Bioinformatics: a Promising field for Case-Based Reasoning

Fernando Orduña Cabrera and Miquel Sànchez-Marrè
Technical University of Catalonia (UPC)
Campus Nord-Building Omega Software Dept. (LSI)
Jordi Girona 1-3, E08034, Barcelona, Spain
{forduna, miquel}@lsi.upc.edu

Abstract. Case Based Reasoning has been applied in different fields such as medicine, industry, tutoring systems and others, but in the CBR there are many areas to explore. Nowadays, some research works in Bioinformatics are attempting to use CBR like a tool for classifying DNA genes. Specially the microarrays have been applied increasingly to improve medical decision-making, and to the diagnosis of different diseases like cancer. This research work analyzes the Microarrays structure, and the initial concepts to understand how DNA structure is studied in the Bioinformatics' field. In last years the CBR has been related to Bioinformatics and Microarrays. In this report, our interest is to find out how the Microarrays technique could help in the CBR field, and specially in the Case-Based Maintenance policies.

Keywords. Bioinformatics, Microarrays, Case-Based Reasoning.

1. Introduction

DNA microarray technology is attracting tremendous interest both among the scientific community and in industry. With its ability to measure simultaneously the activities and interactions of thousands of genes, this modern technology promises new insights into the mechanisms of living systems. Typical scientific questions addressed by microarray experiments include the identification of co-expressed genes, either as genes expressed throughout a sub-population or as genes always expressed together (sample or gene groups), identification of genes whose expression patterns make it possible to differentiate between biological entities that are otherwise indistinguishable, and the study of gene activity patterns under various stress conditions (e.g., chemical treatment). Although microarrays have been applied in many biological studies, the handling and analysis of the large volumes of data generated is not trivial. Different types of scientific questions require different data analytical techniques. Broadly speaking, there are two classes of elementary data analysis tasks, predictive modeling and pattern-detection. Predictive modeling tasks are concerned with learning a classification or estimation function, whereas pattern-detection methods screen the available data for interesting and previously unknown regularities or relationships. A wide range of sophisticated methods and tools have been developed to address these tasks.

One of the main reasons why researchers are pursuing DNA microarray studies with such intensity, in the full knowledge of their limitations, is the fact that protein expression and modification studies are still very expensive, and often involve highly specialized and delicate techniques.

CBR has been applied successfully in different fields, but in the field of Bioinformatics has been applied relatively less frequently. Actually there are a few papers in this field. Molecular Biology domain is a natural application for CBR systems, since CBR can perform remarkably well on complex and poorly formalized domains. CBR is specially used in different areas which incorporate the fields of biology, computer science among others. One of the applications in this research line is the sequence analysis where involves the study of organism's DNA, to understand its molecular structure which has led to the discovery of mutation in DNA and chromosomal abnormalities indicative of diseases such cancer. In this field of cancer diseases there are research work such as [2,3,4] where CBR is applied to help to the studies of DNA. Specially in research [2], has as main task the gene finding by applying CBR by employing a case library of nucleotide segments that have previously been categorized. While in the research work [3] presents an approach to performing the case-based paradigm in the presence of noisy DNA sequences (case boundaries), applied in the domain of molecular biology. And GENE-CBR application to study cancer treatment is presented in [4], where presents a reduction algorithm based on notion of fuzzy codification for gene expression level employed to implement the retrieval step.

Others research works are focused in the use of a CBR classifier attribute as a classifier in microarray data field. For instance, in the research works [7,9] propose a maintenance technique that integrates an ensemble of CBR classifiers with spectral clustering to improve classification accuracy of CBR on (ultra) high-dimensional biological data sets. The main challenge of [7,9] is to interpret the molecular biology data, and find similar samples to eventually use them in case-based medicine, and to identify genes with meaningful relationships.

The aim of this research is to describe in short terms, how the Bioinformatics field is becoming interesting to the CBR community, and to show some research works such as there cited before. That involves the CBR, Bioinformatics and the Microarrays into novelties applications. To understand and using microarrays analysis techniques requires a basic understanding of the fundamental mechanism of gene expressions, for that the research begins with a brief description of it. One time, with the knowledge of DNA the microarray data analysis is explained in the Bioinformatics section. Microarrays as CBR have many topics of data mining, for it an introduction of each topic is introduced in the same section of Bioinformatics. In features sections the task of find genes with CBR is described taking into account some research works. The CBR community has been working in systems that could help making advices in the treatment of the cancer disease, for that some research works are described, and finally the use of microarray with CBR are described based on the research works of Niloofar Arshadi.

2. Fundamentals Topics of Bioinformatics

Understanding and using microarray analysis techniques requires a basic understanding of the fundamental mechanisms of gene expression itself. Describing gene expression starts necessarily with deoxyribonucleic acid (DNA), the very stuff genes are made from, and ribonucleic acid (RNA). Both DNA and RNA are polymers, that is, molecules that are constructed by sequentially binding members of a small set of subunits called nucleotides into a linear strand or sequence. Each nucleotide consists of a base, attached to a sugar, which is attached to a phosphate group. The linear strand consists of alternate sugars and phosphates, with the bases protruding from the sugars. In DNA, the sugar is deoxyribose and the bases are named guanine, adenine, thymine, and cytosine; in RNA the sugar is ribose and the bases are guanine, adenine, uracil, and cytosine [8]. The sugar phosphate backbone can, for the purposes of informatics, be considered as straight (though actually it has all sorts of twists, kinks and loops (higher order structures) that are of interest to those who care about such things). The bases that protrude from the backbone are far more informative. They can form pairs, via hydrogen bonds, with bases in other nucleic acid strands: adenine binds to thymine (or uracil) and guanine to cytosine, by the formation of two and three hydrogen bonds respectively. Such base pairing allows DNA to be organized as a double-stranded polymer whose characteristic three-dimensional helix structure has become famous. The two DNA strands are complementary to each other, meaning that every guanine in one strand corresponds to a cytosine in the other (complementary) strand. The same mechanisms apply to the complementary DNA nucleotides adenine and thymine.

Within the DNA, genes are unique sequences of variable length. The genes within a cell comprise the cell's genome: it contains the information necessary for synthesizing (constructing) proteins, which do all that a cell needs. The genome also contains the information that controls which proteins are synthesized in a given cell under particular circumstances. Implicit in the structure of a cell's genome are mechanisms for self replication and for transforming gene information to proteins. The gene to protein transformation constitutes the "central dogma of molecular biology"; it is described by a two-step process (Expression): ***Step 1: Transcription: Gene(DNA) makes RNA, Step 2: Translation: RNA makes protein.***

That is, the information represented by the DNA sequence of genes is transferred into an intermediate molecular representation, an RNA sequence, using part of a DNA strand as a template for assembling the RNA. The information represented by the RNA is then used as a template for constructing proteins, according to a code in which each amino acid is represented by three bases in the RNA. The RNA occurring as intermediate structure is referred to as messenger RNA (mRNA). The term transcription is commonly used to describe Step 1 and the term translation for Step 2. Collectively, the overall process consisting of transcription and translation is known as gene expression. Notice, in most organisms only a small subset of genomic DNA is capable of being transcribed to mRNA or expressed as proteins. Some regions of the genome are devoted to control mechanisms, and a substantial amount of the genomes of higher-level organisms appears to serve no informational function at all. These DNA sections are also known as junk DNA.

Hybridization

This section will illustrate the biochemical principles involved in measuring transcripts with microarrays. A number of techniques have been developed for measuring gene expression levels, including northern blots, differential display, and serial analysis of gene expression. DNA and oligonucleotide microarrays are the latest in this line of methods. They facilitate the study of expression levels in parallel. All these techniques exploit a potent feature of the DNA duplex (the sequence complementarity of the two strands). This feature makes hybridization possible. Hybridization is a chemical reaction in which single-stranded DNA or RNA molecules combine to form double stranded complexes (see schematic illustration in Figure 1). The famous DNA double helix is an example of such a molecular structure. The hybridization process is governed by the base-pairing rules: specific bases in different strands form hydrogen bonds with each other. For DNA the matching pairs are adenine-thymine and cytosine-guanine. Hybridization is a nonlinear reaction. Yield the number or concentration of nucleic acid elements binding with each other in the resulting double-stranded molecule depends critically on the concentration of the original single-stranded polymers and on how well their sequences align or match. It is this yield that is measured in a microarray experiment. In order to selectively detect and measure the amount of mRNA that is contained in an investigated sample, we must label the mRNA with reporter molecules. The reporters currently used in microarray experiments include fluorescent dyes (fluors), for example, cyanine 3 (Cy3) and cyanine 5 (Cy5). Assume having two samples of transcribed mRNA from two different sources, sample 1 and sample 2. Both samples may consist of multiple copies of many genes. They also a probe, which is a specific nucleic acid sequence, perhaps a gene, or a characteristic subsequence of a gene, or a short, artificially composed nucleotide sequence. Like the two samples, the probe will contain many copies of the sequence in question. This is important because sufficient amounts are needed to get the hybridization reaction going, and to be able to detect and measure the various concentrations. If want to find out is the relative abundance of the mRNA complementary to the probe sequence within sample 1 and sample 2. Sample 1, for example, may contain three times as many copies of sequences complementary to the probe as sample 2, or they may not be contained in either sample at all. To find out the exact answer, proceed as follows (see also Figure 2):

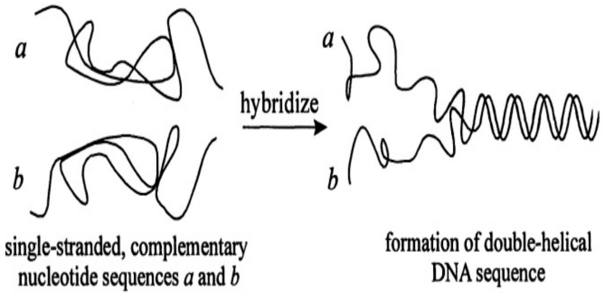
a
hybridize
a
b
b
single-stranded, complementary nucleotide sequences *a* and *b*
formation of double-helical DNA sequence

Figure 1. Hybridization of two single-stranded nucleic acid sequence to double-stranded, helical complex.

1. Prepare a mixture consisting of identical probe sequences. In this scheme of things the probe is a kind of "sitting duck", awaiting hybridization.
2. Label sample 1 with green-dyed reporter.
3. Label sample 2 with red-dyed reporter.

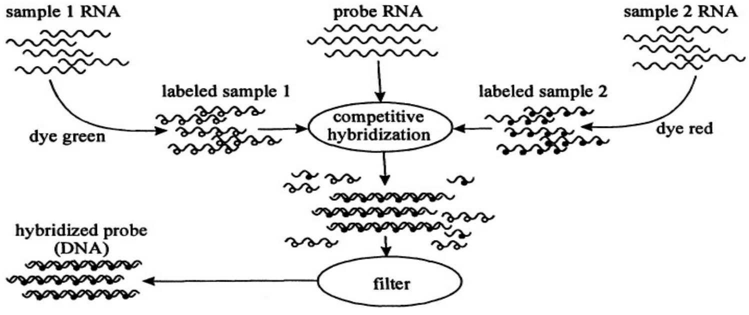

Figure 2. Hybridization of two single-stranded nucleic acid sequence to double-stranded, helical complex.

4. Simultaneously give both sample mixtures the chance to hybridize with the probe mixture. Here, sample 1 and sample 2 are said to compete with each other in an attempt to hybridize with the probe.
5. Gently stir for five minutes.
6. Filter the mixture to retain only those probe sequences that have hybridized, that is, formed a double-stranded polymer.
7. Measure the amount or intensity of green and red in the filtered mixture, and compare the amounts to determine the relative abundance of the probe sequence.
8. Jot down the result, add a little salt, and enjoy.

The Matrix and Data

Microarray data analysis involves methodologies and techniques from life science fields and biotechnology on one hand, and from computer science and statistics on the other. With these broad disciplines comes a lot of heavy baggage in the form of terminology.

The term gene expression profile is commonly used to describe the expression values for a single gene across many samples or experimental conditions, and for many genes under a single condition or sample. They suggest the following terms to distinguish these types of gene expression profiles (see Figure 3):

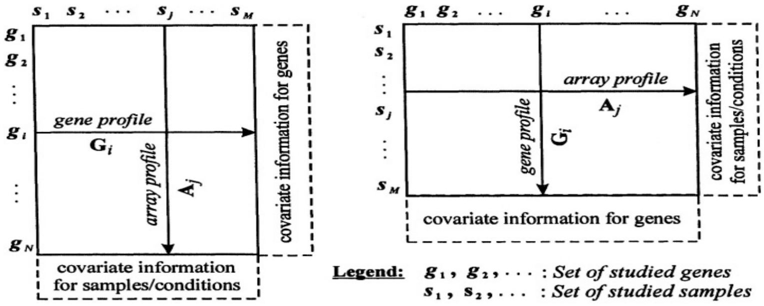
s_1 s_2 ... s_j ... s_M
g_1 g_2 ... g_i ... g_N
gene profile
$\mathbf{G}_i$
array profile
$\mathbf{A}_j$
covariate information for genes
covariate information for samples/conditions
g_1 g_2 ... g_i ... g_N
s_1 s_2 ... s_j ... s_M
array profile
$\mathbf{A}_j$
gene profile
$\mathbf{G}_i$
covariate information for samples/conditions
covariate information for genes
Legend: $g_1, g_2, \dots$: Set of studied genes
$s_1, s_2, \dots$: Set of studied samples

Figure 3. Typical gene expression data matrix formats. The solid-line boxes contain the actual numerical values representing the measured expression levels

One gene over multiple samples. A gene profile is a gene expression profile that describes the expression values for a single gene across many samples or conditions.

Many genes over one sample. An array profile is a gene expression profile that describes the expression values for many genes under a single (condition or) sample.

The diagram in Figure 3 depicts two commonly employed data formats for the integrated gene-expression data matrix. Left part in Figure 3. Concentrating on the gene expression part only, this format can be described by an *N x M* expression matrix *E*, as defined in Equation (1)

$$\mathbf{E} = (\mathbf{X_{ij}}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix} \tag{1}$$

where X_{ij} denotes the expression level of sample *j* for gene *i*, such that *j* = 1, . . . *M*, and *i* = 1, . . . N (Dudoit et al., 2000).

The expression matrix **E** is a convenient format to represent the expression profiles of *N* genes over *M* samples. Whether column vectors or row vectors in this matrix are interpreted as variables (or observations) is not pre-determined by the matrix format. However, the matrix does nail down the interpretation in terms of gene and array profile. With Equation (1) define the i^{th} *gene profile* of expression matrix **E** by the row vector $\mathbf{G}_i$, and the j^{th} *array profile* of **E** by the column vector $\mathbf{A}_j$, as follows: $\mathbf{G}_i = (X_{i1}, X_{i2}, \dots, X_{iM})$ and $\mathbf{A}_j = (X_{j1}, X_{j2}, \dots, X_{jN})$.

The Microarrays

Based in the definition of a microarray of [16], "a microarray is an ordered array of microscopic elements on a planar substrate that allows the specific binding of genes or gene products. Microarray is a new scientific word derived from the Greek word *mikro* (small) and the French word *arayer* (arranged). Microarrays (also know as biochips, DNA chips, and gene chips) contain collections of small elements or spots arranged in rows and columns. To qualify as a microarray, the analytical device must be (1) ordered, (2) microscopic, (3) planar, and (4) specific. Devices that fullfill only a sub set of these criteria do not afford the advantages of microarrays, do not qualify as microarrays, and should not be considered as such".

DNA microarrays analyse (1) the mRNAs in a cell, to reveal the expression patterns of proteins; or (2) genomic DNAs, to reveal absent or mutated genes. 1. For an integrated characterization of cellular activity, if we want to determine what proteins are present, where and in what amounts. To find the expression pattern of a cell's genes, measure the relative amounts of many different mRNAs. Hybridization is an accurate and sensitive way to detect whether any particular nucleic acid sequence is present. The key to high-throughput analysis is to run many hybridization experiments in parallel.

DNA microarrays, or DNA chips, are devices for checking a sample simultaneously for the presence of many sequences.

Microarrays provide an efficient, high-throughput way of carrying out these tests in parallel. They also permit measuring the expression levels of thousands or even tens of thousands of genes in a single sample

To achieve parallel hybridization analysis, a large number of DNA oligomers are affixed to known locations on a rigid support, in a regular two-dimensional array. The mixture to be analysed is prepared with fluorescent tags, to permit detection of the hybrids. After exposing the array to the mixture, each element of the array to which some component of the mixture has become attached bears the tag. Because the sequence of the oligomeric probe in each spot in the array is known, measurement of the positions of the probes identifies their sequences. This analyses the components present in the sample.

Different types of chips designed for different investigations differ in the types of DNA immobilized.

- In an expression chip, the immobilized oligos are cDNA samples, typically 20-80 base pairs long, derived from mRNAs of known genes. The goal of such an experiment is to determine the expression patterns of genes that correspond to the cDNA samples. This is by far the most common application of microarrays. The target sample might be a mixture of mRNAs from normal or diseased tissue.
- In mutation microarray analysis, one looks for patterns of single-nucleotide polymorphisms (SNPs).
- In genomic hybridization, one looks for gains or losses of genes, or changes in copy number. The probe sequences, fixed on the chip, are large pieces of genomic DNA, from known chromosomal locations, typically 500-5000 base pairs long. The probe mixtures contain genomic DNA from normal or disease states. For instance, some types of cancer arise from chromosome deletions, which can be identified by microarrays.

Microarrays are capable of comparing concentrations of components of the sample. This allows quantitative investigation of responses to changed conditions. However, the precision is low. Moreover, mRNA levels, detected by the array, do not always accurately reflect protein levels. Indeed, usually mRNAs are reverse transcribed into more stable cDNAs for microarray analysis; the yields in this step may also be non uniform. Microarray data are therefore semiquantitative, in that a distinction between presence and absence is possible, determination of relative levels of expression in a controlled experiment is more difficult, and measurement of absolute expression levels are beyond the capacity ofcurrent microarray techniques.

Analysis of microarray data

The raw data of a microarray experiment is an image, in which the colour and intensity of the fluorescence reflect the extent of hybridization to alternative probes. The two sets of probes are tagged with red and green fluorophores. If only one probe hybridizes, the spot appears red; if only the other probe hybridizes. The spot appears green. If both hybridize, the colour of the corresponding spot appears red + green = yellow. The initial goal of data processing is a gene expression table. This is a matrix in which the rows correspond to different genes, and the columns to different samples. Different spots in a microarray pattern correspond to different genes. For each gene, results from different sets of samples appear in the red or green channel, respectively (or neither, or both). There is extensive redundancy in the oligos in a microarray each gene may be represented by several spots, corresponding to different regions of the gene sequence; inclusion of controls with a deliberate mismatch allows data verification.

The samples may vary according to experimental conditions and/or physiological states. Or they may be extracted from different individuals, different tissues or different developmental stages. The process of data reduction to produce the gene expression matrix involves many technical details of image processing, checking internal controls, dealing with missing data. Selecting reliable measurements, and putting the results of different arrays on consistent scales. The derived gene expression table indicates relative expression levels. Extraction of reliable biological information from a gene expression table is not Straight-forward. Despite extensive internal controls, there is considerable noise in the experimental technique. In many cases, variability is inherent within the samples themselves, this topics are discussed more in detail in future section. Microorganisms can be cloned; animals can be inbred to a comparable degree of homogeneity. However, experiments using RNA from human sources-for example, a set of patients suffering from a disease and a corresponding set of healthy controls-are at the mercy of the large individual variations that humans present. Indeed, inbred animals, and even apparently identical eukaryotic tissue-culture samples, show extensive variability. Another intrinsic disadvantage-and a severe one-in interpreting gene expression data, is the fact that the number of genes is much larger than the number of samples. Two general approaches to the analysis of a gene expression matrix involve

- Comparisons focussed on the genes; that is, comparing distributions of expression patterns of different genes by comparing rows in the expression matrix: How do gene expression patterns vary among the different samples? Suppose a gene is known to be involved in a disease, or to a change in physiological state in response to changed conditions. Other genes co-expressed with the known gene may participate in related processes contributing to the disease or change in state. More generally, if two rows (two genes) of the gene expression matrix show similar expression patterns across the samples, this suggests a common pattern of regulation, and possibly some relationship between their functions, including but not limited to a possible physical interaction.
- comparisons focussed on samples; that is, comparing expression profiles of different samples by comparing columns of the expression matrix: Comparisons focussed on samples: How do samples differ in their gene expression patterns? A consistent set of differences among the samples may characterize the classes from which the samples originate. If the samples are from different controlled groups (for instance, diseased and healthy animals), do samples from different groups show consistently different expression patterns? If so, given a novel sample, we can assign it to its proper class on the basis of its observed gene expression pattern.

To know more details of the matrix SVD (Singular Value Descomposition) see chapter 2 from the book [15].

How then do we measure the similarity of different rows or columns? Each row or column of the expression matrix can be considered as a vector, in a space of many dimensions. The row-vectors (a row corresponds to a gene), each entry of which refers to the same gene in different samples, has as many elements as there are samples. The column-vectors (a column corresponds to a sample), each entry of which refers to a different gene in a single sample, has as many elements as there are genes reported. It is possible to calculate the 'angle' between different row-vectors, or between different column-vectors, to provide a measure of their similarities. It is then natural to ask whether subsets of the points form natural clusters-points with high mutual similarity-characterizing either sets of genes or sets of samples.

Minimal Information About a Microarray Experiment

Microarray analysis results in the gathering of massive amounts of information concerning gene expression profiles of different cells and experimental conditions. Analyzing these data can often be a quagmire, with endless discussion as to what the appropriate statistical analyses for any given experiment might be. As a result many different methods of data analysis have evolved, the basics of which are outlined in this chapter.

In an effort to standardize the thousands of array experiments, the Microarray Gene Expression Database (MIAME) society established guidelines that require researchers to conform to MIAME guidelines. MIAME describes the minimal information about a microarray experiment that is required to interpret the results of the experiment, and compare it with other experiments from other groups. The checklist for complying with the MIAME guidelines is quite extensive and can be found at $http : //www.mged.org/Workgroups/MIAME/miame_{c}hecklist.html$ In brief, these guidelines include:

- Array design: information regarding the platform of the array, description of the clones and oligomers, and catalog numbers for commercial arrays. This also should include the location of each feature as well as the explanations of feature annotation.
- Experimental design: a description and the goals of the experiment, rationale for cells/tissues and treatment used, quality control steps, and links to any public databases necessary.
- Sample selection: criteria for the selection of samples, description of the procedures used for RNA extraction, and sample labeling.
- Hybridization: conditions of hybridization, including blocking and washing of slides.
- Data analysis: description of the raw data, as well as of the original images, hardware, and software used, and also the criteria used for processing and normalization of data.

Image Acquisition and Analysis

Once the RNA has been isolated and hybridized to the chip, the first stage of data analysis begins. This requires successful acquisition of the fluorescent or radioactive signal bound to the chip or membrane. With radioactive membranes, it is standard procedure to expose the membrane several times and then take an educated average of the best exposures. With fluorescent dyes, it is essential to utilize a high-resolution scanner and that the first scan be performed as quickly and accurately as possible, as the dyes are quickly bleached and multiple scans are not possible. Some salient points of image acquisition are outlined next.

Quality of Scanner

It is important to use a scanner that can detect at a resolution of 10 microns or greater. In addition, the scanner must be able to excite and detect Cy3 (532 nm) and Cy5 fluorescence (633 nm). An adjustable photomultiplier tube to ensure equal scanning, while reducing as much bleaching as possible, is also ideal. Typically, the settings for the photomultiplier tube are around 30

Orientation of Image

The orientation of the image becomes particularly important when combining arrays from one company with a scanner from a different company as images may be inverted depending on the scanner being used. Thus, it is crucial that the array include "landing lights"-control cDNAs or oligonucleotides spotted on the arrays that yield a distinct pattern when the array is in the correct orientation.

Spot Recognition

Often referred to as "gridding," this is the process used to identify each spot on the array prior to extracting information from it. When purchasing arrays and scanners from commercial sources, programs for spot recognition and information extraction are often included. Agilent and Affymetrix both have their ownfeature extractor software, which uses control spots on the array for automated spot recognition and feature extraction. Many other programs require that the user intervene and flag "bad" spots, and realign grids to fit the spots.

Segmentation

Once grids have been placed, information as to the pixel intensity within the spots must be extracted. This process is known as segmentation. Various methods exist to perform this including fixed circle segmentation, adaptive circle segmentation, fixed shape segmentation, adaptive shape segmentation, and seeded region growing method (also known as the histogram-based method).

Analysis of the Quality of the Hybridization

All of these imaging parameters can then be used to analyze the quality of the microarray experiment. Intensities in each channel should ultimately cluster around a central norm in a Gaussian distribution. Background intensity abnormalities can be calculated statistically by computing the average background intensity and using the standard deviation among this intensity to calculate a confidence interval, the upper limit of which is used to assume background correction.

Data Normalization

In order to normalize the information received from a microarray experiment, several methods have been designed and are outlined next.

Data Transformation

After background correction has been performed, the data must be transformed for statistical analysis. The analyses applied to the data (e.g., parametric vs nonparametric) determine the type of transformation that must be performed. Parametric tests are the most commonly utilized, as these tests are much more sensitive and require the data to be normally distributed. This is often achieved by using log transformation of the spot intensities to achieve a Gaussian distribution of the data.

Differential Gene Expression

Differential gene expression is often measured by the ratio of intensity (as a measure of expression level) between two samples. Many early microarray experiments assigned a fold-change cutoff, and considered genes above this fold-change significant. In addition, several statistical analyses can be utilized including maximum-likelihood analysis, F-statistic, ANOVA (analysis of variance), and t-tests. The results of these tests can often be improved by log transformation of data as mentioned previously, and by random permutations of the data.

Pattern Discovery

Often called exploratory or unsupervised data analysis, this approach can encompass a number of different techniques listed next that allow for a global view of the data. These methods often rely on clustering techniques that allow for quick viewing of distinct gene expression patterns within a dataset. Cluster analysis is available free of charge as part of the gene expression omnibus, a site that attempts to catalog gene expression data (16), providing a valuable data mining resource (http://www.ncbi.nlm.nih.gov/geo/). Dimension reduction techniques such as principal component analysis (PCA) and multidimensional scaling analysis can often be used in conjunction with other supervised techniques such as artificial neural networks to provide even more robust data analysis.

PCA

PCA can analyze multivariate data by expressing the maximum variance as a minimum number of principal components. Redundant components are eliminated, thus reducing the dimensions of the input vectors. For information on the mathematical origins of this equation, see http://www.cis.hut.fi/ jhollmen/ dippa/node30.html.

Multidimensional Scaling

This analysis is often based on a pair-wise correlation coefficient and assesses the similarities and dissimilarities between samples and assigns the difference as a "distance" between samples, such that the more similar two samples are, the closer they are together, and vice versa. The multi- as opposed to twodimensional analysis comes into play when not only the degree of difference (distance) but also the spatial relationship of three or more samples to each other (direction) is taken into account. For further mathematical description of this process, see http://www.statsoft.com/textbook/stmulsca.html.

Singular Value Decomposition

Singular value decomposition (SVD) treats microarray data as a rectangular matrix, A, which is composed of n rows (genes) by p columns (experiments). SVD is represented by the mathematical equation.

Hierarchical Clustering

Perhaps the most familiar to biologists, hierarchical clustering presents the data as a gene list organized into a dendrogram, and is a bottom-up analysis. This is obtained by assigning a similarity score to all gene pairs, calculating the Pearson's correlation coefficient, and then building a tree of genes by replacing the two most similar genes with a node that contains the average, then repeating the process for the next closest pair of data points, and then the next. This process is repeated several times (iterative process) to generate the dendrogram or Treeview, as well as heat maps that represent a two-color checkerboard view of the data.

K-Means Clustering

K-means clustering is a top down technique that groups a collection of nodes into a fixed number of clusters (*k*) that are subjected to an iterative process. Each class must have a center point that is the average position of all the distances in that class (representative element), and each sample must fall into the class to which its center is closest. Fuzzy k-means is performed by "soft" assignment of genes to these clusters.

Self-Organizing Maps

These maps are basically two dimensional grids containing nodes of genes in "*K*" dimensional space. These can be represented by sample and weight vectors, which are composed of the data and their natural location. Weight vectors are initialized, and then sample vectors are randomly selected to determine which weight best represents that sample, and these are used to map the nodes into K-dimensional space into which the gene expression data falls. Like the previously mentioned methods, this is also iterative and is often repeated more than 1000 times, and these methods can often be used in combination to generate the best overview of the data.

Class Prediction

Class prediction is based on supervised data analysis methods that impose known groups on datasets. First, a training set is identified-this is a group of genes with a known pattern of expression that is used to "train" a dataset, by comparing the data to the training set and thus classifying it. This particular method is very useful in the subclassification of similar samples, cancer diagnosis, or to predict cell or patient response to drug therapy. In some cases, this type of analysis has also been used to predict patient outcome, allowing for a very clinically relevant use of microarray data. Importantly, gene selection by these methods relies on the assignment of discriminatory weights to these genes, i.e., how often a single gene correlates to a given class or phenotype, often calculated using random permutation tests. Random permutation tests are also used to calculate *p* (probability the weight can be obtained by chance) and (probability of high weight resulting from random classification) values for these weights. Many different statistical methods can be used to find discriminant genes.

Fisher Linear Discriminant Analysis

This theory assumes that a random vector *x* has a multivariate normal distribution between each defined class or group, and the covariance within each group is identical for all the groups. This makes the optimal decision function for the comparison of data a linear transformation of *x*. Variations on this theme include quadratic discriminant analysis, flexible discriminant analysis, penalized discriminant analysis, and mixture discriminant analysis.

Nearest-Neighbor Classification

These methods are based on a measure of distance (e.g., Euclidean distance) between two gene expression profiles. Observations are given a value (x) and the number of observations (k) closest to x is used to choose the class. The value of k can be determined by using cross-validation techniques.

Support Vector Machines

This type of analysis is based on constructing planes in a multidimensional space that separate the different classes of genes, and set decision boundaries using an iterative training algorithm. Data is mapped into the higher dimensional space from its original input space, and a nonlinear decision boundary is assigned. This plane is known as the maximal margin hyperplane, and can be located by the use of a kernel function (a nonparametric weighting function). For further mathematical description, see http://www.statsoft. com/textbook/stsvm.html.

Artificial Neural Networks

Neural networks, or perceptrons, another machine-learning technique, are so named because they model the human brain-they learn by experience. Multilayer perceptrons can be used to classify samples based on their gene expression. Gene expression data for a sample are input into the model, and a response is generated in the next layer, ultimately triggering a response in the output layer. This output perceptron should represent the class to which the sample belongs.

Decision Trees

These are built by using criteria to divide samples into nodes. Samples are divided recursively until they either fall into partitions, or until a termination condition is met. Ultimately the intermediate nodes represent splitting points or partitioning criteria, and the leaf nodes represent those decisions.

Data Validation

As complex and robust as the available analyses for microarray data currently are, there is always room for error, and many inherent problems in the experimental technique. Thus, it is critical that researchers validate their data before drawing any firm biological conclusions from the data. One of the most common techniques for validating array data is the use of real-time PCR. Real-time PCR effectively quantitates differences in transcript levels between different samples, but it must be remembered that the ratios acquired from a microarray experiment are quite likely to be much lower than fold changes seen in real-time PCR, as this method is much more sensitive. Ultimately, protein expression is of course the final confirmation, as most gene expression-profiling experiments, whether of a classifier or exploratory nature seek protein markers, and this is most often confirmed using immunohistochemistry.

3. The Use of CBR in Bioinformatics

Intelligent support is essential for managing and interpreting this overwhelming amount of data, and one of the most important tasks faced currently is the analysis of sequences of nucleotides in order to locate the areas of DNA that actually encode functional biological information [2].

Bioinformatics incorporates the fields of biology, computer science and information technology with the goal of discovering new biological insights and the enhancement of diagnostic and pharmaceutical medicine. Sequence analysis, which involves the study of an organism's DNA in an effort to understand its molecular structure and underlying functionality, is of major importance to the area of gene therapy, which has led to the discovery of mutations in DNA and chromosomal abnormalities indicative of diseases such as cancer. Thus the analysis of nucleotide sequences, in particular the identification of DNA segments that encode functional biological information, can provide the medical profession with invaluable insight into the pathology of disease state and treatment.

Sequence analysis first involves determining the basic molecular structure of an organism's DNA, which is simply a molecule made up of two strands, with each strand comprised of nucleotides from a finite set. There are four different nucleotides-adenine, guanine, thiamine and cytosine, and the first letter of each provides an alphabet {A,G, T,C} for representing DNA. A nucleotide to nucleotide bond holds the two strands together with each nucleotide being bonded to its complementary match, a only bonding to T and C only bonding to G. Therefore, given one strand (i.e., half a DNA molecule), the complementary strand can be reconstructed relatively easily. Determining the basic molecular DNA structure sequence is a well understood task that is providing a deluge of information for interpretation. The task of "gene finding," identification of coding regions in an organism's DNA, is the next essential step in analysing an organism's genome. These coding regions are called exons and when these are put together they form an entire gene. It is the genes that tell the body how to create proteins, and it is these that give rise to biological function. Exons are continuous sequences in a strand that the body uses to replicate proteins; the parts in between these exons do not contribute to protein replication and are called introns, or non coding regions.

- In the research work [2] are interested in applying CBR to the gene-finding problem by employing a case library of nucleotide segments that have previously been categorised as coding (exon) or non-coding (intron), in order to locate the coding regions of a new DNA strand. The work describes the initial research in developing a CBR approach to the problem of finding regions in mammalian DNA that code proteins essential for life.
  An initial system was tested and evaluated the results and conclude it is evident that a simple case-based reasoning approach to the recognition of coding regions is certainly possible.
  They presented their initial work in applying a case-based approach to the problem of gene-finding in mammalian DNA. The results obtained from the approach indicate that it is certainly feasible to do DNA-to-DNA comparisons in order to isolate relevant coding regions. Using DNA sequences, avoids the need for the translation of the sequence to the different protein sequence possible.
- In the research work [3] presents an approach to performing the case-based paradigm in the presence of noisy DNA sequences (case boundaries). The ap-

proach has been fully implemented and applied in the domain of molecular biology; specifically, a successful case-based approach to gene finding is described. They have developed a case-based algorithm for gene finding that is robust in the presence of frame shift errors. The FIND-IT algorithm builds on the BLAST similarity-search program. BLAST efficiently produces approximate matches, but these matches do not extend across frame shift errors. The FIND-IT method described below coherently combines partial matches (to a given protein) in different reading frames, thereby overcoming missing and extra nucleotides in sequenced DNA.

Is presented a case-based approach to gene finding that is robust in the presence of errors both in the input data and in the case libraries. The research addresses the important general question of what makes a case; more specifically, how do parse the "noisy" world into discrete cases for matching against a case library? If the current case is improperly delimited, partial matching with previous cases will fail. The algorithm addresses the problem by producing multiple, partial matches to many cases and then combining some subset of them into a consistent whole. This leads to error detection and correction. The general idea for robust case matching promises to be applicable in other domains involving "continuous" data, such as speech recognition and vision.

- The research work [4] presents a reduction algorithm based on the notion of fuzzy gene expression, where similar (co-expressed) genes belonging to different patients are selected in order to construct a supervised prototype-based retrieval model. This technique is employed to implement the retrieval step in the new gene-CBR system. They propose a fuzzy codification for the gene expression levels of each sample based on the discretization of real gene expression data into a small number of fuzzy membership functions.

  In this work, is presented a fuzzy codification for the gene expression levels of microarray data, based on the discretization of real gene expression data into a small number of fuzzy membership functions. The proposed method aims to find all genes that are significantly expressed between the existing classes in order to obtain a fuzzy representation of the expression levels belonging to those genes that best explain each class in the form of a fuzzy-prototype.
- In [5] presents GENE-CBR, a hybrid model that can perform cancer classification based on microarray data. The system employs a case-based reasoning model that incorporates a set of fuzzy prototypes, a growing cell structure network and a set of rules to provide an accurate diagnosis.

  The hybrid system proposed presents a new synthesis that brings several artificial intelligence subfields together (fuzzy sets, artificial neural networks, and if-then rule-sets).

  GENE-CBR is a model that can perform cancer classification based on microarray data. GENE-CBR employs a CBR model that incorporates a set of fuzzy prototypes for the retrieval of relevant genes, a growing cell structure (GCS) network and a proportional weighted voting algorithm for the clustering of similar patients and the assignation of an initial class, and a set of rules used to formalize the knowledge extraction to justify the results. For the experiments reported they work with a database of bone marrow (BM) cases from forty-three adult patients with acute myeloid leukemia (AML) plus a group of six samples belonging to healthy persons for test purposes. Each case (microarray experiment) stores 22,283 ESTs corresponding to the expression level of thousands of genes. The data consisted of 1,025,018 scanned intensities.

To initially construct the model case base starting from the available patients data, GENE-CBR stores the gene expression levels of each sample in its case base. The system always deals with a fuzzy codification of the values stored. During the retrieval stage, the original case vectors are transformed into a fuzzy microarray descriptors (FMD). Each FMD is a comprehensible descriptor of the sample in terms of a linguistic label for each gene expression level. This transformation is carried out by mean of a fuzzy discretization process.

The retrieval stage in the GENE-CBR system uses the FPs to select the most representative genes given a new patient. The GENE-CBR system selects those FP from its case base which are the nearest to any new case obtained. Then, for each one of the selected FPs, the GENE-CBR system computes its associated discriminant fuzzy pattern (DFP) (a pattern which only includes the genes that are necessary to discriminate the novel instance from other different classes). Finally, the selected genes for the new case are obtained by joining together the genes belonging to the DFPs considered. A GCSs network is trained with the whole case base, only taking the existing cases represented by the genes selected in the previous stage as input. A proportional weighted voting mechanism is applied that ponders the level of similarity with the new FMD.

They are interested in the development of a robust case-based reasoning system that may be employed in the study of cancer treatment. The goal of the decision support tool is to facilitate the construction of therapies, including the level of aggressiveness of treatment, to more closely match the underlying disease, hopefully reducing side effects in low risk cases and increasing cure rates in high-risk cases. Input space reduction is often the key phase in the building of an accurate classifier. Based on the fuzzy discretization method presented, and propose the use of a fuzzy prototype-based retrieval system able to differentiate several kinds of cancer for microarray data.

The proposed method aims to find all genes that are significantly expressed between the existing classes in order to obtain a fuzzy representation of the expression levels belonging to those genes that best explain each class in the form of a fuzzy-prototype. The final goal is the application of the proposed method as a retrieval step for the gene-CBR system.

CBR and Cancer

Recently, knowledge discovery techniques, such as decision trees, neural networks, support vector machines and case-based reasoning, have been applied increasingly to improve medical decision-making and to the diagnosis, prognosis and prescription of cardiovascular diseases, cancer, diabetes, chronic lower respiratory diseases and many other medical problems.

CBR has been used to solve various problems, such as in financial forecasting, credit analysis and medical diagnosis. The advantages of CBR are its ease of learning and explanatory capability. For example, although one of the prevalent knowledge extraction methods, neural networks, has received a great deal of attention because of its high predictability, its lack of an explanatory capability constrains its use in many areas. In contrast, with CBR it is relatively easy to understand how the results are produced and which cases are used for them, so it can be an appropriate method for many real-world areas that need explanation. Another advantage of CBR is that it can still be effective if the knowledge base or domain theory is incomplete. Certain techniques of automated learning, such as explanation based learning, work well only if a strong domain theory exists, whereas CBR can use many examples to overcome the gaps in a weak domain theory, while still taking advantage of the domain theory.

Case-based reasoning (CBR) has been applied relatively less frequently in medicine than other methods. Recently, however, CBR has been used for various medical problems, such as an antibiotics therapy advice system, diet recommendation, care of Alzheimer's disease patients, diagnosing Alzheimer's disease, heart failure and pulmonary disease, and for managing diabetes mellitus.

- In the research work [1] presents two applications for the breast cancer treatment decision helping. The first one is called Casimir/RBR and can be likened to a rule-based reasoning system. In some situations, the application of the rules of this system does not provide a satisfying treatment. Casimir/CBR uses principles of case-based reasoning in order to suggest solutions by adapting the rules of Casimir/RBR. the rules are considered as cases: they are adapted rather than used literally.

  Most important in the framework is about the way the knowledge units -rules or cases- are used:

  - ∗ The rules are (typically) used literally; the conclusion of the rule is simply instantiated.
  - ∗ The cases are adapted and not simply copied or instantiated.

  The difference between a case and a rule is not based on their contents but on the different ways they are used.

  Thus, the same knowledge units can be used to do reasoning from rules or from cases. In particular, a rule $R = (Prem \rightarrow Cclo)$ can be seen as a case whose problem is the premise *Prem* of *R* and whose solution is the conclusion *Cclo* of *R*.

  This is the principle that has been adopted for the two applications Casimir/RBR (which can be seen as a RBR application) (and Casimir/CBR) which can be seen as a CBR application.

  The use of RBR (in fact, of hierarchical classification) in order to model the reasoning of Casimir/RBR is deeply inspired from the system Resyn whose domain is the synthesis planning in organic chemistry.

  The system Resyn/CBR can be seen as an extension of Resyn, the difference being essentially due to the fact that RBR is followed by CBR in Resyn/CBR.

  The part of the application Casimir/CBR which should be the most similar to Resyn/CBR. The technique used for retrieval in Resyn/CBR is smooth classification. For the part of Casimir/CBR dealing with situations, the same technique is scheduled.

  Casimir/RBR and Casimir/CBR, that are dedicated to decision helping in the framework of breast cancer treatment. The second one uses case-based reasoning in an unusual way: the rules of Casimir/RBR are used as cases for Casimir/CBR.

- In research work [6] explores how data mining and knowledge discovery can be applied to medical informatics using human gene information. They applied case-based reasoning to a cancer detection problem using human gene information and SNP analysis because case-based reasoning has been applied in medicine relatively less often than other data mining techniques. Proposed a modified case-based reasoning method that is appropriate for associated categorical variables to use in detecting gastric cancer.
  The work shows how CBR can resolve a cancer detection problem using human gene information and SNP analysis. Propose a modified CBR method, which is appropriate for use with associated categorical variables, to detect cancer using SNP data based on a haplotype analysis. The associated information between genes needs to be considered in a diagnostic model. This modified CBR method can convert the original data into haplotype data because combinations of SNP data can enhance the accuracy.
  They proposed a modified CBR method to fit a specific medical domain, and developed a CBR method that uses associated informative data for gastric cancer detection. previous data mining techniques have used simple gene information. Therefore, the modified CBR method needs to consider the associated information between genes for the medical diagnostic model.

4. The Microarrays and CBR

Molecular biology domain is a natural application for CBR systems, since CBR systems can perform remarkably well on complex and poorly formalized domains. However, due to the large number of attributes in each case, CBR classifiers, similarly as other learning systems, suffer from the "curse of dimensionality". Maintaining CBR systems can improve the prediction accuracy of CBR classifiers by clustering similar cases, and removing "non-informative" features in each group. Focus on integrating clustering and feature selection techniques with CBR systems to enhance the accuracy of CBR classifiers.

In the research work [7], is proposed a maintenance technique that integrates an ensemble of CBR classifiers with spectral clustering and logistic regression to improve the classification accuracy of CBR classifiers on (ultra) high-dimensional biological data sets. They demonstrate the improvement achieved by applying the method to a computational framework of a CBR system called TA3. The maintenance method improves the classification accuracy of TA3 by approximately 20% from 65% to 79% for the leukemia and from 60% to 70% for the lung cancer data set. Two examples used in this paper involve the profiling of gene and protein expression using microarrays and mass spectrometry.

The main challenge is to interpret the molecular biology data to find similar samples to eventually use them in case-based medicine, and to identify those genes whose expression patterns have meaningful relationships to their classification labels whose expression patterns help better understand cancer initiation and progression. Clustering and feature selection techniques have been successfully applied to CBR maintenance; however, they show how those techniques can further improve the prediction accuracy of a CBR classifier when combined with mixture of experts to analyze microarray data sets. The CBR maintenance approach has three main components: ***ensemble of CBR systems,***

clustering, and feature selection. Use an ensemble of CBR systems, called mixture of experts (MOE) to predict the classification label of a given (input) case. A gating network calculates the weighted average of votes provided by each expert. The performance of each CBR expert is further improved by using clustering and feature selection techniques. Apply spectral clustering to cluster the data set into *k* groups, and the logistic regression model is used to select a subset of features in each cluster. Each cluster is considered as a case-base for the *k* CBR experts, and the gating network learns how to combine the responses provided by each expert.

Smyth on his research say that differents CBR maintenance methods can be categorized into two Groups Competence-directed and Efficiency-directed.

Mixture of experts approach is based on the idea that each expert classifies samples separately, and individual responses are combined by the gating network to provide a final classification label.

The goal of the maintenance method (MOE4CBR) is to improve the prediction accuracy of CBR classifiers, when combined with the mixture of experts approach to analyze microarray and mass spectrometry data sets, and at the same time reduce the size of the case-base knowledge container. According to Smyth's categorization, the maintenance method Mixture of Experts for CBR systems (MOE4CBR) is both competence directed, since the range of the problems the system can solve increases and efficiency directed, since the size of case base decreases. The integration approach mixture of experts for case-based reasoning (MOE4CBR) is comprised of three main components:

- An ensemble of CBR systems to predict the classification label for a given input case;
- Clustering to organize data set (case-base) into *k* homogeneous groups; and
- Feature selection to characterize each cluster by a subset of informative features.

The performance of each expert in MOE4CBR is improved by using clustering and feature selection techniques. Based on the initial analysis, they selected spectral clustering for clustering the case-base, and the logistic regression model as a filter feature selection for the TA3 classifier. Given a labeled training data set, the system predicts labels for the unseen data (test set).

The earlier evaluation suggests that spectral clustering outperforms k-means clustering and SOMs. The comparison was based on two criteria:

- Dunn's index, which does not require class labels and identifies how "compact and well separated" clusters are.
- Precision and recall that compare the resulting clusters with pre-specified class labels. Precision shows how many data points are classified (clustered) correctly, and recall shows how many data points the model accounts.

The MOE4CBR maintenance method has two main steps: first, the case-base of each expert is formed by clustering the data set into *k* groups, then each case-base is maintained "locally" using feature selection techniques. Each of the *k* obtained sets will be considered as a case-base for *k* CBR experts. Each expert applies the TA3 model to decide on the class label, and the gating network uses TA3 to assign weights to each classifier.

Fisher criterion and standard t-test are two statistical methods that have been successfully applied to feature selection problem in (ultra) high-dimensional data sets. In order to select a suitable feature selection approach for CBM, they have evaluated performance of Fisher criterion, t-test, and the logistic regression model when used in a CBR classifier. Namely, they have applied the three feature selection techniques to the TA3 classifier, and measured the improvement in accuracy and classification error. Accuracy measures the number of correctly classified data points, and classification error counts the number of misclassified data points. Based on the evaluation, logistic regression applied to feature selection outperforms Fisher and standard t-test techniques.

They contribution can be summarized as follows: First, they have demonstrated that integrating the CBR paradigm with data mining and feature selection improves its prediction accuracy in high-dimensional biomedical domains. Second, they shown that spectral clustering outperforms SOMs and k-means clustering, the two clustering methods widely used in analyzing microarrays and mass spectrometry data sets. Third, reported a novel use of logistic regression as a filter feature selection method for a CBR classifier.

5. Conclusions and Future Work

The Bioinformatics field has been briefly reviewed, and the Microarrays were analyzed more deeply. The microarrays it's a new area in computer science. In fact the literature about the topic shown that there are many research works moving to it. New designs exploiting the characteristics of microarrays, especially innovation and developing new algorithms are seen in this field. For us, the immersion in this information leave us the idea of apply experiments using the matrix of arrays and the structure of cases as backbone to consolidate the research of applying polices based on microarrays, and define a novel policy in the case-based maintenance, based in the mixture of the structure of microarrays and CBR.

References

[1] Jean Lieber and Benoit Bresson, Case-Based Reasoning for Breast Cancer Treatment Decision Helping, LNAI EWCBR,pp. 173-185, 2000.

[2] Edwin Costello and David C. Wilson, A Case-Based Approach to Gene Finding, ICCBR, Computer Science Department, University College Dublin, Dublin, Ireland, 2003.

[3] Jude W. Shavlik, Finding Genes by Case-Based Reasoning in the Presence of Noisy Case Boundaries, Appears in the Proceedings of the DARPA Workshop on Case-Based Reasoning, Morgan-Kaufmann, 1991.

[4] Florentino Fdez-Riverola, Fernando Díaz, M. Lourdes Borrajo, J. Carlos Yáñez, and Juan M. Corchado, Improving Gene Selection in Microarray Data Analysis Using Fuzzy Patterns Inside a CBR System, ICCBR, 3620, pp. 191 - 205, 2005.

[5] FERNANDO DÍAZ, FLORENTINO FDEZ-RIVEROLA, JUAN M. CORCHADO, GENE-CBR: A CASE-BASED REASONIG TOOL FOR CANCER DIAGNOSIS USING MICROARRAY DATA SETS, Computational Intelligence, volume = 22, number 3/4, pp. 254 - 268, 2006.

[6] Se-Chul Chun, Jin Kim, Ki-Baik Hahm, Yoon-Joo, Park and Se-Hak Chun, Data mining technique for medical informatics: detecting gastric cancer using case-based reasoning and single nucleotide polymorphisms, Expert Systems, Volume 25, number 2, pp. 163-172, 2008.

[7] Niloofar Arshadi and Igor Jurisica, Maintaining Case-Based Reasoning Systems: A Machine Learning Approach, ECCBR 2004, pp. 17 - 31, 2004.

[8] Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J.D. Molecular biology of the cell. New York: Garland Publishing,(1989).

[9] Niloofar Arshadi and Igor Jurisica, Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Volume. 17, Number 8, pp. 1127-1137, 2005.

[10] Arthur M. Lesk, Book Introduction to Bioinformatics, Second Edition, The Pennsylvania State University, 2005.

[11] Ashani T. Weeraratna and Dennis D. Taub, Chapter Book Microarray Data Analysis An Overview of Design, Methodology, and Analysis, 2007.

[12] Helen C. Causton, John Quackenbush, Alvis Brazma, Book Microarray Gene Expression Data Analysis A Beginner's Guide, 2004.

[13] Werner Dubitzky, Martin Granzow, C. Stephen Downes, Daniel Berrar, INTRODUCTION TO MICROARRAY DATA ANALYSIS, Chapter of Book "A PRACTICAL APPROACH TO MICROARRAY DATA ANALYSIS" Daniel P. Berrar, Werner Dubitzky, Martin Granzow, 2003.

[14] Soumya Raychaudhuri, Patrick D. Sutphin, Jeffrey T Chang and Russ B. Altman, Basic microarrays analysis: grouping and feature reduction. TRENDS in Biotechnology Vol. 19 No 15,pp. 189 - 193, 2001.

[15] Daniel P. Berrar, Werner Dubitzky, Martin Granzow, Book A PRACTICAL APPROACH TO MICROARRAY DATA ANALYSIS, 2003.

[16] Mark Schena, Book Microarray analysis. 2003.