# Cancer Identification and Gene Classification using DNA Microarray Gene Expression Patterns

**Chhanda Ray**
**RCC Institute of Information Technology**
**Kolkata, INDIA**

**Abstract**

DNA microarray gene expression patterns of several model organisms provide a fascinating opportunity to explore important abnormal biological phenomena. The development of cancer is a multi-step process in which several genes and other environmental and hormonal factors play an important role. In this paper, a new algorithm is proposed to analysis DNA microarray gene expression patterns efficiently for huge amount of DNA microarray data. For better visibility and understanding, experimental results of DNA microarray gene pattern analysis are represented graphically. The shape of each graph corresponding to a DNA microarray gene expression pattern is determined by using an eight-directional chain code sequence, which is invariant to translation, scaling, and rotation. The cancer development is identified based on the variations of DNA microarray gene expression patterns of the same organism by simultaneously monitoring the expression of thousand of genes. At the end, classification of cancer genes is also focused based on the distribution probability of codes of the eight-directional chain code sequences representing DNA microarray gene expression patterns and the experimental result is provided.

***KEYWORDS:*** *Biological, Cancer development, Chain code, DNA microarray gene expression, Pattern analysis.*

## 1. Introduction

Almost all human genetic diseases such as cancer and developmental abnormalities are characterized by the presence of genetic variations. The discovery and analysis of gene expression patterns of several model organisms represents a fascinating opportunity to explore important normal and abnormal biological phenomena. DNA plays a fundamental role at different bio-chemical processes of living organisms.

The past few years have witnessed tremendous interests in investigating genetic signatures for cancer progression and cancer research using high-throughput gene expression experiments [1, 6, 8, 9, 12, 13]. A non-hierarchical method for clustering large amount of DNA sequences focusing on the oligonucleotide frequency is introduced in [1]. In [8], a method for automatic classification of Pubmed articles related to human genome epidemiological research at Center for Disease Control and Prevention using support vector machine has been discussed. CpG-Discover, a machine learning approach for CpG islands identification from human DNA sequence, which plays a fundamental role in genome analysis as genomic markers and tumor markers is focused in [6]. Paper[12] introduces a pipeline method for cis-Regulatory element prediction in mammalian genomes. An alternative approach, namely, a multicategory logit method, to identify novel genes that show significant correlations across multiple stages of prostate cancer progression has been proposed in [9]. In [13], another novel procedure to identify human cancer genes acting in a recessive manner has been represented.

In this context, DNA microarray is a landscape-changing tool in molecular biology, which provides researchers with an approach to monitor genome-wide expression systematically. It provides unprecedented opportunity to explore the biological processes underlying human diseases by providing a comprehensive survey of the cell's transcriptional landscape. In recent years, DNA microarray technology has been widely used for cancer research [2, 3, 4, 7, 14]. In [2], microarray-based comparative genomic hybridization techniques are used to map and measure DNA copy number variations with high-resolution for the analysis of human genome associated with behavior abnormalities. A self-adaptive and incremental neural approach to identify normal and diffuse large B-cell lymphoma (DLBCL) patients using cDNA microarrays data has been introduced in [3]. Paper[4] proposes meanshift clustering technique for analyzing DNA microarray data. In [7], unveiling fuzzy associations between breast cancer prognostic factors and gene expression data is focused. A newly designed data mining model to store microarray experimental data in a systematical organization, and to provide an efficient way for researchers to mine the database and populate it in a reasonable manner for research progresses is represented in [14]. However, DNA microarray analysis is very complex due to large number of genes and the noise that affects the whole process. Further, the main hurdle and challenge is how to make use of the tremendous

amount and ever-growing microarray experimental data to precisely explain the cancer mechanism.

The contribution of the paper is as follows. Initially, a new algorithm for efficient pattern analysis of DNA microarray gene expression data is introduced. In order to provide better visual representation, experimental result of DNA microarray gene pattern analysis is illustrated graphically. The shape of the graph corresponds to a DNA microarray gene expression pattern is determined in terms of eight-directional chain code sequence. The cancer development is identified based on the variations of DNA microarray gene expression patterns of the same organism by simultaneously monitoring the expression of thousand of genes. Finally, the classification of cancer genes is proposed based on the distribution probability of codes of the eight-directional chain code sequences representing DNA microarray gene expression patterns. A preliminary version of the pattern matching algorithm described in this work has applied in statistical estimation in [10] while the shape determination of an object using eight-directional chain code sequence is introduced in [5].

The paper is organized in the following way. Section 2 introduces an efficient algorithm for analysis of DNA microarray gene expression patterns. In Section 3, the experimental results of DNA microarray pattern analysis have been illustrated graphically. The shape of the graph representing DNA microarray gene expression pattern is expressed by using eight-directional chain code sequence in Section 4. In Section 5, the cancer development of an organism is identified based on the variations of DNA microarray gene expression patterns of the same organism. An approach for classification of cancer genes based on the distribution probability of eight-directional unique chain code sequences of graphs is proposed in Section 6 and Section 7 provides a conclusion to the work.

## 2. An Algorithm for DNA Microarray Gene Expression Pattern Analysis

DNA (deoxyribonucleic acid) is a nucleic acid that contains the genetic instructions used in the development, reproducing and functioning of all living organisms. DNA strand consists of a large sequence of nucleotides, or bases. One nucleotide consists on one nitrogen base, one sugar molecule (deoxyribose) and one phosphate. Four nitrogen bases are denoted by one of the letters A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). A linear chain of DNA is paired to a complementary strand. The complementary property stems from the ability of the nucleotides to establish specific pairs A-T and G-C. The pair of complementary strands then forms the double helix. Each strand therefore carries the entire information and the biochemical machinery guarantees that the information can be copied over and over again even when the original molecule has long since vanished.

A gene is primarily made up of sequence of triplets of the nucleotides (exons). DNA microarray results represent the gene expression pattern and based on the genetic code the linear string of DNA is translated into a linear string or amino acids, that is, a protein via mRNA (messenger RNA). For example, the DNA sequence GAACTACACACGTGTAAC codes for the amino acid sequence ELHTCN as follows.

| GAA | CTA | CAC | ACG | TGT | AAC |
|---|---|---|---|---|---|
| E | L | H | T | C | N |

In this section, an efficient pattern matching algorithm is presented in order to detect and count the number of occurrences of all possible patterns composed of the four characters of nitrogen bases A, C, G, and T taking three at a time including the repetitions in a given DNA microarray gene expression data. The total number of possible patterns is 64. Hence, for the sake of simplicity, three characters of nitrogen bases is considered at a time instead of taking four characters of A, C, G, and T. The detailed description of the algorithm is as follows.

INPUT: Given DNA microarray gene expression data, say x.
OUTPUT: Frequency of occurrence of each of possible 64 patterns of length 3 in the given DNA microarray gene expression data x.

```
DNA microarray gene expression data = x, Pattern = y
Begin
m:= length of x,   c:= length of y
i:=0
for s:= 0 to c-1
        for i:= 0 to m-1
        ch:= y[s]
        switch(ch)
                case ch is 'A' : a[i]:= 0
                case ch is 'C' : a[i]:= 1
                case ch is 'G' : a[i]:= 2
                case ch is 'T' : a[i]:= 3
                default: i--
        end switch
        i++
        s++
        end for
        n:=0, j:= 0
        for k:= m-1 to 0
        n:=  n + (4^j) x a[k]
        k--
        j++
        end for
        count[n]++
```

```
        for j:= 0 to m-2
        a[j]:= a[j + i]
        end for
        i--
        pat[]:= {'A', 'C', 'G', 'T'}
        for i:= 0 to 4^m – 1
        y:= q % 4
        arr[j] = pat[y]
        if (q != 0)
                q:= q/4
        end if
        j--
        print: arr[j]
        end for
        print: count[i]
        print: /n
end for
END
```

The pattern matching algorithm for DNA microarray gene data introduced in this work is different from other existing algorithms in a major way. Like pattern matching algorithm, this method does not require any pattern generation scheme and hence the algorithm is implemented as pattern recognition opposed to matching. In this approach, the text is scanned into chunks of 'm' characters at a time and it is encoded to a unique number, and the corresponding counter of the specific pattern is subsequently incremented. Thus, the number generated by each pattern acts as the key and the encoding is done according to a number system of base 4 since there are only 4 characters in DNA microarray gene expression database. Consequently, the complexity of the algorithm is significantly low in comparison with other existing pattern matching algorithms and it has a big impact for handling huge amount of DNA microarray gene expression data.

## 3. Graphical Representation of DNA Microarray Gene Patterns

In this section, the experimental result of DNA microarray gene pattern analysis is presented graphically for better visibility and understanding. For graphical representation of DNA microarray gene expression data, each quadrant of a circle is partitioned into 16 fragments and a radius is drawn for each fragment since 64 patterns are possible by taking three at a time from four characters of nitrogen bases A, C, G, and T. The radius of the circle is fitted to a high scale value considering the maximum frequency value of patterns (that can occur). The frequency value of each pattern of DNA microarray gene expression data is plotted on the corresponding radius representing the particular pattern. The graph has been drawn by connecting the corresponding points.

For a given DNA microarray gene expression data, the experimental result of pattern analysis is graphically represented in the following figure.

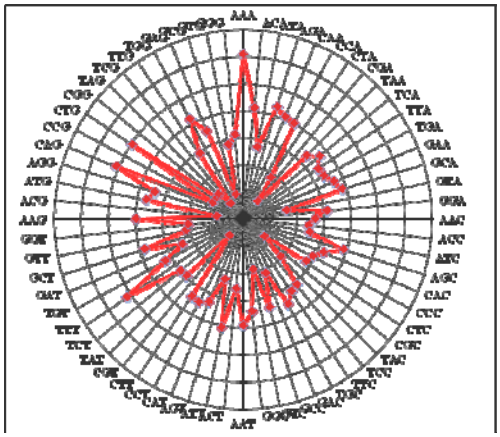


**Figure 1: Graphical Representation of DNA Microarray Gene Patterns**

## 4. Determination of Chain Code Sequence for the Shape of the Graph

This section introduces the shape determination of the graph corresponds to a DNA microarray gene expression patterns using chain code sequence. Hence, one conventional eight-directional chain code (Fig. 2) is used to represent the shape of the graph which is invariant to translation, scaling and rotation. The edges of the shape (graph) are partitioned into line segments by the frequency value of each pattern within a DNA microarray gene expression data. The chain code of a line segment depends on the previous line segment and it is determined by counting the number of direction changes in a counterclockwise direction that separate two adjacent elements of the code as shown in Fig. 2.

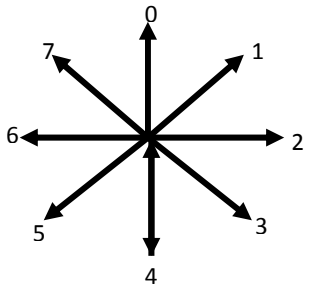



**Figure 2 : Eight-directional Chain Code**

The chain code sequence of a shape also depends on the starting point which can be normalized by treating it as a circular sequence of direction numbers and redefining the starting point. For example, the code for the line segment AB in Fig. 3 is 2, because JA is the previous line segment and twice the direction has been changed from JA to AB.

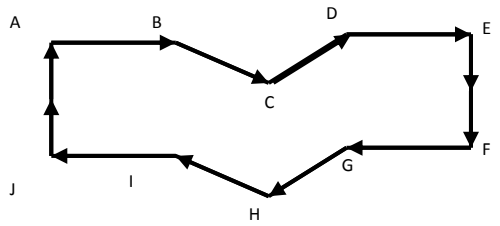



**Figure 3: An Example of Object's Shape**

Using the same coding convention, the chain code sequence for the shape of the graph represented in Fig. 1 is listed below assuming AAA as the starting point.

2 0 6 2 0 2 0 4 0 4 6 0 0 2 6 2 0 6 0 2 0 0 2 4 0 2 0 2 4 4 6 0 2
6 2 6 0 2 2 4 2 6 4 0 4 4 0 6 2 6 2 6 4 4 4 6 2 4 0 4 0 4 2 0

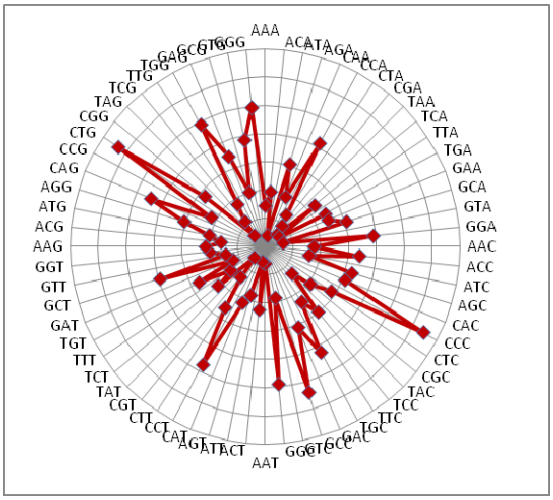
For the sake of simplicity, if all 0's are remove from the chain code sequence then the shape of the graph remains same, because object's shape representation is invariant to scaling, rotation and scaling with rotation.

## 5. Identification of Cancer Development

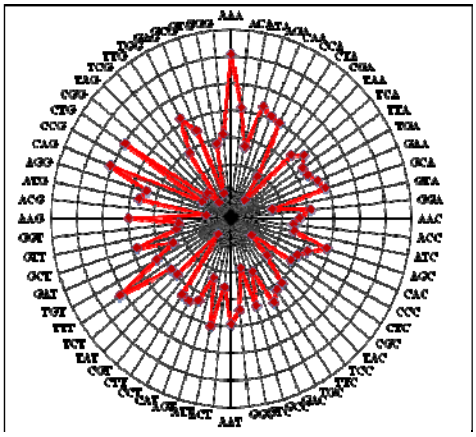
Cancer arises when cells escape normal cell cycle and differentiations, being able to proliferate rapidly, to invade surrounding tissues, and to metastasize to distant sites. The development of cancer is proposed to be a multi-step process in which several genes and other environmental and hormonal factors play a role. One gene or one factor often contributes partly to cancer and its exact role can be demonstrated in specific contexts.

Transcription is the process through which ribonucleic acid (RNA) is formed from DNA (gene) and through translation amino acids are formed from RNA. There are sequences of nucleotides within the DNA that are spliced out progressively in the process of transcription and translation. To detect the cancer development, the biomedical steps that led to the specific cellular malfunction displayed in the cells must be identified. Key features of such process involve the transcription and translation regulatory system, cellular signaling pathways and metabolic pathways that control cell cycle, differentiation, apoptosis and motility.

In [11], it has been shown that specific patterns of gene expression occur during different biological states such as embryogenesis, cell development etc. Generally speaking the expression of a gene provides a measure of "how active" a specific gene under certain biochemical conditions. In this work, cancer development is detected based on the variations of DNA microarray gene expression patterns of the same organism. The graphical representation of DNA microarray gene expression patterns facilitates researchers to simultaneously monitor the expression of thousand of genes of a particular organism, which provides a molecular biomarker to differentiate normal cells from malignant ones. An example of normal DNA microarray gene expression patterns and cancer DNA microarray gene expression patterns is illustrated in the following figure.



**(a)**



**(b)**

**Figure 4: Representation of DNA Microarray Gene Patterns for Normal and Cancer Cell**

The chain code sequences for the shape of the above graphs illustrated at Fig. 4, considering AAA as the starting point are as follows.

6 4 4 2 6 4 0 0 4 4 4 2 0 2 4 4 6 2 6 2 6 4 0 0 4 4 4 4 6 2 6 2 6 2 6 0 0 4 0 0 4 4 4 4 4 4 2 0 2 6 0 0 4 4 4 0 4 0 0 4 0 6 0 2 and

2 0 6 2 0 2 0 4 0 4 6 0 0 2 6 2 0 6 0 2 0 0 2 4 0 2 0 2 4 4 6 0 2 6 2 6 0 2 2 4 2 6 4 0 4 4 0 6 2 6 2 6 4 4 4 6 2 4 0 4 0 4 2 0

The distribution probability for the above chain code sequences is listed in the following table where the values are normalized between 0 and 1.

| **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| 0.2656 (a) | 0.0 | 0.1718 | 0.0 | 0.3906 | 0.0 | 0.1718 | 0.0 |
| 0.2968 (b) | 0.0 | 0.2812 | 0.0 | 0.2343 | 0.0 | 0.1875 | 0.0 |

**Table 1: Distribution Probability of Chain Code Sequences**

In the above example, from the variations of chain code sequences of graphs representing DNA microarray gene expression patterns of the same organism, malignancy can be easily differentiate from normal ones.

## 6. Classification of Cancer Genes

The classification of cancer genes from normal ones plays a crucial role in cancer treatment. In the last few decades, many researchers have proposed various methods to deal with the classification and clustering of cancer genes. In this work, the classification of cancer genes is focused based on the similarity between the distribution probabilities of chain code sequences representing DNA microarray gene expression patterns where the similarity is measured by comparing the vector distance between chain code sequences. In this work, the vector distance between chain code sequences is calculated by using Jeffrey Divergence techniques. Jeffrey Divergence is a modification of Kullback-Leibler Divergence that compares two empirical distributions and the formula for Jeffery Divergence is as follows:

$D(I, J) = \sum f_i(I) \log(f_i(I)/f_i + f_i(J) \log f_i(J)/f_i,$

where $f_i = [f_i(I) + f_i(J)]/2$ and I and J are two different DNA microarray gene expression patterns.

One chain code sequence for normal cell and two chain code sequences for cancer genes corresponds to DNA microarray gene expression patterns is listed in the following.

6 4 4 2 6 4 0 0 4 4 4 2 0 2 4 4 6 2 6 2 6 4 0 0 4 4 4 4 6 2 6 2 6 2 6 0 0 4 0 0 4 4 4 4 4 4 2 0 2 6 0 0 4 4 4 0 4 0 0 4 0 6 0 2 and

2 0 6 2 0 2 0 4 0 4 6 0 0 2 6 2 0 6 0 2 0 0 2 4 0 2 0 2 4 4 6 0 2 6 2 6 0 2 2 4 2 6 4 0 4 4 0 6 2 6 2 6 4 4 4 6 2 4 0 4 0 4 2 0

2 0 6 2 0 2 0 4 0 4 4 0 4 0 6 2 0 0 6 4 4 0 4 6 6 2 0 2 4 4 6 0 2 6 0 2 6 4 0 4 0 0 4 0 6 2 6 0 2 6 0 0 4 4 4 6 2 4 2 2 0 6 2 6

The distribution probability for the above chain code sequences is as follows.

| **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| 0.2656 (N) | 0.0 | 0.1718 | 0.0 | 0.3906 | 0.0 | 0.1718 | 0.0 |
| 0.2968 (C) | 0.0 | 0.2812 | 0.0 | 0.2343 | 0.0 | 0.1875 | 0.0 |
| 0.3125 (C) | 0.0 | 0.2187 | 0.0 | 0.25 | 0.0 | 0.21875 | 0.0 |

**Table 1: Distribution Probability of Chain Code Sequences**

In the above example, it has been shown from the experimental results that distribution probabilities of codes corresponds to DNA microarray gene expression pattern for normal and cancer cells are obviously different. Further, the vector distance between cancer cells are closest to each other whereas the vector distance between cancer cell and normal cell are far away from one another.

## 7. Conclusion

In this paper, a new algorithm for efficient analysis of DNA microarray gene expression patterns is introduced that has a significantly low time complexity, and the experimental result is illustrated graphically. The shape of each graph corresponds to a DNA microarray gene expression pattern is determined by eight-directional chain code sequence. Hence, the cancer development of a particular organism is identified based on the variations of DNA microarray gene expression patterns of the same organism. An approach for classification of cancer genes based on the distribution probability of chain code sequences corresponds to DNA microarray gene expression patterns is also described at the end of the work.

However, the other environmental and hormonal factors for causes of cancer development are not considered in this work.

In future scope, a complex function can be generated based on DNA microarray gene expression pattern and cancer cell can be identified from normal ones depending on the value of the function. The growth of tumors can be measured based on DNA microarray gene expression patterns in the same way in future work. Further, other factors in human gene expression data can be considered to achieve better result in cancer detection and development. Moreover, an efficient algorithm can be developed in order to classify different types of cancer genes from huge amount of DNA microarray gene expression data.

## References

[1] K. Amano, H. Nakamura, "Self-Organizing Clustering: "A Novel Non-Hierarchical Method for Clustering Large Amount of DNA Sequences", Journal of Genome Informatics", Vol. 14, pp 575-576, 2003.

[2] A. K. Alqallaf, A. H. Tewfik, "Signal Processing techniques and statistics for the analysis of human genome associated with behavior abnormalities", proceedings of IEEE International conference on SSP, pp 36-38, 2007.

[3] F. Azuaje, "Gene expression Patterns and Cancer Classification: A Self-Adaptive and Incremental Neural Approach", Journal of IEEE Engineering in Medicine and Biology, pp 308 -313, 2000.

[4] D. Barash, D. Comaniciu, "Meanshift Clustering for DNA Microarray Analysis", Proceedings of the IEEE Conference on Computational Systems, pp 578-579, 2004.

[5] R. Jana, C. Ray, "Image Registration using Object Shape's Chain Code", Proceedings of 2nd International Congress Image and Signal Processing, pp 1-5, ISBN: 978-1-4244-4129-7, 2009.

[6] M. Lan, Y. Xu, L. Li, F. Wang, Y. Zuo, Y. Chen, C.L. Tan, J. Su, "CpG-Discover: A Machine Learning Approach for CpG Islands Identification from Human DNA Sequence", Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, pp 1702-1707, June 14-19 2009.

[7] F.J. Lopez, M. Cuadros, A. Blanco, A. Concha, "Unveiling Fuzzy Associations between Breast Cancer Prognostic Factors and Gene Expression Data", Proceedings of 20th International Workshop on Database and Expert Systems Application, pp 338-342, 2009.

[8] N. Polavarapu, S.B. Navathe, R. Ramnarayanan, A.U. Haque, S. Sahay, Y. Liu, "Investigation into Biomedical Literature Classification using Support Vector Machines", Proceedings of IEEE Conference on Computational Systems Bioinformatics, pp 366-374, 2005.

[9] Y.S. Pyon, J. Li, "Identifying Gene Signatures from Cancer Progression Data Using Ordinal Analysis", Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, pp 136-141, 2009.

[10] C. Ray, S.Tripathi, A. Chatterjee, A. Das, "An efficient Bi-directional String Matching Algorithm for Statistical Estimation", Proceedings of International Symposium on Data, Information, and Knowledge Spectrum, INDIA, pp 73-79, December 2007.

[11] P.J. Russel, "Fundamentals of genetics", Second Edition, San Francisco, Addison Wesly Longman Inc., 2000.

[12] A. Siddiqui, G. Robertson, M. Bilenky, T. Astakhova, O.L. Griffith, M. Hassel, K. Lin, S. Montgomery, M. Oveisi, E. Pleasance, N. Robertson, M.C. Sleumer, K. Teague, R. Varhol, M. Zhang, S. Jones, "Cis-Regulatory Element Prediction in Mammalian Genomes", Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference Workshops, pp 203-206, 2005.

[13] S. Volinia, N. Mascellani, J. Marchesini, A. Veronese, E. Ormondroyd, H. Alder, J. Palatini, M. Negrini, C. M. Croce, "Genome wide Identification of Recessive Cancer Genes by Combinatorial Mutation Analysis", www.plosone.org, Vol. 3, Issue 10, pp 1-13, 2008.

[14] R. Xue, J. Li, D.J. Streveler, "Microarray Gene Expression Profile Data Mining Model for Clinical Cancer Research", Proceedings of the 37th Hawaii International Conference on System Sciences, pp 1-10, 2004.