ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

Classification of Mutated Cancer Genome Using Machine Learning Approaches

Anit V Mathew[1], Jisha Mariyam John[2], Tinu Thomas[3]

PG Student, Computer Science and Engineering, Mangalam College of Engineering, Kottayam, India[1,2]

Assistant Professor, Computer Science and Engineering, Mangalam College of Engineering, Kottayam, India[3]

Abstract: Cancer may be a genetic abnormality derived from genetic changes that end in a loss of control over necessary cellular functions. Identification of mutated cancer gene plays a crucial role in individualizing the treatment of a cancer patient in keeping with his specific tumorigenic profile. Wavelets analysis techniques are capable of extracting each spectral and local information and perform multiscale analysis on DNA/protein sequences. The amino acid index features represent the physicochemical properties of the protein sequences. The wavelet features combined with AAIndex features offer feature vector for classification of mutated driver gene. Machine learning based approaches are utilized in cancer genome analysis to mine patterns from the prevailing data and built mathematical models to learn patterns and make predictions in unanalyzed data. The proposed system deals with scrutiny the performance of varied combinations of Support Vector Machine and Back Propagation Neural Network to identify the mutated driver gene.

Keywords: driver gene; wavelet analysis; AAIndex features; Support Vector Machine ; Back Propagation Neural Network.

I. INTRODUCTION

Cancer represents one of the greatest causes that threaten the human life in the world today. The somatic mutations in the DNA sequence of the cancer cell genome causes all these cancers. Each somatic mutation in a cancer cell genome could be categorized into either of the two classes, driver mutations or passenger mutations. A driver mutation causally involves in the oncogenesis [6]. It has conferred growth advantage on the cancer cell. A driver mutation need not be required for maintenance of the final cancer but it must have been selected at some point along the lineage of the cancer development. A passenger mutation has not been selected and has not conferred growth advantage on cancer cells. Therefore passenger mutations have not contributed to cancer development. Thus identification of driver mutations is the central goal of cancer research. Therefore the key challenge of cancer genome analysis is to distinguish the driver mutations from the passenger mutations.

Biological sequences, consisting of DNA and protein encoded data, could be viewed as one-dimensional signals from the signal processing point of view. As a result, signal processing approaches have been applied to perform analysis on these types of data. The characteristics of most real-world signals are that they vary in both time and frequency domains. Wavelet analysis is capable of performing multiscale analysis on a sequence without any prior knowledge and is also capable of simultaneous localization of time and frequency domains. Therefore it has found many applications in the area of cancer research. The proposed system also uses the wavelet analysis for extracting features of the protein sequences. The AAIndex features helps in understanding the 544 physicochemical properties of the protein sequences such as hydrophobicity, residue volume and so on. The wavelet features are combined with the AAIndex features to provide a better feature vector for driver mutation classification.

The proposed system initially collects the driver and the passenger genes. The protein sequences of these genes are then converted to the numerical representation. The 100 dimensional wavelet features and the 544 dimensional AAIndex features of the protein sequences are then extracted. Both the extracted features are then combined together to form the feature vector. This 644 dimensional feature vector is then given as the input to a classifier to identify the mutated genes. The mutated genes are then given as input to the next classifier to classify the mutated driver and passenger genes.

For the past few years several works related this area are going on. The work in [1] deals with identification of mutated driver gene provided mutated genes are given as test data. This work provided a survey of how wavelet analysis has been applied to cancer bioinformatics. Several approaches regarding representing the biological sequence data numerically and methods of using wavelet analysis on the numerical sequences have been discussed here. The current progress of using wavelets in biological sequences analysis in cancer genome have also been discussed.

The work [2] discussed about an approach for analysis of DNA sequences. According to this work a wavelet based time series approach is used for extracting statistical information from DNA sequences and the variance information in this extracted information is used to construct a feature vector.

The work [9] discussed about the AAindex database which is a flat file database of numerical indices representing various physicochemical and biochemical properties of

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

amino acids and pairs of amino acids. In [10] the researchers discussed about the numerical representation of DNA sequences based on genetic code context within DNA sequences and explore the feasibility of applying this method to identify protein coding regions in genomes.

II. PROPOSED SYSTEM

Identification of cancer genes that carry driver mutations is the central goal of cancer genome analysis. The architecture of the proposed system is shown in Fig 1.The main modules of the proposed system are:

A. Data collection
B. Numerical representation of biological sequence
C. Feature extraction
   1. Wavelet feature extraction
   2. AAIndex feature extraction
D. Identification of mutated genes
E. Identification of mutated driver genes

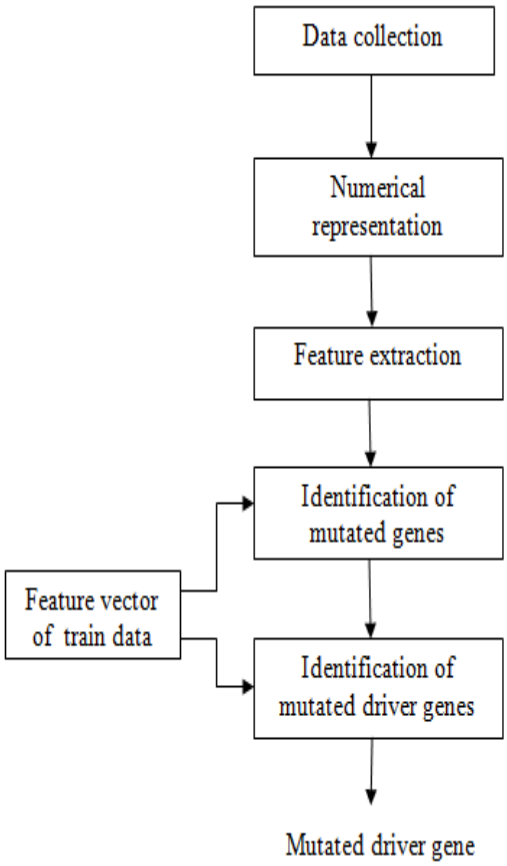

Fig. 1 Architecture of proposed system

In the proposed system first the driver and the passenger genes are collected from the GenBank and then the mutation samples are extracted from these genes using a fixed sized window. Next, the protein sequences are converted into numerical representation inorder to give as input to the wavelet transform functions. Next, the wavelet transforms are applied to the mutation samples to obtain the wavelet coefficients at different scales. These wavelet coefficients are then used as features for driver mutation classification. In addition to the wavelet features the amino acid features that represent the physicochemical properties of proteins are also extracted from the numerically converted mutation sample. Next, combine the wavelet features and the amino acid index features and use as input to the classifier to identify the mutated genes. Then the identified mutated genes are given as input to the next classifier to classify the driver genes and the passenger genes from the mutated gene set.

A. Data collection

The driver and passenger genes can be collected from the GenBank [4]. In the proposed system 50 driver genes and 50 passenger genes are being collected.

B. Numerical representation of biological sequence

The biological sequences are to be encoded into suitable formats thereby making them amenable for being used as input to the wavelet transform functions as well as to the data analysis and data mining tools. This encoding is done by using some mapping schemes [5]. In the proposed system the extracted mutation samples are represented by numerical numbers by using complex number representation. A protein sequence consists of 20 amino acids. In the complex number representation of a protein sequence, each one of these 20 amino acids are represented by a complex number representation with the real and imaginary parts representing different properties of the amino acids. Here the mapping of protein sequence to complex numbers is obtained by using the Table I where the real part represents the hydrophobicity and the imaginary part represents the residue volume [10].

C. Feature extraction

Features are extracted from the protein sequences inorder to perform classification. In the proposed system two methods are used for feature extraction. The methods are as follows:

1. Wavelet feature extraction
2. AAIndex feature extraction

1. Wavelet feature extraction

The origin of the wavelets from harmonic analysis was demonstrated by a French mathematician, Jean Baptiste Joseph Fourier [8]. In wavelet feature extraction, wavelet transform funtions are applied on the mutation samples to extract the wavelet coefficient [3]. The wavelet transform is a tool for carving up functions or data into components of different frequency. Wavelet analysis has proved to be a powerful technique to analyse, classify or process data efficiently. The proposed system perform wavelet analysis using the wavelet toolbox in matlab.

TABLE I
THE COMPLEX REPRESENTATION OF 20 AMINO ACIDS

Amino Acid Name	Symbol	Complex Number Representation
Alanine	A	0.61 + 88.3i
Arginine	R	0.60 + 181.2i
Asparagine	N	0.06 + 125.1i
Aspartic	D	0.46 + 110.8i
Cysteine	C	1.07 + 112.4i

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

Glutamic	E	0.47 + 140.5i
Glutamine	Q	148.7i
Glycine	G	0.07 + 60.0i
Histidine	H	0.61 + 152.6i
Isoleucine	I	2.22 + 168.5i
Leucine	L	1.53 + 168.5i
Lysine	K	1.15 + 175.6i
Methionine	M	1.18 + 162.2i
Phenylalanine	F	2.02 + 189.0i
Proline	P	1.95 + 122.2i
Serine	S	0.05 + 88.7i
Theronine	T	0.05 + 118.2i
Triptophan	W	2.65 + 227.0i
Tyrosine	Y	1.88 + 193.0i
Valine	V	1.32 + 141.4i

Steps followed in wavelet analysis of the protein sequence are [1]:

- Map the protein sequence segment to complex numbers.
- Apply wavelet transform function to the numbers in scales 2 to 200 at a step of 2 to obtain a coefficient matrix.
- Mutate the protein sequence manually.
- Apply the wavelet transform function again on the manually mutated sequence segment to obtain a new coefficient matrix.

In the proposed system daubechies wavelet function is used for feature extraction due to its successful applications in biological sequence analysis. The obtained coefficients are a 100 by 100 matrix, where each row represents a coefficient sequence at a specific scale. A 100 dimensional feature vector is obtained by calculating averages of rows of the matrix.

2. AAIndex feature extraction

Amino acid index (AAIndex) features of the protein sequences are also extracted in addition to the wavelet features. The AAIndex features are extracted from the protein sequence in complex number representation in which the real parts and imaginary parts are from the hydrophobicity properties and residue volumes of the amino acids respectively [10]. The amino acid index features represent the physicochemical properties of the protein sequences. There are 544 physicochemical properties for each protein sequence [9].

After extracting the wavelet coefficients and the AAIndex features , the 100 dimensional wavelet features and the 544 dimensional AAIndex features of the protein sequences are combined together to form a 644 dimensional feature vector which could result in better performance of the classifier.

E. Identification of mutated gene

After combining the wavelet coefficients and the AAIndex features the 644 dimensional feature vector is given as input to the classifier to identify the mutated gene from the given dataset.

F. Identification of mutated driver gene

The mutated genes are of two types : the driver gene and the passenger gene. The mutated driver genes contribute to the cancer development whereas the mutated passenger gene do not contribute to cancer development. Therefore the key challenge of cancer genome analysis is to distinguish the driver mutations from the passenger mutations. Here the feature vector of identified mutated genes are given as input to the second classifier to identify the mutated driver gene which is the main goal of the proposed system.

In the proposed system two classifiers SVM and back propagation neural network are considered for classification purposes. Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data [11], [12]. Support Vector Machines use hypothesis space of a linear function in a high dimensional feature space and is trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. Classification in SVM is an example of supervised learning where the known labels help to indicate whether the system is performing in a right way or not. SVM classification involves identification as which are intimately connected to the known classes. The major strengths of SVM include relatively easy training, no local optimal, unlike in neural networks and it scales relatively well to high dimensional data and the trade-off between classifier complexity and error can be controlled explicitly.

BPNN (Back Propagation Neural Network) is considered as one of the simplest and the most general methods used for supervised training of multilayered neural network [11].Back propagation neural network works by approximating the non-linear relationship between the input and the output by adjusting the weight values internally. A supervised learning algorithm of back propagation is utilized to establish the neural network modeling. A normal back-propagation neural (BPN) model consists of an input layer, one or more hidden layers, and output layer. The learning rate and momentum are required to be defined by the user. The classification performance is affected by the selection and nodes of the hidden layers. The advantages of neural network include high tolerance of noise in data, ability to classify patterns on which they have not been trained, neural network is well suited for continuous valued inputs and outputs and neural network is more suited in cases with little knowledge of relation between attributes and classes.

III. EXPERIMENT ANALYSIS

Here the results of four experiments with different SVM and neural network combinations are analysed. In the first case neural network is used for both mutated gene identification and mutated driver gene identification. In the second case first neural network is used for mutated gene identification followed by using SVM classifier for

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

mutated driver gene identification. In the third case, The SVM classifier is used for identification of mutated gene followed by using neural network to identify the mutated passenger gene. In the fourth case, SVM classifier is used for both mutated gene identification and mutated driver gene identification. The sensitivity (Eq.1), specificity (Eq.2) and accuracy (Eq.3) are used as the performance analysis measures. Here, TP is the total number of true-positive instances, TN is the total number of true-negative instances, FP is the total number of false-positive instances, and FN is the total number of false-negative instances. The training dataset consists of 100 data sets. The test data set consists of mixture of mutated and unmutated driver and passenger genes. The test data used consist of total 40 genes. The performance analysis is tabulated in the following Table 2.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP}+\text{TN}} \quad (2)$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (3)$$

TABLE II
PERFORMANCE ANALYSIS

Machine learning approaches used	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
Neural-Neural	60	94.4	85
Neural-SVM	51.51	76.19	61.11
SVM-Neural	100	85.71	87.5
SVM-SVM	90	87.1	95

The above table shows the sensitivity, specificity and accuracy values obtained when four types of machine learning approaches are performed on the same data set. From the values in the table it could be inferred that the SVM-SVM combination gives the better result.

IV. CONCLUSION

Cancer represents one in every of the greatest causes that threatens the human life in the world nowadays. Identification of mutated cancer gene plays a vital role in individualizing the treatment of a cancer patient keep with his specific tumorigenic profile. This proposed system aims at identifying the mutated cancer gene for which four different combinations of SVM and back propagation neural network are considered. Experimental results shows that SVM-SVM combination outperforms alternative classifier combinations in identifying the mutated driver gene.

ACKNOWLEDGMENT

The authors would like to thank Mr. Vinodh P Vijayan for valuable discussions and comments that improved the quality and clarity of the manuscript.

REFERENCES

[1] Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, S.S. Iyengar, John S. Yordy, and Puneeth Iyengar Abstract—"Wavelet Analysis in Current Cancer Genome Research: A Survey", IEEE/ACM Transactions on computational biology and bioinformatics, Vol. 10, No. 6, November/December 2013

[2] R. Gupta, A. Mittal, K. Singh, P. Bajpai, and S. Prakash, "A Time Series Approach for Identification of Exons and Introns," Proc. 10 th Int'l Conf. Information Technology (ICIT '07), pp. 91-93, Dec. 2007.

[3] A.M. Sarhan, "Wavelet-Based Feature Extraction for Dna Microarray Classification," Artificial Intelligence Rev., vol. 39, no. 3, pp. 237 249, 2013

[4] D.A. Benson, I. Karsch-Mizrachi, K. Clark, D.J. Lipman, J. Ostell, and E.W. Sayers, "Genbank," Nucleic acid research, vol. 39, pp. D32-D37, 2011.

[5] Abo-Zahhad, S.M. Ahmed, and S.A. Abd-Elrahman, "Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques", Int'l J. Information Technology and Computer Science, vol. 4, no. 8, pp. 22-36, July 2012.

[6] M.R. Stratton, P.J. Campbell, and P.A. Futreal, "The Cancer Genome," Nature, vol. 458, no. 7239, pp. 719-724, 2009.

[7] M. Sifuzzaman, M.R. Islam, and M.Z. Ali, "Application of Wavelet Transform and Its Advantages Compared to Fourier Transform," J. Physical Sciences, vol. 13, pp. 121-134, 2009.

[8] C. Gargour, M. Gabrea, V. Ramachandran, and J.M. Lina, "A Short Introduction to Wavelets and Their Applications," IEEE Circuits and Systems Magazine, vol. 9, no. 2, pp. 57-68, Second Quarter 2009.

[9] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: Amino Acid Index Data, Progress Report 2008," Nucleic Acids Research, vol. 36, pp. D202-D205, 2008.

[10] C. Yin and S. Yau, "Numerical Representation of Dna Sequences Based on Genetic Code Context and Its Applications in Periodicity Analysis of Genomes," Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology, pp. 223-227, 2008.

[11] Ming-Chang Lee1 and Chang To,"Comparison of Support Vector and Back Propagation Neural Network in Evaluating the Enterprise Financial Distress", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.3, July 2010.

[12] Vikramaditya Jakkula,"Tutorial on Support Vector Machine (SVM)", School of EECS, Washington State University, Pullman 99164.