altering its expression. The study resulted in agouti mothers being able to produce children with normal weight and no diabetic trait. Later, Eva Jablonka would feed roundworms with a kind of bacteria that would now be "non-fluorescent and non-green" and shall curtail this dumpy appearance for as much as 40 generations. Similarly, fruit flies exposed to a drug called geldanamycin witness unusual growth on their eyes and this abnormality lasted up to 13 generations (no exposure to the drug was prearranged to offsprings from second to 13th generations.) Also, a critical study led by Larry Feig, a biochemist from the Tufts University, cemented that even memorizing and brain nourishment can be improved when apt environment is given. In the premises of research, Feig's team presented mice with genetic memory problems to an environment rich with toys, exercise, and extra attention. The mice showed enhanced long-term potentiation (LTP) which is a neural mechanism, significant from the standpoint of memory-formation. All of the aforementioned facts conceive that epigenetics is far from being a sideshow, overshadowed by DNA.

3.3 The Human Epigenome Project

The Human Epigenome Project (HEP) was initiated to identify and catalogue Methylation Variable Positions (MVPs) in the human genome [32]. Similar to the profile of the Human Genome Project (HGP), HEP is also private/public collaboration.

Conceived by the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/), Epigenomics AG (http://www.epigenomics.com/en/), and Centre National de Génotypage, France (http://www.cng.fr/), HEP is another scientific challenge and an envisionment of a better and healthier tomorrow.

4 Microarray Analysis of Gene Expression

A prodigy of data assimilation is the microarray technology [13]. Genes form a compendium of an organism. They impart the genotype and phenotype to it [organism]. A gene expression is the subset (part of the genetic code studied in a particular regard) of an organism's genetic codon. A microarray is a "virtual-lab" on a chip. It is a 2D array on a solid platform (usually a glass slide) that is studied for biological references. It is used to assay spots (usually called probes or reporters) that are analyzed experimentally under the microscopic vision as the findings are hard to notice by naked eyes. The popularity of the microarray technique notwithstanding, the analysis of the data is far from trivial. Essentially, the information which we capture by using DNA microarray is at the level of transcription (*origin: molecular biology).* The raw microarray data are images, which have to be transformed into *gene expression matrices-*

tables where rows represent genes, columns represent various samples such as tissues or experimental conditions, and numbers in each cell characterize the expression level of the particular gene in the particular sample. These matrices have to be analysed further, if any knowledge about the underlying biological processes is to be extracted. So the inferences and results obtained by such studies have to be validated. More refinement can be achieved in the results by working on multi-tier data. It may lead to putative regulatory signals in the genome sequences.

The microarrays can be categorized under cDNA microarrays and Oligonucleotide arrays (oligochip) [13]. Albeit both behold differences in experimentations, they share a common ground on manufacturing, target preparation, labelling and hybridization, and scanning process. cDNA microarrays and oligochips are collectively, aggregately conceived as microarrays, unless explicitly mentioned. Their *modus operandi* is similar where they draw distinction between control sample and the test sample by comparing the signal intensity ratio, when the same are subjected to superimposition. Albeit, Chinnaiyan et al. [24] (2002) showcased, while working on gene expression data on prostate cancer that aberrantly both oligochips as well cDNA chips resulted in similar outcomes. A cohort of genes was found to be consistently dysregulated in prostate cancer. This study can be extended further to other hard-cured maladies.

Due to improvements in biochemical instrumentation and automation, it has become easier to collect a lot of experimental data in molecular biology. Analysis of such data is extremely important as it leads to knowledge discovery that can be validated by experiments. Analysis of gene expression data leads to cancer identification and classification, which will facilitate proper treatment selection and drug development. Previously, the diagnosis of complex genetic diseases has conventionally been done based on the non-molecular characteristics like kind of tumor tissue, pathological characteristics, and clinical phase. The microarray data can be well accounted for high dimensional space and noise. Same were the reasons for ineffective and imprecise results. Several machine learning and data mining techniques are presently applied for identifying cancer using gene expression data.

5 Gene Therapy

A common adage held with gene therapy is the identification of an abnormal (diseased) gene and then replacing it with a healthy gene (or the healthier version of the same gene from an explicit source) to oversee its functioning. It is the mechanism of *DNA repair*. Somatic (adult kind) cells such as lymphocytes, bone marrow cells, etc. are particularly targeted in the gene therapy [35], [36]. Gene therapy promises to illuminate the loop holes with the traditional clinical methods of treating diseases. Hemophilia B and Primary ImmunoDeficiencies (PIDs) are successfully trialed with gene transfer and the acceptance rate is commendable and irrefuted. But, the complications leading to the selection of the *integrating vectors* in gene therapy have been long known. These need further addressal. Gene therapy promises to be a standard treatment for assured diseases and the important question is its position in the treatment hierarchy. The pricing and practical implementation of this therapy including the infrastructural requisites forms a critical argument, so as to reach a wider domain of patients. It so turns out that documentation of genes and their corresponding pathway occurrences can help map varied diseases as portrayed in Fig. 3.

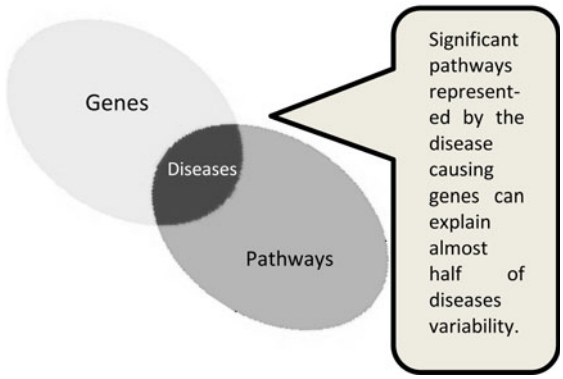

Fig. 3. Representation that genes are intrinsically a very prominent part of a pathway and these pathways again constitute a disease mechanism.

Focusing on cancer in particular, after the high approval of The Cancer Genome Atlas (TGCA) [37], under the aegis of the National Human Genome Research Institutes (NHGRI) and National Cancer Institutes (NCI), The International Cancer Gene Consortium [34], (http://icgc.org/), is an effort to match and compare the sequenced genome of the same subject at the normal and cancerous states. The irony is that we are still using medieval methods to elucidate and rectify cancer abnormalities. This concerted effort aims to stratify aberrations in 50 different cancer types. Its database also provides excellent reference for the scientists and researchers studying cancer genetics and therapeutics.

6 Objective

Efficient use of the large data sets generated by gene expression microarray experiments requires computerized data analysis approaches. It is on the utmost importance to prune irrelevant or even "overlapping" data to meticulous observations. The entwined gene expressions that subsume multiple functionalities pose a greater challenge. Since our focus is the workability and application of Data Mining and statistical techniques, we shall primarily concentrate on classification and clustering. By performing clustering, we may try to identify biological pathways, on which gene in one particular cluster lies. This is the area of Systems Biology. Also, our endeavor is to discover Gene Biomarker, by using compatible method(s) of classification. A biomarker is a representation or an indicator of the severity and presence of some diseased state in a body. Studying biomarker data (cancer oriented) computationally can help identify stages of cancer. The concentration of oncogene (diseased, cancerous gene) can be considered for study making comparisons between a control (normal) sample and a diseased sample. Single Nucleotide Polymorphisms (SNPs are also surrogate representations of markers in the reference genes). Biological pathways represent the biological reactions and

interaction network in a cell. Each reaction is identified with its enzyme, which in turn is coded by certain gene(s). By studying the patterns of gene expression in different experimental conditions, researchers can get an understanding of genes and pathways involved in biological processes. As, no gene operates in an isolated way, it is important to consider information about the complex molecular networks, orchestrating the activity of cells.

So, basically Pasquier et al. [9] identify the following questions for consideration:

- What are the functions of different genes?
- In what cellular processes do they participate?
- How are genes regulated?
- In which cell types and depending on which conditions the genes become active?
- How various diseases or treatments influence the activity of genes?

They also propose *transcriptomics* as a solution for above. In addition, it facilitates global analysis of gene expression and genome-wide expression profiling. Transcriptome represents genes that are actively expressed at any given time. For the profiling, hybridization-based techniques (DNA microarrays) and sequencing-based techniques (*SAGE*: Serial Analysis of Gene Expression, and *MPSS*: Massively Parallel Signature Sequencing) can be implemented in parallel rather than in competition. It is essential to integrate the biological knowledge in all phases of the data mining process to optimize existing knowledge profit. In an endeavour to standardize things, Gene Ontology (GO) has been framed to subsume and preserve the integrity of data as a result of multitude of experimentations. For the proximity to the research purview, only DNA microarrays have been elaborated. They are also the most efficient amongst all sequencing techniques [9].

7 Data Processing

7.1 Pre-Processing Phase

Genes, *cDNA clones*, or *expressed sequence tags* [ESTs] usually constitute the DNA sequences that are scanned by microarray experiments, conditions contingent. They may include time series data of a biological process, e.g., life cycle of a yeast cell, or a collection of varied tissue samples, e.g., normal versus cancerous tissues. For the same, a *gene expression matrix* is obtained, which for obvious reasons, contains gene data, notwithstanding a compendium of noise, missing values and irrelevant data. Data pre-processing is indispensable before any cluster analysis can be performed.

Study on *promoter sequences* [2] and *enhancer sequences* [2] can be staple for deriving transcription factors of an associated gene. Regulation of transcription is the most common form of gene control, and the activity of transcription factors allows genes to be specifically regulated during development and in different types of cells [13].

It is highly warranted to prune irrelevant data while preparing the data sets for further analysis. The process uses gene expression measures to discover co-regulated genes.

To curb "noise" in the data, replication is employed to ensure precipitation of genes whose expression levels mark the outliers [9]. To do so, methods including fold-change and significance analysis of microarrays (SAM) are put to work. While fold-change is a comparatively simpler method, SAM operates on certain statistical assumptions. Fold-change technique works on selecting/eliminating genes with a predetermined threshold level (usually a factor of 2) [16]. It compares this level with the mean level of the gene expression and thence chooses/rejects genes on the basis of the calculation. SAM underlines the use of t-tests for calculating false discovery rate (FDR)—share of false positives amongst sets.

Analyses based on t-tests can determine the probability of difference in gene expression getting monitored by chance. Putatively, the highlight of the problem is that clustering genes in regard of their expression levels is not the sole criteria. It depends on several other factors viz. experimental conditions, external factors (epigenetics), criteria for choosing base platform for genes (common ground), and their [genes] association with other genes. For sure, a microarray consists of vivid groups of co-expressed genes. A common strategy is to commence with the ones that are related to some ordinal biological function and/or out of a preliminary clustering result.

Supervised, semi-supervised learning techniques use existing domain knowledge for filtering the particular genes [16], [21]. Clustering, therefore, is a very condition specific filtering of gene expression profiles.

7.2 Post-Processing Phase

Since, the pre-processing phase aids in precipitating several groups, patterns, correlations of genes at the expression level basis, it becomes almost necessary to re-evaluate and formalize them in a phase called post-processing phase. During this phase, the domain experts analyze and match the extracted patterns to the business objectives and success criteria.

The dogma of pattern management is heterogeneous pattern representation. Since the extracted patterns can be relevant as well as irrelevant; indexing them is a labor intensive task that involves marking and classifying them scrupulously. Predictive Model Markup Language (PMML) and Common Warehouse model for Data Mining (CWM-DM) were designed for genetic data modelling, but they lacked the efficacy to handle and represent specific classes of patterns. As a solution, Rizzi et al. introduced Pattern Base Management System (PBMS) to provide data structure of the patterns describing a model structure based on pattern interestingness. The subject interestingness underlines two criteria: unexpectedness and actionability. According to the actionability criteria, a model is interesting if it can be put to application. Unexpected models are considered interesting because they contradict certain presumptions based on the predetermined beliefs.

Later, Kotsifakos et al. revised the PBMS architecture by enabling support for domain ontologies. After defining a data modelling system, it's vital to design a mechanism to query and extract the required data. For the same, certain APIs namely, SQL/MM DM, Java Data Mining (JDM) API were standardized to handle data as well as the metadata entwining genetic correlational patterns.

In [17], Brisson and Collard propose an improved distinctive schema to infer interestingness based data. KEOPS methodology works on comparing extracted

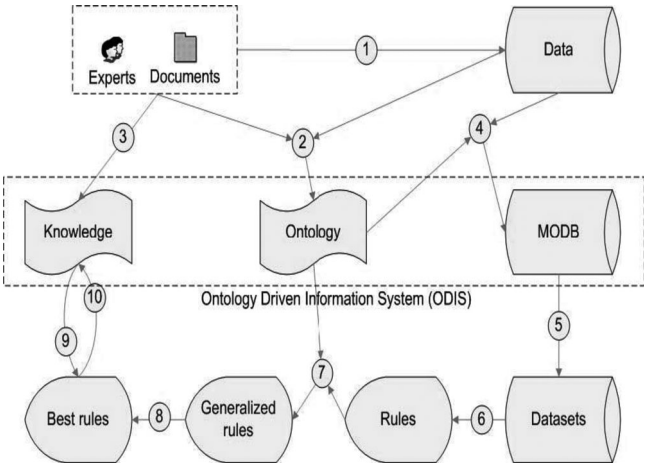

Fig. 4. KEOPS Methodology brings about an integrable and holistic approach in analysing and identifying keys.

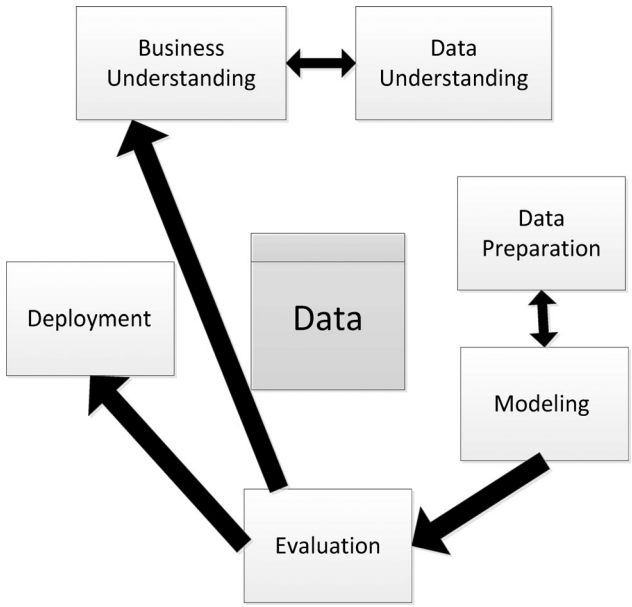

Fig. 5. CRISP- DM (The model functions recursively).

data with expert's knowledge, as elaborated in Fig. 4. Another technique, Cross Industry Standard Process for Data Mining (CRISP-DM) (Fig. 5) is outdone by KEOPS with the inclusion of knowledge in most of the steps of data mining. Ontology formally represents knowledge as a set of concepts within a domain, and the relationships among these concepts. It can be used to reason about the entities within that domain and maybe used to describe the domain. An important issue in ontology-based validation methods is the definition of semantic similarity measures between ontology concepts [17].

Methods that follow are edge counting methods and information-theoretic measures.

Edge counting methods [17] construe the formulation by defining {similarity α (1/distance)}, i.e., the smaller the distance in-between genes, the greater is the similarity, and vice versa. Leacock and Chodorow contemplated the calculation of shortest distance and scaling them to form an "is-a" hierarchy. Zhong et al. consider taxonomical hierarchies to determine weight of the edges. Information-theoretic measures [17], conceptualized by Resnik determine information via lower common ancestors of two underlying concepts. Lin improvised it further to outline distances as well, from the respective ancestor.

Later to it, Jiang introduced its coupled version with the edge counting methods. A comparison by Lord suggested that both of these [edge-counting and information-theoretic] measures were equally efficient. Finally, Schlicker et al. conferred a new technique of similarity between varied GO terms using Resnik's and Lin's definitions. The cyclic nature is due to the way processes run cyclic test-error experiments for the sake of data refinement. KEOPS employs *IMAK interestingness measure* that estimates rule quality via relative confidence value, informative level and certainty of knowledge.

Now, the information is *heterogeneous* in nature. It becomes hard to compare and integrate because of being spread over different sources, represented under different formats, and usually generated with different techniques. It's a challenging task to formalize and standardize such a huge piling of accessible data. Some interpretations show that hybridization-based techniques (microarrays) show greater consistency across platforms than sequencing-based techniques (SAGE, MPSS). To recall, hybridization-based approaches measure ratio of expression changes while sequencing-based approaches produce estimation of the number of transcripts. A solution to this is *MIAME format*.

MIAME format [9] was derived to standardize data from varied microarray experiments and help make comparisons.

It [MIAME format] encompasses the following entities:

- Design of the experiment.
- Microarray layout.
- Preparation of biological samples.
- Protocol used to hybridize the sample.
- The way intensities are quantified.
- The method used to normalize data.

The technology used to derive data from the microarray experiments can illustrate high degree of differences, more than the "actual" distinction in the biological data. To facilitate this, data integration approaches include *light solutions* (link integration, Web 2.0 mashups) and *hardcore methods* (data warehousing, view integration). It is conceived that semantic web technologies can play a major role in sharing and resume biological data between applications by providing a common platform and framework.

Co-clustering techniques envisage distance functions and cluster quality measures for integrating data models and making them indexed. They [co-clustering techniques] generally group the gene expression data with similar expression patterns, i.e., *co-expressed genes* [12]. To prune patterns from the integrated data, it is vital to operate it [data] with classification rules, and clustering algorithms (frequent closed item set). It also focuses on selecting and eliminating ambiguous and redundant rules (frequent item set). Both of these approaches' efficacies are increased significantly when worked in tandem. Dealing with the problem of redundant association rules is incumbent for better results and interpretations. The use of a frequent item set based algorithm to generate classification rules with the form,

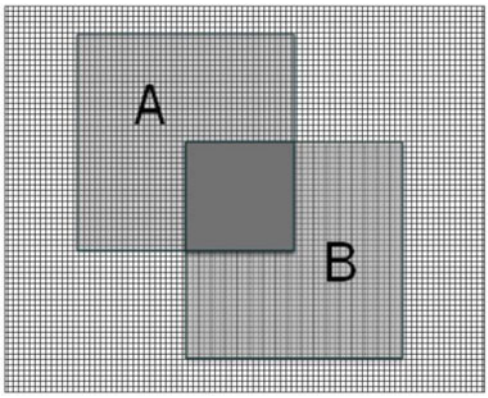

Fig. 6. Illustration of subspace clustering.

"gene expression -> class", from gene expression data of cancerous and healthy tissues has already been studied [9].

Also, Emerging Patterns (EP) has been instantiated to contrast two data classes with particular high dimensionality and construe classifiers from them. They have been proved to be highly indispensable and desirable to sift markers efficiently and aid easy analysis [17].

Parmigiani and Elizabeth [4], [21] elaborate the following data mining approaches of varied clustering and classification techniques for gene expression data.

8 Clustering

8.1 Subspace Clustering

These techniques can be used in microarray analysis to facilitate visual display (mostly preferred by biologists) and interpretation of experimental results and suggest the presence of subgroups of objects (genes or samples) that behave similarly. Often finds itself as the foremost step of data infiltration since it is vital to parry microarray data for noise elimination. Confusion marks with the trait of the genes to participate in multiple pathways that may or may not be coactive under all conditions, so a gene can find its place in multiple clusters or in none at all. Clustering can be *sample-based* and/or *gene-based* by character. A gene-based clustering shall abstract genes as objects and samples as features, while sample-based clustering would perceive vice-versa [13] A third category of clustering type also exists, *subspace clustering*. Subspace clustering (Fig. 6), unlike gene-based or sample-based clustering techniques, is not "global" rather it aims to cluster genes based on their indulgence in any disease, being a part of one or more biological pathways.

8.2 Distance and Similarity

Proximity determination between co-regulated or otherwise genes is momentous to establish any working relationship between them [13]. To determine which objects cluster together, we must have a way of gauging how similar or different both of them are. Different measures reflect different goals, and thus can have a strong influence on the resulting clusters [5], [6]. The correlation coefficient is commonly used as a measure of the divergence of gene expression profiles between different species [20], the core problem being that they [correlation coefficients] tend to cascade and amplify the measurement error [12]. They include Pearson correlation coefficient, euclidean distance, Uncentered correlation, etc.

Euclidean distance is comparatively easier to implement and most commonly used methods to elucidate distance between two data objects [18].

It is formulated as

$$\text{Euclidean}\,(O_i, O_j) = \sqrt{\sum_{d=1}^{p} (O_{id} - O_{jd})^2}.$$

Due to the varied sources of microarray platforms, it becomes logical to make the readings comparable. Liao and Zhang (2006) proposed the computation of *Relative Abundance (RA)*, to leverage the heterogeneity in expression levels coming from distinct processes. The RA is the ratio of gene expression in a particular tissue and its sum of expression levels across all tissues,

$$\text{RA} = \frac{expression_{single}}{\sum_{all\ genes}\ expressions}.$$

Pearson correlation coefficient has proven to be inefficient while dealing with the outlier data. It also assumes that the random variables under consideration are *linearly related* [18]. It is an "over-imposed" version of the euclidean distance formula. It considers calculation of means of the two data objects under consideration, O_i and O_j respectively. The proximity between two objects is measured by a proximity function of corresponding vectors $\rightarrow_{O_i}\ and\ \rightarrow_{O_j}$, where p is the number of dimensions in the space,

$$\text{Pearson}(O_i, O_j) = \frac{\sum_{d=1}^{p} (O_{id} - \mu_{O_i})(O_{jd} - \mu_{O_j})}{\sqrt{\sum_{d=1}^{p} (O_{id} - \mu_{O_i})^2} \sqrt{\sum_{d=1}^{p} (O_{jd} - \mu_{O_j})^2}}.$$

The euclidean distance reflects only the noise present in the data and hence will be small if the noise is small. By contrast, the Pearson distance will have a value close to 1, reflecting the fact that the noise components of different expression levels are independent. Thus the Pearson distance will give the impression that expression divergence is great, but all this apparent divergence is noise [20]. Also, it [Pearson's correlation coefficient] assumes Gaussian distribution of points and may not be suitable for non-Gaussian distributions. These limitations led to the focus on Spearman's rank-order correlation coefficient and the Jackknife correlation as alternative similarity measures [13].

Jackknife $(O_i, O_j) = \min\{\rho_{ij}^1, \ldots, \rho_{ij}^l, \ldots, \rho_{ij}^p\}$, where ρ_{ij}^l is the Pearson's correlation coefficient of data objects O_i and O_j with the lth feature deleted. The Jackknife method is proven efficient for finding the genetic correlation and the associated confidence interval [14], [15]. The Spearman's ranking correlation is derived by replacing the numerical expression level O_{id} with its rank r_{id} among all conditions, and is hence distribution free. The results interpreted by the authors [13] indicate that Pearson's correlation coefficient performed well than Spearman's rank-order correlation.

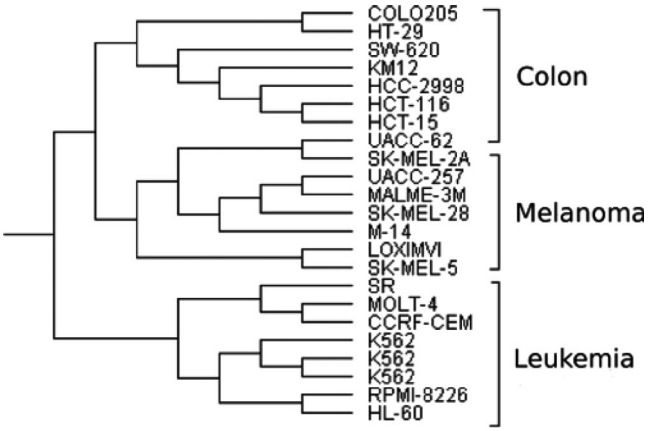

Fig. 7. Dendrogram generated from AH-Cut for the melanoma-colon-leukemia data set [23].

As per the inequality relation given by Daniels (1944), $-1 \leq 3\tau - 2r_s \leq 1$, where r_s is the Spearman's coefficient and τ is Kendall's coefficient.

On the contrary though, authors [18] register the incompetency of correlation coefficients as the measures to exploit gene expression data. It is conveyed that if the measurement error induced during the design of the experiment is large in comparison to the divergence of the gene expression levels, the correlation coefficients (irrespective of the types) shall render accentuated and sporadically-occurring divergence metrics at par with the uniform and "real" differences amongst genes.

8.3 Hierarchical Clustering

It partitions objects into a series of nested clusters [23]. The hierarchy of clusters of samples is displayed using a tree-like structure called dendrogram, as exemplified in Fig. 7. "*Oaks from acorns*" is the catch phrase here. The height at which two clusters are joined represents how similar they are, with low heights representing high similarity [7], [13]. They are of two types: *agglomerative* (bottom-up approach) and *divisive* (top-bottom approach). Agglomerative strategy considers each data object as an individual cluster and while moving forward, clubs the closest pairs until we are left with just one cluster. Conversely, divisive approach starts with initially one cluster containing all data objects and iteratively splits values; at the end just singleton clusters remain.

Eisen et al. (1998) programmed an agglomerative algorithm called *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA), and graphically represented the clustered data set. In this method, each cell of the gene expression matrix is colored with the measured fluorescence ratio and based on dendrogram structure and ordering doctrine, the matrix rows are reordered. Eventually, the matrix is represented by consistent patches of varied colored gene expression patterns that depict groups with certain degree of similarity. Alon et al. (1999) used the divisive approach to formulate an algorithm, called *deterministic-annealing algorithm* (DAA). In this technique, initially centroids of two start-up clusters were randomly chosen to be $C_j, j = 1, 2$. The expression level of gene k was defined via vector $\vec{g}_k$, and the probability of gene k belonging to the cluster j was designated by a two-component Gaussian model:

$$P_j(\vec{g}_k) = \exp(-\beta|\vec{g}_k - C_j|^2) / \sum_j \exp(-\beta|\vec{g}_k - C_j|^2).$$

The cluster centroids were recalculated by,

$$C_j = \sum_k \vec{g}_k P_j(\vec{g}_k) / \sum_k P_j(\vec{g}_k).$$

Later, an iterative process, *the EM algorithm*, was applied to solve P_j and C_j. For $\beta = 0$, there was just one cluster $C_1 = C_2$. On gradually increasing β to a certain limit, two distinctive, converged centroids emerged [13]. The EM algorithm iterates between Expectation (E) steps and Maximization (M) steps. In the E step, hidden parameters are evaluated in correspondence with the given parameters. In the M step, given (model) parameters are estimated so as to maximize the probability of the entire data, given the estimated hidden parameters. The repercussion of the EM algorithm renders each datum to be assigned to a cluster with optimal conditional probability. Visually perceived, data can be retained for a longer period of time, when learning is concerned. The ease of interpretation via graphical representation can be successfully achieved via hierarchical clustering.

Limitation of hierarchical approach, however, is the large computational complexity. While dealing with "big data", partitioning may require exhaustive processing while splitting and merging data into clusters.

8.4 K-Means Clustering and Self Organized Maps (SOM)

It partitions objects into groups that have little variability within clusters and large variability across clusters [7]. The user is required to specify the number K of clusters a priori. Estimation is iterative, starting with a random allocation of objects to clusters, re-allocating to minimize distance to the estimated "centroids" of the clusters, and stops when no further improvements can be made. Its implementation is easy and execution is faster. The time complexity was computed to be O (L^*K^*N), where L is the number of iterations in K clusters [13], [21]

$$E = \sum_{i=1}^{K} \sum_{O \in C_i} |O - \mu_i|^2.$$

In the above expression, O is the data object in cluster C_i and μ_i is the mean of objects in C_i. There are some limitations to K-means algorithm, as well. While the number of clusters is usually unknown in advance, the algorithm is tested multiple times for different values of K to reach an optimal result. This can be cumbersome especially when the magnitude of data is mammoth [gene expression microarray data]. Also, with the huge inherent noise in the data, most of it is "forced" into the clusters; that is questionable for the data integrity and correctness [13]. Solutions to these drawbacks have been considered by certain revised algorithms that control the "quality" of data that is getting clustered.

A closely related, not to mention- more efficient, approach is that of Self-Organizing Map constituted by Kohonen on neural networks [8]. The mapping of input neurons to output neurons in a grid/mesh of sample

neurons was the exemplifying model. The algorithm, on execution, attempts to direct the *reference vectors* to the input vector space. The reference vectors, being trained data sets, "educate" the input vectors to fit the distribution profile [13]. As prerequisites, SOM demands from the user, the number of clusters and the grid-map of the neural network, a priori.

Hierarchical, K-Means, and SOM are categorized under Gene-based clustering.

8.5 Principal Components Analysis (PCA) and Multi-Dimensional Scaling (MDS)

The goal is to reduce the dimensionality of data to facilitate visualization and additional analysis. They are often used as a preliminary step to clustering of large data sets. MDS starts from a distance matrix between objects and finds the locations of these objects in a low dimensional space that best preserves the original distances. These techniques work on *ratio-optimization principle*.

It's almost concomitant of clustering techniques for high dimensional data to be exploratory. Their strength is in providing rough maps and suggesting directions for further study. Also, clustering results are sensitive to a variety of user-specified inputs. The clustering of a large and complex set of objects can be planned in different ways depending on the goals.

9 Classification

These techniques can be used in microarray analysis to predict sample phenotypes based on gene expression patterns. Some of the methods are briefly reviewed [4].

9.1 Dimension Reduction

Because of large number of genes that can be used as potential predictors, it is useful to preselect a subset of genes, or composite variables, likely to be predictive and then investigate in depth the relationship between these and the phenotype of interest. For example, genes with nearly constant expression across all samples can be eliminated.

9.2 Evaluation of Classifiers

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings.

9.3 Predictive Analysis of Microarrays (PAM)

A straightforward approach to classification is the nearest centroid (pivotal point) classifier. This computes, for each class, a centroid given by the average expression levels of the samples in the class, and then assigns new samples to the class whose centroid is nearest. This approach is similar to k-means clustering except clusters are now replaced by known classes. It has been implemented by PAM software [25].

9.4 Top Scoring Pairs (TSP)

In a two-class classification, this looks for pairs of genes such that gene1 is greater than gene2 in class A and smaller in class B, in terms of expression levels. In cancer data, the TSP classifier achieves prediction rates that are as high as those of alternative approaches which use considerably more genes and complex procedures.

9.5 Nearest Neighbor Classifiers

Nearest-neighbor classifiers assign sample to classes by matching the gene expression profile to that of samples whose class is known. The classifiers are robust, simple to interpret and implement, and do not require, although they may benefit from, preliminary dimension reduction.

9.6 Support Vector Machines (SVMs)

SVMs are supervised, machine learning algorithms that seek cuts of the data that separate classes effectively, that is by large gaps. Technically, SVMs operate by finding a hyper surface in the space of gene expression profiles, that will split the groups so that there is largest distance between the hyper surface and the nearest of the points in the groups. More flexible implementations allow for imperfect filtering of groups and promiscuous analysis.

9.7 Discriminant Analysis

Discriminant analysis and its derivatives are approaches for optimally partitioning a space of expression profiles into subsets that are highly predictive of the phenotype of interest, for example by maximizing the ratio between-classes variance to within-class variance.

9.8 Classification Trees

Classification trees recursively partition the space of expression profiles into subsets that are highly predictive of the phenotype of interest. They are robust, easy-to-use, and can automatically sift large data sets, identifying important patterns and relationships. No prescreening of the genes is required. The resulting predictive models can be displayed using intuitive graphical representations.

9.9 Regression-Based Approaches

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence.

9.10 Probabilistic Model-Based Classification

It is based on the specification of probability distribution that describes the variability of the expression values. Model-based approaches are computation-intensive and can be sensitive to assumptions made about the probability model, but can provide a solid formal framework for the evaluation of many sources of uncertainty, and for assessing the probability of a sample belonging to a class.

A wide range of alternative approaches for clustering and classification of gene expression data are available. While differences in efficiency do exist, none of the well-established approaches is uniformly superior to others. Choosing an approach requires consideration of goals of the analysis, the background knowledge, and the specific experimental constraints. The quality of algorithm is important, but is not in itself a guarantee of the quality of a specific data analysis. Uncertainty, sensitivity analysis and, in the case of classifiers, external validation or cross-validation should be used to support the legitimacy of results of microarray data analyses.

10 Association Rules

An important extension (rather it's a new edition of traditional approach towards filtering data) to the clustering and classification techniques, is the employment of association rules [21].

10.1 Limitations of Clustering over Association Rules

In cluster analysis of expression data, the goal is to define each gene as being a part of a self-contained cluster, based on the similarity in the expression pattern of the gene to those of the other genes in the same cluster. Which genes cluster together can vary considerably, both because of the different similarity metrics that can be used to compare any two clusters and because of experimental and biological noise that exists in the expression data.

Another issue with clustering is that a gene can usually be characterized in more than one way, while it can belong to only one cluster (in hierarchical clustering, we have a hierarchy of clusters within clusters, but a gene cannot belong to two unrelated clusters). It warrants further refinement at per tier level.

Determining the interactions that can exist between different genes is not easily done using clustering results, especially as a gene can participate in more than one gene network. So, the "rigidity" observed in the clustering approaches is smoothened out in Association rules. With reference to the market-basket analysis*, while analyzing the gene expression data, the items in an association rule can represent genes that are strongly expressed or repressed, as well as relevant facts describing the cellular environment of the genes. The cognizance of pattern extraction techniques are deployed to apprise correlations and links in the data presented as association rules.

An example of an association rule mined from expression data might be,

$$\{\text{Cancer}\} => \{\text{Gene A} \uparrow, \text{Gene B} \downarrow, \text{Gene C} \uparrow\}.$$

Meaning that, for the data set that was mined, in most profile experiments where the cells used were cancerous, Gene A was measured as being up (i.e., highly expressed), Gene B was down (i.e., highly repressed), and Gene C was up, together. It is important to note, however, that ↑, ↓, or – (neutral) states of the gene interactions have to be measured mathematically for precise analyses.

Another complexity is induced in the data analysis, when values are *binned*. Binning leads to aggravation due to cascading effect of an error getting carried forward. Less precise results are obtained, but this can be categorized "indispensable" because of the voluminous data. May be due to lack of research time only "quick and small" references of data were made.

10.2 Advantages of Implementing Association Rules

They bring about "associativity" amongst genes. That does not imply cause-effect relationship, albeit. A gene can belong to multiple association rules. This will invariably help in establishing a gene function map.

10.3 Applications of Association Rules

Association rules could help in the search for cancerous genes, especially as the case could exist where no single gene might be responsible for the initiation or progression of cancer, but instead certain sets of genes acting together. Search for associations between certain attributes of the medical histories of cancer patients and the genes that might be expressed in their corresponding tumors as a result.

Market basket analysis is a modeling technique based upon the theory that if you have bought or buy a certain group of items, you are more (or less) likely to buy another set of items.

11 Gene Ranking and Cancerous Gene Identification

Owing to the high-dimensional feature space and equi-exploratory noise, microarray data is susceptible to drawing some ambiguous results which could be misleading and render illegitimate conclusions. Several gene ranking techniques, e.g., ANOVA, T-Score were introduced to address the above issues, but drew unsatisfactory inferences [10]. Earlier, cancerous tissues were dealt completely clinically. Now, though due to the implicit efficacies, DNA Microarray method is considered paramount coupled with computational deduction techniques. It helps develop better understanding towards characterizing genes and their affiliation to any disease. A *classifier* must be efficient data sifter as well as it should disconsider irrelevant data that augments *noise*.

To elucidate cancerous limits in a tissue, Revathy and Amalraj proposed a cancer classification using semi-supervised Ellipsoid ARTMAP and Particle Swarm Optimization with gene expression data [10]. With the advancements in microarray technologies, prolific data has emerged and after comparing diseased and control genes, necessary predictions are established. Choice of genes is significant from performance standpoint of the classifier and the correctness of predictions. It also offers better and deeper molecular insights into the treatment. The authors implement the semi-supervised ellipsoid (SSEAM) for multiclass cancer distinction and particle swarm optimization for selecting the particular gene. Alternatively, Huilin et al. propose Optimized Kernel Machines, with an extremely flexible kernel function for optimizing the data kernel [10]; Xiyi et al. presented Sparse Representation using the genetic data. It inhibits the property of finding sparse representations in test samples as in training samples [10]; Runxuan et al. submit Extreme Learning Machine (ELM) for the

same, though it outperforms the iterative learning techniques' underpinnings such as local minima, over-fitting, and improper learning rate.

Revathy and Amalraj (2011) use GCM, Lung, and Lymphoma as the test data sets for diagnosing cancer. These were been derived from online data repositories such as NCBI, EBI, etc.

The proposed methodology is structured in two phases:

- Determining *enrichment scores* using gene importance ranking techniques.
- Classification using *Support Vector Machines* (SVMs)

The enrichment score is posted after the gene importance ranking evaluation. The evaluation of confidence for every enrichment score is formulated using Family Wise Error Rate (FWER) and False Discovery Rate. In the classification phase, support vector machines are used which act as *supervised classification* tools. The results show that gene ranking technique narrowed the search space significantly and hence aided the cancerous gene identification. With the data optimization, the efficiency of the SVM is correspondingly optimized. Earlier to this study, Su et al. [11] (2003) had developed a program called *RankGene* for identification of diagnostic genes by analysing gene expression data and differentiating them in a particular sample. Research by Golub et al. (1999) and Ramaswamy et al. (2001) has successfully demonstrated the utility of DNA microarray-based gene expression data in cancer classification. By comparing the expression levels of genes in a diseased state and normal state, an inference can be drawn regarding the state and stage of the disease, cancer in particular. *Rank of a gene is its ability to distinguish between the classes- diseased, normal*. As an input to the program, data set from a particular sample or tissue is made. The program *rates* every gene and measures its ability to differentiate between varied classes. It assigns to every gene a numerical value – its rank. It outputs best k genes according to this measure, where k is user contingent. However, as a consequence of ranking, a significant amount of information present in the data is lost.

Raza and Mishra [12] (2012) attempt to stratify genes within samples (tissues) by recursively filtering genes on the basis of their expression levels and active indulgence in the disease state. The expression level of genes is proportional to various conditions in an organism. It is incumbent to mark *reference genes* that can be levied as standard for selecting further candidate genes on the basis of a priori criterion. These can be well suited to be potential drug targets and as sites for studying mutations. For such study, gene expression matrix is a good reference source for each gene's expression variance. The algorithm curtails following steps:

1. *Ratio and logarithmic conversion of microarray data.* (The gene regulation is reflected in the fluorescence intensities that illuminate on superimposition of the multivariate genes. Author has limited the regulation levels by defining up-regulation -> [1, ∞] and down-regulation -> [0, 1]. This categorization brings about rigidity in selecting the *acceptable gene expression levels*. Through intensity ratio plot, interpretations can be efficiently visualized.)
2. *Elimination of gene that fails to provide data in majority of experiments.* (Due to several technical issues of improper probing, particular cell orientation of genes, faulty scanner that fails to measure correct expression levels, and due to erroneous manufacturing of microarray chip, gene expression levels can be severely affected. On account of a threshold value of 40 percent, rows holding genes that are not expressed up to the level are abstained from being a part of the experiment in the view of not having significance relevance.)
3. *Analysis of significance of data.* (Use of t-statistics promulgates the compliance of normal distribution in the data that is responsible for certain pattern generation. This [pattern] can be analysed and may be interpreted as a result.)
4. *Replicate handling.* (There should be a single entry for each gene.)
5. *Elimination of gene having less than two-fold change in expression level.* (Irrelevant genes- that do neither show acceptable up-regulation [positive value] nor down-regulation [negative value], are curbed. For convenience of sifting data, genes means of all rows is calculated and genes with $-1<\text{mean}<1$ is selected further.)
6. *Conversion of data sets using Log sigmoid function.* (The Log sigmoid transformation takes range of input values in between [$-\infty$, $+\infty$] and converges them to the range [0, 1]. The function is given by *log sigmoid* $(x) = 1/(1 + e^{(-x)})$.
7. *Elimination of genes that have high variation across the collection of sample.* (The genes with sporadically occurring expression levels are omitted. The process accounts for elimination of genes having more than 36 percent variation due to inconsistency.)

The program was implemented in PERL with Yeast data set having a total of 6,154 genes out of which 48 genes were precipitated out (around 0.78 percent of the sum). This is however a significant proportion of genes getting filtered, but due to the *binning* of values and imposing rigidity as criteria for gene selection (which albeit eases the processing), many key features may have been missed out. On the contrary, the results claim that out of 48 outputted genes, 32 genes were solely responsible for cocooning yeast cells at high temperatures. Additively, eight other genes with obscure functionalities exhibited two-fold change in expression levels at high temperatures as well. These heat tolerant genes are thus supposedly better drug targets [12]. Likewise this, are groping studies that are based on "hit-and-trial" methodology, and render vague results, which are subject to clinical validation and ontology-centred verification. Even software available as free or proprietary, are written to automate various formulations of mathematical and statistical genre, but that is just to make the work easier. They do not guarantee a favoured outcome.

12 Clinical Viability of the Microarray Data

Clinical usefulness of the microarray data is determined by the informational benefits to the patient and the role it plays in successful clinical decision making. One of the reports

under the aegis of the Human Cancer Genome Atlas (HCGA) project suggests that only a handful of mutated genes are causal of breast cancer [45]. Genomic profiling refers to the analysis of the genes that are therapeutically sensitive and clinically interesting. This trait can be derived from the arbitrary nature of their expressions. Some genes (markers or biomarkers, technically) have been profiled for their specific expression patterns or "signatures" and have rightfully marked in the Gene Ontology. Sparano et al. [45] state the embodiment of prognostic and predictive markers; the difference between the two being that while the former is associated with the clinical outcome, the latter defines the genes that clinically respond to therapy, particular or generic.

Simon (2005) [51] proposed a flowchart for adopting a multiparameter marker. It encapsulates phases of:

i. Problem conceptualization.
ii. Clinical development, where training set is determined.
iii. Technical development, to establish reproducibility and reliability of the assay.
iv. Validation of the trained data (clinicopathological data, technically) set in context of its scalability.
v. Application or adoption of the marker for its usability.

Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) has been institutionalized to oversee the procedural deficits and paucity in conformance to the marker analysis. It has a tandem function with MIAME [9], as aforementioned.

Further studies by Kheirelseid et al. [47] for colorectal carcinoma (CRC) and Sveen et al. [48] for colon cancer embark the clinical applicability of the gene expression data. Gauging the efficacy of the microarray technology and its interoperability standardization during its application phase was the vision with which the US Food and Drug Administration MicroArray Quality Control (MAQC)

Project was conceived [46]. It also monitors genome-wide association studies and NGS, which are extension technologies to the DNA microarray. MAQC-II, the next version of the project is notioned to identify the parameters for the clinic-use. Parry et al. attempt to evaluate the MACQ-II data with KNN models (the microarray data analysis that was carried out using KNN methodology). The ease of implementation and scale make KNN, a straightforward choice. Data sets orienting to breast cancer, neuroblastoma and multiple myeloma, were comparatively evaluated with KNN and other clinical parameters (by their performance), and their affect in the results was noted.

13 Next Generation Sequencing (NGS)

The tune of cancer research has been tweaked towards Next Generation Sequencing. Although, the underlying dogma remains the same, i.e., identifying marker genes, but with larger proforma of data now available (single-molecular sequencing), getting closer to drug targets and their clinical validation, with greater stress on clinical trials and people volunteering (its [clinical trial] data available online and at everyone's disposal), is almost achievable. With high

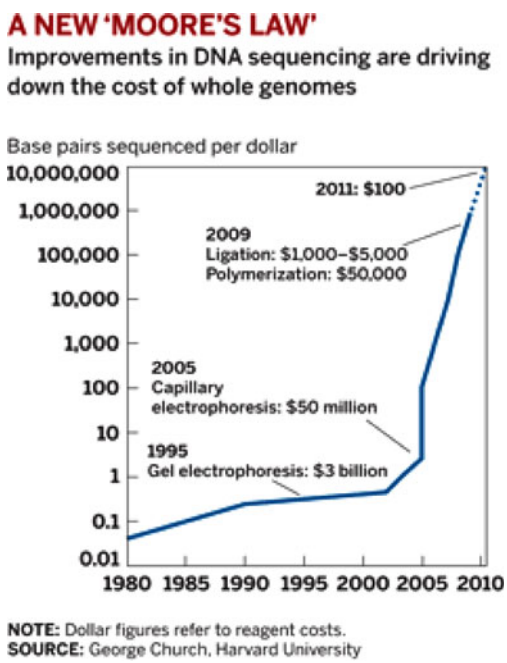

Fig. 8. The graph above shows the plummeting technological costs. Moore's law which states that processing power doubles every year, has to some to extent, fallen short. (Credit: Harvard University).

competency to handle genetic data (as per Fig. 8) in an efficient and precise manner and capability to optimally analyze it, NGS is forging its way towards personalized therapy and curb difficult diseases like cancer and likewise; but a glance at the history of DNA sequencing will be interesting here.

13.1 Background

After Watson and Crick revealed the structure of DNA in 1953 [26], biologists have always sought DNA sequencing with augmented enthusiasm. Before 1976, determining DNA sequences was a pain-staking and laborious task involving traditional chemical methods. That was when Sanger et al.'s "plus and minus" came into picture with some seminal work towards determining a base sequence of bacteriophage ϕ X174. Maxam and Gilbert were teaming up in parallel for some seminal work. Later in 1980, Sanger and Gilbert shared the Nobel Prize in Chemistry for their developments of analytical methods of DNA sequencing. All the concurrent developments of that era warranted that a store be developed for holding the information that was a result of experimentation and make it accessible to scientific and general community for referencial usage. This led to the surrection of GenBank in 1981 and since then the "size" of the repository has forever grown, even exponentially.

The commercialization of DNA sequencing had a major impact on biology with the profuse marketing of instruments on automated capillary-array electrophoresis (CAE). Initially, Sanger's sequencing was a mind-numbing exercise which was error prone and after quite a time taking electrophoresis and audioradiography, the DNA sequence was manually fed into the computer by a technocrat. An improvised version of the Sanger sequencing was reported thereafter which replaced radioactive labesl, audioradiographic detection, and manual data interpretation with fluorescent

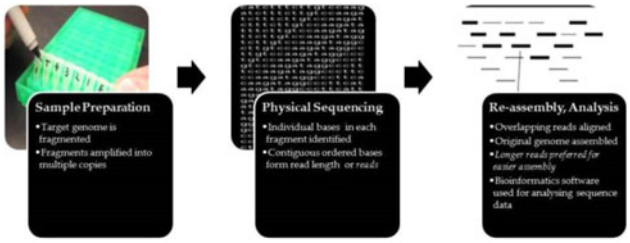

Fig. 9. *DNA Sequencing*. Broadly involving three phases of sample preparation, physical sequencing, and re-assembly, DNA sequencing caters to the feed for genetic data analysis [39].

labels, laser-induced fluorescence detection, and computer-based data analysis. This paradigm shift can well be conceived as the edifice of modern day genetic data analysis. It was cemented by Smith et al. in 1986 [27] and made salable by Applied Biosystems in 1987. Prober et al. at Dupont and Ansorge et al. are known to design improved automated sequencers, later on. These were seminal for the Human Genome Project, which was instituted a year after.

13.2 DNA Sequencing

The mechanism of DNA Sequencing, as expounded in Fig. 9, comprises of three steps mainly, viz.

- *Sample preparation.* The particular genome is fragmented into multiple entities. Further, in accordance to the sample DNA, multiple copies of these *sub-genomes* are created by amplification and using various molecular methods.
- *Physical Sequencing*. The fragments received from the earlier step are sifted for orderly bases. These contiguous chains are called *reads* or read lengths.
- *Reassembly & analysis.* Using sophisticated bioinformatics software, the overlapping reads from various fragments are marked to form a contiguous genome sequence. This is forwarded for auxiliary analysis.

For the analysis of the data, various analysis software were developed. One by Applied Biosystems was subdued by an algorithm designed by Berno [40], that used graph theoretic approach to find the longest sequencial reads that would comply to the entire genome's definition.

13.3 High Throughput DNA Sequencing

NGS embarks high-throughput sequencing technologies that have reduced sequencing and cost expenditures radically. The "Big Data" at disposal has vastly impacted genetic, biological and medical research. Sequencing technologies broadly follows the course of library-building, sequencing and analysis. The efficiency of each [sequencing technology] is dependant on the mechanistic protocols in use.

The sequencing of genetics data has been structurally categorized in the generational context: *First Generation Sequencing*, *Second Generation Sequencing*, and *Third Generation Sequencing* (which we are currently witnessing, also revered as Next Generation Sequencing) [44]. The basic aspects of distinction between Sanger sequencing and the NGS are optimally parallel analysis, greater output threshold, and achieving all of this at a lesser price. Xprize [43] has also been a motivation for rapid advancements in the NGS. There is also an exclusive attempt of sequencing 100 centenarians' genomes, a competition called "100Over100". The medics attempt to identify those very peculiar genes that are disease resistant, support immunity, and can be clue to longetivity and good health. Mulling the history of sequencing, as with any challenging scientific endeavor, we witness a surge of technological advancements through the time and boosted precision [38]. The genetics lobby today is vested with impended longer reads (*read lengths: DNA base pair sequences*), squatter time to output results and lesser overall economics.

Modern day sequencing owes substantial credit to Oxford Nanopore technologies. Previous to this was a transitional day and age of Ion Torrent's semiconductor sequencing, which was later acquired by Life Technologies.

For detailed comparison of varied facets of different sequencing generations, one can refer Table 1 from [38].

TABLE 1
Differences between Clustering and Classification Techniques of Data Mining

Clustering	Classification
Unsupervised learning	Supervised learning
Class discovery	Class prediction
Partitions a set of objects into groups that are relatively similar.	Determines whether an object belongs to a certain class.
Applications	
To generate hypotheses about novel disease subtypes.	Classification of patients into existing disease subtypes or prognostic classes into gene expression information.

14 Conclusion

It is observed that a reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer [3]. By allowing the monitoring of expression levels in cells for thousands of genes simultaneously, microarray experiments may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel approach to cancer classification. Also annotated is that comparing the activity of genes in a healthy and cancerous tissue may give some hints about the genes that are involved in cancer. Albeit, this approach is very limited because many of the genes serve multiple functions and changes in gene expression can be due to factors not directly concerned with the particular experiment. Indeed a microarray data set contains numerous groups of co-expressed genes. Then, a typical strategy for a biologist is to start from genes which are known to be closely related to a biological function and to browse a

preliminary rough clustering result, to focus on a small subset of those genes which are supposed to play a role. Thus, currently biologists follow exploratory strategies by manually selecting potential groups of genes according to their knowledge. So, on experimenting with these "superficial" data and applying various data mining techniques to them, results are rather vague and imprecise. Therefore, the input data has to be concise and close to accurate to obtain the results of the same nature. This will not only be seminal for precipitating strong inferences but will also act as a template for further improvements. This however is beyond the scope of various data mining techniques.

Also, moving towards an era of *personalized* or *precision medication*, NGS and Gene Therapy are making there mark there. Many private players are offering this service and better yet, it can be ordered online. What is still missing is that only few genes are examined in the genetic tests and not the entire genome. The reasons accountable can be time and monetary constraints. The future generations of sequencing can facilitate efficacy in deciphering faulty genes and validating diseases with less hoax.

REFERENCES

[1] *Data and Statistics*. World Health Organization, Geneva, Switzerland, 2006.

[2] PubMedHealth- U.S. Nat. Library Med., (2012). [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002267/

[3] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Amer. Statist. Assoc.*, vol. 97, no. 457, pp. 77–87, Mar. 2002.

[4] G.-M. Elizabeth and P. Giovanni, (2004, Dec.). "Clustering and classification methods for gene expression data analysis." Johns Hopkins Univ., Dept. of Biostatist. Working Papers. Working Paper 70. [Online]. Available: http://biostats.bepress.com/jhubiostat/paper70.

[5] E. Shay, (2003, Jan.). "Microarray cluster analysis and applications" [Online]. Available: http://www.science.co.il/enuka/Essays/Microarray-Review.pdf.

[6] M. B. Eisen, T. P. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 25, pp. 14863–14868, Dec. 1998.

[7] S. Tavazoie, D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, pp. 281–285, 1999.

[8] T. Kohonen, *Self-Organising Maps*. Berlin, Germany: Springer-Verlag, 1995.

[9] N. Pasquier, C. Pasquier, L. Brisson, and M. Collard, (2008). "Mining gene expression data using domain knowledge," *Int. J. Softw. Informat*, vol. 2, no. 2, pp. 215–231, [Online] Available: http://www.ijsi.org/1673-7288/2/215.pdf.

[10] N. Revathy and R. Amalraj, "Accurate cancer classification using expressions of very few genes," *Int. J. Comput. Appl.*, vol. 14, no. 4, pp. 19–22, Jan. 2011.

[11] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, (2003) "RankGene: Identification of diagnostic genes based on expression data," *Bioinformatics*, vol. 19, no. 12, pp. 1578–1579, [Online] Avaialble: http://bioinformatics.oxfordjournals.org/content/19/12/1578.full.pdf.

[12] K. Raza and A. Mishra, "A novel anticlustering filtering algorithm for the prediction of genes as a drug target," *Amer. J. Biomed. Eng.*, vol. 2, no. 5, pp. 206–211, 2012.

[13] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[14] D. A. Roff and R. Preziosi, "The estimation of the genetic correlation: The use of the jackknife," *Heredity*, vol. 73, pp. 544–548, 1994.

[15] T. Scharl and F. Leisch, "Jackknife distances for clustering time-course gene expression data," in *Proc. ASA Biometrics*, 2006, p. 8.

[16] K. M Williams, "Statistical Methods for analysing microarray data: Detection of differentially expressed genes" Inst. Signal Process., Tampere Univ. Technol. Tampere, Finland, Dep. Biology, Univ. York, York, U.K., 2004.

[17] B. Collard, "An ontology driven data mining process" Inst. TELECOM, TELECOM Bretagne, CNRS FRE 3167 LAB-STICC, Technopole Brest-Iroise, France & Univ. Nice Sophia Antipolis, France, 2008.

[18] J. Hauke and T. Kossowski, "Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data," *Quaestiones Geographicae*, vol. 30, no. 2, pp. 87–93, 2011.

[19] B. Collard, "How to semantically enhance a data mining process?" *Lecture Notes Bus. Inform. Process.*, vol. 13, pp. 103–116, 2009.

[20] P. Waxman and E. Walker, "A Problem with the correlation coefficient as a measure of gene expression divergence," *Genetics*, vol. 183, pp. 1597–1600, 2009.

[21] B. Vilo, "MiniReview-Gene expression data analysis," *FEBS Lett.*, vol. 480, pp. 17–24, 2000.

[22] H. Creighton, "Mining gene expression databases for association rules," *Bioinformatics*, vol. 19, no. 1, pp. 79–86, 2003.

[23] R. Mahata and M. Moscato, "Hierarchical clustering using the arithmetic-harmonic cut: Complexity and experiments," *PLoS One*, vol. 5, no. 12, p. e14067, 2010.

[24] D. R Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, "Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Res.*, vol. 62, pp. 4427–4433, 2002.

[25] PAM Software. (2002). [Online]. Available: http://statweb.stanford.edu/~tibs/PAM/

[26] J. D. Watson and F. H. C. Crick, *Nature*, vol. 171, pp. 964–967, 1953.

[27] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood, *Nature*, vol. 321, pp. 674–679, 1986.

[28] J. A. Adam, *A Mathematical Nature Walk*. Princeton, NJ, USA: Princeton Univ. Press, 2011.

[29] J. A. Adam, *Mathematics in Nature: Modeling Patterns in the Natural World*. Princeton, NJ, USA: Princeton Univ. Press, Nov. 2003.

[30] B. Keim, Whew! your DNA isn't your destiny, *Med-Tech Health*, (2005). [Online]. Available: http://www.wired.com/medtech/health/news/2005/08/68468

[31] J. Cloud, "Why your DNA isn't your destiny," *Time Magazine* 2010. [Online]. Avaialble: http://content.time.com/time/magazine/article/091711952313,00.html.

[32] The Human Epigenome Project. (2013). [Online]. Available: http://www.epigenome.org/

[33] G. Kaati, L. O. Bygren, and S. Edvinsson, "Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period," *Eur. J. Human Genetics*, vol. 10, pp. 682–688, 2002.

[34] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, and A. Kasprzyk, "International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data," *Database*, vol. 2011, 2011.

[35] L. Zhang, A. J. Thrasher, and H. B. Gaspar, "Current progress on gene therapy for primary immunodeficiencies," *Mini Rev., Gene Therapy*, vol. 20, pp. 963–969, 2013.

[36] M. I Cancio, U. M. Reiss, A. C. Nathwani, A. M. Davidoff, and J. T. Gray, "Developments in the treatment of hemophilia B: Focus on emerging gene therapy," *Rev., Appl. Clinical Genetics*, vol. 6, pp. 91–101, 2013.

[37] The Cancer Genome Atlas. (2013). [Online]. Available: http://cancergenome.nih.gov/

[38] E. E. Schadt, S. Turner, and K. Andrew, "A window into third-generation sequencing," *Human Molecular Genetics*, vol. 19, no. 2, pp. 227–240, 2010.

[39] M. C. Schatz, A. L. Delcher, and S. L. Salzberg, "Assembly of large genomes using second-generation sequencing," *Genome Res.*, vol. 20, pp. 1165–1173, 2010.

[40] A. J. Berno, "A graph theoretic approach to the analysis of DNA sequencing data," *Genome Res.*, vol. 6, pp. 80–91, 1996.

[41] P. M. Das and R. Singal, "DNA Methylation and Cancer," *J. Clin. Oncol.*, vol. 22, no. 22, pp. 4632–4642, Nov. 2004.

[42] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law, "Comparison of Next-Generation Sequencing Systems," *J. Biomed. Biotechnol.*, vol. 2012, p. 11, 2012.

[43] (2013). [Online]. Available: http://genomics.xprize.org
[44] P. K. Wall, J. Leebens-Mack, A. S. Chanderbali, A. Barakat, E. Wolcott, H. Liang, L. Landherr, L. P. Tomsho, Y. Hu, J. E. Carlson, H. Ma, S. C. Schuster, D. E. Soltis, P. S. Soltis6, N. Altman7, and C. W. dePamphilis, "Comparison of next generation sequencing technologies for transcriptome characterization," *BMC Genomics*, vol. 10, p. 347, 2009.
[45] J. A Sparano, M. Fazzari, and P. A. Kenny, "Clinical application of gene expression profiling in breast cancer," *Surg. Oncol. Clin. North Amer.*, vol. 19, pp. 581–606, 2010.
[46] R. M. Parry, W. Jones, T. H. Stokes, J. H. Phan, R. A. Moffitt, H. Fang, L. Shi, A. Oberthuer, M. Fischer, W. Tong, and M. D. Wang, "k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," *Pharmacogenomics J.*, vol. 10, no. 4, pp. 292–309, Aug. 2010.
[47] E. A. H. Kheirelseid, N. Miller, K. H. Chang, M. Nugent, and M. J. Kerin, "Clinical applications of gene expression in colorectal cancer," *J. Gastrointest Oncol.*, vol. 4, no. 2, pp. 144–157, Jun. 2013.
[48] A. Sveen, A. Nesbakken, T. H Agesen, M. G. Guren, K. M. Tveit, R. I. Skotheim, and R. A. Lothe, "Anticipating the clinical use of prognostic gene expression-based tests for colon cancer stage II and III: Is godot finally arriving?" *Clin. Cancer Res.*, vol. 19, no. 24, pp. 6669–6677, Dec. 2013.
[49] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics," *CA: A Cancer J. Clinicians*, vol. 64, pp. 9–29, 2014.
[50] "International Agency for Research on Cancer (IARC)," WHO, B. W. Stewart and C. P. Wild eds., World Cancer Report, 2014.
[51] R. Simon, "Roadmap for developing and validating therapeutically relevant genomic classifiers," *J. Clin. Oncol.*, vol. 23, pp. 7332–7341, 2005.

Shaurya Jauhari received the BS degree in statistics, mathematics, and computer applications from the University of Lucknow, India, in 2005 and the MS degree in computer applications from Uttar Pradesh Technical University, India, in 2009. He has been a research scholar in bioinformatics at Jamia Millia Islamia, New Delhi, India, since September 2012. Prior to working here, he served as a network assistant and a software engineer, from 2009 to 2012. His research interests include cancer therapeutics, biomedical informatics, and biostatistics.

S.A.M. Rizvi is serving as an associate professor in the Department of Computer Science, Jamia Millia Islamia, New Delhi, India. With a vast academic experience and revered credibility, He is a frontrunner in interdisciplinary research and cross-platform training. His research interests include computer algorithms, artificial intelligence, bioinformatics, and design theory.