International Conference on Recent Innovations in Applied Science, Engineering & Technology
(IEI, Mumbai) The Institution of Engineers, Mumbai, India (AET-2018)
16th-17th June 2018 www.conferenceworld.in ISBN : 978-93-87793-30-2

Detection of Insertion Type Somatic Mutations in Cancer Genome using Principal Component Analysis

[1]Anuradha Choudhary, [2]Aman Sharma, [3]Smriti Dubey

[1]M.Tech Scholar, Dept. of Electronics & Communication Engineering, SATI, Vidisha(M.P), 464001(india)

[2]Assistant Professor, Dept. of Electronics & Communication Engineering, SATI, Vidisha(M.P), 464001(india)

[3]Assistant Professor, Dept. of Electronics & Communication Engineering, SATI, Vidisha(M.P), 464001(india)

ABSTRACT

Human genomic variability occurs at different scales, from single nucleotide polymorphisms (SNPs) to large DNA segments, i.e. Copy number variations (CNVs). These represent a significant part of our genetic heterogeneity and have also been associated with many diseases and disorders. These mutations play an important role in human disease, but these are undetectable in noisy genomics data. Therefore, robust methodologies are needed for their detection. Rapid advances in high throughput sequencing technology are making a strong impact on the detection of somatic mutations associated with cancer. Several computational methods have been developed to assist the experimental methods in identifying the somatic mutations from the sequenced genomic data. This work is oriented toward identifying the locations of insertion type somatic mutations using Principal Component Analysis (PCA, a mathematical technique for dimension reduction). Cancer genome data as obtained from COSMIC database has been subjected to PCA after its numerical mapping.

Keywords- *Cancer, insertion type mutations, somatic mutations, principal component analysis, PCA*

I. INTRODUCTION

Cancer disease represents one of the greatest medical causes of mortality. It is responsible for one in eight deaths in whole world [1]. A thorough understanding of cancer genome is required for better treatment purpose. Cancer occurs because of changes in cells. Human body is made up of billions of cells and inside every cell nucleus is present. Nucleus contains 23 pairs of chromosomes. Chromosome is a long strand of DNA comprising of four nucleotides Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). In human genome approximately 30,000 genes have been detected. Each gene is having some information for cell (i.e. genes tell the cell how to behave, when it will divide & when it will die). Genes have ability to control the functioning of

body & to control the cells division (known as mutation), by passing signals. If these signals are faulty, then cell division occurs in uncontrolled manner & multiple cells are generated resulting in tumor.

Mutation may be harmful, beneficial or having no effect depending upon the position of occurrence. When mutations occur in oncogenes, tumor suppressor genes, suicide genes & DNA repair genes, then there is a great possibility of occurrence of cancer. Genetic mutations associated with cancer fall in two categories. First are acquired mutations & second are germ-line mutations. Acquired mutations occur due to damage of genes during person's life. It is not passed from parent to child. Germ-line mutations are less common. They are passed directly from parent to child and in this situation the mutations can be found in every cell of person's body. In this work we have taken the problem of somatic mutations, which occurs in non-germ-line cell, and it comes under the category of acquired mutations.

Somatic mutations can be further classified as in table 1 –

Table 1

Types, Location, and Effect of various mutations

Type	Location	Effect
Insertion & Deletion	single nucleotide, or series of nucleotides	alters the translation frame & changes the downstream peptide sequence (results in premature termination of protein)
Substitution	substitution of nucleotides in gene sequence by other nucleotides	
Nonsense	termination of codon	nonfunctional protein product
Missense	Alteration of codon	altering the amino acid at that position only
Coding Silent	synonymous substitution	same amino acid as unnamed type codon
Intronic	mutation outside the coding domain	
Complex	mutation that involves multiple insertions, deletions & substitutions	

Huntington's disease and the fragile X syndrome are examples of insertion mutation wherein tri-nucleotide repeats are inserted into the DNA sequence leading to these diseases.

Several somatic mutation (insertions) detection methods have been proposed & applied to genomic data with variable success. They include Mu-Tect (Bayesian Classifier) [4], Hidden Markov Model [5] and support vector machine [6]. These methods have significant computational complexity & take long time to process the data.

Our aim is to detect insertions accurately and in less time. To analyze the cancer genome with somatic mutations, we need database of cancer genome of somatic mutations. Several institutions have been sequencing the genomes of cancer patients. They include - The Cancer Genome Atlas (TCGA), The International Cancer Genome Consortium (ICGC) & The Catalogue of Somatic Mutations in Cancer (COSMIC) [7]. TCGA & ICGC provide information about cancer genome that is related to point mutations, methylation, Copy Number

Variations (CNVs) & structural variations, whereas COSMIC provides information about somatic mutations associated with cancer types. We have used COSMIC online resource for somatic mutation database. It provides somatic mutation data in terms of biological sequences (DNA & Amino Acid).

To analyze the somatic mutations data using computational methods, first we need to encode biological sequences in suitable format. This is done by assigning a numerical value to each symbol that is present in biological sequences. Next we applied a computational efficient algorithm, Principal Component Analysis (PCA), for detection of somatic mutations. PCA algorithm is based on statistical approach. It detected the somatic mutations (insertions) by identifying patterns in the datasets, in terms of similarities & differences between the data points [8] with very high, near perfect accuracy.

II. COSMIC DATABASE

COSMIC provides information about somatic mutations that are associated with cancer types. COSMIC contains curated somatic mutation data, launched in 2004, with few genes data. Now it contains approximately 2.1 million unique somatic variants detected in approx 1 million tumor samples. Currently COSMIC contains 547 genes which have mutations [7]. COSMIC website link is **http://www.sanger.ac.uk/cosmic/**. At COSMIC home page we can search for a gene, cancer type and mutation etc. It graphically summarizes the somatic mutations (substitutions, complex type mutations, insertions, deletions, CNV gain & CNV loss, overexpression & underexpression and hyper methylation & hypo methylation). A sample of 'gene view' page is shown in figure 1. The highlighted circle shows the insertion mutation & second highlighted circle represents the deletion mutation in FOXD3 gene.

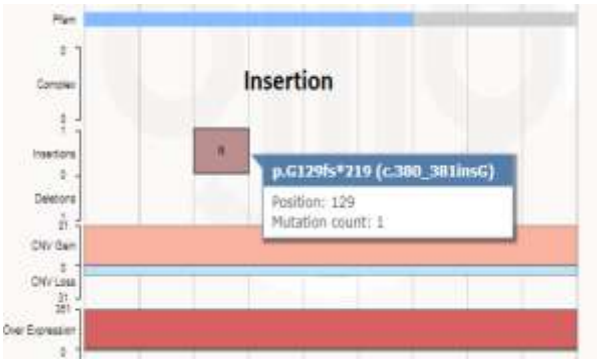


Figure 1: COSMIC 'GeneView' page

In insertion mutation sample X-axis describes the full length of the gene, Y-axis describes the mutation count, In right side of 'gene view' page, filtering options are available. By using these filter options, we can filter the sequences based on parameters like type of mutation, source of mutation, and some graphical properties of geneview page..

Information available on overview page contains:-

i. 'Gene name & its ID'.
ii. 'DNA sequence', and 'protein sequence'.
iii. 'OMIM' provides descriptive information about gene functioning.
iv. 'Transcript' & 'HGNC' links COSMIC with 'ensemble' & 'NCBI' respectively.

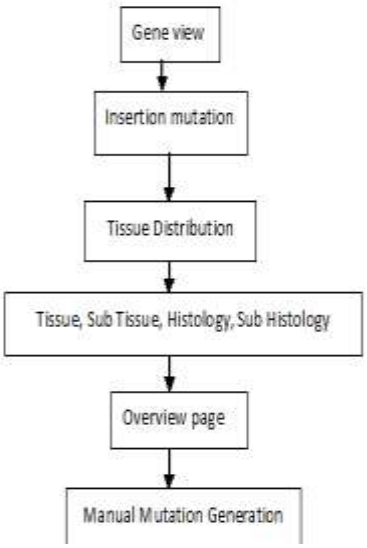


Figure 2: The block representation of data accession from COSMIC

Steps required to get a cancerous sample from COSMIC database

Step 1: Go to 'gene view' page, zoom in to insertion type mutation.

Step 2: Click on any insertion block, a new page open.

Step 3: Click on 'tissue distribution', choose tissue name.

Step 4: We get a new page, this page contains filter that we used for filtering between tissue, subtissue, histology & subhistology. Choose filter section carefully & click on 'go'.

Step 5: Select any gene name which arise after filtering & get DNA sequence from 'over view' page, this DNA sequence is normal gene sequence (take it as a reference sequence).

Step 6: Take the same sequence in separate file and create insertion mutations manually according to 'gene view' histogram page (take it as a cancerous sample).

III. NUMERICAL REPRESENTATIONS OF DNA SEQUENCES

Many approaches are used to encode the DNA sequences as numerical sequences. These approaches are Voss mapping, Z-curve mapping, tetrahedron representation, complex number representation, integer representation & physicochemical property based representation. In our application, we used three methods for numerical representation of DNA sequences. These are-

Table 2

Numerical Mappings

Representation	Type/Condition	Values
Complex Number	G =C* T = A*	A = 1+j T = 1-j C = -1-j G = -1+j
Integer	Sequential	A = 1 T = 2 C = 3 G = 4
Physicochemical Property Based **or** **Electron Ion Interaction Potential (EIIP)**	DNA molecules biochemical properties	A = 0.1260 T = 0.1335 C = 0.1340 G = 0.0806

IV. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a statistical approach, it converts a set of correlated variables, into a set of linearly uncorrelated variables, called Principal Components (PCs), by using orthogonal transform. The number of PCs that we get is less than or equal to original variables. The first PC contains the largest possible variance & other PCs contain the remaning variance. All resulting vectors are uncorrelated and orthogonal to each other.

PCA is used for data analysis. It uses eigenvector decomposition of covariance data matrix, usally after calculating mean & substracting mean from the datasets. The result of PCA shows in terms of feature vector; It is a multiplication of dataset with eigen vectors. Steps for calculation of PCA are-

Step 1: Pre-treatment of data matrix i.e. scaling-

For scaling the data matrix mean centering is used as a scaling method.

i. Mean value of each variable is calculated.
ii. Mean is subtracted from the data.

Step 2: Calculation of covariance matrix-

If the data matrix 'A' has 'm' rows & 'n' columns, covariance matrix S is

$$S = cov(A) = \left(\frac{1}{m-1}\right) * A' * A \quad (1)$$

Where A = column vectors of A1 & A2 = [A1 A2],

A1 = first variable, A2 = second variable

A'= transpose of matrix A.

Let m = 10 & n = 2, then the size of covariance matrix is 2X2.

By the definition of covariance matrix

$$S=\begin{bmatrix} var(A1) & cov(A1,A2) \\ cov(A1,A2) & var(A2) \end{bmatrix} \quad (2)$$

Variance: measure of the spread of data in a given data set

$$var(A)=\sum_{i=1}^{n}[(Ai-\bar{A})*(Ai-\bar{A})]/(n-1) \quad (3)$$

Covariance: It is a multi-dimensional concept & measure of the spread of data between dimensions.

$$cov(A1,A2)=\sum_{i=1}^{n}[(A1i-\bar{A1})*(A2i-\bar{A2})]/(n-1) \quad (4)$$

Where n is the sample number, Ā, Ā1 & Ā2 are the means of the data set A, A1& A2 respectively.

Step 3: Calculation of eigen values & eigen vector of the covariance matrix (S) using the characteristic equation.

$|S-\lambda*I| = 0$, for eigen value calculation (5)

$(S-\lambda*I)*(PC) = 0$, for eigen vector calculation (6)

Where λ = eigen-values and PC is the eigen vectors of A

associated with eigen values λ & $I=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ = Identity matrix

Step 4: Calculation of feature vector

As = A*PC; (7)

Where PC= eigen vector corresponding to eigen values $PC=\begin{bmatrix} eig\ 1 \\ eig\ 2 \end{bmatrix}$

V. IDENTIFICATION OF SOMATIC MUTATIONS USING PCA

In our application, we have big set of data and we want to analyse this set of data in terms of relationship between the individual points in data set. With the help of PCA we can identify patterns in data & express that data by highlighting similarities and differences between data points. We applied PCA for comparative analysis between normal (reference) & mutated (cancerous) sequence. The block representation of PCA model for proposed work is shown in figure 3.

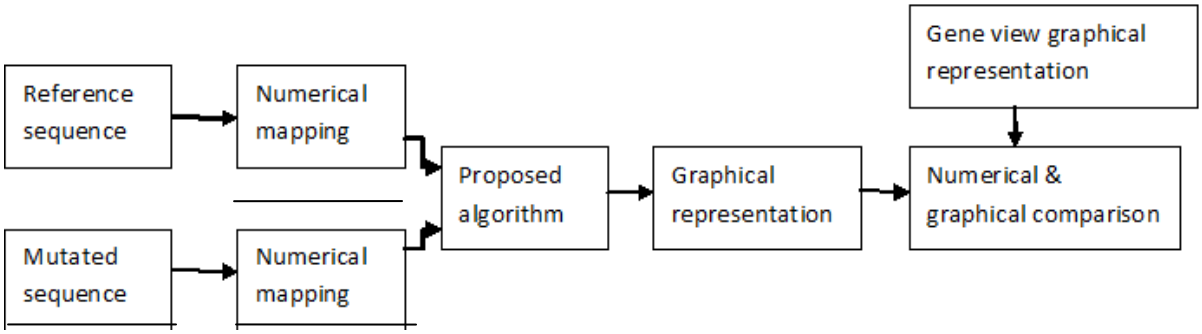


Figure 3: PCA model for comparative study

Algorithm for comparative analysis is illustrated in following steps-

Step 1: acquire non mutated sequence from COSMIC database.

Step 2: manually make insertion in acquired data sequence as per gene view page of COSMIC.

Step 3: perform numerical mapping on both sequence.

Step4: process the signal through proposed algorithm**.**

Step 5: Generate Zero Mean Sequence for both non-mutated and mutated data.

Step 6: make both sequence compatible for PCA Algorithm.

Step 7: Find out correlation, and covariance matrix between both sequences.

Step 8: Calculate the eigen-values and eigen vector matrix of the covariance matrix.

Step 9: Create a composite matrix of both data sequence and multiply with PC (eigen vector) Matrix, obtained in last step. The resultant matrix contains 2 principal components.

Step10: compare graphical output of simulation to gene view chart.

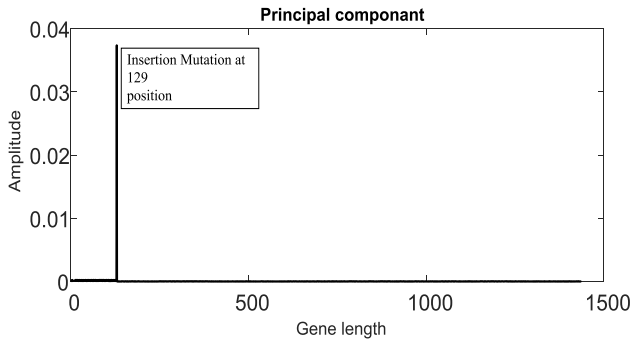


Figure 4: principal component of FOXD3 gene

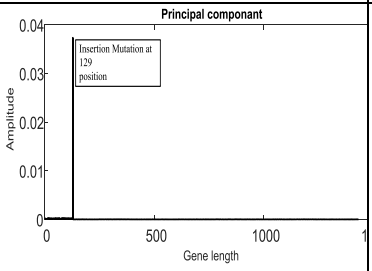
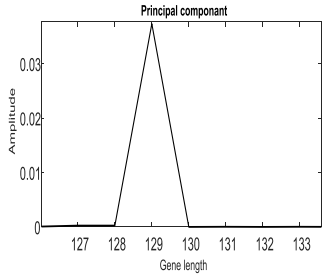
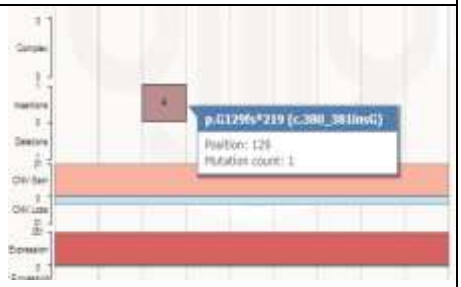
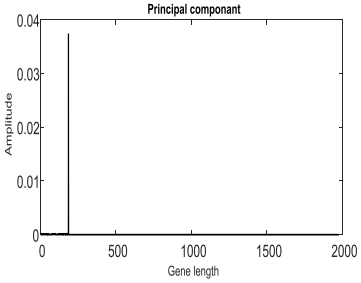
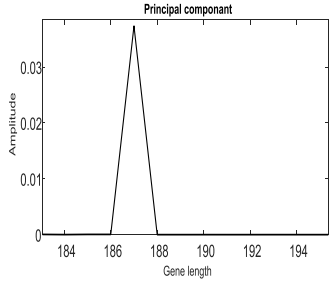
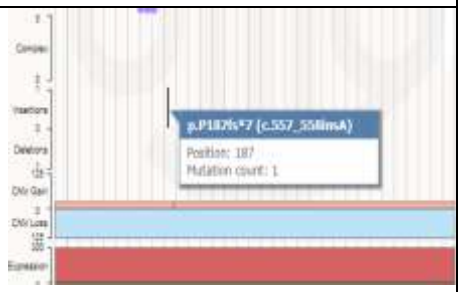
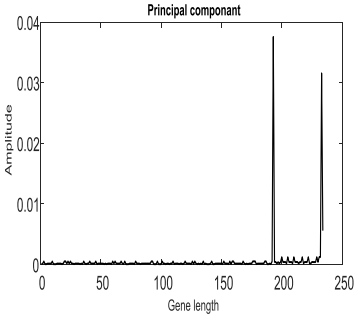
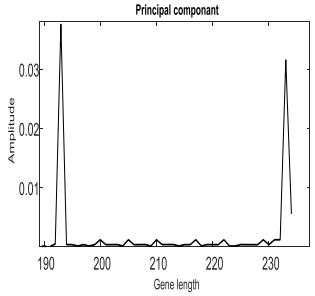
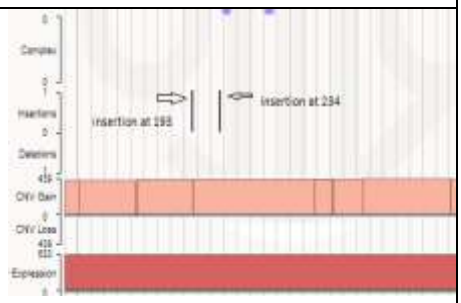
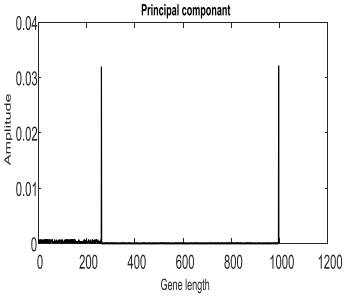
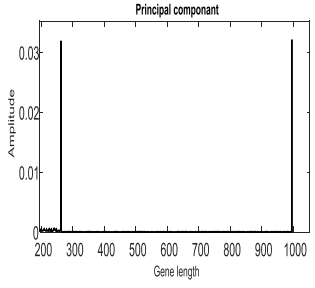
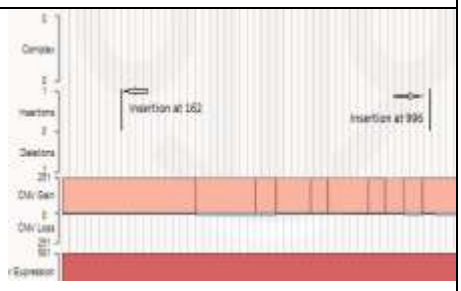
We applied PCA to FOXD3 gene of length 1437 (DNA). Figure 4 show the PCs of FOXD3 genes. It provides the exact position of insertion somatic mutations. In figure 4 spike is generated at position '129', this spike shows the occurrence of insertion mutation at this position.

VI. SIMULATIONS AND EXPERIMENTAL RESULTS

The algorithm is tested on several databases. Which are given in table 3. In table 3, columns 2 & 3, represent the simulation results, with the name of sample data in column 1. Column 4the gene view representation as per COSMIC database.

Table 3

Simulation results

Gene name	Principle Component	Zoomed Principle Component	Zoomed COSMIC 'Geneview' Histogram Page
FOXD 3 (1436)			
BNK (1980)			
Tp63 (2045)			
ERC1 (3353)			

VII. CONCLUSION AND FUTURE WORK

We applied PCA, a computationally efficient methodology to detect insertion type somatic mutations present in cancer. The computational cost of this methodology is very small. The results obtained have high accuracy & are not significantly affected by changes in mapping scheme used. The method works well for small number of insertions. The work can be extended for detecting even large number of insertion as well as deletion type of mutations. In summary, PCA is has proved to be promising approach for detection of insertion type mutation, in human genome.

REFERENCES

[1] Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen, S.S. Lyengar, John S.Yordy and Puneeth Lyengar, "Wavelet Analyses in Current Cancer Genome Research: A Survey", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 6, Nov/Dec 2013.

[2] Lodish H, Berk A, Zipursky SL, "Mutations: Types and Causes", 4th edition. New York: W. H. Freeman; 2000.

[3] 1000 Genomes project C, Abecasis GR, Auton A, Brooks LD, Depristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA; "An integrated map of genetic variation from 1092 human genomes", Nature 2012, 491 (7422): 56-65.

[4] Kristian Cibulskis, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel1, Matthew Meyerson1, Eric S. Lander and Gad Getz; "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples", Nat Biotechnol. 2013 March ; 31(3): 213–219.

[5] Hashem A. Shihab, Julian Gough, David N. Cooper, Ian N. M. Day and Tom R. Gaunt; "Predicting the functional consequences of cancer-associated amino acid substitutions", Bioinformatics vol. 29 no. 12 2013, pages 1504–1510.

[6] Fan Zhang, Tao Lio, Mu Wang, Renee Drabier, "Dual-function biomarkers for detection of breast cancer and its cancer types: invasive versus non invasive", 2013.

[7] Ruitian, Malay K Basu, Emidio Capriotti, "Computational methods and resources for the interpretation of genomic variants in cancer", Tian et al. BMC Genomics 2015, 16 (suppl 8):S7.

[8] Lindsay I Smith, "A tutorial on principal component analysis", February 2002.

[9] C. Cattani, "Complex representation of DNA sequences", Comm.. in Computer and Information Science, vol.13, pp. 528-537, 2008.

[10] M. Abo-Zahhad, S.M. Ahmad, and S.A. Abd-Elrahman, "Genomic analysis and classification of exon and intron sequences using DNA numeric mapping techniques", int'l J. information technology and computer science, vol. 4, no. 8, pp. 22-36, July 2012.

[11] A.S. Nair and S.P. Sreenadhan, "A coding measure scheme employing Electron Ion Interaction Potential (EIIP)", bioinformatics, vol. 1, no. 6, pp. 197-202, 2006.