Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 10 (2013) 37 – 44

Procedia Technology

International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013

Prediction of Prostate Cancer cells Based on Principal Component Analysis Technique

A.Ghosh[a] and S.Barman[b]*

[a,b] Institute of Radio Physics & Electronics, University of Calcutta, 92,APC Road, Kolkata-700009, India

Abstract

Amino acids, the essential building blocks of life are important to study the genetic diseases, modelling of protein structure and also in drug designing. A PCA model along with signal processing technique is used here for differentiating the prostate cancer cells from normal prostate cells. The amino acid sequence of cells is taken as input sample for the PCA technique. The model is successfully tested on 8 normal and 8 cancerous Homo sapiens prostate cells.

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering

Keywords: Genomics; DFT; PCA; Cancer; Disease Diagnosis; Amino acid; DNA.

1. Introduction

Nowadays researchers from various cross fields have concentrated their research in genomic study which is mainly information extraction and data analysis. Generally, two broadly classified research areas of genomics are DNA sequence analysis and disease diagnosis [1]. DNA sequence analysis is a well developed research area used to reveal hidden features present in protein coding regions whereas diagnosis of disease is used to find out abnormalities present in DNA sequence because almost all the genetic diseases, such as Parkinson, Alzheimer, Cancer and development of abnormalities are characterized by the presence of genetic variations. Due to occupying the leading position for causing worldwide death of men and women over few decades, Cancer is able to create great queries among the researchers in the recent time and significant discoveries has made to provide better

* Corresponding Author, Tel.: +91-33-23509115/9116; 9434238666.
E-mail address: barmanmandal@gmail.com.

Selection and peer-review under responsibility of the University of Kalyani, Department of Computer Science & Engineering
doi:10.1016/j.protcy.2013.12.334

understanding of genetic basis of cancer. Amino acids of DNA which are essential to form antibodies to combat bacteria and viruses; they are the part of the enzyme and hormonal system play a significant role in cancer research. Controlled Amino Acid Therapy (CAAT) is an efficient medical treatment used to impair the development of cancer cell [2]. It has been understood from medical research reports, out of the all cancers, prostate cancer is most common among adult men all over the world and Prostate Screening Antigen (PSA) test is used for the patients to measure the protein level produced by prostate gland [3].

It is well known that twenty amino acids of a DNA sequence are responsible for the formation of protein [4]. In this present scenario, the study of amino acids and proteins are important to understand the genetic basis of cancer. The authors in this paper have made a comparative study between prostate cancer and normal prostate Homo sapiens cells based on their amino acids sequence databases. Principal Component Analysis (PCA) , a powerful statistical technique which reduces large input vectors without much loss of information is chosen here for comparative study, as the Homo Sapiens databases consist of long string of amino acids . A Principal Component Analysis (PCA) model has been designed, followed by Discrete Fourier transform of mapped amino acid sequence for the analysis. The method is successfully tested on Homo Sapiens prostate databases available in public domain [18]. The paper is organized into number of sections: Introduction, Brief Background, Methodology, Results and Discussion and Conclusions.

2. Brief Background

The story of life starts from cell and DNA is important to all cells and organisms because an organism could do nothing without DNA [5]. A mutation is the permanent change in the DNA, can arise spontaneously with apparent cause and can lead to cell death, cell alteration, cell formation or in some cases development of cancer cells [6].

A DNA, an informational molecule encoding the genetic information, encoded as a sequence of nucleotide bases: guanine (G), adenine (A), thymine (T) and cytosine (C) and divided into two regions: Genes and Inter-genic spaces. A gene also can be dividing into two sub-regions named Exon (coding region) and Intron (non-coding region) (Fig.1). Exons of a DNA sequence are the most information bearing part because only the exons take part in protein coding while the introns are spliced off during protein synthesis. In exon region the bases are divide into three adjacent bases called codon which translated into amino acid and 64 possible codons generate 20 amino acids [Table.1], are responsible for forming proteins, deficiencies of these may leads to different types of genetic abnormalities[4].

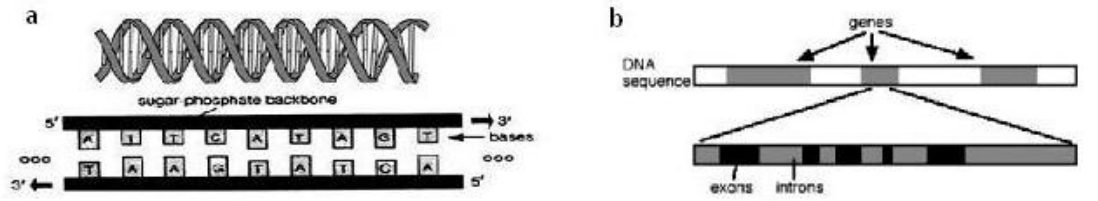

Fig.1. (a) DNA sequence; (b) A DNA sequence showing genes and intergenic regions.

The study of amino acids in a protein sample presents a new horizon for cancer classification and prediction. Some amino acids are essential for the growth of tumor cells and restricting them or inhibiting them may be beneficial for curing cancer patients[4]. Digital Signal Processing (DSP) can be effectively used in genomics study with great accuracy as Genomic or Proteomic information is digital in nature. In this present paper, Discrete Fourier Transform (DFT) is used for spectrum estimation of cells and a PCA model is designed for comparative analysis of normal and cancer cells.

3. Methodology

The application of Fourier transform techniques has found to be very useful for both DNA and amino acids sequences. DSP techniques is applicable only to numerical data but the amino acids consist of alphabets A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y [7,8,9]. Hence, a mapping technique is required to convert the

alphabetic sequence into numerical sequence before applying DSP. Here a well known single sequence electron ion interaction pseudo potential (EIIP) mapping rule [10], based on the distribution of free electron's energy along DNA sequence is used for the conversion. The EIIP values of 20 amino acids are displayed in [Table.1].

Suppose, an amino acid chain of a cell is: *x[n] = [M P I G S K E R P T F D];*
After EIIP mapping using Table 1:
x[n] = [0.0373 0.0198 0.0000 0.0050 0.0829 0.0371 0.0058 0.0959 0.0198 0.0941 0.0946 0.1263];

Table 1. List of 20 amino acids with codons and EIIP values.

Sl. No	Abbreviation		Amino Acid	Codons	EIIP Value
1	A	Ala	Alanine	GCA,GCC,GCG,GCT	0.0373
2	C	Cys	Cystein (has S)	TGC,TGT	0.0829
3	D	Asp	Aspartic Acid	GAC,GAT	0.1263
4	E	Glu	Glutamic Acid	GAA,GAG	0.0058
5	F	Phe	Phenylalanine	TTC,TTT	0.0946
6	G	Gly	Glycine	GGA,GGC.GGG,GGT	0.0050
7	H	His	Histidine	CAC,CAT	0.0242
8	I	Ile	Isoleucine	ATA,ATC,ATT	0.0000
9	K	Lys	Lysine	AAA,AAG	0.0371
10	L	Leu	Leucine	TTA,TTG,CTA,CTC,CTG,CTT	0.0000
11	M	Met	Methionine	ATG	0.0823
12	N	Asn	Asparagine	AAC,AAT	0.0036
13	P	Pro	Proline	CCA,CCC,CCG,CCT	0.0198
14	Q	Gln	Glutamine	CAA,CAG	0.0761
15	R	Arg	Arginine	AGA,AGG,CGA,CGC,CGG,CGT	0.0959
16	S	Ser	Serine	AGC,AGT,TCA,TCC,TCG,TCT	0.0829
17	T	Thr	Threonine	ACA,ACC,ACG,ACT	0.0941
18	V	Val	Valine	GTA,GTC,GTG,GTT	0.0057
19	W	Trp	Tryptophan	TGG	0.0548
20	Y	Tyr	Tyrosine	TAC,TAT	0.0516

After this conversion, spectral estimation of EIIP mapped sequence is obtained by using Discrete Fourier Transform (DFT) technique. PCA, a standard tool for multivariate data analysis [11-13] is used to generate maximum uncorrelated components called Principal Components (PC) between the prostate cancer and normal prostate cells. The block representation of the PCA model for comparative analysis is depicted in Fig.2. and the algorithm for comparative analysis is illustrated in the following steps:

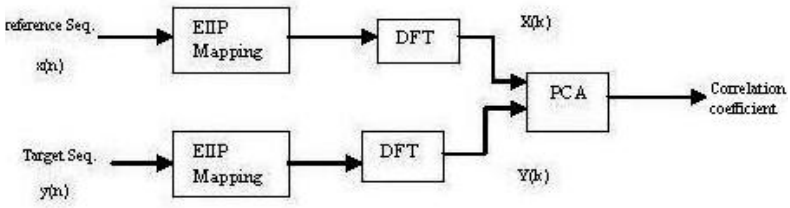

Fig.2. PCA model for comparative study.

Step 1: Convert the scanned amino acid sequence into numerical form using EIIP values and compute DFT of the converted sequence using these equations:
Let, *x(n)* = a normal prostate cell (taken as reference)
y(n) = a normal or cancer prostate cell (taken as target)
The DFT of the EIIP sequence is given by:

$$Xs[k] = \sum_{n} x[n] e^{-j2\pi nk/N} \quad (1)$$

$$X[k] = |Xs[k]| \quad (2)$$

$X_S[k]$ = DFT of reference sequence; $Y_S[k]$ = DFT of target sequence;

$X[k]$ = Amplitude spectrum of $X_S[k]$;

Similarly, $Y[k]$ = Amplitude spectrum of $Y_S[k]$;

N = Length of DNA sequence; $k = 0, 1, 2, \ldots, N-1$; $n = 0, 1, 2, \ldots, N-1$

Step 2: Convert $X[k]$ and $Y[k]$ into column matrix and obtain the mean value of $X[k]$ and $Y[k]$:

$$X = \overline{X[k]} \quad ; \quad Y = \overline{Y[k]} \quad (3)$$

Step 3: For PCA to work properly, subtract the mean from each data dimensions. The mean subtracted is the average across each dimension. This produces a data set whose mean is zero.

$$X_{new} = X[k] - X \quad ; \quad Y_{new} = Y[k] - Y \quad (4)$$

Step 4: To find out correlation, compute covariance matrix between reference and target sequence:

Covariance matrix = $A = \mathrm{cov}(X_{new}, Y_{new})$ (5)

$$= \begin{bmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) \end{bmatrix}$$

Step 5: Calculate the eigen-values and eigen vector matrix of the covariance matrix *(A)* using the characteristic equation:

$$(A - \lambda I)x = 0 \quad (6)$$

Where, λ = eigen-values and x is the eigen vector associated with eigen value λ

and $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ = Identity matrix.

Step 6: Sort λ in ascending order for eigen decomposition. Calculate the eigen vector with highest eigen value which is called the principal component of the data sets and formed a feature vector(matrix of vector) based on this eigen values:

$$Feature\ Vector = M = \begin{bmatrix} eig1 \\ eig2 \end{bmatrix} \quad (7)$$

Step 7: To differentiate cancer and normal cells using feature vectors, a new Final Data matrix *(1 × k)* is created and plot the Final Data with respect to length of the sequence:

$$Final\ Data\ (k) = M^T * [X_{new}, Y_{new}] \quad (8)$$

The algorithm is tested on several databases as shown in [Table 2].

4. Result and Discussion

In this present PCA model, correlation coefficients have been calculated between (8 normal × 8 cancer) prostate cells and (8 normal × 8 normal) prostate cells, are shown in [Table 3] and [Table 4]. PCA analysis shows all positive correlation coefficients, when normal prostate databases compared with other normal prostate databases. Therefore, the correlation between them is high or they are similar cells. Whereas when normal prostate cell compared with the cancerous prostate cells, out of 64 combinations, 63 shows negative value, means they are orthogonal or uncorrelated samples. One cross-correlation shows positive value that means error is only 0.8% in this present comparative study. The model proposed in this article will be suitable for preliminary prediction of cancer and normal cells and the algorithm is tested in MATLAB environment. The correlation coefficients obtained from the

PCA model for normal vs. cancer and normal vs. normal cells are depicted in Table 3 and Table 4 respectively. Due to space constraint, only some of the simulated plots for the comparative analysis are displayed in Fig.3 and Fig.4. Researchers [14-17] are used mostly PCA tool for classification of protein structures or to detect different features of breast cancers. The application of PCA technique for differentiating prostate cancer cells from normal prostate cells is new approach in the recent research.

Table 2. Normal and Cancerous prostate HOMO SAPIENS Genes [18].

Type of cell	Accession Number
Normal Prostate Cell	AF224278.1, AF331165.1, AF462605.1, M15885.1, M24543.1, M24902.1, NM_005984.3, NM_007003.2
Prostate Cancer Cell	AAQ08976.1, AF304370.1, AF338650.1, AF455138.1, AY008445.1, FJ649644.1, NP001035756.1, NP001231873.1

Table 3. Correlation between prostate normal vs. prostate cancerous cells.

			C	A	N	C	E	R	
	Accession no.	AAQ08976.1	AF304370.1	AF338650.1	AF455138.1	AY008445.1	FJ649644.1	NP001035756.1	NP001231873.1
	AF224278.1	-26.0477	-48.4853	-161.8877	-27.7271	-27.8535	-29.0009	-26.0555	-27.8534
N	AF331165.1	-27.7015	-51.6589	-170.885	-29.5669	-28.6867	-30.8476	-27.7092	-29.6863
O	AF462605.1	-26.4612	-49.5525	-165.2082	-28.22	-28.3435	-29.5763	-26.469	-28.3428
R	M15885.1	-27.1346	-50.3156	-167.4919	-28.6605	-28.7839	-29.8534	-27.1425	-28.7852
M	M24543.1	-25.3429	-47.4	-158.5457	-27.0901	-27.2194	-28.4354	-25.3509	-27.2198
A	M24902.1	-26.5532	-49.3816	-163.3887	-28.2532	-28.3769	-29.5725	-26.5608	-28.3768
L	NM005984.3	-26.9842	-50.3914	-166.5609	-28.832	-28.9521	-30.0249	-26.9919	-28.9508
	NM007003.2	-24.8987	-47.0019	-156.3431	-26.5466	-26.6784	27.9601	-24.9068	-26.6785

Table 4. Correlation between prostate normal vs. prostate normal cells.

			N	O	R	M	A	L	
	Accession no.	AF224278.1	AF331165.1	AF462605.1	M15885.1	M24543.1	M24902.1	NM005984.3	NM007003.2
	AF224278.1	*	28.8641	15.1526	15.2976	14.9434	23.0799	17.6268	15.3736
N	AF331165.1	28.8641	*	7.8845	7.1685	16.0727	24.6303	18.96	6.4127
O	AF462605.1	15.1526	7.8845	*	7.4782	15.3553	23.6554	17.8473	7.801
R	M15885.1	15.2976	7.1685	7.4782	*	15.4648	23.8892	18.4618	7.143
M	M24543.1	14.9434	16.0727	15.3553	15.4648	*	22.6466	17.0991	15.5635
A	M24902.1	23.0799	24.6303	23.6554	23.8892	22.6466	*	22.2841	24.0479
L	NM005984.3	17.6268	18.96	17.8473	18.4618	17.0991	22.2841	*	18.399
	NM007003.2	15.3736	6.4127	7.801	7.143	15.5635	24.0479	18.399	*

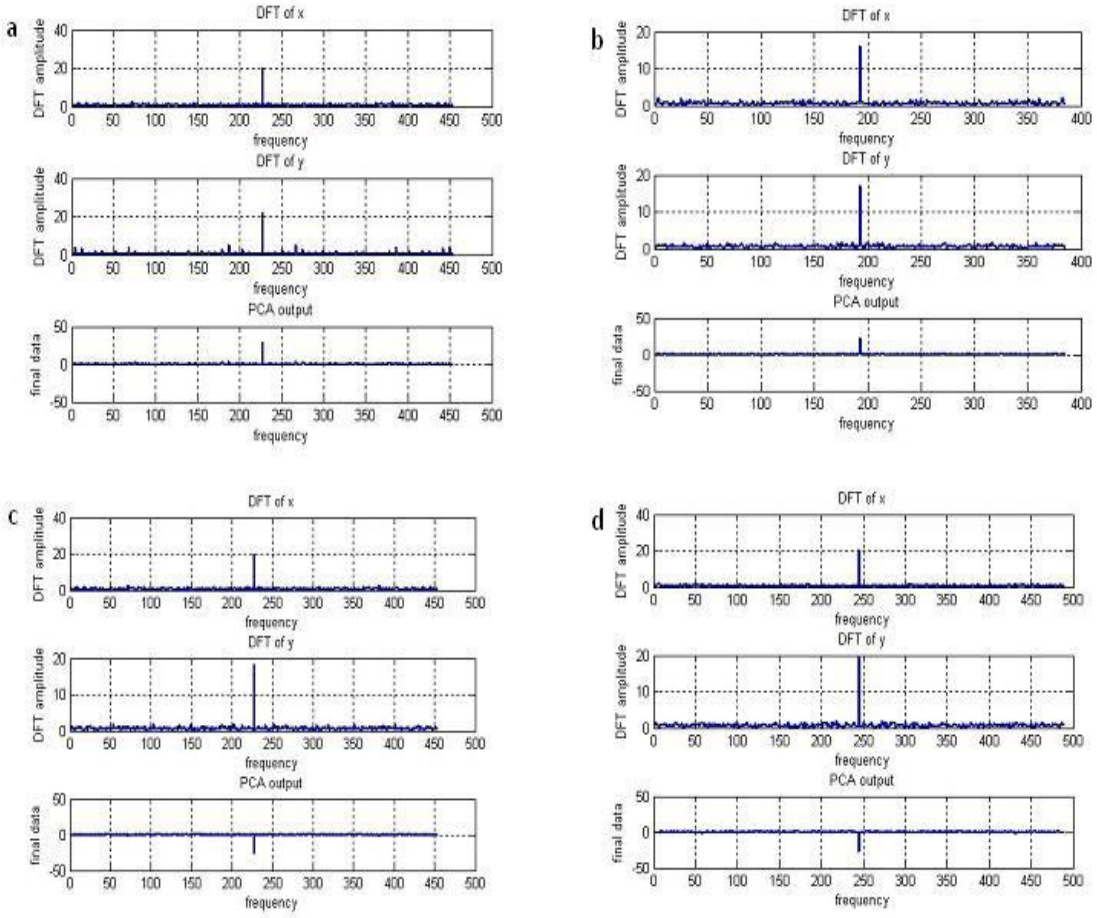

Fig.3. DFT and PCA of (a) AF224278.1 (normal) vs. AF331165.1 (normal); (b) M24543.1 (normal) vs. M24902.1(normal); (c) AF224278.1 (normal) vs. AAQ08976.1(cancer); (d) M24543.1(normal) vs. AY008445.1(cancer).

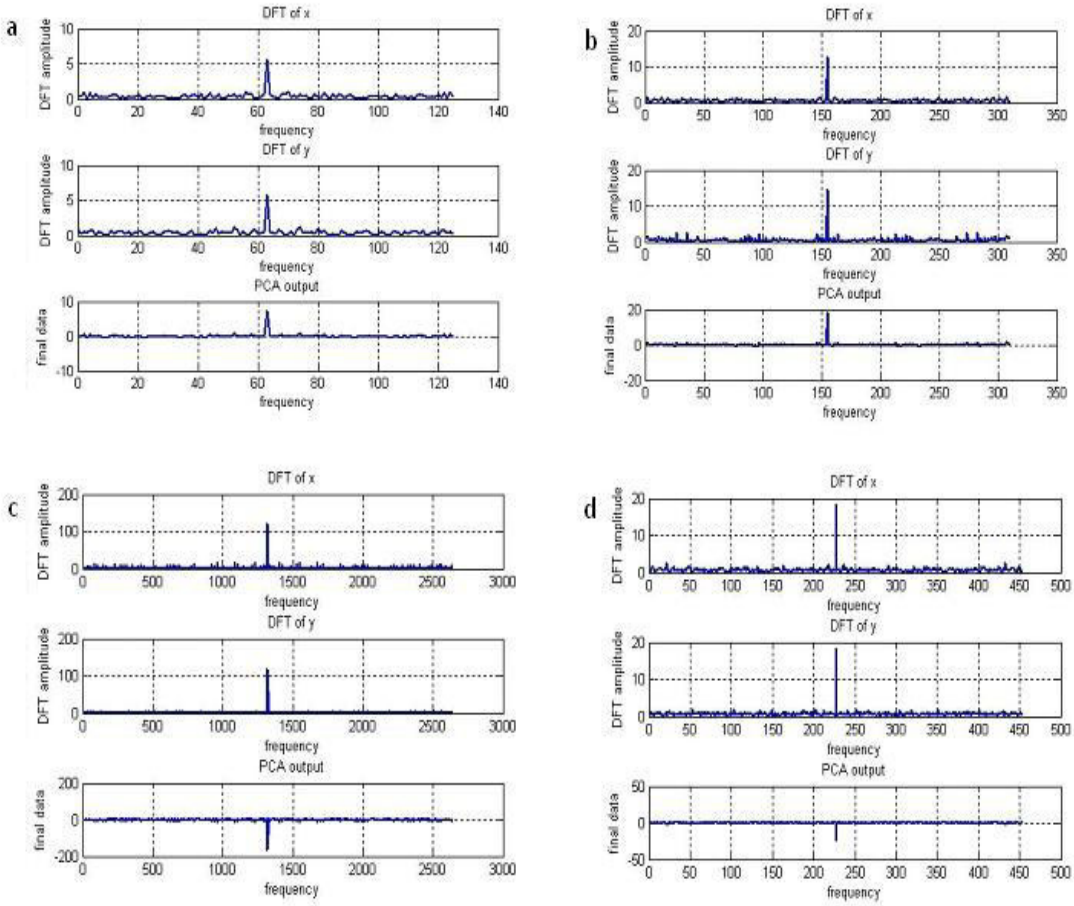

Fig. 4. DFT and PCA of (a) AF462605.1(normal) vs. M15885.1(normal); (b) NM_005984.3(normal) vs. NM007003.2(normal); (c) AF462605.1 (normal) vs. AF338650.1(cancer); (d) NM_005984.3 (normal) vs. NP001035756.1(cancer).

5. Conclusion

The PCA model in this article is tested successfully on prostate cells only. Further cells will be taken as sample databases (e.g. breast, skin, pancreas, lung) in future for prediction of cancers. Canonical Correlation Analysis (CCA) may also be used for future study . But in the present study, we restrict analysis only in PCA because our main target is to separate out cancer associated prostate cells from the normal prostate cells. PCA works better in this perspective as analysis is based on maximum variance between the samples whereas CCA analysis is based on maximum correlation between the datasets. In this present era genomics research is not only limited to wet smelly laboratory. Soft databases are available in public domain [18], scientist from different fields may use their expertise for analysis and diagnosis of genetic diseases.

References

[1] Vaidyanathan PP. Genomics and Proteomics: a signal processor's tour. IEEE circuit and system magazine 2004; 4: 315-319.
[2] A P. John Institute for Cancer Research paper on Controlled Amino Acid Therapy (CAAT) works. Available : http/www.apjohncancerinstitute.org.
[3] www.cancer.gov/cancertopics/types/prostate.
[4] Barman (Mandal) S, Saha S, Mondal A, Roy M.Signal Processing Techniques for the Analysis of Human Genome Associated with Cancer Cells. 2nd Annual international Conf. IEMCON, 11.
[5] Alberts B, Bray D, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Essential cell biology. Garland Publishing Inc.2nd ed. New York; 1998.
[6] Qie P, Wang ZJ, Ray Lie KJ. Genomic Processing for Cancer Classification and Prediction. IEEE Signal Processing Magazine 2007;100: 100-110.
[7] Anastassiou D. Genomic Signal Processing. IEEE Signal Processing Magazine 2001.p. 8-20.
[8] Ali AF, Shawky DM. A Novel Approach for Protein Classification Using Fourier Transform. World Academy of Science, Engineering and Technology 2010;44: 247-251.
[9] Khare A, Nigam A, Saxena M. Identification of DNA Sequences By Signal Processing Tools in Protein-Coding Regions. Search & Research 2011;II:2:44-49.
[10] Roy M, Barman (Mandal) S. Spectral Analysis of Genomic Data by Recursive Winer-Khinchine Theorem using Various Mapping Techniques. Proc. of International Conference on Nanotechnology and Biosensors 2011.
[11] Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. Briefings in bioinformatics 2011; 12: 6: 714-722.
[12] Shlens J. A tutorial on Principal Component Analysis.Version 1. 25 March 2003.
[13] Smith LI. A tutorial on Principal Components Analysis. February 26, 2002.
[14] Melo JCB, Cavalcanti GDC, Guimaraes KS. PCA Feature Extraction for Protein Structure Prediction. IEEE Xplore 2003;2952-2957.
[15] Vipsita S, Shee BK, Rath SK. Protein Superfamily Classification using Kernel Principal Component Analysis and Probabilistic Neural Networks. IEEE Xplore 2011.
[16] Christoyianni I, Koutras A, Dermatas E, Kokkinakis G. Computer Aided Diagnosis of Breast Cancer in Digitized Mammograms. Elsevier Science Ltd.,Computersied Medical Imaging and Graphics 2002; 26: 309-319.
[17] Hasan H, Tahir N. Feature Selection of Breast Cancer Based on Principal Component Analysis. Int. Colloquium on Signal Processing and its Applications 2010; 242-245.
[18] National Centre for Biotechnology Information (NCBI). [Online]. Available: http://www.ncbi.nlm.nih.